A THEORETICAL AND COMPUTATIONAL TWO LEVEL FINITE-STATE APPROACH FOR ARABIC DERIVED BROKEN NOUN CV-METADATA MORPHOLOGY

NASER K. A. K. ALAJMI

UNIVERSITY TEKNOLOGI MALAYSIA
DECLARATION OF THESIS AND COPYRIGHT

Author’s full name : Naser Kamdan Abdullah Kamdan Alajmi
Date of birth : 22-05-1965
Title : A THEORETICAL AND COMPUTATIONAL TWO LEVEL FINITE STATE APPROACH FOR ARABIC DERIVED BROKEN NOUN CV-METADATA MORPHOLOGY
Academic Session : II (2008-2009)

I declare that this thesis is classified as:

☐ CONFIDENTIAL (Contains confidential information under the Official Secret Act 1972)*
☐ RESTRICTED (Contains restricted information as specified by the organization where research was done)*
☒ OPEN ACCESS I agree that my thesis to be published as online open access (full text)

I acknowledged that Universiti Teknologi Malaysia reserves the right as follows:

1. The thesis is the property of Universiti Teknologi Malaysia.
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
3. The Library has the right to make copies of the thesis for academic exchange.

Certified by:

SIGNATURE
002166449 (NEW IC NO. / PASSPORT NO.)

SIGNATURE OF SUPERVISOR
Professor Dr. Safaai Deris

NAME OF SUPERVISOR

Date : 1- July- 2009
Date : 1- July- 2009

NOTES : * If the thesis is CONFIDENTAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction.
“We hereby declare that we have read this thesis and in our opinion this thesis is sufficient in terms of scope and quality for the award of the degree of Doctor of Philosophy (Computer Science)”

Signature : ..............................................
Name of Supervisor : Professor Dr. Safaai Deris
Date : 1st July 2009

Signature : ..............................................
Name of Supervisor : Dr. Salah Rashid Alnajem
Date : 1st July 2009
BAHAGIAN A – Pengesahan Kerjasama*

Adalah disahkan bahawa projek penyelidikan tesis ini telah dilaksanakan melalui kerjasama antara.............................. dengan ........................................

Disahkan oleh:
Tandatangan :..................................................Tarikh :.........................
Nama .................................................................
Jawatan .................................................................
(Cop rasmi)
* Jika penyediaan tesis/projek melibatkan kerjasama.

BAHAGIAN B – Untuk Kegunaan Pejabat Sekolah Pengajian Siswazah

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar: Prof. Dr. Tengku Mohammad bin Tengku Sembok
Sembok
Tembalan Naib Canselor (Akademik & Pengantarabangsaan),
Pejabat Timbalan Naib Canselor (Akademik & Pengantarabangsaan),.................................
Canseleri,
Universiti Pertahanan Nasional Malaysia,
Kem Sungai Besi,
57000 Kuala Lumpur

Nama dan Alamat Pemeriksa Dalam: Prof. Madya Dr. Ali bin Selamat
Sekolah Pengajian Siswazah,
UTM, kudai.

Nama Penyelia Lain (jika ada) : -

Disahkan oleh Timbalan Pendaftar di SPS:
Tandatangan : .................................................. Tarikh : .........................
Nama : KHASIM BIN ISMAIL
A THEORETICAL AND COMPUTATIONAL TWO LEVEL FINITE-STATE APPROACH FOR ARABIC DERIVED BROKEN NOUN CV-METADATA MORPHOLOGY

NASER K. A. K. ALAJMI

A thesis submitted in fulfilment of the requirements for the award of the degree of Doctor of Philosophy (Computer Science)

Faculty of Computer Science and Information System
Universiti Teknologi Malaysia

JULY 2009
I declare that this project entitled “A Theoretical and Computational Two Level Finite-State Approach for Arabic Derived Broken Noun CV-metadata Morphology” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : ..............................................
Name of Candidate : NASER K. A. K. ALAJMI
Date : 1st July 2009
DEDICATION

To my Lord: "The One who created me, and guides me. The One who feeds me and quenches my thirst. And when I get sick, He cures me. The One Who puts me to death, and then brings me back to life. The One Who will hopefully forgive my sins on the Day of Judgment. My Lord, grants me wisdom, and includes me among the righteous."

