Open-source Resources and Standards for Arabic Word Structure Analysis:

Fine Grained Morphological Analysis of Arabic Text Corpora

By

Majdi Shaker Salem Sawalha

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

UNIVERSITY OF LEEDS

The University of Leeds
School of Computing

October, 2011

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Memory

هذه الأطروحة مقدمة لذكرى والدي الحبيب، رحمه الله.

I dedicate this thesis to the memory of the most beloved Father,

Shaker Sawalha
(March 3, 1949 - March 5, 2011)

who lived a life of dignity, courage, wisdom, patience and above all affection, and who brought me up on the true values of life. Father, you will remain my personal hero and my inspiration forever.

May God bless his soul, Amen.

Acknowledgements

I am thanking my GOD *Allāh* for giving me health, patience and strength to write this thesis and all the graces he has granted to me.

I would like to thank my supervisor Dr. Eric Atwell for supervising me during these four years. Thank you very much for your patience, guidance and encouragement. I learnt from how to be a real researcher, how to think differently and how to understand life better.

I would also like to thank the NLP group members for the great seminars we used to enjoy almost every week. Again, it's a great opportunity here to thank Dr. Latifa Al-Sulaiti for her support, encouragement and advice. And I would like to thank all my friends here in the UK and back home in Jordan.

I would like to thank Claire Brierley for being a true friend, and for the discussions, sharing ideas and plans for future research. I am looking forward to producing lots of publications from our great ideas.

To my best friend Dr. Mohammad Haji, thank you very much for being my real friend whom I trust. Your wise advice, encouragement and unending generosity made my research and life in the UK easy and enjoyable. Thank you for being there during the good times and the hard times. I really wish you the best of luck in your life and career.

Finally, I dedicate this thesis to my family who have always supported me in my studies and life. Without your love, care and patience, I would not have achieved this. I would like to thank my eldest brother Rami and his family members: my sister-in-law Dina, my nephew Faris, and my nieces Tala, Layan and Jude. My sister Noor and her family: my brother-in-law Husam, my niece Hadeel, and my nephew Mohammed (who's just born). My sister Dua' and her family: my brother-in-law Mohammed and my nieces Dana and Heba. My sister Eman and her family: my brother-in-law Omar and my niece Hala (who's just born). My youngest brother Mohammed, I wish you the brightest future. My youngest sister Rahma, we are all lucky to have you as our beloved sister. To my beloved Grandma, I wish you prosperity and a long happy life.

The special dedication of this thesis is to the most beloved Mum. Thank you for your patience, care and everything you have done to keep our family gathered in peace and happiness. Thank you for giving us the love we need to survive in this life. I always love you Mum.

Declaration

I declare that the work presented in this thesis, is the best of my knowledge of the domain, original, and my own work. Most of the work presented in this thesis have been published. Publications are listed below:

(*Majdi Sawalha*)

Chapter 3

1- **Sawalha, M.** and E. Atwell (2008). Comparative evaluation of Arabic language morphological analysers and stemmers. Proceedings of COLING 2008 22nd International Conference on Computational Linguistics.

Chapter 4

2- **Sawalha, M.** and E. Atwell (2010). Constructing and Using Broad-Coverage Lexical Resource for Enhancing Morphological Analysis of Arabic. Language Resource and Evaluation Conference LREC 2010, Valleta, Malta.

Chapters 5 and 6

3- **Sawalha, M.** and E. Atwell (Under review). "A Theory Standard Tag Set Expounding Traditional Morphological features for Arabic Language Part-of-Speech Tagging." Word structure journal, Edinburgh University Press.

Chapter 7

4- **Sawalha, M.** and E. Atwell (2011). التحليل الصَّرفي لنصوص اللغة العربية الحديثة والكلاسيكية "Morphological Analysis of Classical and Modern Standard Arabic Text". 7th International Computing Conference in Arabic (ICCA11), Imam Mohammed Ibn Saud University, Riyadh, KSA.

Chapters 8 and 9

5- **Sawalha, M.** and E. Atwell (2009). توظيف قواعد النحو والصرف في بناء محلل صرفي للغة العربية(Adapting Language Grammar Rules for Building Morphological Analyzer for Arabic Language). Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Technology (KACT) and Arabic Language Academy., Damascus, Syria.

6- **Sawalha, M.** and E. Atwell (2009). Linguistically Informed and Corpus Informed Morphological Analysis of Arabic. Proceedings of the 5th International Corpus Linguuistics Conference CL2009, Liverpool, UK.

7- **Sawalha, M.** and E. Atwell (2010). Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text. Language Resource and Evaluation Conference LREC 2010 Valleta, Malta.

Chapter 10

8- **Sawalha, M.** and E. Atwell (2011). Accelerating the Processing of Large Corpora: Using Grid Computing Technologies for Lemmatizing 176 Million Words Arabic Internet Corpus. Advanced research computing open event. University of Leeds, Leeds, UK.

9- **Sawalha, M.** and E. Atwell (2011). Corpus Linguistics Resources and Tools for Arabic Lexicography. Workshop on Arabic Corpus Linguistics, Lancaster University, Lancaster, UK.

Abstract

Morphological analyzers are preprocessors for text analysis. Many Text Analytics applications need them to perform their tasks. The aim of this thesis is to develop standards, tools and resources that widen the scope of Arabic word structure analysis - particularly morphological analysis, to process Arabic text corpora of different domains, formats and genres, of both vowelized and non-vowelized text.

We want to morphologically tag our Arabic Corpus, but evaluation of existing morphological analyzers has highlighted shortcomings and shown that more research is required. Tag-assignment is significantly more complex for Arabic than for many languages. The morphological analyzer should add the appropriate linguistic information to each part or morpheme of the word (proclitic, prefix, stem, suffix and enclitic); in effect, instead of a tag for a word, we need a subtag for each part.

Very fine-grained distinctions may cause problems for automatic morphosyntactic analysis – particularly probabilistic taggers which require training data, if some words can change grammatical tag depending on function and context; on the other hand, fine-grained distinctions may actually help to disambiguate other words in the local context. The SALMA – Tagger is a fine grained morphological analyzer which is mainly depends on linguistic information extracted from traditional Arabic grammar books and prior-knowledge broad-coverage lexical resources; the SALMA – ABCLexicon.

More fine-grained tag sets may be more appropriate for some tasks. The SALMA – Tag Set is a theory standard for encoding, which captures long-established traditional fine-grained morphological features of Arabic, in a notation format intended to be compact yet transparent.

The SALMA – Tagger has been used to lemmatize the 176-million words Arabic Internet Corpus. It has been proposed as a language-engineering toolkit for Arabic lexicography and for phonetically annotating the Qur'an by syllable and primary stress information, as well as, fine-grained morphological tagging.

Contents

Figures

Tables

List of Abbreviations

Abbreviation	*Meaning*
BAMA	Buckwalter's Morphological Analyzer
CCA	The Corpus of Contemporary Arabic
MSA	Modren Standard Arabic
LDC	Linguisic Data Consortium
APT	Khoja's Arabic Part-of-speech Tagger
FST	Finite state transducer
NLTK	Natural Language Toolkit
SALMA-ABCLexicon	Sawalha Atwell Leeds Morphological Analysis – Arabic Broad-Coverage Lexicon
SALMA-Tag Set	Sawalha Atwell Leeds Morphological Analysis – Tag Set
SALMA-Tokenizer	Sawalha Atwell Leeds Morphological Analysis – Tokenizer
SALMA-Lemmatizer & Stemmer	Sawalha Atwell Leeds Morphological Analysis – Lemmatizer and Stemmer
SALMA-Pattern Generator	Sawalha Atwell Leeds Morphological Analysis – Pattern Generator
SALMA-Vowelizer	Sawalha Atwell Leeds Morphological Analysis – Vowelizer
SALMA-Tagger	Sawalha Atwell Leeds Morphological Analysis – Tagger
CML	Croatian Morphological Lexicon
EAGLES	Expert Advisory Group on Language Engineering Standards
SKEL	Software and Knowledge Engineering Laborartory
Le*fff*	Lexique des formes fléchies du français – Lexicon of French inflected forms
LMF	Lexical Markup Framework, the ISO/TC37 standard for NLP lexicons
XML	Extensible Markup Language
ACL SIGLEX	The Special Interest Group on the Lexicon of the Association for Computational Linguistics
COMLEX	COMmon LEXicon
OTA	Oxford Text Archive

AWN	Arabic WordNet
PWN	Princeton WordNet
CLAWS	The Constituent Likelihood Automatic Word Tagging System
BNC	The British National Corpus
AMALGAM	Automatic Mapping Among Lexico-Grammatical Annotation Models
ICE	International Corpus of English
LLC	London-Lund Corpus
LOB	Lancaster-Oslo/Bergen Corpus
SKRIBE	Spoken Corpus Recoddings In British English
PoW	Polytechnic of Wales corpus
SEC	Spoken English Corpus
UPenn	University of Pennsylvania corpus
SALMA Tag Set	Sawalha Atwell Leeds Morphological Analysis – Tag Set
ALECSO/KACST	Arab League Educational, Cultural and scientific Organization / King Abdul-Aziz City of Science and Technology
PADT	Prague Arabic Dependency Treebank
PATB	The Penn Arabic Treebank
MWEs	Multi-Word Expressions
HMM	Hidden Marcov Model

Part I: Introduction and Background Review

Chapter 1
Introduction

"أَنَا البَحْرُ فِيْ أَحْشَائِهِ الْدُّرُّ كَامِنٌ   فَهَلْ سَألُوا الْغَوَّاصَ عَنْ صَدَفَاتِي"

’anā al-baḥru fī ’aḥšā’ihi ad-durru kāmin^{un} *fahal sa’alū al-ğawwāṣ ‘an ṣadafātī*

“Arabic says: I am the sea where pearls are hidden inside. Have they (the people) asked the diver about my seashells?”

Hafiz Ibrahim (1872 – 1932)

Chapter Summary

Morphological analysis for Arabic text corpora is the topic of this thesis. The thesis topic is introduced in the first section of this chapter. This chapter also provides a general definition of computational morphology. It presents Arabic computational morphology and the complexity of Arabic morphology. The motivations and objectives of the thesis, and the original contributions of developed resources, proposed standards and tools are summarized in section 1.5. Finally, this chapter presents the structure of the thesis.

1.1 This Thesis

The topic of this thesis is morphological analysis for Arabic text corpora. Morphological analysis for text corpora is a prerequisite for many text analytics applications, which has attracted many researchers from different disciplines such as linguistics (computational and corpus linguistics), artificial intelligence, and natural language processing, to morphosyntactically analyze text of different languages including Arabic. Recently, several researchers have investigated different approaches to morphological and syntactic analysis for Arabic text. Many systems have been developed which vary in complexity from light stemmers, root extraction systems, lemmatizers, complex morphological analyzers, part-of-speech taggers and parsers. This introduction will detail what is special about morphological analysis for Arabic text corpora. We will introduce computational morphology and the complexity of Arabic morphology that has inspired this research. The motivation and the objectives for this thesis will be discussed. Both research and practical perspectives on the value of carrying out this research will be explained.

We present the argument that the linguistic wisdom in traditional Arabic grammars and lexicons can be utilized (*i.e.* renewed and re-validated) in an Arabic NLP toolkit which is easy to access and implement. We believe that such detailed knowledge is applicable to Modern Standard Arabic and that it can be used to restore orthographic (*e.g.* short vowels) and morphological features which signify important linguistic distinctions. Moreover, fine-grained morphological analysis is possible (*i.e.* achievable) and advantageous. The implemented Arabic NLP toolkit is general-purpose, adherent to standards and reusable, which will fulfil many researchers' and users' needs.

1.2 Computational Morphology

Morphology is the study, identification, analysis and description of the minimal meaning bearing units that constitute a word. The minimal meaning bearing unit of a word is called a morpheme. Categorizing and building a representative structure of the component morphemes is called morphological analysis. Both orthographic rules and morphological rules are important for categorizing a word's morphemes. For instance, orthographic rules for pluralizing English words ending with *–y* such as *party* indicates changing the *–y* to *-i-* and adding *–es*. And morphological rules tell us that *fish* has null plural and the plural of *goose* is formed by a vowel change. Morphological analysis of the surface or input form *going* is the verbal stem *go* plus the *–ing* morpheme VERB-go + GERUND-ing (Jurafsky and Martin 2008); section 2.3 defines morphological analysis in general, while section 2.3.4 redefines morphological analysis for Arabic text.

