Anaphora Resolution for Arabic Machine Translation: A Case Study of *Nafs*

Wafya Hamouda

Submitted for the degree of Doctor of Philosophy

School of English Litearture, Language and Linguistics / Faculty of Humanities and Social Sciences (HASS)

Newcastle University

August 2014

Supervisor**:**

Dr. Hermann Moisl

Dedication

بسم الله الرحمن الرحيم

وَنَفْسٍ وَمَا سَوَّاهَا*فَأَلْهَمَهَا فُجُورَهَا وَتَقْوَاهَا*قَدْ أَفْلَحَ مَنْ زَكَّاهَا*وَقَدْ خَابَ مَنْ دَسَّاهَا

صدق الله العظيم

To the memory of my father the late Ibrahim Hamouda

Acknowledgements

I would like to thank all the people who, during the years in which this work lasted, provided me with assistance and support. Without their tireless care and consideration, this thesis would likely not have matured. First and foremost, my endless thankfulness and unlimited gratitude go unconditionally to Dr. Hermann Moisl, who introduced me to the world of computational linguistics. Without his endless support, I would not have managed to complete this thesis. It was a privilege to work under his supervision, and I am grateful for all our conversations: they made me understand the world and myself in a better way. This leads me to express my thanks to the staff in the school of English Language, Literature and Linguistics in Newcastle University for being consistently cheerful and supportive.

I would like to thank the Egyptian Government that sponsored my PhD scholarship and the Egyptian Cultural Centre in London as well for all the support and help I got during my study in the UK.

I would like to thank those who have worked and continue to work and research in computational linguistics. Special appreciation goes to those who took the effort to put the Arabic language on the map of computational linguistics research. Special thanks goes for Dr. Nizar Habash, Prof. Mona Diab from Columbia University, and Prof. Khaled Shaalan from the British University in Dubai, who all helped me a lot.

Last but not least, I want to express my deep gratitude to my mother, family and brothers (Mohamed and Ahmed) friends especially A.A. Omar for their undivided love and support throughout the years when we have been apart. I conclude, however, by giving a separate word of thanks to my mother whose love and support over the years have always been crucial to me.

Abbreviations

ACE: Automatic Content Extraction

AI: Artificial Intelligence

ANLP: Arabic Natural Language Processing

AR: Anaphora Resolution

BAMA: Buckwalter Arabic Morphological Analyzer

BBC: British Broadcasting Channel

BFP: Brennan, Walker-Friedman, and Pollard algorithm

BNC: British National Corpus

CB: Backward-looking Centre

CEAF: Constrained Entity-Alignment F-Measure

CF: Preferred Centre

CL: Computational Linguistics

CP: Conditional Probability

CP: Preferred Center

DO: Direct Object

DRT: Discourse Representation Theory

ECM: Exceptional Case Marking

EHP: Existential Head Patterns

ERP: Event-Related Potentials

GA: Genetic Algorithm

GB: Government and Binding theory

GPEs: Geopolitical Entities

IP: Infinitive Phrase

IO: Indirect Object

LRC: Left-Right Centering algorithm

MaxEnt: Maximum Entropy Classifier

MSA: Modern Standard Arabic

MT: Machine Translation

MUC: Message Understanding Conferences

NERA: Name Entity Recognition for Arabic

NLP: Natural Language Processing

NLU: Natural Language Understanding

NP: Noun Phrase

POS: Part of Speech

PP: Prepositional Phrase

PSW: Portable Style Writer

RAP: Resolution of Anaphora Procedure

SFA: Semantic Features Acquisition

SL: Source Language

SMT: Statistical Machine Translation

TL: Target Language

TF: Term Frequency

U: Universe of Discourse

WSD: Word Sense Disambiguation

Arabic Transliteration/Encoding Chart

The Buckwalter Transliteration[1] "is a transliteration system that follows the standard encoding choices made for representing Arabic characters for computers. The Buckwalter transliteration has been used in many publications in natural language processing and in resources developed at the Linguistic Data Consortium (LDC). The main advantages of the Buckwalter transliteration are that it is a strict transliteration (i.e., one-to-one) and that it is written in ASCII characters." (Habash 2010:20)

Throughout this thesis the Buckwalter code is used both for citing Arabic words and text in the course of the discussion, and for the representation of the Arabic texts which comprise the corpus on which the proposed anaphora resolution algorithm is based.

Name	UNICODE	Buckwalter	ASMO 449
hamza-on-the-line	\u0621	'	A
madda-on-'alif	\u0622	|	B
hamza-on-'alif	\u0623	>	C
hamza-on-waaw	\u0624	&	D
hamza-under-'alif	\u0625	<	E
hamza-on-yaa'	\u0626	}	F
bare 'alif	\u0627	A	G
baa'	\u0628	b	H

[1] Buckwalter code is adopted from: http://open.xerox.com/Services/arabicmorphology/Pages/translit-chart

taa' marbuuTa	\u0629	p	I
taa'	\u062A	t	J
thaa'	\u062B	v	K
jiim	\u062C	j	L
Haa'	\u062D	H	M
khaa'	\u062E	x	N
daal	\u062F	d	O
dhaal	\u0630	*	P
raa'	\u0631	r	Q
zaay	\u0632	z	R
siin	\u0633	s	S
shiin	\u0634	$	T
Saad	\u0635	S	U
Daad	\u0636	D	V
Taa'	\u0637	T	W
Zaa' (DHaa')	\u0638	Z	X
cayn	\u0639	E	Y
ghayn	\u063A	g	Z
faa'	\u0641	f	a

qaaf	\u0642	q	b
kaaf	\u0643	k	c
laam	\u0644	l	d
miim	\u0645	m	e
nuun	\u0646	n	f
haa'	\u0647	h	g
waaw	\u0648	w	h
'alif maqSuura	\u0649	Y	i
yaa'	\u064A	y	j
fatHatayn	\u064B	F	k
Dammatayn	\u064C	N	l
kasratayn	\u064D	K	m
fatHa	\u064E	a	n
Damma	\u064F	u	o
kasra	\u0650	i	p
shaddah	\u0651	~	q
sukuun	\u0652	o	r
dagger 'alif	\u0670	`	(missing)
waSla-on-alif	\u0671	{	(missing)

Table of Contents

List of Tables

List of Figures

Abstract

In the age of the internet, email, and social media there is an increasing need for processing online information, for example, to support education and business. This has led to the rapid development of natural language processing technologies such as computational linguistics, information retrieval, and data mining. As a branch of computational linguistics, anaphora resolution has attracted much interest. This is reflected in the large number of papers on the topic published in journals such as *Computational Linguistics*. Mitkov (2002) and Ji et al. (2005) have argued that the overall quality of anaphora resolution systems remains low, despite practical advances in the area, and that major challenges include dealing with real-world knowledge and accurate parsing.

This thesis investigates the following research question: can an algorithm be found for the resolution of the anaphor *nafs* in Arabic text which is accurate to at least 90%, scales linearly with text size, and requires a minimum of knowledge resources? A resolution algorithm intended to satisfy these criteria is proposed. Testing on a corpus of contemporary Arabic shows that it does indeed satisfy the criteria.

Introduction

The advent and development of information technology since the mid-twentieth century has generated vast amounts of digitally encoded electronic text in a wide variety of world languages. The most obvious repositories of such texts are the World Wide Web and the increasingly digitally-oriented output from the publishing industry in both academic and leisure spheres. However, at least equally important in terms of volume is text creation in business, government, cultural activity, and personal communication worldwide.

The exploitation of digital text has given rise to a range of research disciplines such as information retrieval (Chowdhury 2003:51**),** data mining (Han et al. 2006; Mucherino et al. 2009; Holmes and Jain 2012), and computational linguistics (Clark et al. 2010), each with its own aims, mathematically and statistically constituted conceptual frameworks, and computational tools. The present thesis is intended as a contribution to computational linguistics.

The historical development of computational linguistics has produced a composite discipline in which ideas from linguistics, computer science, mathematics, and statistics are used to study natural language with a variety of aims (Jurafsky and Martin 2000). For present purposes, these aims can be divided into two broad categories: science and engineering. The science of natural language, that is, linguistics, aims to understand the structure and dynamics of the human language faculty by proposing hypothetical models which can be empirically tested (Allen 1995; Manaris 1998:5)**.** The role of computational linguistics in the science of natural language is, firstly, to provide a basis in the theory of computation for linguistic models, and then to implement such models and to provide tools that make subsets of the worldwide corpus of electronic text available for testing. Natural language engineering, on the other hand, aims to design and implement computational systems which analyze or transform a text corpus for some well-defined practical task without any necessary reference to or implications for linguistic models of the human language faculty (Elhaddad 2006). Concepts from linguistics may or may not be used if relevant, but the primary aim is to carry out the task as efficiently as possible. The present discussion is intended as a contribution to computational linguistics as language engineering.

Machine translation (Hutchins 2005) is a component of contemporary language engineering in the above sense, and is devoted to the design and implementation of computational systems that translate between two or more natural languages as accurately and with as little human intervention as possible. At its most general, this dissertation is concerned with machine translation from Arabic.

A major problem in machine translation has been and continues to be anaphor resolution. An anaphor is understood as a grammatical entity in a text which refers to some other grammatical entity in that text. The problem is due to indeterminacy in anaphor reference (Hirst 1981), where anaphor resolution is a generic term for algorithms which aim to solve that problem (Mitkov 1999; Mitkov 2000; Deoskar 2004)**.** The specific focus of this thesis is anaphor resolution in Arabic with specific reference to the frequently-used anaphor ﻧ فس, which is transliterated into Western orthography as *nafs*.

This thesis comprises an introduction, five main parts, and a conclusion. Part 1 states the aim of the research reported in this thesis, the research question which it addresses, and the methodology it uses. Part 2 outlines the nature of anaphora in general and in Arabic more particularly. Part 3 reviews anaphor resolution factors in general and MSA anaphor resolution in particular. Part 4 reviews anaphora resolution in general and in MSA in particular. Part 5 reviews the grammar of *nafs*. Part 6 proposes an algorithm for the resolution of *nafs*, implements the proposed resolution algorithm, applies it to a MSA corpus, and assesses the results. Part 6 also briefly identifies future work related to the research described in the thesis. The conclusion then summarizes the discussion.

Chapter 1. Aim, Research Question, and Methodology

This chapter introduces the aim of the study, defines the research question and outlines the methodology adopted to answer it. Part 1.1 states the aim of the study, part 1.2 defines the research question, and part 1.3 explains how the research question is addressed.

1.1 Aim

Nafs is a frequently-occurring anaphor in contemporary Arabic (Kremers 1997)**,** and any machine translation system from Arabic will need to be able to resolve it. The aim of this thesis is to design and implement a reliable and efficient resolution algorithm for *nafs* which can be used as a component in a computational system that translates Arabic into some target languages in practical, real-world applications.

- ‘Reliable’ is taken to mean that the algorithm should ideally be able to correctly resolve all instances of *nafs* in any text collection to which it is applied, where the criterion for correctness is based on native speaker competence, or, failing this ideal, that it should be able to resolve *nafs* correctly with an accuracy comparable to that of state-of-the-art anaphor resolution systems for languages such as English, which is currently 90% or a little greater (Mitkov 2002).

- ‘Efficient’ is understood in two senses. In the first sense it is taken to mean that the algorithm should resolve anaphora within

> a time limit that users find acceptable irrespective of the size of the text or text collection to which it is applied. In other words, the algorithm must scale well in terms of computational complexity, and in the ideal case its computational complexity would be O(n), that is, the time required to resolve all instances of *nafs* should grow no more than linearly with text size. In the second sense, ‘efficient’ means financially cost-effective. Existing anaphor resolution systems, as reviewed later in the discussion, require to varying degrees syntactic, semantic, and real-world knowledge provided by, for example, mark-up in the text being processed, parsers, and knowledge representation databases. Such provision is typically labour-intensive and thus expensive; the aim here is to design an algorithm that requires as few of such knowledge resources as possible.

Implicit in the foregoing comments is that the focus of the discussion is on text rather than speech. Speech processing requires a competence that the author cannot claim, though there is no obvious reason why the proposed algorithm should not be adaptable for the resolution of *nafs* in speech.

1.2 Research Question

Based on the above aim, the research question addressed by this thesis is:

Can an algorithm be found for the resolution of nafs in Arabic text which is accurate to at least 90%, scales linearly with text size, and requires a minimum of knowledge resources?

1.3 Methodology

A two-stage methodology is used:

1. Survey the existing anaphor resolution literature. The survey is divided into three main parts. The first part deals with the linguistic and psycholinguistic background. The second part covers data driven approaches which depend on annotated corpora, and the third deals with anaphora resolution in Arabic.

2. Work sequentially through the ranking of approaches until the required 90% accuracy of *nafs* resolution is attained with respect to a test corpus of contemporary Arabic text. Start at the beginning of the ranking with the approach that has the best scaling behaviour and the lowest level of knowledge requirement, and design, implement, and test an algorithm based on that approach. If the implemented algorithm fails to meet the required accuracy, supplement or replace it with the next approach in the ranking. Continue to supplement or replace approaches until the threshold accuracy is reached. If the end of the ranking is reached without the threshold being attained, think of a new approach.

Chapter 2. The Nature of Anaphora

2.1 Introduction

This chapter is concerned with defining and understanding the nature of anaphora and its classification methods as a starting point for thinking about approaches to anaphora resolution. The discussion is not confined to Arabic, however; it discusses anaphora in general terms, giving examples from English, Arabic, and other languages and how they have been approached using different syntactic and grammatical approaches. The chapter reviews the definition of an anaphor, types of anaphora, types of antecedents, relations between antecedents and anaphora, scope of the suggested algorithm, pronominal anaphora in MSA, means of expressing anaphora in MSA, and restrictions on MSA anaphora.

2.2 What is an Anaphor?

The definition of an 'anaphor' proposed is: 'a grammatical entity in a text which refers to some other grammatical entity in that text'. Jurafsky and Martin (2000: 672) define it as 'the reference to an entity that has been previously introduced into the discourse', and Hirst's (1981: 4) definition is 'the device of making in discourse […] an abbreviated reference to some entity in the expectation that the perceiver of the discourse will be able to disabbreviate the reference and thereby determine the identity of the entity. The reference is called anaphor, and the entity to which it refers is its referent or antecedent'. Varieties of anaphora are given below.

More recently, the literature has been concerned with making a distinction between ‘referent’ and ‘antecedent’. Mitkov (1999), for instance, says with respect to any given anaphor that ‘the referent is the object or the state of affairs in the extralinguistic reality to which the referring expression refers, whereas the antecedent is the linguistic realization of this entity’. The present discussion adopts this distinction throughout, and is concerned solely with antecedents.

2.3 Varieties of Anaphora

There are various types of anaphora in natural language text. One of the several ways of classifying anaphora is by form (Leass and Lappin 1994; Mitkov 2002), and this is the classification adopted here. In terms of form, there are three broad classes of anaphora:

- Pronominal anaphor are pronouns, as their name indicates. Example: ‘The man ran into the shop, and there *he* bought a newspaper’, where the anaphor is *he* and the antecedent is *the man*. Note that not all pronouns are anaphora, however. In a sentence such as ‘It is raining’, *it* is referred to as pleonastic.

- Noun-phrase anaphora are, again as the name indicates, noun phrases that refer to antecedents that are themselves noun phrases whose reference is to identical or semantically close concepts. Example: ‘The club has its annual dinner. Members were asked to come alone’. Here, ‘Members’ is the noun-phrase anaphor and ‘The club’ is the antecedent.

- ‘One’ anaphora: Example: ‘If you can’t make this appointment, you can arrange another one’. Here ‘one’ is the anaphor, and ‘this appointment’ is the antecedent.

As the foregoing examples imply, the antecedent of any given anaphor does not have to be in the same sentence as the anaphor. Where it is, the anaphor is said to be intrasentential, and where the antecedent is not in the same sentence it is intersentential. Anaphora are often intrasentential, and, where they are not, the antecedents are often found in the preceding one or two sentences, but antecedents may be far as seventeen sentences away from their anaphora as reported by Mitkov (1999: 3).

The classification of pronominal anaphora depends on three factors: types of existing anaphor, types of existing antecedents, and the relations between each of them. Mitkov (2002) mentioned different kinds of anaphora, including pronominal anaphora, verb and adverb anaphora, noun anaphora and zero anaphora. The current thesis is only concerned with pronominal anaphora. Mitkov (2002) further classified pronominal anaphora depending on the anaphor into three types, which are discussed below.

2.3.1 Nominal Anaphora

According to Mitkov (2002: 8), a nominal anaphor is a ‘referring expression (pronoun, definite noun phrase or proper name) which has a non-pronominal phrase as its antecedent’. This is the most commonly researched type of anaphor in natural language processing (NLP) literature. The most important type of nominal anaphor is the pronominal anaphor. Pronominal anaphora forms are personal pronouns (*he, she, it, they, them, her, him*), possessive pronouns (*his, her, its, their, theirs, hers*), reflexive pronouns (*himself, herself, itself, themselves*), demonstrative pronouns (*this, that, these, those*), or relative pronouns (*who, whom, which, whose*. Sometimes *where* and *when* may be anaphoric as well in cases of locative and temporal anaphora). First and second person singular pronouns are usually deictic in function as in 'can *you* kindly pass me the salt?'. In reported speech, such a function does not commonly occur.

2.3.2 Pleonastic *It*

The pronoun *it* frequently occurs in cases when it is non-anaphoric. For example:

'It is highly unlikely to change the price now'.

Leass and Lappin (1994) name such a use of *it* as being *pleonastic it* while Quirk et al. (1985) call it *prop it*. Mitkov (2002) has summarized some of the instances where the pleonastic *it* occurs:

i. Modal adjective constructions, for example: *It is obvious*, etc.

ii. Cognitive verb constructions, for example: *it is considered to be*, etc.

iii. Temporal constructions, for example: *It is spring*, etc.

iv. Distance-related constructions, for example: *It is far from here*.

v. In idioms, for example: *It's anyone's call.*

vi. Cleft constructions, for example: *It was Mr. Edgar who recruited Prudence Adair* (Mitkov 2002: 10).

The pleonastic *it* is not a clear research area and it is still a matter of debate in linguistics. Consequently, the automatic identification of the pleonastic *it* is still a difficult task.

2.3.3 Zero Pronominal Anaphora

Zero pronominal anaphora occur if the anaphoric pronoun is deleted but is still understood from the general context. Although this case is very rare in English, it is frequent in languages such as Arabic, Chinese, and Spanish. In most cases, zero pronouns in such languages are substituted by overt pronouns in English. Hirst (1981) addressed three problems in classifying pronominal anaphora. He was concerned with differentiating between a *pronoun* and a *noun phrase,* since in classical grammar these were considered to be the same although he proves this to be incorrect. To overcome such problems he uses the term 'pronominally referent' to refer to noun phrases. Hirst states that most pronouns are pronominally referent, such as pronouns marked for gender and number, which makes the process of resolution easy. However, there are cases where such a rule does not apply, for example:

'Who is this Bresson? Is *she* a woman?' (Hirst 1981: 10)

Here, *she* refers to the film director Robert Bresson who is male figure.

A further problem is the use of the singular epicene pronoun, which Hirst defines as: “a genderless plural third-person pronoun referring to a singular third-person of unknown, or deliberately unmarked, gender” (Hirst 1981: 11). For example:

‘The university thanks the students for their patience’.

Such use is accepted in many idiolects while it is rejected in others. An AR algorithm has to be able to accommodate itself to such a use.

The third problem is the use of the expression *same*, which can act as a pronoun but is restricted to referring to a very recent noun phrase.

Hirst classifies pronominal anaphora into three types:

2.3.4 Pronoun Anaphora

These refer to parts of speech such as *he, she*, *it*, *they*, *that*, etc. They are usually marked with number and gender which make the process of resolution easy.

2.3.5 Surface Count Anaphora

These are noun phrases that act as pronominal anaphora. This category includes the constructions *the former* and *the latter*. This type of anaphor requires that both the surface structure of the sentence and the antecedent are retained in the consciousness of the reader or the listener. One major problem of such anaphora is that any designed algorithm faces the problem of determining where to start counting backward in order to find the possible antecedent.

2.3.6 Pronominal Noun Phrases: Epithets

Epithets can act as pronominal anaphors, although as Lakoff and Ross (1976) stated they cannot have pronouns as their antecedents.

2.4 Types of Antecedents

Noun phrases are not the only type of antecedents for pronominal anaphora. Antecedents may be clauses, sentences or situations described by a sequence of sentences. In such cases the antecedent(s) can be referred to using *it* or *this*/*that*, for example:

‘We cooked and ate the quiche in the evening. It was delicious.’

Another type of pronominal antecedent is coordinated noun phrases, a sequence of NPs separated by commas or conjunctions, for example:

‘Nadia and Omar bought their first house a year ago.’

2.5 Relations Between Anaphora and Antecedents

Anaphora can be classified as identity-of-reference-anaphora or identity-of-sense anaphora. Identity-of-reference-anaphora occur when the pronoun and its antecedent refer to the same entity. For example:

‘I saw a bird. It was singing.’

Identity-of-sense-anaphora is another type, where the pronoun shares the sense with its antecedent, for example:

‘Omar bought a house and I bought one too.’

Another factor that affects the identification of the antecedent of an anaphor is the location of each in relation to the other. In the examples above, the pronouns always follow their antecedents in a backward direction; if the direction is reversed it is called a *cataphor*. The usual distance between an anaphor and its antecedent is two-to-three sentences but it can extend up to seventeen sentences, as Mitkov (1996) has noted.

2.6 Scope of the Present Algorithm

This thesis is concerned with pronominal anaphora only, and thus the discussion is limited to Arabic reflexives only represented by *nafs*. In English this represents a limited set consisting of *himself*, *herself*, *itself*, and *themselves*. In modern standard Arabic (MSA), the reflexive *nafs* consists mainly of *nafs* as a base in addition to a clitic pronoun as a suffix, and in some cases it may have also have a prefix. In order to explain how this works, an account of Arabic pronouns is introduced later with special focus on third person pronouns, since *nafs* acts in the same manner. First and second person pronouns are not discussed due to two factors:

1. The infrequent occurrence of first and second person pronouns in newswire texts.

2. First and second person pronouns normally appear in quotations which are considered to have a limited effect on the structure of the discourse.

Demonstratives were excluded as they refer to extralingustic contexts. Furthermore, in MSA they are usually cataphoric, and thus lie outside the scope of the present thesis. The antecedents of relative pronouns that occur within the same sentence are still a parsing problem. The current thesis focuses on resolving the identity-of-reference pronominal anaphora Arabic reflexive *nafs*.

2.7 Pronominal Anaphora in MSA

This section is based on Hammami et al. (2009), who give a typology of MSA pronoun anaphora resolution in a study which is considerably detailed and relevant to the present thesis. In general, Arabic anaphora can be divided into pronominal anaphora, lexical anaphora, verb anaphora and comparative anaphora.

Before explaining pronominal anaphora in Arabic, it is necessary to briefly explain the linguistic situation with regard to Arabic in general, and to give a survey of its pronouns, reflexives and reciprocals since all of these are considered to fall within the pronominal category in Arabic.

In reality there is no single language called ‘Arabic’; however, there is a wide range of different dialects that ought to be considered ‘Arabic’ according to the points of view of its users. To expand on this, in reality there are two types of Arabic; firstly there is a written language that is called ال فصحى *AlfSHY*, which means ‘the eloquent’. Secondly, there is a spoken language called العامية *AlEAmyp*, which means ‘the common’.

The language that was used before the rise of Islam in the fourth century A.D. is called Classical Arabic (Tawfiq 2009). This language underwent many changes as the religion spread geographically and as more people progressively adopted the language of Islam (Classical Arabic). This process had a profound impact on the language, creating a number of dialects as different regions adapted the language in diverse ways.

Classical Arabic, however, remained widely used as a formal written language containing rather archaic verb forms and structures as well as incorporating many new words and structures. The modern variety of classical Arabic is called modern standard Arabic (MSA), which is considered to be an artificial language that children start to learn when they attend school. Although traces of the colloquial languages (dialects, as some would say) can be found, MSA remains widely used in all formal communication and newswire writing. This thesis concentrates on MSA as a variety of Arabic.

After this general introduction to the Arabic language, it would be useful to give a brief explanation of Arabic grammar in order to ease the understanding of Arabic examples used in the thesis. There are two types of sentences in MSA: nominal and verbal. Nominal sentences consist of a noun phrase (NP) and a predicate. The NP may be followed by either another NP, an adjective phrase (AP), a prepositional phrase (PP) or a verb phrase (VP).

A verbal sentence consists of an NP and a VP. The structure of the VP determines the complement type it may take. If the verb is intransitive, for example, it takes no complements. If the verb is transitive it will take complements depending on the sub-categorization nature of the verb.

MSA has distinct forms used to convey features of number, person and gender. For number, there are three forms: singular, dual, and plural. This means that it differs from English, which has only the two forms of singular and plural. This creates a problem when translating from Arabic to English. In the latter, a dual number will be treated as plural, whereas the dual has its own features from the point of view of the assignment of cases. So, in a nominal sentence where all the agreement features must be visible, the number's properties have to dominate over the predicate. Person and gender features, on the other hand, must dominate over each pronoun.

Before discussing MSA pronouns it is important to note that dependent pronouns in Arabic are referred to as either suffixes or enclitics. To solve such an issue Soudi et al. argue that in MSA suffixes can be found (2007: 125) 'in verbal inflexions, nominal cases, the nominal feminine ending ات/a (t)/, ة+ah, etc., while enclitics are complement pronouns some verbs can have a double enclitics as for example علمتمونيها /ElmtmwnyhA/ "you taught me it". ' In MSA enclitics are regarded as suffixed possessive and direct object pronouns while suffixes occur in other positions. The majority of MSA grammar books do not make such a distinction clear and resort to using the word suffix to express both suffixes and enclitics. See, for example, Ryding (2005). Even in ANLP books, for example, Habash (2010:44) argues that enclitics are 'clitics that follow the word (like a suffix).' This researcher chooses to use the term suffix as it is broader and would include enclitics within it. The table below is adapted from Soudi et al. (2007: 161) to show possible MSA pronoun enclitics/ suffixes.

Table 2.1: Possible MSA pronoun enclitics/ suffixes (Soudi et al. 2007:161).

MSA	Transliteration	Number	Gender	Object	Possessive
ي	y	singular	masculine/feminine		my/mine
ني	ny	singular	masculine/feminine	me	
نا	nA	plural	masculine/feminine	ours	our
ك	k	singular	masculine/feminine	you	yours
كما	kmA	dual	masculine/feminine	you	yours
كن	kn	plural	feminine	you	yours
كم	km	plural	masculine/feminine	you	yours
ه	h	singular	masculine	him/it	his/its

ها	hA	singular	feminine	her/it	her/its
هما	hmA	dual	masculine/feminine	their	theirs
هم	hm	plural	masculine	their	theirs
هن	hn	plural	feminine	their	theirs

Pronouns in Arabic are called الضمير AlDmyr, which means ‘something hidden’. The reason behind this name is that the pronoun hides the noun that it refers to. Arabic has two sets of pronouns: independent pronouns and pronominal suffixes (dependent pronouns). In table 2.2 MSA independent pronouns are shown.

Table 2.2: MSA independent pronouns

MSA independent pronouns	Transliteration	English pronouns
أن ا	/>nA/	I (first person, masculine/feminine)
ن حن	/nHn/	We (first person, masculine/feminine)
أنتَ	/>nta/	You (second person,

		masculine)
أنتِ	/>nti/	You (second person, feminine)
أنـ تما	/>ntmA/	You (dual, masculine/feminine)
أنـ تن	/>ntn/	You (third person, feminine plural)
أنـ تم	/>ntm/	You (third person, masculine plural)
هو	/huwa/	He(third masculine singular)
هي	/hiya/	She (third feminine singular)
هما	/huma/	The two of them (third masculine/feminine dual)
هم	/hum/	They (third masculine plural)

هن	/hun/	They (third feminine plural)

Independent nouns are nominative. However, as Arabic is a pro-drop language, when the verb is present the subject pronoun is dropped in most cases and, if added, it would be used for emphasis. Pronominal suffixes that are added as verb suffixes express the accusative, while they occur as noun suffixes and prepositional suffixes to express the genitive, as indicated in table 2.3:

Table 2.3: MSA genitive and accusative case with pronominal suffixes

MSA	Transliteration	English pronouns
ه	/hu/	Third masculine singular
ها	/ha/	Third feminine singular
هما	/huma/	Third masculine/feminine dual
هم	/hum/	Third masculine plural
هن	/hun/	Third feminine plural

كَ	/ka/	Second masculine singular
كِ	/ki/	Second feminine singular

In the examples used in this thesis, a pronoun suffix that is being attached to a verb is used to express the verb object, while a pronoun suffix attached to a noun is used to express the possessor. As for the issue of ‘definiteness’ and ‘indefiniteness’, MSA uses the definite article ‘*al*’ which is attached to nouns regardless of agreement features. However, *al* cannot be attached to pronouns. Indefiniteness is expressed using a nunational marker[2] which is used in cases of segregation.

NPs in MSA play a pivotal role since they can fulfil several syntactic rules as subject, subject complement, object, object complement and object preposition. Some nouns have structures which are characterized by having definite anaphoric relations, and such NPs are called anaphora. In languages like Arabic and English, anaphora is one of the nominal features. In MSA, anaphora is characterized by having co-referential relations with antecedents existing in the same sentence. In order to understand them NPs must be determined by their referents,

[2] With an indefinite noun or adjective a short vowel plus /n/ sound is to be added.

and this is the reason behind the confusion in determining the reference in the structure of NPs.

2.7.1 Pronominal Anaphora

Pronouns in Arabic are characterized by having an empty semantic structure, where the meaning of a pronoun is dependent on its antecedent. This excludes deictic pronouns such as أنا *>nA* ‘I’, أنت *>nt* ‘you’ and نحن *nHn* ‘we’. Pronominal anaphora is subdivided into: nominative disjoint personal pronouns, accusative disjoint pronouns, dative and accusative personal pronouns, nominative joint personal pronouns, and relative pronouns. These are described below.

2.7.1.1 Nominative Disjoint Personal Pronouns

(الضمائر المنفصلة في محل رفع)

As mentioned in the table 2.2, for example:

أكل الأولاد الطعام و هم جالسون بجوار المدفأة

Transliteration: /*>kl Al>wlAd AlTEAm w hm jAlswn*

Glossing: ate the-boys the-food and they sitting-them

bjwAr Almdf>p/

next to radiator.

Translation: ‘The kids ate the food while they were sitting next to the radiator.’

2.7.1.2 Accusative Disjoint Pronouns

(الضمائر المنفصلة في محل نصب)

Table 2.4: A list of Arabic accusative disjoint pronouns

MSA	Transliteration
إياه	/~aAhu/
إياها	/~aAhaA/
إياهما	~aAhumaA/
إياهم	/~aAhumo/
إياهن	/~aAhun~a/

For example:

جمال اللحظة يكمن في أن يشاركنا الآخر إياها

Transliteration: /*jmAl AllHZp ykmn fy >n y$ArknA*

Glossing: beauty moment exists in comp to-share-us

Al|xr ~ aAhaA/

the-other them.

Translation: ‘The beauty of the moment is that someone is sharing it with us.’

2.7.1.2.1 Functions of independent personal pronouns

Independent personal pronouns are used in various ways and may be used as an essential part of a clause or as a non-essential part. The various functions are summarized below.

2.7.1.2.1.1 Emphasizing the subject of the verb

MSA verbs include the subject in their inflections, which consequently makes the personal pronoun unable in most cases to mark the inflection of the verb phrase subject. In addition to the verb, however, the pronoun can be used to emphasize the subject. In the example below, extracted from Ryding (2005), the pronoun can be deleted and the sentence continues to be grammatically correct, but the subject receives less emphasis.

كانت هى نقطة التحول

Transliteration: /*kAnt hY nqTp AltHwl*/

Glossing: was it point the-turning.

Translation: ‘It was the turning-point.’

In the above mentioned example هى ‘it’, which is a singular third person feminine pronoun, can be deleted and the meaning is still conveyed successfully and the sentence remains grammatically correct.

2.7.1.2.1.2 Subject of an equational sentence

An equational sentence is a type of sentence that has no overt verb, but a pronoun can be used as a subject instead. Consequently the pronoun is stated first in the sentence; for example:

هو خبير فى شئون الشرق الأوسط

Transliteration: /*hw xbyr fY $}wn Al$rq Al>wsT*/

Glossing: he expert in affairs the-east the-middle.

Translation: ‘He is an expert in Middle Eastern affairs.’

Although this sentence has no verb, because the pronoun هو ‘he’ is mentioned at the beginning of the sentence, it is therefore grammatically correct.

2.7.1.2.1.3 Predicate of equational sentence

Although it does not commonly occur, there are cases when a pronoun acts as a predicate of an equational sentence; for example:

هذا هو

Transliteration: */h*A how*/

Glossing: this he.

Translation: ‘This is he.’

2.7.1.2.1.4 As a copula

In an equational sentence, the relationship between the subject and a predicate needs to be defined and clarified precisely when the predicate is a definite noun or noun phrase. In such a case, a third person subject pronoun may be added between the subject and the predicate to link them together and to act as the verb ‘to be’ which is then considered to be a copula. For example:

المهم هو العودة

Transliteration: /*Almhm* *hw* *AlEwdp*/

Glossing: the-important it the-return.

Translation: ‘The important [thing] is to return.’

2.7.1.3 Dative and Accusative Personal Pronouns

(الضمائر المتصلة في محل نصب و جر)

As mentioned in the table 2.3 they are:

ه *hu*, ها *ha*, هما *huma*, هم *hum*, هن *hun* only.

For example:

أخي محمد لـ يس لـ ه أخت سواي وولـ دان يـ عـ يـ شان معه

Transliteration: />*xy* *mHmd* *lys* *lh* >*xt* *swAy*

Glossing: brother-me Muhamed not him-for sister me *wwldAn* *yEy$An* *mEh/*

boys-two live-two him-with.

Translation: ‘My brother Muhamed has only got one sister; that is me. He has two boys who live with him.’

It is clear that a dative accusative pronoun cannot begin a sentence, and must therefore be attached to a noun, preposition or verb. In contrast, both nominative and accusative disjoint pronouns can occur at the beginning of a sentence. Disjoint pronouns can have a prefix in order to convey a conjunction, as in و *waw* or ف *fA'*.

2.7.1.4 Nominative Joint Personal Pronouns

(الضمائر المتصلة في محل رفع)

Table 2.5: A list of Arabic nominative joint personal pronouns

MSA	Transliteration
ا	/Alef/
واو	/waw/
نون	/noon/

For example:

أكل الأولاد (VSO)

Transliteration: /*>kl* *Al>wlAd*/

Glossing: ate the-boys.

Translation: ‘The children ate.’

الأولاد أكلوا (SVO)

Transliteration: /*Al>wlAd >klwA*/

Glossing: the-boys ate.

Translation: ‘The children ate.’

The above two examples demonstrate that the nominative joint pronoun behaves in a special manner, as it is always suffixed to a radical verb. This always leads it to take the position of a subject in the SVO sentence, while in a VSO sentence structure we cannot use the pronoun since the subject occurs after the verb.

2.7.1.5 Relative Pronouns

Relative pronouns in Arabic are always anaphoric and refer directly to a previously mentioned noun phrase.

Table 2.6: A list of Arabic relative pronouns

MSA	Transliteration
الذي	/Al*y/
التي	/Alty/
اللذان	/All*An/

اللتان	/AlltAn/
اللذين	/All*yn/
اللتين	/Alltyn/
الاتي	/AlAty/
اللواتي	/AllwAty/
اللائي	/AllA}y/
الذين	/Al*yn/
الآلاء	/Al|lA'/
من	/mn/
ما	/mA/

2.7.2 Lexical Anaphora

Lexical anaphora occur in a sentence when the antecedent is a proper name or a definite description. The aim of using such a type is to increase cohesiveness. For example (Hammami et al. 2009):

ولد ابن خلدون في تونس ثم هاجر العلامة إلى مصر

Transliteration: /*wld Abn xldwn fy twns vm hAjr*

Glossing: born-he Ibn Khaledon in Tunisia then immigrated-he

AlElAmp AlY mSr/

scientist to Egypt.

Translation: ‘Ibn Khaledon was born in Tunisia, then the scientist immigrated to Egypt.’

2.7.3 Comparative Anaphora

In this case the anaphoric expression is introduced or modified by a lexical modifier or a comparative adjective. It aims to make anaphoric relationships more specific. For example (Hammami et al. 2009):

كان قد كان لكم فى فئتين التقتا واحدة فى سبيل الله وأخرى كافرة (Quran)

Transliteration: /*qd kAn lkm fY f}tyn AltqtA*

Glossing: have had to-you in forces-dual met-two

wAHdp tqAtl fb sbyl Allh w >xrY kAfrp/

and one-fights in favour Allah and another against/
Translation: ‘There, you people have had an intellectual lesson to comprehend: two forces met; one fighting in favour of God and the other against God.’

2.7.4 Verb Anaphora

These occur when the ف عل *fEl* ‘verb’ did is used, for example (Hammami et al. 2009):

خلقنا من أجل أن نؤدى واجبتنا و ليس لنا بد من تأديتها فان لم نفعل فنحن وحدنا الملومين.

Transliteration: /*xlqnA mn Ajl An n&dy*
Glossing: created-we from for that achieve-we
wAjbAtnA w lys lnA bd mn t>dythA fAn lm
duties-our and not us must to achieve-it so-that not
nfEl fnHn wHdnA Almlwmyn/

do so-we and-alone-we the-blame-we.

Translation: ‘We live to do our duties and we have to achieve them and if we don’t, we are the only ones reproachable.’[3]

2.8 Means of Expressing Anaphora in MSA

2.8.1 Deletion

If a subject NP, whether a full nominal or a pronoun, is followed by a string of verbs, it is obligatorily deleted after its first appearance. For example (Holes 2002):

هو جاءني و جلس بجانبي و بدء في الحديث

Transliteration: /*hw jA'ny w jls bjAnby w bd' fy*

Glossing: he came-I and sat beside-I and started in

AlHdyv/

the-speech.

Translation: ‘He came and sat by me and began to talk.’

[3] It is important to note that from the above section, it can be concluded that *nafs* becomes a reflexive anaphor when it is followed by a pronominal suffix. That is why it is important to speak about the pronouns’ linguistic behaviour since the pronominal suffix must agree in number and gender with the antecedent of *nafs*.

In the above example, the subject *hw* ‘he’ is mentioned once in the beginning of the sentence and it is not repeated again as the conjunction *w* is used instead.

The same applies if the subject NP of the main clause is the same as the subject NP of the subordinate clause(s). For example (Holes 2002):

أنا أرغب في رؤية الحادثة بنفسي

Transliteration: /*>nA >rgb fy r&yp AlHAdvp bnfsy*/

Glossing: I want to see the-accident by-self-me.

Translation: ‘I want to see the accident by myself.’

In the above example, the subject *>nA* ‘me’ is mentioned at the beginning and not mentioned again , instead a reflexive *bnfsy* is used.

An anaphor is also commonly realized by deletion in conversational exchanges involving answers to questions, follow-on comments from interlocutors, or echo questions; for example (Holes 2002):

ما يقوله معقول

Transliteration: /*mA yqwlh mEqwl*/

Glossing: what say-he sense.

Translation: A: ‘What he is saying is reasonable.’

ولكنه غير مقبول

Transliteration: /*wlknh* *gyr* *mqbwl*/

Glossing: and-however-it not acceptable.

Translation: B: ‘But it is not acceptable’

In the above example, there is a deletion of the pronoun *hw* ‘he’ after the verb *yqwlh* as it can be understood from the conversation. In the second line there is a deletion of the construction *mA yqwlh* as it can be understood from the conversation.

Deletion also occurs if the element concerned is marked on the verb. If they govern several verbs, full nominal subjects are deleted after their occurrence. With or without free subjects, all verbs are marked for person, gender and number. For example (Holes 2002):

المدير وضع السماعة و بدء في الصراخ

Transliteration: /*Almdyr* *wDE* *AlsmAEp* *w* *bd'* *fy*

Glossing: the-boss put the-microphone and started to-

AlSrAx/

the-shout.

Translation: ‘The manager put down the receiver and began screaming.’

In the above example the subject *Almdyr* is only mentioned once and is deleted even after *bd'* as it is understood to be the subject.

2.8.2 Ordinary Person Pronoun

Verbs are morphologically inflected to agree with their subject for gender, person, and number, and so it is not normal to use independent personal pronouns anaphorically in such cases. It is also unnecessary to use independent pronouns where the subjects of co-ordinated clauses are different, or if the subject of the right-hand clause refers back to an element in the left-hand clause; for example (Holes 2002):

قدمت لهم العرض و لكنهم لم يوافقوا

Transliteration: /*qdmt* *lhm* *AlErD* *w*

Glossing: presented-me for-them the-proposal and

lknhm *lm* *ywAfqwA*/

however-they not accepted-they.

Translation: ‘I presented the offer but they didn’t accept.’

In the above example, the verb *ywAfqwA* is masculine plural, which agrees in gender and number with the dependent pronoun in *lknhm* and *lhm,* and although the subjects of the two clauses are different (in the first one it is *me,* and in the second it is *they*) no independent pronoun is needed.

Where a verb in the right-hand clause could theoretically refer either to the subject or the object of the clause, it is interpreted pragmatically as referring to the subject; for example (Holes 2002):

أحمد ضرب على و هرب

Transliteration: />*Hmd* *Drb* *ElY* *w* *hrb*/

Glossing: Ahmad hit Ali and escaped.

Translation: ‘Ahmed hit Ali and fled.’

In the above example, the verb *hrb* is thought to be referring to Ahmad not to Ali. Although Ali agrees in number and gender with it, it is pragmatically understood to refer to Ahmad.

Enclitic pronouns which are used in DO (directly attached to the verb) and IO (normally attached to the preposition) to refer to the nominal are always anaphoric.

2.9 Restrictions on Anaphora

Holes (2002) noted that a general restriction on an anaphor is that it must refer to a backward antecedent and not a cataphoric expression.

2.9.1 Scope of the Anaphor

The anaphor’s scope is limited to:

1. The clause, even if it is a verbal affix.

2. Intraclause reflexivity, where the reflexive element is a verbal affix.

2.9.2 Possible Syntactic Functions of the Antecedent

It may be a subject, for example (Holes 2002):

رأيت رجل يتجرد من ملابسه

Transliteration: /*r>yt rjl ytjrd mn mlAbsh*/

Glossing: saw-me man expose from clothes-his.

Translation: ‘I saw a man stripping off.’

In the above example, the subject of the sentence that is the deleted pronoun *>nA* ‘me’ acts as the antecedent. It is referred to by the pronoun attached to verb *r>y*.

2.9.3 Possible Functions of the Reflexive Markers

Such a marker may be the DO, for example (Holes 2002):

خلع ردائه قبل ان ينام

Transliteration: /*xlE rdA}h qbl An ynAm*/

Glossing: removed garment-his before that sleep.

Translation: ‘He undressed before he went to bed.’

In the above example, the DO *rdA}* is attached to a reflexive marker that is the *h*.

Conversely, it may be one of the two DOs; for example (Holes 2002):

هو تعلم اللغة العربية بروحه/بمفرده

Transliteration: /*hw tElm Allgp AlErbyp brwHh/bmfrdh*/

Glossing: he learned the-language the-Arabic by-self-he.

Translation: ‘He learnt Arabic by himself/he taught himself Arabic.’

In the above example, *brwHh/bmfrdh* acts as the second DO for the verb *tElm.*

Or, it may be an IO; for example (Holes 2002):

كسبت مال كثير

Transliteration: /*ksbt* *mAl* *kvyr*/

Glossing: gained-me money many.

Translation: ‘I gained much wealth.’

And finally it may indicate reciprocity, for example (Holes 2002):

التقي الناس مع بعضهم البعض في القاعة

Transliteration: /*Altqy AlnAs mE bEDhm AlbED fy AlqAEp*/

Glossing: met the-people with themselves them in the-room

Translation: ‘The people assembled in the hall.’

In the above example, the reflexive marker *hm* is attached to *bED* to indicate reciprocity.

2.9.4 Intraclause Positional Possibilities of the Reflexive Pronoun

1. As DO with the subject as an antecedent, for example (Holes 2002):
   ضرب نفسه بعصا

   Transliteration: /*Drb nfsh bESA*/

   Glossing: hit self-he with-stick.

   Translation: ‘He beat himself with a stick.’

In the above example, the reflexive *nfsh* acts as the direct object for the verb *Drb.*

2. As a modifier of the DO with the subject as an antecedent, for example (Holes 2002):
   سمع صوت نفسه

   Transliteration: /*smE Swt nfsh*/

   Glossing: heard sound self-he.

   Translation: ‘He heard his own voice.’

In the above example, the reflexive *nfsh* acts as the modifier of the DO *Swt.*

3. As an IO with the subject as antecedent (zero marking), for example (Holes 2002):
   أعطي نفسي فرصة للنجاح

Transliteration: />*ETy* *nfsy* *frSp* *llnjAH*/

Glossing: give self-me chance for-the-success.

Translation: ‘I’ll give myself the chance to succeed.’

In the above example, the reflexive *nfsy* acts as the IO for the verb >*ETy*.

4. As a modifier of such an IO.
5. As an IO (adposition marking) with the subject as antecedent, for example (Holes 2002):

أعتمد على نفسك

Transliteration: />*Etmd* *ElY* *nfsk*/

Glossing: Depend on self-you

Translation: ‘Depend on yourself.’

In the above example, the reflexive *nfsk* acts as the IO for the verb >*Etmd.*

6. As a modifier of such an IO with the subject as antecedent. In such a case, an ordinary possessive noun is used with a subsequent disjunctive pronoun echo to indicate the self; for example (Holes 2002):

سأعطيها لولدي أنا

Transliteration: /*s>ETyhA* *lwldy* >*nA*/

Glossing: will-give-her to-son-my me

Translation: ‘I will give it to my own son.’

7. As a copular complement with the subject as antecedent.
8. As a modifier of a copular complement with the subject as antecedent, for example (Holes 2002):

أنت والله عدو نفسك

Transliteration: />*nt* *wAllh* *Edw* *nfsk*/

Glossing: you and-Allah enemy self-you

Translation: ‘By God, you are your own worst enemy.’

In the above example, the reflexive *nfsk* acts as the modifier for *Edw* and >*nt* is the subject.

9. As a subject-complement with the subject as antecedent, for example (Holes 2002):

عقب أن تزوج رجع نفسه

Transliteration: /*Eqb* >*n* *tzwj* *rjE* *nfsh*/

Glossing: after that married-he returned self-him.

Translation: ‘After he got married he became himself again.’

In the above example, the reflexive *nfsh* acts as subject-complement and the subject is its antecedent which is a deleted pronoun *hw*.

10. As a modifier of a subject-complement with the subject as antecedent, for example (Holes 2002):

عقب تزوجه صارعدو نفسه

Transliteration: /*Eqb tzwjh SAr Edw nfsh*/

Glossing: after married-he became enemy self-him.

Translation: ‘After he got married, he came to be his own worst enemy.’

In the above example, the reflexive *nfsh* acts as modifier of a subject-complement and the subject is its antecedent which is a deleted pronoun *hw*.

11. As an object-complement with the subject as antecedent, for example in the case 6 above.
12. As a modifier of an object-complement with the subject as antecedent, for example (Holes 2002):

    هم جعلوه عدو نفسه

    Transliteration: /*hm jElwh Edw nfsh*/

    Glossing: they made-him enemy self-him.

    Translation: ‘They have made him the enemy of himself.’

In the above example, the reflexive *nfsh* acts as the modifier of an object-complement *Edw* and the subject is the antecedent.

13. As an object of an adjective with the subject as an antecedent, for example (Holes 2002):

    هو مغرور جداً بنفسه

    Transliteration: /*hw mgrwr jdAF bnfsh*/

    Glossing: he arrogant very with-self-him.

Translation: ‘He’s very much taken with himself.’

In the above example, the reflexive *bnfsh* acts as the object of the adjective *jdAF* and the antecedent of the reflexive is the subject *hw*.

14. A modifier of an object with the subject as an antecedent, for example (Holes 2002):
    عليّ صور نفسه بالكاميرا
    Transliteration: / *ElY~i Swr nfsh bAlkAmyrA*/

    Glossing: Ali pictured self-him with-the-camera.

    Translation: ‘Ali has taken a picture of himself with the camera.’

In the above example, the reflexive *nfsh* functions as the modifier of the object *bAlkAmyrA and* Ali acts the reflexive antecedent.

15. An agent in passive/pseudo-passive/impersonal constructions with the subject as an antecedent, for example (Holes 2002):
    ما أحد خربها هي اتخربت من نفسها
    Transliteration: /*mA >Hd xrbhA hy Atxrbt mn*

    Glossing: no one corrupted-it she corrupted by

    nfshA/

    self-her.

    Translation: ‘No one corrupted her, she corrupted herself.’

In the above mentioned example, the reflexive *nfshA* stands to be an agent for the *hy* in the impersonal construction. The reflexive antecedent acts as the subject.

16. A modifier of such an agent, with the subject as an antecedent.
17. An element in another adpositional phrase or case-marked modifier with the subject as an antecedent, for example (Holes 2002):

    أكملت العمل بروحي/بنفسي

    Transliteration: />*kmlt* *AlEml* *brwHy/bnfsy*/

    Glossing: finished-I the-work with-myself/with-self-me

    Translation: 'I completed the work by myself.'

In the above mentioned example, the reflexive *brwHy/bnfsy* acts as a modifier for *AlEml* and the subject of the sentence is its antecedent.

18. As a modifier of such an element with the subject as an antecedent.
19. Other possibilities for the use of reflexives include their use within nominalized clauses where the reflexive is an indirect object (IO) or is in an adpositional phrase; for example (Holes 2002):

    ما تفعله بنفسك غير صحيح

    Transliteration: /*mA* *tfElh* *bnfsk* *gyr* *SHyH*/

    Glossing: what do-you with-self-you not right.

Translation: ‘ What you are doing for yourself is not considered to be right’

In the above mentioned example, the reflexive *bnfsk* acts as the IO for the verb *tfElh.*

The reflexive may appear as the direct object (DO) of a verbal noun whose subject is expressed as a pronominal enclitic. Here, an obligatory *li* must be inserted, which is part of the rule for forming complex NP formations.

Reflexive pronouns do not freely combine with other nouns to form construct NPs, although it is normal for *nafs* to appear in nomalized clauses; for example (Holes 2002):

علىِّ رفض تعيين نفسه مدير

Transliteration: /*ElY~i rfD tEyyn nfsh mdyr*/

Glossing: Ali refused appoint self-him boss.

Translation: ‘Ali refused to appoint himself as boss.’

2.10 Conclusion

In this chapter a definition of an anaphor is provided. The chapter discusses varieties of anaphora, types of antecedents, and relations between anaphora and antecedents. The chapter states the scope of the current algorithm which is Arabic pronominal anaphora. In stating the algorithm’s scope the chapter discusses types of pronominal anaphora is in MSA, means of expressing anaphora in MSA and restrictions when using anaphora in MSA.

The next chapter discusses anaphora resolution techniques, factors of anaphora resolution generally, and in MSA in particular.

Chapter 3. Anaphora Resolution

3.1 Introduction

Chapter 3 describes anaphora resolution factors in general. The discussion develops to discuss anaphora resolution constraints in MSA in particular.

To resolve an anaphor embedded in a given text is to make the connection between it and its antecedent in that text. Studies of how this connection is made can be divided into two types. Scientific anaphora resolution (henceforth AR) is an aspect of linguistics that aims to understand how the human language faculty resolves anaphora (Jurafsky and Martin 2000). In contrast, technological AR aims to develop algorithms for the resolution of anaphora in practical applications such as machine translation systems without any necessary reference to or implications for scientific AR (Jurafsky and Martin 2000). The present discussion is concerned with technological AR.

Although the nature of AR is easily stated, its implementation in practical natural language processing systems has turned out to be a difficult problem; many approaches have been developed, but none has thus far been entirely successful. This section briefly outlines the nature of the problem and solution factors proposed so far in terms of their accuracy, computational complexity, and knowledge requirements. This part of the discussion is necessarily focused on English because most of the work done on AR relates to this language; however, because the language of interest in this thesis is Arabic, a survey of existing Arabic AR is also included.

3.2 The Nature of the Problem

Given an anaphor, the AR problem is to identify its antecedent. For a human with native speaker competence this is usually unproblematic, but not invariably so. One example of a resolution that any human would find impossible on account of its inherent ambiguity is ‘Jenny put the cup on a plate and broke it’ (Mitkov 1999: 6), where the antecedent might be either ‘cup’ or ‘plate’ and there is no way of deciding which it is without some additional information. However, for an engineering AR system that lacks the innate grammar, semantics, logic, and real-world knowledge which together comprise native speaker competence, a correct identification can present varying degrees of difficulty. It has already been noted that the antecedent of an anaphor in a given sentence can be found anywhere from the same sentence or up to – according to current knowledge- the seventeen preceding sentences. The scope of this possible backward reference typically generates numerous candidates for the antecedent. The problem is how to choose the correct antecedent from among all the candidates, where correctness is determined by human judgement based on native speaker competence.

3.3 Existing Approaches to Anaphor Resolution

Modern anaphor resolution has a history in natural language processing research that goes back as far as the 1960s (Mitkov 2002). Since then various approaches to the problem have appeared in the literature. Though reasonable in principle, the amount of work to be covered makes a review of this literature an onerous undertaking in practice. Three simplifying conditions render it more tractable:

- Much of the literature is concerned with the identification of anaphora; that is, of determining whether, say, a pronoun or a noun phrase in a text is or is not an anaphor. None of this concerns the present discussion because its focus, *nafs*, is always an anaphor. This discussion can, in other words, assume that the anaphor of interest has been found and concentrate on ways of identifying the antecedent.

- Most of the AR systems in the literature are designed to deal with the range of types of anaphora listed in section 2.3 above. *Nafs* is, however, a pronominal anaphor, and as such the details of how these systems deal with types of anaphora other than pronominals are irrelevant to the present discussion.

- The survey is not exhaustive in the sense that it includes everything ever written on anaphor resolution. Instead the concentration is on recent work since 2000 (Poesio et al. 2010), while earlier work can be reviewed in Hirst (1981) and Mitkov (2002).

The following survey of the relevant AR literature begins by identifying and describing the various types of techniques used to resolve anaphora in the literature referred to for convenience as ‘anaphor resolution factors' (Mitkov 2002), and then goes on to show how these techniques are used by various researchers.

3.3.1 Anaphor Resolution Factors

Anaphora have been approached in many different ways, including from the perspectives of gender and number agreement, syntactic constraints, semantic consistency, centering, domain-specific and real-world knowledge, psycholinguistics, and mathematical and statistical models. These are discussed as follows.

i. Gender and Number Agreement

In both English and Arabic the pronominal anaphor must agree in number and gender with its antecedent. In 'Jane told the boys that she was leaving', for example, the third person feminine pronoun *she* agrees in gender and number with *Jane* but not with *boys*. In 'John went to university with Sarah in Newcastle and he worked in Durham', according to the gender and number matching rule, the noun phrase *John* is selected as the antecedent of the pronominal anaphor and the remaining candidates *Sarah*, *Newcastle*, and *Durham* are discounted on the basis of gender and number.

Gender agreement in English is a useful criterion when the candidates for the anaphora are:

- Proper masculine or feminine names such as 'Catherine', 'John', 'George'.

- Human being nouns such as ‘man’, ‘woman’, ‘son’, ‘daughter’, etc.

- Gendered animals such as ‘ox’, ‘chicken’, etc.

- Words such as ‘country’ or ‘school’, which can be referred to by either ‘she’ or ‘it’ but not ‘he’.

ii. Syntactic Constraints

The rules governing the syntax of the language of interest can be used to eliminate grammatically incorrect anaphor resolutions (Mitkov 1999). Syntax plays an important role in providing information about the clause and noun phrase boundaries. This then helps in the formation of the rules in the resolution process whereby unacceptable antecedents are eliminated. Some examples are as follows:

Reflexivization: in ‘Nadia says that Sue is knitting a sweater for her’ (Hirst 1981: 43), the antecedent of *her* must be *Nadia* or some other feminine but it cannot be *Sue* because, in English syntax, the reflexive *herself* would be used if Sue were the antecedent.

C-command constraints: play a vital role in discarding impossible candidates for antecedents of anaphors that are not of reflexive pronouns. They help in selecting antecedents of reflexive anaphors. C-command constraints are discussed by Mitkov (1999):

- A non-pronominal NP anaphor cannot overlap in reference with any NP that c-commands it. For example: in ‘He told him about John'. John appears as the object of a preposition which is c-commanded by the subject *he* and the direct object *him*. The pronouns *he, him* are disjoint from *John*.

- The antecedent of a bound anaphor must c-command it. For example in 'John likes pictures of <u>himself</u>', the underlined reflexive pronoun *himself* appears as a prepositional object, and the c-commanding subject *John* is a possible antecedent.

A personal pronoun cannot overlap in reference to an NP that c-commands it. In cases such as 'John told Bill about him', the pronoun under consideration here is *him*, which always appears in the position of an object or a prepositional object. The pronoun is disjoint to the c-commanding subject, which here is *John*, and the c-commanding object, which here is Bill.

- Preference is given to antecedents with the same syntactic function as their anaphora. Consider, for example:

  (a) 'The programmer successfully combined Prolog with C but he had combined it with Pascal last time.'

  (b) 'The programmer successfully combined Prolog with C but he had combined Pascal with it last time.'