(Prophet Abraham peace be upon him)
ABSTRACT

Arabic morphology is a nonconcatenative type of morphology. It depends on manipulating root letters in a nonconcatenative manner by using patterns. This morphology type poses the inheritance problems of Arabics derived irregular noun morphology such as allomorphic, gemination of radical roots, Arabic orthography and syllabifications. These problems may cause stems to undergo various modifications and redundancies in the presence of different syntactic features and certain stem consonants and vowels. As a result, the major problem is the large number of redundant variants that are captured and implemented in a prefect irredundant manner. The existing computational and theoretical approaches of pattern “sharing” and “non-sharing” techniques have redundant modifications because of the absence of general rules and pattern ontology feature format working on all patterns. This thesis aims to address these redundancy problems. They are resolved by capturing the derivational aspect of the approach based on the generalizations, dependencies and syncretism that rule the Arabics derived irregular noun patterns. This capturing is used in the two-level finite-state Consonant Vowel (CV)-metadata morphology to enable the removal of absolute and internal redundancy. In order to do so, the CV-pattern in the CV-metadata is partitioned into five fundamental parts. The prefix part, left part and infix part are concatenated from left to right, providing no branching to constitute the left side. Similarly, the right part and suffix part are concatenated as the right side to allow branching. Finally, the resulting left side and right side are concatenated to construct an irredundant CV-pattern. Each irredundant CV-pattern is equipped with a tagging system that defines the CV-pattern ontology features. This technique has been implemented using Xerox finite-state tool and compared against the "sharing" and "non-sharing" approaches. The comparative factors are size, states and edges of the CV-metadata. The results show that this approach has the highest level of redundancy reduction effect on all the factors before and after roots population of the CV-metadata.
ABSTRAK

TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DEDICATION</td>
<td>iii</td>
</tr>
<tr>
<td></td>
<td>ABSTRACT</td>
<td>iv</td>
</tr>
<tr>
<td></td>
<td>ABSTRAK</td>
<td>v</td>
</tr>
<tr>
<td></td>
<td>TABLE OF CONTENTS</td>
<td>vi</td>
</tr>
<tr>
<td></td>
<td>LIST OF TABLES</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>LIST OF FIGURES</td>
<td>xiii</td>
</tr>
<tr>
<td></td>
<td>LIST OF APPENDICES</td>
<td>xv</td>
</tr>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.2 Problem Background</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.3 Problem Statement</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1.4 Significance of Research</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1.5 Research Objectives</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1.6 Research Scope</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1.7 Organization of Research</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>LITERATURE REVIEW</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>2.1 Introduction</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>2.2 Finite State Automata and Transducer Basics</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>2.3 Finite State Automata Morphology</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>2.4 Finite State Transducer Morphology</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>2.5 Two Level Technology</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>2.5.1 Xerox Finite State</td>
<td>16</td>
</tr>
</tbody>
</table>
2.6 Morphology 17
2.7 Arabic Morphology 18
2.8 Arabic Language 20
2.9 Arabic Language Word Derivation System 21
2.10 Arabic Verb Derivation 26
   2.10.1 Arabic Perfective Active Verb Structures 27
   2.10.2 Arabic Imperfective Active Verb Structures 32
   2.10.3 Arabic Imperfective Passive Verb Structures 32
   2.10.4 Arabic Future Imperfective Active and Passive Verb Structures 34
   2.10.5 Pronominal Suffixes 34
2.11 Arabic Noun Derivation 36
   2.11.1 Nouns Number 37
   2.11.2 Singular Nouns 37
   2.11.3 Dual Nouns 38
   2.11.4 Plural Nouns 38
   2.11.5 Masculine Plural Sound (regular) Noun 39
   2.11.6 Feminine Plural Sound (regular) Noun 40
   2.11.7 Irregular Plural (Irregular) 40
2.12 Theoretical Approach to Arabic Morphology 42
   2.12.1 Arabic Grammarian’s Approach: 42
   2.12.2 CV-Skeletal approach 42
   2.12.3 Prosodic Morphology Approach: 47
   2.12.4 Affixation Analysis Approach 50
2.13 Computational Approach to Arabic Morphology 51
   2.13.1 Kay's Four Tape Finite State Model 51
   2.13.2 Beesley Two Level Model 52
   2.13.3 Kiraz Multi-Tape Two Level Model 53
   2.13.4 Arabic Morphology Inheritance-Based Approach 57
   2.13.5 Arabic Morphology Sharing-Based Approach 59
2.14 Summary 65
3 METHODOLOGY
3.1 Introduction 66
3.2 Framework of the Study 67
3.3 Linguistic Data Collection and formalization 67
3.4 Theoretical Two level Finite State Phase 68
3.5 The computational Two-level Finite State phase 70
3.6 The Experiments and Comparison phase 72
3.7 Summary 73