Computational morphology is a branch of computational linguistics (*i.e.* natural language processing or language engineering). The main concern of computational morphology is to develop computer applications (*i.e.* toolkits) that analyze words of a given text and deal with the internal structure of words such as determining their part-of-speech and morphological features (*e.g.* gender, number, person, case, mood, voice, etc) (Kiraz 2001); see sections 2.3 and 2.3.4.

Morphological analysis has many applications throughout speech and language processing. In web searching for morphologically complex languages, morphological analysis enables searching for the inflected form of the word even if the search query contains only the base form. Morphological analysis gives the most important information for a part-of-speech tagger to select the most suitable analysis for a given context. Dictionary construction and spell-checking applications rely on a robust morphological analysis. Machine translation systems rely on highly accurate morphological analysis to specify the correct translation of an input sentence (Jurafsky and Martin 2008). Lemmatization is an aspect of morphological analysis. Google's search facilities use lemmatization to produce hits of all inflectional forms of the input word. Statistical models of language in machine translation and speech recognition also use lemmatization. Lexicographic applications use lemmatizers as an essential tool for corpus-based compilation (Pauw and Schryver 2008). Morphological analysis techniques form the basis of most natural language processing systems. Such techniques are very useful for many applications, such as information retrieval, text categorization, dictionary automation, text compression, data encryption, vowelization and spelling aids, automatic translation, and computer-aided instruction (Al-Sughaiyer and Al-Kharashi 2004); see also section 2.3.3.

1.3 Arabic Computational Morphology

Arabic is a living language that belongs to the Semitic group of languages. The Semitic group of languages include other living languages such as: Modern Hebrew, Amharic, Aramaic, Tigrinya and Maltese (Haywood and Nahmad 1965).

The main characteristic feature of Semitic languages is their nonconcatenative morphology where words are derived from their basis of mostly triliteral consonantal roots. Roots of Semitic languages carry the basic conceptual meanings, while varying the vowelling of the simple root and adding prefixes, suffixes and infixes to produce the different variations in shade of meaning (Haywood and Nahmad 1965). For example, from the Arabic root كتب *k-t-b* 'wrote' we can derive the following words by filling in the vowels: كِتاب *kitāb* 'book', كُتُب *kutub* 'books', كَاتِب *kātib* 'writer', كُتَّاب *kuttāb* 'writers', كَتَبَ *kataba* 'he wrote', يَكْتُبُ *yaktubu* 'he writes', etc. Sections 1.4 and 2.3.4.1 discuss in detail the complexity of Arabic morphology.

Arabic is classified into Classical Arabic (*e.g.* the Qur'an); Modern Standard Arabic (*e.g.* newspapers and magazines); and Spoken or Colloquial Arabic. Modern Standard Arabic varies in idiom and vocabulary from Classical Arabic. However, the grammar of the 6th century Classical Arabic still applies largely to modern written Arabic. This is because Classical Arabic was the vehicle of God's Revelation in the Qur'an (Haywood and Nahmad 1965).

The study of traditional Arabic grammar started in the 8th century. The main reason for Arabic linguistic studies was to preserve the original Arab language due to the wide expansion of the Islamic community that included many non-Arabic native speaking Muslims who spoke Arabic to perform daily worship. The first Arabic order for establishing traditional Arabic grammar language was given by the fourth Khalifa Imam Ali bin Abi Talib الإمَام عَلِي بِنْ أَبِي طَالِب *al-'imām 'alī bin 'abī ṭālib* to Abu Al-Aswad Ad-Du'aly أَبُو الأَسْوَدْ الدُّؤَلِي *'abū 'al-'aswad ad-du'alī* to write the fundamentals of Arabic grammar. Early scholars such as Abū Amr bin Al-Ala' أَبُو عَمْرو بِن العَلاء *'abū 'amr bin al-'alā'* established the relations between language and its grammar rules; and the connections of Qur'an recitation styles. Al-Khalil bin Ahmad Al-Farahidi الخَلِيْل بِنْ أَحْمَد الفَرَاهِيدي *al-ẖalīl bin 'aḥmad al-farāhīdī* is the founder of Arabic grammar as a discipline where he defined its rules, regulations, documentation methodologies. These methodologies allowed Sibawayh سِيْبَوَيْه *sībawayh* to write the first comprehensive traditional Arabic grammar book called Al-Kitab الكِتَاب *al-kitāb* 'The Book' (Wlad Abah 2008).

Present-day Arabic language scholars are still interested in studying traditional Arabic grammar books. These interests include rewriting and verifying manuscripts and studying the life of their authors and their methodologies. Among the recent interests of Arabic linguists is the study of new international linguistic knowledge and its application to Arabic. Moreover, researchers are interested in connecting the results of modern linguistic studies applied to Arabic with the findings and conclusions of the early Arabic traditional grammar scholars (Wlad Abah 2008).

Modern linguistic theories of Arabic morphology have studied the derivation process of Arabic words from two points of view: root-based and stem-based (or word-based). The theory of Prosodic Morphology (McCarthy and Prince 1990b; McCarthy and Prince 1990a) defines the basic character of phonological structure and its consequences for morphology. The true templatic morphology is represented by the derivational categories of the Arabic verbs. Using multiple levels of representation, Arabic verbs have three auto-segmental tiers: consonantal tier (*i.e.* the root), CV skeleton (*i.e.* patterns) and vocalic melody (*i.e.* short vowels).

Benmamoun (1999) studied the nature and role of the imperfective verb in Arabic. The imperfective verb is not specified for tense. Hence, it is the default form of the verb that does not carry temporal features. This feature of unmarked status for imperfective verbs is consistent with its central role in word formation which allows for a unified analysis of nominal and verbal morphology. In conclusion, a word-based approach for Arabic word formation is more important than root-based.

Morphological analysis for Arabic entails computer applications that analyze Arabic words of a given text and deal with the internal structure. It involves a series of processes that identify all possible analyses of the orthographic word. These processes are both form-based and function-based (Thabet 2004; Hamada 2009a; Habash 2010; Hamada 2010). Morphological analyzers for Arabic text are required to develop processes that deal with both the *form* and the *function* of the word. These processes include tokenization, spell-checking, stemming and lemmatization, pattern matching, diacritization, predicting the morphological features of the word's morphemes, part-of-speech tagging and parsing.

Many morphological analyzers for Arabic text were developed using a range of methodologies. These methodologies are: Syllable-Based Morphology (SBM), which depends on analyzing the syllables of the word; Root-Pattern Methodology, which depends on the root and the pattern of the word for analysis; Lexeme-based Morphology, where the stem of the word is the crucial information that needs to be extracted from the word; and Stem-based Arabic lexicons with grammar and lexis specifications (Soudi, Cavalli-Sforza and Jamari 2001; Soudi, Bosch and Neumann 2007).

Morphological analyzers are different in their methodologies and their tasks. **Stemmers** are responsible for extracting the stem/root of words (Khoja 2001; Al-Sughaiyer and Al-Kharashi 2002; Al-Shalabi, Kanaan and Al-Serhan 2003; Khoja 2003; Al-Shalabi 2005; AlSerhan and Ayesh 2006; Boudlal et al. 2011). **Lemmatizers** identify the canonical form, dictionary form, or citation form, which is also called the lemma for words (Dichy 2001; Al-Shammari and Lin 2008). **Pattern matching algorithms** generate the templatic form (*i.e.* patterns) and vocalism of the analysed words. However, the representation of the templatic forms and vocalism might vary from one algorithm to another (Dichy and Farghaly 2003; Al-Shalabi 2005; Alqrainy 2008; Yousfi 2010). General purpose **morphological analyzers** generate all possible analyses of the words out of their contexts. Key morphological analyzers for Arabic text are: Xerox system (Beesley 1996; Beesley 1998), Buckwalter's Morphological Analyzer (BAMA) (Buckwalter 2002; Buckwalter 2004), ElixirMF (Smrz 2007), AlKhalil (Boudlal et al. 2010), MORPH2 (Hamado, Belghayth and Sha'baan 2009; Kammoun, Belguith and Hamadou 2010) and MIDAD (Sabir and Abdul-Mun'im 2009).

1.4 The Complexity of Arabic Morphology

Arabic is a highly inflectional language which makes processing tasks for Arabic text extremely hard. Morphological analysis of Arabic text is not an easy task and it affects higher level applications such as part-of-speech tagging and parsing.

Due to the rich "root-and-pattern" non-concatenative (or nonlinear) morphology and the highly complex word formation process of root and patterns, hundreds of words can be derived from a single root by following certain patterns and conjoining affixes and clitics to the word. The attachment of affixes and clitics significantly increases the number of derived words.

Ambiguity in Arabic text is a major challenge for processing. Ambiguity is due to the absence of short vowels for most Arabic texts and the interaction between affixes or clitics letters and the original letters that compose the root especially if one or two long vowels are part of the root letters.

Clitics and affixes of Arabic words are productive. Therefore, storing word forms in a dictionary and doing morphological analysis by dictionary lookup is not possible, as we cannot list all morphological variants of every Arabic word. Thus, morphological analysis done dynamically is unavoidable. A word such as بِوَالِدَيْهِ *bi-wālidayhi* 'in his parents' consists of four morphemes بِ *bi* 'in' is a preposition, وَالِدَ *wālida* 'parent' is the noun stem morpheme, يْ y 'two' is a dual letter, and هِ *hi* 'his' is object relative pronoun. The proclitic بِ *bi* 'in' and the enclitic هِ *hi* 'his' are productive clitics.

The root letters can be hard to guess and increase text ambiguity if one or two root letters are long vowels or belong to the affixes and clitics' letters. The absence of short vowels can make morphological analysis even harder. For example, the word ولدينا *wldynā* has two possible morphological analyses, see figure 1.1. First, وَلَدَيْنَا *waladaynā* 'Our two sons' has the root ولد *w-l-d* 'descendant, offspring, child, son' and has three morphemes وَلَدَ *walada* 'son or boy', يْنَ *yna* 'dual letters', and ا *ā* 'our' nominative suffixed pronoun. Second, وَلَدَيْنَا *wa-ladaynā* 'and we have got' of the root لدى *l-d-y* has three morphemes; وَ *wa* 'and' is a conjunction proclitic, لَدَيْ *laday* 'have got' a perfect verb stem, and نَا *nā* 'we' a genitive suffixed pronoun. In this example, the interaction between the clitic letter and the underlying letter of the word increases the complexity of morphological analysis for Arabic text. The first letter of the word و *wa* is one of the underlying letters of the word in the first analysis and it can be analyzed as a conjunction letter as shown in the second analysis. Section 2.3.4.1 discusses the challenges of complex Arabic morphology. Sections 5.5 and 8.3.1.4 define our approach to defining the word's morphemes.

ولدينا *wldynā*	ا + يْنَ + وَلَدَ = وَلَدَيْنَا *waladaynā* ‘Our two sons’ has the root ولد *w-l-d* ‘descendant, offspring, child, son’
	نَا + لَدَيْ + وَ = وَلَدَيْنَا *wa-ladaynā* ‘and we have got’ of the root لدى *l-d-y*

Figure 1.1 Example of ambiguous Arabic word

Gemination is one of the orthographic issues that the morphological analyzer has to deal with correctly. Other orthographic issues of Arabic such as short vowels (َ ُ ِ) and gemination *šadda*^{h} (ّ) are: *hamza*^{h} (ء إ أ ؤ ئ), *tā’ marbūṭa*^{h} (ة) and *hā’* (ه), *yā’* (ي) and *’alif maqṣūrā* (ى) and *madda*^{h} (آ) or extension which is a compound letter of *hamza*^{h} and *’alif* (اأ). Chapter 2 discusses the morphological complexity of Arabic text.

1.5 Motivation and Objectives for this Thesis

Our research into morphological analysis of Arabic text corpora involves original scientific research, and focuses on the question of how to widen the scope of Arabic morphological analyses, to develop an NLP toolkit that can process Arabic text in a wide range of formats, domains, and genres, of both vowelized and non-vowelized Arabic text.

The inspiration behind this research is centuries-old linguistic wisdom and knowledge captured and readily available in traditional Arabic grammars and lexicons. The knowledge can be utilized in an Arabic NLP toolkit which can be accessed, standardized, reused and implemented in Arabic natural language processing. The detailed knowledge is applicable to both Classical and Modern Standard Arabic and can be used to restore orthographic (*e.g.* short vowels) and morphological features which signify important linguistic distinctions. Fine-grained morphological analysis is possible, achievable and advantageous in processing Arabic text. Enriching the text with linguistic analysis will maximize the potential for corpus re-use in a wide range of applications. We foresee the advantage of enriching the text with part-of-speech tags of very fine-grained grammatical distinctions, which reflect expert interest in syntax and morphology, but not specific needs of end-users, because end-user applications are not known in advance.