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

This is part of syntactic parallelism, which can be helpful in the absence of other constraints or when such constraints or preferences are not able to resolve an ambiguous antecedent. Noun phrases have the same syntactic function as the anaphor. In (c) above, the anaphor *it* and its antecedent *the program* each have a different syntactic function; however *it* and *Prolog* have the same syntactic function.

iii. Semantic Consistency

The anaphor and its antecedent must be semantically consistent. For example:

> (a) 'Vincent removed the disk from the computer and then disconnected it'
>
> (b) 'Vincent removed the disk from the computer and then copied it'.

In (a) the antecedent of the anaphor 'it' must be 'computer' because computers can be disconnected whereas disks cannot; in (b) the antecedent must be 'disk' because disks can be copied but computers cannot (or at least not in the intended sense).

For example:

> (a) 'Vincent gave the disk to Sody. Kim also gave him a letter.'
>
> (b) 'Vincent gave the disk to Sody. He also gave Kim a letter.'

Preference is given to antecedents which share the same semantic category as their anaphora in order to establish a relation between the anaphoric noun and its potential antecedent. Semantic consistencies are based on the following criteria:

i. Number consistency, where the anaphoric expression and its antecedent must be consistent in number as singular or plural.

ii. Sort consistency, where the anaphoric expression sort must be either equal to or subsume the antecedent sort.

iii. Modifier consistency (Christodoulakis 2000), which is a factor allowing the incorporation of semantic constraints in parsing. This means that specific semantic features are to be added for each object meaning. Each feature is to denote parts of the universe to which the object belongs (Ferrández et al. 1998).

iv. Centering

Centering involves the identification of an antecedent candidate that is most salient with respect to the anaphor to be resolved. To exemplify this, Mitkov (1999) uses a sentence quoted above to illustrate the kind of ambiguity which prevents even humans from resolving an anaphor: 'Jenny put the cup on a plate and broke it'. The only way to select between 'cup' and 'plate' is to examine the textual context in which the sentence occurs. If, for example, the preceding text is all about cups with no reference to plates, 'cup' is more salient than 'plate' as an antecedent; it is the focus or centre of the discourse, and is thus preferred.

v. Domain-specific and Real-world Knowledge

Semantic consistency and centering typically require access to some representation of domain-specific and real-world knowledge in a format amenable to computational processing. With reference to the foregoing examples, the evaluation of these criteria requires system knowledge of such things as the characteristics and interrelationships of disks, computers, letters, cups, and plates in the real world.

vi. Mathematical and Statistical Criteria

This factor depends on collecting statistics from the corpus examined. A lot of research in this area depends on the work of Ge et al. (1998). The main procedure used in such research depends on the decomposition of a probability condition upon several features that depend on product conditional (Gasperin 2009). Statistical anaphora resolution is a branch of statistical NLP that relies on large corpora of training data to determine statistical relationships between words for the purpose of gauging the relationship between pronouns and antecedents in the absence of any higher level expert knowledge of the language.

In their landmark paper, ‘A Statistical Approach to Anaphora Resolution’ Ge, Hale, and Charniak (1998) describe a probabilistic architecture for considering written works and identifying the antecedents that the pronouns therein refer to. The algorithm that they present for doing so approximates the probability that a candidate antecedent is associated with a particular pronoun.

3.4 Anaphor Resolution Constraints in MSA

3.4.1. Arabic Diglossia

Ferguson (1959: 435) defines diglossia as ‘a relatively stable language situation in which, in addition to the primary dialects of the language, (which may include a standard or regional standards), there is a very divergent, highly codified (often more grammatically complex) superposed variety, the vehicle of a large and respected body of written literature, either of an earlier period or in another speech community, which is learned largely by formal education and is used for most written and formal spoken purposes but is not used by any sector of the community for ordinary conversation.’

Each of these varieties is used for a specific purpose that the user must be aware of. This is different from the case of a dialect that is used informally but where users would switch to the formal language when communicating formally.

Farghaly (2005) showed that Arabic has three language varieties that are used alongside each other in everyday life. Classical Arabic is used in religious discourse and daily prayers conducted by Muslims. MSA is used in formal communication and the media, and regional or colloquial dialects are used among friends and family. The factors that lead to the existence of such a unique situation include:

a. Suitability of purpose, where in some situations it is only appropriate to speak MSA whereas in other contexts the local dialect is used; for example, when speaking to friends and family members.

b. To demonstrate the social and educational level of the speaker, symbolized by the language he/she uses/speaks. This is the case, for example with MSA which is used to indicate a person's prestige. Classical Arabic, meanwhile, is related mainly to religion.

c. Each Arabic diglossia has its own literature and an audience who enjoys it.

d. The method by which the language is acquired. Classical Arabic and MSA are products of education which children start to learn when they go to school, while the local dialect is learned at home with no explicit grammar rules being taught.

e. For Classical Arabic there is an established grammar system, dictionaries, texts, etc. For MSA and colloquial dialects such grammar systems vary or may not exist, which makes it harder for non-Arabic speakers to learn them since all that is available to them is MSA which is hardly used outside academic classroom situations. To learn a colloquial style, there are hardly any formal sources that one can use.

For example, to briefly compare Classical Arabic and the Egyptian local dialect, the former has three end case marking suffixes which are completely absent in the latter. In Classical Arabic, and MSA as well, the main sentence structure is VSO, while in the Egyptian local dialect it is SVO. Wh-constructions are fronted in classical Arabic and MSA, while in the Egyptian local dialect they are not. The important question is therefore how NLP can deal with Arabic diglossia, which is discussed within the following section.

3.4.2. ANLP and Arabic Diglossia

In order for an NLP system to try to solve the problem of Arabic diglossia, it has to take into consideration several factors:

i. It will not be able to address all Arabic varieties in one application due to their differences in morphology, lexicon, and grammar; regardless of their common factors, they still differ a lot.

ii. It has to be aim-oriented, meaning that it should have a clear aim and be aware of the linguistic characteristics of the variety it is intended to deal with. It should also be accompanied by an understanding of the Arabic sociolinguistic situation.

Due to the above factors most of the tools developed so far are focused on written texts that are mainly written in MSA. However, some researchers, such as Habash and Rambow (2005), have tried to extract and categorize the grammatical features of a dialect and then apply it to MSA NLP tools. Another attempt was made by Shaalan and Abo Bakr (2007) to build up something similar to MSA Treebanks (Farghaly and Shaalan 2009). This is called Dialect Treebanks, and it was intended to transform Egyptian Arabic words into MSA via a lexical transfer approach. This approach changed Egyptian Arabic sentences from SVO into the MSA order VSO, as well as adapting Buckwalter's morphological analyzer to transform Egyptian Arabic words into MSA words.

There is a need to consider what Fargahly (2005) called 'inter-Arabic grammar' which would form a phase between classical Arabic, MSA and colloquial forms and allow the development of a line of research that would aim to explore intelligibility among all Arabic speakers. This would help enormously in addition to Dialect Treebanks in the development of ANLP tools.

What makes the problem of diglossia more complicated is that few resources are available. LDC has built an Egyptian, Levantine, and Iraqi Arabic corpora in order to try to solve this problem (Farghaly and Shaalan 2009). Columbia University are also trying to build a Dialect Treebank using MSA resources and mapping (Farghaly and Shaalan 2009).

3.4.3 Arabic Script

Arabic has no dedicated letters to represent short vowels. It also undergoes changes in the forms of letters due to their position in the word, and lacks capitalization and strict punctuation rules.

Due to the absence of dedicated letters to represent short vowels, diacritics have been used instead. Diacritics are marks that appear above and under the letter, but they are hardly in common use these days. It is difficult for ANLP to process texts correctly without diacritics which would indicate what is a verb and what is a noun. Non-native speakers also find it difficult to learn the language when these diacritics are absent.

The forms of Arabic letters change with their position in the word, although such changes are governed by rules and Arabic word-processors adhere to such rules, which makes it simple to select the correct shape. In order to choose the correct shape, each letter has only one key and the coding rules must be able to recognize the context and consequently the correct shape can be chosen. However, there are still problems with morphological processors. For example, the *hamza* letter undergoes changes during morphological and syntactic generations of an inflected word. So, the letter ي *y* indicates that something is mine, but when it is added to the irregular plural نساء *nsA’* which means ‘women’, it produces نسائي *nsA}y* ‘my-women’ instead of نساءي[4] *nsA'y*. Shaalan and Raza (2009) argue that Arabic, Chinese, Japanese and Korean scripts do not have capitalization or strict punctuation rules. Consequently, NLP applications such as machine translation, information retrieval, clustering, and classification tasks become more difficult as they are unable to split running text correctly into sentences as, for example, in the case of the English language or Latin script-based languages.

In Arabic, sentences are coordinated using the coordinators *wa*, and *fa*. In Arabic discourse it is common to use coordinators frequently and to write complete paragraphs without a single full stop. The lack of capitalization and strict punctuation rules makes the process of named entity recognition (NER) (Shaalan and Raza 2009) hard and the results are far from adequate, as well as complicating the process of information extraction (IE).

[4] This is not a correct form of a word.

3.4.3.1 Arabic Script Normalization

This problem arises due to the inconsistent use of diacritics and certain letters. Some Arabic letters have the same form and only differ in the addition of certain signs such as a dot, a *hamza*, or a *madda* above or below the letter. *Alif*, for instance, has three different forms depending on the position of the *hamza*, or *madda*. MSA tends not to use diacritics, and, as a consequence, most ANLP tools and systems normalize the text, as Larkey and Connell (2002) do. The Stanford Arabic Statistical Parser and The SYSTRAN Arabic-to-English machine translation system also incorporate normalization.

Normalization seems to solve the problem of letter recognition but, as Farghaly (2010) argued, it also increases problems of ambiguity.

3.4.4 NLP and Ambiguity in Arabic Texts

Arabic has many levels of ambiguity, as Attia (2008) and Farghaly and Shaalan (2009) show. Researchers developing the SYSTRAN Arabic-to-English machine translation system have found that ambiguity exists in Arabic at every level, as follows:

i. Homographs: words that have the same orthographic form but mean different things or belong to different syntactic categories.

ii. Internal word ambiguity: complex words could cause misinterpretation if not segmented correctly.

iii. Syntactic ambiguity: this arises when an internal analysis of the sentence is not available, especially with prepositional attachments.

iv. Semantic ambiguity: this arises due to the different possible interpretations of a sentence.

v. Constituent boundary ambiguity: different phrase boundaries may be established within the construction.

vi. Anaphoric ambiguity: which can arise from varying analyses of the deep structure of the sentence.

All of these factors, in addition to the nature of the Arabic language as a pro-drop language, its lack of capitalization, and strict punctuation rules, and complex word structure, make it very hard to be processed by NLP tools. The absence of short vowels is a further major factor in making Arabic hard to process, because without them no case markers can be assigned to word endings. Even parts of speech may give no clues as in the case of *mn* since it can be used as a preposition or as a wh-phrase.

Some researchers see tokenization as a solution to such problems, but because of the nature of the Arabic language it has been proven by Attia (2007) that such a process is very difficult and time consuming. This is because even a single word may have up to four tokens and therefore complex linguistic knowledge would be required to analyse it.

3.4.5 Arabic Morphology

Shaalan and Raza (2009) claimed that Arabic grammarians define the morpheme as a language word block that is meaningful. The roots stand for semantic fields while vocalism represents a grammatical case. This has led researchers such as McCarthy (1981) to suggest that Arabic words should be analysed as tiers, while Farghaly (1987) suggested a three-tier morphology. Early ANLP benefited linguistic research in Arabic since most of it was focused on morphological analysis, whereas much of the computational work in Arabic linguistics focuses on recovering the roots of Arabic words.

The Buckwalter Arabic Morphological Analyzer (BAMA) was developed in the 1980s and became commercially available in 2000. It consists of three tables: one for stems, one for prefixes and one for suffixes. It includes the constraints of adding prefixes and suffixes for words, and is widely used, including by non-Arabic developers, due to its bidirectional transliteration schema from Arabic to Latin script. It is a stem-based approach to Arabic morphology which helped in the development of ANLP systems and, later MT engines. BAMA provides users with access to Arabic roots, English glosses, and noun case endings. MADA (used in the current thesis) takes that work further by providing a disambiguation module to provide the correct POS tags in a natural text.

3.4.5.1 Systran's Stem-Based Morphological Generator

Developed by Farghaly and Senellart in 2003, this differentiates between two types of affix that can be added to an Arabic root. One is used to represent subject-verb agreement (which represent different parts of speech), while the other is produced by the morphological generator. It is considered to be an example for rule-governed affixes.

3.4.5.2 Morphological Processing and the Dialects

There is a need to develop ANLP tools to process Arabic dialects. The main barrier to this is that there is a lack of ANLP resources and there are no parallel MSA-dialect NLP resources. MAGEAD is a morphological analyser that was developed by Habash and Owen in 2005 in order to explore common points between MSA dialects. It still needs a lot of work as it operates without a lexicon, which causes a lot of problems, and phonological and orthographical representations still need to be developed to make the work more useful.

3.4.6 Arabic as a Pro-drop Language

Arabic has a complex word structure, which makes it a language where affixes and clitics represent parts of speech. In addition to the morphological nature of the language, this makes Arabic very hard to process. In MSA, a word may be analysed to constitute four parts of speech which, consequently, requires deep morphological analysis as well as tagging and tokenization. Attia (2007) suggests that words should be tokenized as a pre-processing task, especially since affix attachment is governed by syntactic rules. However, this is still not easy due to the ambiguity of the language (Attia 2008). Arabic allows subject pronouns to be deleted or to be freely dropped. This makes the NLP task difficult because a sentence must be understood as a native speaker would otherwise it will be analysed incorrectly.

Several researchers have tried to provide a solution by developing morphological analysers. In the 1990s Ken Beesley developed the

Xerox Arabic Morphological Analyzer which uses finite state technology to provide analysis and generation. Tim Buckwalter subsequently developed a morphological analyzer which uses a stem-based approach and it is used widely. The reason behind the wide usage of the Buckwalter's analyser is that it uses a single lexicon of all prefixes, short vowels and diacritics and a unified corresponding lexicon for suffixes (Sawalha and Atwell 2008). Other analyzers use numerous lexicons of prefixes and suffix morphemes which cause processing problems. An important factor as well is that it is available freely over the web while other analyzers are not.

3.4.7 Arabic Language Syntactic Structure

The main word order in MSA is VSO, although SVO is allowed in newspapers, for instance. All Arabic language variants allow subjectless sentences. To form a question, the wh-phrase is placed at the beginning of the question even though the Egyptian dialect does not do this. Arabic pronouns have a resumptive nature in order to refer to the relative clause head.

The agreement system in Arabic is quite complex, having twenty-four features compared to ten in the English language. A noun and its modifier have to agree in number, gender, and definiteness. In the SVO structure the verb and the subject must agree in number, gender and person. In other sentence structures such as VSO or OVS this is not the case, but the noun and its quantifier must agree in gender.

One of the problems that ANLP faces is that the grammar system of Classical Arabic is the one that is currently being adapted to MSA, but it cannot account for all the linguistic phenomena associated with MSA. This represents a problem of consistency, since an ANLP tool needs an established grammar which can be depended on in the analysis process, and it is especially important that the surface structure of MSA grammar can be used easily.

Badawi et al. (2004) provided a starting point in describing MSA grammar. Although this has not been computationally adapted, there have been attempts to build up Arabic corpora such as that provided by LDC. For analysis there is also the Prague Arabic Dependency Treebank, and the Arabic Treebank at Columbia University.

3.5 Conclusion

Chapter 3 discusses anaphora resolution techniques in general. The chapter shows how these techniques are used by various researchers. The chapter describes the problems of Arabic anaphora resolution and suggested resolution methods by various researchers.

The next chapter will survey approaches to anaphora resolution developed over the last forty years.

Chapter 4 Approaches to Anaphora Resolution

4.1 Introduction

Research in anaphora and anaphora resolution (AR), which is also known as coreference resolution, has instigated important developments in theoretical and computational linguistics. In the field of theoretical linguistics the dynamic models of language interpretation resulted from such research, while in computational linguistics various theories were developed to detect local and global salience. AR is closely related to information extraction, summarization and entity disambiguation.

The present chapter surveys approaches to anaphora resolution developed over the last forty years. The survey is divided into three main parts: the first deals with the linguistic and psycholinguistic background, the second covers the data driven approaches depending on annotated corpora, and the third deals with anaphora resolution in Arabic.

A definition of AR is needed to avoid misunderstanding. There are various definitions, but the one used throughout this discussion is adopted from the Message Understanding Initiative (MUC) and is used by various scholars, including: Aone and Bennett (1995), McCarthy and Lehnert (1995), Kehler (1997), Vieira and Poesio (2000), Soon et al. (2001), Ng and Cardie (2002b), Yang et al. (2003), Luo et al. (2004), and Hoste (2005). AR is ‘the task of identifying which parts of a text refer to the same discourse entity’ (Poesio et al. 2010: 1). The following example demonstrates this:

1. Sarah likes makeup, she buys lots of it but her choice of colours is horrible.

In the above mentioned example, *Sarah*, *she* and *her* refer to the same entity and so do *makeup* and *it*. In natural language (NL), AR is pervasive and is considered to be one of the major elements of semantic interpretation. This is why it has been studied in detail in linguistics, psycholinguistics and computational linguistics (CL) (Poesio et al. 2010).

4.2 The Linguistics of Anaphora Resolution

4.2.1 Context Dependence

The interpretation of noun phrases (NPs) depends on the surrounding context, specifically on the linguistic context entities which have been previously mentioned. Pronoun interpretation, in particular, depends entirely on linguistic context entities. Also, NPs and nouns may depend for their interpretation on visual context. This is classified by Clark and Marshall (1981) in terms of visual deixis; that is, the discourse situation which includes the linguistic context and its surroundings and participants. According to Kamp and Reyle (1993), the set of entities introduced in the discourse situation are called "*U*" (the Universe of Discourse). The main focus of such theory (DRT) (Discourse Representation Theory) is to explain how natural language utterances are context dependent, where the meaning of an utterance depends on its context. In addition, it should be noted that there is a reciprocal interaction between the context and the utterance. In general, the domain of interpretation controls the interpretation of a given noun phrase, depending on shared knowledge of the topic being discussed.

For instance, proper noun interpretation is domain-dependent because proper nouns refer directly to constants/objects which are encoded in their semantics. It would be inappropriate if the interpretation domain of proper nouns did not specifically identify the targeted object. This makes the process of interpreting proper nouns completely different from that of pronouns and nominals (Poesio et al. 2010). It has to be taken into consideration that, due to advances in CL research work, the process of disambiguating direct references to the domain of interpretation is now considered to be easier. For example, Wikipedia makes use of objects' identifiers which consequently facilitates direct reference disambiguation which is domain-dependent. In addition, CL identifying systems can link named entities indirectly by proper noun referencing; all noun interpretation systems, however, still use the context-modifying effect of proper nouns to provide pronoun and nominal antecedents.

The choice of domain of interpretation has an effect on the nominal's quantification domain, which is 'the set of objects of the type specified by the nominal complex which are included in the domain of interpretation' (Cooper 1996: 70). The quantification domain can be identified through the linguistic context as well.

4.2.2 Types of Context-dependent Expression

It is important to point out that nominals are not the only kind of expression whose interpretation depends on the domain of discourse. For example, pronouns with a verbal interpretation domain that can be considered as analogues, as well as ellipsis, also depend for their interpretation on the domain of discourse. As pronouns are characterized among nominals by being context-dependent, full verbal expressions have a context-dependent component that is pragmatically determined by the discourse (Kamp and Reyle 1993).

The study of ellipsis received much attention during the early years of CL, but currently the focus is on the use of corpus-based studies to interpret anaphoric expressions. The reason for this shift of interest is due to the lack of annotated resources. In theoretical linguistics nominal expressions have four semantic functions, which are (Poesio et al. 2010):

- Referring, which is concerned with noun phrases that introduce new entities in the discourse, or refer to previously introduced entities.

- Quantification, which expresses the relations between the objects that are denoted by the nominal complex and objects denoted by the verbal phrase.

- Predication, which expresses the properties of objects. For example, in *Omar is a journalist*, the noun phrase *a journalist* expresses a property of Omar.

- Expletives, used with verbal arguments in syntactic constructions. In most cases these are semantically vacuous as with *it* and *there;* for example: <u>It</u> is hot.

It not an easy task even for humans to draw clear distinctions between these functions, as noted by Poesio et al. (1998). Everything depends on one's theoretical assumptions; for instance, to consider whether a noun phrase is referring or quantificational. In some theories all nominals are quantifiers while in others definites and indefinites are not considered to be nominals. For instance, van Deemter and Kibble (2000) argue that the MUC annotation scheme treats the NPs of copular clauses and appositions as referential, which is considered to be problematic. In contrast, many linguistic theories assume that NPs of copular clauses and appositions are considered to be predictive which may not always be the case. It should be noted that predicative noun phrases are independent of the universe of discourse, *U*, while other types of nominals can depend on context. In the current research, predicative NPs, unlike other types of nominal phrases, are less dependent on the universe of discourse. Predicative NPs are considered vital in the current thesis since many types of NPs can be used either referentially or predicatively. The domain of quantificational NPs is contextually specific. In the current thesis the focus is on referring expressions and on the process of selecting the antecedent they are associated with.

Referring noun phrases have various forms which vary according to the rules governing their anaphoric behaviour, as stated in Reinhart (1976), Chomsky (1981), Gundel et al. (1993), Garrod (1993) and Garnham (2001). Varieties of referring noun phrases include:

i- Reflexives (known in Binding theory as anaphors), for example: Omar hurt himself.

ii- Pronouns, which are subdivided into :

a) Definite pronouns

Ross bought a {a radiometer/ three kilograms of after-dinner mints} and gave {it/them} to Nadia for her birthday.

(Hirst 1981)

b) Indefinite pronouns

Kim bought a t-shirt so Robin decided to buy one as well. (Webber 1979)

c) Demonstrative pronouns

Can you give me that cup on the table over there?

iii- Nominals, which are NPs with a noun as a head such as a girl or a boy. *A boy* and *a girl* walked together. The boy wore a blue t-shirt.

iv- Proper names

Omar and *Aly* in; Omar and Aly are good students.

Kaplan (1977) argued that proper names directly refer to pronouns or nominals as well as demonstratives rather than referring to an entity introduced in the linguistic context. In linguistics, differences between reflexives and personal pronouns have been intensively studied and discussed. Such differences were researched in depth in terms of generative syntax, which resulted in a whole new Chomskyan paradigm called 'Government and Binding' (Reinhart 1976; and Chomsky 1981). For example:

2. Omar considered <u>himself</u> lucky to play with <u>him</u>.

In example (2) *himself* must corefer with *Omar* but *him* cannot.

The factors affecting the choice of multiple linguistic forms were researched by Ariel (1990), Almor (1999) and Poesio (2000), in order to study in detail the differences between personal and demonstrative pronouns. Linde (1979) and Passonneau (1993) used corpus data to search for such differences, whereas Garrod (1993) studied the differences between definites and pronouns and between definites and proper names.

Poesio et al. (2010) argue that referring expressions have no constant referring form or constant context dependence. Expletives, as discussed earlier, are a clear example that even pronouns can sometimes be non-referring.

4.2.3 The Relation of Referring Expressions to their Context

Referring expressions introduce discourse-new entities (i.e. entities that have not been mentioned before), which differ from expressions referring to discourse-old entities (i.e. those already mentioned). Poesio et al. (2010) argued that discourse-new entities can be differentiated as expressions that are completely new to the hearer and entities that the hearer is expected to know which can be called hearer-old.

When discourse-new entities are related indirectly to the linguistic context, they are considered to be anaphoric. For example the indefinite pronouns *one* and *another* have identity of sense relations with their antecedents as they refer to a different object of the same type. For example:

3. Omar liked Aly's suit, so he bought one for his wedding.

Paycheck pronouns, which are definite pronouns used in the same way as the above mentioned indefinite pronouns, are used similarly; for example:

4. The man who gave his paycheck to his wife is wiser than the man who gave it to his mistress. (Hirst 1981)

Bound anaphora occur when the antecedent is a quantified expression. The relationship between the pronoun and the antecedent can be described as a variable in a procedure. The variable is repeatedly called over elements under the restriction of the quantifier. In such a case the antecedent and the pronoun have no identity relation (Poesio et al.

2010). This can be readily identified when the quantifier is entailing, for example:

5. No kid ever believes that Santa got him the right toy.

An associative anaphor occurs when the context-dependent nominal is related to its antecedent by a part-of relation. It requires a bridging inference in order to identify the antecedent, for example:

6. The university buildings are nice. The labs are tidy but the toilets are dirty.

Creating clear distinctions between discourse-old and discourse-new expressions is not easy. Poesio at al. (1998) argued that readers can distinguish between them, but there is no agreement about the distinctions. Poesio at al. (1998) and Poesio et al. (2005) argued that, even when an expression is anaphorically related, it is still hard to define the antecedent and declare what kind of relationship there is between anaphor and antecedent.

7. We saw a flat yesterday. The kitchen is very spacious but the garden is very small.

(Vieira 1998)

4.2.4 Discourse Models

The development of discourse models by Karttunen (1976), Heim (1982), and Garnham (2001) has made the relationship between the context and anaphora more specific. These authors argue that the interpretation of context-dependent expressions is carried out with respect to a dynamically built-up discourse model. The interpretation is carried out while the discourse is being processed, including objects that are being mentioned in *U* (the universe of the discourse, as mentioned above). The importance of the discourse model hypothesis arises from its assertion that:

a) The context on which an utterance is dependent for interpretation is always updated. The updating potential itself also needs to be modelled.

b) Objects included in *U* are not restricted to those explicitly mentioned. They may include objects that can be inferred or constructed from explicitly mentioned objects. Those explicitly mentioned objects can be used as antecedents of sets of objects, or prepositions and abstract objects. Grosz (1977) called these implicitly mentioned objects as the 'implicit focus' of discourse (Poesio et al. 2010).

Karttunen (1976) originally formulated the idea of a discourse model hypothesis. Sanford and Garrod (1981) and Garnham (2001) developed it further in psycholinguistics. Kamp (1981) and Heim (1982) developed it more formally in theoretical linguistics and Webber (1979) applied it to computational linguistics. Kamp (1981) and Heim (1982, 1983) called their framework 'Discourse Representation Theory' (DRT), which deals with the semantics of anaphora and is used as a basis for the linguistic treatment of anaphora.

The main contributions of the dynamic theories of anaphora are:

- The ability to demonstrate the discourse model constructions in a formal way.
- The production of resulting interpretation semantics that can be used to interpret other semantic phenomena.

The discourse model construction is considered to be highly idiosyncratic (Poesio et al. 2010) but when combined with formal semantics it leads to the development of discourse model construction theory (Groenendijk and Stokhof 1991; Muskens 1996). Discourse model construction approaches revolve around the idea of the card file. Heim (1983) described this as a collection of cards, each of which introduces information about a new discourse entity that is introduced in the discourse. Recent versions of DRT interpret referring expressions as follows (Poesio et al. 2010): ‘Indefinite (a P, some P): a new file card x_i is added to the discourse model and asserted to be of type p. This update is formally written as $[x_i, |\mathrm{p}(x_i)]$.

- Proper nouns: as a result of a reference to object b via a proper name, a new file card x_i is added to the discourse model and asserted to be identical with b. This update is formally written $[x_i, |x_i = b]$.
- Pronouns: a new file card x_i is added to the discourse model and noted as needing resolution via the condition x_i =?. This update is

formally written $[x_i, |x_i =?]$. Resolution leads to this condition being replaced with equality with the file card of the anchor.

- Definite nominal (the P, that P): this is a type of referring expression about which there is the least agreement. Most researchers propose that definite descriptions have a uniqueness presupposition: the existence of an object of type P is presupposed instead of asserted, and furthermore this object is meant to be unique (Barker 1991; Roberts 2003). The semantics can be translated as follows: a new file card x_i is added to the discourse model and asserted to be identical with the unique object of type p (in the context). This update is formally written $[x_i, |x_i = \iota y.\mathrm{p}(y)]$.’

In the 1980s and 90s work on anaphora resolution depended on the notion of file cards or discourse entities (Poesio and Kabadjov 2004). Later on, single anaphor antecedent links were the predominant notion in anaphora resolution but currently the former idea is being revived.

The crucial character of DRT is that it provides logical representations that have their own truth conditions. Logical representations are different, but in the meantime equivalent to first-order logic, which consequently allow inferences to be made. As many cases of anaphora resolution require complex inference, the use of a deductive system for such representations is crucial (Poesio et al. 2010).

DRT is used for a range of anaphoric phenomena to reference events, plurals or abstract objects as prepositions; for example:

8. Omar saw Ahmed. <u>That</u> happened at 4 o’clock.

9. Omar met Ahmed. <u>They</u> had gone to the cinema together.

10. Omar saw Ahmed. <u>This incident</u> made him look pale…

In contrast, Kamp and Reyle (1993) based their analysis of plurals on the resolution of references via bridging inferences which enlarge the discourse model with new objects. However, prepositional references require the introduction of new prepositional variables by making inferences on the discourse model.

Based on encoding the results of rich inference, mental models can be formalized. Mental models (instead of discourse models) are based on the work of Bransford et al. (1972) and Garnham (2001). Such models deal with the results of rich inferences, making them very different from language models introduced in computational and theoretical linguistics.

4.2.5 Statistics About Anaphora from Corpora

To obtain a quantitative estimate of the types of nominal anaphoric phenomena and their importance, anaphorically annotated corpora have been developed. Anaphora and degree of anaphoricity in written formats have been studied by various scholars whose work is discussed in what follows.

Various studies discuss pronouns, definites and proper names as types of anaphoric expressions. Studies focusing on the anaphoricity of pronouns (or its lack) and relevant statistics have shown the following. Evans (2001) analysed the SUSANNE and BNC corpora and obtained

3171 examples of *it*. It was found that 67.9% of cases were examples of nominal anaphoric relations, while 26.8% were examples of expletives, 2.2% of idioms, 2% of discourse topic mentions, 0.8% of clause anaphors and 0.1% cataphors. Similar results were reported by Boyd et al. (2005) using the British National Corpus (BNC). Muller's (2008) study is considered to be the most comprehensive study concerning the distribution of the third-person pronouns *it*, *this,* and *that*.

Kabadjov's (2007) study of the relative frequency of nominal types used the GNOME corpus, and the Vieira-Poesio corpus showed that the most frequent NPs used were: bare-np, the-np, the-pn, pers-pro, pn and a-np. The anaphoric relations were mainly (56%) identity relations, and the other 44% were bridging relations.

Passonneau (1993), Byron (2002) and Gundel et al. (2002) studied the referents of pronouns and their distribution and whether they were introduced directly or indirectly. Byron reported that 16% of pronouns in the corpus had non-NP antecedents. Gundel et al. (2002) reported that 16% of the sample antecedents had no NP antecedents. Poesio and Vieira (1998) carried out a study of the definite descriptions used in the first mention compared to the anaphorically used ones. The results showed that around 50% of the definite descriptions were first mention, around 40% were anaphoric, and the rest were bridging.

4.3 The Interpretation of Anaphoric Expressions in Corpora and Psycholinguistics

Resolving anaphoric expressions demands the use of a combination of types of information. For example, gender is considered to be one of the strongest resolving factors. Syntactic constraints, common sense, and other factors act as preferences rather than constraints (Poesio et al. 2010). In the development of computational models of anaphora resolution the differentiation between constraints and preferences plays an important role, as standard expositions such as Mitkov's (2002) have argued. Poesio et al. (2010), however, argued that there is no conclusive evidence about the existence of two distinct mechanisms. In what follows, resolution constraints and preferences are discussed as well as the psychological evidence that supports their importance.

4.3.1 Constraints

Much early work on anaphora resolution depended on the identification of morphological and syntactic constraints. Agreement constraints (syntactic and semantic) and binding are the best known forms of constraint. Types of constraint can be summarized as:

a) Agreement/morphological constraints: These include gender, number and person constraints. Psychological studies such as those by Garnham et al. (1995) and Arnold et al. (2000) have shown that gender helps in anaphora resolution. The differences in gender use in semantic gender languages such as English or syntactic gender languages such as Italian or Spanish are used at an early stage to disambiguate anaphoric relations. The majority of modern anaphora resolution systems tend to incorporate agreement constraints, where the problem with gender as a constraint is consistency in witness cases; for example :

11. To get a customer's 110 parcel-a-week load to its doorstep (Poesio et al. 2010)

The error in the above mentioned example is due to the erroneous use of the pronoun *it*.

Errors may occur when pronouns refer to entities that are to be referred to using uncommon proper names, for example:

12. a. Maja arrived to the airport. (Maja is a man) <u>He</u>…

b. John brought Maja to the airport. (Maja is a small dog) <u>It</u>…

This problem was partially addressed by Ge et al. (1998) and Bergsma (2005), who attempted to infer the gender of unknown names; generally however, the gender can be inferred from context.

As for the use of number as a constraint, there have not been many studies of its use in anaphora resolution. There have, however, been studies (for example, Gordon et al. 1999) which compare the difficulty of anaphora resolution using plural and singular references. Clifton and Ferreira (1987) showed that the plural pronoun *they* is easily interpreted when it occurs after a conjoined noun phrase as in 'Ahmed and Omar' rather than when it occurs after syntactically divided antecedents as in 'Ahmed met Omar'. This suggests that the antecedents of plural pronouns are to be found in a discourse model rather than in a syntactic representation. The main problems with numbers in computational linguistics occur due to nouns which are syntactically singular but semantically plural, as in the case of *the government*. This is shown in the example below:

13. The government said that they will not allow immigrants to come into the country unless truly needed.

b) Syntactic constraints: anaphoric reference constraints are important in generative linguistics to the extent that the best-known paradigms are named after them: Government and Binding (GB) theory (Chomsky, 1981). The aim of GB theory is to explain why *her* in (14a) cannot corefer with *Nagwa* whereas *herself* must obligatorily be referring to *Nagwa* in (14b).

14. a. Nagwa loves her.

b. Nagwa loves herself.

Based on the relation between nodes in a syntactic tree, Langacker (1969) called this a ‘command’. Lasnik (1976) and Reinhart (1976) provided a definition of the c-command relation as follows:

Definition 1 Node A c-commands node B if

1. A≠B

2. A does not dominate B and B does not dominate A, and

3. Every X that dominates A also dominates B.

The c-command relation symbolizes the core of what is now called binding theory, which revolves around three main principles. Principle A deals with constraints imposed on reflexives and reciprocals. It states that ‘reflexives and reciprocals must have a c-commanding antecedent in their governing category; that is the smallest clause or noun phrase in which they are included.’ Principle B states that ‘pronouns cannot have an antecedent in this governing category.’ Both principles A and B claim that the distributions of pronouns and reflexives are complementary. Principle C states that ‘R-expressions as proper names and nominal cannot have c-commanding antecedents.’

GB theory underwent considerable development in order to overcome the limitations of the 1981 version; and in 1986 Chomsky introduced the alternative notion of the m-command. In 1994, Pollard and Sag introduced the alternative definition of the c-command, which is based on argument structure rather than phrase structure. These proposals were trying to account for picture NPs; for example:

> 15. John was going to get even with Mary. That picture of <u>himself</u> in the paper would really annoy her, as would the other stunts he had planned. (Poesio et al. 2010)

In 1993, Reinhart and Reuland proposed a major development of GB theory. They proposed that some reflexives are logophors, and thus have discourse-antecedents; for example:

> 16. Bill told us that Elisabeth had invited Charles and himself.

Over the years there have been numerous experimental tests of binding constraints. For instance, Nicol and Swinney (1989) proposed using priming techniques, while Gordon and Hendrick's (1997) results supported Principles A and B of binding theory while little support was found for Principle C. Runner et al.'s (2003) study showed that many reflexives behave as logophors when they are found in picture NPs.

c) Semantic constraints might also be called scope constraints. Karttunen (1976) argued that semantic constraints prevent anaphoric reference to introduced antecedents existing in downward-entailing operators. In recent psycholinguistics studies semantic constraints have gained importance as event-related potentials (ERP) in experiments using anaphoric reference as an example of violation-effects.

4.3.2 Preferences

Constraints cannot stand as the main and only factor that eliminates anaphoric ambiguity. Much research has been carried out in order to determine the factors which affect preferences among interpretations. Such factors are discussed in what follows.

a) Commonsense knowledge: this includes plausibility as a main factor. Sidner (1979) reported examples of the effect of plausibility. Implicit causality effects are one type of plausibility that has been studied extensively, for example studies by Garvey and Caramazza (1974), Stevenson et al. (1994), and Kehler et al. (2008) who discussed various relevant issues. The Garvey and Carmazza study showed that, when a sentence needs to be completed as in (17), it tends to continue in a consistent manner that matches *he* being Bill, in order to explain why Bill is to be blamed.

17. John blamed Bill because he… (Poesio et al. 2010)

Stevenson et al. (1994) showed that such preferences are affected by the verb thematic structure, where agent-patient verbs behave in a different manner than experience-stimulus verbs.

Kehler et al. (2008) showed that discourses which have one semantically coherent interpretation tend to choose that interpretation and ignore any other salient factors in the meantime. If both possible interpretations scored equal in terms of plausibility, the choice of an interpretation would then depend on general salience.

Selectional restrictions are another form of preference carried out with verbs, where a restriction is imposed on the type of argument the verb may have. Mitkov (2002) showed such an effect using minimal pairs.

Due to such studies as the ones mentioned above, anaphora resolution models focused on theories of commonsense reasoning such as Wilks (1975), and Hobbs et al. (1993). One can, however, argue that commonsense was not the only factor, and that other factors are at play as well.

b) Syntactic preferences: corpus statistics show that 60-70% of English pronouns occur in the subject position and about 70% of those have an antecedent that also occupies the position of a subject. This kind of relation and preference is called subject assignment and has been the focus of various psycholinguistics studies such as those by Broadbent (1973) and Crawley et al. (1990).

Preference for object pronouns referring to antecedents in the object position was studied for example by Kameyama (1985). Smyth's (1994) results suggested that, whenever the syntactic function is closer, the greater the effect it has; while Stevenson et al.'s (1995) results implied a similar effect but subject pronouns had a stronger effect than object pronouns. These researchers, among others, have hypothesized that parallelism is semantic rather than syntactic, an idea which Hobbs and Kehler (1997) developed.

c) Salience: With its simplest form as recency, this plays an important role in anaphora resolution. In Hobbs' corpus (1978), it was found that 90% of pronoun antecedents existed in the same sentence, while 98% existed in the previous sentence. In every referential distance study, the importance of the existence of antecedents in the same sentence has been highlighted regardless of reported frequencies. Givon's (1992) study proposed that 25% of definite antecedents were in the same clause while 60% of the definite antecedents existed in the previous 20 clauses and the rest were further away. This study, as well as others, showed that distance is not important in the resolution of other anaphoric expressions.

Studies such as Tetreault's (2001) have argued that choosing the nearest possible antecedents would lead to only moderate success. However, there are other studies, such as Gordon et al. (1993), which argue that the first mention advantage is the best choice.

In contrast to the above mentioned contradictory results, there has been a strong claim that differences between salience entities have an effect on the interpretation of anaphoric expressions. Linde (1979) and Sanford and Garrod (1981) carried out various studies to show that linguistic focus has a vital role in the anaphora resolution process as well, while Gundel et al. (1993) argued that it has an effect on the production and choice of the form of the referring expression.

In 1986, Grosz and Sidner proposed a framework with two levels: global and local focus. Global focus is concerned with identifying the articulation of discourse into segments, while local focus is concerned with identifying how the relative salience of utterances changes utterance by utterance. Discourses are classified by topics or episodic organization, as in Anderson et al. (1983). Grosz and Sidner (1986) added to this idea another factor: that this classification is hierarchical and dependent on the intentional structure of discourse. In addition, they proposed that global focus is stacked, while Walker (1998) argued for a cache model. Knott et al. (2001) argued that Grosz and Sidner's model was suitable only for task-oriented dialogue.

As for local focus, various researchers such as Grosz and Sidner (1986) and Sanford and Garrod (1981) argued that in every conversation or readable text there are some entities which are more salient than others. This makes some antecedents preferred for pronominalization, while others are preferred for anaphoric reference. Sidner (1979) argued that local focus can be verified according to two types of focus: discourse focus and actor focus.

Discourse focus, according to Reinhart (1981) and Vallduvi (1993), can be explained in terms of the notion of discourse topic; while Sidner argues that actor focus gains its effect through subject assignment. Complex algorithms would thus be needed to detect both types of foci as the focus may change after each sentence.

Grosz et al.'s (1995) centering hypothesis appeared in reaction to Sidner's theory. It soon became a theory in its own right and a main paradigm for the understanding of salience in computational linguistics, psycholinguistics and corpus linguistics. Centring theory argues that every utterance increases and updates the local focus, which is achieved via the introduction of new forward-looking centre. Each new forward-looking centre updates the focal structure and is ranked, which gives each utterance a most highly ranked entity called the preferred centre (CF), which is similar to Sidner's actor focus. The object which acts as the discourse topic is called the backward-looking centre.

Various researchers have tried to verify the applicability of this hypothesis. For instance, Poesio et al. (2004) carried out a corpus-based study which revealed that the degree of entity coherence between utterances is much less than that predicted where the majority of the utterances have no CB (backward-looking centre). Gundel et al. (1993) argued that there are factors which affect the choice of NPs in the salience theory as well as in the centring theory. Among such factors is the cognitive status of the referred entities. Gundel et al. also identified the lexical acquaintance levels of 'givenness' including: 'in focus', activated, familiar and lexical acquaintance levels. Gundel et al. (1993) provided definitions of their terms as follows:

In focus: the addressee can associate with the entity a unique representation that is in the current focus of attention. For example, I couldn’t sleep last night. It kept me awake.

Activated: The addressee can associate with the entity a unique representation that is in current working memory. This includes speech participation as well as other entities in the immediate discourse context. For example, I couldn’t sleep last night. That kept me awake.

Familiar: The addressee can associate with the entity a unique representation that is somewhere in the memory, perhaps long-term memory. For example, I couldn’t sleep last night. That dog (next door) kept me awake.

The ‘in focus’ level is related to the notions of CB and CP (preferred centre) but it can have more than one entity in focus or it can have no entity at all in focus. Activation level, however, is nearly equivalent to Grosz and Sidner’s implicit focus. Activation models have been examined by researchers such as Alshawi (1987), Leass and Lappin (1994), Strube (1998) and Tetreault (2001). There have been models that integrate salience and commonsense knowledge, such as in Carter (1987). In psychology, Gordon and Scearce (1995) studied the interaction of centering theory with commonsense preferences, and revealed that pronouns are to be interpreted according to centring theory rules before commonsense rules are applied.

4.4 Early Computational Models

Many computational models of anaphora resolution were developed in the 1980s and 1990s. These attempted to implement the syntactic, commonsense, and discourse theories discussed in the previous section.

The main differences between the theoretical assumptions in these models were that some regarded the process of anaphora resolution as entirely a commonsense matter, while others regarded it as a purely syntactic informational matter. In addition, the importance of level of formality is a significant difference between such models as some are linguistically and formally based while others are pragmatically based. However, the models shared the following characteristics:

i. ‘No large scale evaluation was attempted: the models were either purely theoretical, or the implementation was a proof of concept’ (Poesio et al. 2010: 28), and

ii. ‘Development was guided near-exclusively by the researcher’s own intuitions, rather by annotated texts from the targeted domain.’ (Poesio et al. 2010: 28)

The next sections review the development of anaphora resolution models and how researchers have tried to overcome all of the early limitations.

4.4.1 Syntax-based Models and the Hobbs Algorithm

The previous section described the role that information about syntactic role such as constraints, preferences, commonsense knowledge, and salience plays in types of filtering interpretation (gender and binding constraints) and defining preferred interpretations (subject assignment, and parallelism). Different algorithms have been developed to incorporate such information in anaphora resolution.

One of the best-known syntax-based algorithms was proposed by Hobbs (1978) using pronoun resolution. Hobbs' algorithm is still used as a baseline, which are a set of reference algorithms for pronoun resolution, which is unsophisticated and domain-independent. Until the development of Soon et al.'s algorithm, Hobbs' naïve algorithm was considered to be the standard baseline, as it goes beyond the surface parse tree breadth. To look for an antecedent that matches the pronoun in gender and number, it goes back one sentence at a time. The algorithm makes use of binding theory by applying syntactic constraints and preferences, specifically the use of subject and preference for first-mentioned entities. The algorithm makes sure not to choose an antecedent NP that lies within the same binding domain as the pronoun, and also establishes a relation/node between the top node and any candidate.

Table 4.1: Hobbs' algorithm (Poesio et al. 2010)

Hobbs' Algorithm

1: Begin at the NP node immediately dominating the pronoun.

2: Go up the tree to the first NP or S node encountered. Call this node X, and call the path used to reach it p.

```
3: Traverse all branches below node X to the left of path p in a left-to-right, breadth-first fashion. Propose as the antecedent any NP node that is encountered which has an NP or S node between it and X.
4:if node X is the highest node in the sentence then
5:     traverse the surface parse trees of previous sentences in the text in order of recency, the most recent first; each tree is traversed in a left-to-right, breadth-first manner, and when an NP is encountered, it is proposed as antecedent
6: else
7:     (X is not the highest node in the sentence) continue to step 9.
8: end if
9: From node X, go up the tree to the first NP or S node encountered. Call this new node X, and call the path traversed to reach it p.
10: if X is an NP node and if the path p to X did not pass through the N node that X immediately dominates then
11:     propose X as the antecedent
12: end if
13: Traverse all branches below node X to the left of path p in a left-to-right, breadth-first manner. Propose any NP node encountered as the antecedent.
14: if X is an S node then
15:     traverse all branches of node X to the right of path p in a left-to-right, breadth-first manner, but do not go below any NP or S node encountered.
16:     Propose any NP node encountered as the antecedent.
17: end if
18: Go to step 4
```

An important feature of Hobbs' work is that he was the first researcher to attempt a formal evaluation of his algorithm. He evaluated it manually using 100 pronoun examples extracted from three different genres (a history book, a novel, and a news article). It scored an accuracy rate of 88.3%. After the addition of selection restrictions, the algorithm scored 91.7% accuracy. Several researchers have tried to apply large-scale evaluations using syntactically hand-annotated corpora; the results indicated improvement in the pre-seen results. Lappin and Leass (1994) tried to implement the algorithm using 360 pronouns extracted from a corpus of computer manuals and reported over 82% accuracy. Tetreault (2001) used Ge's et al.'s news text corpus extracted from the Penn Treebank and reported a 76.8% accuracy rate compared to 80.1% for fictional texts.

4.4.2 Commonsense Knowledge: Charniak, Wilks, Hobbs' Abductive Model

Charniak (1972) Winograd (1972) and Wilks (1975) were among the pioneers to carry out research concerning the effect of commonsense knowledge on computational models of anaphora resolution. Between the mid-1970s and mid-90s such research flourished and researchers such as Carter (1987), Alshawi (1992) and Gardent and Konrad (2000) labelled this the knowledge-based years of artificial intelligence (AI). Some of these studies, such as Charniak's (1972), argued that there is no need to use syntactic information to carry out anaphora resolution. Charniak's study was based on the frame theory of commonsense knowledge developed by Minsky (1975). Alshawi (1987) initiated the trend of anaphora resolution using frame and semantic network information. Later on, Poesio et al. (1997) and Harabagiu and Moldovan (1998) developed WordNet, which is widely used in anaphora resolution.

Wilks (1975) developed a semantic interpretation theory which was applied to anaphora resolution. Wilks' semantic interpretation theory revolved around preference semantics, however, semantics played only a limited role in the process. Wilks specified all meanings in nearly 70 primitive semantic units, such as entities and actions. To resolve the ambiguity of a targeted sentence, the interpretation which satisfies the greatest number of preferences is the one to be chosen. To fill any gaps, commonsense reasoning and specific casual reasoning is used.

Between 1975 and 1995 commonsense inference was widely studied and used in anaphora resolution and it resulted in formal frameworks for inference. Researchers such as Hobbs et al. (1993), Asher and Lascarides (1998), Gardent and Konrad (2000), and the SRI Cambridge group who developed the Core Language Engine (Alshawi 1992) developed systems that can be used in real-world applications.

Hobbs used abduction as a basis for a theory of semantic interpretation. Abduction is 'reasoning from effects to (the most plausible) causes: e.g., to conclude a friend must have woken up late in order to explain the observable fact that he hasn't showed up in time to go jogging in the morning' (Poesio et al. 2010: 31). Abduction was used to interpret problems such as noun-noun compounds for example, chessboard and woodboard, or word sense disambiguation and anaphora resolution. In abduction theories, in order to understand a discourse an explaining bond between the first utterance and the second utterance is to be established. If such a bond is not easily explained or detected, the juxtaposition is regarded to be felicitous. The explanation includes an assumption that the second utterance is the reason for the first one in what can be called a reason-rhetorical relation. The antecedents are chosen depending on the lowest cost explanation, as each assumption has a cost (Poesio et al. 2010).

4.4.3 Salience: Discrete and Activation-based Models

The salience work discussed earlier formed the basis for computational models incorporating theories of salience.

i. Sidner’s algorithm is considered to be the best developed model for anaphora resolution using salience, although it was never subjected to substantial evaluation, which leaves its accuracy rate unclear. The two main structural components of Sidner’s algorithm are:

- The organization of entities in a semantic network inspired by the work of Charniak, and
- Building data structures to keep track of which entities are currently most in focus. This aspect of the theory is the one which has had the greatest influence on subsequent research, in particular on the development of the Centering theory (Poesio et al. 2010: 32).

The three main data structures in Sidner's theory are: discourse focus; actor focus; and lists of previous discourse foci, actor foci, and sentence foci.

'Discourse focus: is introduced by special syntactic constructions or by serving as theme (in the thematic role sense) of a sentence. Agents of sentences serve as preferred antecedents for pronouns that also fill the agent role' (Sidner 1979: 50).

'Actor focus: is an animate object which may function as the agent of a particular verb' (Sidner 1979: 152).

Sidner's theory proposed a bottom-up anaphora interpretation as proposed by psycholinguists and such algorithms should be classified according to the anaphoric expressions, anaphoric semantic positions, personal pronouns in agent positions, non-agent positions, and possessive positions on which they operate. Sidner's theory was not evaluated, although studies were carried out to investigate how it works with various examples. Carter tried to conduct one such evaluation, which is discussed later on.

ii. Centering theory was developed by Grosz et al. (1995), and it formed the theoretical foundation for various anaphora resolution algorithms. Two of the most important are those of Brennan et al. (1987)[5] and Strube and Hahn (1999), which discussed below.

[5] Henceforce called BFP

The BFP (Brennan, Walker-Friedman, and Pollard) (1987) algorithm is influential as its features are based on solid empirical evidence. Poesio et al. (2004c) argued that there is sound empirical evidence for some of its features; for example, a preference for pronominalizing the CB (backward-looking centre) against any other entity being pronominalized. Other characteristics, however, are not grounded in solid verification, as Gordon et al. (1993) argued regarding preferences among transitions. The original algorithm was never evaluated by its original authors. Walker (1989), however, manually evaluated its performance compared to that of Hobbs' algorithm. The BFP results were slightly better than Hobbs' when it was evaluated using narrative texts (90% versus 88% accuracy). The performance of Hobbs' algorithm was better when using task-oriented dialogues (51% versus 49% accuracy), and it scored even better when using news data (89% versus 79%). Poesio et al. (2010) argued that Hobbs' algorithm scored better as it dealt with intrasentential antecedents while BFP dealt more with intersentential antecedents. Tetreault (2001) carried out an extensive evaluation which suggested that Hobbs' algorithm performed better than BFP in cases of both fictional texts and news articles.

Strube and Hahn (1999) proposed an algorithm in which grammatical function is replaced by functional ranking. Functional ranking is based on Prince's (1981) taxonomy of given-new information. The taxonomy proposed that the hearer old-entities (anaphoric entities and entities referred to using proper names) are more highly ranked than mediated (bridging) references, which consequently are more highly ranked than hearer-new entities. Functional ranking showed better results than grammatical function. Such results were confirmed by Poesio et al. (2004c), who found that functional ranking parameter configuration best supports the centring hypothesis.

iii. Graded Salience Models (Leass and Lappin 1994) are based on the notion of activation. Activation-based anaphor resolution models are based on the idea that each discourse entity has a given activation level which can be measured using a graded scale. The activation level is updated after each new utterance, which determines the prospect of that entity being referred to. Poesio et al. (2010) argued that although activation-based models have been discussed less often, they are widely used in anaphor resolution systems compared with discrete models of salience.

Lockman and Kloppholz (1980) proposed the first activation-based model, but MEMORY, a system proposed by Alshawi (1987), is considered to be the best-known activation-based model. Leass and Lappin's (1994) pronoun resolution algorithm is based on Alshawi's algorithm with the addition of several expletives treatments and binding constraints.

Leass and Lappin's algorithm (the Resolution of Anaphora Procedure, RAP) is classified as a generate-filter-rank anaphora resolution model. RAP depends for its input on the output of a full parser, and it uses syntactic information and binding constraints to filter antecedents. It categorizes antecedents as:

a. Antecedents of non-reflexives, when the pronoun occurs in the adjunct or NP domain of the potential antecedent, and

b. Non-pronominal antecedents, which occur within the pronoun governing phrase (Poesio et al. 2010).

Binding criteria decide how to resolve reflexive pronoun antecedents. Possible candidates have to pass the syntactic filter and agree in number and gender with the pronoun, and then the one with the highest salience weight is selected. This method overcomes the closest antecedent principle. For every newly introduced mention, it is assigned an initial salience weight that consists of sentence recency weight, additional weights for mentions occurring in the correct position, grammatical roles parallelism, cataphora (which are treated as a penalty) and a weight for grammatical functions.

In order to evaluate their algorithm, Leass and Lappin used 360 examples extracted from computer manuals. The RAP got 310 pronoun antecedents which formed 86% of the total number being examined. If salience, grammatical function, and parallelism function are removed, the algorithm's scores significantly decrease. Other factors such as coreference chains and the cataphora penalty have a limited effect on scores. When implemented with the same data, Hobbs' algorithm scored 82% accuracy. More deep linguistic information is used at three positions:

a. To define restrictions and incompatibility in the case of reflexive resolution.

b. Using grammatical functions as a base to assign salience weights.

c. To assign the gender for a full noun phrase using a parser’s lexicon (Poesio et al. 2010).

Kennedy and Boguraev (1996) introduced the usage of Constraint Grammar parsers in order to assign morphological tags and grammatical functions, and to identify NP chunks. Its rate of accuracy was 75% for news text; errors were due to direct speech and insufficient gender information.

Strube (1998) and Tetreault (2001) were inspired by the centring theory to propose an algorithm. The algorithm, as Poesio et al. (2010) argued, should be considered as an example of activation models where activation scores (a partial order) are replaced by a list (a total order). In table 4.2 below Tetreault’s left-to-right centring (LRC) algorithm is stated. The algorithm is a combination of CFs from centring theory and some ideas from Hobbs’s algorithm. Tetreault evaluated his algorithm using a corpus of news articles and fictional texts. The algorithm scored 80.4% accuracy for the news articles and 81.1% for the fictional texts.

Table 4.2: Tetreault's LRC Algorithm (Poesio et al. 2010)

Tetreault's LRC Algorithm
1: for all U_n do

```
parse $U_n$
for all $CF_i$ in the parse tree of $U_n$ traversed breadth-first, left-to-right do
    if $CF_i$ is a pronoun then
        search intrasententially in CF-partial($U_n$), the list of CFs found so far in $U_n$, an antecedent that meets feature and binding constraints
        if found matching antecedent then
            move to the next pronoun in $U_n$
        else
            search intersententially in CF($U_{n-1}$) an antecedent that meets feature and binding constraints.
        end if
    else
        add $CF_i$ to CF-partial($U_n$)
    end if
end for
```

4.4.4 SPAR: Putting Syntactic, Commonsense and Focusing Preference Together

In 1987 Carter proposed the SPAR system. This is considered to be the most fully developed proposal for pronoun resolution before the data-driven methods that are discussed later on. Carter’s main contribution was in creating a combination of existing proposals. SPAR used Sidner’s pronoun rules to resolve intersentential anaphora, while

Hobbs' algorithm was employed to produce the ranking used in resolving intrasentential anaphora, and Wilks' preference semantics were used to encode the semantic types of mentions and causal reasoning. The algorithm's input is generated using Boguraev's (1979) English analyser.

Carter carried out an evaluation of SPAR using a corpus consisting of sixty stories, each of which is two-three sentences long. He reported a 100% accuracy rate in the stories written by him and 93% accuracy with the other stories. There is no evidence that any other attempt at evaluation was carried out. However, many of Carter's ideas were adopted later on by Alshawi (1992) in the Core Language Engine.

The foregoing section gave a brief overview of the linguistic background of anaphora and anaphora resolution. It has summarized the early models of anaphora resolution where preferences, constraints, and required information were hand-coded. The next section discusses how the broad empirical study of anaphora resolution was affected by the creation of large, modern, digital corpora, which led to the development of data-driven methods. These methods require techniques to reliably and automatically extract morpho-syntactic knowledge, commonsense knowledge, and large repositories of lexical knowledge. In the early days of data-driven methods, such techniques were not available and so, simple approximations were used to deal with constraints and preferences. Since then more complicated techniques have been developed and become available, enabling such methods to be applied to large numbers of texts.

4.5 Towards an Empirical Approach to Anaphora Resolution: Developing an Experimental Setting

In the 1990s, there was a shift in focus in anaphora resolution research towards greater empiricism, largely as a result of the development of the field of information extraction. The first medium-sized annotated corpora were created, which made the creation and development of data-driven resolution procedures and machine learning approaches a possibility.