4 LINGUISTIC DATA COLLECTION
4.1 Introduction 74
4.2 Converting Arabic Measure Balance Form to CV-skeleton 75
4.3 Broken Nouns 76
4.3.1 Agent Noun (AN) 77
4.3.2 Exaggeration Form (EF) 78
4.3.3 Semi-Agent Noun Adjective (SANA) 79
4.3.4 Patient Noun (PN) 80
4.3.5 Instrument Noun (IN) 81
4.3.6 Time and Place Noun (T&PN) 82
4.4 Summary 82

5 THEORETICAL TWO LEVELFINITE STATE APPROACH FOR ARABIC DERIVED BROKEN NOUN CV-METADATA MORPHOLOGY
5.1 Introduction 84
5.2 Theoretical Sharing Approach of ADBNM 86
5.3 Capturing CV-skeletons Syncretism 86
5.4 Capturing CV-skeletons Generalizations and Dependencies 90
5.4.1 Capturing CV-skeletons Generalizations 90
5.4.2 Capturing CV-skeletons Dependencies 93
5.5 Summary 99
A COMPUTATIONAL TWO LEVEL FINITE STATE APPROACH FOR ARABIC DERIVED BROKEN NOUN CV-METADATA MORPHOLOGY

6.1 Computational Sharing Approach

6.2 Universal Mini Rewrite Rules Transducers Character-Based

6.3 A Universal CV-skeleton Transducer Based On Rewrite Rules

6.4 Irredundant CV-metadata Morphology Network Construction of Derived Broken Nouns

6.5 Population of the Irredundant CV-Metadata-Morphology Network

6.6 Summary

RESULT AND DISCUSSION

7.1 Introduction

7.2 Experiment Linguistic Data and Formalization

7.3 Implementation Issues

7.4 Comparison Pattern Formula Issues

7.5 Experiments

7.6 Discussion

7.7 Limitations

7.8 Conclusion

CONCLUSION AND FUTURE WORK

8.1 Conclusion

8.2 Contributions

8.3 Future work

REFERENCES

APPENDICES A-B
<table>
<thead>
<tr>
<th>TABLE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Closure properties for regular languages and regular relations</td>
<td>11</td>
</tr>
<tr>
<td>2.1</td>
<td>Examples of derivational morphology</td>
<td>19</td>
</tr>
<tr>
<td>2.2</td>
<td>Examples of inflectional morphology</td>
<td>19</td>
</tr>
<tr>
<td>2.3</td>
<td>The example of word derivation in English language</td>
<td>22</td>
</tr>
<tr>
<td>2.4</td>
<td>The example of word derivation in Arabic language</td>
<td>22</td>
</tr>
<tr>
<td>2.5</td>
<td>Illustration of the function of the augmentation letters</td>
<td>23</td>
</tr>
<tr>
<td>2.6</td>
<td>The example of root notion of Arabic language</td>
<td>25</td>
</tr>
<tr>
<td>2.7</td>
<td>The classes of Arabic verb perfective and imperfective</td>
<td>26</td>
</tr>
<tr>
<td>2.8</td>
<td>The aspects of English verb past tense</td>
<td>27</td>
</tr>
<tr>
<td>2.9</td>
<td>Expressing verbal aspects in Arabic language compared to English language</td>
<td>28</td>
</tr>
<tr>
<td>2.10</td>
<td>Basic structures of perfective verb active voice</td>
<td>28</td>
