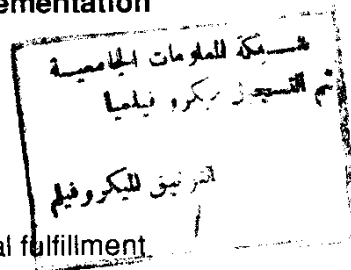


AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
COMPUTERS AND SYSTEMS ENGINEERING DEPARTMENT

**Computer Based Arabic Language Understanding
Analysis, Design And Implementation**

د. عبد الله



A Thesis submitted in partial fulfillment

of the requirements for the degree of

Master of Science

in Computer Engineering

By

Tarek Thabet

4559

Supervised by

Prof. Dr. Osman Badr Professor of Computer Engineering

Prof. Dr. Abd El-Monaim Wahdan Professor of Computer Engineering

1993

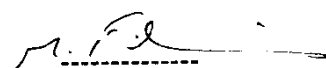


Examiners' Committee


Name, Title, Affiliation

Signature

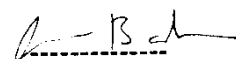
1- Prof. Dr. Magdy Fikry Mohammed Ragaie
Faculty of Engineering, Cairo University



2- Prof. Dr. Mohammed Adeeb Reyad Ghonaimy
Faculty of Engineering, Ain Shams University



3- Prof. Dr. Osman Abd El-Latif Badr
Faculty of Engineering, Ain Shams University






Statement

This dissertation is submitted to Ain Shams University for the degree of Masters of Science in Computers Science.

The work included in this thesis was carried out by the author in the department of Computers and Systems Engineering, Ain Shams University.

No part of this thesis has been submitted for a degree or a qualification at any university or institution.

Date: 31/1/93

Signature: 

Name: Tarek Thabet

Acknowledgement

I owe a special debt of gratitude to Prof. Dr. Osman A. Badr for his valuable guidance and unique encouragement. I would like to thank Dr. Osman Badr for his patience, the time spent with me and most of all, his technological vision. Without Dr. Osman Badr advice and supervision, this thesis would not have come into light.

ABSTRACT

This thesis gives all the needed fundamentals to develop a Computer Based Arabic Language Understanding System. This system could be used both as a user front end for expert systems, which was the original purpose for this thesis, as well as a stand-alone language analyser. The thesis also presents a complete Arabic lexicon model that incorporates morphology, syntax and semantic features. This lexicon has been designed to allow all kinds of syntax and semantic analysis.

The thesis presents an overview of all types of basic grammars that could be used for morphology and syntax analysis. The issue of augmentation and language added features to those grammars has been addressed with emphasis on added Arabic features.

A thorough study of the different parsing methods that could be used for the different kinds of grammars has been presented. Top-Down parsers, Bottom-Up parsers and mixed mode parsers were compared and used to analyse Arabic sentences.

A specific attention has been given to use Prolog as an Arabic language analyser. Prolog has been used to implement all different techniques, augmentations and parsing methods. The Prolog implementation facilitates a good comparison between different parsing methods. Definite Clause Grammar (DCG) has been selected to be the base for all Arabic understanding modules. DCG was used to represent Arabic syntax rules.

Arabic word classification has been addressed from the computation point of view and Arabic word formation has been explained. Different strategies to build an efficient Arabic lexicon has been examined. The possible Arabic lexicon structures has been compared and the best was selected to better serve all syntax and semantic analysis. An incremental approach has been used to explain the built of a complete Arabic lexicon. A complete morphology analyser has been developed using Prolog to serve this purpose.

TABLE OF CONTENTS

Chapter 1

Introduction

1.1 What is Natural Language Processing?	5
1.2 The need for NLP	6
1.2.1 The machine ability to process linguistic jobs	6
1.2.2 The machine ability to interact naturally with humans	7
1.3 The feasibility of Natural Language Processing	7
1.4 The need for Arabic Natural Language Processing	8
1.5 Natural Language Understanding	8
1.5.1 The organization of Actual NLU systems	9
1.5.2 Phonological Knowledge (Phonetics)	10
1.5.3 Morphological Knowledge	11
1.5.4 Syntactic Knowledge (Syntax)	11
1.5.5 Semantic Knowledge (Semantics)	11
1.5.6 Pragmatic Knowledge (Context)	11
1.5.7 World Knowledge	11
1.5.8 An illustrative example	11
1.6 Natural Language Understanding and Ambiguity	13
1.6.1 Ambiguity On the word level (Morphological ambiguity)	14
1.6.2 Sentence Ambiguity	15
1.6.3 Context Ambiguity	16
1.7 Arabic Language vs. English Language	16
1.8 Thesis Objective	17
1.9 Thesis Strategy	18
1.10 Thesis Organization	19