The objective of the thesis has been achieved through developing a novel language-engineering toolkit for morphological analysis of Arabic text, the SALMA – Tagger. The SALMA – Tagger combines sophisticated modules that break down the complex morphological analysis problem into achievable tasks which each address a particular problem and also constitute stand-alone units. These modules are:

- **The SALMA – Tokenizer** which tokenizes the input text files and identifies the Arabic words, spell-checks and corrects the words, and identifies the word’s parts or morphemes.

- **The SALMA – Lemmatizer and Stemmer** which extracts the lemma and the root of the analysed word.
- **The SALMA – Pattern Generator** which is responsible for matching the word with its pattern.
- **The SALMA – Vowelizer** which is responsible for adding the short vowels to the analysed words.
- **The SALMA – Tagger module** that predicts the fine-grained morphological features for each of the analysed word's morphemes.

These modules are useful as stand-alone tools which users can select and/or customise to their own applications.

The previously mentioned original Arabic NLP toolkit depends on two novel and original resources and proposed standards developed throughout this project. These are:

- **The SALMA – Tag Set,** the theory informing the morphological features tag set, and developed in this thesis, is to base the tag set on traditional morphological features as defined in long-established Arabic grammar, in a notation format intended to be compact yet transparent.
- **The SALMA – ABCLexicon**, a novel broad-coverage lexical resource constructed by extracting information from many traditional Arabic lexicons, constructed over 1200 years, of disparate formats.

An additional resource resulting from the construction the SALMA – ABCLexicon is the Corpus of Traditional Arabic Lexicons. The Corpus of Traditional Arabic Lexicons is a special corpus of Arabic which is compiled from the text of 23 traditional Arabic lexicons that cover a period of 13-hundred years and shows the evolution of Arabic vocabulary. It contains about 14 million word tokens and about 2 million word types.

In summary, this research has contributed to Arabic NLP in three dimensions: resources, proposed standards and tools (*i.e.* practical software). The following is a list of the contributions classified into the three dimensions:

A. Resources

1. The SALMA – ABCLexicon.
2. The Corpus of Traditional Arabic Lexicons.
3. The morphological lists of the SALMA – Patterns Dictionary and the SALMA – Clitics and Affixes lists.
4. The several linguistic lists that are used by the SALMA – Tagger such as: function words list, named entities lists, broken plural list, conjugated and non-conjugated verbs list, and transitive verbs lists.
5. The Lemmatized version of the Arabic Internet Corpus.

B. Proposed Standards

6. The SALMA – Tag Set.
7. The SALMA – Gold Standard for evaluating morphological analyzers for Arabic text.
8. The MorphoChallenge 2009 Qur'an Gold Standard.
9. Proposed standards for developing morphological analyzers for Arabic text.
10. Proposed standards for evaluating morphological analyzers for Arabic text.

C. Tools (practical software)

11. The SALMA – Tagger
12. The SALMA – Tokenizer
13. The SALMA – Lemmatizer and Stemmer
14. The SALMA - Vowelizer
15. The SALMA – Pattern Generator

Finally, a potential future application of using these contributions is as a language-engineering toolkit for Arabic lexicography to construct Arabic monolingual and bi-lingual dictionaries (Section 10.3).

1.6 Thesis Structure

This thesis is organized into five parts. Part I: Introduction includes Chapter 1. Part II: Background Review includes Chapters 2, 3, 4 and 5. Part III: Standards for Arabic Morphological Analysis includes Chapters 6 and 7. Part IV: Tools and Applications for Arabic Morphological Analysis includes Chapters 8, 9 and 10. Part V: Conclusions and Future Work includes Chapter 11. The following highlights the thrust of the work presented in this thesis:

- **Part I: Introduction** and Background Review includes:
  - **Chapter 1: Introduction** where the previous sections have given an introduction to the problems associated with studying morphological analysis in general and for Arabic text in particular. Section 1.5 discussed the motivations and objectives for this thesis. It also summarized the original contributions to the Arabic NLP field of study.**Chapter 2: Literature Review: Morphological Analyses of Arabic Text** presents coverage of background and literature surveys relevant to the research. First, a survey of Arabic text corpora is discussed in section 2.2. Second, a literature survey of morphological analysis in general and morphological analysis for Arabic text in particular is discussed in section 2.3. This section presents the general methodologies of morphological analysis and those which have been applied to Arabic text. It also surveys the existing key

morphological analyzers for Arabic text and discusses their attributes. Third, a survey of part-of-speech taggers for Arabic text is presented in section 2.4. It comparatively evaluates existing part-of-speech taggers for Arabic text.

- **Part II: Background Analysis and Design** includes:
  - **Chapter 3: Comparative Evaluation of Arabic Morphological Analyzers and Stemmers** surveys stemming algorithms for Arabic text used in the comparative evaluation in section 3.2. Then it discusses four different fair and precise evaluation experiments using a gold standard for evaluation in sections 3.4 and 3.5. Finally, it presents an analytical study of the triliteral Arabic roots in section 3.7.
  - **Chapter 4: The SALMA-ABCLexicon: Prior-Knowledge Broad-Coverage Lexical Resource to Improve Morphological Analyses** surveys morphological lexicons for Arabic and other languages in section 4.1. Traditional Arabic lexicons and lexicography are presented in section 4.2. Twenty-three traditional Arabic lexicons are listed and and classified according to their ordering methodology in section 4.3. The construction methodology of the SALMA – ABCLexicon using the traditional Arabic lexicons and its evaluation are discussed in sections 4.4 and 4.5. The Corpus of Traditional Arabic Lexicons is described in section 4.6.
  - **Chapter 5:** The survey of Arabic Morphosyntactic Tag Sets and Standards for Designing the SALMA Tag Set presents existing part-of-speech tagging systems and tag sets for Arabic text in sections 5.2 and 5.3. Section 5.4 discusses the morphological features in Tag Set design criteria.
- **Part III: Proposed Standards for Arabic Morphological Analysis** includes:
  - **Chapter 6: The SALMA Tag Set** analyzes 22 morphological features of Arabic word morphemes. It defines the attributes of each morphological feature by identifying their characteristics and deciding which attributes are used for the analysis of specific morphological categories.
  - **Chapter 7: Applying the SALMA Tag Set** explores the evaluation methodologies of the SALMA – Tag Set in section 7.3. A practical application of the SALMA – Tag Set has been achieved by mapping from the Quranic Arabic Corpus morphological tag set in section 7.4. The evaluation of the mapping process is reported in section 7.5 and discussed in section 7.6.
- **Part IV: Tools and Applications for Arabic Morphological Analysis** includes:
  - **Chapter 8: The SALMA Tagger for Arabic Text** discusses morphological analysis for Arabic text. It presents standards for developing a robust morphological analyzer for Arabic text based on our experiences in participating in two contests for developing morphological analyzers for Arabic text: the ALECSO/KACT initiative and MorphoChallenge 2009 competition (section 8.2).

The SALMA – Tagger algorithm is described in section 8.3. The SALMA – Tagger is decomposed into sophisticated modules that break down the complex morphological analysis problem into achievable tasks so they solve particular problems and are useful in their own right. These modules are: The SALMA – Tokenizer; the SALMA – Lemmatizer and Stemmer; and the SALMA – Pattern Generator. A rule-based system for predicting the morphological features of Arabic word morphemes is discussed in section 8.4. Finally, standard output formats of the SALMA – Tagger are described in section 8.5.

  - **Chapter 9: Evaluation for the SALMA – Tagger** depends on developing agreed standards for evaluating morphological analyzers for Arabic text, based on our experiences and participation in two evaluation contests: the ALECSO/KACT initiative for developing and evaluating morphological analyzers; and the MorphoChallenge 2009 competition, section 9.2. The construction of a reusable general purpose gold standard (the SALMA – Gold Standard) for evaluating the SALMA – Tagger and morphological analyzers for Arabic text in general is described in sections 9.4 and 9.5. Sections 9.6 and 9.7 discuss the process of evaluating the SALMA – Tagger using gold standards. Evaluation metrics are discussed and the results of the evaluation reported. The discussion of the results analyzes the prediction process, the challenges and suggestions for improvement for each morphological feature category in section 9.8.
  - **Chapter 10: Practical Applications of the SALMA Tagger** describes two practical applictions for applying the resources, standards, and tools developed in this thesis. The first application was achieved by lemmatizing the 176-million word Arabic Internet Corpus, section 10.2, and an exemplar for using the resources, standards and tools is as a language-engineering toolkit for Arabic lexicography to construct Arabic monolingual and bi-lingual dictionaries, in section 10.3.
- **Part V: Conclusions and Future Work** includes:
  - **Chapter 11: Conclusions and Future Work** summarizes the conclusions of this thesis. It reviews the motivations and objectives for this thesis and lists the main contributions and their impact on Arabic NLP. The second part of the chapter discusses future work that can be done to improve the developed resources, standards and tools. It also shows example projects of higher NLP applications that can benefit directly from our contributions and from our research interests.

Chapter 2
Literature Review: Morphosyntactic Analysis of Arabic Text

2.1 Introduction

This chapter surveys existing morphosyntactic analysis systems for text corpora. The survey studies these systems in three dimensions. First, it explores Arabic text corpora as a background prerequisite for morphosyntactic analysis. Second, it studies morphological analysers for text corpora concentrating on methodologies, challenges, examples of existing morphological analysers, and evaluation standards. Third, it surveys part-of-speech tagging technology and existing part-of-speech taggers for Arabic text.

Arabic corpora started to appear in the late 1980s. Most of the existing Arabic corpora are of MSA written text, mainly newspaper text. Only two corpora are open-source and available to download. These are the Corpus of Contemporary Arabic (CCA) (Al-Sulaiti and Atwell 2006) and the Quranic Arabic Corpus (QAC) (Dukes, Atwell and Sharaf 2010; Dukes and Habash 2010). The CCA represents MSA and contains 1 million words of raw text, and the QAC represents Classical Arabic and consists of the Qur'an text of about 80,000 words. The QAC is enriched with morphological and syntactic annotation layers. Section 2.2 surveys existing Arabic corpora.

Several morphological analysers for Arabic text exist. Morphological analysis is an important pre-processing step for many text analytics applications. The aim of morphological analysis is to define words in a corpus in terms of morphosyntactic information such as: (i) information about the word structure (*i.e.* root, affixes, clitics, patterns and vowelization); (ii) part-of-speech of the word (*i.e.* noun, verb and particle) (iii) part-of-speech subcategories of the word (*e.g.* gerund, noun of place, active participle, generic noun, proper nouns, pronouns, perfect verb, imperfect verb, imperative verbs, prepositions, etc.); and (iv) the morphological features of the word (*e.g.* Gender, Number, Person, Case or Mood, Transitivity, Rational, Number of root letters, etc.). The information resulting from morphological analysers can be used in different levels of NLP applications. Section 2.3 surveys morphological analysis of text corpora focusing on its approaches, applications, the specific definition of morphological analysis for Arabic text, challenges of Arabic morphology, and morphological analysis of both Classical and MSA text. It also surveys state of the art morphological analysers and evaluation methodologies.

Morphological analysers are designed to generate all possible analyses of the analysed words out of their context. Disambiguating the analysis to suit the context is done by using part-of-speech taggers. Section 2.4 surveys part-of-speech technology. It lists state of the art part-of-speech taggers for English, the tagged corpora and the standards. The section surveys existing part-of-speech taggers for Arabic text. It briefly lists existing part-of-speech taggers, their development approaches and their accuracy as reported by their developers.