</tr>
<tr>
<td>2.11</td>
<td>A summary of basic form &quot;I&quot; verb active voice based on &quot;dynamic&quot;, &quot;stative&quot;, &quot;transitive&quot;, and &quot;intransitive&quot; verb</td>
<td>30</td>
</tr>
<tr>
<td>2.12</td>
<td>A summary of the fourteen perfective active verb structures</td>
<td>31</td>
</tr>
<tr>
<td>2.13</td>
<td>A summary of the fourteen imperfective active verb structures</td>
<td>32</td>
</tr>
<tr>
<td>Section</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>2.14</td>
<td>Alternating rule that formed form &quot;I&quot; passive perfective structures from form &quot;I&quot; active perfective structures</td>
<td></td>
</tr>
<tr>
<td>2.15</td>
<td>A summary of the fourteen imperfective passive verb structures corresponding to the fourteen perfective active structures</td>
<td></td>
</tr>
<tr>
<td>2.16</td>
<td>The future tense particles used for the imperfective verb structure in Arabic</td>
<td></td>
</tr>
<tr>
<td>2.17</td>
<td>Examples future tense particles suffixation of the verb &quot;go&quot;</td>
<td></td>
</tr>
<tr>
<td>2.18</td>
<td>The full conjugation scheme which can be applicable to all verbs whose roots do not have weak letters</td>
<td></td>
</tr>
<tr>
<td>2.19</td>
<td>Conjugation of 3rd person singular perfective verbs</td>
<td></td>
</tr>
<tr>
<td>2.20</td>
<td>Conjugation schema for the singular noun with its gender suffixation</td>
<td></td>
</tr>
<tr>
<td>2.21</td>
<td>Dual noun with its gender suffixation based on grammatical two cases</td>
<td></td>
</tr>
<tr>
<td>2.22</td>
<td>Conjugation schema for the dual noun with its gender suffixation based on grammatical two cases</td>
<td></td>
</tr>
<tr>
<td>2.23</td>
<td>The masculine plural sound noun with its gender suffixation based on grammatical two cases</td>
<td></td>
</tr>
<tr>
<td>2.24</td>
<td>Conjugation schema for the masculine plural sound noun with its gender suffixation based on grammatical two cases</td>
<td></td>
</tr>
<tr>
<td>2.25</td>
<td>The feminine plural sound noun with its gender suffixation based on grammatical two cases</td>
<td></td>
</tr>
<tr>
<td>2.26</td>
<td>Conjugation schema for the masculine plural sound noun with its gender suffixation based on grammatical two cases</td>
<td></td>
</tr>
<tr>
<td>2.27</td>
<td>Examples of Arabic irregular plural nouns pluralization</td>
<td></td>
</tr>
<tr>
<td>2.28</td>
<td>Examples Arabic noun could be pluralized by more than one way</td>
<td></td>
</tr>
</tbody>
</table>
2.29 The heavy syllable always transfers its final syllable to the final syllable