Chapter 2

Arabic Syntactic Processing

2.1 Language, Grammar and parsing	21
---	----

2.2 The four types of formal grammars	22
2.3 Regular grammar	23
2.3.1 Simple Transition Network(STN)	23
2.4 Context Free Grammar (CFG)	25
2.5 Recursive Transition Network(RTN)	26
2.6 Logical grammars	28
2.7 The history of logical grammars	29
2.7.1 Metamorphosis Grammars	29
2.7.2 Definite Clause Grammars	30
2.7.3 Extraposition Grammars	30
2.7.4 Definite Clause Translation Grammars	31
2.8 Testing a grammar	31
2.9 Features And Augmented Grammars	32
2.9.1. Augmented Transition Network (ATN)	32
2.9.2 Checking features with tests.....	33
2.9.3 Useful Feature Systems	36
2.9.3.1 Features agreement (Number, Gender, Person and others)	37
2.9.3.2 Verb complements	38

Chapter 3

Parsing Methods

3.1 Top-Down Parsing methods	41
3.1.1 Top- Down parser for CFG	42
3.1.2 Top-Down parser for Horn Clause Theorem Prover	45
3.1.3 Top Down Parser for RTN	47
3.1.4 Remarks on the Top-Down Parsers	50
3.2 Bottom-UP Parsing Method	50
3.3 Mixed mode parsers	56
3.3.1 A Top-Down CFG Parser with a chart	56
3.3.2 A mixed mode RTN Parser	61
3.3.3 A mixed mode logical grammar Parser	61

Chapter 4

Arabic NLU using Prolog

4.1 Using Prolog as an Arabic language grammar	62
4.1.1 Prolog as a simple grammar	63
4.1.2 Building the parsing tree(Recording the building structure)	65
4.1.3 Reading text from input	67
4.1.4 Adding extra features(Augmentation)	68
4.2 Different parsing algorithms using Prolog	71
4.2.1 Top-Down Parser using Prolog	71
4.2.2 Bottom-Up Parser using Prolog	71
4.2.3 A mixed mode parser using Prolog	76
4.3 Definite Clause Grammar(DCG) and Prolog	81
4.4 Building a DCG grammar using Prolog	84

Chapter 5

Arabic lexicon

5.1 Introduction	90
5.2 Arabic word classification	91
5.3 Verbs	91
5.3.1 The verbs built	92
5.3.2 Other morphology features	94
5.4 Nouns	95
5.4.1 Non- Conjugated Nouns	95
5.4.2 Conjugated Nouns	99
5.4.3 Noun structure	102
5.4.4 Extended Derivative Nouns	102
5.5.5 Conjugated Noun features	105
5.5 Particles	106
5.6 Arabic word composition	107
5.6.1 Prefixes and Suffixes	108
5.7 Remarks on existing Arabic lexicons	111

5.8 Design strategies for Arabic lexicon	120
5.8.1 Results of the morphology rules are new entries in the lexicon	120
5.8.2 The morphology rules themselves in the lexicon	121
5.9 Morphology and Arabic lexicons	122
5.10 What is to be stored in the lexicon ?	124
5.11 Arabic lexicon design	128
5.11.1 Lexicon objectives	129
5.11.2 The design strategies	129
5.11.3 The morphology analyser	131
5.11.4 Prefixes and Suffixes analysis	131
5.11.5 A simple morphology analyser	136
5.11.6 The complete morphology analyser level 3	143
5.12 The output from the morphology analyser	147
 Chapter 6	
Conclusion and future work	
6.1 Conclusion	153
6.2 Future work	153
 References	155

Chapter 1

Introduction

1.1 What is Natural Language Processing?

The Natural Language Processing (NLP) science as its name implies is the science that allows computers to both understand and generate Natural Language (like English, Arabic, etc.), the same way as humans do. Computers cannot perform many of the tasks people do every day until they, as human, share the ability to use language. Although a three-year old child, who cannot play a legal game of chess, can speak and understand his native language, nobody has still made a computer program whose overall linguistic performance rivals that of people.