The Message Understanding Conferences (MUC) project was behind the changes, which is a DARPA-funded initiative which aimed to compare the qualities of information extraction systems using annotated corpora. The funding agencies hosted several coreference resolution systems, such as MUC-6 (Grishman and Sundheim 1995) and MUC-7 (Chinchor 1998), where annotated corpora were provided. As a result, guidelines for the annotation of coreference were created and standard evaluation metrics to be used in the comparison process were developed. This made training and testing of anaphora resolution systems using the same datasets possible. These changes had a strong influence on the anaphora resolution field specifically and on the field of evaluation in general, which is still in progress in the Automatic Content Extraction (ACE) initiative (Poesio et al. 2010). Some researchers consequently classify research in the field as conducted in the pre-MUC or post-MUC periods.

4.5.1 Annotation Schemes for Anaphora

The design of an annotation scheme is a crucial component of data-driven methods. The coreference information is used in:

a. The performance evaluation of coreference resolvers; and

b. Supervised systems training, which is directly related to machine learning approaches (Poesio et al. 2010).

The annotation scheme mission is to define coreference problems and to specify what data can be learned from the linguistic phenomena. The following discussion explains the MUC decisions, initiatives, controversies, and subsequent developments.

The MUC annotation scheme is considered to be one of the most important annotation schemes as it has defined the focus of research during the fifteen years since it was developed by Hirschman in 1998. The focus of the annotation scheme is on nominal coreference. Coreference is defined in the scheme as 'the identity of reference'; that is, when two nouns phrases refer to the same set, object, or activity. All coreference relations involving two NPs or a noun phrase and a nominal modifier were annotated; any other types of relations were ignored (Poesio et al. 2010).

The MUC annotation scheme brought to the attention of researchers the problem of defining an anaphora coding scheme, or 'which text constituents to choose as mentions of the entities' (Poesio et al. 2010: 39). The scheme depends on syntactic and semantic factors;

syntactically, the coders need to mark the full noun phrase with all its post-modifiers. MUC coders marked the maximal span of NPs while the head of each NP was marked separately using a MIN attribute. This made the evaluation process easier as scores were given to matching heads and minimal spans while the full set of modifiers could be recovered at a later stage using another category of syntactic information. In subsequent stages the annotators had to annotate the NP with all its modifiers (Poesio et al. 2004; Pradhan et al. 2007).

From the semantic perspective, coders had to annotate mentions of all entity types, or only a subset of them. For a small number of semantic classes coreference resolution is important. The early models such as those of MaCarthy and Lehnert (1995) and Aone and Bennett (1995) mainly focused on organisations and persons. This focus on a small group of well-defined semantic classes makes identity determination easier, whereas this would have been difficult in cases of non-defined objects. The ACE evaluation, consequently, limited the coreference task so that it would only consider persons, organizations, geopolitical entities, locations, vehicles, and weapons. The ACE simplifies the coreference task by creating an application-oriented setting but it does attend to entities mentioned in other domains. In order to overcome such problems, Poesio et al. (2004) developed GNOME, whose domain included museum objects as well.

MUC was criticized for its tendency to annotate apposition and copula constructions which were not usually seen as cases of coreference. Van Deemter and Kibble (2000) argued that the annotation of intensional descriptions (as the predicates in a copula construction) led to unnecessary effects. Poesio et al. (2004) and Pradhan et al. (2007), who developed the MATE and OntoNotes annotation schemes, tried to overcome this problem by distinguishing between transitive coreference links and directed, non-transitive ones. Other schemes, such as the one developed by Artstein (2008), tried to annotate other anaphoric relations.

Specifying which markables to annotate is a difficult problem, as Poesio et al. (2010) argued, especially in the treatment of metonymy and particularly with geopolitical entities. For example, Washington may mean the city of Washington or the country or government of the USA as a geographical entity. Each annotation scheme treated such structures differently. For example, the ACE resolved them by developing a semantic class called 'geopolitical entities' (GPEs), while OntoNotes distinguishes such entities from other uses of an NP.

Annotating coreference relations is problematic as it requires quantitative agreement between annotators. There were early attempts such as with MUC to try to score agreement in terms of a scoring metric, but later studies did not include such quantification. Poesio and Vieira (1998) and Poesio and Artstein (2008) studied agreement in anaphoric annotations as part of the GNOME and ARRAU corpora. These studies showed that agreement can be detected via the distinction between old discourse and new discourse. These studies also argued that the identification of subset bridging relations is essential for annotating bridging reference to be possible.

Recent coding schemes, including the GNOME corpus developed by Poesio (2004), ARRAU developed by Poesio and Artstein (2008), OntoNotes developed by Pradhan et al. (2007), and ANCORA developed by Recasens and Martí (2009), differ from MUC/ACE schemes as only a few types, rather than all, NPs are annotated. In such modern schemes, the annotation of associative relations, types of discourse deixis, and all modifiers, as well as the ability to distinguish between identity and predication, are all available.

Table 4.3 below (adapted from Poesio et al. 2010) gives a summary of the available anaphorically annotated corpora, with information about publications and sites, some of which are available in compatible mark-up formats as part of the Anaphoric Bank initiative.

Table 4.3: A summary of anaphorically annotated corpora (Poesio et al. 2010)

Language	Name	Reference	Size (words)
Arabic	ACE-2005	Walker et al. (2006)	100k

	OntoNotes3.0	Weischedel et al. (2008)	200k
Catalan	AnCora-CO-Ca	Recasens and Martí (2009)	300k
Chinese	ACE-2005 OntoNotes3.0	Walker et al. (2006) Weischedel et al. (2008)	≈200k 1224k
Dutch	COREA	Hendrickx et al. (2008)	325k
English	MUC-6 MUC-7 GNOME	Grishman and Sundheim (1995) Chinchor (1998) Poesio (2004) Walker et al.	30k 30k 50k

	ACE-2005	(2006)	400k
	NP4Events	Hasler et al. (2006)	50k
	OntoNotes 3.0	Weischedel et al. (2008)	1150k
	ARRAU 1.0	Poesio and Artstein (2008)	300k
French	DEDE (definite descriptions)	Gardent and Manuẻlian (2005)	50k
German	Potsdam Commentary Corpus	Stede (2004)	33k
	TüBa-D/Z	Hinrichs et al. (2005b)	600k

Italian	Venex	Poesio et al. (2004a)	40k
	i-Cab	Magnini et al. (2006)	250k
	LiveMemories1.0	Rodriguez et al. (2010)	250k
Japanese	NAIST Text Corpus	Iida et al. (2007b)	38k sentences
Spanish	AnCora-CO-Es	Recasens and Martí (2009)	300k
Tibetan	Tusnelda (B11)	Wagner and Zeisler (2004)	<15k

4.5.2 Evaluating Coreference Resolution Systems

Poesio et al. (2010) argued that that a persisting question is how algorithms and systems of anaphora resolution work in comparison to each other. The earlier models of pronoun resolution depended on accuracy as an evaluation measure. Accuracy is the ratio of correctly resolved anaphora incidents to the total number of anaphora incidents.

Mitkov (2000) and Byron (2001) established criteria for judging an evaluation method:

a. ‘Does the evaluation compute the performance of the resolution algorithm only (i.e. assuming perfect pre-processing, including agreement features like number or gender) or rather of the whole system, where pre-processing steps such as parsing and determination of gender features are done automatically?

b. Does the evaluation include or exclude difficult cases such as first-person pronouns (which may not be resolvable to an antecedent), cataphora, cases of expletive pronouns, or pronouns and demonstratives that refer to clauses instead of noun phrases?

c. What type of texts is the evaluation carried out on, as technical manuals seem to be easier to treat with pronoun resolution than newspaper text?’ (Poesio et al. 2010: 44).

The latter two points become less problematic when adopting the MUC and ACE standard corpora. Quantitative results still pose a problem even when using standard datasets, as a variety of evaluation metrics and conditions are used. Various researchers, such as Stoyanov et al. (2009), showed that marked-up NPs in an annotated corpus cause many inadequate results when compared with the anaphora resolution systems that treat automatically extracted markables. Glaser (2011) argued that ‘a markable is a linguistic expression that may refer to another linguistic expression. Usually, markables are noun phrases. In ACE terminology, a markable is called a mention.’ Each markable noun phrase, together with the anaphor, forms a negative training instance. The next section discusses the most important evaluation measures that have been developed. These are classified into three main classes: link-based measures, set-based measures, and alignment-based measures.

4.5.2.1 Link-based Measures

The simplest way to evaluate an anaphora resolution algorithm is to let the module choose an antecedent for each pronoun and then calculate the accuracy of such choices depending on how many correct incidents are resolved. Until recently, most anaphora resolution systems were mention-pair models, as the algorithm has to decide if two noun phrases refer to the same discourse entity. The simplest method here is called link-based and entails checking whether the mention chosen by the system as the last mention of the same entity is in fact the last mention in the gold standard. Burch et al. (2003) claimed that ‘a gold standard is a manually crafted set of examples, against which the results are compared’. This measure of evaluation is unsatisfactory in many respects (Poesio et al. 2010).

Link-based evaluation gives unsatisfactory performance at many levels, such as in information retrieval where inflated accuracy assessments are produced due to the fact that only 30-40% of the markables are anaphoric. Accuracy rates do not yield a very clear picture of system performance since expressions may be anaphoric or non-anaphoric, as in the case of definite noun phrases. For example, definite NPs like ‘the town’ may refer to an introduced entity 50% of the time, or may be introducing a new entity the rest of the time, as Poesio and Vieira

(1998) argued. As a consequence, one system may regard the definite NP as an anaphor and start to look for its antecedents, whereas another system may regard it as non-anaphoric. Each choice has its advantages and disadvantages, so there was a need to replace the measure of accuracy with two more reliable performance measures:

i. ‘Precision: the ratio of the number of correctly resolved anaphoric links to the total number of links that a system resolves, and

ii. Recall: the ratio of the number of correctly resolved anaphoric links to the total number of anaphoric links in the annotated gold standard’ (Poesio et al. 2010: 46).

$$Precision = \frac{\#\text{correct}}{\#\text{resolved}} \qquad Recall = \frac{\#\text{correct}}{\#\text{wanted}}$$

Both precision and recall are usually merged into one evaluation measure; which is called the F-measure (F_1). The F-measure was introduced by van Rijsbergen (1979) as a measure of evaluation in information retrieval.

$$F = \frac{2}{1/\text{Precision} + 1/\text{Recall}} = \frac{2 \times \text{Precision} \times \text{Recall}}{Precision + Recall} = \frac{2 \times \#\ \text{correct}}{\#\ \text{resolved} + \#\ \text{wanted}}$$

The arithmetical mean of the two numbers when they are close to each other indicates harmony; a large difference shows that harmony is closer to the minimum of the two numbers.

4.5.2.2. Set-based Measures

The calculation of precision and recall with the early MUC versions was carried out using comparisons and gold-standard links. This proved to be an inaccurate method as the system is required to reproduce links which are annotated in the gold-standard. Vilain et al. (1995) proposed precision and recall statistics over equivalence classes in order to overcome this problem. This method was called the MUC evaluation measure at the beginning of the MUC-6.

4.5.2.3 Alignment-based Measures

Vilain et al.'s (1995) evaluation method was regarded as an optimistic generalization of link-based measures used with coreference sets. The reason for this is that the MUC's scores are considered to be attainable for the decomposition of the system's links and gold-standard partitions. Poesio et al. (2010: 47) pointed out that 'This leads to counterintuitive effects on the small scale (misclassifying one markable into the wrong coreference set counts as one precision and one recall error, while completely merging two coreference sets counts as a single recall error) which are compound when evaluating the system response on true (gold) mentions, where all singletons and non-referring mentions are removed. In this case, just merging all coreference chains simply incurs a number of precision errors of the number of coreference chains (minus one), whereas the number of correct links is evaluated as the total number of gold mentions (minus one) [.…] with 100% recall and about 80% precision on the MUC-6 and MUC-7 datasets.'

Trouilleux et al. (2000) and Luo (2005) proposed methods that aggressively overcome overmerging methods. The idea of alignment was proposed in such studies, which aims to work between gold and system partitions by selecting links which satisfy the following conditions:

> i. 'Every coreference chain in the system's response corresponds to at most one coreference chain from the gold standard, and vice versa, and
>
> ii. The highest weight among these assignments is reached' (Poesio et al. 2010: 47).

Trouilleux et al. (2000) tried to calculate the weights of the alignment links. Poesio et al. (2010) argued that there were initiatives to: create an alignment between gold partitions and system partitions assuming that: a) every system coreference chain corresponds to a chain in the gold standard and vice versa, and b) reaching the highest weight among such alignments is main requirement. The sum of weights is equal to the score of 1, even in cases of names, common noun phrases and pronouns, where the weighting is different. The summed score resulting from the number of correct links that are in common with the aligned coreference chains using is to be compared with:

i. ‘The link count for the system’s coreference chain, to get the precision, and

ii. The link count for the coreference chains in the gold standard, to obtain the number for the recall’ (Poesio et al. 2010: 48).

Luo (2005) proposed a similar measure, the Constrained Entity-Alignment F-Measure (CEAF) metric, which calculates the alignment and then carries out a comparison between the mention sets in the systems (for precision) or the gold standard coreference chains resulting from the alignment. Each mention has to occur both in the system and the gold-standard coreference chains that the alignment links together. Luo argued that the weighting emphasises named entities and de-emphasises pronouns, which means that the name matching is overemphasized and that pronoun resolution is under-scored.

4.5.2.4 Comparing the Metrics

As an example of set-based metrics, MUC gives credit for a system if it recognizes part of a coreference set or if it misses it. Alignment-based methods, in contrast, depend on determining if the system succeeds in discriminating between the various coreference chains in the global view (Poesio et al. 2010).

Table 4.4 below shows a comparison between MUC scores, the CEAF alignment-based metric, and ‘purity’ (Solomonoff et al. 1998), which is an evaluation metric used in document clustering systems. The table shows that the CEAF’s results overwhelmingly disagree from the point of view of recall and precision. MUC’s results show a slight decrease in precision while purity shows a greater decrease, while both MUC and purity recall scores remain the same.

Table 4.4: A comparison between MUC scores, CEAF alignment-based metric, and ‘purity’ (Solomonoff et al. 1998)

GOLD A_1 A_2 A_3 A_4 A_5	MUC PRF_1	Purity PRF_1	CEAF PRF_1
System 1 A_1 A_2 A_3 A_4 A_5	3/4 3/3 0.86	3/5 5/5 0.75	3/5 3/5 0.60
System 2 A_1 A_2 A_3 A_4 A_5	2/3 2/3 0.67	4/5 4/5 0.80	4/5 4/5 0.80

4.6. Modern Computational Approaches

Klavans and Resnik (1996) claimed that coreference resolution researchers tended to use large quantities of linguistic data. This tendency leads to similar results as those achieved in other areas of CL research. The coreference resolution researchers learned from their work that using linguistic and ontological information and sources of errors is a difficult process, especially in an automatic system that would generate analyses of unrestricted text. This apprehension led to the usage of ‘knowledge-poor’ methods. Knowledge-poor methods count on structures/features that are easy and reliable to get. These models were developed since earlier models were dependent on domain knowledge or deep syntactic analysis as in the case of Hobbs' naïve algorithm. Domain knowledge models are considered to be expensive in terms of time and effort as they require the analysing and encoding of relevant facts, especially when adapted to a different domain. Meanwhile syntactic analysis models require accurate automatic parsing, which was not available during the mid-1990s. For example, Leass and Lappin's (1994) algorithm used an automatic parser, and its results needed to be edited to overcome errors.

In other NLP tasks, the use of simpler types of information such as morpho-syntactic contextual features, and shallower methods such as data-driven supervised learning, has become popular. This encouraged AR researchers to adopt such methods; although recently, with the ease of use of robust statistical parsing methods and the availability of annotated semantic information, there have been studies that try to couple shallow methods with sources of information in modelling syntactic heuristics and commonsense reasoning. The re-introduction of syntactic and semantic analysis is encouraged especially for the features of coreference classifiers which are automatically extracted from linguistic data.

Poesio et al. (2010) argued that the right establishment of priorities within anaphora resolution process proved to be difficult as parsing for other CL aspects. Bod et al. (2003) claimed that machine learning techniques solved problems of the establishment of priorities; in addition, probabilistic techniques are used to solve problems concerning the combination of evidence.

4.6.1 Resolution Architectures

Computational linguistics defines coreference chains as the construction of equivalence sets of mentions of discourse entities (Poesio et al. 2010: 50). Identifying coreference chains requires the identification of 'links' between mentions or between mentions and entities. The links, in addition, need to be clustered in equivalence classes.

i- Hand-coded versus machine learning: Soon et al.'s (2001) seminal proposal used machine learning techniques as well as a reasonable amount of hand coding for feature extraction. Anaphora resolution methods proposed in the 2000s used supervised learning in conjunction with hand-annotated resources, while others such as Ng's (2008) used unsupervised learning.

ii- Single versus multiple classifiers: algorithms developed by Hobbs (1978), Carter (1987) and Sidner (1979) all focused on one type of NP, where a different algorithm is developed for each NP type. Machine learning systems usually develop a single model that deals with all types of NPs, as in the case of Soon et al. (2001), although that of Hoste (2005) was an exception to this trend.

iii- Serial versus parallel: many algorithms, such as Winograd's, choose antecedents by going backwards from the anaphor.

Sidner's algorithm detects suitable antecedents by following the order dictated by focus rules, in addition to the LRC (Left-Right Centering) algorithm, and Soon et al.'s algorithm considered one antecedent at a time. This method of the choice of antecedents makes it difficult to compare alternatives. Where several competing hypotheses are considered, parallel and ranked algorithms may be considered as an alternative, depending on preference scores.

In psycholinguistics and computational linguistics, the early disambiguation algorithms were serial in order to explain incremental effects such as garden paths. More recent algorithms are parallel, such as that of MacDonald et al. (1994). Hobbs used heuristically calculated weights, going through to the abduction based resolution developed by Hobbs et al. (1993) and then the use of statistics. Parallel models are used widely in AR algorithms, as in the case of Brennan et al.'s (1987) BFP, or in the ranking algorithm developed by Ng and Cardie (2002b) or the tournament models proposed by Iida et al. (2003a) and Yang et al. (2003). Antecedent ranking models that have to deal with the intricacies of the anaphora resolution task are also called global models, such as the ones proposed by Ng (2005), Denis and Baldridge (2007b), and Rahman and Ng (2009), as well as the unsupervised models of Haghighi and Klein (2007) and the document-level models of Culotta et al. (2007), and Daumé III and Marcu (2005).

iv- Generate-filter-rank: The algorithms of Sidner (1979), Leass and Lappin (1994), Mitkov (1998), and Ng and Cardie (2002b) all belong to this category. The main feature of these algorithms is that there is a distinction between constraints and preferences. The main three components are that:

- In order to extract antecedent candidates from the preceding text, one or more generators are needed.
- In order to use hard linguistic constraints like binding and agreement constraints, a filter is needed.
- A ranker is needed to choose between antecedent candidates; the antecedent that scores the highest salient score is to be chosen. The ranking is carried out depending on surface form and configuration information. When the ranking is predictable, ranked candidates can be generated by choosing them after they pass the filter. For generate-filter-rank approaches the antecedents are chosen after filtering and the ranking of all anaphoric mentions in a sentence. For centring-based approaches, where each pronoun in an utterance is resolved simultaneously, machine learning approaches treat constraints and preferences as features.

    v- Clustering-based approaches take a global view in constructing coreference chains. They use a kind of uncertainty reasoning as constraint propagation, as in the case of Klenner and Ailloud's (2008) algorithm, or in the probabilistic approach of Culotta et al. (2007). Cluster approaches do not depend on single antecedent decisions but rely on the larger contexts to overcome any drawbacks of the single antecedent decisions. Cluster approaches make use of the generate-filter-rank model results as input, by incorporating them as features, as in Lin’s (1995) algorithm.

4.6.2 Heuristic Approaches to Pronoun Resolution

In the 1990’s, there was a tendency to develop heuristic approaches which used poor-quality information extracted from corpora. This section describes the main approaches of this kind.

i. MARS was developed by Mitkov (1998) using heuristic rules to assign a score to each antecedent candidate and to select the candidate with the highest score. The approach was evaluated using technical manuals, and it avoided knowledge-intensive features. Candidates that score the same are collected and subjected to a set of heuristics (each heuristic or preference has a certain weight and awards certain points to every anaphor-antecedent relationship); and then the sum of individual scores of heuristics is calculated. The heuristics are as follows (Poesio et al. 2010: 53):

   • Definiteness: since definite noun phrases are more likely to be discourse-old, and thus salient, indefinite NP antecedent candidates get a -1 score.

• Givenness: the first NP in a sentence gets a score of +1 on the grounds that it is more likely to represent given information.

• Indicating Verbs: the objects of verbs such as *discuss, present, illustrate, summarise, examine* etc. are given a +1. Mitkov (2002) argued that empirical evidence showed that noun phrases following the previously mentioned verbs would carry more salience.

• Lexical iteration: if a noun phrase head occurs more than once within the paragraph, this is taken to be an indication that the entity is especially salient and the corresponding noun phrases are given a +1 (two occurrences in the paragraph) or +2 (more than two occurrences) score.

• Section heading preference: Aa noun phrase that occurs in the header to the current section gets a +1 score.

• "Non-prepositional" noun phrases: noun phrases embedded in PPs are not preferred (on the grounds of grammatical salience) and given a -1 score.

• Collocation pattern preference: noun phrases that occur as a subject/object of the same verb as the anaphor are preferred and get a +2 score.

• Immediate reference: in a coordinated construction of the form "V_1 NP and V_2 *it*", a resolution of *it* to the noun phrase in NP is preferred as it usually expresses strong parallelism. The noun phrase in parallel position (NP) gets a +2 score. Mitkov (2002) argued that immediate preference can be regarded as a modification of collocation preference. The importance of immediate preference arises from it being highly genre-specific and with high occurrence in imperative constructions. For example, 'To print the paper, you can stand *the printer* up or lay *it* flat' (Mitkov 2002: 148).

• Referential distance: nearby antecedent candidates in the information source are preferred over distant ones. In complex clauses, noun phrases in the previous clause get a +2 score. Otherwise, noun phrases one, two or more than two sentences back get scores of +1, 0, or -1, respectively.

• Term preference: candidate noun phrases are checked against a list of nouns that are part of the domain's terminology, and get a +1 score if they are such terms.

Priority is given to immediate reference, collocation pattern preference, and indicating verbs scores in that order to calculate the highest scores, and selecting the highest scoring candidate or choosing the most recent candidate if all else fails. The approach was evaluated using technical manuals, where gender, chunks and clauses were manually checked. The results showed that it scored 89.7% accuracy. Mitkov's approach was compared to Baldwin's (emulating) approach which scored 75% accuracy (manually calculated) or 66%

when compared to selecting the most recent matching candidate.

ii. Heuristics for high-precision resolution were developed by Baldwin (1997). In order to extract mentions and utterances, his system uses NP and clause chunking. Shallow patterns are used to determine a number of cases that can be resolved. Once the partial order is established by the shallow information that is available, a single preferred antecedent is chosen and the system applies the following rules (Poesio et al. 2010: 54):

- Unique in Discourse: if there is a single compatible antecedent in the prior discourse, resolve to that antecedent.

- Reflexive: resolve reflexive to nearest possible antecedent.

- Unique in Current+ Prior: if the preceding noun groups of the current sentence and those in the previous sentence yield exactly one compatible antecedent, resolve to that antecedent.

- Possessive Pro: in the case of a possessive pronoun in "his X", if the previous sentence contains one exact match for "his X", resolve to that possessive pronoun as an antecedent.

- Unique Current Sentence: if there is a single compatible antecedent in the preceding noun groups of the current sentence, resolve to that antecedent.

- Unique Subject/Subject Pronoun: if the anaphor is the subject of the current sentence, and the subject of the prior sentence contains a single possible antecedent, then resolve to that antecedent. In the case of coordinated noun phrases, Baldwin counts the conjuncts as multiple subjects.

To resolve all pronouns in the text, Baldwin proposes two additional rules:

- Cb-Picking: motivated by concepts from centring theory, this rule resolves some cases that the subject/subject rule does not cover. If the anaphor is in a non-subject position and the subject of the utterance is a compatible pronoun (i.e. the Cb), pick that pronoun as the antecedent.

- Pick most recent: picks the most recent compatible antecedent.

The corpus that was used for the evaluation of Mitkov's algorithm consisted of three stories in which gender was manually annotated. The results showed that Baldwin's algorithm scored 92% precision, and 60% recall. When the high-precision rules were applied, it scored 77.9% accuracy while Hobbs' algorithm scored 78.8%.

In the MUC-6 evaluation, a modified version of this system was used. The system used WordNet look-up in order to determine gender, and Collins' parser was used to determine clause chunks. In order to process first-person pronouns in quoted speech, a special measure was used, while possessive pro, Cb-picking and pick most recent were removed, and the subject-same clause rule was applied in addition to the automatic detection of non–referential *it* pronouns. The rate of recall was 75%, and precision 73% using MUC-6 data.

4.6.3 Early Machine Learning Models

In the previous section Mitkov's and Baldwin's approaches were discussed. It was shown how the production of the final clustering of markables into coreference chains depends on heuristics and how much weight each would score. Poesio et al. (2010) argue that one of the main drawbacks of such approaches is that the process of ordering and weighting heuristics is time-consuming and prone to errors. This led to the development of machine learning methods, since these can carry out such tasks automatically and can make use of training data to learn constraints and preferences. The automatic usage of training data allows machine learning approaches to explore new features more easily and in depth than rule-based heuristic approaches. In what follows, the main machine learning approaches are briefly discussed.

i. Aone and Bennett (1995) designed a machine learning approach that is based on decision trees extracted from Quinlan's (1993) model. It is applied to the Japanese language, and targets anaphoric pronouns, anaphoric definite noun phrases and name coreference for persons and organizations. In the training corpus, features such as zero pronouns and anaphoric definites were manually marked up. In the training data, each anaphor was paired with previous members of its coreference chain to act as a positive example, while negative examples were made by pairing the anaphor with mentions that are not coreferent with it. For each instance pair, feature vectors and semantic information are created to be used as an input for the classifier. Within the resolution process each anaphoric expression is paired with a possible antecedent and feature vectors are created for each anaphor-antecedent pair. The classification of each pair is dependent on the decision tree that results from the training data. The antecedent that is positively marked and has the highest confidence score is chosen.

ii. RESOLVE was developed by McCarthy and Lehnert (1995) as part of the MUC-5 information extraction task. They built a decision-tree-based coreference resolver called RESOLVE which makes use of domain independent features such as name substrings and mention types, in addition to domain-specific features. The evaluation was carried out manually by annotating texts extracted from the MUC-5. The results showed that the recall results of the decision trees were higher than those of Lehnert's et al. (1992) rule-based system, while it made only a very slight change to the precision results. RESOLVE makes a record of every pair of template-relevant noun phrases.

In the MUC-6 coreference task a more fully developed version was evaluated. Features such as string match and sharing a common semantic type were used. The results showed that RESOLVE scored 44% for recall and 51% for precision, which is considered to be low compared with rule based systems such as that of Kameyama (1997).

iii. Vieira and Poesio (1997, 2000) aimed to resolve definite noun phrase anaphora in unrestricted texts. The system represents an early attempt to provide solutions using lexical and commonsense knowledge. Vieira and Poesio developed hand-coded and machine-learned versions of decision trees. Consequently, these were used to compare hand-coded algorithms with machine-learned ones. Vieria and Poesio's algorithm is interesting as it proposes a solution for discourse-old versus discourse-new identification. Vieria and Poesio's work tries to choose possible antecedents for discourse-old descriptions by integrating decision trees with heuristics that are relevant.

Vieria and Poesio's algorithm developed a typology of definite noun phrases. The main obstacle that it faced is that not all definite noun phrases are anaphoric, as Loebner (1987) argued, since half of the definites mentioned in a corpus are considered to be discourse-new descriptions. Another obstacle is that some associative descriptions may denote an object which itself may be discourse-new while it may be associated within an already introduced identity. The 1998 algorithm succeeded in making the distinction between old and new discourse descriptions. It managed as well to be able to choose the compatible antecedent suitable for the anaphor. Decision trees included heuristics suitable to deal with unique descriptions that could be discourse-old as well.

The algorithm dealt with direct anaphora by identifying all the noun phrases having the same head as the definite noun phrase. Possible candidates undergo a check using modification heuristics.

The process of head matching may result in producing spurious antecedents. This happens when, in an earlier part of discourse, a certain type of entity is used and later a different entity belonging to the same type as the first one is mentioned. In such a case it is recommended to use segmentation heuristics in order to exclude potential antecedents that can be possible candidates for the definite noun phrase. Considering only the most recent same-head noun phrase and limiting the distance to the antecedent can work well, which is why Vieira and Poesio developed a loose segmentation heuristic that limits the search of the possible antecedents within a four-sentence window or which are discourse-old or identical to the definite noun phrase.

The algorithm includes a number of heuristics for detecting discourse-new descriptions where syntax is an important source of information. The algorithm tries to detect certain syntactic configurations or copula constructions. In order for a predicate to be functional, the algorithm looks for functional heads or modifiers that make predicates functional. Such definites need to be licensed to be anaphoric through semantic uniqueness.

In the case of bridging descriptions, the use of lexical resources like WordNet is allowed to resolve cases where the antecedent’s head suggests a possible coreference relationship which can be hypernymy or synonymy or part of a relationship that can be classified as being associative bridging. Categorized named-entities lists are used as helpful tools to resolve instance relations.

The sources of information listed above are combined to determine discourse-new descriptions and resolve anaphoric relations via two methods: a hand-coded decision tree and the learned decision tree developed by Quinlan (1986). The hand-coded tree is similar to the one developed by Baldwin for the COGNIAC system. As for the machine learned decision tree, it starts by attempting to resolve same-head anaphora, then high precision discourse-new heuristics use lower precision information. An incremental resolution strategy is then applied by assigning a file card for every noun phrase it encounters. For dealing with a definite nominal in order to determine its classification, and also to try to find an antecedent, a decision tree is used. A serial resolution is applied that goes right-to-left until it locates a suitable antecedent or it reaches the boundary of the segment.

Using twenty texts adopted from the Penn Treebank, the two decision trees (hand-coded and machine-learned) were developed and trained. The texts contained 6831 NPs out of which only 1040 were definite descriptions. The evaluation of the hand coded system took place using fourteen texts with 2990 NPs and 464 definite descriptions. The system scored 53% for recall and 76% for precision. All unresolved definites marked as discourse-new in the hand-coded version were compared with the machine-learned decision tree on a subset of the previous evaluation with a set of 200 definite descriptions which were hand-annotated. The hand-coded system scored an F-measure of 77% while the machine-learned system scored an F-measure of 75%. The precision score was low because of the attempt to interpret bridging references while the score for recall improved to a F-measure of 62%.

4.6.4 Anaphora Resolution: A Probabilistic Formulation

All the fields and subfields of computational linguistics were affected by the rise of statistical empiricism during the 1990s and 2000s. Anaphora resolution from such a probabilistic perspective can be summarized by the following quotation from Poesio et al. (2010: 58):

Given mention m_j, anaphora resolution is the problem of finding entity e_i belonging to the universe of discourse U for which it is most likely that m_j is a mention of e_i. In probabilistic terms, this means finding entity e_i

such that the probability of m_j being a mention of e_i is maximal, given the context the C of m_j.

$$\text{argmax}_{e_i \in U} P(m_j \text{ mention-of } e_i | C)$$

A completely general formulation should also cover the possibility that m_j is discourse-new; that is introduces a new entity e_{new} - or non-referring (i.e. an expletive). This can be done by allowing m_j to be a mention of a new entity e_{new} not included in U, and introducing a pseudo entity: we write that m_j is a mention of pseudo entity to mean that m_j is not referring. This leads to the following more general formulation:

$$\text{argmax}_{e_i \in E} P(m_j \text{ mention-of } e_i | C), E = \{U \cup \{e_{new}\} \cup \{\ \}\}.$$

The formulation above suggests that evidence combination techniques from probability could be used. E.g., viewing context C as a set of features f_k, applying Bayes' rule, and making the Naive Bayes assumption, we can compute the desired probability as follows:

$$\text{P}(m_j \text{ mention-of } e_i | C) =$$

$$P(C) \cdot P(C | m_j \text{ mention-of } e_i) =$$

$$P(f_1) \cdot P(f_1 | m_j \text{ mention-of } e_i) \cdot \ldots P(f_m) \cdot P(f_m | m_j \text{ mention-of } e_i).$$

In practice, systems estimate the probability that an indicator variable L, which is

1 if m_j is a mention of e_i and 0 otherwise, is 1 (e.g., see (Yang et al. 2008)):

$$\text{argmax}_{e_i \in U} P(L|m_j, e_i).$$

In the case of so-called mention-pair models, this probability is approximated to classify links between mentions:

$$\text{argmax}_{m_i} P\,(L|m_j,m_i).”$$

4.6.5 Early Probabilistic Approaches

i. Ge et al. (1998) tried to develop a generative statistical system that is able to use statistics for the addition of gender identification, selectional preferences, and a mention-count-based measure of saliency that is related to Hobbs' algorithm. The formula below shows the method of calculating the probability distribution over plausible antecedents.

P (m_j mention-of $e_i|C) \propto P$ ($d_H|e_i)P$ (m_j is-pronoun$|e_i$) $P(e_i|h; t; l)/ P(e_i|t) P(e_i|m_i)$

Ge et al. (1998) later presented a more developed version of the algorithm in which automatically resolved anaphor-antecedent pairs extracted from a large corpus were used. This addition resulted in a small improvement in the overall results.

The algorithm was evaluated using texts taken from the Penn TreeBank, where mentions were manually coded and cases of the expletive *it* were also manually removed. The later version scored 84.2% accuracy compared with the older version that used Hobbs' distance, which scored 65.3% accuracy.

ii. Kehler (1997) aimed to calculate the probability that two mentions co-refer and he developed two approaches to convert such probabilities into a probability distribution over partitions of mentions. The first approach is called the 'evidential reasoning approach' using the pairwise classification of all mention pairs adopted from the maximum entropy (MaxEnt) classifier (Berger et al. 1996). For inconsistent partitions, the approach assigns a non-zero probability distribution as a means of normalization. The second approach is called 'merging decisions', and regards a coreference set as a chain of decisions with every mention being regarded as part of an existing set; otherwise a new set would be created. Depending on how close a mention is from a set, the coreference probability factor decides whether to merge a mention with an existing set of mentions, or to create a new set.

Training examples are generated in accordance with the approach adopted. In the evidential reasoning approach, an example is generated for every pair in the training data. In the merging decisions approach, the most recent mention of a coreference is paired with a mention. In order to measure the compatibility between any two mentions, Kehler used a function of template representations; that is, either using identical slot values or one template properly subsuming the other or otherwise being consistent. The other features are classified into five classes that result from rule-based coreference models depending on the form of the noun phrase and the distance in number of characters between anaphor and antecedent. The system tries to show whether or not a preferred potential antecedent would be the choice in the case of a rule-based module. In the case of a rule-based module, the potential antecedent would be included among a list of possible antecedents and would not be marked as the highest possible one, or it may be classified as being unsuitable for a rule-based module.

In trained models, a positive value is given for two or more common slot fillers as well as when an antecedent is preferred by a rule-based system.

The system was evaluated using cross-entropies of test data of exact matches. The evidential reasoning in terms of cross-entropy and perfect matches gave superior results compared to the merging decisions approach.

4.6.6 The Mention-Pair Model of General Coreference

This model was proposed by Soon et al. (2001) and developed further by Ng and Cardie (2002b). The model aimed to shift away from the single NP type with restricted domain, and became the standard statistical formulation in AR. It regarded a resolved anaphor m_j as a classification task; the task of finding mention m_i which maximizes the probability according to the following function (for more details see section 4.6.4):

$$\text{argmax}_{m_i} P\,(L|m_j,\, m_i)$$

i. Soon et al. (1999, 2001) developed an algorithm that is decision-tree-based for coreference resolution using the evaluation corpora of the MUC-6 and MUC-7. The algorithm tries to handle the problem of pre-processing unrestricted texts in order to identify and analyse markables which the coreference classifier could deal with. The pre-processing stage includes a flow of sequence taggers that are standard statistical learning rules based on hidden Markov models, part-of-speech tagging, noun chunk identification, and named entities recognition. The module tried to merge spans and adjusted phrase boundaries and added the use of two extra modules that extract possessive premodifiers and premodifying nouns that the MUC-6 allows to co-refer with other mentions.

   These modifications allowed the usage of standard off-the-shelf components, which ensures portability across languages and domains. Consequently, the level of recall in retrieving potentially coreferring candidates is augmented due to such combinations.

   The generated training examples are divided as follows:

   a. Positive examples are created by pairing each markable with the most recent antecedent in the gold-standard coreference chain.

b. Negative examples are created by pairing the anaphor with other markables existing in between the anaphor and the most recent antecedent.

Soon et al.'s model used feature vectors to train a decision tree classifier. Table 4.5 below adopted from Poesio et al. (2010) shows the twelve features used by the system. Features include the form of the noun phrase, while other features deal with agreement, distance, string matching, and alias features.

Table 4.5: The 12 features used in the system from Soon et al. (2001)

Feature	Value	Description
Distance Feature	Integer	The distance in sentences between m_i and m_j
NP type features		
I PRONOUN	Boolean	1 if m_i a pronoun 1 if m_j
J PRONOUN	Boolean	a pronoun
DEF NP	Boolean	1 if m_j a definite NP

DEM NP	Boolean	1 if m_j a demonstrative NP
Agreement features		
STR MATCH	Boolean	1 if m_i and m_j string match
ALIAS	Boolean	1 if m_j an alias of m_i
GENDER	Boolean	1 if m_i and m_j gender match 1
NUMBER	Boolean	If m_i and m_j number match
SEMCLASS	Boolean	1 if m_i and m_j match semantically
NUMBER	Boolean	1 if m_i and m_j number match
PROPER NAME	Boolean	1 if m_i and m_j both proper names

Syntactic position		
APPOSITION	Boolean	1 if m_j in appositive position

A list of previously identified mentions are organised in document order and then processed from left to right during testing. To create a test instance, each mention is to be paired with any preceding one. A serial resolution model is used as the algorithm stops once a test instance is marked as positive. A feature vector, which is based on the features mentioned above in table 4.5, is produced and passed to the classifier that is to decide if the mentions are coreferent or not. If the classifier finds that the mention pair is coreferent, the resolution algorithm shifts its focus to the next anaphor in the list, and if not it iteratively pairs the examined anaphor with the preceding candidate antecedent until it reaches one that it finds can be coreferent with it. If the classifier decides that pairs of mentions are coreferent then a partitioning is applied to the document. The collection of mentions is regarded as a disjoint set while coreferent pairs are classified into separate, non-overlapping sets. Soon et al.'s system is considered to be simpler than those of Aone and Bennett (1995) and McCarthy and Lehnert (1995), since the generate-rank-filter is applied at an earlier stage.

For the coreference classifier to work efficiently, an in-depth analysis is needed in order to prioritize features according to their usage. The decision tree that is adapted from the MUC-6 gives the system a tendency to choose the closest antecedent which:

a. shares the same surface form, or

b. is detected as a name alias of the anaphor, or

c. exists in the same sentence as the pronoun anaphor and is gender-matched with it (Poesio et al. 2010).

Generally speaking the system scored a MUC F_1 of 62.6% on MUC-6, and for the MUC-7 it scored 60.4%. Poesio et al. (2010: 63) claimed that the reason for such performance levels is ‘the identification of mentions in text as a necessary preprocessing step […] they explicitly assess the influence of the preprocessing component responsible for automatically identifying the markables to be classified as coreferent’.

ii. Ng and Cardie (2002b) developed a system that extends those suggested by Soon et al. in two main respects; the use of:

    a. ‘Best-first clustering: Instead of stopping at the first antecedent for which P (Llm$_i$, m$_j$) is greater than a given threshold (i.e. > 0.5), their system computes the probability for all antecedents and selects the one with the highest coreference probability value from among all antecedents with coreference class values above 0.5.

    b. Feature set expansion: The effects of using a much larger feature set are investigated in detail. This extension explores the effect of including 41 additional

> features to the original feature set from Soon et al., which include a variety of knowledge sources for the coreference resolution classifier such as lexical, grammatical, semantic features, as well as the result of a 'naive' external pronoun resolver’ (Poesio et al. 2010: 65).

Ng and Cardie’s system scores a MUC F_1 of 70.4% on MUC-6 and 63.4% on MUC-7. Its success is attributed to coupling best-first clustering with a manually created list of 27 features; it also discarded features that caused the precision tree to score low when dealing with common noun resolution. The decision tree seems not to be able to successfully select features, although the 27 features include 9 that are adopted from Soon et al.’s system.

4.6.7 Beyond Mention-Pair Models

Researchers subsequently developed more sophisticated models that reflect a more in-depth view of anaphora resolution than the original systems developed by Soon et al. (2001) and Ng and Cardie (2002b).

Iida et al. (2003a) and Yang et al. (2003) proposed an approach in which a machine learning classifier carries out the ranking using tournament-based scoring. Another main research direction was to abandon the use of local models in determining the probability of links between mentions. Instead global models are used based on the probability that a mention refers to a given entity. This inclined these approaches more towards the discourse model-based theories of anaphora resolution used in psycholinguistics, as mentioned earlier. This research shift was due to the fact that systems resolving an anaphor of an antecedent without taking into consideration any foregoing linking decisions involving the examined antecedent are liable to make implausibility errors. An example of implausibility errors is choosing the pronoun *she* to refer to *Michelle Obama* where *Obama* was previously linked to the mention of *President Obama*. This shift was proposed to maintain global consistency across anaphoric chains, but it created new problems:

- ‘As observed by Kehler (1997), using only information about members of a coreference chain without the notion of antecedence blurs certain important notions such as recency.
- Inconsistencies in the coreference chains could derive from any decision in the sequence of those performed for a single document. This means that the algorithm has to keep track of multiple alternatives (and their scores) in a search space which increases exponentially with the number of markables in a document’ (Poesio et al. 2010: 50).

The global consistency of coreference has to be ensured in order to process coreference chains effectively. Luo et al. (2004), Daumé III and Marcu (2005), and Rahman and Ng (2009) proposed combining an entity-based model with a ranking algorithm, and this is discussed briefly in the following section.

Yang et al. (2003) made use of a classifier where each anaphoric expression is paired with two previously mentioned candidates and the classifier’s outcome expresses a preference for one of the two candidates.

The preliminary selection of candidates to be presented as input for the coreference classifier is crucial for a ranking-based approach:

- In the training set, class imbalance must be maintained or the classifier’s results would be biased towards the first or the second candidate.

- In the training data, a training pair is produced by linking a positive candidate with a negative one. This dictates that the test data are generated differently according to various NP kinds in order to maintain the class balance.

- For evaluation purposes a Soon et al. (2001) classifier is used to filter the candidates, which ranks all candidates that are positively classified by the classifier.

The original system proposed by Yang et al. (2003) was developed by Yang et al. (2005) in order to identify discourse-new, i.e. non-anaphoric definite NPs generated by the tournament model. In the new model, non-anaphoric non-pronouns are determined by integrating their classification into the tournament model being used for ranking. This gives the classifier the chance to declare that neither of the two candidates is suitable.

Yang et al. make use of discourse-new mentions in the gold-standard and randomly pair them with selected previous mentions to train the model. Of these candidate-pair instances, a sub-sample is added to the training data with the appropriate classification model. The candidate's score either increases or decreases during the tournament classification testing, or scores for both mentions decrease. The best scoring candidate is chosen if its score is more than 0. This alternation in the model leads to an increase in the precision score at the cost of the recall score, which improves the F-measure score as well.

i. Luo et al. (2004) designed an entity-based system in which training is carried out over clusters. The resolution algorithm looks for the highest probable partition of a mentions set. The search is structured according to the Bell tree (Bell 1934), with each leaf including a candidate partition of the mentions. Each mention existing in a document is taken into consideration by the entity-mention model which processes it from left to right.

   A binary classifier is trained to process either anaphor-antecedent pairs or anaphor-coreference set pairs. The highest scoring candidate antecedent is chosen if its score is higher than the optimal threshold found in the development data set.

   In the mention-pair model, Luo et al. (2004) modified the features they used in the entity-mention model such as string matching and quantized edit sentence. This modification required the calculation of the minimum string distance across the mentions in a given coreference chain in addition to the surface distance to the closest mentions. It was reported that the entity-mention model gives slightly lower scores than the mention-pair model. It is worth noting that the mention-pair model uses 20 times more features than the entity-mention model. The latter, however, tries to overcome errors arising from clustering the masculine pronoun and feminine pronoun as the same entity.

ii. Daumé III and Marcu (2005) proposed an entity model based on online learning. The model tries to overcome the problem of non-optimal local decisions by using multiple partial solutions and neglecting partial solutions once they prove to be inconsistent later on in the document.

The model resolves anaphora by aggregating the scores for pairing each anaphor with every antecedent in a single coreference set using various strategies such as max-link (choosing the highest score), min-link (scoring the lowest score), average-link (taking the average score) or the nearest-link (taking the score of the nearest antecedent of the coreference set). The model proposed the use of intelligent-link, which is an aggregation method which considers different mentions separately:

- Proper names undergo a matching process with the most recent document the model dealt with. If it does not match with such most recent document, it is matched against the last nominal or the model resorts to using the highest-scored link.

- Nominals are matched with the previous chain highest-score nominal. Those that do not match are matched against the most recent name or the model resorts to using the highest-scored link.

d. Pronouns are resolved using the average-link against all pronouns or names and if pronouns do not match the model resorts to using the highest-scored link.

The use of mention clusters allows the model to deal with 'decayed destiny'[6], which is a hypothesized entity similar to Leass and Lappin's (1994) salience measure. It captures some entities that are referred to consistently across a given document, while others are mentioned in short segments. This is because, as with the salience measure, some entities are central to a document while some pronominal coreferences are very local.

iii. Rahman and Ng (2009) use a cluster-ranking algorithm which incorporates improvements of the early statistical models of anaphora resolution. The coreference chain that scores the highest is chosen as the antecedent of the mention. The model proposes to relate discourse-novel mentions and anaphora resolution.

[6] It is of a hypothesized entity, it is computed as $\Sigma_{m=e}\ 0.5^{d(m)}/\ \Sigma_m\ 0.5^{d(m)}$ where (m) ranges over all previous mentions (constrained in the numerator to be in the same coreference chain as per mention) and d(m) is the number of entities away from this mention.

4.6.8 Discourse-new Detection

Not all definite noun phrases are considered to be anaphoric; consequently, not all anaphoric noun phrases would have a coreferring antecedent. Coreference resolution systems can benefit from perfect or near-perfect information by deciding which definite noun phrases require to be resolved to a coreferent antecedent and which ones do not. This information helps the resolution system to decide which techniques to adopt in order to deal with common-sense knowledge for resolving definite noun phrases. The information helps in resolving to an antecedent but it does not benefit the system in deciding whether or not a definite noun phrase needs an antecedent.

The information helps in differentiating between discourse-new and discourse-old as well as defining and specifying true anaphoric definite noun phrases by considering ones previously introduced in the discourse. Noun phrases that uniquely specify can occur as discourse-new mentions, and when they occur as a repeated mention the variation is recognized by the surface form between the subsequent mentions.

Vieira and Poesio (1997, 2000) were among the first researchers to use syntactic heuristics in order to differentiate between discourse-old and discourse-new definite noun phrases. Features such as restrictive post modification, capitalization-based heuristics, hand-crafted lists of special nouns, and modifiers indicating uniqueness are used for resolution.

Bean and Riloff (1999) argued that a hand-crafted list of nouns cannot cover all cases, and so they proposed an approach that creates such lists by unsupervised learning. This approach is based on the idea that definite noun phrases in most cases occur with a definite article, whereas anaphoric noun phrases occur in the indefinite variant form.

Bean and Riloff made use of another fact: that the first sentence's mentions are properly nonanaphoric. They made use of this heuristic to help them in compiling a list of nouns that occur as definites in the first sentence of a text. They tried to generalize such lists for the purpose of creating patterns where the presence of the head noun with premodifiers would indicate that a matching noun phrase was uniquely referring. Such patterns would be extended to the longest suffix of a noun phrase that would usually occur as a head in order to increase the specificity of such patterns. Such patterns are called existential head patterns (EHP).

Another fact is the relative frequency of indefinite and definite variants of a noun phrase. This heuristic helps in specifying unique noun phrases which only occur in the definite form and non-unique noun phrases which occur in indefinite form. The advantage of such a heuristic is: full noun phrases and heads that occur five times or more in the training corpus are used to form a list of 'definite-only' noun phrases. The definite/indefinite ratio of a NP is linked to a threshold: if it is above the threshold, the NP is to be considered as always definite. If the noun phrase is below the threshold it would be considered as uniquely specifying, especially if it occurs in the first three sentences of the text.

Ng and Cardie (2002a) use a machine learning classifier for a discourse-new classification. The results of the model are integrated with their Soon et al.-style coreference system. They use features to indicate the existence of a possible antecedent, such as string-matching or head-matching. The pattern-based indicators of the form deal with pre- and post-modification, in addition to the mention's location, whether it be in the first sentence, first paragraph or in the header. Where a mention is not resolved when the results are integrated with the coreference classifier, this is used as an indication that such a mention is discourse-new, which is reflected in an increase in precession that is accompanied by a decrease in recall. When the system starts to resolve string-matching or alias antecedents it is able to compensate for the decrease in recall while the precision rate is maintained.

4.7 Anaphor Resolution in Arabic

Anaphor resolution is a relatively new topic among Arabic linguists, and not much work has yet been done on it.

Before introducing AR in Arabic it is important to understand the position of Arabic as a language with regard to natural language processing (NLP), which AR is part of, as discussed in detail by Farghaly and Shaalan (2009). The Arabic language presents an interesting challenge for NLP. It is interesting because it is a language whose classical form has remained unchanged for more than fifteen centuries spoken by 330 million people who occupy a region extending from the Gulf area to the Atlantic Ocean. The challenges represented by the Arabic language arise from its linguistic nature. This linguistic nature can be described as complex (Attia 2008) due to its diglossia (Diab and Habash 2007) and as a language where morphology plays a vital role (Attia 1999; Beesley 2001; Buckwalter 2004).

NLP applications face complex problems when dealing with the Arabic language in particular (Habash 2007). For instance, Arabic is written from right to left, it has no capitalization, letters change their format according to their position within the word, and short vowels have no orthographic representation in modern standard Arabic (MSA), which demands homographic resolution and word sense disambiguation (WSD). NLP also has to deal with the nature of Arabic being a pro-drop language where the subject can be deleted. Any NLP system dealing with Arabic must take into account such problems and try to resolve them.

Farghaly and Shaalan (2009) claim that Arabic natural language processing (ANLP) has lately gained increased attention and many applications have been developed, such as machine translation (MT), information retrieval (IR), text-to-speech, and document categorization. As most ANLP methods have been developed in the Western world, they tend to focus on enabling non-Arabic speakers to understand Arabic language texts. Most of the tools developed so far have used machine learning approaches which are fast, cheap and do not require complex linguistic knowledge. Machine learning tools usually give good results, especially when the training data is similar to the testing data. ANLP tool developers have had to face problems such as the lack of a corpus for Arabic-named entities, which is a significant tool in NLP research since it allows the identification of proper nouns in open-domain (unstructured) text. However, some trials, such as the LDC in May 2009, implemented an entity translation training test for Arabic, English, and Mandarin Chinese, but there is still a lot to be done. Another problem that Shaalan at al. (2008) noted is the translated and transliterated named entities within Arabic texts. In their research they tried to recognize and extract the ten most important named entities (person names, locations, companies, dates, times, prices, measurements, phone numbers, ISBNs, and file names) in Arabic script. They developed a system called NERA (Name Entity Recognition for Arabic) that is rule-based. NERA included a dictionary of names, a grammar, and regular expression form, in order to be able to recognize the named entities. The evaluation process resulted in satisfactory results in terms of precision, recall, and the F-measure.

The adaptation of Western language tools to Arabic is quite a difficult task, as Choukri (2009) noted, which led the MEDAR consortium to begin an initiative in cooperation with the EU and Arabic-speaking countries to develop ANLP tools and resources (Farghaly and Shaalan 2009).

ANLP applications developed in the Arab world use rule and machine learning approaches. The main aims of such tools in the Arab world are as follows (Farghaly and Shaalan 2009):

i. Knowledge and technology transfer to the Arab world. It is important for Arabic readers and consumers to access science and technology publications published in English or any other language. Human translators are not sufficient in number and their capacities are limited with respect to the translation of such huge amounts of data; ANLP tools help in reducing the time wasted in translation, IR, and text summarizing.

ii. The modernization of the Arabic language; translation into Arabic involves the coinage of new words, and the Arabization of western words. Such linguistic processes help to fulfil commercial needs and renew the language by adding new words to its lexicon and using old words in a new way.

iii. The modernization of Arabic linguistics; MSA requires a more modern grammar than the traditional one; that is, one more in line with current western linguistic theory. This process has two aspects: to preserve the Arabic language heritage, and at the same time provide tools to fulfil modern needs.

iv. Availability of NLP tasks for MT, IR, and text summarization for end users; any technological gaps between the Arab world and the rest of the world can be overcome by making information accessible to the younger Arab generations.

The following sections briefly describe the main problems of anaphora resolution with Arabic along with some of the suggested solutions.

4.7.1 Mitkov (1998)

Mitkov (1998) appears to be the first researcher to have specifically addressed Arabic anaphor resolution. His aim was to develop an AR algorithm that meets the demands of NLP systems operating in real-world and knowledge-poor environments as an alternative to knowledge-based approaches such as those described in the preceding section which have proven to be expensive to develop in terms of both time and money. Mitkov's algorithm relies on a list of preferences known as antecedent indicators. The algorithm 'works from the output of a text processed by a part-of-speech tagger and an NP extractor, locates noun phrases which precede the anaphor within a distance of two sentences, checks them for gender and number agreement with the anaphor and then applies the indicators to the remaining candidates by assigning a positive or negative score (2, 1, 0, or-1). The noun phrase with the highest composite score is proposed as antecedent' (Mitkov 2002: 145).

The algorithm has two main stages:

i. The pre-processing stage includes the use of a sentence splitter, a part-of-speech tagger and noun phrase grammar rules to enable the extraction of the NP in the targeted sentence and the two preceding ones. In later versions of the algorithm the sentence search scope was varied although no complex or embedded clauses were considered.

ii. The resolution stage starts with the sentence being processed by invoking a gender and number filter. This takes into consideration that certain collective nouns in English such as ‘team’ or ‘government’ can be referred to by using ‘they’ whereas plurals such as ‘data’ can be referred to using ‘it’. Then antecedent indicators are applied to successful NPs acting in either a boosting or impeding capacity. Indicators are genre-independent and coherence-related, while with other algorithms they are genre-specific.

The boosting indicators are as follows (Mitkov 2002: 146):

- First noun phrases: a score of +1 is assigned to the first NP in a sentence.

- Indicating verbs: a score of +1 is assigned to those NPs immediately following a verb which is a member of a predefined set (including verbs like ‘analyse’, ‘examine’, ‘discuss’, etc.).

- Lexical reiteration: a score of +2 is assigned to those NPs repeated twice or more in the paragraph in which the pronoun appears, and a score of +1 is assigned to those NPs repeated once in that paragraph.

- Section heading preference: a score of +1 is assigned to those NPs that also occur in the heading of the section in which the pronoun appears.

- Collocation match: a score of +2 is assigned to those NPs that have an identical collocation pattern to the pronoun.

- Immediate reference: a score of +2 is assigned to those NPs appearing in the construction of the form '(You) V1 NP . . . *con* (you) V2 it (con (you) V3 it)', where *con* ∈ {and/or/before/after/until . . . }. This is considered to be a modification of the collocation preference which is highly genre-specific and occurs in imperative constructions, for example:

  'To print the paper, you can stand *the printer* up or lay *it* flat.'

- The noun phrase that is awarded the highest score according to the immediate reference indicator emerges as the correct antecedent. The noun phrase after the V_1 is most properly the antecedent of the pronoun *it*.

- Sequential instructions: a score of +2 is applied to NPs in the NP_1 position of constructions of the form: 'To V_1 NP_1, V_2 NP_2. (Sentence). To V_3 it, V_4 NP_4' where the noun phrase NP_1 is the likely antecedent of the anaphor *it* (NP_1 is assigned a score of 2). For example:

‘To turn on the *video recorder,* press the red button. To programme it, press the ‘Programme’ key’.

- Term preference: a score of +1 is applied to NPs identified as representing domain terms. A small term bank is developed to represent terminology for programming languages and computer hardware. For MARS (Mitkov’s Arabic AR algorithm) it obtains those terms automatically using TF * IDF (term frequency) *(inverse document frequency) (Mitkov 2002)**.**

- Indefiniteness: indefinite NPs are assigned a score of -1.

- Prepositional noun phrases: NPs appearing in prepositional phrases are assigned a score of -1.

If two candidates have the same score, then the candidate with the higher score for immediate reference is selected. Otherwise, the collocational pattern would be the criterion for selection, and, failing that, the candidate with the higher score for indicating verbs and then the most recent candidate is chosen.

Mitkov’s algorithm is claimed to be practical since it does not depend on semantic knowledge or statistical evidence, using only limited syntactic knowledge provided by part-of-speech tagging to give results that match those of the knowledge-based approaches outlined earlier. It was developed and tested with reference to English, but when adapted to Arabic (Mitkov 2002**)** it required only minimal modification and achieved a good success rate.