2.30 The number of syllables in broken plural depends on the number of syllables of the broken singular

2.31 Affixation analysis with parsing function

2.32 Two-level analysis of /janaadib/ and /salaatiin/

2.33 Two-level analysis of /jundub/ and /sultaan/

4.1 Agent noun structures

4.2 Exaggeration form noun structures

4.3 Semi-agent noun adjective structures

4.4 Patient noun adjective structures

4.5 Instrument noun adjective structures

4.6 Instrument noun adjective structures

5.1 Absolute redundancy removed by syncretism

5.2 Absolute redundancy removed by syncretism

5.3 Partitioning the CV-skeleton

5.4 Grouping the five parts of CV-skeletons into left and right sides

5.5 Removing the internal redundancies by capturing the dependencies

5.6 Removed internal redundancy by generalization and dependency

7.1 CV-metadata size before population

7.2 CV-metadata size after the population of thirty roots
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>A simple finite-state automaton accepting the language</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>A simple FST that computes the relation language</td>
<td>11</td>
</tr>
<tr>
<td>2.3</td>
<td>Finite state network model that describe the entities and relations</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>between the different states of a given set of words</td>
<td></td>
</tr>
<tr>
<td>2.4</td>
<td>A finite state transducer which translates from Arabic to English</td>
<td>14</td>
</tr>
<tr>
<td>2.5</td>
<td>Abstract example of a tow-level rule</td>
<td>15</td>
</tr>
<tr>
<td>2.6</td>
<td>Example of how the two-level morphology applied its rules</td>
<td>15</td>
</tr>
<tr>
<td>2.7</td>
<td>A cascade of replace rules composed into a single FST</td>
<td>16</td>
</tr>
<tr>
<td>2.8</td>
<td>A set of two level rules intersected into a single transducer</td>
<td>16</td>
</tr>
<tr>
<td>2.9</td>
<td>Acknowledgement of the two kinds of processes in English</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>language with theirs corresponding morphology types</td>
<td></td>
</tr>
<tr>
<td>2.10</td>
<td>Root-Pattern-Morphology</td>
<td>19</td>
</tr>
<tr>
<td>2.11</td>
<td>Syllabic typology</td>
<td>48</td>
</tr>
<tr>
<td>2.12</td>
<td>Prosodic hierarchy structure</td>
<td>48</td>
</tr>
<tr>
<td>2.13</td>
<td>Prosodic hierarchy and plural template</td>
<td>49</td>
</tr>
<tr>
<td>2.14</td>
<td>The Prosodic hierarchy and deriving</td>
<td>49</td>
</tr>
<tr>
<td>2.15</td>
<td>Illustrates how the stem &quot;daras&quot; is constructed by Beesley</td>
<td>52</td>
</tr>
</tbody>
</table>
# LIST OF APPENDICES

<table>
<thead>
<tr>
<th>APPENDIX</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>The transliteration table</td>
<td>150</td>
</tr>
<tr>
<td>B</td>
<td>The 30 roots transliteration table</td>
<td>152</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 Introduction

As Natural Language Processing (NLP) systems grow more sophisticated, they need larger, more comprehensive and detailed morphology. This seems to become all the more true as NLP systems get bigger and move out of the laboratory into applications; consequently, that require more information in each lexical entry of the system. This research involves one of the most important areas of the attempt to make computer systems available in many Arabic NLP applications. The Arabic broken noun morphology is an important, and productive integrated part for the whole Arabic morphology because, it is the lowest layer in many NPL applications like information retrieval, parsing, text proofing, text generation applications and other applications in natural language.

1.2 Problem Background

Most verbs in the Arabic language follow clear rules that define their morphology and generate their paradigms. However, broken nouns do not seem to follow a similar set of well-defined rules which cause the allomorphic problem; instead, there are groups that show pattern family resemblances. In addition, Arabic noun morphology has an overt generalizations, dependency and syncretism; they are absent in the computational approaches introduced in the literature; except, Alnajem (1998, 2004, 2006) mentions that Arabic morphology has a number of
generalizations, dependency and syncretism which give details over the aspects of regularity in its derivational and inflectional sides. Moreover, the major problems of Arabic morphology are allomorphic, the geminating of radical roots as well as the zero, partial and full realization of short vowels. As a result, broken nouns should be analyzed and formalized in CV-metadata broken noun morphology which depends on their CV-skeletons. In order to make it available, in many Arabic natural language applications, in a compact and non-redundant form, the generalization, dependency and syncretism must be captured. The generalization is a statement about the facts of language morphemes which holds true in all word structure or in nearly all structures. The dependency in context of this work is the case in which all the non-redundant valid forms is derived from another original forms through applying changes to that original forms by concatenated their language morphemes in a manner with no redundancy. The syncretism is the case in which two or more features which are morphs syntactically distinct appear identical in the form.