2.2 Arabic Corpora

Arabic corpora started to appear in the late 1980s; the following list of Arabic corpora developed from (Al-Sulaiti and Atwell 2006) outlines their size, type, purpose of development and the materials used to develop them:

- **Buckwalter Arabic Corpus** (1986-2003) consists of about 3 million words of public resources on the web to be used in lexicography.
- **Leuven Corpus** (1990-2004) developed at the Catholic University of Leuven, Belgium, consists of about 3 million words of written and spoken text from internet sources, radio and TV and primary school books, to be used in the development of Arabic-Dutch /Dutch-Arabic learner's dictionaries.
- **Arabic Newswire Corpus** (1994) developed at the University of Pennsylvania LDC, consists of 80 million words of written text collected from Agence France Presse (AFP), Xinhua News Agency, and Umma Press, to be used in education and the development of technology.
- **CALLFRIEND Corpus** (1995) developed at the University of Pennsylvania LDC. This corpus comprises 60 telephone conversations by Egyptian native speakers, to be used in the development of language identification technology.
- **Nijmegen Corpus** (1996) developed at Nijmegen University consists of over 2 million written words collected from magazines and fiction, to be used in Arabic-Dutch / Dutch-Arabic dictionaries.
- **CALLHOME Corpus** (1997) developed at the University of Pennsylvania LDC, consists of 120 telephone conversations of Egyptian native speakers, to be used in telephony and speech recognition.
- **CLARA** (1997) developed at Charles University, Prague, consists of 50 million words collected from periodicals, books, internet sources from 1975-present, to be used for lexicography.
- **Egypt** (1999) developed at John Hopkins University, a parallel corpus of the Qur'an in English and Arabic to be used in machine translation.
- **Broadcast News Speech** (2000) developed at University of Pennsylvania LDC, consists of more than 110 News broadcasts from the Voice of America radio station, to be used in speech recognition.

- **DINAR Corpus** (2000) developed at Nijmegen University and SOTETEL-IT, in co-ordination with Lyon2 University, consists of 10 million words, to be used in lexicography, general research, and NLP.
- **An-Nahar Corpus** (2001) developed by ELRA, consists of 140 million words of written text collected from An-Nahar newspaper (Lebanon), to be used in general text research.
- **Al-Hayat Corpus** (2002) developed by ELRA consists of 18.6 million words of written text collected from Al-Hayat newspaper (Lebanon), to be used for language engineering and information retrieval applications.
- **Arabic Gigaword** (2002) developed at the University of Pennsylvania LDC, consists of around 400 million words collected from Agence France Press (AFP), Al-Hayat news agency, An-Nahar news agency and Xinhua news agency, to be used in natural language processing, information retrieval and language modelling.
- **E-A Parallel Corpus** (2003) developed at the University of Kuwait, consists of 3 million words of written text collected from publications from Kuwait National Council, to be used in teaching, translation and lexicography.
- **General Scientific Arabic Corpus** (2004) developed at UMIST, UK, consists of 1.6 words of written text, to be used in investigating Arabic compounds.
- **Classical Arabic Corpus** (CAC) (2004) developed at UMIST, UK, consists of 5 million words of written text, to be used in lexical analysis.
- **Multilingual Corpus** (2004) developed at UMIST, UK, consists of 11.5 million words of written text including 2.5 million words in Arabic, collected from IT-specialized websites-computer system and online software help-one book, to be used in translation studies.
- **SOTETEL Corpus** developed at SOTETEL-IT, Tunisia, consists of 8 million words of written text collected from literature, academic and journalistic materials, to be used in lexicography.
- **Corpus of Contemporary Arabic (CCA)** (2004) developed at the University of Leeds, consists of 1 million words of written and spoken data, collected from websites and online magazines, to be used in language teaching and language technology.
- **DARPA Babylon Levantine Arabic Speech and Transcripts** (2005) developed at the University of Pennsylvania LDC, consists of about 2000 telephone calls collected from Fisher style telephone speech collection, to be used in machine translation, speech recognition and spoken dialogue systems.
- **The Penn Arabic Treebank** (2001) Part 1 consists of 166,000 words of written Modern Standard Arabic newswire from the Agence France Presse corpus; and Part 2 consists of 144,000 words from Al-Hayat distributed by Ummah Arabic News

Text, to be used in computational linguistics. New features of annotation in the UMAAH (UMmah Arabic Al-Hayat) corpus include complete vocalization (including case endings), lemma IDs, and more specific part-of-speech tags for verbs and particles. The Arabic Treebank corpora are annotated for morphological information, part-of-speech, English gloss (all in the "part-of-speech" phase of annotation), and for syntactic structure (Maamouri and Bies 2004).

- **The Quranic Arabic Corpus** (2009) contains the classical Arabic source text of the Quran, the holy book of Islam. The text consists of nearly 80,000 words, divided into numbered chapters and verses. The text is being enriched with morphological analysis, Part-of-Speech tagging, dependency parsing, coreference resolution, and other linguistic markup, via a collaborative web-based project. The annotated corpus is online, used by Quranic scholars, linguists, and the general public with an interest in Islam.

Nearly all these corpora have been collected by Arabic corpus linguistics research groups for their own purposes, and are not freely downloadable. The Corpus of Contemporary Arabic (CCA) developed at the University of Leeds (Al-Sulaiti and Atwell 2004; Al-Sulaiti and Atwell 2005; Al-Sulaiti and Atwell 2006), is the only freely available corpus on the web which has been widely reused for linguistic research. But it has not been annotated by part-of-speech tags. The only annotated corpus of the Arabic language used widely in computational linguistics research is the Penn Arabic Treebank (Maamouri and Bies 2004) developed at the University of Pennsylvania and distributed (at cost) by LDC Linguistic Data Consortium. The Quranic Arabic Corpus, developed recently, is starting to be used in tagging and parsing research.

2.3 Morphological Analysis for Text Corpora

Morphology is the study, identification, analysis and description of the minimal meaning bearing units (morphemes) that constitute a word. Morphological analysis is the process of categorizing and building a representative structure of the component morphemes where both orthographic rules and morphological rules are important for categorizing a word's morphemes. For instance, the plural of *party* is *parties* where orthographic rules indicate changing the *–y* to *-i-* and adding *–es*. And morphological rules tell us that *fish* has null plural (Jurafsky and Martin 2008).

Automatic morphological analysis started in the 1950s to support machine translation systems. The Porter stemmer (Porter 1980) is an example early morphological analysis system which is widely used in information retrieval applications. Automatic morphological analyses are beneficial for many early developed applications such as spelling correction, text input systems and text-to-speech synthesis. There was little interest in evaluating the correctness of results obtained by morphological analysers in early applications. The concern was on the soundness of the results rather than the methods (Roark and Sproat 2007).

Finite-state methodology has been dominant since the 1980s. The Finite-state approach for automatic morphological analysis was originally investigated at Xerox and the first practical application was due to Koskenniemi (Koskenniemi 1983); this has been used to develop wide-coverage morphological analysers for several languages. Two main approaches for computational morphology are: explicitly finite-state approaches which are based on a finite-state model and morphotactics, and integrating finite-state morphology and phonology, with unification of morphosyntactic features (Roark and Sproat 2007).

Morphological analyzers have been developed for a wide range of languages; the following are some examples. *EMERGE*[1] is a morphological analyzer for Spanish. It analyzes words and shows their canonical form, grammatical category and the inflection or derivation they come from. *ExtraLink* is an information extraction (IE) system and automatic hyperlinking that uses ontologies to define the relationships. Its IE system is *SProUT*[2], a generic multilingual shallow analysis platform, which can process English, German, Italian, French, Spanish, Czech, Polish, Japanese, and Chinese. It has modules for tokenization, morphological analysis, and named entity recognition. *FLEMM*[3] is a rule-based program (lemmatizer) for French that performs flexional morphological analysis for a tagged text using the Brill Tagger or TreeTagger, and extracts the lemma of words. It uses a small lexicon of 3,000 entries to handle exceptions. *FreeLing*[4] is a library that provides language analysis services for Spanish, English, and Catalan such as tokenizing, sentence splitting, morphological analysis, NE detection, date/number/currency recognition, PoS tagging, and chart-based shallow parsing. *POSTAG*[5] is morphological analysis plus part-of-speech tagging with morpheme dictionary for Korean. *ROSANA*[6] (RObust Syntax-based ANAphor resolution) is a coreference resolution system for English text. It identifies co-referring of anaphoric expressions such as third person pronouns, possessives, reflexives, common nouns, and names. *TWOL*[7] is a two-level morphological analysis tools for English, German, Swedish, Finnish, Danish, and Norwegian. *XeLDA*[8] is a framework that provides a general-purpose

[1] EMERGE http://protos.dis.ulpgc.es/morfolog/morfolog.htm
[2] SProUT http://sprout.dfki.de/
[3] FLEMM http://www.univ-nancy2.fr/pers/namer/Telecharger_Flemm.htm
[4] FreeLing http://www.lsi.upc.edu/~nlp/freeling
[5] POSTAG http://nlp.postech.ac.kr/DownLoad/k_api.html
[6] ROSANA http://www.stuckardt.de/rosana.htm
[7] TWOL http://www.lingsoft.fi/
[8] XeLDA http://www.mkms.xerox.com/

text retrieval system which includes several language processing operations such as: language identification; tokenization; morphological analysis; part-of-speech disambiguation; noun phrase extraction; contextual dictionary lookup; idiomatic expression recognition; relational morphology; and shallow parsing. It supports processing for text of several languages (Dutch, English, French, German, Italian, Portuguese, Spanish, Czech, Hungarian, Polish, Russian, Danish, Swedish, Finnish Norwegian, and Chinese) and other languages in development (Czech, Arabic, Japanese and Korean). It also includes bilingual dictionaries of English, French and German to English, French, German, Italian and Spanish.

2.3.1 Approaches to Morphological Analysis

The two-level formalism is the most widely used theoretical approach to morphological analysis. It is based on construction of a collection of finite-state transducers which each implement a particular morphological rule. The transducers attempt to map between the surface and the lexical realizations of a given morpheme. The main drawbacks of this approach are: it is language dependent and it needs manual construction of the transducers for each language which makes developing a morphological analyzer very costly and time consuming (Pauw and Schryver 2008). The minimum requirements for building a morphological analyzer using the two-level formalism approach are as follows. First, it requires a lexicon of stems and affixes together with basic information about them. Second, it is informed by morphotactics where the model of morpheme ordering is explained and the relations between morpheme classes inside a word are determined. Third, orthographic rules that govern the spelling of the word are used to model the changes that occur in a word (Jurafsky and Martin 2008).

Corpus-based approaches to morphological analysis use morphologically annotated corpora to build a morphological database rather than depending on linguistic knowledge. For example, CELEX is a lexical database for English, Dutch and German. It contains detailed information on orthography and phonology such as phonetic transcription of variant pronunciations, syllable structure and primary stress. CELEX morphology includes derivational and compositional structure and inflexional paradigms. Syntactic information includes word class, word class-specific subcategorizations and agreement structure. It also contains information about word frequency such as word and lemma counts based on representative text corpora (Baayen, Piepenbrock and Rijn 1995).

Corpus-based approaches to building morphological analysis can be used to provide a morphological database that is used in statistical processing and machine-learning techniques to morphological analysis. Statistical processing and machine-learning techniques are language independent, so in principle they can be ported to new domains and languages. Moreover, data-driven approaches to morphological analysis can outperform manually constructed rule-based analyzers (Pauw and Schryver 2008).

Recently, unsupervised approaches to morphological analysis have been explored, based on using minimum-distance edit metrics and pattern-matching techniques to automatically guess the morphological properties of a language on the basis of raw, unannotated text (Pauw and Schryver 2008). The unsupervised morpheme analysis contest MorphoChallenge is a challenge to design a statistical machine-learning algorithm for morphological analysis. The challenge has been run 5 times since 2005. The next section gives more detail about MorphoChallenge 2009 in particular.

2.3.2 MorphoChallege Competition

The MorphoChallenge task is to develop an unsupervised learning algorithm which can return the morpheme analyses of each word given lists of words of several languages; for Morphochallenge 2009 these were Arabic, English, Finish, German and Turkish. The preferred algorithm needs to be as language independent as possible. All words in the training corpus occur in sentences, so the algorithm might utilize information about word context (Kurimo, Virpioja and Turunen 2009).

The training corpora were 3 million sentences for English, Finnish and German, and 1 million sentences for Turkish in plain unannotated text files. The training corpus for Arabic was the Quran, which is a small corpus consisting of only 78K words. The text of the Qur'an corpus is available in both vowelized and non-vowelized formats. For Arabic, the participants could test their algorithms using the vowelized words or the unvowelized, or both. The algorithms were separately evaluated against the vowelized and the non-vowelized gold standard analyses. For all Arabic data, the Arabic writing scripts were provided as well as the Roman script (Buckwalter transliteration), see figure 9.1. However, only the morpheme analysis submitted in Roman script, was evaluated (Kurimo et al. 2009).