In Arabic, agreement rules for gender and number filter out antecedent candidates, as in English, but these rules differ in a few respects from those of English. For example, a non-human set of items may be referred to using a singular feminine pronoun. However, agreement rules in Arabic are different from those in English. For instance, Arabic pronouns may appear as suffixes of verbs, nouns and prepositions. The only additional indicator that was used for Arabic was the relative pronoun indicator which depends on the fact that the 'first anaphor following a relative pronoun refers exclusively to the most recent NP preceding it' (Mitkov 2002: 154). The indefiniteness indicator was modified slightly since in Arabic definiteness occurs in a richer variety of forms. The prepositional noun phrase indicator also had to be adapted, because in Arabic the antecedent and the anaphor can belong to the same prepositional phrase, so it was modified as follows: if an NP belongs to a prepositional phrase which does not contain the anaphor, it is penalised by -1, otherwise it is not assigned any score. The referential distance indicator was modified as well, since an anaphor in Arabic tends to refer to the most recent NP. Therefore it would score 2, but if it refers to the one that precedes it, it would score 1, otherwise it scores zero. Mitkov's algorithm was evaluated using two methods: the first method used his robust approach without any modifications made for Arabic. The second method incorporated the modified antecedent indicator mentioned earlier, used to capture specific aspects of MSA. The evaluation was based on a corpus of technical manuals (Minolta Photocopier, Portable Style-Writer (PSW), Alba Twin Speed Video Recorder, Seagate Medalist Hard Drive, Haynes Car Manual, and Sony Video Recorder). Mitkov's original approach achieved a success rate of

77.9% based on 148 out of 190 anaphors being correctly resolved (Mitkov 2002). Mitkov's improved version for Arabic achieved 95.8% success based on 182 out of 190 anaphors being correctly resolved (Mitkov 2002).

4.7.1.1 Evaluation of Mitkov's Original Approach

The approach was evaluated using a success rate that was computed depending on the ratio of correctly resolved anaphora to the number of all anaphora in the corpus (Mitkov 2002) using the texts processed by the POS tagger and NP identifier. The input was manually edited in order to make sure that the input to the algorithm was correct. The English language version was assessed using various technical manuals containing a total of 223 anaphoric pronouns. The algorithm successfully resolved 200 of the anaphora, representing a success rate of 89.7%. Success rates were measured for each technical manual, which proved that results may vary even within the same genre, and indicating that more data needed to be tested. The following table shows the results for each manual.

Table 4.6: Success rates of the knowledge-poor approach on different manuals (Mitkov 2002)

Manual	Number of anaphoric pronouns	Success rate in %

Minolta Photocopier	48	95.8
Portable Style-Writer (PSW)	54	83.8
Alba Twin Speed Recorder	13	100.0
Seagate Medalist Hard Drive	18	77.8
Haynes Car Manual	50	80.0
Sony Video Recorder	40	90.6
All manuals	223	89.7

The critical success rate of the approach was 82% as measured for the Portable Style Writer (PSW) manual, which is represented in table 4.7:

Table 4.7: Comparative evaluation and critical success rate based on the PSW corpus (Mitkov 2002)

Approach	Number of anaphoric pronouns	Success rate in %	Critical success rate

Knowledge-poor approach PSW	54	83.8	82
Baldwin’s CogNIAC	54	75	-
Hobbs’ naïve algorithm	54	71	-

The approach’s critical success rate is 82%. This rate applies to anaphors with more than one candidate for an antecedent after applying number and gender filters. The high success rates indicates that antecedent indicators are efficient with difficult anaphors (having more than one candidate for the antecedent) compared to other models. Table 4.8 below shows the results in comparison to those of other approaches.

Table 4.8: Comparison of the success rates of Mitkov’s knowledge-poor approach with two baseline models (Mitkov 2002)

Approach	Number of anaphoric pronouns	Success rate in %
Knowledge-poor approach	223	89.7
Baseline Most Recent	223	65.9
Baseline Subject	223	48.6

4.7.2 MARS (Mitkov 2002)

MARS is a re-implemented and improved fully automatic version of the algorithm described in the preceding section. It makes use of a functional dependency grammar parser whose purpose is to help prevent the algorithm from treating as anaphora pronouns which are either not anaphoric or fall outside the rules of the algorithm.

Mitkov's robust, knowledge-poor approach was implemented and fine–tuned by Richard Evans (Orasan and Evans 2007), and he subsequently called it MARS (Mitkov's Anaphora Resolution System). MARS depends on its fully automatic FDG (functional dependency grammar) parser. The main improvement in MARS is that it does not depend on pre-edited input which most of the other algorithms do as in the cases of Hobbs (1976, 1978), Dagan (1990, 1995) Mitkov (1998) and Ferrández et al. (1998).

Mitkov (2002) claimed that the development of MARS and the re-implementation of Baldwin's algorithm in addition to Kennedy and Boguraev's (1996) approaches proved that automatic anaphora resolution is a difficult process. Anaphora resolution in the real world requires difficult pre-processing requirements such as POS tagging, named entity recognition, NP extraction, and parsing. These difficulties decrease the success rates of anaphora resolution algorithms.

Conexor's FDG parser was implemented in MARS. This parser provides information concerning dependency relations between words, which helps in the extraction of complex NPs. The syntactic roles of words and information about lemmas are also provided. This resulted in the algorithm being able to recognize non-anaphoric pronouns such as the pleonastic *it*, and occurrences of cataphora or anaphora that do not fall within the scope of the algorithm. Accuracy rates consequently increased as antecedents were not assigned to such pronouns.

The differences between MARS and the original approach are twofold:

1. The addition of three new indicators.

- Boost pronouns which allow pronouns (acting as NPs) to be among the candidates for other pronouns. The advantages of employing pronominal candidates are two-fold. 'Firstly, pronominalised entities tend to be salient. Secondly, the NP corresponding to an antecedent may be beyond the range of the algorithm, explicitly appearing only prior to the two sentences preceding the one in which the pronoun appears' (Mitkov 2002: 166). Consequently, the problem of the correct antecedent existing beyond the scope of the previous two sentences is solved. In the translation process, salient pronouns are often omitted, and by using such an indicator the procedure would not have any effect on the coherence of the translation output. However, such an indicator requires that the algorithm would have access to the antecedent of the pronoun in a transitive manner so that an NP would always be the antecedent of the pronoun. In order to access such information, one or more intervening pronouns must be accessed. As pronominal mentions may reflect their antecedents' salience, pronouns are awarded a bonus of +1.

- Syntactic parallelism is achieved by determining which NP has the same syntactic role as a candidate pronoun, which would then act as its antecedent, by adding a boosting score of +1.
- Within the framework of a document, frequent candidates may occur, and consequently antecedents would be repeated frequently and calculation would be based on such occurrences. In this case, frequent candidates would act as a discussion topic of the document. The three with the highest scores are then boosted with a +1 bonus score.

2. Different preprocessing tools were used, as five of the original indicators were implemented differently.

The first implementation of MARS terms were obtained by identifying words with the ten highest TF*IDF scores (Mitkov 2002). If the antecedent candidates included any of these words it was awarded a score. However, in the latest version of MARS the use of the preference indicator means that the ten NPs with the greatest frequency in a given text are awarded the score if any of them is an antecedent candidate.

MARS is able to distinguish the pleonastic from a non-pleonastic *it*. The successful classification rate is 78.74%, and table 4.9 gives details of the accuracy of this classification.

Table 4.9: The characteristics of the texts used for evaluation of MARS (Mitkov 2002)

Text	Words	Anaphoric pronouns	Non-nominal anaphoric / Pleonastic *it*	Classification accuracy for *it*
ACC	9,753	157	22	81.54
CDR	10,453	83	7	92.86
BEO	7,456	70	22	83.02
MAC	15,131	149	16	89.65
PSW	6,475	75	3	94.91
WIN	2,882	48	3	97.06
SCAN	39,328	213	22	95.32
GIMP	155,923	1 468	313	83.42
Total	247,401	2 263	408	85.54

More recently, MARS has included Kennedy and Boguraev’s (1996) syntax filters. These are applied before activating the antecedent indicators and after the gender and number agreement tests.

MARS operates in five steps (Mitkov 2002).

In step 1, the text is processed using Conexor's FDG Parser (Tapanainen and Järvinen 1997) which determines the POS, lemmas, grammatical number and, most importantly, the dependency relations between words in the text.

Step 2 uses the machine learning method developed by Richard Evans in 2000. Here the identification of anaphoric pronouns is carried out and non-anaphoric and non-nominal instances of *it* are filtered.

In step 3 candidates are extracted from the related NP for each pronoun identified as anaphoric. The candidates then undergo syntactic and morphological filtering. Candidates have to adhere to criteria for several characteristics in order to be selected as possible candidates: they must agree in number and gender with the pronoun and satisfy the syntactic constraints.

Step 4 applies a total of 14 boosting and impeding indicators to the candidate sets. Each indicator assigns a score to each candidate, indicating the algorithm’s confidence in it as a suitable or unsuitable candidate for the anaphor.

In step 5 the candidate with the highest score is selected as the anaphor’s antecedent.

4.7.2.1 Optimisation of MARS

Success rates in Mitkov's original approach are empirically driven, and it has been considered that such results need to be optimised in order to achieve the best success rates. In MARS, the antecedent indicators were optimised using a genetic algorithm developed by Constantin Orasan (Mitkov 2002). The following function is used to calculate the score:

$$score_k = \sum_{i=1}^{i=14} x_{k_i}$$

where *score*$_k$ is the composite score assigned to the candidate *k,* and x_{k_i} is the score assigned to the candidate *k* by the indicator *i* (Mitkov 2000).

The aim of an optimisation process is to look for the set of indicators that scores the maximum. Memory-based learning and perception methods were used to optimise MARS, but it did not perform well, and yielded lower success rates than the optimised version. It was found that a genetic algorithm (GA) is more suitable for the optimisation process. Orasan et al. (2000: 5) claimed that GA are 'search algorithms that imitate the principles of natural evolution as a method to solve parameter optimisation problems where the problem space is large, complex and contains possible difficulties like high dimensionality and noise'.

4.7.2.2 Evaluation of MARS

The MARS corpus consists of eight files taken from software and hardware technical manuals. It has a total of 27,401 words with 2,263 anaphoric pronouns. The latter were classified as 1,709 intrasentential anaphora and 554 intersentential anaphora.

Overall, MARS had a success rate of 59.35%. The use of the genetic algorithm developed by Orasan et al. in 2000 (which Mitkov called the optimised version) increased the rate to 61.55%. There were 238 cases where the antecedents did not exist in the list due to pre-processing errors. The success rate is calculated as a ratio of the anaphora successfully resolved by MARS against the overall number of anaphora that exist in the text. Table 4.10 below gives a detailed account of the MARS evaluation process.

Table 4.10: Success rates for the different versions of MARS (Mitkov 2002)

Files	Old (2000)	MARS								MAX		Baseline	
		Standard				'Optimised'				Sct	Ptl	Recent	Random
		Default	w/o *it* filter	w/o num / gender agr	w/o syn constr	Default	w/o *it* filter	w/o num / gender agr	w/o syn constr				
ACC	33.33	51.59	52.87	35.67	49.04	55.41	55.41	43.31	43.31	73.88	96.18	28.02	26.75
BEO	35.48	60.00	60.00	45.71	60.00	67.14	64.28	50.00	67.14	81.43	95.71	35.71	22.86
CDR	53.84	67.47	68.67	51.81	67.47	75.90	74.69	54.22	74.69	78.31	95.18	36.14	43.37
GIMP	-	57.15	60.42	17.57	57.63	57.83	60.83	18.94	57.22	79.70	91.69	37.80	30.72
MAC	53.93	71.81	69.79	60.40	71.14	75.84	77.85	67.11	76.51	83.89	96.64	51.68	44.97
PSW	64.55	82.67	84.00	80.00	82.67	86.67	90.67	80.00	89.33	92.00	97.33	49.33	45.33
SCAN	-	61.50	62.44	46.48	60.56	63.85	64.79	51.64	63.85	79.81	87.32	32.39	30.52
WIN	33.32	52.08	62.50	39.58	52.08	68.75	66.67	60.42	68.75	81.25	87.50	37.50	18.75
TOTAL	45.81	59.35	61.82	29.03	59.35	61.55	63.68	32.04	60.41	80.03	92.27	37.78	31.82

The MAX column records the maximum success rate that MARS can obtain. The column *Sct* indicates the maximum success rate in resolving a pronoun if the NP representing it is selected, where the maximum reached was 92% due to various factors such as pre-processing errors. The column *Ptl* records partial matching. Two baseline models (unsophisticated basic models; until Soon et al.'s algorithm, Hobbs' naïve algorithm was considered as the standard baseline) were evaluated and recorded in the *Baseline* column. In one model, the most recent candidate was selected as the antecedent, whereas for the other, the antecedent was selected randomly and in both models agreement restrictions were applied. In the *Old* column the results of the implementation of the fully automatic original, though slightly modified, version were recorded.

MARS underwent four different configurations in order to be evaluated. In the *Default* column, the full version of the algorithm was applied without using any filters, constraints of number and gender or identification of pleonastic/non-nominal instances of *it*. The comparison of these results shows that MARS gained around 30% in success rate due to the application of number and gender constraints. Syntactic constraints surprisingly did not increase performance, because of problems with parsing accuracy. The *Standard* column displays the results of each configuration with each text and the success rate achieved. The *Optimised* column records the upper limit of the performance of MARS when the optimal indicator scores were applied. Performance decreased when the recognition module for pleonastic/non-nominal *it* recognition was applied. This was the result of the inaccuracy of the classification required in the application of a new performance measure (Mitkov 200).

4.7.3 Al-Sabbagh (2008)

A thesis by Al-Sabbagh investigated pronominal anaphora resolution in Arabic and English machine translation systems. The motivation for the study was the poor performance of some current MT systems such as: Sakhr, which is a dictionary-based system; Google, which is a statistical machine translation system (SMT) system; and SYSTRAN, which is also an SMT system for Arabic and English AR. Al-Sabbagh attributed the poor performance to the differences between the pronominal systems of English and Arabic regarding gender, number, morphology and grammatical cases.

She then proposed an AR algorithm using a statistical, corpus-based approach that can be described as knowledge-poor, for four distinct reasons:

- Firstly, it uses tokenization for corpus pre-processing and POS tagging is provided by the SVM package designed by Diab et al. (2004).

- Secondly, there is only a minimal use of semantic information manifested in semantic features such as gender, number, rationality and collocational associations between the pronoun agent and its antecedent. Collocational association depends on the relationship between the pronoun agent and the possible antecedent, on condition that it is a noun that semantically matches the pronoun. The semantic features are gender, number and person and these are extracted using monolingual and bilingual semi-automatic algorithms.

- Thirdly, no syntactic information is needed or used; a word-based search space is used instead. It only uses recency, which is an easily depicted discourse-based feature. Al-Sabbagh uses word bands which are considered to be groups of words but not a complete linguistic unit.

  Al-Sabbagh (2008: 152) argued that ‘The minus-20-word search space is found to be the most suitable search space for Arabic AR. Using bands is intended to limit the search space from –20, to –10, to –5, to –2 and to –1, respectively, according to the following algorithm:

  1. The -20 words are divided into two bands of –10 words each. These bands are not necessarily complete linguistic units.

  2. A score is calculated for each minus-10-word band. The score of the band is the summation of the conditional probabilities of the bigrams of the band; each bigram consists of the carrier of the pronoun and a candidate antecedent.

  3. The band of the highest score is chosen to the next step as it is further divided into minus-5-word bands.

  4. The score of each minus-5-word band. The score of the band is the summation of the conditional probabilities of the

bigrams of the band; each bigram consists of the carrier of the pronoun and a candidate antecedent.

5. The band of the highest score is chosen to the next step as it is further divided into 4 bigrams.

6. The score of each bigram is calculated. The score of the band is the summation of the conditional probabilities of the bigrams of the band; each bigram consists of the carrier of the pronoun and a candidate antecedent’.

Al-Sabbagh faced two problems during the development of her AR algorithm.

Firstly, she overcame the sparseness of her data using a linguistically-based approach with the Web as the corpus in order to determine the frequencies of the bigrams and thus to measure the conditional probability (CP) of each bigram (a bigram consists of the pronoun agent and a candidate antecedent) (Al-Sabbagh 2008). CP is related to the problem of the sum total of words in the Arabic documents in the web, and Al-Sabbagh used Kilgarriff and Grefenstette's (2006):

$$\text{Web size} = \frac{The\ Size\ of\ the\ known\ corpus\ *\ Web\ frequencies\ of\ function}{\text{Frequencies of function words in the corpus of known size}}$$

She thereby determined that the total size of Arabic Web documents uploaded in the search engines she used was approximately 4,500,000,000 Arabic words (Al-Sabbagh 2008). Al-Sabbagh used collocational association and conditional probabilities and thus avoided the problem of sparseness of data.

Secondly, there is a scarcity in Arabic of semantic feature taggers and non-pleonastic pronoun identifiers. Al-Sabbagh used monolingual and bilingual bootstrapping algorithms based on Arabic and English cues respectively. These achieved a coverage rate of 59% of the nouns in *Al-Ahram* (an Egyptian newspaper) corpus as a sample of MSA**.** As for the problem of non-pleonastic pronoun identifiers, she used a rule-based algorithm to extract them from the AR input. The algorithm managed to exclude 16% of non-pleonastic pronouns based on Arabic grammatical rules. Al-Sabbagh used no training model, so the output of the algorithm could not be evaluated against it. Instead she used a gold standard evaluation set. This consists of 5,000 pronouns which are manually annotated for anaphoric relations, which is used to evaluate AR-related features and the entire AR algorithm**.** The gold standard is what a native language speaker would consider to be correct. The algorithm achieved a success rate of 87.4%.

The subsequent analysis of errors showed that, firstly, they could be attributed to limitations of search space, POS tagger output and web frequencies. Secondly, the minus-20 window size led the algorithm to cover only 88% of the nouns tested. The window size was designed in such a manner so that it was thought that it would be suitable to cover the previous two sentences prior to the sentence where the anaphor would occur. To overcome this problem Al-Sabbagh tried to increase the window size but found that precision rate decreased. Thirdly, the POS tagger yielded 5% error which decreased to 2% when Al-

Sabbagh’s tokenizer was used. Finally, the web frequencies calculated proved not to be very accurate, as they caused 3% of the errors due to the inability to measure pronoun bands correctly.

4.7.4 Hammami et al. (2009)

Hammami et al. (2009) tried to resolve one of the main AR problems in Arabic, which is the annotation of Arabic corpora so that they can be used in the evaluation and training of AR algorithms. The authors tried to accomplish the annotation of the co-referential chain, which is considered to be very difficult without an appropriate tool. They designed a customized XML-tool which they called AnAtAr, and tested it with a corpus of 77,457 words extracted from newspapers articles, technical manuals, a book on education and a novel. The scheme they used was adopted from Tutin et al. (2000) which is compatible with the MUC scheme. Their proposed tool has the advantage of the automatic detection of Arabic pronouns and it allows human annotators to select several anaphoric pronouns that one antecedent may have.

4.8 Conclusion

This chapter surveys approaches to anaphora resolution developed over the last forty years. The linguistic and psycholinguistics background of various approaches is described. Data driven approaches are discussed. The chapter discusses previous work in Arabic anaphora resolution.

Chapter 5. The Grammar of Arabic *Nafs*

5.1 Introduction

This chapter describes the grammar of the Arabic reflexive *nafs*. Mashharawi (2012) claims that only two detailed studies exist: the first is a small booklet by Nahla (1990), and the second is an M.A. dissertation by Mashharawi (2010) herself. In both, the authors admit that there is a scarcity of resources concerning Arabic reflexives in general and *nafs* in particular. In Arabic grammar textbooks reflexives are explained in a very abridged way. Kremers (1997) is considered the best non-Arabic language account.

5.2 The General Nature and Function of *Nafs*.

Nafs is a feminine noun whose literal meaning is 'soul' and it is used as such in many cases, for example:

لعلك تجد بينها شفاء لنفسك الحائرة

Transliteration: /*lElk tjd bynhA $fA' lnfsk*

Glossing: might-you find between remedy for-self-you

AlHA}rp/

the-worries.

Translation: 'You might find a remedy for your troubles among them' (Kremers 1997: 44).

When *nafs* is used as a noun it may be replaced by a pronoun, and in such a case it would be a third person feminine pronoun. This happens, as Kremers (1997) reports, when *nafs* is used as a reflexive expression, for example:

عاش بين لومه لنفسه و اعتذارها

Transliteration: /*EA$ byn lwmh lnfsh w*

Glossing: lived between blame-him to-self-you and

*AEt*ArhA/*

excusing-it.

Translation: ‘He lived between half-blaming and half-excusing himself’.

In the above example, *nafs* as a reflexive is referred to by the feminine suffix pronoun ها hA that is attached to the noun اعتذار AEt*Ar ‘apology’.

Nafs may be used in such a way as to resemble the English reflexive *himself*, meaning that it may emphasize a noun to denote the meaning of *itself* or *same*. There are two ways of doing this;, firstly, as an appositive to the noun that needs to be emphasized where a suffix is attached to the *nafs* case. Secondly, *nafs* is used with the preposition *bi*’ by, with, in which case *nafs* would mean ‘by himself’ or ‘in person’.

In addition to *nafs* with the meaning of 'soul' being used to paraphrase a personal pronoun, there are also other uses for *nafs* to emphasize a noun's meaning. There are two methods of doing this, which lead to differences in meaning. When *nafs* is attached to a bound pronoun it is used as a reflexive. This is the subject of this thesis.

5.3 The Forms of *Nafs*

Table5.1: The forms of *nafs*

All the forms in table 5.1 can be used with ب /b/, ك/K/ and ل/l/

Basic Form	***Nafs*** /nafs/ /nafos/ /nfs/		
Personal Pronoun, including forms	**Singular**	**Dual** **Common**	**Plural**
First person	نفسي (Masculine and	نفسينا (Masculine and feminine)	أنفسنا/نفسنا (Masculine and feminine)

	feminine) /náfsi/ /nafosiy/ /nfsy/	/nafsínā/ /nafosinaA/ /nfsynA/	/nafsínā/ /nafosinaA/ /nfsynA/
Second person	نفسك (Masculine) /náfsak/ /nafosak/ /nfsk/ نفسكِ (Feminine) /náfsik/ /nafosik/ /nfsk/	نفسكما/نفسيكما (Masculine and feminine) /nafsukumā/ /nafosukumaA/ /nfskmA/	أنفسكم/نفسكم (Masculine) /nafsúkum/ /nafosukum/ /nfskm/ أنفسكن/نفسكن (Feminine) /nafsukúnna/ /nafosukun~/ /nfskn/
Third Person	نفسه (Masculine)	نفسيهما/نفسهما (Masculine and feminine)	أنفسهم/ نفسهم (Masculine) /nafsúhum/

	/náfsuhu/ /nafosuhu/ /nfsh/ نفسها (Feminine) /nafsáhā/ /nafosahaA/ /nfshA/	/nafsuhumā/ /nafosuhumaA/ /nfshmA/	/nafosuhum/ /nfshm/ أنفسهن/نفسهن (Feminine) /nafsuhúnna/ /nafosuhun~/ /nfshn/

5.4 The Uses of *Nafs*

When *nafs* is attached to a bound pronoun it is used as a reflexive pronoun, which is the subject of this thesis as noted above. In MSA, a pronominal suffix attached to a noun may refer to the verb agent and, consequently, it may have a reflexive meaning.

In order to have a reflexive meaning the word *nafs* is used as the object combined with an appropriate genitive suffix.

In MSA, reflexive markers are generally used less often in the first and second persons since there is a very limited risk of misinterpretation, while the use of *nafs* is possible in such constructions where the subject of both the main clause and the subclause is in the first person. Forms of *nafs* are often used after prepositions.

MSA verbs have several forms. The finite form is the most common but nominal infinitives and participles do occur occasionally. All three forms can take a reflexive object, though participles rarely do. As *nafs* is a feminine noun meaning ‘soul’, it has no reflexive meaning in some cases, and can be substituted for a pronoun just like any other name. As mentioned earlier, the pronominal suffix attached to a noun may have a reflexive meaning when it refers to the agent of a verb, where the type of verb would act as a constraint or a marker in order to help in the reference process. *Afal al-qulub* or the ‘perception/cognition verbs’ (such as *raa, ‘*to see’, *wajada* ‘to find’ or ‘perceive’), for instance, have a reflexive meaning when a normal object suffix can refer to the subject. Such verbs take two objects and usually the first is a noun and the second may be a noun, adjective, or a verbal sentence. In the nominal case both objects receive an accusative case, while if the first object is a pronoun it takes the form of a pronoun suffix attached to the main verb. A clause, which acts as a subclause to the main verb, is formed by the two objects. There the first object acts as the subject while the second acts as the predicate. If the subject of the subclause is identical to that of the main verb, an object pronoun suffix is attached to the latter and in this case the object pronoun cannot be reflexive and *nafs* is not used.

In general, reflexive markers are used less often with the first and second persons. Reflexive verbs indicate that the subject is directly affected by the action or indirectly affected by the side-effects of the action.

5.5 The Use of *Nafs* with Finite Verbs

MSA verb objects take the accusative case, but certain verbs take specific prepositions associated with the objects. For example, the verb *raa*, ‘to see’, takes a noun in the accusative case, but the verb *naara*, ‘to look at’, would require the preposition *ila*, ‘to’. The same happens in English, but not in all cases as some Arabic verbs may be assigned the accusative in English and vice versa. The problem of misinterpretation is not likely to occur when the antecedent is local, and so a pronoun is allowed. When the pronoun is not locally interpreted, a reflexive is required. Reflexives that are arguments to finite verbs are bound by a co-argument of that verb. This is identical to the role of reflexives, as it indicates that two arguments are identical if they share the same predicate. Verbs that require a prepositional object rather than an accusative object are often associated with reflexives, especially if they have two identical arguments. MSA allows locally-bound pronouns, since a preposition can introduce an optional argument.

5.6 The Use of *Nafs* with Infinitives

The Arabic infinitive form is comparable to the English gerund, since it is nominal. It can also take a definite article and the positions the noun can take replace the object subclause. In most cases the infinitive verb subject is not expressed, but is considered to be identical to the finite verb subject. If needed, the subject is expressed by adding it to the infinitive in the genitive. This is similar to the situation in the English language where the subject of the gerund can be expressed by the same method. In the case of transitive verbs, the object may be added to the infinitive by modifying the latter to become genitive.

An important point about MSA infinitives needs to be noted before discussing the use of *nafs* with MSA infinitives. MSA infinitives are nominal in form and not verbal, which is similar to the English gerund. This makes the infinitive decline as a noun; taking all of the positions a noun can take in addition to its ability to take a definite article.

In MSA, a noun can be modified by a one-genitive constituent. In order to express the subject and the object of the infinitive it is usually the subject that is put in the genitive form, while the object takes the accusative form or is preceded by the *li* ‘to’ preposition.

When a verb uses a preposition to express its object, the prepositional object can be added to the infinitive verb. Non-obligatory prepositions can be added as well.

When translating a MSA infinitive, care has to be taken that it is translated using a gerund or a verb only, as it tends to have a nominal meaning. For example **hAb* does not only mean ‘to go’ but also ‘to depart’.

When infinitives are used with reflexives, they assign an accusative case to their objects. Infinitive verbs that take a preposition their object occur after the preposition. In cases where the subject is omitted, the infinitive takes a definite article, and not a noun, as modifier. If the subject of the infinitive is omitted, it is assumed that it is the same subject as that of the verb governing the infinitive. This depends on the verb used. Reflexive infinitives can be arguments for other infinitives or if the antecedent of the reflexive is an argument for an infinitive, it may occur in a higher clause, although a distinct subject may occur in between them. With the infinitives it is possible for the reflexive antecedent to be the object.

The uses of *nafs* with infinitives can be summarized as follows. The MSA infinitive form is comparable to the English gerund since it is nominal. It can also take a definite article and the positions the noun can take replace the object subclause. In most cases the infinitive verb subject is not expressed, but is considered to be identical to the finite verb subject. If needed, the subject is expressed by adding it to the infinitive in the genitive. This is similar to the English language, where the subject of the gerund can be expressed by the same method. In the case of transitive verbs, the object may be added to the infinitive by modifying it so as to become genitive.

The MSA noun can be modified by a genitive constituent. Therefore, in order to express the subject and the object of the infinitive, the subject is usually put in the genitive form while the object takes the accusative form or is preceded by the *li* (to) preposition, and the prepositional object can be added to the verb. Arabic infinitives often have a more nominal meaning, although they can be translated as gerunds or verbs. Infinitives allow reflexive use by assigning the accusative to their objects.

For verbs that select a preposition for their objects, the object is placed after the preposition and not added to the infinitive. The infinitive is not noun-modified in the genitive when no subject is expressed, and takes a definite article.

The problem of the usage of reflexives with infinitives is that the reflexive object of the infinitive is not identical to the subject of the governing finite verb. Instead it refers to other arguments of the finite main verb or to the arguments of another infinitive. Also, the reflexive antecedent can be in a higher clause, although a distinct subject may intervene.

5.7 Use of *Nafs* with Participles

Arabic verb participles may be either active or passive, with no distinction being made between past and present participles as in the English language. Arabic participles have three main uses: firstly, as predicative or attributive adjectives; secondly, as nouns in the form of lexicalized participles; and thirdly, as an *al*-accusative when adjoined to the sentence so as to express the state of the action of the main verb. *Al* in that case may refer to both the object and the subject, and it takes the accusative case. The use of a reflexive with a participle is quite rare but may occur. The objects of the participle refer back to their subject, which implies that participles are reflexive predicates. This can be further explained as follows.

i. When the participle is an attributive adjective, it is to be translated using a relative subclause since Arabic involves constructions that do not exist in English.

ii. Participles often occur as lexicalized nouns in Arabic. For instance the English 'nomen agents' such as those endings as in 'reader', 'singer', 'editor', etc. are translated into Arabic using active participles.

iii. When the participle acts as an accusative of state, using the *al*, the latter refers to the object and the subject. Here the participle is joined to the sentence in order to express the condition or the state in which the main verb action is performed. This is equivalent to the predicative adjunct or secondary predicate in the English language. The use of reflexives with participles is rare but may occur.

To summarize the uses of *nafs* with participles, Arabic verb participles are divided into active and passive, with no distinction between past and present participles as in the English language. Arabic participles have three main uses: firstly as adjectives (predicative or attributive); secondly as nouns in the form of lexicalized participles; and thirdly as al-accusatives when a participle is adjoined to the sentence to express the state of the action of the main verb. *Al* in that case may refer to both the object and the subject and it takes the accusative case. The use of a reflexive with the participles is quite rare but it may occur. The participle objects refer back to their subject, which implies that participles are reflexive predicates.

5.8 The Use of *Nafs* with *Afal Al-qulub* (perception/cognition verbs)

Perception/cognition verbs are used to 'signify an act that takes place in the mind' (Kremers 1997). Examples of these kinds of verbs are *raa* 'to see', and *wajada* 'to find' or 'to perceive'. Perception/cognition verbs take two objects, where the first must be a noun and the second may be a noun, an adjective, or a verbal sentence. If the objects are nominal they take the accusative case, and if the first object is a pronoun, it is usually in the form of a suffix to the main verb. The two objects form a small clause which should be considered as a subclause to the main verb. In such cases the first object acts as the subject while the second object acts as its predicate.

When used with *nafs*, perception/cognition verbs do not usually have pronouns since the reflexive takes the position of the object. When *nafs* is used with perception/cognition verbs, the second object can be a sentence. *Nafs* may occur with an infinitive of a perception/cognition verb, and in such cases *nafs* occupies the position of an object.

The use of *nafs* with perception/cognition verbs constructions is common. *Nafs* takes the position of the object for the infinitive, and it follows in the genitive. The replacement of *nafs* with pronouns is possible, but native speakers would consider such sentences to belong to classical rather than modern Arabic.

5.9 The Impersonal Use of *Nafs*

Nafs may occur without the pronoun suffix, in which case it receives a definite article. *Nafs* in such a case indicates an impersonal reference thing with the use of infinitives. In other words, *nafs* usually occurs with a pronominal suffix attached to it. There are cases when *nafs* occurs without such a pronominal suffix, but it would then have a definite article indicating the meaning of an impersonal reflexive; as for example in (Kremers 1997):

لكن لا جدوى من مخادعة نفسي

Transliteration: /*lkn lA jdwY mn mxAdEp nfsy*/

Glossing: but not avail from deceiving self-me.

Translation: ‘But there is no use in deceiving oneself’.

In the above mentioned example, the *nafs* case occurs with the possessive ي *y* which still indicates it is a reflexive.

5.10 In All Other Contexts

Nafs can be the predicate of a nominal sentence, and it will then be bound to the subject of the sentence. Alternatively, it can be an argument of a noun in the form of a genitive, or may occupy the position of a prepositional object. *Nafs* can occur as a predicate of a nominal sentence, in which case the reflexive will be bound to the subject of the sentence.

Nafs can occur in the position of a noun argument, as a genitive or as a prepositional object. *Nafs* can also occur as an argument of an adjective, for example in (Kremers 1997):

فسأله عن أحب أغانيه إلي نفسه

Transliteration: /*fs>lh En >Hb >gAnyh Aly nfsh*/

Glossing: so-ask-him about favourite song-his to self-him.

Translation: ‘And he asked him which of his songs he liked most himself’.

In this example, the reflexive *nfsh* is an argument to the adjective *>Hb*.

Nafs can occur without having an accompanying antecedent in the same clause.

5.11 General Summary

The uses of *nafs* can therefore be summarized as follows.

Arabic verbs have several forms. The finite form is the most common, but nominal infinitives and participles do occur occasionally. All three forms can take a reflexive object though participles rarely do. Since *nafs* is a feminine noun, meaning ‘soul’, and has no reflexive meaning in some cases, it can be substituted for a pronoun just as any other proper noun can. As mentioned earlier, the pronominal suffix attached to a noun may have a reflexive meaning when it refers to the agent of a verb, and here the type of verb would act as a constraint or marker which helps in the reference process. *Afal al-qulub* or the ‘perception/cognition verbs’ (e.g. *raa* ‘to see’, *wajada* ‘to find’ or ‘perceive’, etc.), for instance, have a reflexive meaning when a normal object suffix can refer to the subject. Such verbs take two objects, the first of which is usually a noun and the second may be a noun, adjective, or a verbal sentence. In nominal cases both objects receive an accusative case, while the first object if it is a pronoun takes the form of a pronoun suffix attached to the main verb. A clause which acts as a subclause to the main verb is formed by the two objects, in which the first acts as the subject while the second acts as the predicate. If the subject of the subclause is identical to the subject of the main verb, an object pronoun suffix is to be attached to the main verb which means that the object pronoun cannot be reflexive and *nafs* is not used.

In general, reflexive markers are used less often with the first and second person, consequently misinterpretation cannot occur. Reflexive verbs indicate that the subject is directly affected by the action or indirectly affected by the side-effects of the action.

5.12 Restrictions on the Use of *Nafs*

When the subject is coreferent with one of the arguments, then a reflexive has to be used, for example[7]:

قتل الرجل نفسه

Transliteration: /*qatala r-rajul-u nafs-a-?u*/

Glossing: kill the-man self-him.

Translation: ‘The man killed himself’.

[7] In this part, all transliterations and translations are adopted from Tsukanova and Nikolaeva (2008).

In the above example, the subject الرجل 'the man' and the argument are identical so a reflexive is used.

Arabic reflexives cannot occupy the subject position as, for example, in Tsukanova and Nikolaeva (2008):

قتل نفسه الرجل

Transliteration: /*[8]*qatala* *nafs-u-hu* *r-rajul-a*/

Glossing: kill self-him the-man.

Translation: *Himself killed the man. (Tsukanova and Nikolaeva 2008)

In the above example, it is incorrect because the reflexive نفسه 'himself' cannot act as the subject of the sentence.

Research into Arabic reflexives is relatively scarce. Most studies are concerned with the asymmetry of Arabic anaphora and interaction problems between the c-command and the precedence that determines the distribution of Arabic pronouns (Kremers 1997).

Nafs may be interchangeable with pronominals in some contexts, which raises the problem of defining the binding domains for pronominals and anaphora.

PPs

[8] An * indicates that the sentence is incorrect.

Arabic PPs tend to behave like the English ones, so that when a PP is a complement rather than a reflexive *nafs* can be used. The farther the PP is from a complement, the less it needs a reflexive *nafs*.

Complement PPs are semantically empty. The place to look for PPs is still a complement, but the preposition is empty.

NPs

NPs have their own domains and so pronominals are only allowed there.

5.13 Conclusion

The chapter reviews the various uses of *nafs* including the various forms of *nafs*. The chapter reviews the various cases where *nafs* would occur as a reflexive as with finite verbs, infinitives, participles, verbs of perception, and the impersonal use of *nafs*. The next chapter discusses the algorithm developed by the researcher. It contains the results and interpretation of results, and the conclusion of the thesis.

Chapter 6.The *Nafs* Resolution Algorithm

6.1 Introduction

This chapter proposes an algorithm for the resolution of *nafs* in contemporary Arabic text, referred to for convenience in what follows as NRA (for '*Nafs* Resolution Algorithm'). Given the success of Mitkov's anaphor resolution system for Arabic, it is reasonable to ask why an additional algorithm for *nafs* is required. The answer is that the NRA deals with *nafs* that Mitkov's algorithm did not deal with

The discussion in this chapter comprises five parts. The first part describes the format of the text input and the second the dictionary used by NRA. The third part specifies NRA itself in terms both the abstract algorithm and its implementation using the programming language Delphi. The fourth part tests the implementation of NRA on a corpus of contemporary Arabic and reports the results. The fifth and final part interprets the results.

6.2 NRA Text Input

Input to NRA is assumed to be a collection T of *m* text documents, where:

- each document T_i (for i = 1..m) consists of n strings, where n ranges from 1 to unbounded but finite number.
- each string n_j (for j = 1..n) consists of an arbitrary number of words terminated by a full stop.
- each word consists of a contiguous sequence of alphanumeric characters demarcated by a space character at the beginning and end of each sequence or by a space character at the beginning of the sequence and some form of standard punctuation such as a full stop or a comma at the end.

The documents comprising T are assumed to be transliterated from Arabic orthography into standard Western. The transliteration is essential for two reasons.

Firstly, in contemporary Arabic orthography vowels are not represented but left implicit for the reader to supply using his or her native speaker competence and the utterance context. To humans this is not a problem, but for NRA it imposes an insuperable level of ambiguity. Classical Arabic orthography (Joshi and Aaron 2006) is a cursive script written horizontally from right to left. There are 29 consonant symbols and 3 long-vowel symbols; short vowels are indicated by diacritics placed above or below the consonant symbols. In addition, other diacritics indicate gemination, the indefinite suffix, and various phonetic features. In MSA text only consonants and long vowels are represented in the orthography. Diacritics are omitted, rendering many orthographic forms ambiguous among several lexical types. Disambiguation depends on the reader's knowledge of Arabic and the semantic context provided by the text being read. For example, depending on the context, the word ملك can be read as *mulk*, 'reign', *malik*, 'king', or *malak*, 'angel'. This ambiguity is a significant problem for the computational processing of Modern Arabic text, since the disambiguating information, and semantic context more particularly, are not easily provided in current computational systems. For this reason, work on Arabic NLP such as in machine translation, morphological analysis, stemming, and part-of-speech tagging (Beesley 1996; Abduljaleel and Larkey 2003) has used Arabic text transliterated into Western orthography.

Secondly, the process of transliteration makes the boundaries of the Arabic words explicit and can be easily dealt with. Since the present analysis depends on being able to identify words, a Western transliterated text greatly simplifies the analysis.

When transliterating, it has to be kept in mind that Arabic language has a number of phonemes which have no equivalent in English or other European languages. Transliteration from Arabic to Western orthography is therefore not entirely straightforward. Several transliteration methods have been proposed to represent Arabic characters in various applications -- for example, Al-Misbar and Ajeeb. There is no accepted transliteration standard at present; current methods typically combine two or more Western symbols to approximate the pronunciation of the corresponding Arabic symbol. Alternatively, Western symbols are enhanced in some way.

The obvious approach to the digital representation of Arabic cursive characters is to render them in Unicode, and while there is no problem with this in principle, it would make the implementation of NRA complicated because the programming language used for implementation of the NRA, as described in due course, does not support Unicode. So, only standard ASCII codes are used for the following transliteration scheme.

The scheme used in this study is the Buckwalter scheme that was mentioned earlier on page viii. Table 6.1 gives an example of Arabic text transliterated using the Buckwalter scheme.

6.1 A sample of MSA text transliterated using the Buckwalter scheme

MSA	Transliteration	Translation by Google translate
هل هل هذا ممكن؟ وكيف، ومن الذي يقدر على أن يفعل ذلك؟ وهل بوسع "أنضوني" نفسه أن ينسلخ هكذا، ويخلق عالمه الخاص، أم أن وطاة الذاكرة ، فردية كانت أم جماعية، ستطغى في النهاية على العقل، وتسيطر عليه وتدفعه في اتجاه "التذكر"، والمزيد من التذكر، وبالتالي المعاناة التي تولد صداعا دائما هو ذلك الصداع (الحقيقي والمجازي) الذي يدفع	hl h*A mmknØŸ wkyfØŒ wmn Al*y yqdr ElY >n yfEl *lkØŸ whl bwsE ">nDwny" nfsh >n ynslx hk*AØŒ wyxlq EAlmh AlxASØŒ >m >n wTAp Al*Akrp ØŒ frdyp kAnt >m jmAEypØŒ stTgY fy AlnhAyp ElY AlEqlØŒ wtsyTr Elyh wtdfEh fy AtjAh	Is this possible? And how, who is able to do so? Could "Andoni" itself so that the sheds, and creates his own world, or that the impact of memory, whether individual or collective, in the end to dominate the mind, and controlled and

المخرج (الذي يقوم بدوره الحقيقي في الفيلم) للبحث عن المساعدة لدى طبيب نفسي، وخوض تجربة الجلوس بين يدي هذا الطبيب خلال ثماني عشرة جلسة من جلسات العلاج النفسي الذي يقوم أساسا، على منهج فرويد المعروف في "التحليل النفسي"	"Alt*kr"ØŒ wAlmzyd mn Alt*krØŒ wbAltAly AlmEAnAp Alty twld SdAEA dA}mA hw *lk AlSdAE (AlHqyqy wAlmjAzy) Al*y ydfE Almxrj (Al*y yqwm bdwrh AlHqyqy fy Alfylm) llbHv En AlmsAEdp ldY Tbyb nfsyØŒ wxwD tjrbp Aljlws byn ydy h*A AlTbyb xlAl vmAny E$rp jlsp mn jlsAt AlElAj Alnfsy Al*y yqwm >sAsAØŒ ElY mnhj frwyd AlmErwf fy "AltHlyl Alnfsy	protected in the direction of "Remembrance", and more memory, and thus generate the suffering is always a headache that headache (and figuratively) to be paid director (who is the real turn in the film) to search for help by a psychiatrist, and experience to sit in the hands of the doctor during the eight session of the ten sessions of psychological treatment that is primarily on the approach known in Freud's "analysis psychological "?

6.3 Dictionary

NRA requires access to a dictionary that lists the gender and number of every noun in T. The compilation of such a dictionary for use with the proposed algorithm is a once-only exercise, after which it can be used indefinitely in any application involving anaphora resolution using Arabic plain text and NRA. An excerpt from the dictionary used later in the discussion for testing of NRA is given in Table 6.2 by way of example.

Table 6.2: Dictionary sample

Word	Gender	Number
A$m}zAz	m	s
A$tbAkAt	f	p
A$tbAkhm	m	s
…	…	…

The dictionary is a list of Arabic noun types, giving its gender and number, for each noun. This gender and number information is used by NRA. Morphological variants of words are listed separately to expedite looking up words. For example, كتب /ktb/ ‘to write’ and its morphological variants are separate entries in the dictionary:

كاتب /kAtb/ ‘writer’

مكتبة /mktbp/ ‘library or stationary’

مكتب /mktb/ ‘office or desk'

كتابة /ktAbp/ ‘writing’

كتب /ktb/ ‘books’

كتاب /ktAb/ ‘book’

6.4 NRA

To resolve anaphora in transliterated Arabic plain text, NRA uses two sources of information:

i. the lexical positioning of candidate antecedents in the surface string; and
ii. gender/number agreement between anaphor and candidate antecedents.

The algorithm is as follows, stated as programming language pseudo-code for clarity and precision; the actual code is specified and discussed in the implementation section later in this chapter.

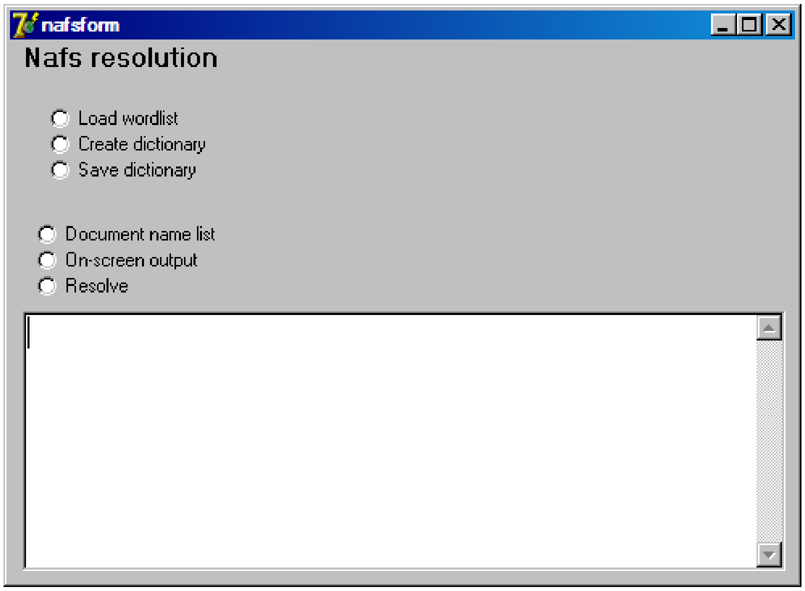
```
For each document $T_i$ in succession, where i = 1..m and m is the number of documents in T
    Begin
        For each string $S_j$ in $T_i$, where j = 1..n and n is the number of strings in $T_i$
            Begin
                For each word $W_k$ in $S_j$, where k = 1..p and p is the number of words in $S_j$
                    Begin
                        If $W_k$ is one of the forms of nafs then
                            Begin
                                Search all the words preceding nafs in the current string for candidate
                                antecedents, that is, nouns compatible in gender and number with the
                                current form of nafs;
                                If one or more candidate antecedents is found then
                                    select the candidate that is lexically furthest from nafs in the string
                                else
                                    Begin
                                        If the current string is not the first in the document, search all the
                                        words in the string preceding the current one for candidate
                                        antecedents;
                                        If one or more candidate antecedents is found then
                                            select the candidate that is lexically furthest from nafs in the string
                                        else
                                            the resolution fails;
                                    End;
                            End;
                    End;
        End;
    End.
```

This algorithm is linear in the length of the document collection to which it is applied. Each document in the collection is read through sequentially once, and for each case of *nafs* the string in which it occurs, and if necessary the string preceding, is read again. In the worst case, therefore, each document is read three times, and its computational complexity is thus O(3*n*), where *n* is the number of strings in the collection. To this must be added a dictionary search for each case of *nafs*, but the dictionary is structured as a binary search tree in order to avoid a computationally intensive sequential search, so that the computational complexity is *O(3n+c)*, where *c* is a constant representing dictionary lookup. The expression *3n+c* has the form of a first degree linear polynomial, which justifies the claim that NRA is linear in text length *n* and thereby that is satisfies the requirement specified out the outset of this discussion: that the proposed *nafs* anaphora resolution algorithm must be efficient in this sense.

The software implementation of NRA used for testing is written in DELPHI, a general-purpose programming language, developed from the teaching language PASCAL**.** DELPHI was selected for two reasons. Firstly, one of the aims of the author of this thesis is to become familiar with computer programming, and DELPHI is ideal for this. It is based on PASCAL, a language explicitly designed for teaching the fundamentals of programming. The researcher is aware of other programming languages such as R and Java that can be used for control mechanisms, primitive data constructions, low-level tasks like data input and output. Secondly, the present author already had some prior knowledge of PASCAL on which the following DELPHI implementation could be built.

The following account of the DELPHI NRA implementation is given in high-level functional terms. Implementation details are provided as part of the program listing in Appendix 1A. User access to the program's operation is via the graphical user interface shown in Figure 6.1

Figure 6.1: Graphical user interface for the *nafs* resolution implementation



Each button in the user interface invokes a separate procedure in the program, as follows:

- ‘Load wordlist’ reads a text file containing a list of nouns, each with associated gender and number information, and stores it in a list data structure.
- ‘Create dictionary’ transforms the word list into a dictionary with a binary tree structure for efficient subsequent searching.
- ‘Save dictionary’ outputs the binary tree structure to a text file in the form of a sorted table for visual inspection where this is convenient or necessary.
- ‘Document name list’ reads a text file containing a list of the filenames of the documents to be processed.

- ‘On-screen output’ writes various types of information into the text box during program execution.
- ‘Resolve’ carries out anaphor resolution on the specified documents.

All but the last of these is generic in the sense that they involve standard text processing computational procedures, and therefore do not require any further discussion. ‘Resolve’, however, does require a description because it implements NRA; implementation details are available in the full program listing in Appendix 1B.

Procedure ‘resolve’:

1. *Parameters*

- Current sentence
- Previous sentence
- Current nafs form
- Lexical dictionary containing gender and number information

2. Output: the current nafs form and its referent, or notification of failure to resolve

3. *Algorithm*

```
For each sentence in the current text
  begin
    Store the sentence preceding the current one in case it's necessary for resolution;
    Read sequentially through the current sentence, allowing for the possibility that there might be more than one instance of nafs;
    When an instance of nafs is found
      begin
```

```
        Assign the necessary grammatical information to the current nafs form using the dictionary;
        Assuming left-to-right processing, look backwards through the current sentence starting with the word left of the current nafs form until the start of the current sentence;
        If no match was found in the current sentence, try looking in the previous sentence using the same procedure as above;
        If no reference was found either in the current or in the preceding sentence,
            write a note to this effect to output
        else
            write the nafs form and its referent to output;
    end;
end.
```

6.5 NRA testing

This section tests the performance of NRA relative to a corpus of contemporary Arabic text. The discussion is in three parts: the first part describes the text corpus and how it was pre-processed, the second part describes the compilation and the structure of the dictionary, the third part tests the NRA on the corpus and reports the results of the testing.

6.5.1 The Corpus

The test corpus C is a collection of texts covering the period 2005-2010 taken from BBC Arabic and Aljazeera websites. The aim was to test the NRA algorithm on a representative sample of MSA newswire. The selected texts cover a range of topics such as politics, the economy, religion and sport. The language variety throughout C is Modern Standard Arabic (MSA).

BBC Arabic is a news portal for TV and radio broadcasts, targeting audiences from the Middle East and North Africa. The service was started in Cairo in 1936 with the explicit aim of offering an Arabic news and current affairs radio service independent of the contemporary Arabic-language British broadcasting, which was held to be biased and propagandistic. In 1996 BBC Arabic was closed due to problems with the Kingdom of Saudi Arabia. In 2008, BBC resumed its work and it launched an Arabic-language satellite channel.

Aljazeera, a television and web-based news and current events service with headquarters in Doha, Qatar is regarded as the BBC's successor. It was launched in 1996 following the closure of BBC Arabic in the wake of Middle Eastern and more specifically Saudi outrage at the inclusion of Hebrew-speaking Israelis in its selection for the first time. Since then Aljazeera has grown in stature as an international news and current affairs outlet focussed on Arabic and more broadly Middle Eastern views of current world events. It was, for example, the only international news network to have correspondents in Iraq during Operation Desert Fox 1998, and has since received several awards and accolades.

C encompasses 1030 texts containing a total of 680,512 words. These texts range in length from shorter reports and essays with an average length of approximately 140 words to longer ones with an average length of 3566 words. Table 6.3 gives a summary of the various categories of text together with average length intervals for each category.

Table 6.3: A summary information of various text categories and their average length intervals in C.

Category	Average length intervals
Politics	700
Economy	650

Sports	500
Religion	400
Art	200

6.5.2 Transliteration

The texts comprising C are in Arabic orthography. These were transliterated using the Buckwalter scheme described earlier. The transliteration was carried out using MADA (Morphological Analysis and Disambiguation for Arabic). MADA is a tool developed by Nizar Habash and MADA operates in stages; one of the stages is to transliterate texts using Buckwalter. The researcher used this tool to transliterate all the texts in corpus C. Habash (2010) argues that MADA's transliteration tool achieves 99.4% accuracy rate. For MADA to process the C corpus texts, all texts had to be converted from Microsoft Word format to UTF. For MADA's transliteration tool to work properly, numbers, diacritics (if any existed), punctuation marks, and Out-of-Vocabulary (OOV) words were removed. It should be noted that MADA adds vowel diacritics which affects the error rate (Diab et al. 2007). Diab et al. noted that a full diacritization scheme performs significantly worse than no diacritization while partial diacritization schemes do not significantly vary in performance from no-diacritization baselines.

It has been argued that MADA is 96% accurate on lemmatization and basic morphological choice; consequently, NRA chooses to use MADA that contains ALMORGEANA morphological analyser to return all nouns included in C. This is quite similar to MARS, which uses Conexor's FDG parser (Mitkov 2002) to return parts of speech morphological lemmas, syntactic functions, and grammatical number, etc. However, to maintain the highest possible accuracy rates the generated noun list is reviewed by the researcher in order to remove words such as *mE* (with) and *byn* (between) that are considered nouns in Arabic. That is why they appear in the noun list.

Habash (2010) claims that MADA is a morphological disambiguation system as it adds lexical and morphological information in one operation while tokenization and stemming are done in a later stage if needed, using TOKAN tool. Habash (2010) notes that MADA differentiates between morphological analysis problems handled by ALMORGEANA analyser and morphological disambiguation in its approach. In the current thesis, the first phase of MADA was the only phase used as no pre-processing beyond that was applied to C which makes NRA a knowledge-poor algorithm. Knowledge-poor in the current thesis follows Mitkov's definition of knowledge-poor that 'avoids complex syntactic, semantic, and discourse analysis' (Mitkov 2002); instead it uses eliminative or preferential techniques.

Figure 6.2 shows a sample of how a text looks after transliteration

Arabic text before transliteration	Text after transliteration
وقد أقر شهزاد (31 عاما) بأنه مذنب في جميع الاتهامات الموجهة له، واعترف للسلطات بأنه تلقى تدريبا على صنع القنابل من حركة طالبان باكستان وتلقى تمويلا منها لتنفيذ الهجوم في ساحة تايمز سكوير. ورد شهزاد على الحكم مطلقا صيحات التكبير وقائلا "استعدوا، لان الحرب مع المسلمين بدات لتوها. هزيمة الولايات المتحدة باتت وشيكة وستحصل في وقت قريب". وقضت المحكمة على شهزاد بأقصى	wqd >qr $hzAd (31 EAmA) b>nh m*nb fy jmyE AlAthAmAt Almwjhp lh، wAEtrf llslTAt b>nh tlqY tdrybA ElY SnE AlqnAbl mn Hrkp TAlbAn bAkstAn wtlqY tmwylA mnhA ltnfy* Alhjwm fy sAHp tAymz skwyr. wrd $hzAd ElY AlHkm mTlqA SyHAt Altkbyr wqA}lA "AstEdwA، lAn AlHrb mE Almslmyn bdAt ltwhA. hzymp AlwlAyAt AlmtHdp bAtt w$ykp wstHSl fy wqt qryb".

عقوبة ممكنة في جميع التهم الموجهة له. من جانبه وصف الادعاء العام الامريكي شهزاد بانه "ارهابي ولم ينتابه الشعور بالندم وخان الوطن الذي اقسم الولاء له وقد نال اليوم العقاب الذي يستحقه".	wqDt AlmHkmp ElY $hzAd b>qSY Eqwbp mmknp fy jmyE Althm Almwjhp lh. mn jAnbh wSf AlAdEA' AlEAm AlAmryky $hzAd bAnh "ArhAby wlm yntAbh Al$Ewr bAlndm wxAn AlwTn Al*y Aqsm AlwlA' lh wqd nAl Alywm AlEqAb Al*y ystHqh".

The motivation for compiling a new Arabic-language corpus is the inadequacy of existing ones for the present purposes. In the field of Arabic NLP, corpus-building has had a low priority historically (Alansary et al. 2007; Parkinson and Farwaneh 2003), though, as the latter have pointed out, Arabic corpus-based linguistic research has recently become more prominent. For example, the 15th annual symposium on Arabic linguistics in 2001 (Parkinson and Farwaneh 2003) included four research papers on Arabic corpus linguistics. Although there has been a significant increase in research interest in corpus-based Arabic linguistics, it remains one of the poorly researched languages from the corpus linguistics point of view (Farghaly and Shaalan 2009).

Due to the lack of a suitable corpus that suits the needs of the research currently carried out in the present thesis, the researcher had to compile a new corpus to suit that need.

6.5.3 The Dictionary

The dictionary was created by abstracting all the nouns from C, creating an alphabetically ordered list, and attaching the associated gender and number information to each noun. The initial stage of abstraction was carried out using the

MADA software created by Nizar Habash at Columbia University (Habash et al. 2010). Table 6.4 gives a sample of MADA output.