1.3 Problem Statement

Natural language processing (NLP) systems need a large amount of explicit information for each single vocabulary known to the system. Different applications, in turn, require different information in each lexical entry. The different information is called lexical features. The lexical features are grouped based on their functions on the layer they belong to inside the NLP system. Pronoun feature (first person, second person, and third person) is the examples of the broken noun inflectional morphology feature; noun class (e.g. agent, patient), number (e.g. singular, plural) and gender (e.g. male, female, neutral) are the broken noun derivational morphology case of point. Derivational morphology features are one of the difficulties in the broken noun morphology because there are no clues to figure out them by the broken noun stems. As far as it is known there is no computational Arabic broken noun morphology for computers available at this time, Kiraz (1995) believed that nobody has dealt computationally with the challenging problem of the Arabic broken plural. And Xu & Weischedel (2002) said that broken plurals are not handled by most of current Arabic stemmers. Alnajem (1998, 2004, 2006) stated that there are generalizations,
dependencies in the derivational system of Arabic morphology which have not been formalized and implemented effectively. The focuses of this research are on the theoretical and computational formalization of capturing the generalizations, dependency and syncretism. Moreover, the second focus is their implementation in the Arabic broken noun morphology CV-metadata which yields a compact and irredudant construction. Last focus is to solving the inheritance problems of Arabic derived broken noun morphology such as allomorphic, the gminating of radical roots and the complexity of standard Arabic orthography between the lexical and surface strings that caused by zero, partial or full realization of short vowels.

1.4 Significance of Research

The main significance of this work is to formalize and implement a theoretical and computational compact finite state CV-metadata morphology for the Arabic broken derived noun. This approach attempts to capture generalizations, dependency and syncretism existing in Arabic broken noun morphology to handles generating and analysis processes in a compact non-redundant manner, when it is implemented in many useful natural language applications. Focusing on capturing the dependencies, generalizations and syncretism in Arabic morphology is absent in the computational systems and approaches introduced in the literature. This capturing yields linguistic motivations in the theoretical and computational formalization for Arabic morphology. It also saves this formalization from the redundancy. It also makes the patterns share once the general and dependent internal parts rather than repeating these internal parts to waste memory. Moreover, it is solving inheriting Arabic morphology such as allomorphic, the gminating of radical roots and the zero, partial and full realization of short vowels which are the major problems of Arabic morphology with universal rewrite rules rather than specific-pattern rules.

1.5 Research Objectives

The main objectives of this work are as follow:
1. To collect and formalize explicitly the Arabic standard balance forms from the traditional Arabic linguistic literature then converting them to the CV-skeleton for the purpose of theoretical and computational implementations.

2. To construct and formalize a theoretical and computational, compact irredundant CV-metadata network morphology for the Arabic derived broken noun by capturing generalization, dependency and syncretism; with the capability of solving the inheriting broken noun morphology problems and burdens-free such as specific-patterns rules and unnecessary excessive tags.

3. To implement CV-metadata network morphology of broken derived noun by the Xerox finite state machine tool.

1.6 Research Scope

This work considers the derivational part of Arabic derived broken noun morphology while the inert noun is out of the scope; hence it is the majority of the broken nouns. It is also, considering the two-level finite-state theoretical and computational approach, in which I attempt to capture generalizations, dependency and syncretism in derivational broken derived noun patterns. The inflectional part of the approach was considered in Alnajem (2003, 2006). The orthographic part of the approach was considered in Alnajem (2005a, 2005b). The approach will be implemented using Xerox finite-state tools and use the CV-skeleton approach introduced by McCarthy to handle word structure. The considered data in this work are all the trilateral roots of the Arabic morphology balance forms of the six classes of the derived broken nouns.

1.7 Organization of Research

This thesis is organized into 8 chapters. A brief description of the contents of each chapter is given as follows:
1. Chapter 1 describes the problems, significance, objective, and scope of the study.

2. Chapter 2 reviews main subjects used in the thesis that include basic finite state automata, morphology, Arabic language, Arabic language word formation, theoretical and computational approaches in Arabic morphology.

3. Chapter 3 adopted the operational framework to achieve the objective of the study including the experiments and comparison outlines.

4. Chapter 4 describes the data collection and formalization.

5. Chapter 5 describes a theoretical approach of the CV-metadata two-level finite state morphology networks construction.

6. Chapter 6 describes a computational approach of the CV-metadata two-level finite state morphology networks construction.

7. Chapter 7 describes and explain the experiment data, the experiment patterns representation formulas and the experiments of the construction of the CV-metadata network morphology before and after its roots population for each approach of the three approaches: this research approach which is called Naser sharing approach, Alnajem (2004) Sharing approach, and the "Others" non-sharing approach to judge the redundancy level effect according to four factors: the size on the disk, and the number of states, edges and paths.

8. Chapter 8 draws general conclusions about the achieved results and presents the contributions of this research and future works of the study.