In Competition 1 the proposed unsupervised morpheme analyses were compared to the correct grammatical morpheme analyses called here the linguistic gold standard. The gold standard morpheme analyses were prepared in exactly the same format as the result file the participants were asked to submit: alternative analyses separated by commas. For Arabic the gold standard had in each line: the word, the root, the pattern and then the morphological and part-of-speech analysis (Kurimo et al. 2009). Section 9.3 discusses the MorphoChallenge competition as a standard for evaluating morphological analyzers.

Twelve algorithms were evaluated against the Arabic Qur'an gold standard. The evaluation results for Arabic turned out to be quite surprising, because most algorithms gave rather low recall and F-measure and the simple "letters" reference outperformed all other participating algorithms; see section 9.3.1 for the definitions of the accuracy measures. “Promodes” and “Ungrade” methods scored clearly better than the rest of the participants in Arabic. Tables 2.1 shows the evaluation results for the twelve algorithms compared to the gold standards of non-vowelized as reported by (Kurimo et al. 2009).

Table 2.1 The submitted unsupervised morpheme analysis compared to the Gold Standard in non-vowelized Arabic (Competition 1).

AUTHOR(S)	METHOD	PRECISION	RECALL	F-MEASURE
-	letters	70.48%	53.51%	60.83%
Spiegler et al.	PROMODES 2	76.96%	37.02%	50.00%
Spiegler et al.	PROMODES committee	77.06%	36.96%	49.96%
Spiegler et al.	PROMODES	81.10%	20.57%	32.82%
Golénia et al.	UNGRADE	83.48%	15.95%	26.78%
Virpioja & Kohonen	Allomorfessor	91.62%	6.59%	12.30%
-	Morfessor Baseline	91.77%	6.44%	12.03%
Bernhard	MorphoNet	90.49%	4.95%	9.39%
Monson et al.	ParaMor-Morfessor Union	93.72%	4.81%	9.14%
Monson et al.	ParaMor-Morfessor Mimic	93.76%	4.55%	8.67%
Lavallée & Langlais	RALI-ANA	92.40%	4.40%	8.41%
Tchoukalov et al.	MetaMorph	95.05%	2.72%	5.29%
Monson et al.	ParaMor Mimic	91.29%	2.56%	4.97%
Lavallée & Langlais	RALI-COF	94.56%	2.13%	4.18%

2.3.3 Applications of Morphological analysis

Morphological analysis has many applications throughout speech and language processing. Morphological analysis techniques form the basis of most natural language processing systems (Kiraz 2001; Al-Sughaiyer and Al-Kharashi 2004; Jurafsky and Martin 2008; Pauw and Schryver 2008). Such applications are:

- **Searching the Web:** In web searching for morphologically complex languages, morphological analysis enables searching for the inflected form of the word even if the search query contains only the base form.
- **Part-of-speech taggers:** Morphological analysis gives the most important information for a part-of-speech tagger to select the most suitable analysis for a given context.
- **Dictionaries and Spell-checkers:** Dictionary construction and spell-checking applications rely on a robust morphological analysis.
- **Machine translators:** Machine translation systems rely on highly accurate morphological analysis to specify the correct translation of an input sentence (Jurafsky and Martin 2008).
- **Lemmatizers:** lemmatization is part of morphological analysis. Google’s search facilities use lemmatization to produce hits of all inflectional forms of the input word. Statistical models of language in machine translation and speech recognition also use

lemmatization. Lexicographic applications use lemmatizers as an essential tool for corpus-based compilation (Pauw and Schryver 2008).

- **Other applications:** morphological analysis is useful for many applications, such as information retrieval, text categorization, dictionary automation, text compression, data encryption, vowelization and spelling aids, automatic translation, and computer-aided instruction (Al-Sughaiyer and Al-Kharashi 2004).

2.3.4 Morphological Analysis for Arabic Text

Morphological analysis is the process of assigning the morphological features of a word such as: its root or stem, the morphological pattern of the word, the morphological attributes of the word (part-of-speech of the word whether it is noun, verb or particle). It also involves specifying the number of the word (singular, dual or plural), and the case or mood (nominative, accusative, genitive or jussive). Moreover, it identifies the internal structure of the word such as prefixes, suffixes, clitics and the root or stem (Thabet 2004); see sections 1.2 for general definition of morphology and morphological analysis.

Hamada (2009), also Hamada (2010) defined morphological analysis of Arabic text as a series of processes. Morphological analysis for Arabic text includes extracting the root of the analyzed word, deriving all possible derivatives of a given root, analyzing the words into their morphemes, distinguishing the stem of the word by separating its prefixes and suffixes and stripping the conjugated or inflectional affixes of the word.

Habash (2010) distinguished between two types of approaches to morphology: form-based morphology and functional morphology. The morpheme as the smallest meaningful unit in a language is the central concept in form-based morphology. However, the central concept of functional morphology is the study of words and morphemes in terms of their morpho-syntactic and morpho-semantic behaviour in context. (Habash 2010) defined morphological analysis as the process of determining all possible morphological analyses of the orthographic word. This process includes identifying the main part-of-speech of the analyzed word. The morphological analysis is either form-based where the word's morphemes are identified or based on functional morphology where the functions (grammatical features) of each morpheme are determined.

The previous definitions of morphological analysis for Arabic text agree with the general definition of computational morphology in section 1.2. A pragmatic definition of morphological analysis for Arabic is computer applications that analyze Arabic words of a given text and deal with their internal structure. This involves a series of processes that identify all possible analyses of the orthographic word. These processes are both form-based and function-based. Orthographic words can be fully-vowelized, partially-vowelized or non-vowelized. They also can be Classical Arabic or Modern Standard Arabic.

Form-based analysis deals with the orthographic word to identify its morphemes. These processes include tokenization, spell-checking, stemming and lemmatization, pattern matching and diacritization. Function-based processes deal with identifying the morphosyntactic features and functions of the word. These processes include predicting the morphological features of the word's morphemes, part-of-speech tagging and parsing.

The following subsections survey Arabic morphological analysis. The first subsection explores the challenges for Arabic morphological analysers. The second subsection defines basic related concepts which are used throughout this thesis. The third and fourth subsections discuss morphological analysis of Classical and Modern Standard Arabic respectively. The fifth subsection surveys the approaches for morphological analysis development. The sixth subsection discusses the requirements of developing Arabic morphological analysers. The seventh subsection surveys existing morphological analysis systems for MSA text. The last subsection gives an example of a community-based approach for evaluating Arabic morphological analysers, the ALECSO/KACST initiative for developing and evaluating morphological analysers for Arabic text; see also section 8.2.

2.3.4.1 Challenges of Arabic Morphology

Arabic is a morphologically complex and highly inflectional language. Its root-pattern nonconcatenative (*i.e.* nonlinear) morphology makes both theoretical and computational processing tasks for Arabic text extremely hard. Morphological analysis of Arabic text affects higher level applications such as part-of-speech tagging and parsing. It affects both syntactic and phonological levels of analysis (Beesley 1996; Al-Sughaiyer and Al-Kharashi 2004; Smrz 2007; Soudi et al. 2007; Attia 2008; Habash 2010). Chapter 8 discusses practical solutions for these challenges as implemented in the SALMA – Tagger. Here is a list of major challenges that face Arabic morphological analysis:

1- **The orthography of Arabic**: the orthography of Arabic is based on standard Arabic script. The Arabic alphabet consists of: 25 consonants; 6 vowels divided into three long vowels (ا ، و ، ي) (*ā*, *w*, *y*) and three short vowels written as diacritics (◌َ ، ◌ُ ، ◌ِ) (*a*, *u*, *i*); and a glottal stop *hamza^h*. In addition, the writing system for Arabic contains other shapes of letters such as *'alif maqṣūra^h* (ى). Arabic letters change their shape according to their position in the word as Arabic script requires connection of the word's letters. Other orthographic issues in Arabic are the use of diacritics above or below letters. These diacritics include *sukūn* (◌ْ) to mark silent letters (*i.e.* absence of short vowel); and gemination or incorporation[9] *šadda^h* (◌ّ) to indicate a doubled letter; and *tanwīn* (◌ً ، ◌ٌ ، ◌ٍ) the syntactic case mark of indefinite singular nouns. *hamza^h* has 5 shapes (ء أ إ ؤ ئ). *tā' marbūṭa^h* (ة) shares phonetic properties of the two consonants *tā'* (ت) and *hā'* (ه) and is used to mark feminine singular nouns. *madda^h* (آ) or extension is a compound letter of *hamza^h* and *'alif* (ءا).

2- **Nonconcatenative nature**: the rich "root-and-pattern" nonconcatenative (or nonlinear) morphology results in a highly complex word formation process of roots and patterns. Hundreds of words can be derived from a single root by following certain patterns. These patterns are abstract templates where root radicals (*i.e.* mostly triliteral roots) and vocalism (*i.e.* short vowels) are inserted in certain positions within the pattern. The pattern also has *prefixed* letters appearing before the position of the first root radical; *suffixed* letters appearing after the position of the last root radical; and *infixed* letters appearing between the root radicals. Patterns transmit morphological and semantic features to the derived words. During the derivation process changes might occur to the original root letters such as assimilation, elision and gemination. Broken plurals exemplify the nonconcatenative nature of Arabic (Clark 2007). For example, the plural form of the word قَلْب *qalb* 'heart' is قُلُوب *qulūb* 'hearts' and this is formed by adding the letter و *wāw* as an infix between the second and the third radicals. And the plural form of the word مِصْبَاح *miṣbāḥ* 'light' is مَصَابِيْح *maṣābīḥ* which is formed using the special pattern of broken plural مَفَاعِيْل *mafā'īl* that re-arranges the root radicals and the infixes. This "root and pattern" morphology also

[9] Gemination or incorporation are used in the thesis to indicate a doubled letter which usually marked by *šadda^h* (◌ّ) in vowelized text. *šadda^h* does not appear in non-vowelized text. Therefore, the absence of *šadda^h* represents a challenge to morphological analyzers for Arabic text.

brings problems for western linguistic terminology. A “morpheme” in Western traditions is an indivisible “atomic” lexical unit, and the “stem” is the core morpheme of a word. In Arabic, the “stem” combines root and pattern. In this thesis, we refer to stem as a morpheme, but purists may argue a stem is really 2 morphemes – root and pattern.

3- **Arabic clitics**: clitics and affixes of Arabic words are productive. Clitics are conjunctions, prepositions, particles, and genitive suffix-pronouns that are attached to the beginnings and at the ends of words. According to our classification into clitics or affixes as explained later in sections 8.3.1.4 and 8.3.1.5, the definite article is classified as a proclitic rather than a prefix because the definite article is not part of the pattern even though it cannot appear as a stand-alone word. Therefore, storing word forms in a dictionary and doing morphological analysis by dictionary lookup is not possible, as we cannot list all morphological variants of every Arabic word. Thus, morphological analysis done dynamically is unavoidable. A word such as بِوَالِدَيْهِ *bi-wālidayhi* ‘in his parents’ consists of four morphemes بِ *bi* ‘in’ is a preposition, وَالِد *wālida* ‘parent’ is the noun stem, يْ y ‘two’ is a dual letter, and هِ *hi* ‘his’ is object relative pronoun. The proclitic بِ *bi* ‘in’ and the enclitic هِ *hi* ‘his’ are productive clitics.