Table 6.4: A sample of MADA output

Frequency of the word	Transliterated Noun	Part of speech	Gender	Number	Other forms of the word	English Translation
1208	hw it/he	pron	m	s	huwa	it; he
1142	gyr	noun	m	s	gayor	not; other
1076	Al*yn	pron_rel	m	p	Al~a*iy	who; whom
1047	AlHkwmp	noun	f	s	Hukuwmap	government; administration
1033	AlSHyfp	noun	f	s	SaHiyfap	newspaper
1029	Al>mrykyp	adj	f	s	>amoriykiy~	American
1113	>nfshm	noun	m	p	nafos	selves

In table 6.4, the MADA output shows the frequency; i.e. how many times each word is repeated in the corpus. The table shows the transliteration of each word using the Buckwalter scheme and the different forms a word can be transliterated into. It gives the part of speech of each word; which enables the extraction of nouns to form the dictionary. MADA also provides the English translation for each word

MADA also outputs statistics on occurrences of various parts of speech. For C these are shown in table 6.5:

Table 6.5: MADA statistics for C

Category	Number
Noun	255399
Verb	80396
Prep	72771
Adj	70853
Punc	67139
Noun prop	35478

Conj sub	22789
Pron rel	12440
Noun num	10503
Pron dem	7872
Part neg	5438
Noun quant	5182
Conj	4713
Pron	4583
Part verb	4545
Verb pseudo	3681
Adj comp	3606
Adj num	2883

Adv	2441
0 punc na na	2085
Abbrev	1917
Adv rel	1223
Part focus	557
Part	477
Pron interrog	351
Part restrict	338
Part interrog	226
Adv interrog	203
Part det	184
Part fut	174

Part voc	31
Interj	29
Pron exclam	5

Output from MADA was, in turn, abstracted using a small utility program provided by my supervisor to retain only those features relevant to present purposes, that is, the lexical item and its gender and number. The abstracted MADA output is shown in Figure 6.3 below. This is the word-list used by NRA to create the dictionary.

Figure 6.3: A sample of output from MADA modified using the utility program

MADA output without any modification	MADA output after being modified using the utility program
Gyr m s	$&wn n m s
>nfshm m s	$&wnh n m s
AlHkwmp f s	$&wnhm n m s
AlSHyfp f s	$>fp n f s

In Figure 6.3 the first column shows a sample of MADA output. It is clear that MADA output does not put together words that are under the same root. The second column shows MADA output after using the utility tool; now the nouns are arranged alphabetically and under related roots with the number and gender of each item displayed next to it. The second column is used as an input to the NRA implementation.

6.5.4 Testing

C and the grammatically annotated word list abstracted from it were the input to the implementation of NRA described above. Figure 6.4 gives a sample of the output.

Figure 6.4: Sample output from anaphor resolution of C

```
Document C1

Document C2

Sentence: 30

wlknnA nstmd $rEytnA mn AlEmAl >nfshm wlys w*lk <lY >n ytm
AntxAb Hkwmp tmvl mSAlHnA nHn wlys mSAlH Al<mbryAlyp .
wy&kd Hsn jmEp EwD >n AlnqAbp Alty yr>shA mstqlp En >y Hzb
syAsy wyDyf >n mEZm AlnqAbAt fy bryTAnyA lA tErf swY nqAbp
why AlAtHAd AlErAqy llnqAbAt AlEmAlyp wAlty yr>shA rAsm
whw fy nfs Alwqt nA}b r}ys AlwzrA' >yAd ElAwy AlmfrwD mn .
wyqwl r}ys nqAbp EmAl AlnfT fy AlbSrp <n AlnqAbp brhnt >nhA
qAdrp ElY Alwqwf fy wjh <HdY >kbr $rkAt AlnfT lqd tSdynA l$rkp
kylwj brAwn |nd Alty ttbE $rkp EndmA HAwlt AlAstylA' ElY mqAr
EmlnA bAlAstEAnp bAlqwAt . wyDyf Hsn EwD >n AlnqAbp >jbrt
Al$rkp Alkwytyp AlmtEAqdp mn AlbATn >n tstbdl mn EmAlhA
```

Al>jAnb b|xryn ErAqyyn

Nafs form: >nfshm

Referent: AlEmAl

Document C3

Sentence: 12

. wmE AjtyAz H$wd AlmHtflyn $wArE bgdAd qAm AlbED bDrb >nfshm bslAsl Hdydyp k<HdY AlEAdAt Al$yEyp xlAl EA$wrA' . wtblg *rwp h*A AlAHtfAl fy fbrAyr $bAT whw Alywm Al*y mn AlmtwqE An ttjmE fyh H$wd Dxmp fy krblA' wbgdAd

Nafs form: >nfshm

Referent: AlbED

Document C4

Sentence: 9

wyjd AlnybAlywn >nfshm fy Ezlp En *wyhm w>SdqA}hm bynmA tst>nf AlslTAt AEtqAlAthA

Nafs form: >nfshm

Referent: AlnybAlywn

Document C5

```
Sentence: 14

. yjd Alkvyrwn >nfshm bdwn >w fy >Hsn Al>HwAl yqblwn bwZA}f
lA ttnAsb w$hAdAthm

Nafs form: >nfshm

Referent: Alkvyrwn

Sentence: 38

. w>Sybt nAhd bmrD nfsy HAd bsbb h*A AlwDE Al*y wjdt nfshA fyh

Nafs form: nfsy
Referent: bmrD
```

Each string in each document in the sequence C1 – C1030 is searched for instances of *nafs* and, where found, an attempt is made to identify the antecedent. As shown in Figure 6.4, document C1 contains no instances of *nafs*. Document C2 contains one instance of *nafs* in sentence 30. The sentence in which *nafs* occurs and, the one preceding it, are written in the output to provide a context. This is to enable an assessment of whether the resolution is correct or not to take place. Below the sentences are written the *nafs* form in use and the proposed antecedent. In document C5 there are two instances of *nafs*, and in both cases the antecedents are identified in the sentences in which they occur, so the preceding sentence is not written. This procedure continues to the final document C1030. A complete sequence of the output of NRA for C is given in Appendix 1a.

Each instance in the output sequence was assessed for correctness by direct inspection using the present author's native-speaker competence in Arabic. The results were as follows:

Table 6.6 Resolution results and success rate

Category	Results
Total number of texts in C	1030
Total number of texts with *nafs* instance in it	954
Nafs instances occurrence	1678
Total correct *nafs* resolutions	1535= 91.4% (1448 correct with no exception, 44 with adjectives, 12 with the genitive case, 26 with a conjunction, 5 with number specification)
Total incorrect resolutions	143=8.5%
Success rate	91.4%

A sample listing of results is given in Appendix 3.

6.6 Results interpretation

The aim of this thesis, as stated in the Introduction, has been to design and implement a reliable and efficient resolution algorithm for the anaphor *nafs*, which can be used as a component in a computational system that translates Arabic into some target language in practical, real-world applications. The efficiency of the proposed system, NRA, has already been addressed in the earlier discussion. It remains to assess NRA's reliability. The Introduction took 'reliable' to mean 'that the algorithm should ideally be able correctly to resolve all instances of *nafs* in any text collection to which it is applied, where the criterion for correctness is based on native speaker competence, or, failing this ideal, that it should be able to resolve *nafs* correctly with an accuracy comparable to that of state of the art anaphor resolution systems for languages such as English, which is currently 90% or slightly greater (Mitkov 2002). Table 6.6 shows a success rate of 91.4% for NRA, where the success rate is calculated as a ratio of the successfully resolved instances of *nafs* to the total number of *nafs* occurrences in the corpus. In terms of the stated benchmark for reliability, NRA scores well. Although MARS is a broad-coverage anaphor resolution system, but it does not perform on *nafs* so consequently it is impossible to compare its results with NRA's results. Another important factor for making such comparison impossible is that the published results of the MARS' are no longer available (Al-Sabbagh 2008).

It remains to look at the various types of anaphor structure which NRA was able to resolve successfully in detail, and to identify the structures for which it failed, together with reasons for the failures.

NRA resolved 1448 cases with no exceptions at all. In the correct cases the NRA looked at the dictionary and found the nearest antecedent to *nafs*. The antecedent had to agree in number and gender with *nafs*. The matching between the antecedent and the *nafs* case depends on the Arabic grammar rules where the noun/ adjective agree in number, gender, case and definiteness with the head noun. *Nafs* follows the same rule in the current thesis as its antecedent agrees with it in number and gender. The examples below show how such a rule is applied in C corpus.

1026.buck.txt

Sentence: 15

.rAfq Emr wAldh <lY mydAn AlHrb fy AfgAnstAn bnyp AlgzAp wlknh Al|n bEd snwAt qDAhA wrA' AlqDbAn wAl>slAk wbEd >n blg mn AlEmr SAr mn mdmny qrA'p Alktb wmn bynhA qSS jy ky rAwlynz En tlmy* mdrsp bryTAny yjd nfsh fy mEmEp mErkp Dd qwY Al$r

Nafs form: nfsh

Referent: tlmy*

In the above example NRA succeeds in identifying the antecedent that is _tlmy*_ ‘student’ with the *nafs* case *nfsh* ‘himself’. NRA deals with *nfsh* which is masculine and singular so it looks to the nearest noun that agrees in number and gender with it. It chooses _tlmy*_ because it agrees in number and gender with it.

In the following example, NRA resolves correctly the *nafs* form by referring it to the correct antecedent that is a collective noun. NRA deals with the *nfsha* ‘herself’ which is feminine and singular. NRA chooses _Alm$AEr_, ‘feelings’, that agrees in number and gender with it. This reflects the accuracy of the noun list formed from MADA output which helped in making NRA a success.

1009.buck.txt

Sentence: 12

.whnA yZhr >n >wbAmA yHrS fy xTAbh ElY t>kyd <ymAnh b>n >myrkA hy |xr w>fDl |mAl Al>rD >w Alb$ryp wb*lk yg*y Alm$AEr Alqwmyp Al>myrkyp w$Ewr Al>myrkyyn bAlrsAlp >n Al>myrkyyn $Eb xAS lh rsAlp qdryp t&hlh lqyAdp AlEAlm wtTAlbh b*lk why Alm$AEr nfshA Alty >sA' AlmHAfZwn Aljdd AstglAlhA xlAl AlsnwAt Al>xyrp

Nafs form: nfshA

Referent: Alm$AEr

In addition to the 1448 cases, there are cases which are considered to be correct since Mitkov (2002: 171) stated that ‘a pronoun was considered to be correctly resolved if only part of the NP which represented its antecedent was identified’. NRA successfully resolves 12 cases where the antecedent is part of *idafa* construction, or, ‘genitive construction’, which in Arabic consists of two parts (consecutive and cannot be separated).When the algorithm spots one part it is considered correct as the two parts form one entity. For example:

11.buck.txt

Sentence: 17

. kmA *kr Aljy$ >yDAF >n Almtmrdyn qAmwA bnhb mwAd ElY Alrgm mn >n wkAlAt Al<gAvp nfshA lm tublg En wqwE >y m$Akl

Nafs form: nfshA

Referent: Al<gAvp

In the above example, NRA deals with the *nafs* case *nfsha* ‘herself’ which is feminine and singular and tries to find the nearest noun that agrees in number and gender with it. NRA chooses *Al<gAvp*, ‘aid’, which is part of the construction ‘relief aid’.

Following Mitkov’s principle that if a part of the antecedent is identified it will be considered as a correct incident of resolution, there are 44 cases where NRA identifies the adjective that modifies the antecedent noun as the antecedent of the nafs case. Adjectives in MSA are required to agree in number, gender, case and definiteness with their head nouns. Therefore, they are regarded as one entity. This affects many cases when the selection of the antecedent as a noun and adjective in MSA may have the same orthographical form, unless diacritics are used to show case endings. This might explain why the algorithm in the current thesis sometimes chooses the adjective of the noun as the antecedent for the anaphor as both the noun and the adjective look the same. For example:

43.buck.txt

Sentence: 20

. wyqwl AlkAtb <nh fy kAlyfwrnyA $nt mjmwEp mHAfZp tTlq ElY nfshA mjmwEp AldfE b>mrykA Hmlp <ElAmyp lH$d AldEm Trd Al>mm AlmtHdp mn AlwlAyAt . wyDyf AlkAtb >n AlAntqAdAt ElY Alrgm mn *lk lA t>ty mn AlwlAyAt AlmtHdp fAlrAfDwn llHrb ElY AlErAq y$Erwn bxybp Al>ml lEjz Al>mm AlmtHdp En <yqAf tlk fAlkvyr mn AlbldAn t$tky mn >n Al>mm AlmtHdp nAd tsyTr Elyh Aldwl Algnyp wlA yEb> kvyrA bm$Akl Aldwl fymA yErb n$TA' Hqwq Al<nsAn En Sdmthm lEjz Al>mm AlmtHdp En wqf EmlyAt Alqtl wAsEp AlnTAq fy dArfwr

Nafs form: nfshA

Referent: mHAfZp

In the above example, the *nafs* case is *nfshA* which is feminine and singular. NRA searches and finds the nearest possible antecedent that agrees in number and gender with it. This is *mHAfZp*, ‘conservative’, which is an adjective in Arabic, chosen by NRA because it is feminine and singular. The word *mHAfZp* as an adjective modifies the noun *mjmwEp*, ‘group’, so together they mean a ‘conservative group’. The word *mHAfZp* can also mean governorate with the same orthographical form as the adjective that means ‘conservative’. It only differs in diacritics which are not used since C is written in modern standard Arabic.

Another example is:

170.buck.txt

Sentence: 2

. gyr >n AlmHllyn yqwlwn <n AlmbAlg Alty ytwqE >n ttEhd bhA AljhAt AlmAnHp stkwn >ql mn *lk bkvyr Hyv yEtrf Alms&wlwn Al>fgAn >nfshm b>nhm sykwnwn sEdA' AlHZ lw HSlwA ElY nSf h*A Almblg

Nafs form: >nfshm

Referent: Al>fgAn

Here the *nafs* case is *>nfshm*, ‘themselves’, which is masculine and plural. NRA searches for the nearest antecedent that agrees in number and gender and it chooses *Al>fgAn*, ‘Afghani’. This is an adjective that modifies the noun *Alms&wlwn*, ‘officials’. *Al>fgAn* can be used as a an adjective or it can be used as a noun, which is why NRA chooses it, as it cannot decide if it is used as an adjective or a noun. As previously explained, because no diacritics are used, both the noun and the adjective looks the same. Since the adjective in MSA follows the noun in gender and number, NRA chose it.

There are 5 cases where the algorithm referred *nafs* to a conjunction construction. In MSA the conjunction occurs between two nouns or two verbs or two sentences. So if the algorithm spots one of the two conjunct nouns as the antecedent, it is to be considered as being correct as they represent one identity, albeit in two parts. Since Mitkov (2002) argued that identifying part of the antecedent is considered a correct incident of resolution, therefore the researcher considered NRA's choice to be correct. For example:

233.buck.txt

Sentence: 9

.fy gDwn qAlt jmAEAt Hqwq Al<nsAn <n <dAnp AlqwSy lA tDfy b>y HAl mn Al>HwAl $rEyp ElY mHkmp jwAntnAmw Alty twAjh $kwkA wtHdyA mn jAnb jmAEAt Hqwq Al<nsAn wAlmHAmyn Almdnyyn wAlmEtqlyn >nfshm

Nafs form: >nfshm

Referent: wAlmEtqlyn

In the example above the *nafs* case is *>nfshm*, 'themselves', which is masculine and plural. NRA searches for the nearest possible antecedent and it chooses *wAlmEtqlyn*, 'detainees', which agrees in number and gender with the nafs case. The conjunction و *w* 'and' is attached to the noun *AlmEtqlyn*. The noun *AlmEtqlyn* is joined with the noun and adjective *wAlmHAmyn Almdnyyn*, 'civil lawyers', (masculine and plural) which is joined to the noun genitive construction *jmAEAt Hqwq Al<nsAn* 'human rights organizations' (as an inanimate identity it is considered as male and plural). In MSA the conjunction parts must agree in number and gender with each other so if NRA selects part of the

conjunction structure the researcher considers it correct as the conjunction structure is treated as single identity.

There are 5 cases where the NRA related *nafs* with *tamyiz construction* ‘number specification’. In MSA, number specification agrees with the noun it quantifies, which in such a case is considered correct. NRA can recognize an accusative of specification and comparison and measurement (*tamyiz construction*) which occurs with numbers, as such constructions would agree in number and gender with nafs, but in certain cases this does not work. For example:

367.buck.txt

Sentence: 4

. w*krt wkAlp AnbA' $ynxwA AlSynyp >n AlhjmAt wqEt qbyl Alfjr fy bldp kwjA jnwby $ynjyAnj wbd>t btfyjr qnblp mHlyp AlSnE wbEd *lk fjr >rbEp AntHAryyn >nfshm msthdfyn mkAtb Hkwmyp

Nafs form: >nfshm

Referent: AntHAryyn

In the above example the *nafs* case is *>nfshm*, ‘themselves’, which is masculine and plural. NRA searches for the antecedent that agrees in number and gender with it and selects *AntHAryyn*, ‘suicidal’, which is a number specification for the MSA number *>rbEp* ‘four’. As in MSA, the number and its number specification is considered as one identity which is the reason why it is considered to be correct.

NRA failed to resolve 143 cases. The reasons behind such failures are various and will be discussed in detail in the following section.

The majority of failures (66 cases) occur because MSA nouns can occur as a sequence (using conjunctions between them) or as a chain after each other with no barriers (without any conjunctions). This makes the process of determining the antecedent noun very difficult. For example:

114.buck.txt

Sentence: 3

. wtqwl AlSHyfp <n ElAwy xShA bmqAlp qbyl tslm AlslTp rsmyA lHkwmth mn Al<dArp Almdnyp Al>mrykyp Al>rbEA' wsEY fyhA <lY >n yn>Y bnfsh En AlzEymyn Al*yn yqdmAn AldEm lh whmA twny blyr r}ys AlwzrA' AlbryTAny wjwrj bw$ Alr}ys Al>mryky

Nafs form: bnfsh

Referent: Al>rbEA'

In the above mentioned example the *nafs* case is *bnfsh* ‘by himself’ which is masculine and singular. NRA looks for the nearest possible antecedent and selects *Al>rbEA'*, ‘Wednesday’, which in MSA is masculine singular. NRA does not recognize proper nouns and names. NRA could not realize that the correct antecedent is further back *ElAwy*. In another example:

Sentence: 9

.wybdw >n AltAryx fy AlTryq <lY >n yEyd nfsh kmA ybdw >n sbyl Alxrwj mn Alm>zq msdwd >kvr mn Ay wqt mDY

Nafs form: nfsh

Referent: AlTryq

In the above mentioned example the *nafs* case is *nfsh*, ‘himself’, which is masculine and singular. NRA searches for a possible antecedent that agrees in number and gender with the *nafs* case. NRA selects *AlTryq*, ‘way’, that agrees in number and gender and it does not realize that the correct antecedent is *AltAryx*, ‘history’. This problem could be solved by having more linguistic information as parsing which would require more time and effort.

Another form of failure occurred when verb and noun forms were identical (21 instances). The corpus C is in MSA, which does not use diacritics. If diacritics had been used they would have been removed at the pre-processing stage. Therefore, verbs and nouns can look the same, such as the verb *slm*, ‘surrender’ and the noun *slm*, ‘ladder, peace’. To resolve this problem further semantic analysis must be undertaken which makes the AR more time consuming. For example:

102.buck.txt

Sentence: 5

. wkAn fAyz Alx$mAn hw rAbE mn Hyv slm nfsh msA' Alxmys fy mdynp AlTA}f

Nafs form: nfsh

Referent: slm

As corpus C is extracted from news wire it contains quotations and interviews. Consequently direct speech occurs using *nafs* forms such as *nfsy* and *nfsk*. In such cases the antecedent is the elliptic personal pronoun. NRA cannot identify this, as discussed by researchers such as Chalabi (2004). The researcher suggests that the resolution of this special case, in which a pronoun can be attached to the verb, requires further research. NRA failed in 27 cases to determine the correct antecedent because they were cases of direct speech, for example:

33.buck.txt

Sentence: 10

. kAn Al*hAb llHmAm yEd m$klp kAn ynbgy Elyk >n tntZr <*A >rdt AlHmAm >w

mnthY Alfxr bnfsy l>ny AstTEt twfyr mnzl >wsE . wtsmH h*h AlwZyfp lt$Ay bAlH
>fDl lhA wl>TfAlhA sykwn bAmkAny >n >rslhm <lY >fDl AlmdArs AlxASp w>n y
mnzl wsyArp txSnA nHn . wbynmA tskn t$Ay fy mnzl mn TAbqyn ybdw h*A
Almnzl kAlqSr bAlnsbp lZrwf bw wAlty tEy$ mE fy <HdY qrY Al>kwAx fy AlEASm
Alkmbwdyp bnwm bnh

Nafs form: bnfsy

Referent: Alfxr

In the above mentioned example the *nafs* case is *bnfsy* ‘by myself’. NRA starts to search for a suitable antecedent it selects *Alfxr*, ‘pride. NRA could not realize that the antecedent is a hidden pronoun that is ‘I’ or ‘me’.

There are 11 cases which NRA fails to determine the correct antecedent as the antecedent is part of the *kl mn* structure. To overcome this problem, another algorithm could be developed in order to realize structures as *kl mn* or structures that act as collective identity. For example:

497.buck.txt

Sentence: 18

lkn kl mn yEml ldY Al>mrykyyn yErD nfsh lnfs AlxTr

Nafs form: nfsh

Referent: ldY

In the above mentioned example the *nafs* case is *nfsh* ‘himself’ which is masculine and singular. NRA starts to look for a possible antecedent it chooses *ldY* ‘with’. NRA could not realize that is *kl mn*, ‘each one’, is the correct antecedent.

There are 7 incidents of failure that are due to the plural condition of the antecedent. In Arabic, the feminine plural of inanimate objects can be referred to using plural masculine anaphors. In this case the algorithm could not detect the correct antecedent due to the gender difference. The broken plural in MSA does not abide by the normal laws of plurals. Such cases needed to be altered in the dictionary to allow the algorithm to recognize them as possible candidates. For example:

63.buck.txt
Sentence: 5
. w>$Ar AtHAd AlSlyb Al>Hmr Aldwly <lY >n t$jyE AlmjtmEAt Almnkwbp ElY
AlqyAm bmbAdrAt l<EAnp >nfshm >vnA' AlkwArv >w bEdhA ymvl EnSrA
>sAsyA fy Altxfyf mn wT>p AlkwArv
Nafs form: >nfshm
Referent: Almnkwbyn

In the above example the *nafs* case is *>nfshm*, ‘themselves’, which is masculine and plural. NRA searches for a possible candidate and selects *Almnkwbyn*, ‘affected’. NRA does not realize that the correct antecedent is *AlmjtmEAt Almnkwbp*, ‘affected communities’. *AlmjtmEAt*, ‘communities’, ends with the feminine plural ending and is considered by MADA as a feminine plural therefore it is not a possible candidate. In MSA the inanimate feminine plural can be associated and expressed by using masculine reflexives, nouns and adjectives.

There are 11 cases in which NRA could not find the antecedent. The reasons behind this include differences in number and gender from the *nafs* case, or the antecedent did not exist in the same sentence or the previous sentence. In the case of broken plurals, adjectives are singular in form with an *ad hoc* form-based gender, which explains cases where the algorithm could not find the antecedent in the sentence even though it did exist. However, it differed in number and gender from the antecedent. Often the adjectives of broken plural nouns are feminine singular. For example:

185.buck.txt
Sentence: 5
. w>DAft >nh ytEyn >yDA mnAq$p tlk AlqDAyA bSrAHp byn Al$Ewb AlErbyp >
Nafs form: >nfshA

No referent found

In the above example the *nafs* case is *nfshA*, ‘herself’, which is feminine and singular. NRA could not find a suitable antecedent in the sentence or the sentence preceding it. NRA could not determine that the correct antecedent is *Al$Ewb AlErbyp*, ‘Arabic nations’. The reason for this is that the MSA noun *Al$Ewb* is a collective noun which takes the form of the singular, which can be expressed using feminine singular reflexives, nouns and adjectives.

6.7. Conclusion

This thesis addressed the following research question:

> *Can an algorithm be found for the resolution of nafs in Arabic text which is accurate to at least 90%, scales linearly with text size, and requires a minimum of knowledge resources?*

In order to address this question, a two-stage methodology was used. First, a survey of the existing anaphor resolution literature was conducted where the various approaches found were discussed regarding their computational complexity, where complexity was assessed in terms of the accuracy, scaling behaviour, and knowledge requirements specified in the research question. Second, an algorithm was built and tested with a corpus of contemporary Arabic text. This chapter summarizes the findings and limitations of the study and suggests recommendations for further research.

The answer to the research question is positive:

- The proposed algorithm, NRA, yielded resolutions of antecedents of pronouns attached to *nafs* in a corpus of contemporary Arabic with a 91.4% success rate. This success rate exceeds the 90% rate widely accepted as a benchmark in the anaphor resolution literature.
- NRA scales linearly with text size.

- The only knowledge resources required by NRA in addition to the surface strings of the corpus being processed are transliteration from Arabic to Western orthography. They include insertion of the vowels which the former omits, and a compilation of a dictionary listing gender and number information for lexical entities in the corpus.

In terms of success rate, scaling, and knowledge resources,NRA achieves a success rate of 91.4%. It is worth mentioning that MARS deals with Arabic pronouns but does not cover *nafs* that NRA covers, which makes the comparison between the two systems unfair.

Al-Sabbagh's algorithm tries to resolve Arabic pronouns. She uses a statistical, corpus-based approach. Al-Sabbagh's algorithm achieves a performance rate of 87.4%. Al-Sabbagh's algorithm did not deal with *nafs*. Al-Sabbagh uses newswire as a corpus as in the case of the current thesis.

NRA would be regarded as knowledge-poor algorithm for three valid reasons. First, it uses the least linguistic resources. It only uses the output of MADA as an input for the corpus preprocessing stage. Second, it requires the least semantic knowledge that can be represented in the semantic features of gender and number. Third, no syntactic knowledge is needed since it uses an abstracted dictionary of nouns. In other words, no knowledge-rich features are used.

Test results have identified several problems with NRA.

- A further problem might be that the referent might be in a sentence preceding the current one or the one before it, earlier in the text. A simple solution for such a problem is to expand the scope of the search to include more preceding sentences.
- Pronouns and anaphora: MSA has a larger system of pronouns than English. This reflects on the problem of translating dual pronouns such as *they* and *we* into English. The problem can be partially resolved by number and gender specifications provided by MADA.

- Proper names need to be distinguished from other nouns. In MSA, NRA does not recognize some proper names if they are not distinguished from nouns or prepositions. NRA mistakes the proper name for an adjective or a preposition; for example على ElY, ‘over’, and علىّ ElY~, ‘Ali’, which is a proper name. This has to be manually edited in some cases. A possible solution is to create a proper name database which includes gender specification.
- Common nouns and anaphor: MADA’s output does not correctly specify the gender of the noun. This has to be corrected manually. MSA contains a number of nouns and variants have to be dealt with carefully when specifying gender.

The NRA algorithm is, to the researcher’s knowledge, the first to deal specifically with the resolution of the grammatically important particle *nafs* in Arabic. The problems identified while testing it on a corpus of contemporary Arabic are in principle amenable to resolution with further development. NRA is therefore a substantial contribution to Arabic natural language processing.

Apart from the refinement of the NRA algorithm by resolution of the problems discussed above, a potentially productive direction for further work on anaphora resolution in Arabic is to see whether the approach which underlines NRA, that is, lexical positioning in surface strings without recourse to grammatical knowledge apart from gender and number, can be more generally applied to the problem.

References

Abdelali, A., Cowie, J., and Soliman, H. 2005. 'Building a modern standard Arabic corpus'. Paper presented at *the Workshop on Computational Modeling of Lexical Acquisition, The Split Meeting*. Croatia, 25 - 28th of July 2005.

AbdulJaleel, N. , and Larkey, L. S. 2003. 'Statistical transliteration for English-Arabic cross language information retrieval'. Paper presented at *Conference on Information and Knowledge management (CIKM-03)*. USA, 3 - 8th November 2003.

Abdul-Mageed, M. 2009. *Natural Language Analyzer and Processor (NLPA).* Available online at: http://ella.slis.indiana.edu/~mabdulma/NLAnalyzer/scripts/AraUTF8ToBuck.html [Accessed: 9 November 2013].

Abdul-Rauf, M. 1977. *Arabic for English Speaking Students*. Cairo: Shorouk Interantional.

Adams, R. 2003. *Perceptions of innovations: exploring and developing innovation classification*, unpublished PhD thesis, School of Management, Cranfield University.

Ahn, D., Jijkoun, V., Mishne, G., Muller, K., Rijke, M., and Schlobach, S. 2004. 'Using Wikipedia at the TREC QA track'. *Proceedings of the Thirteenth Text Retrieval Conference.* Gaithersburg, MD. 16 - 19th November 2004.

Ahn, K., Bos, J., Curran, J., Kor, D., Nissim, M., and Webber, B. 2005. 'Question answering with QED at TREC-2005'. *Proceedings of* *the Fourteenth Text Retrieval Conference*. USA, Gaithersburg, MD. 15-18th November 2005.

Al-Ahram. Available online at: http://www.ahram.org.eg/ [Accessed: 9 November 2013].

Al-Ansari, A. 1987. *Awdah Al-masalik ila Alfiyat Ibn Malek*. Beirut: Dar Ihya' Al-Ulum.

Alansary, S., Nagi, M. and Adly, N. 2007. 'A semantic-based approach for multilingual translation of massive documents'. Paper presented in *the 7th International Symposium on Natural Language Processing (SNLP)*. Thailand, Pattaya, 13-15th December 2007.

Alfiky, S. 2000. *Elm Al-Lugha Al-Nasi Baiyn Al-Nazari'a w Al-tatbeeq (Text Linguistics: Theory and Practice: A Case Study of Mecca Suras)*. Cairo: Daar Qiba'a.

Algilayyeny, M. 2003. *JameE Aldruws AlErabyp (Arabic Lessons Collection)*. Beirut: Almaktebh Alesri'a.

Al-Hafez, M., Clarke, M., and Vella, A. 1994. 'A semantic knowledge-based computational dictionary'. P*roceedings of Machine Translation: Ten Years on the Second International Conference.* England, Cranfield University, 12 - 14th November 1994. pp 1-13.

Alhashemy, A. 2000. *Al-Quawaed Al-Asasiy'a Lillugha Al-Arabia (Basic Rules of the Arabic Language)*. Beirut: Daar Al-Kitab Al-Alamyia.

Aljazeera. Available online at: www.aljazeera.net/ [Accessed: 9 November 2013].

Allen, J. 1995. *Natural Language Understanding*. Wokingham: Benjamin/Cummings.

Almisbar. Available online at: http://www.almisbar.com/salam_trans.html [Accessed: 9 November 2013].

Almor, A. 1999. 'Noun-phrase anaphora and focus: the informational load hypothesis'. *Psychological Review* 106:748-765.

Al-Sabbagh, R., and Elghamry, K. 2007. 'Arabic anaphora resolution: A distributional, monolingual and bilingual approach'. *Proceedings of the Information and Communication Technologies International Symposium (ICTIS'07).* Morocco, Fez.

Al-Sabbagh, R. 2008. *Pronominal anaphora resolution in Arabic/English machine translation systems*, published MA thesis, Ain Shams University.

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

Alshawi, H. 1990. 'Resolving quasi-logical forms'. *Computational Linguistics* 16:133-144.

Alshawi, H. (ed.) 1992. *The Core Language Engine*. Cambridge: MIT Press.

Altintas, K., Can, F., and Patton, J. 2007. 'Language change quantification using time-separated parallel translations'. *Literary and Linguistic Computing* 22:375-393.

Anderson, A., Garrod, S., and Sanford, A. 1983. 'The accessibility of pronominal antecedents as a function of episode shifts in narrative text'. *The Quarterly Journal of Experimental Psychology* 35:427-440.

Anirban, D., Petros, D., Boulos, H., Vanja, J. and Michael, W. 2007. 'Feature selection methods for text classification'. *Proceedings of the on Knowledge Discovery and Data Mining*. USA, California, San Jose, 12 - 15th August 2007. pp 230-239.

Aone, C., and McKee, D. 1993. 'A language-independent anaphora resolution system for understanding multilingual texts'. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. USA, 22 - 26th June 1993. pp 156-163.

Aone, C., and Bennett, S. 1995. 'Evaluating automated and manual acquisition of anaphora resolution strategies'. *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*. USA, Massachusetts, 26 - 30th June 1995. pp 122-129.

Ariel, M. 1990. *Accessing Noun-Phrase Antecedents.* London: Routledge.

Arnold, D. 1994. *Machine Translation : An Introductory Guide*. Manchester: NCC Blackwell.

Arnold, J., Eisenband, J., Brown-Schmidt, S., and Trueswell, J. 2000. 'The immediate use of gender information: eyetracking evidence of the time-course of pronoun resolution'. *Cognition* 76:13-26.

Artstein, R. , and Poesio, M. 2008. 'Inter-coder agreement for computational linguistics'. *Computational Linguistics* 24:555-596.

Asher, N. , and Lascarides, A. 1998. ‘Bridging’. *Journal of Semantics* 15:83-113.

Attia, M. 1999. *A large scale computational processor of Arabic morphology and application,* published MA thesis, Cairo University.

Attia, M. 2006. ‘An ambiguity-controlled morphological analyzer for modern standard Arabic modelling finite state networks’. Paper presented at *Challenge of Arabic for NLP/MT Conference 2006*. UK, London, October 2006.

Attia, M. 2007. ‘Arabic tokenization system’. *ACL-Workshop on Computational Approaches to Semitic Languages*. Czeck Republic, Prague.

Attia, M. 2008. ‘A unified analysis of copula constructions in LFG’. In *13th International LFG Conference*. USA, Stanford: CSLI Publications.

Badawi, S. 1973. *MstwyAt AlErbyp AlmuEASirp fy Misr (Modern Arabic Levels in Egypt)*. Cairo: Dar Al-Maaref.

Badawi, E., Carter, M. G., and Gully, A. 2004. *Modern Written Arabic : A Comprehensive Grammar*. London: Routledge.

Baker, M. 1992. *In Other Words: A Course Book in Translation*. London: Routledge.

Baldwin, B., Reynar, J., Collins, M., Eisner, J., Ratnaparkhi, A., Rosenzweig, J., and Sarkar, A. 1995. ‘University of Pennsylvania: description of the University of Pennsylvania system used for MUC-6’. *Proceedings of the 6th Message Understanding Conference*. USA, Maryland, Columbia, 6 - 8th November 1995.

Baldwin, B. 1997. ‘CogNIAC: high precision coreference with limited knowledge and linguistic resources’. Paper presented at *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*. Spain, Madrid, 11th July 1997. pp 38-45.

Barker, C. 1991. *Possessive Descriptions*. Stanford University: CSLI Publications

Barwise, J., and Perry, J. 1983. *Situations and Attitudes.* Cambridge: MIT Press.

British Broadcasting Channel. Available online at: www.bbc.co.uk/arabic/ [Accessed: 9 November 2013].

Bean, D., and Riloff, E. 1999. ‘Corpus-based identification of non-anaphoric noun phrases’. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.*USA, College Park, MD., 20 - 26th June 1999.

Bean, D., and Riloff, E. 2004. ‘Unsupervised learning of contextual role knowledge for coreference resolution’. Paper presented at *Proceedings of the Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL).*USA, Boston.

Beesley, K. 1996. ‘Arabic finite-state morphological analysis and generation’. Paper presented at *Proceedings of COLING-96, the 16th International Conference on Computational Linguistics*. Denmark, Copenhagen.

Beesley, K. 2001. ‘Finite-state morphological analysis of Arabic at Xerox research: status and plans in 2001’. *Proceedings of the Workshop on Arabic Natural Language Processing at the 39th*

Annual Meeting of the Association for Computational Linguistics (ACL'01). France, Toulouse, 6 - 11th July 2001. pp 1-8.

Bell, E. 1934. 'Exponential numbers, Amer'. *Math* 4:411-419.

Bengtson, E, and Roth, D. 2008. 'Understanding the value of features for coreference resolution'. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. USA, Hawaii, Honolulu. pp 294-303.

Berger, A., Pietra, S., and Pietra, V. 1996. 'A maximum entropy approach to natural language processing'. *Computational Linguistics* 22:39-72.

Bergsma, S. 2005. 'Automatic acquisition of gender information for anaphora resolution'. *Proceedings of 18th Conference of the Canadian Society for Computational Studies of Intelligence*. Canada, Victoria, B.C. pp 342-353,

Berk, L. 1999. *English Syntax: From Word to Discourse*. Oxford: Oxford University Press.

Berland, M., and Charniak, E. 1999. 'Finding parts in very large corpora'. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. USA, College Park, MD., 20 - 26th June 1999. pp 57-64.

Beun, R., and Cremers, A. 1998. 'Object reference in a shared domain of conversation'. *Pragmatics and Cognition* 6:121-152.

Biber, D., Susan C., and Reppen, R. 1998. *Corpus Linguistics Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Blum, A., and Mitchell, T. 1998. ‘Combining Labeled and Unlabeled Data with Co- training’. *Proceedings of the Workshop on Computational Learning Theory*. USA, Wisconsin. pp 92-100

Bobrow, D., and Berkeley, E. 1964. *The Programming Language LISP: Its Operation and Applications*. Cambridge: MIT Press.

Bod, R., Hay, J., and Jannedy, D. (eds.) 2003. *Probabilistic Linguistics*. Cambridge: MIT Press.

Boguraev, B. 1979. *Automatic resolution of linguistic ambiguities*, published PhD thesis, University of Cambridge.

Bond, F., Ogura, K., and Kawaoka, T. 1995. ‘Noun phrase reference in Japanese-to-English machine translation’. *Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI '95).* Belgium, Leuven, pp 1-14.

Boot, P. 2006. ‘Decoding emblem semantics’. *Literary and Linguistic Computing* 21:15-27.

Borg, I., and Groenen, P. 2005. *Modern Multidimensional Scaling: Theory and Applications*. USA: Springer.

Bowker, L. 2002. *Computer-Aided Translation Technology : A Practical Introduction*. Ottawa: University of Ottawa Press.

Boyd, A., Gegg-Harrison, W., and Byron, D. 2005. ‘Identifying non-referential it: a machine learning approach incorporating linguistically motivated patterns’. *Proceedings of the ACL Workshop on Feature Selection for Machine Learning in NLP*. USA, Ann Arbor. pp 40-47.

Branco, A. 2002. *Anaphora Processing: Linguistic Cognitive and Computational Modelling*. Amsterdam: John Benjamins Publishing.

Branco, A. 2007 (ed.) 'Anaphora: analysis algorithms and applications'. *Proceedings of 6th Discourse Anaphora and Anaphor Resolution Colloquium, (DAARC'07)*. Portugal.

Bransford, J., Barclay,J., and Franks, J. 1972. 'Sentence memory: a constructive vs. interpretive approach'. *Cognitive Psychology* 3:193-209.

Brashi, A. 2005. *Arabic Collocations: Implications for Translation*. Published PhD thesis, University of Western Sydney.

Bravo, J. (ed.) 2004. *A New Spectrum of Translation Studies*. Valladolid: Universidad De Valladolid.

Brennan, S., Friedman, M., and Pollard, C. 1987. 'A centering approach to pronouns'. *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*. USA, Stanford, California, 6 - 9th July 1987. pp 155-162.

Broadbent, D. 1973. *In Defence of Empirical Psychology*. London: Methuen.

Brown, P., Pietra, S. , and Mercer, R. 1993. 'The Mathematics of Statistical Machine Translation: Parameter Estimation'. *Computational Linguistics* 19:263-311.

Buckwalter, T. 2002. *Buckwalter Arabic Morphological Analyzer Version 2.0*. Philadelphia: Linguistic Data Consortium.

Buckwalter, T. 2004. 'Issues in Arabic orthography and morphology analysis'. Paper presented at *The Workshop on Computational*

Approaches to Arabic Script-based Languages, (COLING-2004). Switzerland, Geneva, 23 - 27th August 2004.

Buitelaar, P. 1988. *CoreLex: Systematic Polysemy and Underspecification*, published PhD dissertation, Department of Computer Science, Brandeis University.

Bunescu, R. 2003. 'Associative anaphora resolution: a web-based approach'. *In Proceedings of the EACL 2003 Workshop on the Computational Treatment of Anaphora*. Hungary, Budapest, April 2003.

Bunescu, R, and Pasca, M. 2006. 'Using encyclopedic knowledge for named entity disambiguation'. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. Italy, Trento, 3 - 7th April. pp 9-16.

Burch, C., and Osborne, M. 2003. 'Statistical Natural Language Processing'. In A. Farghaly (ed.) *A Handbook for Language Engineers*. Center for the Study of Language and Information. 1-3.

Bussmann, H. 1996. *Routledge Dictionary of Language and Linguistics*. London; New York: Routledge.

Butler, C. 1992. *Computers and Written Texts*: *Applied language Studies*. Oxford, UK ; Cambridge: B. Blackwell.

Byron, D. 2001. 'The uncommon denominator: A proposal for consistent reporting of pronoun resolution results'. *Computational Linguistics* 27:569-577.

Byron, D. 2002. 'Resolving pronominal reference to abstract entities'. *Proceedings of the 40th Annual Meeting of the Association for*

Computational Linguistics. USA, Philadelphia, Pennsylvania, 7-12th July 2002. pp 80-87.

Carbonell, J., and Brown, R. 1988. ‘Anaphora resolution: a multi-strategy approach’. *Proceedings of the 12th International Conference on Computational Linguistics (COLING'88).* Hungary. pp 96-101.

Carletta, J. 1996. ‘Assessing agreement on classification tasks: the kappa statistic’. *Computational Linguistics* 22:249-254.

Carlson, G. 1977. *Reference to kinds in English*, published PhD thesis, University of Massachusetts.

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

Chalabi, A. 2001. ‘Sakhr web-based Arabic-English MT engine’. Paper presented at *Proceedings of the Association for Machine Translation in the Americas (AMTA'98)*. France, Toulouse.

Chalabi, A. 2004. ‘Elliptic personal pronoun and MT in Arabic’. Paper presented at *JEP-2004-TALN 2004 Special Session on Arabic Language Processing-Text and Speech*. Morocco, Fez.

Chamberlain, J., Poesio, M., and Kruschwitz., U. 2008. ‘Phrase detectives-a web-based collaborative annotation game’. *Proceedings of International Conference of Semantic Systems (I-Semantics 08).* Graz.

Champollion, L. 2006. ‘On the (ir)relevance of psycholinguistics for anaphora resolution’. *Ambiguity in Anaphora Workshop Proceedings ESSLLI.* Spain, Málaga, 7 - 11th August 2006. pp 13-21.

Charniak, E. 1972. *Towards a model of children's story comprehension*, Published PhD Thesis, MIT.

Charniak, E. 1997. ‘Statistical parsing with a context-free grammar and word statistics’. *Proceedings of the 14th National Conference on Artificial Intelligence*. Providence, R.I., 27 - 31st July 1997.

Chen, H. 1992. ‘The transfer of anaphors in translation’. *Lit Linguist Computing* 7:321-328.

Cheng, H. 2001. *Modelling aggregation motivated interactions in descriptive text generation*, published PhD thesis, Edinburgh University.

Chinchor, N. 1998. ‘Overview of MUC-7/MET-2’. In B. Sundheim (ed.) *Proceedings of the 7th Message Understanding Conference (MUC-7)*. USA, Morgan Kaufmann, San Mateo, CA. Available online at: http://www.itl.nist.gov/iaui/894.02/related.projects/muc/-procceedings/muc_7_procceedings/ overiew.html [Accessed: 9 November 2013]. ORRR USA,Virginia, 29 April - 1st May 1998

Chomsky, N. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, N. 1986. *Barriers*. Cambridge: MIT Press.

Choukri, K. 2009. ‘MEDAR: Mediterranean Arabic language and speech technology: inventory of the HLT products, players, projects and language resources’. *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR’09)*. Egypt, Cairo, 22-23rd April 2009.

Chowdhury, G. 2003. ‘Natural language processing’. *Annual Review of Information Science and Technology* 37:51-89.

Cimiano, P., Schultz, A., Sizov, S., Sorg, P., and Staab, S. 2009. ‘Explicit vs. latent concept models for cross-language information retrieval’. *Proceedings of the 21st International Joint Conference on Artificial Intelligence*. USA, Pasadena, 14 - 17th July 2009. pp 1513-1518.

Clark, H. 1977. ‘Bridging’. In N. Johnson-Laird and P. Wason (eds.) *Thinking: Readings in Cognitive Science*. London and New York: Cambridge University Press. 411-420.

Clark, H., and Marshall, C. 1981. ‘Definite reference and mutual knowledge’. In A. Joshi, B. Webber and I. Sag (eds.) *Elements of Discourse Understanding*. New York: Cambridge University Press.

Clark, A., Fox, C., and Lappin, S. 2010. *The Handbook of Computational Linguistics and Natural Language Processing*. Oxford: Wiley-Blackwell.

Clear, J. 1993. ‘The British national corpus’. In P. Delany and G. P. Landow (eds.) *The Digital Word: text-based computing in the humanities*. Cambridge, Mass: MIT Press. 163-187.

Clifton, C., and Ferreira, F. 1987. ‘Discourse structure and anaphora: some experimental results’. In M. Coltheart (ed.) *Attention and Performance XII: The Psychology of Reading*. Hove, UK: Lawrence Erlbaum. 635-654.

Clopper, C., and Paolillo, J. 2006. ‘North American English vowels: A factor-analytic Perspective’. *Literary and Linguistic Computing* 21 (4):445-462.

Cohen, P. 1984. ‘The pragmatics of referring and the modality of communication’. *Computational Linguistics* 10:97-146.

Collins, M. 1997. ‘Three generative, lexicalised models for statistical parsing’. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain, 7 - 12th July 1997. pp 16-23.

Cooper, R. 1996. ‘The role of situations in generalized quantifiers’. In S. Lappin (ed.) *Handbook of Contemporary Semantic Theory*. UK, Blackwell. 65-86.

Cornish, F. 1986. ‘Anaphoric pronouns: under linguistic control or signalling particular discourse representations?’. *Journal of Semantics* 53:233-260.

Coulthard, M., and Baubeta, P. (eds.) 1996. *The Knowledges of the Translator : From Literary Interpretation to Machine Classification*. Lewiston: E. Mellen Press.

Crawley, R., Stevenson, R., and Kleinman, D. 1990. ‘The use of heuristic strategies in the comprehension of pronouns’. *Journal of Psycholinguistic Research* 19:245- 264.

Csomai, A., and Mihalcea, R. 2008. ‘Linking documents to encyclopedic knowledge’. *IEEE Intelligent Systems Special issue on Natural Language Processing for the Web* 23:34-41.

Cucerzan, S. 2007. ‘Large-scale named entity disambiguation based on Wikipedia data’. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learnin.* Czech Republic, Prague, 28-30th June 2007. pp 708-716.

Culotta, A., Wick, M., and McCallum, A. 2007. ‘First-order probabilistic models for coreference resolution’. *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*. USA, Rochester, 22 - 27th April 2007. pp 81-88.

Dagan, I. , and Itai, A. 1990. ‘Automatic processing of large corpora for the resolution of anaphora references’. *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90).* Finland, Helsinki. pp 330-332.

Dagan, I., Justeson, J., Lappin, S., Leass, H., and Ribak, A. 1995. ‘Syntax and lexical statistics in anaphora resolution’. *Applied Artificial Intelligence* 9:633-644.

Dale, R. 1992. *Generating Referring Expressions*. Cambridge, MA: MIT Press.

Dalrymple, M., Shieber, S., and Pereira, F. 1991. ‘Ellipsis and higher-order unification’. *Linguistics and Philosophy* 14:399-452.

Darwish, K. , and Oard, D. 2003. ‘CLIR experiments at Maryland for TREC-2002: evidence combination for Arabic-English retrieval’. *Proceedings of the 11th Text Retrieval Conference (TREC 2002).* Available online at: http://trec.nist.gov/pubs/trec11/paper/umd.darwish.pdf [Accessed: 9 November 2013].

Daumé III, H. , and Marcu, D. 2005. ‘A large-scale exploration of effective global e-features for a joint entity detection and tracking model’. *Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical*

Methods in Natural Language Processing. Canada,Vancouver, B.C., 6 - 8th October 2005. pp 97-104.

Delphi. Available online at: www.delphi.com [Accessed: 9 November 2013].

Denis, P, and Baldridge, J. 2007a. ‘A ranking approach to pronoun resolution’. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. India, Hyderabad, 6 - 12th January 2007. pp 1588-1593.

Denis, P, and Baldridge, J. 2007b. ‘Joint determination of anaphoricity and coreference resolution using integer programming’. *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*. USA, Rochester, 22 - 27th April 2007. pp 236-243.

Deoskar, T. 2004. Techniques for Anaphora Resolution: A Survey.Cornell University. Available online at: http://www.cs.cornell.edu/courses/cs674/2005spprojects/tejaswini-deoskar.doc [Accessed: 9 November 2013].

Diab, M., Hacioglu, K. , and Jurafsky, D. 2004. ‘Automatic tagging of Arabic text: from raw text to base phrase chunks’. In S. Dumas, D. Marcus and S. Roukos (eds.) *HLT-NAACL 2004: Short Papers*. Boston, USA: Association for Computational Linguistics. 140-152.

Diab, M. and Habash, N. 2007. ‘Arabic dialect tutorial’. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational*

Linguistics (NAACL'07). USA, New York, Rochester, April 2007.

Diab, M., Ghoneim, M., and Habash, N. 2007. 'Arabic diacritization in the context of statistical machine translation'. *Proceedings of the MT Summit X1*. Denmark, Copenhagen, 10-14th September 2007. pp 143-149.

Dickins, J., Hervey, S., and Higgins, I. 2002. *Thinking Arabic Translation: A Course in Translation Method: Arabic to English*. London: Routledge.

Dorothy, K. 2001. *Lexis and Creativity in Translation: A Corpus Based Study*. Manchester: St Jerome Publishing.

Dowty, D. 1986. 'The effects of aspectual class on the temporal structure of discourse: Semantics or pragmatics?'. *Linguistics and Philosophy* 9:37-61.

Dwivedi, V., Phillips, N., Laguë-Beauvais, M., and Baum, S. 2006. 'An electrophysiological study of mood, modal context, and anaphora'. *Brain Research 11* 17:135-153.

Eckert, M., and Strube, M. 2001. 'Dialogue acts, synchronising units and anaphora resolution'. *Journal of Semantics* 17:51-89.

Ehrlich, K., and Rayner, K. 1983. 'Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing'. *Journal of Verbal Learning and Verbal Behavior* 22:75-87.

Eid, M., and Holes, C. (eds.) 1993. *Perspectives on Arabic linguistics V*. (Current Issues in Linguistic Theory, 101) Philadelphia: John Benjamins.

Elghamry, K, El-Zeiny, N., and Al-Sabbagh, R. 2007. 'Arabic anaphora resolution using the web as corpus'. In *Proceedings of the 7th Conference of Language Engineering.* Cairo, Egypt December 2007.

Evans, R. 2001. 'Applying machine learning toward an automatic classification of it'. *Literary and Linguistic Computing* 16:45-57.

Farghaly, A. 1981. *Topics in the Syntax of Egyptian Arabic*, published PhD thesis, University of Texas.

Farghaly, A. 1982. 'Subject pronoun deletion rule'. *In Proceedings of the 2nd English Language Symposium on Discourse Analysis (LSDA'82)*. pp 110-117.

Farghaly, A. 1987. 'Three level morphology for Arabic'. *In Proceedings of the Arabic Morphology Workshop (AMW'87)*.

Farghaly, A. 2005. 'A case for inter-Arabic Grammar'. In Eligbali, A., (ed.) *Investigating Arabic: Current Parameters in Analysis and Learning*. Boston: Brill.

Farghaly, A. 2008. 'Arabic NLP: overview, the state of the art: challenges and opportunities'. *In Proceedings of the International Arab Conference on Information Technology (ACIT'08*). Tunisia, Hammamet, 16 - 18th 2008.

Farghaly, A. 2010. *Arabic Computational Linguistics*. Stanford, CA: CSLI Publications.

Farghaly, A. and Senellart, J. 2003. Intuitive coding of the Arabic lexicon. *In Proceedings of the MT Summit IX, the Association for Machine Translation in the Americas (AMTA'03)*. USA, Louisiana, New Orleans, 23 - 27th September 2003.

Farghaly, A., and Shaalan, K. (2009). ‘Arabic natural language processing: challenges and solutions’. *ACM Transactions on Asian Language Information Processing (TALIP)* 8 (4):1-22.

Fassi Fehri, A. 1993. *Issues in the structure of Arabic clauses and words*. Dordrecht ; London: Kluwer Academic.

Fellbaum, C. (ed.) 1998. *WordNet: An Electronic Lexical Database*. Cambridge, Mass: MIT Press.

Ferguson, C. (1959). ‘Diglossia’. In R. Schjerve (eds.) *Diglossia and Power Language Policies and Practice in the 19th Century Habsburg Empire*. Berlin: Walter de Gruyter GmbH and Co. KG.

Ferrández. A, Palomar, M. , and Moreno, L. 1998. ‘A computational approach to pronominal anaphora, one-anaphora and surface count anaphora’. *Proceedings of the 2nd Colloquium on Discourse Anaphora and Anaphora Resolution (DAARC’98).* UK, Lancaster. pp 117-128.

Filippova, K. 2005. *A memory-based learning approach to pronominal anaphora resolution in German newspaper texts*, published MA thesis, Universitat Tubingen.

Fisher, D., Soderland, S., McCarthy, J., Feng, F., and Lehnert, W. 1996. ‘Description of the UMass system as used for MUC-6’. *Proceedings of the 6th Message Understanding Conference (MUC-6).* USA, Maryland, Columbia, 6 - 8th November 1995. pp 127-140.

Fleischman, M., Hovy, E., and Echihabi, A. 2003. ‘Offline strategies for online question answering: Answering questions before they are asked’. *Proceedings of the 41st Annual Meeting of the*

Association for Computational Linguistics. Japan, Sapporo, 7-12th July 2003. pp 1-7.

Fraurud, K. 1990. ‘Definiteness and the processing of NPs in natural discourse’. *Journal of Semantics* 7:395-433.

Freeman, A. 2001. ‘Brill's POS Tagger and a Morphology Parser for Arabic’. *Proceedings of the 39th Annual Meeting of Association for Computational Linguistics and 10th Conference of the European Chapter, Workshop on Arabic Language Processing: Status and Prospects*. France, Toulouse.

Fujii, A. , and Ishikawa, T. 2000. ‘Utilizing the World Wide Web as an encyclopedia: extracting term descriptions from semi-structured text’. *Proceedings of 38th Meeting of the Association for Computational Linguistics*.Hong Kong. pp 488- 495.

Gaber, M. 1980. *Al-Damayer fy Allugha AlArabia (Pronouns in the Arabic Language)*. Cairo: Daar Al-Maeref.

Gabrilovich, E., and Markovitch, S. 2006. ‘Overcoming the brittleness bottle-neck using Wikipedia: enhancing text categorization with encyclopedic knowledge’. *Proceedings of the 21st National Conference on Artificial Intelligence*. USA, Boston, Massachusetts, 16 - 20th July 2006. pp 1301-1306.

Gabrilovich, E., and Markovitch, S. 2007. ‘Computing semantic relatedness using Wikipedia-based explicit semantic analysis’. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. India, Hyderabad, 6 - 12th January 2007. pp 1606-1611.

Gardent, C. , and Konrad, K. 2000. ‘Interpreting definites using model generation’. *Journal of Language and Computation* 1:193-209.

Gardent, C., and Manuélian, H. 2005. ‘Création d'un corpus annoté pour le traitement des descriptions définies’. *Traitement Automatique des Langues* 46:I.

Garera, N, and Yarowsky, D. 2006. ‘Resolving and generating definite anaphora by modeling hypernymy using unlabeled corpora’. *Proceedings of the 10th Conference on Computational Natural Language Learning*. USA, New York, 8 - 9th June 2000. pp 37-44.

Garnham, A. 1982. *On-line Construction of Representations of the Content of Texts*. USA:Indiana University Linguistics Club.

Garnham, A., Oakhill, J. V., Ehrlich, M. F., and Carreiras., M. 1995. ‘Representation and process in the interpretation of pronouns’. *Journal of Memory and Language* 34:41-62.

Garnham, A. 2001. *Mental Models and the Interpretation of Anaphora*. Hove: Psychology Press.

Garrod, S. C. 1993. ‘Resolving pronouns and other anaphoric devices: the case for diversity in discourse processing’. In L. Frazier C. Clifton and K. Rayner (eds.) *Perspectives in Sentence Processing*. UK, Hove: Lawrence Erlbaum.

Garvey, C., and Caramazza, A. 1974. ‘Implicit causality in verbs’. *Linguistic Inquiry* 5:459-464.

Gasperin, C, Gamallo, P., Agustini, A., Lopes, G., and Lima, V. 2001. ‘Using syntactic contexts for measuring word similarity’. *Proceedings of the ESSLLI 2001 Workshop on Semantic Knowledge Acquisition and Categorisation*. Finland, Helsinki.

Gasperin, C., and Vieira, R. 2004. ‘Using word similarity lists for resolving indirect anaphora’. *ACL'04 Workshop on Reference*

Resolution and its Applications. Spain, Barcelona, 25 - 26th July 2004.

Gasperin, C. 2009. ‘Statistical anaphora resolution in biomedical texts’. *Technical Report*. Cambridge: Cambridge University.

Ge, N., Hale, J., and Charniak, E. 1998. ‘A statistical approach to anaphora resolution’. *Proceedings of the 6th Workshop on Very-large Corpora*. Canada, Montréal. pp 161-170.