Natural Language Processing is one of the sciences that most proved the idea that "The practical need is the main motivation for any progress in any science and not its technological feasibility".

In fact, the need for NLP has existed for a long time, trying to motivate this science even before the existence of the technological tools needed for the research. This explains the long period that the research in this field has exist without practical results.

The idea of, having a computer that can simulate people behaviour and can understand and talk, was always the dream of human being -and still is- since the introduction of the first computer. But unfortunately, among four generations of computers, this dream could not become true and the use of computers has always kept limited to its great ability to store, process and retrieve information; plus its enormous speed in execution of arithmetic operations.

1.2 The need for NLP

The need for computers that can deal with natural languages can be divided into two main categories; each has its own motivations and area of application. The two motivations are:

- The machine ability to process linguistic jobs
- The machine ability to interact naturally with humans

1.2.1 The machine ability to process linguistic jobs

1. For several decades, man has always wanted to have a machine that will automatically do linguistic jobs for him. Maybe the most important need in this field was the automatic translation machine which could have a great impact on technology transfer and the evolution of all sciences.

Although the development of a complete machine translation system is not successful yet, there is a lot of limited systems that perform some translation-related tasks.

2. The need for a machine that automatically correct and analyse written and spoken language, was another area that motivated NLP in the field of linguistics. This ability can be used to build a spelling checker, a writing style checker and even a full dictating system.

3. The exponential increase of available/used information which needs an efficient automatic way to organize, store, retrieve and process. This efficient way cannot exist without the ability to analyse text, meanings and logical reasoning of context.

4. The introduction of Expert Systems which simulate expert behaviours and should have the ability to gain experience and process knowledge that requires logical reasoning of text.

1.2.2 The machine ability to interact naturally with humans

1. The spread of home and personal computers as a result of the introduction of micro-computers and its result in changing the nature of the user of the computers who is not, any more, necessarily a computer expert. This has raised the need for having a better interface with the user. So, instead of teaching the user computer languages, the computer should learn human languages and let the user concentrate on the field of the application he is using the computer for.
2. The introduction of expert systems which deal with experts in different specialized fields - and who are not computer experts- requires the ability to communicate with them in both directions, using human languages with a very high degree of efficiency.
3. The recent trend to use computers as an education tool to cover the limitation of books towards the complexity of different education curriculums. The computer education courses, which are called courseware require a high degree of interaction with the student which involves also the usage of human languages with a very high degree of efficiency.

1.3 The feasibility of Natural Language Processing

The research work to turn the above mentioned human dream into reality has been exponentially accelerated due to the progress in the sciences needed for NLP which could be summarized as follows:

1. The progress in different linguistic sciences and the ability to express some of the language features into mathematical, logical and statistical rules.
2. The reasonable spread in using Artificial Intelligence techniques which is useful for NLP; and the progress in the development of Operating Systems and logical computer languages which make developing NLP systems more feasible.
3. The introduction of fast micro-computer systems which made it feasible to

implement the above mentioned Artificial Intelligence techniques on Personal Computers economically and practically.

The bad need for NLP with the existence of some of the tools that are needed to build NLP systems made this field one of the hot fields in the computer/linguistic researches.

1.4 The need for Arabic Natural Language Processing

Now coming to Arabic language, we find that the need for this science is even more important. Not only we have all the above mentioned needs that promote the research for foreign languages, but we also have our own extra reasons. These reasons could be summarized as follows:

1. The bad need for automatic translation: Since most of the current sciences are available in English and other foreign languages, it is clear how beneficial would be to have an automatic translator from English/ other languages to Arabic.
2. Although Arabic language is very rich, there are not enough studies about the logical, arithmetical and statistical approaches for describing Arabic language. The research in Arabic language understanding will certainly invoke such linguistic researches.
3. We have always been accused that our language is difficult to learn. This accusation has ignored what we called the richness of Arabic. The NLP research could be used in teaching Arabic language.

1.5 Natural Language Understanding

The Natural Language Processing science can be mainly divided into two branches:-

- Natural Language Understanding.
- Natural Language Generation.