4- **High degree of ambiguity**: Arabic also has a high degree of ambiguity for many reasons such as:

   a. **Assimilation or elision of vowels**: the presence of long vowels in some root radicals causes these weak radicals to be deleted or changed during the derivation process. For example, the weak radical و *wāw* of the root قول *q-w-l* is changed into another vowel or is deleted according to vocalic environment. It is changed into ا *’alif* in the past verb قَالَ *qāl* ‘he said’; and into ي *yā’* in the passive past verb قِيلَ *qīla* ‘it is said’; and deleted in the first person past verb قُلْتُ *qultu* ‘I said’.

   b. **Interaction between affix or clitic letters and the root radicals**: word affixes and clitics can be homographic with the underlying letters of the word which means the morphological analyzer must deal with words whose clitics and affixes interact with the underlying letters by producing all possible analyses of these words. For example, the word بِطَاقَات *biṭāqāt*; can have two possible analyses. One way is to treat the first letter of the word as a prepositional proclitic بِ *bi* "with", where the root is ط-و-ق *ṭ-w-q* and it means 'with the abilities'.The other way is to treat the first letter as an underlying letter where the root is ب-ط-ق *b-ṭ-q* and it means 'cards', where it has no clitic or prefix. Section 8.2.3.2 gives more examples.

c. **Tokenization[10] (*i.e.* segmentation) of words into their morphemes** where word tokens out of context can be segmented into different sequences of morpheme tokens. Therefore, morphological analyzers need to investigate all possible variants correctly for words out of context. Morphemes such as ت *tā'* can be attached to verbs to indicate second person masculine subject or second person feminine subject. For example, the ت *tā'* morpheme of the word فرمت *frmt* can be analyzed as: فَرَمْتَ*faramta* 'you (2MS) chopped'; or فَرَمْتِ *faramti* 'you (2FS) chopped'. The same form can involve one morpheme فَرْمَتَ *farmata* 'he formatted' which represents a foreign word; or three morphemes فَرُمْتَ = ف + رم + ت *farumta* 'you (2MS) desired' which has the root روم *r-w-m*; or فَرَمَتْ = ف + رم + ت *faramat* 'she (3FS) threw' from the root رمي *r-m-y*.

d. **Extracting the root letters of the word**: root letters can be hard to extract or predict and increase the text ambiguity if the one or two root letters are long vowels or belong to the affixes and clitics letters. For example, the form يسر *ysr* involves two roots: يسر *y-s-r* where the word يَسِر *yasir* means 'ease or prosperity'; and سرر *s-r-r* where the word يَسِرُّ *yasirru* means 'he tells a secret'. Moreover, assimilation or elision occurring on root radicals or affix letters increases the complexity of root extraction algorithms especially those that assume letters which are not shared with clitic and affix letters are original root radicals. For example, the letter ط *ṭa*^{h} of the word اصْطَدَمَ *'iṣṭama* 'impact' which has the root صدم *ṣ-d-m*, will be treated as a root radical, where it has changed from the underlying letter ت *tā'* of the pattern افْتَعَلَ*'ifta'ala*.

[10] Tokenization refers to both word tokenization and morpheme tokenization throughout the thesis

e. **The omission of short vowels especially in MSA text**: will affect the functional behaviour and the part-of-speech classification of words. For example, ورد*wrd*: can be وَرْدٌ *ward^{un}* “roses” representing a noun or وَرَدَ *warada* “to come” representing a verb; رب *rb*: رَبٌّ *rubb^{un}* “God” is a noun, while رُبَّ *rubba* “many” is a particle;. A non-vowelized word can be noun, verb and particle. Thus بل *bl*; بَلٌّ *ball^{un}* “moistening” is a noun; بَلَّ *balla* “to moisten, wet, make wet” is a verb; بَلْ *bal* “nay, -rather …, (and) even, but, however, yet” is a particle.

5- **Phonology, morphology and syntax**: morphology interacts with phonology and syntax. Phonology deals with phonemes which are sound units smaller than morphemes, and syntax deals with rules of composing sentences by combining words. Phonological processes cannot be separated from morphology. Therefore, morphological analyzers need to deal with the different kinds of phonological processes such as assimilation, syncope or deletion, epenthesis or insertion, and gemination or doubling. Syllabification is a well-studied phonological phenomenon in English dictionaries, but it is not established in Arabic dictionaries. On the other hand, syntax interacts significantly with morphology such that many words require contextual knowledge to solve their morphological ambiguities. In conclusion, morphological analysis modules must account for phonology and syntax which increases the complexity of developing morphological analysis systems for Arabic text (Kiraz 2001).

6- **Punctuation**: punctuation has been introduced recently into the Arabic writing system. MSA text is characterized by inconsistency and irregularity in the use of punctuation marks. In addition to the late introduction of punctuation to MSA text, the absence of a comprehensive treatment of punctuation in Arabic grammar books increases the problem of inconsistency in the use of punctuation in MSA text. Moreover, the use of punctuation in Arabic text is prescriptive rather than based on a linguistic description of actual usage in authentic written samples (Khafaji 2001; Attia 2008). Punctuation plays a significant part in phrase break prediction for English, and serves as an input to the classifier along with POS tags in both rule-based (Liberman and Church 1992) and probabilistic (Taylor and Black, 1998; Ingulfsen et. al, 2005) approaches.

2.3.4.2 Basic Concepts of Arabic Morphological Analysis

This section defines the basic concepts related to Arabic morphological analysis. These terms will be used in this thesis according to these definitions. Some of them are drawn from Wikipedia, as although Wikipedia is not an authoritative academic source, it is a widely-used explanatory source.

- **Tokenization or segmentation**: is the process of defining the word’s morphemes. These morphemes can be classified into 5 types: proclitics, prefixes, stem, suffixes and enclitics. A word must have at least one stem morpheme. Combinations of clitics and affixes can be attached to the word. A morphological analyzer is responsible for defining all possible variations of segmenting a word into its morphemes.
- **Stemming**: is the process of assigning morphological variants of words to equivalence classes, such that each class corresponds to a single stem. It is also defined as reducing inflected words to their stem, base, or root form[11]. For example words such as *writing, write, writer* and *written* are reduced to the root *write*. For distinguishing between stem and root in Arabic – see note 2 on section 2.3.4.1.
- **Lemmatization**: is the process of grouping a set of words into the canonical form, dictionary form, or citation form which is also called the *lemma*. *E.g.*, in English, *run, runs, ran* and *running* are forms of the same lexeme, with *run* as the lemma[12]. The lemma is usually also the stem.
- **Root:** is the smallest lexical unit. An Arabic root usually consists of three letters (*i.e.* radicals) which carries the aspects of semantic contents[13]. Both root and pattern are used to derive Arabic words. In the derivation process the root radicals are inserted into their positions in the pattern. These positions are not necessarily consecutive.
- **Morpheme:** is the minimal meaning bearing unit that for constituting a word. The principal difference between morpheme and word is that morphemes may or may not be standalone units, while a word is a meaningful freestanding unit[14].
- **Patterns:** are the templates of combinations of consonants and vowels. The consonants represent slots for the root radicals to be inserted and the vowels represent the vocalism. The pattern is represented by sequences of Cs representing the consonants and Vs representing vocalism. The CV approach for representing patterns is widely used across languages (McCarthy and Prince 1990b; McCarthy and Prince 1990a; Smrz 2007; Attia 2008; Habash 2010). The original representation of patterns was proposed by Arabic grammar scholars as الميزان الصرفي *al-mīzān aṣ-ṣarfī*

[11] Wikipedia explanation, http://en.wikipedia.org/wiki/Stemming

[12] Wikipedia explanation of Lemma, http://en.wikipedia.org/wiki/Lemma_(linguistics)

[13] Wikipedia explanation of Root, http://en.wikipedia.org/wiki/Root_(linguistics)

[14] Wikipedia explanation of Morpheme, http://en.wikipedia.org/wiki/Morpheme

‘the morphological scale’ which uses the past verb فَعَلَ ‘did’ to represent the root radicals (Ali 1987; al-Saydawi 2006).

- **Pattern matching**: is the process of matching words with their possible patterns, either morphosyntactic patterns or morphophonemic patterns. The pattern matching algorithm must deal with three types of changes: incorporation or assimilation, substitution and deletion of vowel letters.
- **Function words**: are words with little semantic content meaning. They serve as important elements in the structure of sentences. They define grammatical relationships with other words within the sentence. They also signal the structural relationships that words have with one another. Function words are pronouns, prepositions, determiners, conjunctions, auxiliary and modal verbs (Baker, Hardie and McEnery 2006). In some languages, some function words are not free-standing, but clitics attached to content words.
- **Diacritization or vowelization**: is the process of adding the correct short vowels and diacritics to words. Vowelization is an important characteristic of the Arabic word. Vowelization helps in determining some morphological features of words. The presence of the short vowel on the last letter helps in determining the case or mood of the word. And the presence of a vowel on the first letter determines whether the verb is active or passive. The presence of other diacritics such as *šadda^h* and *madda^h* (extension) solve some ambiguities of words.
- **Part-of-speech tagging**: is the process of assigning part-of-speech grammatical category labels to the words of a corpus. Tagging is done automatically using part-of-speech tagger programs, and manual proofreading to content errors.
- **Parsing**: is the process of analysing the grammatical structure of a sequence of words or tokens. Parsing is automatically accomplished by using syntactic parser programs which output the syntax trees of the analysed text.

2.3.4.3 Morphological Analysis of Classical Quranic Arabic Text

The Quranic Arabic Corpus is a newly available resource enriched with multiple layers of annotation including morphological segmentation and part-of-speech tagging. The motivation behind this work is to produce a resource that enables further syntactic and semantic analysis of the Qur’an; a genre difficult to compare with other forms of Arabic, since the vocabulary and the spelling differs from Modern Standard Arabic (Dukes and Habash 2010). The Quranic Arabic Corpus uses the old Arabic script called the Othmani script; this is the same script used in writing the first copies of the Qur’an about 1,400 years ago. In addition, dots, short vowels and diacritics were added to the same word skeletons of the first written Qur’an.

Buckwalter's Arabic Morphological Analyzer (BAMA) was used to generate the initial tagging. The analyzer was adapted to work with Quranic Arabic text. After that, the annotated corpus was then put online to allow for collaborative proofreading and correction of the annotation (Dukes and Habash 2010).

Mapping was required to convert from the Modern Standard Arabic BAMA tag set to the classical grammar model used in the Quranic Arabic Corpus tag set. Manual disambiguation was required for some cases, where one-to-one mapping was not applicable such as particles. In order to adapt BAMA to process the Quranic Arabic Corpus text, three main modifications were made. First, spelling of the Qur'an differs from MSA. The differences involve orthographic variations of *hamza*^{h}, *'alif* and the long vowel *ā*. Second, the multiple diacritized analyses produced by BAMA for the processed words were ranked in terms of their edit-distance from the Qur'anic diacritization, with closer match ranked higher. Finally, filtering is done by choosing the highest rank analysis part-of-speech as a solution (Dukes and Habash 2010).

Manual annotation involves adding some parts of the morphological analysis, such as missing verb voice (active/passive), the energetic mood for verbs, the interrogative *alif* prefix, identifying particles, verb forms, and disambiguating *lām* prefix (Dukes and Habash 2010). Figure 2.1 shows a sample of the morphological and part-of-speech tags of the Quranic Arabic Corpus taken from chapter 29.

Index	Word	QAC morphological tag
29 | 1 | 1	الٓمٓ	POS:INL
29 | 2 | 1	أَحَسِبَ	A:INTG+ POS:V PERF ROOT:Hsb 3MS
29 | 2 | 2	ٱلنَّاسُ	Al+ POS:N LEM:<insa`n ROOT:Ans MP NOM
29 | 2 | 3	أَن	POS:SUB LEM:>an
29 | 2 | 4	يُتْرَكُوٓا۟	POS:V IMPF PASS ROOT:trk 3MP MOOD:SUBJ
29 | 2 | 5	أَن	POS:SUB LEM:>an
29 | 2 | 6	يَقُولُوٓا۟	POS:V IMPF ROOT:qwl 3MP MOOD:SUBJ
29 | 2 | 7	ءَامَنَّا	POS:V PERF (IV) ROOT:Amn 1MP
29 | 2 | 8	وَهُمْ	wa+ POS:PRON 3MP
29 | 2 | 9	لَا	POS:NEG LEM:laA
29 | 2 | 10	يُفْتَنُونَ	POS:V IMPF PASS ROOT:ftn 3MP

Figure 2.1 Sample of the morphological and part-of-speech tags of the Quranic Arabic Corpus taken from chapter 29

The automatic algorithm produced an analysis for 67,516 out of 77,430 words, followed by manual annotation done by native Arabic speakers. In the first stage the annotators corrected 21,550 words (28%) including 9,914 words missed by the analyzer and 11,636 corrections to existing analyses. In the second stage, another annotator made changes to 1,014 words (1.38% of all words). In the final stage, the corpus was put online for community volunteer correction, resulting in over 2,000 (2.6%) approved corrections to words (Dukes and Habash 2010).

The Quranic Arabic Corpus tag set adapts traditional Arabic grammar leading to morphological annotation that uses familiar terminology. This terminology enables people with Quranic syntax experience to participate in the online annotation to be verified against existing recognized standard textbooks on Quranic Grammar (Dukes and Habash 2010).

2.3.4.4 Four Approaches to Morphological Analysis for MSA Arabic Text

Generally, there are four main methodologies for developing robust morphological analysers. Arabic morphological analysis techniques include two-level and finite-state morphology (Al-Sughaiyer and Al-Kharashi 2004). The four main methodologies used for Arabic morphological analysis are:

- **Syllable-Based Morphology (SBM)**, which depends on analysing the syllables of the word.
- **Root-Pattern Methodology**, which depends on the root and the pattern of the word for analysis. Using this method, the root of the word is extracted by matching the word with lists of patterns and affixes.
- **Lexeme-based Morphology**, where the stem of the word is the crucial information that needs to be extracted from the word.
- **Stem-based Arabic lexicon with grammar and lexis specifications**, where stem-grounded lexical databases with entries associated with grammar and lexis specifications, is the most appropriate organization for the storage of Arabic lexical information.