Gentzler, E. 1993. *Contemporary Translation Theories*. London: Routledge.

Gernsbacher, M., and Hargreaves, D. 1988. ‘Accessing sentence participants: the advantage of first mention’. *Journal of Memory and Language* 27:699-717.

Geurts, B. 1997. ‘Good news about the description theory of names’. *Journal of Semantics 14(4):*. 14:319-348.

Giles, J. 2005. ‘Internet encyclopedias go head to head’. *Nature* 438:900-901.

Giuseppe, P., Massimo, R., and Domenico, T. 2008. ‘Advanced semantic search and retrieval in a collaborative peer-to-peer system’. In *Proceedings of the 3rd International Workshop on Use of P2P, Grid and Agents for the Development of Content Networks*. USA, Boston, MA.

Givon, T. (ed.) 1983. *Topic Continuity in Discourse : A Quantitative Cross-Language Study*. Amsterdam and Philadelphia: J. Benjamins.

Givon, T. 1992. ‘The grammar of referential coherence as mental processing instructions’. *Linguistics* 30:5-56.

Glaser, A. 2011. Feature design and evaluation for a coreference resolution system for English, published MA thesis, Universität Stuttgart. Available online at: www.ims.uni-stuttgart.de/institut/mitarbeiter/glaseraa/.../diplomarbeit-glaser.pdf [Accessed: 9 November 2013].

Google. Available online at: http://translate.google.com/# [Accessed: 9 November 2013].

Gordon, P., Grosz, B., and Gillion, L. 1993. ‘Pronouns, names, and the centering of attention in discourse’. *Cognitive Science* 17:311-348.

Gordon, P., and Searce, K. 1995. ‘Pronominalization and discourse coherence, discourse structure and pronoun interpretation’. *Memory and Cognition* 23:313-323.

Gordon, P., and Hendrick, R. 1997. ‘Intuitive knowledge of linguistic coreference’. *Cognition* 62:325-370.

Gordon, P., Hendrick, R., Ledoux, K., and Yang, C. 1999. ‘Processing of reference and the structure of language: an analysis of complex noun phrases’. *Language and Cognitive Processes* 14:353-379.

Grefenstette, G., Semmar, N., and Elkateb, F. 2005. ‘Modifying a natural language processing system for European languages to treat Arabic information in information retrieval applications’. *Proceedings of the ACL Computational approaches to Semitic Languages*. USA, Ann Arbor, 29th June 2005. pp 31-38.

Grishman, R., and Sundheim, B. 1995. ‘Design of the MUC-6 evalutation’. *Proceedings of the 6th Message Understanding*

Conference (MUC-6). USA, Maryland, Columbia, 6 - 8th November 1995.

Groenendijk, J., and Stokhof, M. 1991. 'Dynamic predicate logic'. *Linguistics and Philosophy* 14:39-100.

Grosz, B. 1977. *The representation and use of focus in dialogue understanding*, published PhD Thesis, Stanford University.

Grosz, B., Joshi, A., and Weinstein, S. 1983. 'Providing a unified account of definite noun phrases in discourse'. *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*. USA, Cambridge, Massachusetts, 15 - 17th June 1983. pp 44-50.

Grosz, B., and Sidner, C. 1986. 'Attention, intention, and the structure of discourse'. *Computational Linguistics* 12:175-204.

Grosz, B., Joshi, A., and Weinstein, S. 1995. 'Centering: a framework for modeling the local coherence of discourse'. *Computational Linguistics* 21:202-225.

Gundel, J. 1974. *The role of topic and comment in linguistic theory*, published PhD thesis, University of Texas at Austin. Reprinted by Garland Publishing, New York and London, 1988.

Gundel, J. 1988. 'Centering theory and the givenness hierarchy: towards a synthesis'. In M. A. Walker, A. K. Joshi and E. F. Prince (eds) *Centering Theory in Discourse*. Oxford University Press. 183-198.

Gundel, J., Hedberg, N., and Zacharski, R. 1993. 'Cognitive status and the form of referring expressions in discourse'. *Language* 69:274-307.

Gundel, J., Hedberg, N., and Zacharski, R. 2002. ‘Pronouns without explicit antecedents: how do we know when a pronoun is referential?’. *Proceedings of the 2nd Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC’02).* UK, Lancaster.

Habash, N. 2004. ‘Large scale lexeme based Arabic morphological generation’. Paper presented at *Proceedings of Session Traitement Automatique de l'Arabe (JEP-TALN 2004).* Morocco, Fez.

Habash, N. 2007. ‘Syntactic preprocessing for statistical machine translation’. *Proceedings of the Machine Translation Summit (MT-Summit).* Denmark, Copenhagen.

Habash, N. 2010. *Introduction to Arabic Natural Language Processing*. USA: Morgan and Claypool.

Habash, N. , and Rambow, O. 2005. ‘Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop’. Paper presented at *proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics.* USA. Michigan.

Habash, N., Owen R. and Ryan R. MADA+TOKAN Manual. CCLS technical report CCLS-10-01. 2010. Available online at: http://www1.ccls.columbia.edu/~cadim/MADA [Accessed: 9 November 2012].

Haghighi, A., and Klein, D. 2007. ‘Unsupervised coreference resolution in a non-parametric bayesian model’. *Proceedings of the 45th Annual Meeting of the Association for Computational*

Linguistics. Czech Republic, Prague, 23 - 30th June 2007. pp 848-855.

Haghighi, A., and Klein, D. 2010. ‘Coreference resolution in a modular, entity-centered model’. *Proceedings of Human Language Technologies 2010: The Conference of the North American Chapter of the Association for Computational Linguistics*. USA, California, Los Angeles, 1 - 6th June 2010. pp 385-393.

Halliday, M. , and Hasan, R. 1976. *Cohesion in English*. London: Longman.

Halliday, M., and Matthiessen, C. 2004. *An Introduction to Functional Grammar*. London: Arnold.

Hammami, S., Belguith, L., and Hamadou, A. 2009. ‘Arabic anaphora resolution: corpora annotation with coreferential links’. *The International Arab Journal of Information Technology*. 6(5): 480-488.

Han, J., Kamber, M., and Pei, J. 2006. *Data Mining : Concepts and Techniques*. Waltham, MA: Morgan Kaufmann.

Harabagiu, S. 1998. ‘WordNet-based interface of textual context: cohesion and coherence’. *Proceedings of the 11th International FLAIRS Conference.* Available online at: http://aaaipress.org/Papers/FLAIRS/1998/FLAIRS98051.pdf. [Accessed: 9 November 2013].

Harabagiu, S., and Moldovan, D. 1998. ‘Knowledge processing on extended WordNet’. In C. Fellbaum (ed.) *WordNet: An Electronic Lexical Database*. MIT Press. 379-405.

Harabagiu, S., Razvan C., and Steven J. 2001. ‘Text and knowledge mining for coreference resolution’. *Proceedings of the 2nd*

Conference of the North American Chapter of the Association for Computational Linguistics. USA, Pittsburgh, 2 - 7th June 2001. pp 55-62.

Hardt, D. 1997. 'An empirical approach to VP ellipsis'. *Computational Linguistics* 23:525-541.

Hartrumpf, S. 2001. 'Coreference resolution with syntactico-semantic rules and corpus statistics'. *Proceedings of the 3rd Conference on Computational Natural Language Learning*. France, Toulouse, 6 - 7th July 2001. pp 137-144.

Hartrumpf, S., Helbig, H., and Osswald, R.. 2003. 'The semantically based corpus HaGenLex - structure and technological environment'. *Traitement automatique des langues* 44:81-105.

Hary, B. 1992. *Multiglossia in Judeo-Arabic : With an Edition, Translation and Grammatical Study of the Cairene Purim scroll*. Leiden ; New York: E.J. Brill.

Hasan, A. 1999. *AlnaHw AlwAfy (The Comprehensive Grammar)* vol. 4. Cairo: Daar Al-Maeref.

Hasler, L., Orasan, C., and Naumann, K. 2006. 'NPs for events: experiments in coreference annotation'. *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Italy, Genoa, 22 - 28th May 2006.

Hauenschild, C., and Heizmann, S. (eds.) 1997. *Machine Translation and Translation Theory*. Berlin: Mouton de Gruyter.

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

Hearst, M. 1992. 'Automatic acquisition of hyponyms from large text corpora'. *Proceedings of the 15th International Conference on*

Computational Linguistics.France, Nantes, 23 - 28th August. pp 539-545.

Heeman, P., and Allen, J. 1995. The TRAINS-93 dialogues. TRAINS Technical Note TN 94-2.

Heim, I. 1982. The semantics of definite and indefinite noun phrases, published Ph.D. thesis, University of Massachusetts at Amherst.

Heim, I. 1983. 'File change semantics and the familiarity theory of definiteness'. In R. Bauerle, C. Schwarze and A. von Stechow (eds.) *Meaning, Use and Interpretation of Language*. Berlin: de Gruyter.

Helbig, H., and Hartrumpf, S. 1997. 'Word class functions for syntactic-semantic analysis'. *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing*. Bulgaria, Tzigov Chark.

Hendrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A., Vloet, J., and Verschelde, J. 2008. 'A coreference corpus and resolution system for Dutch'. *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Morocco, Marrakech, 26 May - 1st June 2008.

Hinrichs, E., Filippova, K., and Wunsch, H. 2005a. 'What treebanks can do for you: rule-based and machine-learning approaches to anaphora resolution in German'. *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT'05)*. Spain, Barcelona.

Hinrichs, E., Ubler, S., and Naumann, K. 2005b. 'A unified representation for morphological, syntactic, semantic and referential annotations'. *ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*. USA, Michigan, Ann Arbor.

Hirschman, L. 1998. 'MUC-7 coreference task definition, version 3.0'. *Proceedings of the 7th Message Understanding Conference*. USA,Virginia, 29 April - 1st May 1998.

Hirschman, L., Robinson, P., Burger, J., and Vilain, M. 1997. 'Automating coreference: The role of automated training data'. *Proceedings of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*. USA, California.

Hirst, G. 1981. *Anaphora in Natural Language Understanding*. Berlin: Springer Verlag.

Hitzeman, J., and Poesio, M. 1998. 'Long-distance pronominalisation and global focus'. *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*. Canada, 10 - 14th August 1998. pp 550-556.

Hobbs, J. 1976. 'Pronoun resolution'. In *Research Report 76-1*. New York: University of New York.

Hobbs, J. 1978. 'Resolving pronoun references'. *Lingua* 44:339-352.

Hobbs, J. 1979. 'Coherence and coreference'. *Cognitive Science* 3:67-90.

Hobbs, J., and Shieber, S. 1987. 'An algorithm for generating quantifier scopings'. *Computational Linguistics* 13:47-63.

Hobbs, J., Stickel, M., Martin, P., and Edwards, D. 1993. 'Interpretation as abduction'. *Artificial Intelligence Journal* 63:69-142.

Hobbs, J., and Kehler, A. 1997. 'A theory of parallelism and the case of vp ellipsis'. *Proceedings of the 35th Annual Meeting of the*

Association for Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics. Spain, Madrid, 7 - 12th July 1997. pp 394-401.

Holes, C. 2002. *Colloquial Arabic of the Gulf and Saudi Arabia*. USA: Routledge.

Holmes, D., and Jain, L. 2012. *Data Mining : Foundations and Intelligent Paradigms*. Berlin ; London: Springer.

Hoste, V. 2005. *Optimization issues in machine learning of coreference*, published PhD Thesis, University of Antwerp.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. 2006. ‘Ontonotes: the 90% solution’. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. USA, New York, 4 - 9th June 2006.

Hudson-D’Zmura, S., and Tanenhaus, M. 1998. ‘Assigning antecedents to ambiguous pronouns: the role of the center of attentian as a default assignment’. In M. Walker, A. Joshi and E. Prince (eds.) *Centering Theory in Discourse*. Oxford: Clarendon Press. 273-291.

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H., and Wilks, Y. 1998. ‘Description of the lasie system as used for MUC-7’. *Proceedings of the 7th Message Understanding Conference.* USA,Virginia, 29 April - 1st May 1998.

Hutchins, W. 1986. *Machine Translation Past, Present and Future.* Chichester: Ellis Horwood.

Hutchins, W. 1987. 'Prospects in machine translation'. In *Proceedings of MT Summit manuscripts and program*. Japan, 17 - 19th September 1987. pp 48-52.

Hutchins, W., and Somers, H. 1992. *An Introduction to Machine Translation*. London: Academic Press.

Hutchins, W. 1997. 'From first conception to first demonstration: the nascent years of machine translation 1947-1954: a chronology'. *Machine Translation* 12:195-252.

Hutchins, J. (ed.) 2000. *Early Years in Machine Translation : Memoirs and Biographies of Pioneers*. Amsterdam; Philadelphia: J. Benjamins.

Hutchins, J. 2005. 'Towards a definition of example-based machine translation'. *Proceedings of the 2nd Workshop on Example-Based Machine Translation*. Thailand, Phuket.

Iida, R., Kentaro I., and Yuji M. 2007a. 'Zero-anaphora resolution by learning rich syntactic pattern features'. *ACM Transactions on Asian Language Information Processing (TALIP)* 6(4):1-22.

Iida, R., Inui, K., and Matsumoto, Y. 2003b. 'Zero-anaphora resolution by learning rich syntactic pattern features'. *ACM Transactions on Asian Language In formation Processing (TALIP)* 6:1-22.

Iida, R., Inui, K., Takamura, H., and Matsumoto, Y. 2003a. 'Incorporating contextual cues in trainable models for coreference resolution'. *Proceedings of EACL Workshop on the Computational Treatment of Anaphora*. Hungary, Budapest, April 2003.

Iida, R., Komachi, M., Inui, K., and Matsumoto, Y. 2007b. 'Annotating a Japanese text corpus with predicate-argument and coreference relations'. *Proceedings of the ACL-07 Linguistic Annotation Workshop*. Czech Republic, Prague, 28 - 29th June 2007.

Imad A., Ibrahim A. 2004. 'Arabic morphological analysis techniques: a comprehensive survey'. *Journal of the American Society for Information Science and Technology* 55 (3):189-213.

Jakob, N., and Gurevych, I. 2010. 'Using anaphora resolution to improve opinion target identification in movie reviews'. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Sweden, Uppsala, 11 - 16th July 2010. pp 263-268.

Ji, H., Westbrook, D., and Grishman, R. 2005. 'Using semantic relations to refine coreference decisions'. *Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processing*. Canada, Vancouver, B.C., 6 - 8th October 2005. pp 17-24.

Joshi, M. , and Aaron, P. 2006. *Handbook of Orthography and Literacy*. New Jercy: Lawrence Erlboum Associates

Jurafsky, D., and Martin, J. 2000. *Speech and Language Processing; An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. New Jersey: Prentice Hall Ltd.

Jurasky, D., and Martin, J. 2009. *Speech and Language Processing*. New Jersey: Prentice Hall Ltd.

Kabadjov, M. 2007. *Task-oriented evaluation of anaphora resolution*, published PhD thesis, University of Essex.

Kameyama, M. 1985. *Zero Anaphora: The Case of Japanese*. Stanford University.

Kameyama, M. 1997. ‘Recognizing referential links: an information extraction prespective’. *ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*. Spain, Madrid, 11th July 1997. pp46-53.

Kamp, H. 1979. ‘Events, instant and temporal reference’. In R. Bauerle, U. Egli and A. von Stechow (eds.) *Semantics from Different Points of View*. Berlin: Springer-Verlag. 376-41.

Kamp, H. 1981. A theory of truth and semantic representation. In . J. Groenendijk, T. Janssen and M. Stokhof (eds.) *Formal Methods in the Study of Language*. Amsterdam: Mathematical Centre.

Kamp, H., and Reyle, U. 1993. *From Discourse to Logic*. Dordrecht: D. Reidel.

Kaplan, D. 1977. *Demonstratives. an essay on the semantics, logic, metaphysics and epistemology of demonstratives and other indexicals*, Unpublished manuscript, University of California, Los Angeles.

Karamanis, N. 2003. *Entity coherence for descriptive text structuring*, published PhD thesis, Edinburgh University.

Karamanis, N., Poesio, M., Oberlander, J., and Mellish, C. 2009. ‘Evaluating centering for information ordering using corpora’. *Computational Linguistics* 35:1-17.

Karttunen, L. 1976. ‘Discourse referents’. In J. McCawley (ed.) *Syntax and Semantics 7 - Notes from the Linguistic Underground*. New York: Academic Press. 363-385.

Kehler, A. 1997. ‘Probabilistic coreference in information extraction’. *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*. USA, Providence, 1 - 2nd August 1997. pp 163-173.

Kehler, A., Appelt, D., Taylor, L., and Simma, A. 2004. ‘The (non)utility of predicate-argument frequencies for pronoun interpretation’. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. USA, Boston, Massachusetts, 2 - 7th May 2004. pp 289-296.

Kehler, A., Kertz, L., Rohde, H., and Elman, J. 2008. ‘Coherence and coreference revisited’. *Journal of Semantics* 25:1-44.

Kelleher, J., Costello, F., and Genabith, J. 2005. ‘Dynamically updating and interrelating representations of visual and linguistic discourse’. *Artificial Intelligence* 167:62-102.

Kennedy, C., and Boguraev, B. 1996. ‘Anaphora for everyone: Pronominal anaphora resolution without a parser’. *Proceedings of the 16th International Conference on Computational Linguistics*. Denmark ,Copenhagen, 5 - 9th August 1996. pp 113-118.

Khoja, S. 2001. ‘APT: Arabic part-of-speech tagger’. In *Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics*.USA, PA.

Khoja, S., Garside, R., and Knowles, G. 2001. ‘A Tagset for the Morpho-Syntactic Tagging of Arabic’. *Proceedings of the*

Corpus Linguistics 2001 Conference. UK, Lancaster, 29 March-2nd April 2001.

Kibble, R., and Power, R. 2000. ‘An integrated framework for text planning and pronominalization’. *Proceedings of the International Conference on Natural Language Generation (INLG)*.Israel, Mitzpe Ramon.

Kilgarriff, A., and Grefenstette, G. 2006. ‘Web as corpus’. Paper presented at *Proceedings of Corpus Linguistics 2001 Conference* . UK, Lancaster, 29 March- 2nd April 2001.

Klapholz, D., and Lockman, A. 1975. ‘Contextual reference resolution’. *American Journal of Computational Linguistics*, microfiche 36.

Klavans, J., and Resnik, P. (eds.) 1996. *The Balancing Act*. London: MIT Press.

Klenner, M., and Ailloud, E. 2008. ‘Enhancing coreference clustering’. *Second Bergen Workshop on Anaphora Resolution (WAR II)*. Norway, Bergen, 29 - 31st August 2008.

Knott, A., Oberlander, J., O' Donnell, M., and Mellish, C. 2001. ‘Beyond elaboration: The interaction of relations and focus in coherent text’. In T. Sanders, J. Schilperoord and W. Spooren (eds.) *Text representation: linguistic and psycholinguistic aspects*. Amsterdam and Philadelphia: John Benjamins. 181-196.

Korhonen, A. 2002. ‘Subcategorization Acquisition’. *Technical Report* University of Cambridge.

Kouchnir, B. 2004. ‘A machine learning approach to German pronoun resolution’. *Proceedings of the ACL'04 Student Research Workshop*. Spain, Barcelona, 25 - 26th July 2004.

Kremers , Joost. 1997. *When Arabs talk to each other about themselves,* unpublished MA thesis, University of Nijmegen.

Kripke, S. 1972. ‘Naming and necessity’. In D. Davidson and G. Harman (eds.) *Semantics of Natural Language*. Dordrecht: Reidel. 253-355.

Krovetz, R. 1993. ‘Viewing morphology as an inference process’. *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. USA, Pittsburgh, 27 June - 1st July 1993. pp 191-202.

Lakoff, G., and Ross, J. 1976. ‘Why you can’t do so into the kitchen sink’. In James D. McCawley (ed.) *Syntax and Semantics*. New York: Academic Press. 101-111.

Landragin, F., Angeli, A., Wolff, F., Lopez, P., and Romary, L. 2002. ‘Relevance and perceptual constraints in multimodal referring actions’. In K. van Deemter and R. Kibble (eds.) *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*. Stanford: CSLI Lecture Notes. 395-413

Langacker, R. 1969. ‘Pronominalization and the chain of command’. In D. Reibel and S. Schane (eds.) *Modern Studies in English*. Englewood Cliffs: Prentice-Hall.

Lappin, S. 2005. ‘A sequenced model of anaphora and ellipsis resolution’. In B. Antonio, T. McEnery and R. Mitkov (eds.) *Anaphora Processing: Linguistic, Cognitive and Computational Modeling*. Amsterdam; Philadelphia: J. Benjamins. 3-17.

Larkey, L., Ballesteros, L., and Connell, M. 2002. ‘Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis’. *Proceedings of the 25th ACM SIGIR*

Conference on Research and Development in Information Retrieval. Finland, Tampere, 11 - 15th August 2002. 269-303.

Larkey, L., and Connell, M. 2002. ‘Arabic information retrieval at UMass in TREC- 10 (TREC 2001)’. *Proceedings of the 10th Text Retrieval Conference*. USA, Gaithersburg, MD. 562-570.

Larson, M. 1998. *Meaning-Based Translation: A Guide to Cross-Language Equivalence*. Lanham: University Press of America.

Lasnik, H. 1976. ‘Remarks on coreference’. *Linguistic Analysis* 2:1-22.

Lasnik, H., and Uriagereka, J. 1988. *A Course in GB Syntax : Lectures on Binding and Empty Categories*. Cambridge, Massachusetts; London: MIT.

Leacock, C., and Chodorow, M. 1998. ‘Combining local context and WordNet similarity for word sense identification’. In C. Fellbaum (ed.) *WordNet. An Electronic Lexical Database*. Cambridge: Mass: MIT Press. 265-283.

Leass, H., and Lappin, S. 1994. ‘An algorithm for pronominal anaphora resolution. *Computational Linguistics* 20:535-561.

Lehnert, W., Cardie, C., McCarthy, J., Riloff, E., and Soderland, S. 1992. ‘University of Massachusetts: Description of the CIRCUS system as used for MUC-4’. *Proceedings of the 4th Message Understanding Conference (MUC-4).* USA, San Mateo. pp 223-233

Liddy, E. 2001. Natural Language Processing. In M. Drake (ed.) *Encyclopedia of Library and Information Science*. New York: Marcel Decker, Inc.

Lin, D. 1995. ‘University of Manitoba: Description of the PIE system used for MUC-6’. *Proceedings of the 6th Message*

Understanding Conference. USA, Maryland, Columbia, 6 - 8th November 1995. pp 113-126

Lin, D. 1998. ‘Automatic retrieval and clustering of similar words’. *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*. Canada, Montréal, 10-14th August 1998. pp 768-774.

Lin, D. 1998. ‘Using collocation statistics in information extraction’. *Proceedings of the 7th Message Understanding Conference*. USA,Virginia.

Linde, C. 1979. ‘Focus of attention and the choice of pronouns in discourse’. In T. Givon (ed.) *Syntax and Semantics Vol. 12: Discourse and Syntax.* New York/San Francisco/London: Academic Press. 337-354.

Lo, K., and Lam, W. 2006. ‘Using semantic relations with world knowledge for question answering’. *Proceedings of the 15th Text Retrieval Conference*. Gaithersburg, MD., 14 - 17th November 2006.

Lockman, A., and Klappholz, A. 1980. ‘Toward a procedural model of contextual reference resolution’. *Discourse Processes* 3:25-71.

Loebner, S. 1987. ‘Definites’. *Journal of Semantics* 4:279-326.

Lu, X. 2005. ‘Hybrid methods for POS guessing of Chinese unknown words’. *Proceedings of the Student Research Workshop at the 43rd Annual Meeting of the Association of Computational Linguistics*. USA, Michigan, Ann Arbor, 27 - 27th June 200. pp 1-6.

Lund, K., Atchley, R., and Burgess, C. 1995. ‘Semantic and associative priming in high-dimensional semantic space’. *Proceedings of the 17th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ, 22 - 25th July 1995. pp 660-665.

Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., and Roukos, S. 2004. ‘A mention-synchronous coreference resolution algorithm based on the Bell Tree’. *Proceedings of the 42nd Annual Meeting of the Association for Computa- tional Linguistics*. Spain, Barcelona, 21 - 26th July 2004. pp136-143.

Luo, X. 2005. ‘On coreference resolution performance metrics’. *Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processin*. Canada, Vancouver, B.C., 6 - 8th October. pp 25-32.

LuperFoy, S. 1992. ‘The representation of multimodal user interface dialogues using discourse pegs’. *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*. USA, Newark, Delaware , 28 June - 2nd July 1992. pp 22-31.

Lust, B. (ed.) 1986. *Studies in the Acquisition of Anaphora*. Dordrecht ; Lancaster: Reidel.

Müller, M. 2008. *Fully automatic resolution of it, this and that in unrestricted multi-party dialog*, published PhD thesis, University Tübingen.

MacDonald, M., Pearlmutter, N., and Seidenberg, M. 1994. ‘Lexical nature of syntactic ambiguity resolution’. *Psychological Review* 101:676-703.

MADA. Available online at: http://www1.ccls.columbia.edu/MADA/ [Accessed: 9 November 2013].

Maghalseh, M. 1991. *Al-nahu al-shafi*. Amman: Dar Al-Basheer.

Maghalseh, M. 2007. *Al-nahu al-shafi al-shamel*. Amman: Dar Al-Masira.

Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Lenzi, V. Bartalesi, and Sprugnoli, R. 2006. 'I-CAB: the italian content annotation bank'. *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Italy, Genoa, 22 - 28th May 2006.

Maguitman, A., Menczer, F., Roinestad, H., and Vespignani, A. 2005. 'Algorithmic Detection of Semantic Similarity'. Paper presented at the 14th *International World Wide Web Conference Committee (WWW' 05)*. Japan, Chiba. pp 107-116.

Malmkjaer, K. 1999. *Contrastive Linguistics and Translation Studies: Interface and Differences*. Utrecht: Platform Vertalen and Vertaalwetenschap.

Manaris, B. (ed.) 1998. 'Natural language processing: a human-computer interaction perspective'. *Advances in Computers*. 47:1-66.

Mani, N. 2004. 'The role of prosody in parsing ambiguous sentences'. In B. Bel and I. Marlien (eds.) *Proceedings of Speech Prosody*. Japan, Nara, 23 - 26th March 2004.

Manning, C., and Schütze, H. 2002. *Foundations of Statistical Natural Language Processing*. London: The MIT Press.

Manning, C., Raghavan, P., and Schütze, H. 2008. *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Marcus, M., Santorini, B., and Marcinkiewicz, M. 1993. ‘Building a large annotated corpus of English: the Penn treebank’. *Computational Linguistics* 19:313-330.

Markert, K., and Nissim, M. 2005. ‘Comparing knowledge sources for nominal anaphora resolution’. *Computational Linguistics* 31:367-402.

Martin, H., Claes, W., and Thomas, T. 2005. ‘Experimental context classification: incentives and experience of subjects’. In *Proceedings of the 27th International Conference on Software Engineering*. USA, St. Louis, 15 - 21st May 2005.

Mashharawi, H. 2012. *Reflection phenomenon of conscience in Arabic: descriptive study,* published MA thesis, Islamic University in Palestine.

Matthews, A., and Chodorow, M. 1988. ‘Pronoun resolution in two-clause sentences: effects of ambiguity, antecedent location, and depth of embedding’. *Journal of Memory and Language* 27:245-260.

May, R. 1985. *Logical Form in Natural Language*. Cambridge: The MIT Press.

Mayr, E. 1982. *Systematics and the Origin of Species*. New York ; Guildford: Columbia University Press.

McCarthy, J. 1981. ‘A prosodic theory of nonconcatenative morphology’. *Linguistic Inquiry* 12:373-418.

McCarthy, J., and Lehnert, W. 1995. ‘Using decision trees for coreference resolution’. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Canada, Montréal 20 - 25th August. pp 1050-1055.

McCarthy, J. 1996. *A trainable approach to coreference resolution for information extraction*, published PhD thesis, University of Massachusetts.

McEnery, T., and Wilson, A. 2001. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, A., and Xiao, R. (eds.) 2005. 'Parallel and comparable corpora: what are they up to?*'* In *Incorporating Corpora: Translation and the Linguist*. UK: Clevedon.

Mihalcea, R. 2007. 'Using Wikipedia for automatic word sense disambiguation'. *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*. USA, Rochester, 22 - 27th April 2007. pp 196-203.

Miller, G., and Hristea, F. 2006. 'WordNet nouns: classes and instances'. *Computational Linguistics* 32:1-3.

Milne, D., and Witten, I. 2008. 'An effective, low-cost measure of semantic relatedness obtained from Wikipedia links'. *An Evolving Synergy at AAAI-08 , Proceedings of the Workshop on Wikipedia and Artificial Intelligence*. USA, Chicago, 13th July. pp 25-30.

Milne, D., and Witten, I. 2008. 'Learning to link with Wikipedia'. *Proceedings of the ACM 17th Conference on Information and Knowledge Management*. USA, California, Napa Valley, 26 - 30th October. pp 1046-1055.

Miltsakaki, E. 2002. 'Towards an aposynthesis of topic continuity and intrasentential anaphora'. *Computational Linguistics* 28:319-355.

Minsky, M. 1975. ‘A framework for representing knowledge’. In P. H. Winston (ed.) *The Psychology of Computer Vision*. New York: McGraw-Hill. 211-277.

Mitkov, R. 1994a. ‘An integrated model for anaphora resolution. *Proceedings of the15th International Conference on Computational Linguistics (COLING'94)*. Japan, Kyoto, 5 - 9th August 1994. pp 1170-1176.

Mitkov, R. 1994b. ‘A new approach for tracking center’. *Proceedings of the International Conference New Methods in Language Processing (NeMLaP-1*). UK, Manchester, 14 - 16th September 1994.

Mitkov, R. 1995. ‘An uncertainty reasoning approach for anaphora resolution’. *Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'95)*. Korea, Seoul, December 1995. pp 149- 154.

Mitkov, R. 1996. ‘Anaphora and machine translation’. *Machine Translation Review* 4:6-16.

Mitkov, R. 1997. ‘Factors in anaphora resolution: they are not the only things that matter: a case study based on two different approaches’. Paper presented at *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution.* Spain, Madrid, 11th July 1997.

Mitkov, R. 1998. ‘Robust pronoun resolution with limited knowledge’. Paper presented at *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98*

Conference. Canada, Montréal, 10 - 14th August 1998. pp 869-875.

Mitkov, R. 1999. '*Anaphora Resolution: The State of the Art'. Technical Report based on COLING'98 and ACL'98 Tutorial on Anaphora Resolution*: University of Wolverhampton.

Mitkov, R. 2000. 'Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems'. *Proceedings of the 3rd Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC'03).* UCREL Technical Papers Volume 12 Special Issue.

Mitkov, R. 2002. 'Outstanding Issues in Anaphora Resolution'. *Proceedings of the 2nd International Conference on Computational Linguistics and Intelligent Text Processing*. USA, Mexico City. pp 110-125.

Mitkov, R. 2002. *Anaphora Resolution.*UK: Longman.

Mohammad, M. 2000. *Word Order, Agreement, and Pronominalization in Standard and Palestinian Arabic*. Amsterdam ; London: J. Benjamins Publishing Company.

Morton, T. 2000. 'Coreference for NLP applications'. *Proceedings of the 38th ACL*. Hong Kong, 1 October - 8th October 2000.
pp 173-180.

Mucherino, A., Papajorgji, P., and Pardalos, P. 2009. *Data Mining in Agriculture*. Dordrecht ; New York: Springer Verlag.

Munday, J. 2001. *Introducing Translation Studies: Theories and Applications*. London: Routledge.

Muskens, R. 1996. 'Combining montague semantics and discourse representation'. *Linguistics and Philosophy* 19:143-186.

Mustafawi, E., and Mahfoudhi, A. 2002. 'The development of binding principles: new findings'. *Cahiers Linguistics D'Ottawa* 30:91-111.

Nabhan, A. 2005. *Arabic to English morphology based statistical machine translation system*, unpublished MA thesis, Cairo University.

Nahla, M.1990. *AlDmAAr AlmnEksp fy Allgp AlErbyp* (*Reflexive Pronouns in Arabic)*. Beirut: Dar Al Alum Al Arabia

Nastase, V. 2008. 'Topic-driven multi-document summarization with encyclopedic knowledge and activation spreading'. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. USA, Hawaii, 25 - 27th October 2008.
pp 763-772.

Nelken, R. , and Shieber, S. 2005. 'Arabic diacritization using Weighted finite-state transducers'. *Proceedings of the 2005 Association for Computational Linguistics Workshop on Computational Approaches to Semitic Languages*. USA, Michigan. pp 79-86.

Newton, John (ed.) 1992. *Computers in Translation*. London: Routledge.

Ng, V. 2005. Machine learning for coreference resolution: From local classification to global ranking. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. USA, Ann Arbor, Michigan, 25 - 30th June. pp 157-164.

Ng, V. 2007. 'Shallow semantics for coreference resolution'. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. India, Hyderabad, 6 - 12th January.
pp 1689-1694.

Ng, V. 2008. ‘Unsupervised models for coreference resolution’. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. USA, Honolulu, October 2008. pp 640-649.

Ng, V., and Cardie, C. 2002a. ‘Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution’. *Proceedings of the 19th International Conference on Computational Linguistics.* Taiwan, Taipei, 24 August - 1st September 2002. pp 730-736

Ng, V., and Cardie, C. 2002b. ‘Improving machine learning approaches to coreference resolution’. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.* USA, Philadelphia, Pennsylvania, 7 - 12th July 2002. pp 104-111.

Nicol , J., and Swinney, D. 1986. ‘The psycholinguistics of anaphora’. In Barbara Lust (ed.) *Studies in the Acquisition of Anaphora*. Dordrecht ; Lancaster: Reidel. 72-105.

Nicol, J., and Swinney, D. 1989. ‘The role of structure in coreference assignment during sentence comprehension’. *Journal of Psycholinguistic Research, Special Issue on Sentence Processing* 18:5-19.

Nida, E. 1964. *Toward a Science of Translating, With Special Reference to Principles and Procedures Involved in Bible Translating*. Brill: Leiden.

Nida, E. 2001. *Contexts in Translating*. Amsterdam ; Philadelphia: John Benjamins Pub.

Nirenburg, S (ed.) 1987. *Machine Translation: Theoretical and Methodological Issues*. Cambridge: Cambridge University Press.

NIST. 2002. The ACE 2002 evaluation plan. Available online at: ftp://jaguar.ncsl.nist.gov/ace/doc/ACE-EvalPlan-2002-v06.pdf. [Accessed: 30 August 2013].

Nunan, D. 1993. *Introducing Discourse Analysis*. London: Penguin Books Ltd.

Nwesri, A., Tahaghoghi, S., and Scholer, F. 2007. 'Capturing out-of-vocabulary words in Arabic text'. *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*. Czech Republic, Prague, June 2007. pp 285-266

Odisho, E. 2005. *Techniques of teaching comparative pronunciation in Arabic and English*. New Jersy: Gorgias Press LLC.

Olohan, M. 2003. *Introducing Corpora in Translation Studies*. London: Routledge.

Olteanu, M., and Moldovan, D. 2005. 'PP-attachment disambiguation using large context'. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Canada, Vancouver, October 2005. pp 273-280.

Onaizan, Y., and Knight, K. 2002. 'Translating Named Entities Using Monolingual and Bilingual Resources'. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*. USA, Philadelphia, 7 - 12th July 2012. pp 400-408.

Orasan, C., Evans, R. and Mitkov, R. (2000) 'Enhancing preference-based anaphora resolution with genetic algorithms'. *Proceedings of 2nd International Conference on Natural Language Processing*. Greece, Patras, 2 - 4th June 2000. pp 185-195.

Orasan, C., and Evans, R. 2007. 'NP animacy identification for anaphora resolution'. *Journal of Artificial Intelligence Research* 29:79-103.

Ozgur, Y. 2006. *Empirical selection of nlp-driven document representations for text categorization*, published PhD thesis, Syracuse University.

Parkinson, D., and Farwaneh**,** S. (eds.) 2003. *Papers from the 15th Annual Arabic Linguistics Symposium*. USA, University of Utah, 2-3rd March 2003.

Partee, B. 1972. 'Opacity, coreference, and pronouns'. In D. Davidson and G. Harman (eds.) *Semantics for Natural Language*. Dordrecht, Holland: D. Reidel. 415-441

Partee, B. 1973. 'Some structural analogies between tenses and pronouns in English'. *Journal of Philosophy* 70:601-609.

Partee, B. 1995. 'Quantificational structures and compositionality'. In E. Bach, E. Jelinek, A. Kratzer and B. H. Partee (eds.) *Quantification in Natural Languages.* Kluwer: Springer.

Partner, P. 1988. *Arab Voices : The BBC Arabic Service, 1938-1988*. London: British Broadcasting Corporation.

Pascal. Available online at: www.Pascal.com [Accessed: 9 November 2013].

Passonneau, R. 1993. 'Getting and keeping the center of attention. In M. Bates and R. M. Weischedel (eds.) *Challenges in Natural*

Language Processing. UK:Cambridge University Press. 179-227.

Pecina, P., and Schlesinger, P. 2006. 'Combining association measures for collocation extraction'. *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Australia, Sydney, July 2006. pp 651-658.

Pedersen, T. 2008. 'Computational Approaches to Measuring the Similarity of Short Contexts : A Review of Applications and Methods'. University of Minnesota Supercomputing Institute Research Report UMSI 2010/118, October 2010. Available online at: http://www.bibsonomy.org/bibtex/0bbd7800d7590387c63c598542115def [Accessed: 9 November 2013].

Perner, P. 2011. 'Machine learning and data mining in pattern recognition'. *Proceedings of 7thInternational Conference, MLDM 2011*. USA, New York, NY, 30August - 3rd September 2011.

Peter, J., Clemens, H., and Helmut, P. 2004. 'Distribution results for low-weight binary representations for pairs of integers'. *Theoretical Computer Science* 319:307-331.

Poesio, M. 1993. 'A situation-theoretic formalization of definite description interpretation in plan elaboration dialogues'. In P. Aczel, D. Israel, Y. Katagiri and S. Peters (eds.) *Situation Theory and its Applications*. Stanford: CSLI. 339-374.

Poesio, M. 1994a. *Discourse Interpretation and the Scope of Operators*, published PhD thesis, University of Rochester.

Poesio, M. 1994b. ‘Weak definites’. In M. Harvey and L. Santelmann (eds.) *Proceedings of the 4th Conference on Semantics and Linguistic Theory, SALT-4*. USA: Cornell University Press. 282-299.

Poesio, M. 2000. ‘Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results’. *Proceedings of the 2nd LREC*. Greece, Athens, May 2000. pp 211-218.

Poesio, M. 2004. The MATE/GNOME scheme for anaphoric annotation revisited. *Proceedings of SIGDIAL*. USA, Boston, April 2004.

Poesio, M., and Vieira, R. 1998. ‘A corpus-based investigation of definite description use’. *Computational Linguistics* 24:183-216.

Poesio, M., and Kabadjov, M. 2004. ‘A general-purpose, off the shelf anaphoric resolver’. *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Portugal, Lisbon, 26 - 28th May 2004. pp 653-656.

Poesio, M., and Artstein, R. 2005a. ‘The reliability of anaphoric annotation, reconsidered: taking ambiguity into account’. *Proceedings of ACL Workshop on Frontiers in Corpus Annotation*.USA, Michigan, Ann Arbor, June 2005. pp 76-83.

Poesio, M., and Modjeska, N. 2005b. ‘Focus, activation, and this-noun phrases: An empirical study’. In A. Branco, R. McEnery and R. Mitkov (eds.) *Anaphora Processing*. USA: John Benjamins. 429-442.

Poesio, M., and Artstein, R. 2008. ‘Anaphoric annotation in the ARRAU corpus’. *Proceedings of the 6th International*

Conference on Language Resources and Evaluation. Morocco, Marrakech, 26 May - 1st June.

Poesio, M., Patel, A., and Eugenio, B. 2006. 'Discourse structure and anaphora in tutorial dialogues: an empirical analysis of two theories of the global focus'. *Research in Language and Computation, Special Issue on Generation and Dialogue* 4:229-257.

Poesio, M., Vieira, R., and Teufel, S. 1997. 'Resolving bridging references in unrestricted text'. *Proceedings of the ACL Works Workshop on Operational Factors in Robust Anaphora Resolution*. Spain, Madrid, 11th July 1997.

Poesio, M., Walde, S., and Brew, C. 1998. 'Lexical clustering and definite description interpretation'. *AAAI Spring Symposium on Learning for Discourse*. USA, Stanford, March 1998. pp 82-89.

Poesio, M., Ponzetto, S., and Versley, Y. 2010. 'Computational models of anaphora resolution: a survey'. Available online at: clic.cimec.unitn.it/massimo/Publications/lilt.pdf [Accessed: 30 August 2013].

Poesio, M., Ishikawa, T., Walde, S., and Vieira, R. 2002. 'Acquiring lexical knowledge for anaphora resolution'. *Proceedings of the 3rd International Conference on Language Resources and Evaluation* Las Palmas. Spain, Canary Islands, 29 - 31st May. pp 1220-1225.

Poesio, M., Mehta, R., Maroudas, A., and Hitzeman, J. 2004b. 'Learning to resolve bridging references'. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*.Spain, Barcelona, 21 - 26th July 2004. pp 143-150.

Poesio, M., Stevenson, R., Eugenio, B., and Hitzeman, J. 2004c. 'Centering: A parametric theory and its instantiations'. *Computational Linguistics* 30:309-363.

Poesio, M., Delmonte, R., Bristot, A., Chiran, L., and Tonelli, S. 2004a. 'The VENEX corpus of anaphoric information in spoken and written Italian'. Available online at http://cswww.essex.ac.uk/staff/poesio/publications/VENEX04.pdf. [Accessed: 30 August 2013].

Poesio, M., Uryupina, O., Vieira, R., Alexandrov-Kabadjov, M., and Goulart, R. 2004d. 'Discourse-new detectors for definite description resolution: A survey and a preliminary proposal'. *Proceedings of the ACL Workshop on Reference Resolution*. Spain, Barcelona, July 2004. Available online at: http://clair.eecs.umich.edu/aan/paper.php?paper_id=W04-0707 [Accessed: 30 August 2013].

Poesio, M., Alexandrov-Kabadjov, M., Vieira, R., Goulart, R., and Uryupina, O. 2005. 'Does discourse-new detection help definite description resolution?'. *Proceedings of the 6th International Workshop on Computational Semantics (IWCS-6)*. Netherlands, Tilburg, January 2005.

Pollard, C., and Sag., I. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Ponzetto, S., and Strube, M. 2006. 'Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution'. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for*

Computational Linguistics. USA, New York, 4 - 9th June 2006. pp 192-199.

Ponzetto, S., and Strube, M. 2007a. 'Deriving a large scale taxonomy from Wikipedia'. *Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence*. Canada, Vancouver, B.C., 22 - 26th July 2007. pp 1440-1445.

Ponzetto, S., and Strube, M. 2007b. 'Knowledge derived from Wikipedia for computing semantic relatedness'. *Journal of Artificial Intelligence Research* 30:181-212.

Ponzetto, S. 2010. *Knowledge Acquisition from a Collaboratively Generated Encyclopedia*, published PhD thesis, University of Stuttgart

Ponzetto, S., and Navigli, R. 2010. 'Knowledge-rich word sense disambiguation rivaling supervised system'. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.Sweden, Uppsala, 11 - 16th July 2010. pp 1522-1531.

Pradhan, S., Ramshaw, L., Weischedel, R., MacBride, J., and Micciulla, L. 2007. 'Unrestricted coreference: Identifying entities and events in ontonotes'. *Proceedings of the IEEE International Conference on Semantic Computing (ICSC)*. USA, California, Irvine, 17 - 19th September 2007.

Prince, E. 1981. 'Toward a taxonomy of given-new information'. In P. Cole (ed.) *Radical Pragmatics*. New York: Academic Press. 223-256.

Prince, E. 1992. 'The ZPG letter: subjects, definiteness, and information status'. In S. Thompson and W. Mann (eds.) *Discourse*

Description: Diverse Analyses of a Fund-Raising Text. USA: John Benjamins. 295-325.

Qiu, L., Kan, M., and Chua, T. 2004. ‘A public reference implementation of the RAP anaphora resolution algorithm’. *Proceedings of the 4th International Conference on Language Resources and Evaluation*.Portugal, Lisbon, May 2004. pp 26-28.

Quinlan, R. 1986. ‘Induction of decision trees’. *Machine Learning* 1:81-106.

Quinlan, J. 1993. *C4.5: Programs for Machine Learning*. San Mateo, California: Morgan Kaufman.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman

Rada, R., Mili, H., Bicknell, E., and Blettner, M. 1989. ‘Development and application of a metric to semantic nets’. *IEEE Transactions on Systems, Man and Cybernetics* 19:17-30.

Radford, A. 1981. *Transformational Syntax : A Student's Guide to Chomsky's Extended Standard Theory*. Cambridge: Cambridge University Press.

Radford, A. 1988. *Transformational Grammar : A First Course*. Cambridge: Cambridge University Press.

Radford, A. 1990. *Syntactic Theory and The Acquisition of English Syntax : The Nature of Early Child Grammars of English*. Oxford: Blackwell.

Radford, A. 1997a. *Syntactic Theory and The Structure of English : A Minimalist Approach*. Cambridge: Cambridge University Press.

Radford, A. 1997b. *Syntax : A Minimalist Introduction*. Cambridge: Cambridge University Press.

Radford, A. 1999. *Linguistics : An Introduction*. Cambridge: Cambridge University Press.

Radford, A. 2004a. *Minimalist Syntax : Exploring The Structure of English*. Cambridge: Cambridge University Press.

Radford, A. 2004b. *English Syntax : An Introduction*. Cambridge: Cambridge University Press.

Rahman, A., and Ng, V. 2009. 'Supervised models for coreference resolution'. *Proceedings of Empirical Methods in Natural Language Processing Conference (EMNLP)*. Singapore, 6-7th August 2009. pp 968-977.

Ramos, J. 2003. 'Using TF-IDF to Determine Word Relevance in Document Queries'. *Proceedings of the First Instructional Conference in Machine Learning*. USA, Piscataway, 3 - 8th December 2003.

Ratnaparkhi, A. 1999. 'Learning to parse natural language with maximum entropy models'. *Machine Learning* 34:151-178.

Ravichandran, D., Pantel, P., and Hovy, E. 2005. 'Randomized algorithms and NLP: using locality sensitive hash function for high speed noun clustering'. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. USA, Michigan, Ann Arbor, 25 - 30th June 2005. pp 622-629.

Recasens, M., and Martí, M. 2009. 'Ancora-co: coreferentially annotated corpora for Spanish and Catalan'. *Language Resources and Evaluation* 44 (4):315-345.

Recasens, M., Marquez, L., Sapena, E., Martí, M. Antonia, T., Mariona, Poesio, M., and Versley, Y. 2010. 'Semeval-2010 task 1: coreference resolution in multiple languages'. *Proceedings of the 5th International Workshop on Semantic Evaluation Association for Computational Linguistics*. Sweden, Uppsala, 15 - 16th 2010 . pp 1-8.

Redington, M., Crater, N., and Finch, S. 1998. 'Distributional information: A powerful cue for acquiring syntactic categories'. *Cognitive Science: A Multidisciplinary Journal* 22 (4):425-469.

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

Reinhart, T. 1976. *The syntactic domain of Anaphora*, published PhD thesis, Massachusetts Institute of Technology.

Reinhart, T. 1981. 'Pragmatics and linguistics: An analysis of sentence topics'. *Philosophica* 27 (1) 53-94.

Reinhart, T., and Reuland, E. 1993. 'Reflexivity'. *Linguistic Inquiry* 24:657-720.

Resnik, P. 1999a. 'Mining the Web for Bilingual Texts'. *Proceedings of the 37th Annual Meeting of the Association Computational Linguistics*. USA, College Park, Maryland, 20 - 26th June 1999. pp 527-534.

Resnik, P. 1999b. 'Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of ambiguity in Natural Language'. *Journal of Artificial Intelligence Research* 11:95-130.

Rihoux, B. 2003. 'Bridging the gap between the qualitative and quantitative Worlds? a retrospective and prospective view on qualitative comparative analysis'. *Field Methods* 15 (4):351-365.

Roberts, C. 1989. 'Modal subordination and pronominal anaphora in discourse'. *Linguistics and Philosophy* 12:683-721.

Roberts, C. 2003. 'Uniqueness presuppositions in English definite noun phrases'. *Linguistics and Philosophy* 26:287-350.

Rodriguez, K., Delogu, F., Versley, Y., Stemle, E., and Poesio, M. 2010. 'Anaphoric annotation of Wikipedia and blogs in the live memories corpus'. *Proceedings of the 7th International Conference on Language Resources and Evaluation.* Malta, Valletta, 19 - 21st May 2010.

Rooth, M. 1987. 'Noun phrase interpretation in Montague grammar, file change semantics, and situation semantics'. In P. Gärdenfors (ed.) *Generalized Quantifiers*. Dordrecht, The Netherlands: D. Reidel. 237-268.

Roth, D. , and Yih, W. 2004. 'A linear programming formulation for global inference in natural language tasks'. *Proceedings of CONLL 2004*. USA, Boston, 6 - 7th May 2004. pp 1-8

Runner, J., Sussman, R., and Tanenhaus, M. 2003. 'Assignment of reference to reflexives and pronouns in picture noun phrases: evidence from eye movements'. *Cognition* 81:1-13.

Ryding, K. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge: Cambridge University Press.

Sadat, F., and Habash, N. 2006. 'Combination of Arabic preprocessing schemes for statistical machine translation'. *Proceedings of the 21st International Conference on Computational Linguistics and*

44th Annual Meeting of the Association for Computational Linguistics. Sydney, July 2006. pp 1-8.

Sag, I., and Hankamer, J. 1984. 'Toward a theory of anaphoric processing'. *Linguistics and Philosophy* 7:325-345.

Sakhr. Available online at:http://www.sakhr.com/ [Accessed: 9 November 2013].

Sanchez-Graillet, O., Poesio, M., Kabadjov, M., and Tesar, R. 2006. 'What kind of problems do protein interactions raise for anaphora resolution? A preliminary analysis'. *Proceedings of SMBM*. Germany, Jena, 9 - 12th April 2006. pp 109-112.

Sandra, W., Harvey, M., and Preston, K. 1996. 'Rule-based reference resolution for unrestricted text using part-of-speech tagging and noun phrase parsing'. *Proceedings of the International Colloquium on Discourse Anaphora and Anaphora Resolution (DAARC'96).* UK, Lancaster, July 1996. pp 441-456.

Sanford, A., and Garrod, S. 1981. *Understanding Written Language*. Chichester: Wiley.

Sarkar, A., and Roeck, A. 2004. 'A framework for evaluating the suitability of non-English corpora for language engineering'. *Proceedings of Language Resources and Evaluation Conference (LREC).* Portugal, Lisbon, 26 - 28th 2004.

Sasano, R., Kawahara, D., and Kurohashi, S. 2008. 'A fully-lexicalized probabilistic model for Japanese zero anaphora resolution'. *Proceedings of the 22nd International Conference on Computational Linguistics*. UK, Manchester, 18 - 22nd August. pp 769-776.

Sauper, C., and Barzilay, R. 2009. 'Automatically generating Wikipedia articles: a structure-aware approach'. *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*. Singapore, 2 - 7th July 2009. pp 208-216.

Sawalha, M. and Atwell, E. (2008). 'Comparative evaluation of Arabic language morphological analysers and stemmers'. *Proceedings of COLING 2008 22nd International Conference on Computational Linguistics*. UK, Manchester ,18 - 22nd August 2008. pp 107 - 110.

Schütze, H., and Pedersen, J. 1995. 'Information retrieval based on word senses'. *Proceedings of the 4th Symposium on Document Analysis and Information Retrieval*. USA, Las Vegas, 24 - 25th April 1995. pp 161-175.

Schiehlen, M. 2004. 'Optimizing algorithms for pronoun resolution'. *Proceedings of the 20th International Conference on Computational Linguistics (COLING-2004).* Switzerland, Geneva, 23 - 27th August 2004. pp 390-396.

Seco, N., Veale, T., and Hayes, J. 2004. 'An intrinsic information content metric for semantic similarity in WordNet'. *Proceedings of the 16th European Conference on Artificial Intelligence.* Spain, Valencia, 22 - 27th August 2004. pp 1089-1090.

Seki, K., Fujii, A., and Ishikawa, T. 2002. 'A probabilistic method for analyzing Japanese anaphora integrating zero pronoun detection and resolution'. *Proceedings of the 19th International*

Conference on Computational Linguistics. Taiwan, Taipei, 24 August - 1st September 2002.

Shaalan K. (2005a). 'An intelligent computer-assisted language learning system for Arabic learners'. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)* 18 (1/2): 81-108.

Shaalan, K. (2005b). 'Arabic GramCheck: a grammar checker for Arabic: research articles'. *Software-Practice and Experience* 35:643-665.

Shaalan K., and Rafea, A. (2004). 'Machine translation of English noun phrases into Arabic'. *International Journal of Computer Processing of Oriental Languages*. 17 (2): 121-134.

Shaalan, K., and Abdel Monem, A. (2006). 'Arabic morphological generation from interlingua: a rule-based approach'. In Z. Shi, K. Shimohara, and D. Feng (eds.*) Intelligent Information Processing*. USA: Springer. 441-451.

Shaalan, K., and Abo Bakr, H. (2007). 'Transferring Egyptian colloquial into modern standard Arabic'. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'07)*. Bulgaria, Borovets. pp 525-529.

Shaalan, K., and Raza H. (2008). 'Arabic named entity recognition from diverse text types'. In B. Nordström, and A. Ranta, (eds.) *Proceedings of the 6th International Conference on Natural Language Processing (GoTAL'08*). Sweden, Gothenburg, 25 - 27th August 2008.

Shaalan, K., and Raza, H. (2009). ‘NERA: named entity recognition for Arabic’. *Journal of the American Society for Information Science and Technology (JASIST)* 60 (7): 1-12.

Shedeh, F., Hamdan, J., Amayreh, M., and Anani, M. 2006. *Moqadima fi Allughawiyat Al-Mucasira*. Amman, Jordan: Dar Wael li Al-Nashir.

Sheldon, A. 1974. ‘The role of parallel function in the acquisition of relative clauses in English’. *Journal of Verbal learning and Verbal behavior* 13:272-281.

Sidner, C. 1979. *Towards a computational theory of definite anaphora comprehension in English discourse*, published PhD thesis, Massachusetts Institute of Technology.

Smadja, F., McKeown, K., and Hatzivassiloglou, V. 1996. ‘Translating collocations for bilingual lexicons: a statistical approach’. *Computational Linguistics* 22:1-39.

Smyth, R. 1994. ‘Grammatical determinants of ambiguous pronoun resolution’. *Journal of Psycholinguistic Research* 23:197-229.

Solomonoff, A., Mielke, A., Schmidt, M., and Gish, H. 1998. ‘Clustering speakers by their voices’. *IEEE International Conference on Acoustics, Speech, and Signal Processing*. USA, Seattle, 12 - 15th May 1998. pp 757-760.

Somers, H. (ed.) 2003. *Computers and Translation: A Translator’s Guide*. Amesterdam: John Benjamins Publishing Company.

Soon, W., Ng, H., and Lim, D. 1999. ‘Corpus-based learning for noun phrase coreference resolution’. *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural*

Language Processing and Very Large Corpora (EMNLP/VLC-99). USA, Maryland , 21 - 22nd June 1999. pp 285-291.

Soon, W., Ng, H., and Lim, D. 2001. ‘A machine learning approach to coreference resolution of noun phrases’. *Computational Linguistics* 27:521-544.

Soudi A., Bosch A. and Neumann G. (eds.) 2007. *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. USA: Springer.

Stageberg, N. 1981. *An Introductory English Grammar*. New York ; London: Holt, Rinehart and Winston.

Stede, M. 2004. The Potsdam commentary corpus. *Proceedings of the ACL'04 Workshop on Discourse Annotation*. Spain, Barcelona, 25 - 26th July 2004.

Steinberger, J., Poesio, M., Kabadjov, M., and Jezek, K. 2007. ‘Two uses of anaphora resolution in summarization, special issue on summarization’. *Information Processing and Management* 43:1663-1680.

Stevenson, R., Crawley, R., and Kleinman, D. 1994. ‘Thematic roles, focus, and the representation of events’. *Language and Cognitive Processes* 9:519-548.

Stevenson, R., Nelson, A., and Stenning, K. 1995. ‘The role of parallelism in strategies of pronoun comprehension’. *Language and Cognitive Processes* 38:393- 418.

Stevenson, R., Knott, A., Oberlander, J., and McDonald, S. 2000. ‘Interpreting pronouns and connectives: Interactions among focusing, thematic roles and coherence relations’. *Language and Cognitive Processes* 15:225-262.

Stone, M., and Doran, C. 1996. ‘Paying heed to collocations’. *Proceedings of the 8th International Workshop on Natural Language Generation (INLG'96).*UK, Sussex, Herstmonceux Castle, 13 - 15th June 1996. pp 91-100.

Stoyanov, V., Gilbert, N., Cardie, C., and Riloff, E. 2009. ‘Conundrums in noun phrase coreference resolution: making sense of the state-of-the-art’. *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing.* Singapore, 2 - 7th July. pp 656-664.

Strube, M. 1998. ‘Never look back: an alternative to centering’. *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics* Canada, Montréal, 10 - 14th August 1998. pp 1251-1257.

Strube, M., and Hahn, U. 1999. ‘Functional centering-grounding referential coherence in information structure’. *Computational Linguistics* 25:309-344.

Strube, M., Rapp, S., and Christoph, M. 2002. ‘The influence of minimum uller edit distance on reference resolution’. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*. USA, Pennsylvania, 6 - 7th July 2002. pp 312-319.

Strube, M., and Ponzetto, S. 2006. ‘WikiRelate! computing semantic relatedness using Wikipedia’. *Proceedings of the 21st National*

Conference on Artificial Intelligence. USA, Boston, 16 - 20th July 2006. pp 1419-1424.

Stuckardt, R. 2004. ‘Three algorithms for competence-oriented anaphor resolution’. *Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2004)*. Portugal, Azores, 23 - 24th September 2004. pp 157-163.

Suri, L., and McCoy, K. 1994. ‘RAFT/RAPR and centering: a comparison and discussion of problems related to processing complex sentences’. *Computational Linguistics* 20:301-317.

Systran. Available online at: http://www.systran.co.uk/ [Accessed: 9 November 2013].

Tanaka, T. 2002. ‘Measuring the similarity between compound nouns in different languages using non-parallel corpora’. *Proceedings of the 19th International Conference on Computational Linguistics (COLING’02)*. Taiwan, Taipei, 24 August - 1st September 2002. pp 981-987.

Tapanainen, P. and Järvinen, T. (1997) ‘A non-projective dependency parser’. *Proceedings of the 5th Conference of Applied Natural Language Processing (ANLP-5)*. USA, DC, Washington, 31 March - 3rd April 1997. pp 64-71.

Tawfiq, R. 2009. *Anaphors in modern standard Arabic syntax with reference to modern syntax theories*, published MA thesis. Middle East University for Graduate Studies.

Tetreault, J. 2001. ‘A corpus-based evaluation of centering and pronoun resolution’. *Computational Linguistics* 27:507-520.

Thatcher, G. 1911. *Arabic Grammar of the Written Language*. Heidelberg: Julius Groos.

Trouilleux, F., Gaussier, E., Bies, G., and Zaenen, A. 2000. 'Coreference resolution evaluation based on descriptive specificity'. *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*. Greece, Athens, 31 May - 1st June 2000. pp 1315-1324.

Tsukanova, V., and Nikolaeva, L. 2008. 'Ways to express reflexive in Arabic. published manuscript'. Russian State University. Available online at: http://people.umass.edu/partee/RGGU.../**Tsukanova**_Nikolaeva_**arabic**_refl.doc [Accessed 9 November 2013].

Tutin A., Trouilleux F., Clouzot C., Gaussier E., Zaenen A., Rayot S., and Antoniadis G. 2000. 'Annotating a large corpus with anaphoric links'. *Proceedings of the Discourse Anaphora and Reference Resolution Conference.* UK, Lancaster, 16 - 18th November 2000. pp. 134-137.

Uryupina, O. 2003. 'High-precision identification of discourse new and unique noun phrases'. *Proceedings of the ACL Student Workshop*. Japan, Sapporo, 7 - 12th July 2003. pp 80-86.

Uryupina, O. 2006. 'Coreference resolution with and without linguistic knowledge'. *Proceedings of Language Resources and Evaluation Conference (LREC 2006)*. Italy, Genoa, 22 - 28th May 2006. pp 893-898.

Vallduvi, E. 1993. 'Information packaging: a survey'. *Research Paper RP-44* University of Edinburgh, HCRC.

van Deemter, K., and Kibble., R. 2000. 'On coreferring: coreference in MUC and related annotation schemes'. *Computational Linguistics* 26:629-637.

Van Hoek, Karen. 1997. *Anaphora and Conceptual Structure*. Chicago ; London: University of Chicago Press.

van Rijsbergen, C. and Keith, J. 1979. *Information Retrieval*. London: Butterworths.

Versley, Y. 2006. 'A constraint-based approach to noun phrase coreference resolution in German newspaper text'. *Proceedings of KONVENS 2006.* Germany, Konstanz, 4 - 7th October 2006. pp143-150.

Versley, Y. 2008. 'Vagueness and referential ambiguity in a large-scale annotated corpus'. *Research on Language and Computation* 6(3-4):333-353.

Versley, Y., Ponzetto, S., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., and Moschitti, A. 2008. 'BART: A mod- ular toolkit for coreference resolution'. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*. USA, Ohio, Columbus, 15 - 20th June 2008. pp 9-12.

Vieira, R. 1998. *Definite description resolution in unrestricted texts*, published PhD. Thesis, University of Edinburgh.

Vieira, R., and Poesio, M. 1997. 'Processing definite descriptions in corpora'. In S. Botley and M. McEnery (eds.) *Corpus-based and Computational Approaches to Discourse Anaphora*. London: UCL Press.

Vieira, R., and Poesio, M. 2000. 'An empirically based system for processing definite descriptions'. *Computational Linguistics* 26:539-593.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. 1995. 'A model-theoretic coreference scoring scheme'.

Proceedings of the 6th Message Understanding Conference (MUC-6). USA, Maryland, Columbia, 6 - 8th November 1995.

Villasenor-Pineda, L., Montesy Gomez, M., Perez-Coutino, M., and Vaufreydaz, D. 2003. 'A corpus balancing method for language model construction'. *Fourth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2003)*. Mexico, Mexico City, 16 - 22nd 2003. pp 393-401.

Volk, M. 2001. 'Exploiting the WWW as a corpus to resolve PP attachment ambiguities'. *Proceedings of Corpus Linguistics 2001*. UK, Lancaster, 29 March - 2nd April 2001. pp 601-607

von Ahn, Luis. 2006. 'Games with a purpose'. *Computer* 39:92-94.

Wagner, A., and Zeisler, B. 2004. 'A syntactically annotated corpus of Tibetan'. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*. Portugal, Lisbon, 26 - 28th May 2004. pp 1141-1144.

Walker, M. 1989. 'Evaluating discourse processing algorithms'. *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*. Canada, Vancouver, B.C., 26 - 29th June 1989. pp 251-261.

Walker, M. A., Iida, M., and Cote, S. 1994. Japanese discourse and the process of centering. *Computational Linguistics* 20:193-232.

Walker, M. 1998. 'Centering, anaphora resolution, and discourse structure'. In M. A. Walker, A. K. Joshi and E. F. Prince (eds.) *Centering in Discourse.* UK*:* Oxford University Press. 401-435.

Walker, M., Joshi, A., and Prince, E. (eds.) 1998. *Centering Theory in Discourse*. Oxford: Clarendon Press.

Walker, C., Strassel, S., Medero, J., and Maeda, K. 2006. ACE 2005 multilingual training corpus. *LDC2006T06, Philadelphia, Penn.: Linguistic Data Consortium*.

Wang, W. 2007. ‘An empirical study on hierarchical text categorization, published MA thesis, the University of Guelph.

Webber, B. 1979. *A Formal Approach to Discourse Anaphora*. New York: Garland.

Webster, J., and Ku, C. 1992. ‘Tokenization as the initial phase in NLP’. *Proceeding of International Conference of Computational Linguistics (COLIG '92)*. France, Nantes, 23 - 28th August 1992. pp 1106-1110.

Weischedel, R., Pradhan, S., Ramshaw, L., Palmer, M., Xue, N., Marcus, M., Taylor, A., Greenberg, C., Hovy, E., Belvin, R., and Houston, A. 2008. Ontonotes release 2.0. LDC2008T04, Philadelphia, Penn.: Linguistic Data Consortium. .

Werth, P. 1999. *Text Worlds: Representing Conceptual Space in Discourse*. New York: Longman

Wilks, Y. 1975. ‘An intelligent analyzer and understander of English’. *Communications of the ACM* 18:264-274.

Williams, S., Harvey, M., and Preston, K. 1996. ‘Rule-based reference resolution for unrestricted text using part-of-speech tagging and noun phrase parsing’. *Proceedings of the International Colloquium on Discourse Anaphora and Anaphora Resolution (DAARC’96)*. UK, Lancaster, July 1996. pp 441- 456.

Winograd, T. 1972. *Understanding Natural Language*. Edinburgh: Edinburgh University Press.

Woods, W., Kaplan, R., and Nash-Webber, B. 1972. *The lunar sciences natural language information system: Final report*: Report 2378, BBN, Cambridge, Mass.

Wright, W. 1964. *A Grammar of the Arabic Language.* Cambridge: Cambridge University Press.

Wu, Z., and Palmer, M. 1994. 'Verb semantics and lexical selection'. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*. USA, New Mexico, Las Cruces, 27-30th June 1994. pp 133-138.

Xu, J., Fraser, A., and Weischedel, R. 2002. 'Empirical studies in strategies for Arabic retrieval'. *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002)*. Finland, Tampere, 11-15th August 2002. pp 269-274.

Yang, X., Zhou, G., Su, J., and Tan, C. 2003. 'Coreference resolution using competition learning approach'. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Japan, Sapporo, 7 - 12th July 2003. pp 176-183.

Yang, X., Zhou, G., Su, J., and Tan, C. 2005. 'A twin-candidate model of coreference resolution with non-anaphor identification capability'. *Proceedings of the 2nd International Joint Conference on Natural Language Processing*. South Korea, Jeju Island, 11 - 13th October 2005. pp 719-730.

Yang, X., and Su, J. 2007. 'Coreference resolution using semantic relatedness information from automatically discovered patterns'. *Proceedings of the 45th Annual Meeting of the Association for*

Computational Linguistics. Czech Republic, Prague, 23 - 30th June 2007. pp 528-535.

Yang, X., Su, J., Lang, J., Tan, C., Liu, T., and Li, S. 2008. 'An entity-mention model for coreference resolution with inductive logic programming'. *Proceedings of ACL-08:HLT*. USA, Ohio, Columbus, June 2008. pp 843-851.

Yoshimi, T. 2001. 'Improvement of translation quality of English newspaper headlines by automatic pre-editing'. *Machine Translation* 16:219-232.

Appendix 1

A. Data Structures

```
const                         {Constants}

 maxfilenamelength = 24;      {maximum length of file names in file
name list}

 maxfilenamelistlength = 551;        {maximum number of input files /
documents to be processed}

 maxwordlength = 24;          {maximum word length}

 maxsentencelength = 500;     {maximum sentence length}

 maxtextlength = 10000;              {maximum document length}

 maxwordlistlength = 40000;   {maximum dictionary length}



type                          {data types}

 Tfilename = packed array [0..(maxfilenamelength - 1)] of char;
        {names of document files}

 Tfilenamelist = record
        {list of document file names}

            list : array [0..(maxfilenamelistlength - 1)] of Tfilename;
```

```
                 length : longint;

                end;



  Tword = packed array [0..(maxwordlength - 1)] of char;
            {word}

  Tentry = record
                                          {word with associated

                word : Tword;
             grammatical information}

                gender : char;

                number : char;



 Tsentence = record
            {sentence}

              s : array [0..(maxsentencelength - 1)] of Tentry;

              length : longint;

             end;

  Ttext = record
            {document}
```

```
      t : array [0..(maxsentencelength - 1)] of Tsentence;

      length : longint;

     end;


Tpointer = ^Tnode;
       {tree-structured dictionary}

Tnode = record

      entry : Tentry;

      left : Tpointer;

      right : Tpointer;

     end;


Tnafsform = class(TForm)
       {GUI type definitions}

  StaticText1: TStaticText;

  Memo1: TMemo;

  OpenDialog1: TOpenDialog;

  RadioButton5: TRadioButton;
  RadioButton6: TRadioButton;
  RadioButton7: TRadioButton;
  radiobutton1: TRadioButton;
  RadioButton4: TRadioButton;
  RadioButton3: TRadioButton;
  procedure radiobutton1Click(Sender: TObject);
  procedure RadioButton3Click(Sender: TObject);
  procedure FormCreate(Sender: TObject);
  procedure RadioButton4Click(Sender: TObject);
  procedure RadioButton5Click(Sender: TObject);
  procedure RadioButton6Click(Sender: TObject);
  procedure RadioButton7Click(Sender: TObject);
  procedure RadioButton2Click(Sender: TObject);
private
  { Private declarations }
public
  { Public declarations }
```

```
end;
```

iii. Variables

```
filenamelist : Tfilenamelist; {list of filenames of documents to be
processed}

currenttext : Ttext;          {the document currently being processed}

dictionary : Tpointer;                 {the dictionary}

newnode : Tpointer;           {the dictionary}

rootnode : Tpointer;          {the dictionary}
```

B. Program

```
unit Naf;


interface


uses

  Windows, Messages, SysUtils, Variants, Classes, Graphics, Controls, Forms,Dialogs, StdCtrls, ExtCtrls;


const

 maxfilenamelength = 24;                  {max length of file names in file name list}

 maxfilenamelistlength = 551;        {max nr of input files}

 maxwordlength = 24;

 maxsentencelength = 500;

 maxtextlength = 10000;

 maxwordlistlength = 40000;


type

 {Input file name list}

 Tfilename = packed array [0..(maxfilenamelength - 1)] of char;

 Tfilenamelist = record
```

```
                 list : array [0..(maxfilenamelistlength - 1)] of Tfilename;
                 length : longint;
                end;

Tword = packed array [0..(maxwordlength - 1)] of char;
Tentry = record
          word : Tword;
          person : longint;
          gender : char;
          number : char;
          pos : char;
          match : boolean;
         end;
Tentrylist =  record
               e : array [0..(maxwordlistlength - 1)] of Tentry;
               length : longint;
              end;
Tsentence = record
             s : array [0..(maxsentencelength - 1)] of Tentry;
             length : longint;
            end;
Ttext = record
```

```
      t : array [0..(maxsentencelength - 1)] of Tsentence;
      length : longint;
     end;

Tpointer = ^Tnode;
Tnode = record
      entry : Tentry;
      left : Tpointer;
      right : Tpointer;
     end;

Tnafsform = class(TForm)
  StaticText1: TStaticText;
  Memo1: TMemo;
  OpenDialog1: TOpenDialog;
  RadioButton5: TRadioButton;
  RadioButton6: TRadioButton;
  RadioButton7: TRadioButton;
  radiobutton1: TRadioButton;
  RadioButton4: TRadioButton;
  RadioButton3: TRadioButton;
  procedure radiobutton1Click(Sender: TObject);
```

```
    procedure RadioButton3Click(Sender: TObject);
    procedure FormCreate(Sender: TObject);
    procedure RadioButton4Click(Sender: TObject);
    procedure RadioButton5Click(Sender: TObject);
    procedure RadioButton6Click(Sender: TObject);
    procedure RadioButton7Click(Sender: TObject);
    procedure RadioButton2Click(Sender: TObject);
  private
    { Private declarations }
  public
    { Public declarations }
  end;

var
  nafsform: Tnafsform;


  filenamelist : Tfilenamelist;
  currenttext : Ttext;
  entrylist : Tentrylist;
  dictionary : Tpointer;
  onscreenoutput : boolean;
  newnode : Tpointer;
  rootnode : Tpointer;

  infile : textfile;
  inbuffer: array[1..8192] of char;
  outfile : textfile;
  outbuffer: array[1..8192] of char;

implementation

{$R *.dfm}

procedure makenode (var root : Tpointer;
                        entry : Tentry);
begin
 {nafsform.Memo1.lines.add (content.lexis + content.lemma); }
 new (root);
 root^.entry := entry;
 root^.left := nil;
 root^.right := nil;
end;
```

```
procedure insertnode (var parentnode : Tpointer;
                          newnode : Tpointer);
begin
 if parentnode = nil then
  parentnode := newnode
 else
  if newnode^.entry.word <> parentnode^.entry.word then
   if newnode^.entry.word < parentnode^.entry.word then
    insertnode (parentnode^.left, newnode)
   else
    insertnode (parentnode^.right, newnode);
end;

procedure outputinorder (var root : Tpointer);
var
 i,j : longint;
begin
 if root <> nil then
  begin
   outputinorder (root^.left);
   writeln (outfile, root^.entry.word, ' ', root^.entry.pos, ' ',
root^.entry.gender, ' ', root^.entry.number);
```

```
   outputinorder (root^.right);
  end;
end;

procedure writesentencetomemo (sentence : Tsentence);
var
 str : packed array [0..499] of char;
 currententry : Tentry;
 strindex : longint;
 i,j : longint;
begin
 for strindex := 0 to 499 do
  str[strindex] := ' ';
 strindex := 0;
 for i := 0 to (sentence.length - 1) do
  begin
   currententry := sentence.s[i];
   {nafsform.memo1.Lines.add (currentword.w);}
   j := 0;
   while (currententry.word[j] in ['A'..'Z', 'a'..'z', '.', '?', '!', chr(39), '}', '>',
'<', '&','*', '~', '$', '/', chr(124), '/', '_']) and (j <= maxwordlength) do
    begin
```

```
     str [strindex] := currententry.word[j];
     j := j + 1;
     strindex := strindex + 1;
    end;
   str[strindex] := ' ';
   strindex := strindex + 1;
  end;
 nafsform.Memo1.lines.add (str);
end;

procedure readtext (    nr : longint);
var
 currententry : Tentry;
 currentsentence : Tsentence;
 endofword : boolean;
 endofsentence : boolean;
 ch : char;
 i : longint;
begin
 assignfile (infile, filenamelist.list [nr]);
 reset (infile);
 system.settextbuf (infile, inbuffer);
 currenttext.length := 0;
 while not eof(infile) do
  begin
   endofsentence := false;
   currentsentence.length := 0;
   while (not endofsentence) and (not eof(infile)) do
    begin
     while (not (ch in ['A'..'Z', 'a'..'z', '.', '?', '!', chr(39), '}', '>', '<', '&','*',
'~', '$', '/', chr(124), '/', '_'])) and (not eoln (infile)) and (not eof(infile))
do
      read(infile, ch);
     if not (eof(infile)) then
      begin
       if eoln (infile) then
        readln (infile)
       else
        begin
         endofword := false;
         for i := 0 to (maxwordlength - 1) do
          currententry.word[i] := ' ';
         currententry.person := 0;
         currententry.gender := 'x';

         currententry.number := 'x';

         currententry.match := false;

         currententry.word[0] := ch;

         i := 1;

         while (ch in ['A'..'Z', 'a'..'z', '.', '?', '!', chr(39), '}', '>', '<', '&','*', '~', '$', '/', chr(124), '/', '_']) and (not endofword) do

          begin

           read (infile,ch);

           if ch in ['.', '?', '!'] then

            begin

             currentsentence.s[currentsentence.length] := currententry;

             currentsentence.length := currentsentence.length + 1;

             currenttext.t[currenttext.length] := currentsentence;

             currenttext.length := currenttext.length + 1;

             endofword := true;

             endofsentence := true;

            end;

           if ch = ' ' then

            begin

             currentsentence.s[currentsentence.length] := currententry;

             currentsentence.length := currentsentence.length + 1;
```

```
             endofword := true;
            end;
           if ch in ['A'..'Z', 'a'..'z', '.', '?', '!', chr(39), '}', '>', '<', '&','*', '~', '$',
'/', chr(124), '/', '_'] then
            begin
             currententry.word[i] := ch;
             i := i + 1;
            end;
          end;
        end;
      end;
    end;
  end;
 nafsform.memo1.lines.add (' ');
 nafsform.Memo1.lines.add (filenamelist.list [nr]);
 closefile (infile);
end;


function isnafs (entry : Tentry) : boolean;
 begin
  if (entry.word = 'nafsi                ') or
     (entry.word = 'nafosiy              ') or
```

```
(entry.word = 'nfsy                     ') or

(entry.word = 'nafsina                  ') or

(entry.word = 'nafosinaA                ') or

(entry.word = 'nfsynA                   ') or

(entry.word = 'nafsak                   ') or

(entry.word = 'nafosak                  ') or

(entry.word = 'nfsk                     ') or

(entry.word = 'nafsik                   ') or

(entry.word = 'nafosik                  ') or

(entry.word = 'nfsk                     ') or

(entry.word = 'nafsukuma               ') or

(entry.word = 'nafosukumaA             ') or

(entry.word = 'nfskmA                   ') or

(entry.word = 'nafsukum                 ') or

(entry.word = 'nafosukum                ') or

(entry.word = 'nfskm                    ') or

(entry.word = 'nafsukunna              ') or

(entry.word = 'nafosukun~               ') or

(entry.word = 'nfskn                    ') or

(entry.word = 'nafsuhu                  ') or

(entry.word = 'nafosuhu                 ') or

(entry.word = 'nfsh                     ') or
```

```
(entry.word = 'nafsaha              ') or
(entry.word = 'nafosahaA             ') or
(entry.word = 'nfshA                ') or
(entry.word = 'nafsuhuma             ') or
(entry.word = 'nafosuhumaA           ') or
(entry.word = 'nfshmA               ') or
(entry.word = 'nafsuhum             ') or
(entry.word = 'nafosuhum             ') or
(entry.word = 'nfshm                ') or
(entry.word = 'nafsuhunna            ') or
(entry.word = 'nafosuhun~            ') or
(entry.word = 'nfshn                ') or


(entry.word = 'bnafsi               ') or
(entry.word = 'bnafosiy             ') or
(entry.word = 'bnfsy                ') or
(entry.word = 'bnafsina             ') or
(entry.word = 'bnafosinaA            ') or
(entry.word = 'bnfsynA              ') or
(entry.word = 'bnafsak              ') or
(entry.word = 'bnafosak             ') or
(entry.word = 'bnfsk                ') or
```

```
(entry.word = 'bnafsik          ') or
(entry.word = 'bnafosik         ') or
(entry.word = 'bnfsk           ') or
(entry.word = 'bnafsukuma        ') or
(entry.word = 'bnafosukumaA       ') or
(entry.word = 'bnfskmA          ') or
(entry.word = 'bnafsukum         ') or
(entry.word = 'bnafosukum        ') or
(entry.word = 'bnfskm           ') or
(entry.word = 'bnafsukunna        ') or
(entry.word = 'bnafosukun~        ') or
(entry.word = 'bnfskn           ') or
(entry.word = 'bnafsuhu          ') or
(entry.word = 'bnafosuhu         ') or
(entry.word = 'bnfsh           ') or
(entry.word = 'bnafsaha          ') or
(entry.word = 'bnafosahaA        ') or
(entry.word = 'bnfshA           ') or
(entry.word = 'bnafsuhuma        ') or
(entry.word = 'bnafosuhumaA       ') or
(entry.word = 'bnfshmA          ') or
(entry.word = 'bnafsuhum         ') or
```

```
(entry.word = 'bnafosuhum              ') or
(entry.word = 'bnfshm                  ') or
(entry.word = 'bnafsuhunna             ') or
(entry.word = 'bnafosuhun~             ') or
(entry.word = 'bnfshn                  ') or


(entry.word = 'fnafsi                  ') or
(entry.word = 'fnafosiy                ') or
(entry.word = 'fnfsy                   ') or
(entry.word = 'fnafsina                ') or
(entry.word = 'fnafosinaA              ') or
(entry.word = 'fnfsynA                 ') or
(entry.word = 'fnafsak                 ') or
(entry.word = 'fnafosak                ') or
(entry.word = 'fnfsk                   ') or
(entry.word = 'fnafsik                 ') or
(entry.word = 'fnafosik                ') or
(entry.word = 'fnfsk                   ') or
(entry.word = 'fnafsukuma              ') or
(entry.word = 'fnafosukumaA            ') or
(entry.word = 'fnfskmA                 ') or
(entry.word = 'fnafsukum               ') or
```

```
(entry.word = 'fnafosukum              ') or
(entry.word = 'fnfskm                  ') or
(entry.word = 'fnafsukunna             ') or
(entry.word = 'fnafosukun~             ') or
(entry.word = 'fnfskn                  ') or
(entry.word = 'fnafsuhu                ') or
(entry.word = 'fnafosuhu               ') or
(entry.word = 'fnfsh                   ') or
(entry.word = 'fnafsaha                ') or
(entry.word = 'fnafosahaA              ') or
(entry.word = 'fnfshA                  ') or
(entry.word = 'fnafsuhuma              ') or
(entry.word = 'fnafosuhumaA            ') or
(entry.word = 'fnfshmA                 ') or
(entry.word = 'fnafsuhum               ') or
(entry.word = 'fnafosuhum              ') or
(entry.word = 'fnfshm                  ') or
(entry.word = 'fnafsuhunna             ') or
(entry.word = 'fnafosuhun~             ') or
(entry.word = 'fnfshn                  ') or

(entry.word = 'lnafsi                  ') or
```

```
(entry.word = 'lnafosiy          ') or
(entry.word = 'lnfsy            ') or
(entry.word = 'lnafsina           ') or
(entry.word = 'lnafosinaA          ') or
(entry.word = 'lnfsynA            ') or
(entry.word = 'lnafsak           ') or
(entry.word = 'lnafosak           ') or
(entry.word = 'lnfsk            ') or
(entry.word = 'lnafsik           ') or
(entry.word = 'lnafosik          ') or
(entry.word = 'lnfsk            ') or
(entry.word = 'lnafsukuma           ') or
(entry.word = 'lnafosukumaA           ') or
(entry.word = 'lnfskmA             ') or
(entry.word = 'lnafsukum           ') or
(entry.word = 'lnafosukum           ') or
(entry.word = 'lnfskm             ') or
(entry.word = 'lnafsukunna          ') or
(entry.word = 'lnafosukun~          ') or
(entry.word = 'lnfskn            ') or
(entry.word = 'lnafsuhu           ') or
(entry.word = 'lnafosuhu           ') or
```

```
(entry.word = 'lnfsh            ') or
(entry.word = 'lnafsaha           ') or
(entry.word = 'lnafosahaA           ') or
(entry.word = 'lnfshA            ') or
(entry.word = 'lnafsuhuma           ') or
(entry.word = 'lnafosuhumaA           ') or
(entry.word = 'lnfshmA             ') or
(entry.word = 'lnafsuhum           ') or
(entry.word = 'lnafosuhum           ') or
(entry.word = 'lnfshm             ') or
(entry.word = 'lnafsuhunna          ') or
(entry.word = 'lnafosuhun~          ') or
(entry.word = 'lnfshn            ') or


(entry.word = 'l>nafsi           ') or
(entry.word = 'l>nafosiy          ') or
(entry.word = 'l>nfsy            ') or
(entry.word = 'l>nafsina          ') or
(entry.word = 'l>nafosinaA          ') or
(entry.word = 'l>nfsynA            ') or
(entry.word = 'l>nafsak           ') or
(entry.word = 'l>nafosak           ') or
```

```
(entry.word = 'l>nfsk                  ') or
(entry.word = 'l>nafsik                ') or
(entry.word = 'l>nafosik               ') or
(entry.word = 'l>nfsk                  ') or
(entry.word = 'l>nafsukuma             ') or
(entry.word = 'l>nafosukumaA           ') or
(entry.word = 'l>nfskmA                ') or
(entry.word = 'l>nafsukum              ') or
(entry.word = 'l>nafosukum             ') or
(entry.word = 'l>nfskm                 ') or
(entry.word = 'l>nafsukunna            ') or
(entry.word = 'l>nafosukun~            ') or
(entry.word = 'l>nfskn                 ') or
(entry.word = 'l>nafsuhu               ') or
(entry.word = 'l>nafosuhu              ') or
(entry.word = 'l>nfsh                  ') or
(entry.word = 'l>nafsaha               ') or
(entry.word = 'l>nafosahaA             ') or
(entry.word = 'l>nfshA                 ') or
(entry.word = 'l>nafsuhuma             ') or
(entry.word = 'l>nafosuhumaA           ') or
(entry.word = 'l>nfshmA                ') or
```

```
(entry.word = 'l>nafsuhum               ') or
(entry.word = 'l>nafosuhum              ') or
(entry.word = 'l>nfshm                  ') or
(entry.word = 'l>nafsuhunna             ') or
(entry.word = 'l>nafosuhun~             ') or
(entry.word = 'l>nfshn                  ') or


(entry.word = '>nafsi                   ') or
(entry.word = '>nafosiy                 ') or
(entry.word = '>nfsy                    ') or
(entry.word = '>nafsina                 ') or
(entry.word = '>nafosinaA               ') or
(entry.word = '>nfsynA                  ') or
(entry.word = '>nafsak                  ') or
(entry.word = '>nafosak                 ') or
(entry.word = '>nfsk                    ') or
(entry.word = '>nafsik                  ') or
(entry.word = '>nafosik                 ') or
(entry.word = '>nfsk                    ') or
(entry.word = '>nafsukuma               ') or
(entry.word = '>nafosukumaA             ') or
(entry.word = '>nfskmA                  ') or
```

```
  (entry.word = '>nafsukum          ') or
  (entry.word = '>nafosukum         ') or
  (entry.word = '>nfskm             ') or
  (entry.word = '>nafsukunna        ') or
  (entry.word = '>nafosukun~        ') or
  (entry.word = '>nfskn             ') or
  (entry.word = '>nafsuhu           ') or
  (entry.word = '>nafosuhu          ') or
  (entry.word = '>nfsh              ') or
  (entry.word = '>nafsaha           ') or
  (entry.word = '>nafosahaA         ') or
  (entry.word = '>nfshA             ') or
  (entry.word = '>nafsuhuma         ') or
  (entry.word = '>nafosuhumaA       ') or
  (entry.word = '>nfshmA            ') or
  (entry.word = '>nafsuhum          ') or
  (entry.word = '>nafosuhum         ') or
  (entry.word = '>nfshm             ') or
  (entry.word = '>nafsuhunna        ') or
  (entry.word = '>nafosuhun~        ') or
  (entry.word = '>nfshn             ') then
isnafs := true
```

```
  else
   isnafs := false;
end;


{procedure match (var root : Tpointer;
                     candidateentry : Tentry;
                 var entry : Tentry;
                 var found : boolean);
begin
 found := false;
 if root <> nil then
  begin
   nafsform.Memo1.lines.add (candidateentry.word + ' ' + root^.entry.word);
   if root^.entry.word = candidateentry.word then
    begin
     entry := root^.entry;
     found := true;
    end
   else
    if candidateentry.word < root^.entry.word then
     match (root^.left, candidateentry, entry, found)
    else
     match (root^.right, candidateentry, entry, found);
  end;
end;}

procedure match (var root : Tpointer;
                      candidateentry : Tentry;
                  var entry : Tentry;
                  var found : boolean);
var
 i : longint;
begin
 found := false;
 i := 0;
 while (i < entrylist.length) and (not found) do
  begin
   if candidateentry.word = entrylist.e[i].word then
    begin
     entry := entrylist.e[i];
     found := true
    end
   else
```

```
    i := i + 1;
  end;
end;

procedure resolve;
var
 previoussentence : Tsentence;
 currentsentence : Tsentence;
 previousentry : Tentry;
 currententry : Tentry;
 currentnafs : Tentry;
 nafsindex : longint;
 startatindex : longint;
 candidate : Tentry;
 dictionaryentry : Tentry;
 previousdictionaryentry : Tentry;
 referent : Tentry;
 nafsfound : boolean;
 dictionaryentryfound : boolean;
 referentfound : boolean;
 i,j,k,m,n : longint;
```

```
begin
 {For each sentence in the current text}
 for i := 0 to (currenttext.length - 1) do
  begin
   {Keep track of the preceding sentence in case it's necessary for resolution}
   if i > 0 then
    previoussentence := currentsentence;
   {Get the sentence to be examined for possible resolution}
   currentsentence := currenttext.t [i];
   if onscreenoutput then
    nafsform.memo1.lines.Add('Current sentence: ' + inttostr(i+1));
   {Process the current sentence; there might be more than one instance of nafs}
   startatindex := 0;
   j := 0;
   while j < currentsentence.length do
    begin
     nafsfound := false;
     {Keep looking until an instance of nafs is found}
     while (j < currentsentence.length) and (not nafsfound) do
      begin
```

```
{Look at each word, here entry, in turn, keeping track of the previous word}

previousentry := currententry;

currententry := currentsentence.s [j];

{If the current word / entry is one of the many nafs forms}

if isnafs (currententry) then

 begin

  {Write some relevant output both to the screen and to the output file}

  if onscreenoutput then

   begin

    nafsform.memo1.lines.add (' ');

    nafsform.memo1.lines.add ('Sentence ' + inttostr (i + 1));

    writesentencetomemo(currenttext.t[i]);

   end;

  writeln (outfile, 'Sentence: ', (i+1));

  for k := 0 to (currentsentence.length - 1) do

   begin

    m := 0;

    while currentsentence.s[k].word[m] in ['A'..'Z', 'a'..'z', '.', '?', '!', chr(39), '}', '>', '<', '&','*', '~', '$', '/', chr(124), '/', '_'] do

     begin

      write (outfile, currentsentence.s[k].word[m]);
      m := m + 1;
     end;
    write (outfile, ' ');
   end;
  writeln (outfile);
  {Having done the output housekeeping, proceed with the resolution}
  currentnafs := currententry;
  nafsfound := true;
  nafsindex := j; {where nafs is in the sentence}
  {nafsform.Memo1.lines.add (inttostr(nafsindex));}
  writeln (outfile, 'Nafs form: ', currentnafs.word);
  {Now assign the necessary grammatical information to the current nafs form}
  if (currentnafs.word = 'nafsi                    ') or
     (currentnafs.word = 'bnafsi                   ') or
     (currentnafs.word = 'lnafsi                   ') or
     (currentnafs.word = 'fnafsi                   ') or
     (currentnafs.word = 'nafosiy                  ') or
     (currentnafs.word = 'bnafosiy                 ') or
     (currentnafs.word = 'lnafosiy                 ') or
     (currentnafs.word = 'fnafosiy                 ') or
     (currentnafs.word = 'nfsy                    ') or
     (currentnafs.word = 'bnfsy                   ') or
     (currentnafs.word = 'lnfsy                   ') or
     (currentnafs.word = 'fnfsy                   ') then
   begin
    currentnafs.person := 1;
    currentnafs.gender := 'c';
    currentnafs.number := 's';
   end;
  if (currentnafs.word = 'nafsina                  ') or
     (currentnafs.word = 'bnafsina                 ') or
     (currentnafs.word = 'lnafsina                 ') or
     (currentnafs.word = 'fnafsina                 ') or
     (currentnafs.word = 'nafosinaA                  ') or
     (currentnafs.word = 'bnafosinaA                 ') or
     (currentnafs.word = 'lnafosinaA                 ') or
     (currentnafs.word = 'fnafosinaA                 ') or
     (currentnafs.word = 'nfsynA                   ') or
     (currentnafs.word = 'bnfsynA                  ') or
     (currentnafs.word = 'lnfsynA                  ') or
     (currentnafs.word = 'fnfsynA                  ') then
   begin
```

```
  currentnafs.person := 1;
  currentnafs.gender := 'c';
  currentnafs.number := 'p';
 end;
if (currentnafs.word = 'nafsak               ') or
   (currentnafs.word = 'bnafsak              ') or
   (currentnafs.word = 'lnafsak              ') or
   (currentnafs.word = 'fnafsak              ') or
   (currentnafs.word = 'nafosak              ') or
   (currentnafs.word = 'bnafosak             ') or
   (currentnafs.word = 'lnafosak             ') or
   (currentnafs.word = 'fnafosak             ') or
   (currentnafs.word = 'nfsk                 ') or
   (currentnafs.word = 'bnfsk                ') or
   (currentnafs.word = 'lnfsk                ') or
   (currentnafs.word = 'fnfsk                ') then
 begin
  currentnafs.person := 2;
  currentnafs.gender := 'm';
  currentnafs.number := 's';
 end;
if (currentnafs.word = 'nafsik               ') or
```

```
   (currentnafs.word = 'bnafsik                    ') or
   (currentnafs.word = 'lnafsik                    ') or
   (currentnafs.word = 'fnafsik                    ') or
   (currentnafs.word = 'nafosik                    ') or
   (currentnafs.word = 'bnafosik                   ') or
   (currentnafs.word = 'lnafosik                   ') or
   (currentnafs.word = 'fnafosik                   ') or
   (currentnafs.word = 'nfsk                        ') or
   (currentnafs.word = 'bnfsk                       ') or
   (currentnafs.word = 'lnfsk                       ') or
   (currentnafs.word = 'fnfsk                       ') then
 begin
  currentnafs.person := 2;
  currentnafs.gender := 'f';
  currentnafs.number := 's';
 end;
if (currentnafs.word = 'nafsukuma                  ') or
   (currentnafs.word = 'bnafsukuma                 ') or
   (currentnafs.word = 'lnafsukuma                 ') or
   (currentnafs.word = 'fnafsukuma                 ') or
   (currentnafs.word = '>nafsukuma                 ') or
   (currentnafs.word = 'nafosukumaA                ') or
   (currentnafs.word = 'bnafosukumaA          ') or
   (currentnafs.word = 'lnafosukumaA          ') or
   (currentnafs.word = 'l>nafosukumaA         ') or
   (currentnafs.word = 'fnafosukumaA          ') or
   (currentnafs.word = '>nafosukumaA          ') or
   (currentnafs.word = 'nfskmA                ') or
   (currentnafs.word = 'bnfskmA               ') or
   (currentnafs.word = 'lnfskmA               ') or
   (currentnafs.word = 'l>nfskmA              ') or
   (currentnafs.word = 'fnfskmA               ') or
   (currentnafs.word = '>nfskmA               ') then
 begin
  currentnafs.person := 2;
  currentnafs.gender := 'c';
  currentnafs.number := 'd';
 end;
if (currentnafs.word = 'nafsukum              ') or
   (currentnafs.word = 'bnafsukum             ') or
   (currentnafs.word = 'lnafsukum             ') or
   (currentnafs.word = 'fnafsukum             ') or
   (currentnafs.word = '>nafsukum             ') or
   (currentnafs.word = 'nafosukum             ') or
   (currentnafs.word = 'bnafosukum          ') or
   (currentnafs.word = 'lnafosukum          ') or
   (currentnafs.word = 'l>nafosukum         ') or
   (currentnafs.word = 'fnafosukum          ') or
   (currentnafs.word = '>nafosukum          ') or
   (currentnafs.word = 'nfskm               ') or
   (currentnafs.word = 'bnfskm              ') or
   (currentnafs.word = 'lnfskm              ') or
   (currentnafs.word = 'l>nfskm             ') or
   (currentnafs.word = 'fnfskm              ') or
   (currentnafs.word = '>nfskm              ') then
 begin
  currentnafs.person := 2;
  currentnafs.gender := 'm';
  currentnafs.number := 'p';
 end;
if (currentnafs.word = 'nafsukunna          ') or
   (currentnafs.word = 'bnafsukunna         ') or
   (currentnafs.word = 'lnafsukunna         ') or
   (currentnafs.word = 'fnafsukunna         ') or
   (currentnafs.word = '>nafsukunna         ') or
   (currentnafs.word = 'nafosukun~          ') or
   (currentnafs.word = 'bnafosukun~            ') or
   (currentnafs.word = 'lnafosukun~            ') or
   (currentnafs.word = 'l>bnafosukun~          ') or
   (currentnafs.word = 'fnafosukun~            ') or
   (currentnafs.word = '>nafosukun~            ') or
   (currentnafs.word = 'nfskn                  ') or
   (currentnafs.word = 'bnfskn                 ') or
   (currentnafs.word = 'lnfskn                 ') or
   (currentnafs.word = 'l>nfskn                ') or
   (currentnafs.word = 'fnfskn                 ') or
   (currentnafs.word = '>nfskn                 ') then
begin
  currentnafs.person := 2;
  currentnafs.gender := 'f';
  currentnafs.number := 'p';
 end;
if (currentnafs.word = 'nafsuhu                ') or
   (currentnafs.word = 'bnafsuhu               ') or
   (currentnafs.word = 'lnafsuhu               ') or
   (currentnafs.word = 'fnafsuhu               ') or
   (currentnafs.word = 'nafosuhu               ') or
   (currentnafs.word = 'bnafosuhu              ') or
   (currentnafs.word = 'lnafosuhu               ') or
   (currentnafs.word = 'fnafosuhu               ') or
   (currentnafs.word = 'nfsh                    ') or
   (currentnafs.word = 'bnfsh                   ') or
   (currentnafs.word = 'lnfsh                   ') or
   (currentnafs.word = 'fnfsh                   ') then
 begin
  currentnafs.person := 3;
  currentnafs.gender := 'm';
  currentnafs.number := 's';
 end;
if (currentnafs.word = 'nafsaha                 ') or
   (currentnafs.word = 'bnafsaha                ') or
   (currentnafs.word = 'lnafsaha                ') or
   (currentnafs.word = 'fnafsaha                ') or
   (currentnafs.word = 'nafosahaA               ') or
   (currentnafs.word = 'bnafosahaA              ') or
   (currentnafs.word = 'lnafosahaA              ') or
   (currentnafs.word = 'fnafosahaA              ') or
   (currentnafs.word = 'nfshA                   ') or
   (currentnafs.word = 'bnfshA                  ') or
   (currentnafs.word = 'lnfshA                  ') or
   (currentnafs.word = 'fnfshA              ') then
 begin
  currentnafs.person := 3;
  currentnafs.gender := 'f';
  currentnafs.number := 's';
 end;
if (currentnafs.word = 'nafsuhuma           ') or
   (currentnafs.word = 'bnafsuhuma          ') or
   (currentnafs.word = 'lnafsuhuma          ') or
   (currentnafs.word = 'fnafsuhuma          ') or
   (currentnafs.word = '>nafsuhuma          ') or
   (currentnafs.word = 'nafosuhumaA         ') or
   (currentnafs.word = 'bnafosuhumaA        ') or
   (currentnafs.word = 'lnafosuhumaA        ') or
   (currentnafs.word = 'l>nafosuhumaA       ') or
   (currentnafs.word = 'fnafosuhumaA        ') or
   (currentnafs.word = '>nafosuhumaA        ') or
   (currentnafs.word = 'nfshmA              ') or
   (currentnafs.word = 'bnfshmA             ') or
   (currentnafs.word = 'lnfshmA             ') or
   (currentnafs.word = 'l>nfshmA            ') or
   (currentnafs.word = 'fnfshmA             ') or
   (currentnafs.word = '>nfshmA                ') then
 begin
  currentnafs.person := 3;
  currentnafs.gender := 'c';
  currentnafs.number := 'd';
 end;
if (currentnafs.word = 'nafsuhum              ') or
   (currentnafs.word = 'bnafsuhum             ') or
   (currentnafs.word = 'lnafsuhum             ') or
   (currentnafs.word = 'fnafsuhum             ') or
   (currentnafs.word = '>nafsuhum             ') or
   (currentnafs.word = 'nafosuhum             ') or
   (currentnafs.word = 'bnafosuhum            ') or
   (currentnafs.word = 'lnafosuhum            ') or
   (currentnafs.word = 'l>nafosuhum           ') or
   (currentnafs.word = 'fnafosuhum            ') or
   (currentnafs.word = '>nafosuhum            ') or
   (currentnafs.word = 'nfshm                 ') or
   (currentnafs.word = 'bnfshm                ') or
   (currentnafs.word = 'lnfshm                ') or
   (currentnafs.word = 'l>nfshm               ') or
   (currentnafs.word = 'fnfshm                ') or
   (currentnafs.word = '>nfshm                ') then
 begin
  currentnafs.person := 3;
  currentnafs.gender := 'm';
  currentnafs.number := 'p';
 end;
if (currentnafs.word = 'nafsuhunna            ') or
   (currentnafs.word = 'bnafsuhunna           ') or
   (currentnafs.word = 'lnafsuhunna           ') or
   (currentnafs.word = 'fnafsuhunna           ') or
   (currentnafs.word = '>nafsuhunna           ') or
   (currentnafs.word = 'nafosuhun~            ') or
   (currentnafs.word = 'bnafosuhun~           ') or
   (currentnafs.word = 'lnafosuhun~           ') or
   (currentnafs.word = 'l>nafosuhun~          ') or
   (currentnafs.word = 'fnafosuhun~           ') or
   (currentnafs.word = '>nafosuhun~           ') or
   (currentnafs.word = 'nfshn                ') or
   (currentnafs.word = 'bnfshn               ') or
   (currentnafs.word = 'lnfshn               ') or
   (currentnafs.word = 'l>nfshn              ') or
   (currentnafs.word = 'fnfshn               ') or
   (currentnafs.word = '>nfshn                ') then
 begin
  currentnafs.person := 3;
  currentnafs.gender := 'f';
  currentnafs.number := 'p';
 end;
if onscreenoutput then
 nafsform.Memo1.lines.add ('Nafs form: ' + currentnafs.word + currentnafs.gender + currentnafs.number);
{Start looking for the referent of nafs}
referentfound := false;
{start looking backwards through the sentence starting with the word left of the nafs form}
k := nafsindex - 1;
{While no referent has been found and the start of the sentence has not been reached (note that}
{ the referent might be in the preceding sentence)}
while (not referentfound) and (k >= startatindex) do
 begin
  currententry := currentsentence.s [k];
  if onscreenoutput then
   nafsform.Memo1.lines.add ('Candidate referent: ' + currententry.word);
```

```
        match (rootnode, currententry, dictionaryentry,
dictionaryentryfound);

        if dictionaryentryfound then

         begin

          if onscreenoutput then

           nafsform.Memo1.lines.add ('Dictionary entry found: ' +
dictionaryentry.word + dictionaryentry.gender +
dictionaryentry.number);

          if (dictionaryentry.pos = 'v') and ((dictionaryentry.gender =
currentnafs.gender) or (currentnafs.gender = 'c')) and
(dictionaryentry.number = currentnafs.number) then

           begin

            referent := dictionaryentry;

            referentfound := true;

            if onscreenoutput then

             begin

              nafsform.memo1.lines.add ('Referent: ' + referent.word +
referent.gender + referent.number);

              nafsform.Memo1.lines.add (' ');

             end;

            writeln (outfile, 'Referent: ', referent.word);

            writeln (outfile);

           end

          else
          begin

           {If the grammatical features match or the special case obtains}

           if (((dictionaryentry.gender = currentnafs.gender) or
(currentnafs.gender = 'c')) and (dictionaryentry.number =
currentnafs.number)) or

              ((currentnafs.word = 'nfshA                    ') and
(dictionaryentry.gender = 'f') and (dictionaryentry.number = 'p')) then
{special case}

            begin

             {See if there's a word preceding the current one, in which
case that preceding word is the one required}

             previousdictionaryentry := dictionaryentry; {save the entry
already found in case the following condition doesn't hold}

             match (rootnode, previousentry, dictionaryentry,
dictionaryentryfound);

             if (((dictionaryentry.gender = currentnafs.gender) or
(currentnafs.gender = 'c')) and (dictionaryentry.number =
currentnafs.number)) or

               ((currentnafs.word = 'nfshA                    ') and
(dictionaryentry.gender = 'f') and (dictionaryentry.number = 'p')) then

              begin

               referent := dictionaryentry;

               referentfound := true;

               if onscreenoutput then

                begin

                 nafsform.memo1.lines.add ('Referent: ' + referent.word +
referent.gender + referent.number);
```

```
         nafsform.Memo1.lines.add (' ');
        end;
       writeln (outfile, 'Referent: ', referent.word);
       writeln (outfile);
      end
     {If the preceding word didn't match}
     else
      begin
       referent := previousdictionaryentry;
       referentfound := true;
       if onscreenoutput then
        begin
         nafsform.memo1.lines.add ('Referent: ' + referent.word + referent.gender + referent.number);
         nafsform.Memo1.lines.add (' ');
        end;
       writeln (outfile, 'Referent: ', referent.word);
       writeln (outfile);
      end;
    end;
  end;
end;
```

```
           k := k - 1;

          end;


         {If no match was found in the current sentence, try looking in the
previous sentence using the same procedure as above}

         if not referentfound then

          begin

           {If the current sentence is the first in the text then there's no
previous sentence,

            so this test can't apply}

           if i > 0 then

            begin

             k := previoussentence.length;;

             while (not referentfound) and (k >= 0) do

              begin

               currententry := previoussentence.s [k];

               if onscreenoutput then

                nafsform.Memo1.lines.add ('Candidate referent: ' +
currententry.word);

               match (rootnode, currententry, dictionaryentry,
dictionaryentryfound);

               if dictionaryentryfound then

                begin
```

```
if onscreenoutput then
 nafsform.Memo1.lines.add ('Dictionary entry found: ' + dictionaryentry.word + dictionaryentry.gender + dictionaryentry.number);
if (dictionaryentry.pos = 'v') and ((dictionaryentry.gender = currentnafs.gender) or (currentnafs.gender = 'c')) and (dictionaryentry.number = currentnafs.number) then
 begin
  referent := dictionaryentry;
  referentfound := true;
  if onscreenoutput then
   begin
    nafsform.memo1.lines.add ('Referent: ' + referent.word + referent.gender + referent.number);
    nafsform.Memo1.lines.add (' ');
   end;
  writeln (outfile, 'Referent: ', referent.word);
  writeln (outfile);
 end
else
 begin
  {If the grammatical features match or the special case obtains}
```

```
if (((dictionaryentry.gender = currentnafs.gender) or (currentnafs.gender = 'c')) and (dictionaryentry.number = currentnafs.number)) or

    ((currentnafs.word = 'nfshA                ') and (dictionaryentry.gender = 'f') and (dictionaryentry.number = 'p')) then {special case}

 begin

  {See if there's a word preceding the current one, in which case that preceding word is the one required}

  previousdictionaryentry := dictionaryentry; {save the entry already found in case the following condition doesn't hold}

  match (rootnode, previousentry, dictionaryentry, dictionaryentryfound);

  if (((dictionaryentry.gender = currentnafs.gender) or (currentnafs.gender = 'c')) and (dictionaryentry.number = currentnafs.number)) or

      ((currentnafs.word = 'nfshA                ') and (dictionaryentry.gender = 'f') and (dictionaryentry.number = 'p')) then

   begin

    referent := dictionaryentry;

    referentfound := true;

    if onscreenoutput then

     begin

      nafsform.memo1.lines.add ('Referent: ' + referent.word + referent.gender + referent.number);

      nafsform.Memo1.lines.add (' ');

     end;
```

```
         writeln (outfile, 'Referent: ', referent.word);
         writeln (outfile);
        end
       {If the preceding word didn't match}
       else
        begin
         referent := previousdictionaryentry;
         referentfound := true;
         if onscreenoutput then
          begin
           nafsform.memo1.lines.add  ('Referent:  '  +  referent.word + referent.gender + referent.number);
           nafsform.Memo1.lines.add (' ');
          end;
         writeln (outfile, 'Referent: ', referent.word);
         writeln (outfile);
        end;
      end;
    end;
  end;
 k := k - 1;
end;
```

```
          end;
         end;
        if not referentfound then
         begin
          if onscreenoutput then
           nafsform.Memo1.lines.add ('No referent found');
          writeln (outfile, 'No referent found');
          writeln (outfile);
         end;
       end;
      j := j + 1;
     end;
   end;
  end;
end;


procedure Tnafsform.radiobutton1Click(Sender: TObject);
 {Read file name list from external file}
var
 ch : char;
 i,j : longint;
begin
```

```
nafsform.opendialog1.execute;
assignfile (infile, nafsform.opendialog1.filename);
reset (infile);
system.settextbuf (infile, inbuffer);
i := 0;
while not eof (infile) do
 begin
  j := 0;
  while (not eoln (infile)) and (not eof (infile)) do
   begin
    read (infile, ch);
    {convert to lower case if necessary}
    if ch in ['A'..'Z'] then
     ch:= chr(ord(ch) + 32);
    filenamelist.list [i,j] := ch;
    j := j + 1;
   end;
  if not eof (infile) then
   readln (infile);
  {nafsform.Memo1.lines.add (filenamelist.list [i]); }
  i := i + 1;
 end;
 filenamelist.length := i;
 nafsform.memo1.lines.add ('File name list read: length ' + ' ' + inttostr (filenamelist.length));
 nafsform.memo1.Lines.add (' ');
 closefile (infile);
end;


procedure Tnafsform.RadioButton3Click(Sender: TObject);
var
 i : longint;
begin
 assignfile (outfile, 'resolution.txt');
 rewrite (outfile);
 system.settextbuf (outfile, outbuffer);

 for i := 0 to (filenamelist.length - 1) do
  begin
   writeln (outfile, filenamelist.list [i]);
   readtext(i);
   resolve;
   writeln (outfile);
  end;
 closefile (outfile);
 nafsform.memo1.lines.add ('Resolution complete');
end;

procedure Tnafsform.FormCreate(Sender: TObject);
begin
 onscreenoutput := false;
end;

procedure Tnafsform.RadioButton4Click(Sender: TObject);
begin
 onscreenoutput := true;
end;

procedure Tnafsform.RadioButton5Click(Sender: TObject);
{Read word list}
 var
 ch : char;
 i,j : longint;
begin
 nafsform.opendialog1.execute;
 assignfile (infile, nafsform.opendialog1.filename);
 reset (infile);

 system.settextbuf (infile, inbuffer);

 i := 0;

 while not eof (infile) do

  begin

   ch := '£';

   while not (ch in ['0'..'9']) do

    read(infile, ch);

   while ch <> ' ' do

    read (infile, ch);

   while not (ch in ['A'..'Z', 'a'..'z', '.', '?', '!', chr(39), '}', '>', '<', '&','*', '~',
'$', '/', chr(124), '/', '_']) do

    read(infile, ch);

   for j := 0 to (maxwordlength - 1) do

    entrylist.e[i].word[j] := ' ';

   entrylist.e[i].word[0] := ch;

   j := 1;

   while ch <> ' ' do

    begin

     read (infile, ch);

     if ch <> ' ' then

      entrylist.e [i].word [j] := ch;
     j := j + 1;
    end;

   while ch <> 'n' do
    read(infile, ch);
   while ch <> ' ' do
    read (infile, ch);
   while not (ch in ['M','F','m','f']) do
    read (infile, ch);
   entrylist.e [i].gender := ch;
   while not (ch in ['S','P','s','p', 'D', 'd']) do
    read (infile, ch);
   entrylist.e [i].number := ch;

   {nafsform.memo1.lines.add (inttostr (i) + entrylist.e [i].word + entrylist.e [i].gender + entrylist.e [i].number); }

   while (not eoln (infile)) and (not eof(infile)) do
    read(infile, ch);
   if not eof (infile) then
    readln (infile);
   i := i + 1;
```

```
  end;
 entrylist.length := i;
 nafsform.memo1.lines.add ('Word list read: length ' + ' ' + inttostr (entrylist.length));
 nafsform.memo1.Lines.add (' ');
 closefile (infile);
end;

procedure Tnafsform.RadioButton6Click(Sender: TObject);
{Build sort tree}
var
 currententry : Tentry;
 i : longint;
begin
 rootnode := nil;
 nafsform.Memo1.lines.add (inttostr(entrylist.length));
 for i := 0 to (entrylist.length - 1) do
  begin
   currententry := entrylist.e[i];
   makenode (newnode, currententry);
   {nafsform.memo1.lines.add (newnode^.entry.word + newnode^.entry.gender + newnode^.entry.number);}
   insertnode (rootnode, newnode);
  end;
 nafsform.Memo1.lines.add ('Dictionary created');
end;

procedure Tnafsform.RadioButton7Click(Sender: TObject);
var
 i : longint;
begin
 assignfile (outfile, 'dictionary.txt');
 rewrite (outfile);
 system.settextbuf (outfile, outbuffer);
 outputinorder(rootnode);
 closefile (outfile);
 nafsform.Memo1.lines.add ('Dictionary saved');
end;

procedure Tnafsform.RadioButton2Click(Sender: TObject);
{Read dictionary}
var
 i,j,k : longint;
 ch : char;
 currententry : Tentry;
begin
 nafsform.opendialog1.execute;
 assignfile (infile, nafsform.opendialog1.filename);
 reset (infile);
 system.settextbuf (infile, inbuffer);
 i := 0;
 while not eof (infile) do
  begin
   ch := '£';
   while not (ch in ['A'..'Z', 'a'..'z', '.', '?', '!', chr(39), '}', '>', '<', '&','*', '~', '$', '/', chr(124), '/', '_']) do
    read(infile, ch);
   for j := 0 to (maxwordlength - 1) do
    entrylist.e[i].word[j] := ' ';
   entrylist.e[i].word[0] := ch;
   j := 1;
   while ch in ['A'..'Z', 'a'..'z', '.', '?', '!', chr(39), '}', '>', '<', '&','*', '~', '$', '/', chr(124), '/', '_'] do
    begin
     read (infile, ch);
     entrylist.e[i].word[j] := ch;
     j := j + 1;
    end;
   ch := '£';
   while (ch <> 'n') and (ch <> 'v') do
    read(infile, ch);
   entrylist.e[i].pos := ch;
   ch := '£';
   while (ch <> 'm') and (ch <> 'f') and (ch <> 'c') do
    read(infile, ch);
   entrylist.e[i].gender := ch;
   ch := '£';
   while (ch <> 's') and (ch <> 'p') and (ch <> 'd') do
    read(infile, ch);
   entrylist.e[i].number := ch;

   {nafsform.memo1.lines.add (inttostr (i) + entrylist.e [i].word +
entrylist.e [i].pos + entrylist.e [i].gender + entrylist.e [i].number);}

   while (not eoln (infile)) and (not eof(infile)) do
    read(infile, ch);
   if not eof (infile) then
    readln (infile);
```

```
   i := i + 1;
  end;
 entrylist.length := i;
 nafsform.Memo1.lines.add ('Dictionary entries read: ' + inttostr (entrylist.length));
 closefile (infile);


 {rootnode := nil;
 for i := 0 to (entrylist.length - 1) do
  begin
   currententry := entrylist.e[i];
   makenode (newnode, currententry);
   insertnode (rootnode, newnode);
  end;
 nafsform. ('Dictionary read');}
end;


end.
```

Appendix 2

Some samples of the test corpus

Topic :Politics

مع اقتراب مرور سنة على باراك أوباما رئيساً للولايات المتحدة الأميركية لا يستطيع أحد أن يسجّل في مصلحته إيفاءً بوعوده التغييرية، ولا حتى ببعضها القليل.