All these methodologies (Al-Sughaiyer and Al-Kharashi 2004; Soudi et al. 2007) use pre-stored lists of root, stems, patterns and affixes and grammar and linguistic information encoded with the analysers. A fifth methodology is using tagged corpora and computer algorithms to extract a morphological database of the tagged words.

Machine learning algorithms do not really apply given the absence of morphologically tagged corpora and the absence of tractable learning algorithms.

Moreover, other challenges that face the application of machine learning algorithms to solve Arabic morphological analysis problems are: the encoding differences of Arabic text samples coded in Unicode and systems which only accept text coded in ASCII; the nature of Arabic as a highly inflected language; its variable word order of (VSO) for morphologically rich languages could lead to greater contextual ambiguity. Therefore it would require a higher-order model than languages like English and it would require a larger training corpus (Sánchez León and Nieto Serrano 1997; Hardie 2004); and the large tag set size used.

2.3.4.5 Requirements for Developing Morphological Analysers for Arabic Text

A robust and well-designed morphological analyzer for Arabic text has to meet the following conditions. First, it can correctly divide the analysed word into morphemes such as proclitics, prefixes, stem or root, suffixes and enclitics and specify the morphological features for each morpheme. Second, it can generate the correct pattern of the word and specify whether the generated pattern is a noun pattern, verb pattern or both. Third, it can extract the correct root of the word, whether it is a tri-literal root or quadriliteral root. Fourth, it can deal with unambiguous words (inert or stop words), irregular words, rare words and borrowed words. Fifth, it can specify the rules of transitive and intransitive verbs. Sixth, it can specify the derivation rules of past verbs, progress verbs and imperative verbs. Finally, it can deal with the orthographic aspects of the words such as vowelizing, incorporation, substitution and the writing of *hamza*^{h}, which helps in correcting spelling mistakes (Al-Bawaab 2009; Hamada 2009a). Section 8.2 discusses the requirements and specifications for developing an Arabic morphological analyser.

2.3.4.6 Morphological Analysers for Modern Standard Arabic Text

In this section, we will survey existing morphological analysers of Arabic text. Each morphological analyzer is studied in terms of the approach used to build it, the definition of a word’s morphemes, the database used to support morphological analysis, the morphological features that the analyzer can determine and the tag set used to encode these features.

1- Xerox Arabic Finite-State Morphological Analysis and Generation System (1998)

Xerox deals with Modern Standard Arabic text. It accepts input text which is fully-vowelized, partially-vowelized or non-vowelized, and outputs root, pattern, and affixes of the analysed word with feature tags such as: part-of-speech, person, number, mood, voice and aspect. The Xerox system aims to solve three challenges of Arabic: morphotactics, short vowels and Arabic lexicon lookup. The Xerox system is based on a lexicon of root-pattern representation of 5000 roots and 400 phonologically distinct patterns. It is based on the large two-level morphological analyzer for Arabic ALPNET. Xerox finite-state calculus was used to insert roots into their patterns and effectively generated 85,000 valid stems. The lexicon transducer also contains suitable prefixes and suffixes which are added to stems in the normal concatenative way. The result of the analysis returns back the upper-side string as root base-form followed by relevant morphosyntactic features of the analysis (Beesley 1996; Beesley 1998).

The advantages of the Xerox system are its large coverage; the reconstruction of short vowels; and the English glossary provided for each word. However, it has disadvantages such as lack of specification for multiword expressions (MWEs) and improper spelling relaxation rules. The major disadvantages of Xerox are: over-generation in word derivation due to uneven distribution of patterns for roots; the coarse-grained classification of words which is limited to 4 part-of-speech tags (verbs, nouns including adjectives and adverbs, particles and function words); and the high-level of ambiguity where it produces many analyses for most words (Attia 2008).

2- ElixirFM Functional Arabic Morphology (2007)

ElixirFM is an implementation of a novel computational model of the morphological processes in Modern Written Arabic. It is still in active development and related to the Prague Arabic Dependency Treebank (PADT) project (Hajič et al. 2004; Smrž et al. 2008). The system includes two essential components, namely a multipurpose programming library promoting clear style and abstraction in the model, and a linguistically refined, yet intuitive and efficient, morphological lexicon.

ElixirFM provides the user with four different modes of operation:

- **Resolve** provides tokenization and morphological analysis of the inserted text, even if one omits some symbols or does not spell everything correctly (Smrz 2007; Smrž 2009). The tokenization decision follows the conventions of PADT and PATB. For example the word للكتب *lil-kutub* ‘for the books’ has the following analyses (Habash 2010):

  - P--------- li ‘l’ ‘li’
  - N-----P2D al-kutub ‘k t b’ al >| FuCuL | << ‘i’

- **Inflect** transforms words into the forms required by context.
- **Derive** converts words into their counterparts of similar meaning but different grammatical category, specified via natural language descriptions or morphological tags. Word forms are encoded using morphophonemic patterns pertaining to morphological stem and reflect their phonological qualities.
- **Lookup** can lookup lexical entries by the citation form and nests of entries by the root. The lexicon of ElixirFM is derived from the open-source Buckwalter lexicon which contains about 40,000 entries that are grouped into about 10,000 nested entries.

Word forms are encoded via carefully designed morphophonemic patterns that interlock with roots or literal word stems. ElixirFM implements the comprehensive rules that draw the information from the lexicon and generate the word forms given the appropriate morphosyntactic parameters. ElixirFM also implements derivation, in any direction, between verbs, active or passive participles, and *masdar*s (*i.e.* de-verbal nouns). ElixirFM effectively exploits the inflectional invariant during the resolution of word forms from its root. ElixirFM presents the results of tokenization and morphological analysis in form of MorphoTrees which introduce intuitive hierarchies over the tokens and their readings that can be further pruned and disambiguated (Smrz 2007; Smrž 2009).

The advantages of the ElixirFM are the use of morphophonemic patterns that avoid the design of special rules to avoid the challenges of assimilation, gemination and deletion and listing the forms for each lexical item. However, the lexicon size of the morphophonemic patterns in the system is 4,290, which might suffer from coverage problems. Moreover, use of the open-source Buckwalter lexicon which contains about 40 thousands entries, inherits the disadvantages to the system such as the lack of specification for MWEs; improper spelling relaxation rules; and the lack of grammar-lexis specifications.

3- AlKhalil Morpho Sys (2010)

Alkhalil Morpho Sys is a morphological analyzer for Standard Arabic text. *Alkhalil* processes non-vowelized, partially vowelized and fully-vowelized MSA text. It is based on modeling a very large set of Arabic morphological rules, and on integrating linguistic resources that are useful to the analysis, such as (i) the root database; (ii) vowelized morphophonemic patterns associated with roots, (iii) and proclitic and enclitic lists. The outputs of analyzing Arabic words are presented in a table which shows: the fully-vowelized stem; its grammatical category and morphosyntactic features in natural language phrases; its possible roots associated with corresponding patterns; and its proclitics and enclitics (Boudlal et al. 2010).

The lists of noun patterns and verb patterns were obtained using *Sarf* (Arabic Morphology System) (ALECSO 2008b) and NEMLAR corpus (Attia et al., 2005). These lists contain a large number of about 28,000 morphophonemic patterns with full vowelization. *Alkhalil* contains about 7000 roots obtained from *Sarf* where each root is connected with specific derivation patterns used to derive words of that root (Mazroui et al. 2009; Boudlal et al. 2011). Matching the roots with their vowelized pattern gives the analyzer control over the derivations of that root, which solves the over-generation problem. However, using morphophonemic patterns has the shortcoming of under-generation. Moreover, *Alkhalil* inherited the limitations of Sarf of uncovering all derivatives such as broken plurals and non-derived words.

Alkhalil processes words by segmenting the words into (proclitics + stem + enclitics) then matches the stem with the non-derived words list. Then it treats the word as a derived word in the second phase and identifies the possible roots and patterns by analyzing the clitics and matching the words with the patterns. The system classifies nouns into 5 categories: gerund, active participle, passive participle, noun of place and time, and instrumental noun. It identifies morphological features of gender, number and syntactic form. Verbs are classified into perfect, imperfect and imperative. The morphological features of voice, syntactic form, number of root letters, conjugation, person and transitivity are identified for analyzed verbs. Particles are classified into their subcategories (Mazroui et al. 2009; Boudlal et al. 2011).

No evaluation was reported due to the unavailability of a test corpus. A basic evaluation was carried out to show the ability of the system to analyze words, by examining the outputs of *Alkhalil* on a sample of the Qur'an – chapter 20, which has about 1000 words. The outputs of *Alkhalil* showed that about 13.37% (132 words out of 987word of the sample) have no analysis. Most of the non-analyzed words belong to the function word and proper nouns categories.

4- MORPH2: A Morphological Analyzer for Arabic Text (2006-2010)

MORPH2 is a morphological analyzer for Arabic text and it is an extension to MORPH (Hadrich and Chaâben 2006). The focus of the improvement was adding a new step of vocalization and validation. MORPH2 uses a standard model of Arabic morphology. The model interprets all possible rules that govern the derivation of a word from its morpheme (root). MORPH2 takes into account the orthographic issues of Arabic words such as incorporation, substitution, vowelization and omission. The inputs are either fully vowelized words, partially vowelized words or non-vowelized words. The outputs are stored in an XML file and .xsl stylesheet in a structured format. MORPH2 depends on a pre-stored list of patterns and generated patterns to deal with substitution and vowelization cases. The analysis of words is carried out by following 5 steps:

- **Tokenization step:** is based on contextual exploration of punctuation that divides the text into sentences, then detection of words within sentences.
- **Morphological pre-processing step:** extracts clitics of the analysed words. Then, a filter process classifies the stem of the analysed word into particle, number, date or proper noun.
- **Affix analysis step:** identifies the basic elements of the word, namely: root and affixes. This process is accomplished following a five-stage process of (i) prefix and suffix identification; (ii) candidate affix identification; (iii) lexical filtering; (iv) association control of root radicals and affixes; and (v) transformation recognition.
- **Morphological analysis step:** determines all possible morphosyntactic features which are made in three stages: (i) identification of the part-of-speech of the word (*i.e.* noun, verb and particle); (ii) identification of the morphological features (i.e. gender, number, time and person); and (iii) filtering of the feature lists.
- **Vocalization and validation step:** depends on the previous two steps of affix and morphological analysis. The vowelization of the analysed word is done according to the morphosyntactic features and by matching the analysed word with its pattern. The validation process deals with transformation, omission and assimilation operations which occur for the analysed words.

MORPH2 contains many XML lexicons that provide necessary information for each step. Such lexicons are: the lexicon of proclitics, enclitics, and particles; lexicon of affixes and roots; and lexicon of derived and primitive nouns. The most important lexicon is the triliteral and quadriliteral roots of 5,754 entries, where patterns are connected with their corresponding roots. This combination provides 15,212 verbal stems and 28,024 nominal stems (Kammoun et al. 2010).

The evaluation of MORPH2 is done by calculating the recall and precision of analysing 23,121 word types of the test corpus which has all possible analyses of each word without taking into account the context of the words. The reported average recall and precision are 89.77% and 82.51% respectively. The limitation of the system is failure to detect relation nouns and non-derived (primitive) nouns (Hamado et al. 2009; Kammoun et al. 2010).

5- MIDAD Morphological Analyzer for Arabic Text (2009)

MIDAD applies linguistic knowledge of Arabic morphology to develop computer algorithms and rules that simulate human methods for deriving and analyzing words. The analyzer uses a database of Arabic roots and irregular words that need special processing. This database can be used to generate a larger database which includes most Arabic vocabulary. The use of the roots and irregular words database makes the program small, fast and robust (Sabir and Abdul-Mun'im 2009).

6- Application Oriented Arabic Morphological Analyzer (2009)

The analyzer depends on a novel algorithm that classifies the word's letters into letters belonging to affixes or underlying letters. The algorithm applies rules governing the relations between the word's letters. The algorithm does not depend on any pre-stored dictionaries. The analyzer depends on this algorithm to extract the root or stem, the affixes and the pattern of the analysed word. The inputs are either fully vowelized words, partially vowelized words or non-vowelized words. The outputs show all possible roots, affixes and patterns of the analysed word. They report an accuracy rate of 97.7% and they claim that the analyzer is five times faster than any existing analyser. As reported, the analyzer can be integrated into other applications and parts of the analyzer might be re-used (Sonbul, Ghnaim and Dusouqi 2009).