فإذا لم يستطع أوباما طوال عام كامل إغلاق سجن غوانتانامو وهو أمر يجب أن يكون من حيث المبدأ عملاً روتينياً عادياً. ولعل هذا ما جعله يضع إغلاق هذا السجن على رأس أولوياته. ولكنه فشل في تحقيق ذلك خلال العام 2009. الأمر الذي يدّل على سوء تقدير للموقف من حيث حسابات المعوّقات التي تحول دون إغلاقه في عام.

بيد أن الأهم فقد فشل خلال عام أيضاً في أن يمرّر مشروعه الخاص بالضمان الصحي. وقد أثبتت الوقائع أنه أخطأ حتى في حساب موقف بعض أعضاء حزبه نفسه مما جعله يتعثر حتى الآن في أروقة الكونغرس. علماً أن تغييرات ومساومات كثيرة أُدخلت على المشروع الأصلي ومع ذلك ما زال ينتظر الفرج.

على أن إدارة بوش نفسها ومنذ عهدها الثاني تراجعت عملياً عن تلك السياسة بعد مسلسل الإخفاقات التي مُنيت بها، أو بعد مسلسل المآزق التي أدخلت أميركا في أتونها ولا سيما في العراق وأفغانستان.

من هنا لا يستطيع أحد أن يسجّل تغييراً واحداً أحدثه أوباما على سياسات بوش في عهده الثاني والذي أحدث فيه بوش نفسه تغييراً عن سياساته في السنوات الثلاث الأخيرة من عهده الأول.

لقد وجد باراك أوباما نفسه في المأزق أو المأزق مع هبوط شعبيته في استطلاعات الرأي العام وتأزم وضعه العسكري في أفغانستان، إلى جانب ارتباك موقفه في موضوع البرنامج النووي الإيراني بسبب الاقتراح الإيراني لحل الإشكال من خلال تبادل اليورانيوم المخصّب 5% أو أقل، بيورانيوم مخصّب بحدود 20% ليستخدم لأغراض علمية.

. وذلك ليقولوا إن أميركا حين تخطئ تصحّح نفسها بنفسها فالخير فيها لا محالة. –

أعلن وزير الخارجية البريطاني السابق ديفيد ميليباند أنه سوف يخوض غمار السباق للفوز بزعامة حزب العمال الذي خسر السلطة بعد هزيمته في انتخابات الخميس الماضي ليدشن بذلك معركة التنافس على المنصب الذي سيقود من يشغله المعارضة المقبلة.

وبإعلانه هذا يكون ميليباند أول من يرشِّح نفسه رسميا لزعامة حزب العمال بعد استقالة رئيس الوزراء السابق وزعيم الحزب جوردون براون إثر فشله باجتذاب حزب الديمقراطيين الأحرار الذي اختار التحالف مع حزب المحافظين لتشكيل الحكومة الجديدة.

وفي مؤتمر صحفي أمام مجلس العموم يوم الأربعاء قال ميليباند إن حزب العمال يحتاج إلى إعادة بناء نفسه كقوة إصلاح في السياسة البريطانية.

وأشاد ميليباند برئيس الوزراء السابق جوردون براون كما وجَّه التمنيات الطيبة لحكومة ديفيد كاميرون الجديدة.

واعتبر وزير الخارجية في الحكومة العمالية السابقة نفسه قادرا على إعادة بناء حزب العمال كقوة من أجل التغيير الاجتماعي والاقتصادي في البلاد خلال وجوده في مقاعد المعارضة.

وأضاف قائلا إن العماليين حققوا الكثير عبر حكوماتهم السابقة لكن هنالك ثمة مرحلة جديدة بأخطار وفرص وإمكانيات جديدة.

كما اعتبر التحالف بين حزبي المحافظين والديمقراطيين الأحرار بمثابة لحظة الزخم في السياسة البريطانية لكنها في الوقت ذاته تلقي مسؤوليات جساما على عاتق حزب العمال ليصبح بذلك قوة لتوحيد كل أطياف الوسط ويسار الوسط في البلاد.

لكن من المتوقع أن يواجه ميليباند منافسة من قيادات حزبية أخرى مثل جون كروداس عضو مجلس العموم الذي قال إنه يفكر بشكل جديٍّ بخوض السباق وإد بولز وزير شؤون المدارس السابق وآندي بيرنهام بالإضافة إلى إد ميليباند شقيق ديفيد ميليباند نفسه.

وقال كروداس إنه يفكِّر مليًّا بخوض سباق زعامة الحزب لكنه يفضل أن يترك الآن كل الخيارات والاحتمالات الأخرى مفتوحة أمامه.

ففي كلمة ألقاها أمام مجموعة كومباس التي تنتمي إلى يسار الوسط قال كروداس إنه تلقَّى العديد من الرسائل الالكترونية والنصية التي تحثه على ترشيح نفسه لزعامة الحزب.

وختم بقوله لن أقول أي شيء عن ذلك لطالما من شأن ذلك أن يضع الحصان أمام العربة.

أمَّا وزير الداخلية السابق ألان جونسون فقد استبعد ترشيح نفسه من السباق على زعامة حزب العمال.

وقال جونسون الذي فشل في عام بالحصول على أصوات كافية لانتخابه لمنصب نائب زعيم الحزب الذي ذهب إلى هاريت هارمان القائمة حاليا بأعمال زعيم الحزب إن

حزبه بحاجة إلى تجديد نفسه بعد عاما من وجوده في السلطة.

ودعت كوبر من ينوون ترشيح أنفسهم لزعامة الحزب إلى الخلود إلى فترة من التفكير القصير قبل إعلان ترشيحاتهم.

ومن بين الذين استبعدوا ترشيح أنفسهم لزعامة الحزب أيضا كارولاين فلينت الوزيرة السابقة وبيتر هين الوزير السابق لشؤون ويلز وهازل بليرز وزيرة الدولة السابقة.

فقد أعلنت فلينت تأييدها الصريح لترشيح ميليباند لشغل المنصب بينما قال هين إن الحزب يحتاج إلى بعض التفكير الجديد حول قضايا الإصلاح السياسي والعدالة الاجتماعية والمساواة والبيئة.

---

literature:cTopi

قبل عامين ونيف عندما كان بصري مشدوداً إلى منظر مدينة طبرية وبحيرتها الساحر وكانت كلَّ ذرة من كياني تتفاعل مع تلك اللوحة الرائعة وبينما كنت سارحاً فيما قاله أو سيقوله عاشق طبريا أنيس صايغ عن مدينته دقّ جرس الهاتف النقال ليقطع سكون المكان حيث أقف على مرتفعات أم قيس شرقي الأردن.

يا الله إنه أنيس صايغ نفسه يتصل من بيروت مهنئاً بعيد الفطر. وكأنه التقط اللحظة الأروع والأنسب للمشاركة في حب فلسطين.

لا أدعي علاقة قديمة بأنيس فعلاقتي لا تزيد على بضع سنوات هي فترة إقامتي في بيروت وهناك المئات من تلامذته ومريديه ومحبيه ممن سبقوني إليه.

ولكن وجوده معنا في الهيئة الاستشارية لمركز الزيتونة ومستشاراً للتقرير الإستراتيجي الفلسطيني ومشاركته الدائمة ودعمه الدائم لأنشطتنا الأكاديمية

ومؤتمراتنا جعلتنا على قرب منه.

كان مستعداً لدفع ثمن مواقفه وقد فعل وحتى يكون حرّاً صادقاً مع نفسه فضَّل أن يعيش عيشة بسيطة في شقة مستأجرة على الرغم من أنه خريج الدكتوراه من جامعة كمبردج كان يملك مواهب وإمكانات وعلاقات واسعة تفتح له آفاقاً لا يحلم بها الكثيرون.

فضّل أنيس أن يقدم نموذج المثقف الحر الذي لو حذا حذوه مثقفونا المعاصرون لربما حدثت حركة نهضوية في أمتنا تقود الجماهير وتصوب مسيرة الأنظمة والحكومات بدلاً من العمل أذناباً للسلاطين ومحامين عن الشياطين.

كتب في فترة دراسته في الجامعة الأميركية ببيروت مقالات كثيرة في مجال تخصصه التاريخ نُشر معظمها في جريدة الحياة. في العلوم السياسية وفي التخصص الذي يحبه

وكان يُسلم مقالاته خُفية لئلا يعرف صاحب الجريدة كامل مروة عمره الحقيقي فيوقف النشر.

وكان مروة يظن أن الكاتب يحمل الدكتوراه، ولا يضع حرف الدال قبل اسمه تواضعاً منه، فكان مروة يتبرع بنفسه بإضافة هذا اللقب

كان كتاب لبنان الطائفي هو أول كتاب نشره سنة لأنه بحسب تعبيره هاله التعصب الطائفي في لبنان وأنا القادم من فلسطين لا نعرف للطائفية معنى .

ثمّ أصدر سوريا في الأدب المصري القديم .

وأجهد نفسه لسنوات في إعداد كتاب العلاقات السورية المصرية وقابل عبد الناصر وحصل منه على مقدمة للكتاب لكنه لم ينشره لأنه كان يريد أن يخرجه بشكل أفضل بعد عودته من الدراسة والتدريس في جامعة كمبردج وهو ما لم يحدث لانشغاله بأمور أخرى.

وفي أثناء وجوده في كمبردج اهتم بتاريخ العائلة الهاشمية فجمع ما استطاع من مصادر ووثائق حيث أصدر بعد عودته لبيروت كتاب الهاشميون وقضية فلسطين وكتاب الهاشميون والثورة العربية الكبرى وقد راجا رواجاً كبيراً غير أنهما منعا من النشر في الأردن والعراق وسوريا كما مُنع أنيس صايغ نفسه من دخول الأردن .

وعاد إلى بيروت مغموماً.

ليعلم بعد ذلك أن أول من ذكرها من العرب هو المحامي أنطوان كنعان فسافر أنيس باحثاً عنه في مصر حتى وجده فاعترف له أن رجلاً هندياً ركب معه في طائرة حدّثه عن شيء من هذا كان قد قرأ عنه.

فلا هو ولا جاره الهندي اطلعا عليها.

ولذلك فهي لم تثبت من الناحية العلمية.

وهكذا منع أنيس صايغ في كل ما يشرف عليه وعلى نشره ذكر هذه الوثيقة كحقيقة ثابتة.

يُعد المغربي محمد عابد الجابري من بين المفكرين العرب ذوي المشاريع النظرية الأكثر لفتا للانتباه واجتذابا للنقاش والجدل في اللحظة الراهنة .

ففي المدرسة الابتدائية وكذلك في السنتين اللتين قضيتهما في المدرسة الثانوية والكلام للجابري كان الأستاذ يملي علينا في نهاية الدرس تمارين الحساب والهندسة ثم الجبر كواجبات منزلية وكثيرا ما كنت أكتفي بالاستماع إليه وكتابة المعطيات الرقمية بدل كتابة نصوص التمارين بأكملها حتى إذا انتهى من الإملاء كنت قد هيأت الجواب ! وفي المنزل كنت أشتغل على تمارين مماثلة في الكتب الفرنسية ! وكانت تستغرق جل أوقات الفراغ عندي .

بهذا العشق لعلم الخوارزمي سافر الجابري إلى سوريا سنة ليتم دراسته في تخصص الرياضيات وقد كان عمره آنذاك واحدا وعشرين عاما بيد أن أمرا طريفا كان سيرغمه على تبديل اختياره التعليمي فقبل أن يسجل نفسه في كلية العلوم أخذ الكتب المدرسية السورية لاستطلاعها والمقارنة بينها وبين ما تعلمه في المغرب .

ويقول المفكر المغربي شعرت أنه سيكون علي أن أقوم بدور المترجم لنفسي من العربية إلى الفرنسية ومنها إلى اصطلاحنا في المغرب . فكان هذا شيئا مثبطا تماما .

وقرر الجابري ترك الرياضيات والتخصصات العلمية على حد سواء والارتحال هاربا من استغلاق الرموز العلمية التي تحفل بها كتب الرياضيات والفيزياء إلى حقل معرفي آخر هو القانون ولكنه بعد أن راجع كتب القانون لاحظ أنها تعتمد بالأساس على الحفظ والذاكرة .

وقرر الطالب التغيير مرة أخرى فسجل نفسه في كلية الآداب بالسنة الأولى في تخصص كان يسمى الثقافة العامة على أن يختار بالسنة الثانية شعبة الفلسفة ودرس الجابري السنة الأولى في دمشق ونجح بتفوق فكان ترتيبه السادس من بين ما ينيف على خمسمائة طالب .

وإذا كانت مجرد مسألة شكلية تتعلق بالرمز الرياضي قلبت دفة حياة الجابري الطالب من الرياضيات إلى الفلسفة فإن ابن خلدون كان الدافع إلى توجيه الدفة مرة أخرى ولكنها هذه المرة داخل التخصص نفسه أي الفلسفة إذ جذب الاهتمام بالمقدمة الخلدونية تفكير الجابري إلى دراسة التراث .

وعلاقة الجابري بابن خلدون تعود إلى الباكالوريا الثانوية العامة حين يحكي عن نفسه قائلا في سنة كان قد مضى على استقلال المغرب نحو سنة ونصف السنة وكنت آنذاك أهيئ البكالوريا .

وبعد عودته من دمشق وبعد حصوله على الإجازة في الفلسفة وتحضيره للتسجيل في دبلوم الدراسات العليا سيجد الجابري نفسه من جديد أمام ابن خلدون فيقول ألح علي

الأستاذ المشرف الدكتور محمد عزيز الحبابي أن يكون موضوع بحثي آراء ابن خلدون في كتابة التاريخ والنظر في الكتابات التاريخية المغربية المعاصرة في ذلك الوقت إن كانت قد استفادت من نقد ابن خلدون للمؤرخين .

يقول الجابري أذكر أنني حين كنت أكتب أو أقرأ عن ابن خلدون كنت منشغلا بهاجس أساسي وهو أن التأويلات المعاصرة والمتعددة لابن خلدون تخفيه عنا. لذلك قررت أن أنسى كل ما كتب عن صاحب المقدمة. وأن لا أتخذ لي مرجعا آخر غير نصوص ابن خلدون نفسه . هكذا قررت أن أكتب عن ابن خلدون وكأن أحدا لم يكتب عنه قبلي .

هذا المنهج في القراءة المرتكز على العودة إلى النص وتخطي مختلف التأويلات المعاصرة هو ما سيطبقه الجابري في دراساته اللاحقة حين شارك في العراق سنة في ندوة عن الفارابي متعاملا مع هذا الفيلسوف بالطريقة نفسها أي أن لا ينشغل بما كتب عنه .

ثم كانت دراسات أخرى واحدة عن ابن رشد وأخرى عن ابن سينا بالمنظور المنهجي نفسه أي العودة إلى المتون والنصوص ذاتها وصرف النظر عن قراءات الآخرين لها.

وعلى ضوء ذلك فإن مشروع الجابري في قراءة التراث لم يكن من أجل مطلب يتعلق بفهم الماضي حصرا بل كان أساسا من أجل فهم الحاضر فهو نفسه في بداية كتابه يقول يتناول هذا الكتاب موضوعا كان يجب أن ينطلق القول فيه منذ مائة سنة. إن نقد العقل جزء أساسي وأولي من كل مشروع للنهضة.

Topic :Religion

طارق أوبرو هو اليوم إمام مسجد الهدى بمدينة بوردو الفرنسية من أبرز الأئمة والمرشدين الفرنسيين المسلمين ينحدر من المغرب لأبوين اشتغلا بالتعليم بمدينة تارودانت المجاورة لأغادير جنوبي المغرب سافر إلى فرنسا لمتابعة دراسة الطب وهو على مشارف العشرين من العمر لكن القدر كان يعده لأمر آخر كما يحكي عن نفسه إذ انتابته رغبة شديدة في التدين دون تأثير من أحد أو من جماعة فقال غمرتني موجة

عارمة من الإيمان والانفتاح الروحي لم أفهمها إطلاقا ووجدت نفسي متشبثا بسلوك ديني قوي والتزام أخلاقي شديد.

. ولم يكن لذلك أي علاقة بما كان يحدث للشبان الآخرين في مثل سني

وليس طارق أوبرو خريجا لجامعة من الجامعات الدينية العتيقة أو العصرية ولكنه إمام عصامي دأب على تكوين نفسه بنفسه وعين أول الأمر إماما للصلاة بمسجد الهدى بمدينة بوردوالتابع لجمعية مسلمي الجيروند الموالية لاتحاد المنظمات الإسلامية بفرنسا المؤيدة للتنظيم الدولي للإخوان المسلمين.

كما جاء على لسان نيكولا ساركوزي في الوزارة نفسها بعد عشر سنوات كفى من إنزال الأئمة القادمين من الخارج.

ويعتبر طارق أوبرو واحدا من نسبة من الأئمة ذوي الجنسية الفرنسية ومن بين حوالي ألف إمام بفرنسا منهم مغاربة وجزائريون وأتراك و تونسيون و من البلدان الأفريقية جنوب الصحراء والمشرق العربي حسب تحقيق وزارة الداخلية الفرنسية المنشور في جريدة لوموند يوم الفاتح من يوليوتموز تحت عنوان تكوين الأئمة تحد جديد للإسلام الفرنسي.

وباستثناء مسجد باريس الذي تديره الجزائر فإن باقي المساجد بالمدن الفرنسية تديرها جمعيات مدنية إسلامية مختلفة.

ويوضح الإمام أنه يجتهد في الملاءمة بين الإسلام والعلمانية الفرنسية، ليعيش المسلم في تناغم مع نفسه ومجتمعه دون أن يعني ذلك التخلي عن جوهر دينه.

ويصف نظريته قائلا إنها تقدم مجموعة من الوسائل التي تتيح متابعة التطور الحاصل في المجتمع الفرنسي وتقترح أدوات منهجية مستنبطة من أصول الفقه لكنها تأخذ بعين الاعتبار الإبيستيمولوجيا العالمية المعاصرة أضفت إليها بعض المفاهيم التي ابتكرتها شخصيا

أعمال التيار الأول تشكلت حركة النسوية الإسلامية الحقيقة التي تجعل الإسلام نفسه منطلقا ومرجعية أولية لها وتوجه عملها وفق منهج إعادة القراءة للنصوص الدينية وللتاريخ الإسلامي

حيث تمثل فاطمة المرنيسي عالمة الاجتماع المغربية أول من أبرزنسوية إسلامية قائمة على التأويل وإعادة قراءة النصوص والواقع التاريخي.

ففي أطروحاتها الأساسية، التي تميزت بجرأة المعالجة وبأصالة المنهج، حاولت البحث عن المسوغات والأسباب التي أدت إلى استبعاد النساء من المجال العام والسياسي، محاولة تطبيق منهج جديد يرتكز على أسس علمية ومعرفية، تتعلق بالانطلاق من مرجعية الإسلام نفسه في سياقه التاريخي والاجتماعي، ومراجعة جميع الأحاديث النبوية التي جاءت في شأن النساء.

ولذلك فإن هذا التمييز كان ضروريا ليضع الحدود الفاصلة، بين ما هو إلهي وما هو إنساني، ويحول دون تحميل الدين نفسه بعض الأوضاع التي لا تطاق.

يعرض الكتاب مفاهيم الحرية في الفكر الإسلامي التراثي، ويبحث موقف الفقهاء والفلاسفة والمتصوفة المسلمين من قضايا الحرية والرق والسخرة والسجن، ويقارن بين المفاهيم الإسلامية للحرية والرؤية الفلسفية اليونانية التي اشتغل بها المسلمون فترة طويلة واستوعبها في العلوم والأفكار الإسلامية.

يعود المؤلف في معالجة مفهوم الحرية إلى عصور ما قبل التاريخ حتى العصر الحاضر وفي النطاق القانوني، ويلاحظ أن مفهوم الحرية هو أهم محرك تاريخي عرفه العالم، وأن "الحرية" استطاعت أن تعتق نفسها من إطار قيود التعريفات، وأن تتطور إلى مصطلح ليس له وجود خارجي يمكن تحديده إلا ما يعطيه لها العقل الإنساني.

ثم يصل الفيلسوف إلى القول إن الإنسان لا يفكر ولا يجيل رأيه في الشيء الممتنع إنما يفكر ويجيل رأيه في الشيء الممكن التي تخص بالفعل الإنساني وإذا كان الفعل مما ينظر فيه على طريق الإضافة أن يكون طاعة لمن تجب طاعته أو معونة لمن تجب

معونته أو غير ذلك من وجوه الإضافات الواجبة ثم امتنع من الفعل فهو ملوم غير معذور لأنه قادر ومتمكن ولأجل ذلك تلحقه الندامة، من نفسه والعقوبة من غيره أو العيب والذم وهذه الجهة التي تخص الإنسان من جهات الفعل المتعلقة بالفكر وإجالة الرأي المسمى بالاختيار هي ثمرة العقل ونتيجته ولولا هذه الجهة لما كان لوجود العقل فائدة بل يصير وجوده عبثا ولغوا ونحن نتيقن أن العقل أجل الموجودات أشرف ما من الله تعالى به ووهبه للإنسان.

والحرية الأخلاقية تعني رغبة الإنسان في أن يكون طيبا، وضبط الإنسان بنفسه هو شرط ضروري لسيطرته على غيره إن الحرية تعطي الإنسان العاقل القدرة على تحرير نفسه من قيود بيئته الطبيعية وعاداته الرتيبة وبذلك تجعل منه إنسانا حكيما.

وسئل من أحق الناس أن يؤتمن على تدبير المدينة؟ فقيل من كان في تدبير نفسه حسن المذهب فالحر النفس هو سيد لناموس الطبيعة ومن شروط العالم والمفكر الحقيقي أن يولد حرا وقد جعل أبقراط في وصيته الحرية بالمولد لطالب الطب أما إذا نظرنا إلى الحرية من الناحية السلبية فهي التحرر من عوامل الإرغام ومن أعباء الحياة اليومية أما العالم رشيد الدين بن خليفة فيرى أن الحرية هي الحياة الخيرة ويعرف الحرية بالسلب على أنها التحرر من الشر ومن العوامل التي تعيق الإنسان عن بلوغ الهدف الحقيقي لإنسانيته إن تجنب الشرور التي ترتكبها الكائنات البشرية عادة هو الحرية الحقيقية فعسير على الإنسان أن يكون حرا وهو ينصاع للأفعال القبيحة الجارية مجرى العادة .

وقد وصل الفكر السياسي المتعلق بالحرية إلى المسلمين بالطريقة نفسها التي وصلت بها كثير من الأعمال السياسية لأفلاطون وأرسطو التي ترجمت إلى العربية فقد فسر ابن رشد جمهورية أفلاطون بأن على كل إنسان أن يقتنع بأنه حر وأن شكل الدولة التي تمثل الحرية هو الديمقراطية ويبقى هنا على كل حال احتمال تحول الزيادة في الحرية لغير صالحها.

فالإنسان المنحط يستطيع أن يدعي لنفسه الحرية المطلقة ويعطي نفسه الحق في التخلي أو في رفض كل القيود الأخلاقية وبالمثل فإن دولة الحرية يمكن أن تفقد شخصيتها.

Economy:Topic

الواقع أن انتشار الرأسمالية المتشددة لم يكن مع انهيار الشيوعية وتفكك المعسكر الشرقي، إنما بلغ آنذاك محطة متقدمة جديدة، أما بداية انتقال رأسمالية آدم سميث إلى مرحلة التشدد أو عودتها إلى التشدد فكانت في أواخر السبعينيات ومطالع الثمانينيات من القرن الميلادي العشرين في عهد رونالد ريغان في الولايات المتحدة وتزامن ذلك مع انتشار أفكار المسيحية الصهيونية والمحافظين الجدد.

وقد شملت سياسته فيما شملت القضاء على بقايا قوة النقابات العمالية والتأمينات الصحية والاجتماعية ورفع بقايا القيود على رؤوس الأموال وتصعيد نفقات التسلح حتى عرف ذلك النهج الاقتصادي من بعده بالسياسة الريغانية .

آنذاك لم تكن جميع الدول الأوروبية الغربية في أوضاع تسمح باتباع النهج نفسه فكلما كانت الدولة الغربية أقرب جغرافيا إلى الحدود الفاصلة بين المعسكرين كانت أكثر حذرا في ممارسة الضغوط الرأسمالية على الطبقات الفقيرة لا سيما العمال خشية ازدياد انتشار الشيوعية غربا والتي كانت بغض النظر عن مساوئها متميزة بالضمانات الاجتماعية على حد أدنى شامل للسكان.

وقد كانت القاعدة الذهبية للرأسمالية منذ عهد آدم سميث تقوم على حرية رأس المال بمعنى إعفائه من كل ضابط قانوني ناهيك عن الأخلاقي في ميدان تشغيله وأن كل ما عدا ذلك كالأسعار.

أي تكاليف الحياة على العامة أو كسوق اليد العاملة أي طلب الرزق بالجهد البشري أو كالتطور التقني والعلمي.

أي إيجاد خدمات ومنتجات جديدة.

جميع ذلك يتحقق أو يفترض تحقيقه من خلال التنافس بين مالكي رؤوس الأموال وسعي كل منهم لتحقيق كسب مادي لنفسه أكثر من الآخر وهذا ما يعنيه التعبير الشائع في الغرب السوق تنظم نفسها بنفسها.

افتقاد الضوابط كان من وراء تصرف عدد محدود من المضاربين الماليين عبر ثروات مالية كبرى بمصائر بعض الدول كما كان مع جنوب شرق آسيا قبل سنوات وكما كادت تتعرض لمثيله آنذاك شبكة العلاقات المالية في أوروبا الغربية نفسها أثناء بحث الولايات المتحدة عن سبب بديل لاستمرار الهيمنة على القارة بعد سقوط السبب الرئيسي في الحرب الباردة أي ما كان يرمز إليه تعبير المظلة النووية الواقية وكانت الأزمة النقدية الأوروبية آنذاك من أسباب التعجيل في إنشاء منطقة اليورو الموحدة، بديلا عما كان يسمى نظام الأفعى المالية بمعنى ربط أسعار أهم العملات الأوروبية بعضها ببعض ارتفاعا وانخفاضا في نطاق نسبة مئوية متدنية

إن الخطة الموضوعة وما سبقها من خطوات مبدئية ليست ابتكارا جديدا فكثيرا ما جرى إنقاذ مصرف مالي أو مؤسسة أو شركة كبرى بالأسلوب نفسه إنما لم يكن ذلك في يوم من الأيام بحجم ما بلغه الآن دفعة واحدة.

تقول الخطط المعنية بوضوح ما محوره قيام الدولة بتخليص المصارف المالية من الصفقات والعقود الخاسرة وترك المضمونة الرابحة منها للمصارف نفسها ويعني هذا واقعيا

تعني كلمة تُعفى هنا إعفاء أصحاب الثروات المالية الحقيقيين من المحاسبة أيضا فهؤلاء لا تحاول الدولة أصلا تحميلهم المسؤولية وبالتالي لا تصل إليهم أيدي المحاسبة الجزئية التي تصل إلى مدراء الأعمال التنفيذيين أي الموظفين واقعيا ممن يخسرون أمكنة عملهم فيعين أصحاب الثروات سواهم سواء كان ذلك في المنشآت المالية نفسها بعد إنقاذها أو من خلال إقامة بدائل عنها.

الدولة تقترض من أصحاب رؤوس الأموال أنفسهم

لقد كان وما يزال أهم عنصر في الرأسمالية المتشددة وغير المتشددة هو تخفيف الأعباء على الشركات والمصارف المالية، لأن الفكر الرأسمالي نفسه يقوم على أن الاقتصاد هو المعاملات الجارية بين المصارف والشركات الكبرى فإن حقق مالكوها الرأسماليون أرباحا متزايدة أقدموا على مشاريع جديدة يفترض أن توجد أماكن عمل جديدة لتخفف العبء المعيشي عن الطبقة المتوسطة فتتمكن من تسديد ما عليها.

مثل القروض العقارية والرسوم الضرائبية المتزايدة وبالتالي يستمر وجودها عصبا لاستمرار حركة الإنتاج والاستهلاك في الدولة.

وجميع ذلك لم يؤثر على انتشار السلعنة في ظل الرأسمالية على حد تعبير عبد الوهاب المسيري رحمه الله مترجما لظاهرة يكتب عنها بعض مفكري الغرب وفلاسفته فباتت صناعة السلع والخدمات وترويجها وتسويقها هدفا بحد ذاته أو جزءا من هدف تحقيق مزيد من العائدات بأي وسيلة وأي ثمن بما في ذلك الحروب وبالتالي لتحقيق مزيد من الهيمنة المالية، دون وضع حقيقة الاحتياجات البشرية بعين الاعتبار ناهيك عن أي درجة من الحرص على عدالة اجتماعية أو مادية.

بل حتى أصبحت قيمة الإنسان نفسه مرتبطة بسلعنته بمعنى تصويره سلعة والتعامل معه على هذا الأساس

عندما سارت الدول العربية في طريق الاستدانة الوعر كانت تظن نفسها قادرة على تحقيق معادلة صعبة طرفها الأول هو الحصول على الديون واستغلالها في برامج التنمية المختلفة وطرفها الثاني هو سداد هذه الديون وفوائدها. لكن بعد مرور سنوات طويلة على السير في هذه الاتجاه وجدت نفسها في حيرة.. فلا هي حققت التنمية المطلوبة ولا هي أصبحت قادرة على سداد ديونها الخارجية أو الداخلية حتى أصبح مجموع هذه الديون مجتمعة 560 مليار دولار يدفع للقسم الخارجي منها فقط كل عام 40 مليارا.

وأمام العجز عن سداد الديون واستجابة لضغوط المؤسسات الدولية مثل البنك الدولي

وصندوق النقد الدولي لجأت الدول العربية إلى مزيد من الاستدانة أو إعادة جدولة ديونها وفقا لشروط الدائنين الجدد في نادي باريس، مما يثير تساؤلات عن مدى الحاجة إلى اللجوء لمثل هذه الإجراءات.

لكن التساؤل هنا.. هل كانت هذه الاختلالات الهيكلية في الاقتصاديات العربية سببا للاستدانة أم نتيجة لها؟ على أية حال وسواء أكانت سببا أم نتيجة فإن الدول العربية المدينة وجدت نفسها مضطرة إلى الرضوخ لـ شروط المؤسسات المالية المانحة سواء الدولية منها كـ صندوق النقد الدولي والبنك الدولي اللذان يلزمان الدول المدينة باتباع سياسات اقتصادية واجتماعية معينة. أو المؤسسات المالية العربية التي لم تكتف بإخضاع الدول العربية المقترضة للشروط نفسها التي يطلبها صندوق النقد والبنك الدوليين فحسب وإنما أضافت إليها شروطها الخاصة والتي تحرص فيها عادة على ألا تتخذ الدول المقترضة مواقف سياسية تتعارض مع سياسات الدول الدائنة مما خلق معايير مختلفة في التعامل مع الدول العربية المقترضة كما حدث مع مصر وسوريا والأردن بعد حرب الخليج الثانية.

Social Affairs:Topic

أشارت إحصائية مصرية حديثة صادرة عن مركز الدراسات والبحوث الاجتماعية والجنائية إلى أن هناك قرابة ألف حالة زواج مسجلة في سجلات مأذوني مصر وافق فيها الزوج على أن تكون العصمة بيد الزوجة وهو ما يعني أن من حقها تطليق الزوج شرعا.

كما أشارت إلى تزايد نسبي في هذا النوع من الزواج بالمقارنة مع عقود سابقة ووجود حالة زواج من هذا النوع بين كل ثمانين حالة زواج تقريبا تزيد في بعض الأحيان إلى حالة زواج واحدة بين كل حالة يوافق فيها الزوج على إعطاء المرأة حق تطليق نفسها عن طريق التنازل لها عن العصمة وذلك في بعض المناطق.

ويعتقد هؤلاء الخبراء أن السبب وراء تزايد حالات الزواج بعصمة الزوجة هو رغبة المرأة المتزوجة في تأمين نفسها كي لا يتزوج عليها زوجها أو في حالة المرأة سيدة الأعمال التي تتزوج أحد مساعديها أو العاملين عندها أو في حالات التخوف من ماضي الزوج وربما رغبته في السيطرة على أموالها.

الجدير بالذكر أن بعض المذاهب الإسلامية تجيز إعطاء الزوج حق العصمة إلى الزوجة بشرط النص على ذلك في عقد النكاح باعتبار أن العقد شريعة المتعاقدين أو باتفاق لاحق بحيث إذا فوضها أو وكلها واختارت الطلاق أي أرادت أن تطلق نفسها فينبغي أن تطلق نفسها ولا يجوز أن تطلق زوجها كأن تقول له أنت طالق .

وقد جاء في فتاوى المجلس الأوروبي للبحوث والإفتاء أن المجلس قرر بعد بحث مستفيض في هذه المسألة أنه يمكن أن تطلق المرأة نفسها إذا اشترطت ذلك في عقد الزواج أو إذا فوضها زوجها بذلك بعد العقد .

يذكر أن سجلات الزواج في مصر تشير أيضا إلى تزايد معدلات الطلاق التي وصلت إلى .

لا شك في أن الوعي والإدراك والاقتناع تعد المقومات الأساسية التي تدفع أي جماعة إلى الحركة في اتجاه أي قضية من قضاياها، فإذا توفرت كل هذه المقومات لدى غالبية أفراد الجماعة ولم تنتج عن ذلك محاولات القيام بأفعال جماعية، فإن هذا على الأغلب يعني أن هناك عطلا في الحركة الذاتية لهذه الجماعة بسبب غياب روح الفعل الجماعي لدى غالبية أفرادها، وبالتالي فهي تعتمد فقط على العوامل والأطراف الخارجية لتحريك واقعها.

إن تعطل الحركة السياسية للجماعة في اتجاه طموحاتها لا يعني الانعدام الكلي للحركة داخلها، فهناك دائما أقلية تملك روح الاستعداد للتضحية والمخاطرة، وذلك يدفعها إلى الفعل بصرف النظر عن طبيعة هذا الفعل وكيفيته في محاولة منها لإنابة نفسها عن الجماعة المستقيلة من دورها.

فالأقلية في المجتمعات العربية تعتبر نفسها مضطلعة بالدفاع عن جماعة تفتقد القدرة على الدفاع عن مطالبها ومصالحها في مواجهة السلطة والغرب معا.

إن السلطة العربية التي لا تستجيب لرغبة ومطالب أغلبية المجتمع، إلى جانب الهيمنة الأجنبية التي لا تتوقف عند السيطرة على المقدرات، بل تتعداها إلى الغزو والاحتلال والتدخل في مختلف التفاعلات داخل الجماعة، كل ذلك كان من الممكن أن يتولد عنه فعل جماعي من قبل الأغلبية، يؤدي على الأقل إلى توازن المصالح أي تنازلات من قبل السلطة والغرب، كما ينتج عنه تثبيت الحدود والخطوط الحمراء التي لا ينبغي تجاوزها في اتجاه ما تعتبره الغالبية مقدسا ومحرما وحيويا.

ونظرا لغياب ردة فعل من الأغلبية التي يفتقد أفرادها روح الفعل الجماعي فإن الأقلية هي التي تقوم بردة الفعل معتبرة نفسها نائبة عن الجماعة في التعبير عن المطالب والقضايا التي تكاد تجمع حولها الأغلبية الصامتة والعاجزة.

ومن هنا فإن هذه الأقلية لا تعتبر نفسها أقلية من حيث اعتقادها ورؤيتها للمصالح والمطالب والطموحات التي تتبناها، بل هي أقلية فقط من حيث عدد الأفراد القادرين على القيام بردة الفعل للدفاع عن قضايا تقاسمهم الأغلبية العظمى من الجماعة الإيمان بها، وبالتالي فإن هذا الاتفاق في النظرة بينها وبين الأغلبية حيال تلك القضايا تعتبره بمثابة تفويض ضمني.

لكن هذه الأقلية التي أنابت نفسها عن الأغلبية تجد نفسها أمام ضرورة الإجابة عن سؤال جوهري يتعلق بطبيعة وحجم الفعل الذي ينبغي أن تقوم به للتعويض عن غياب فعل الجماعة الذي كان من المفترض أن تقوم به الأغلبية بكل ما تمثله من قوة التكتل والحجم والزخم والتواصل.

بعد أن تقوم الأقلية بإنابة نفسها عن الأغلبية المستقيلة من دورها تقوم بتحديد نوع وطبيعة الفعل الذي تعتقد أنه يسد الفراغ الذي تركه غياب فعل الأغلبية في بعده السياسي والاجتماعي.

Appendix 3

Some samples of results :

Correct ones with no exceptions

11.buck.txt

Sentence: 2

. wmn Al|n fSAEdA ynbgy ElY EmAl Al<gAvp wAlSHAfyyn tsjyl >nfshm ldY AlslTAt AlAndwnysyp fy bAndA |t$yh EASmp w<ETA' <xTAr msbq En >y xTT llsfr xArj Almdyntyn Alr}ysytyn fy Al<qlym

Nafs form: >nfshm

Referent: wAlSHAfyyn

Sentence: 6

. wqd >bdt bED Aldwl AlbArzp bAlnAdy dEmhA lxTp tjmyd gyr >n Als&Al Al*y yTrH nfsh ytElq bAl$rwT Alty stwDE ElY >y AtfAq mn h*A AlnwE

Nafs form: nfsh

Referent: Al*y

12.buck.txt

Sentence: 7

wrfD Alms&wl Al>mryky tSwr synAryw yzdAd fyh AlEnf bsbb <HsAs Alsnp b>nhm hm$wA fy fqAl fy AstjwAb fy nyrwby Alty HDr fyhA twqyE AtfAq AlslAm AlswdAny ywm mn yhm$ AlmslHwn >nfshm mn yfEl . wlm nstbq snrY ywm ynAyr <n kAn Alsnp rADwn wlhm frSp . wlmA Tlb mnh tEryf AlnjAH fy qAl <nh AntxAb Hkwmp tmvl kl AlErAqyyn wt$kyl qwAt >mn qAdrp ElY HmAyp AlblAd mn AlmslHyn wmn AlqwAt Al>jnbyp

Nafs form: >nfshm

Referent: AlmslHwn

14.buck.txt

Sentence: 6

. wymDy AlkAtb qA}lA <n fkrp wSwl qyAdp $yEyp mntxbp <lY qmp AlslTp fy AlErAq tvyr mxAwf Al>nZmp AlErbyp swA' tlk Alty ywjd byn skAnhA $yEp >w lA . wTbqA llkAtb f<n AlEAhl Al>rdny kAn AlzEym AlErby AlwHyd Al*y Ebr En tlk AlmxAwf ElAnyp bynmA AHtfZ bhA Al|xrwn l>nfshm

Nafs form: l>nfshm

Referent: Al|xrwn

Sentence: 7

. lknh ynql En mElqyn Erb qwlhm <n mA yxyf Al>nZmp AlErbyp Hqyqyp lys wSwl Al$yEp <lY AlHkm fy AlErAq w<nmA AldymqrATyp nfshA Alty ymkn >n tnt$r <lY Aldwl wAl$Ewb AlErbyp AlmjAwrp llErAq

Nafs form: nfshA

Referent: AldymqrATyp

15.buck.txt

Sentence: 12

<*A kAnt <srA}yl Alty tEAdynA fk>nmA nryd >n nhAjm swryA lkn lA nryd Alswryyn >n yHmwA >nfshm

Nafs form: >nfshm

Referent: yHmwA

- Correct (exception as a conjunction)

1-34.buck.txt

Sentence: 7

. w>wDHt AlbyAnAt Alty k$f EnhA qAnwn Hryp AlmElwmAt >n nZAm tjnyd >frAd Aljy$ AlbryTAny Zlt Alsryp AltAmp tktnfh HtY En wzrA' wms&wly AlHkwmp >nfshm

Nafs form: >nfshm

Referent: wms&wly

2-56.buck.txt

Sentence: 3

. w<*A kAn AlmwATn AlEAdy yErb En An$gAlh mn hymnp Allwn AlAHmr ElY AlAjwA' AlAntxAbyp wAlHzbyp wAlsyAsyp fy twns f<n Alnxb wqAdp >HzAb AlmEArDp >nfshm yqrwn bAlxll AlwADH fy myzAn AlqwY AlsyAsy wAlHzby fy twns HAlyA lSAlH AlHzb Al*y yntmy Alyh >glb kwAdrAldwlp

Nafs form: >nfshm

Referent: wqAdp

3-80.buck.txt

Sentence: 5

. yqwl >nA >tnAwl TEAmy b$kl EAdy vlAv wjbAt EAdyp wlA >tnAwl h*h AlbrwtynAt Almrkzp <lA fy AlmEskrAt wmA >ql h*h kmA >nny >tdrb bSHbp mdrb ErAqy >$rf ElY tdryby mn* >n knt . wlm y$Ark mHmd Ebd AlmnEm Ely <lA fy mEskr xArjy wlmdp >sbwEyn fy swryA qbl dwrp Al>lEAb AlErbyp wkAn qblhA qd AnDm

<lY mEskr tdryby fy lknh qTE qbl whw sEyd bh*h AlmEskrAt Alty ytElm fyhA Alkvyr kmA ElY Alrgm mn >n AlrbAEyn Alswryyn wAlmSryyn >nfshm AstEdwA lldwrp AlErbyp nfshA bmEskrAt tdrybyp fy Almjr wblgAryA Astmrt fy bED Al>HyAn <lY >rbEp >$hr

Nafs form: >nfshm

Referent: Alswryyn

4-139.buck.txt

Sentence: 9

. AntHAry yqtl xmsp bynhm DAbTAn bArzAn bAl$rTp wTflp fy AlHAdyp fy hjwm ElY mbnY Hkwmy wmjmwEp <slAmyp mt$ddp tTlq ElY nfshA ktA}b AlHrmyn tEln Alms&wlyp En Alhjwm

Nafs form: nfshA

Referent: wmjmwEp

5-184.buck.txt

Sentence: 23

. wkAn AlHAkm Al>mryky fy AlErAq bwl brymr qd SrH AlAvnyn b>n >tbAE AlSdr wDEwA bAlfEl >nfshm xArj nTAq gyr >n AlSdr rd ElY *lk bAlqwl <nh bAEtbAr AlwlAyAt AlmtHdp lh xArjA En AlqAnwn

Nafs form: >nfshm

Referent: wDEwA

6-228.buck.txt

Sentence: 13

.<lA~a >n mn yqrr fy nhAyp AlmTAf mA yun$r ElY AlmwqE hw fryq mn AlxbrA' Al*yn yqwmwn bmrAjEp wtqyym bAl<DAfp <lY mtTwEyn mn wsA}l <ElAm kbrY wr}ysyp fy wSHfyyn wmwZfy wykylyks >nfshm

Nafs form: >nfshm

Referent: wmwZfy

7-230.buck.txt

Sentence: 2

.wqAl AyhAb AlHsyn AlnATq bAsm wzyr AldAxlyp fy Hkwmp HmAs AlmqAlp An Al$rTp Atx*t h*A AlqrAr lAnh lA ytmA$Y mE AlEAdAt wAltqAlyd .wqAl bED mAlky wmdyry AlmqAhy Almnt$rp ElY $AT} gzp lwkAlp AlAnbA' Alfrnsyp Anhm fwj}wA xlAl AlAyAm AlAxyrp bqrAr Al$rTp mnEhm mn tqdym Al$y$p wbEd AtDH An h*A AlmnE hw ElY tqdym Al$y$p llnsA' .wqAl Abw AHmd Al*y ymlk mqhY ElY Al$AT} An AljmyE y&ydwn mnE tqdym Al$y$p llqASryn wlkn lA yjb mnE AlnsA' mn wbxASp A*A kn ydxn fy AldAxl wlys fy .AmA n$AT whw mAlk AHd AlnwAdy AlbHryp fy gzp fqd qAl Anh Astmr btqdym Al$y$p wlknh xsr bAlm}p mn zbA}nh bsbb qrAr Al$rTp mnE tqdym Al$y$p .mn qAl DAbT fy $rTp gzp lwkAlp AlAnbA' Alfrnsyp An mA HSl EndmA mnE bED rjAl Al$rTp ASHAb AlmqAhy mn tqdym Al$y$p b$kl kAml kAn bmvAbp sw' tfAhm HtY AwDHt lhm AlslTAt AlmEnyp An AlAmr ytElq bAlnsA' .wfy bED rdwd AlfEl Al$Ebyp ElY tqwl snA' why TAlbp fy AljAmEp wrbp mnzl wAm lTflyn tEtbr nfshA gyr lknhA tltzm bAlqlyl mn tEAlym AnhA Dd Al$y$yp wtdxynhA swA' llftAp >w llrjl fy >y mkAn

Nafs form: nfshA

Referent: wrbp

8-233.buck.txt

Sentence: 9

.fy gDwn qAlt jmAEAt Hqwq Al<nsAn <n <dAnp AlqwSy lA tDfy b>y HAl mn Al>HwAl $rEyp ElY mHkmp jwAntnAmw Alty twAjh $kwkA wtHdyA mn jAnb jmAEAt Hqwq Al<nsAn wAlmHAmyn Almdnyyn wAlmEtqlyn >nfshm

- Correct with Number (tamez)

1-357.buck.txt

Sentence: 21

. fElY sbyl qAm AvnAn mn byn kl vlAvp nAxbyn fy flwrydA btsjyl >nfshm ElY >nhm dymqrATyyn

Nafs form: >nfshm

Referent: nAxbyn (number)

2-367.buck.txt

Sentence: 4

. w*krt wkAlp AnbA' $ynxwA AlSynyp >n AlhjmAt wqEt qbyl Alfjr fy bldp kwjA jnwby $ynjyAnj wbd>t btfyjr qnblp mHlyp AlSnE wbEd *lk fjr >rbEp AntHAryyn >nfshm msthdfyn mkAtb Hkwmyp

Nafs form: >nfshm

Referent: AntHAryyn

3-548.buck.txt

Sentence: 20

. qAm sbEp >$xAS fy qryp sAn bydrw kwtwd b$mAl Alflbyn bdq >nfshm bAlmsAmyr <lY SlbAn fy tqlyd snwy tEbyrA En . wfymA tErb Alknysp En AstyA}hA tjAh tlk <lA >nhA tjt*b Alkvyryn lm$AhdthA

Nafs form: >nfshm

Referent: >$xAS

4-809.buck.txt

Sentence: 18

. wkAnt mHAwlAt Altfjyr Alty $hdthA lndn qd jA't bEd >sbwEyn mn <qdAm >rbEp AntHAryyn ElY AlqyAm btfjyr >nfshm fy wsA}l Alnql AlEmwmy fy AlEASmp AlbryTAnyp mmA >sfr En mqtl $xSA

Nafs form: >nfshm

Referent: AntHAryyn

5-810.buck.txt

Sentence: 4

. wkAnt tfjyrAt lndn Alty wqEt fy ywlyw tmwz AlmADy qd >wdt bHyAp $xSA bmn fyhm Almfjrwn Al>rbEp >nfshm w>Syb $xS bjrAH

Nafs form: >nfshm

- Eroor case of plural (total number 7)

63.buck.txt

Sentence: 5

. w>$Ar AtHAd AlSlyb Al>Hmr Aldwly <lY >n t$jyE AlmjtmEAt Almnkwbp ElY AlqyAm bmbAdrAt l<EAnp >nfshm >vnA' AlkwArv >w bEdhA ymvl EnSrA >sAsyA fy Altxfyf mn wT>p AlkwArv

Nafs form: >nfshm

Referent: Almnkwbyn

504.buck.txt

Sentence: 6

. w$hd mxym nhr AlbArd lylp AljmEp A$tbAkAt mtqTEp bAl>slHp Alxfyfp fy AlmnATq nfshA

Nafs form: nfshA

Referent: bAl>slHp

601.buck.txt

Sentence: 15

. wy$dd AlEb~Ar ElY lA twjd hnAk HmAyp mA}p fy . wyrdf hy TbEA klhA ttx*hA swA' Alm&ssAt >w Al>frAd lHmAyp >nfshm mn AljrA}m . whw yrY bArqp >ml fy kwn AlwEy bjrA}m tqnyp AlmElwmAt y$hd mtzAyd fy mnTqp Al$rq . hnAk <HSA}yp fy h*A AlmwDwE tuZhir >n AlwlAyAt AlmtHdp tEtbr Alrqm EAlmyA fy Al<nfAq ElY AlHmAyp wylyhA mnTqp Al$rq . wfy Al$rq Al<mArAt hy mn >kvr Aldwl Alty tnfq swA' km&ssAt EAmp >w $rkAt xASp lHmAyp >nfshA mn AljrA}m Al<lktrwnyp >w mn AlhjmAt . wyErb AlEb~Ar En AEtqAdh b>n hnAk sbAqA byn Almjrmyn wbyn HmAp fy mjAl AljrA}m AlmElwmAtyp

Nafs form: >nfshm

Referent: mjrmy

604.buck.txt

Sentence: 7

. wkAnt qd sbqt jlsAt Alt$Awr AHtqAnAt syAsyp <* >Eln Hzb Allh wEwn AstEdAdhmA llnzwl <lY Al$ArE fy HAl Edm tlbyp whw mA rd~ Elyh fryq Al>kvryp bAlt>kyd >n Al$ArE syqAblh $ArE . wbynmA tuEtbr h*h AljlsAt bmvAbp AlfrSp AlAxyrp >mAm AlHwAr tqwl mrAslp by by sy fy byrwt ndY Ebd AlSmd <n jlsAt Alt$Awr stkwn HAsmp fy tHdyd AtjAhAt Al>mwr fy AlmrHlp . yu*kr >n EddA mn jlsAt mA sum~y HwAr qd AnEqd byn AlqAdp >nfshm bhdf AlAtfAq ElY AlmsA}l AlxlAfyp fy wtwqft bfEl AlHrb Alty $nthA <srA}yl ElY w*lk qbl >n ynjH AlHwAr fy AltwSl <lY Hl l>kvr AlqDAyA Al$A}kp why mSyr Hzb Allh

Nafs form: >nfshm

Referent: EddA

-- Error(too many candidates)

77.buck.txt

Sentence: 2

. ftHt EnwAn mjzrp lA qAlt AltAymz <n qtl TflA ErAqyA ElY wt$wyh Alkvyr gyrhm bfEl syArtyn mlgmtyn >vnA' tjmE llAHtfAl btd$yn wHdp jdydp llSrf AlSHy bbgdAd yEd >b$E Al>fEAl Albrbryp mn* bdAyp Hrkp Altmrd . wqAlt AlSHyfp <n AlHzn wAly>s Al*y y$Er bh >qArb AlDHAyA yEbr En Al<HbAT wAlgDb AlEAm mn AlEnf Al*y >sfr h*A Al$hr wHdh En mqtl ErAqyA ElY Al>ql whw AlgDb Al*y Atjh >HyAnA <lY qwAt AltHAlf gyr >nh ynbgy >n ytHwl <lY wAlkvyrwn mnhm mn xArj Al*yn thdf >fEAlhm <lY >HdAv >kbr qdr mmkn mn sfk AldmA' wAldmAr . w>DAft <n hdf >bw mSEb AlzrqAwy Al<rhAby Al>rdny Al*y yxTT >glb AltfjyrAt wEmlyAt AlxTf wAl*y qtl bnfsh Edp rhA}n grbyyn hw jEl AlErAq gyr qAbl w<yqAEh fy dA}rp lA nhAyp lhA mn Almwt wAxtTAf AlrhA}n wAlHylwlp dwn <jrA' AlAntxAbAt Almqrrp fy ynAyr/kAnwn . w>mA SHyfp Al<ndbndnt fqd Hmlt fy tgTythA Aldwlyp Swrp lsyArp tbdw |vAr AldmAr wElY jAnb AlSwrp Tfl yDE ydh ElY Zhr Tfl >Sgr wqd bdt ElY AlSgyr ElAmAt wqr> AltElyq >sfl AlSwrp Alm$hd bEd gArp >mrykyp ElY mdynp AlSdr >fqr >HyA' bgdAd wAlty >sfrt En qtl vmAnyp ErAqyyn ElY . wqAlt AlSHyfp swyt mnAzl bAlArD wA$tElt AlnyrAn fy E$rAt AlsyArAt xlAl Emlyp wqd AHtmY skAn sAmrA' Alty qTEt AlqwAt Al>mrykyp wqwAt AlHkwmp AlErAqyp AlkhrbA' wAlmyAh wqAlt <n Alkvyryn >SybwA fy tbAdl AlnyrAn

Nafs form: bnfsh

Referent: AlxTf

83.buck.txt

Sentence: 9

. wqrrt AlmHkmp >ms Alxmys t>jyl jlsthA b$kl mfAj} bEdmA Trd bhlwl mHAmyh mTAlbA bmnHh AlHq fy AldfAE En nfsh

Nafs form: nfsh

Referent: AldfAE

89.buck.txt

Sentence: 5

. wqd >Sr mylw$yfyt$ ElY AldfAE En nfsh bnfsh fy mHkmp yEtbrhA gyr qAnwnyp

Nafs form: nfsh

Referent: AldfAE

Sentence: 5

. wqd >Sr mylw$yfyt$ ElY AldfAE En nfsh bnfsh fy mHkmp yEtbrhA gyr qAnwnyp

Nafs form: bnfsh

Referent: AldfAE

105.buck.txt

Sentence: 21

. wqAlt fyky <*A knt fy HmAyp qwAt f<n *lk ySnfk ElY >Hd jAnby AlSrAE nHn >TbA' mHAydwn wASTHAbnA lHrs mslHyn lA ygyr h*A . wHtY AlmnZmAt Al<nsAnyp AltAbEp ll>mm Alty ysyr EAmlwhA fy >glb Al>HyAn tHt HmAyp qwAt Al>mm lA t>mn ElY nfshA bEd slslp AlhjmAt Altfjyryp Alty Asthdft mqrAt Al>mm AlmtHdp fy bgdAd fy >gsTs/ |b w>ktwbr/ t$ryn Al>wl fy AlEAm AlmADy

Nafs form: nfshA

Referent: Al>mm

114.buck.txt

Sentence: 3

. wtqwl AlSHyfp <n ElAwy xShA bmqAlp qbyl tslm AlslTp rsmyA lHkwmth mn Al<dArp Almdnyp Al>mrykyp Al>rbEA' wsEY fyhA <lY >n yn>Y bnfsh En AlzEymyn Al*yn yqdmAn AldEm lh whmA twny blyr r}ys AlwzrA' AlbryTAny wjwrj bw$ Alr}ys Al>mryky

Nafs form: bnfsh

Referent: Al>rbEA'

-Error as verb and noun look the same

394.buck.txt

Sentence: 8

. s>lth En AlEA}q AlHqyqy lslAm >jAb >Elm >n AsrA}yl jAhzp wlknny lst mt>kdA >n AlflsTynyyn jAhzwn fElyhm AqAmp nZAm Hkm dAxl AlHrkp AlflsTynyp ykwn lky ykwnwA qAdryn ElY Hkm >nfshm b>nfshm qbl >n yHSlwA ElY AstqlAlhm . Al$Eb Alyhwdy fy h*h AlArD Hkm nfsh qbl snwAt mn HSwlnA ElY AlAstqlAl wlwlA *lk mA knA lnnjH fy EAm vmAnyp w>rbEyn r&yth llHl AlAn ttrkz ElY AEtrAf mtbAdl bAlHqwq lHl AlAzmp wl>n AlwDE lys k*lk AlAn fhAlyfy ElynA AlEml llwSwl <lY tfAhm llHl Twyl AlAmd

wxlAl *lk ykwn llTrfyn >HlAm wmE Alwqt ttlA$Y AlAHlAm wnSl <lY AlwAqEyp h*A Hdv bAsrA}yl wllAsf *lk lm yHdv bEd End AljAnb AlErby w AlflsTyny

Nafs form: nfsh

Referent: Hkm

812.buck.txt

Sentence: 24

. wTAlb AlzEym Al<xwAny bAl>$rAf AlkAml llqDA' AlmSry w<lgA' AlqwAnyn AlAstvnA}yp wAl<frAj En AlmEtqlyn Al*yn AEd bEDhm nfsh llm$Arkp fy AntxAbAt mjls Al$Eb mvl ESAm w<lA f<n AlAntxAbAt stkwn ksAbqAthA >y An yktsHhA AlHzb AlwTny ysmH bwjwd $kly llmEArDp kmA hw AlwDE AlHAly Hyv ywjd EDwA mEArDA fqT bmjls Al$Eb mn mjmwE Akvr mn EDwA

Nafs form: nfsh

Referent: wAl<frAj

876.buck.txt

Sentence: 7

. lqy bwl msAEdp Edd D}yl mn qwAt Al>mm AlmtHdp ldY bd' EmlyAt vm wjd nfsh mrgmA ElY Alljw' <lY Alr$wp w>sAlyb ttsm bAlHylp fy bED Al>HyAn l<nqA* >frAd >srth fy AlbdAyp vm <nqA* >lf wmA}tyn wvmAnyp wstyn $xSA

Nafs form: nfsh

Referent: bd'

909.buck.txt

Sentence: 13

.wqAl jwnswn Al*y f$l fy EAm bAlHSwl ElY >SwAt kAfyp lAntxAbh lmnSb nA}b zEym AlHzb Al*y *hb <lY hAryt hArmAn AlqA}mp HAlyA b>EmAl zEym AlHzb <n Hzbh bHAjp <lY tjdyd nfsh bEd EAmA mn wjwdh fy AlslTp

Nafs form: nfsh

Referent: tjdyd