2.3.4.7 The ALECSO/KACST Initiative of developing and evaluating Morphological Analysers of Arabic text

The Arab League Educational, Cultural and Scientific Organization (ALECSO) and King Abdul-Aziz City of Science and Technology (KACST) have promoted an initiative on morphological analysers for Arabic text which aims to encourage research in developing an open source morphological analyzer for Arabic text which has high accuracy, is easy to develop and which can be integrated into higher levels of applications for processing Arabic text.

Six morphological analysers entered the ALECSO/ KACST competition for evaluating morphological analysers for Arabic text. Table 2.3 lists the names, affiliations and the major contributions of the participants. According to the evaluation methodology, the organizers of the ALECSO/KACST workshop evaluated the results of the morphological analysers. The highest scores were achieved by Mazroui, Meziane et al. (2009), and Boudlal, Lakhouaja et al. (2010). The official results and scores of the ALECSO/KACST competition have not been published for unspecified and unknown reasons. Only specifications for development and evaluation methodology were published (Al-Bawaab 2009; Hamada 2009b; Hamada 2009a; Hamada 2010). Section 9.2 discusses the initiative as guidelines for evaluating Arabic morphological analysers.

Table 2.2 ALCSO/KACST competition participants

Author(s)	Affiliation	Algorithm Name	Methodology
bin Hamdo et al	MIRACL Labs, Tunis.	MORPH	Depends on pre-stored list of patterns and generated patterns
Mazroui et al	University of Mohammed I, Morocco.	Alkhalil	Depends on databases of verbs, derived nouns and original nouns derived using *Sarf* (Arabic Morphology System)
Sabir and Abdul-Mun'im	MIDAD, Egypt.	MIDAD	Depends on rules that simulate the human methods of deriving and analyzing words and a database of Arabic roots and irregular words.
Sawalha and Atwell	University of Leeds, UK.	SALMA	Depends on linguistic knowledge of the language as well as corpora. Broad-coverage lexicon and comprehensive lists of roots, clitics, affixes and patterns.
Sonbul et al	Higher Institute of Applied Science and Technology (HIAST), Syria.	-	Depends on a novel algorithm that classifies the word's letters into letters belong to the affixes or original letters.
Smrz	Charles University in Prague, Czech republic.	ElixirFM	An implementation of a novel computational model of the morphological processes in Modern Written Arabic.

2.4. Part-of-Speech Tagging

Part-of-speech taggers are used to enrich a corpus by adding a part-of-speech category label to each word, showing the broad grammatical class of the word, and morphological features such as tense, number, gender, etc. The list of all grammatical category labels is called the tag set. The design of the tag set is an important prerequisite to this annotation task. The task requires a tagging scheme, where each tag or label is practically defined by showing the words and contexts where each tag applies; and a tagger, a program responsible for assigning a tag to each word in the corpus by implementing the tag set and tagging scheme in a tag-assignment algorithm (Atwell 2008).

Automatic taggers have been used from the early years of Corpus Linguistics. TAGGIT in 1971 achieved an accuracy of 77% tested on the Brown corpus. In the late 1970s, CLAWS1, a data-driven statistical tagger was built to carry out the annotation of the Lancaster/ Oslo-Bergen corpus (LOB), and had an accuracy rate of 96-97%. Later tagger development included systems based on Hidden Markov Models (HMM); HMM taggers have been made for several languages. The Brill tagger (Brill 1995) is an example of data-driven symbolic tagger. The ENGCG and EngCG-2 are based on a framework known as Constraint Grammar (CG) (Voutilainen 2003).

Recently, many new systems based on a variety of Markov Model and Machine Learning (ML) techniques have appeared for many languages. Hybrid solutions have also been investigated (Voutilainen 2003). ACOPOST[15], A Collection of POS Taggers, consists of four taggers of different frameworks: Maximum Entropy Tagger (MET), Trigram Tagger (T3), Error-driven Transformation-Based Tagger (TBT) and Example-based tagger (ET). The SNoW-based Part of Speech Tagger[16] and LBJ Part of Speech Tagger[17] make use of the Sequential Model. NLTK[18], the Natural Language Toolkit, includes Python re-implementations of several POS taggers such as; Regexp Tagger, N-Gram Tagger, Brill Tagger and HMM Tagger; in addition NLTK includes tutorials and documentation on tagging. RelEx[19] provides English-language part-of-speech tagging, entity tagging, as well as other types of tags (gender, date, money, etc.). Spejd[20] - Shallow Parsing and Disambiguation Engine is a tool for simultaneous rule-based morphosyntactic disambiguation and partial parsing. VISL Constraint Grammar[21] is an example of rule based disambiguation.

Enriching the source text samples of corpora with part-of-speech information for each word, as a first level of linguistic enrichment, results in more useful research resources. English corpora have been developed for a long time and for a variety of formats, types and genres. Several English corpora have been enriched with Part-of-Speech tagging, and a variety of different English corpus part-of-speech tag sets have been developed, including: the Brown corpus (BROWN), the Lancaster/ Oslo-Bergen corpus (LOB), the Spoken English Corpus (SEC), the Polytechnic of Wales corpus (PoW), the University of Pennsylvania corpus (UPenn), the London-Lund Corpus (LLC), the International Corpus of English (ICE), the British National Corpus (BNC), the Spoken Corpus Recordings In British English (SCRIBE), etc (Atwell 2008). The AMALGAM[22] multi-tagged corpus amalgamates all these tagging schemes in a common collection of English texts: in the AMALGAM corpus, the different part-of-speech tag sets used in these English general-purpose corpora are applied to illustrate the range of rival English corpus tagging schemes, and the texts are also parsed according to a range of rival parsing schemes, so each sentence has more than one parse-tree, called "a forest" (Atwell et al. 2000). Part-of-speech tag sets and taggers have also been developed for other European languages. The EAGLES, European Advisory Group on Language Engineering Standards project, drew up standards for tag sets, morphological classes and codes for (western) European languages, including: EAGLES recommendations for the morphosyntactic

[15] ACOPOST http://acopost.sourceforge.net/

[16] SNoW-based Part of Speech Tagger http://l2r.cs.uiuc.edu/~cogcomp/asoftware.php?skey=POS

[17] LBJ Part of Speech Tagger http://l2r.cs.uiuc.edu/~cogcomp/asoftware.php?skey=FLBJPOS

[18] NLTK http://www.nltk.org/

[19] RelEx http://opencog.org/wiki/RelEx

[20] Spejd http://nlp.ipipan.waw.pl/Spejd/

[21] VISL Constraint Grammar http://beta.visl.sdu.dk/cg3.html

[22] Automatic Mapping Among Lexico-Grammatical Annotation Models (AMALGAM) http://www.comp.leeds.ac.uk/amalgam/amalgam/amalghome.htm

annotation of corpora (Leech and Wilson 1999); a synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora: a common proposal and applications to European languages (Monachini and Calzolari 1996); and an EAGLES study of the relation between tag sets and taggers (Teufel et al. 1996).

The potential uses of a part-of-speech tagged corpus are key factors in deciding the range and number of part-of-speech tags. Many linguistic analyses use part-of-speech tagged corpora to analyze text and extract information, where part-of-speech tags play an essential role in classifying text and direct search to the actions, events, places, etc are described in the text. The most obvious applications are in lexicography and NLP/computational linguistics. Further applications include using the tags in data compression (Teahan 1998); and as a possible guide in the search for extra-terrestrial intelligence (Elliott and Atwell 2000). Other generic applications that make use of part-of-speech tag information are: searching and concordancing, grammatical error detection in Word Processing, training Neural Networks for grammatical analysis of text, or training statistical language processing models (Atwell 2008). Part-of-Speech tagging is a key technology in discovering suspicious events from text. Part-of-speech tagging is required for partial parsing which is a first step for named entity (NE) recognition as one module of the Information Extraction (IE) pipeline. IE is the main text extraction methodology used for counter-terrorism text analysis tools (Zolfagharifard 2009), and processing Arabic is a key task in discovering these suspicious events.

2.4.1 Part-of-Speech Taggers for Arabic Text

Arabic part-of-speech tagging development started more recently. A range of different techniques have been used to solve the problem of part-of-speech tagging of Arabic. The APT tagger uses a combination of both statistical Viterbi algorithm, and rule-based techniques (Khoja 2001). Brill's "transformation-based" or "rule-based" part-of-speech tagger has been applied for Arabic (Freeman 2001). Harmain (2004) developed a web-based Arabic tagger. Diab, Hacioglu et al. (2004) used Support Vector Machines (SVM), a supervised learning algorithm, to achieve an accuracy of 95%. Habash and Rambow (2005) developed another part-of-speech tagger that uses SVM and Viterbi decoding. HMM has been widely used in part-of-speech tagging for Arabic, with reported accuracy of 97% on LDC's Arabic Treebank of Modern Standard Arabic (Al-Shamsi and Guessoum 2006) and 70% when tested on CallHome Egyptian Colloquial Arabic (ECA) and the LDC Levantine Arabic (Duh and Kirchhoff 2005). Applications of Memory-Based learning to morphological analysis and part-of-speech tagging of written Arabic have been explored (Marsi, Bosch and Soudi 2005). Also, combinations of rule based and machine learning methods for tagging Arabic words (Tlili-Guiassa 2006). A multi-agent architecture was developed to address the problem of part-of-speech tagging of Arabic text with vowel marks (Zibri, Torjmen and Ahmad 2006). A rule-based PoS tagging system, Arabic Morphosyntactic Tagger AMT (Alqrainy 2008), uses two different techniques: the pattern-based technique, which is based on using Pattern-Matching Algorithm (PMA), and lexical and contextual techniques. The AMT tagger makes use of the last diacritic mark of Arabic words to reduce the tagging ambiguity. The accuracy of the AMT tagger reported was 91%.

Nearly all these Arabic part-of-speech taggers were developed by NLP research groups for their own internal use, and are not freely downloadable by other researchers. The taggers use different tag sets, and accuracies are reported on different test corpora. Appendix B compares between these part-of-speech taggers for Arabic text in terms of methodology, corpus used, tag set, evaluation methodology, and evaluations metrics.

2.5 Chapter Summary

This chapter studied existing morphosyntactic analysis systems for text corpora in three dimensions. First, it explored Arabic text corpora as a background prerequisite for morphosyntactic analysis. Second, it studied morphological analysers for text corpora concentrating on methodologies, challenges, examples of existing morphological analysers, and evaluation standards. Third, it surveyed part-of-speech tagging technology and existing part-of-speech taggers for Arabic text.

Arabic corpora started to appear in the late 1980s. Most of the existing Arabic corpora are of MSA written text, mainly newspaper text. Only two corpora are open-source and available to download. These are the Corpus of Contemporary Arabic (CCA) (Al-Sulaiti and Atwell 2006) and the Quranic Arabic Corpus (QAC) (Dukes et al. 2010; Dukes and Habash 2010). A new third open source corpus is the Corpus of Traditional Arabic Lexicons which is discussed in Chapter 4.

Several morphological analysers for Arabic text exist. Morphological analysis is an important pre-processing step for many text analytics applications. The aim of morphological analysis is to define the morphosyntactic information of a corpus words. Automatic morphological analysis started in the 1950s. Finite-state methodology has dominated since the 1980s. It was originally investigated at Xerox and it has been used to develop wide-coverage morphological analysers for several languages. The four main methodologies used for Arabic morphological analysis are: Syllable-Based Morphology (SBM); Root-Pattern Methodology; Lexeme-based Morphology; and Stem-based Arabic lexicon with grammar and lexis specifications. A fifth methodology is using tagged corpora and computer algorithms to extract a morphological database of the tagged words.

This chapter surveyed existing Arabic morphological analysers focusing on the morphological analysers that participated in the ALECSO/KACST competition. These surveyed morphological analysers are: (i) Xerox Arabic Finite-State Morphological Analysis and Generation System (1998); (ii) ElixirFM Functional Arabic Morphology (2007); (iii) Alkhalil Morpho Sys (2010); (iv) MORPH2: A Morphological Analyzer for Arabic Text (2006-2010); (v) MIDAD Morphological Analyzer for Arabic Text (2009); and (vi) Application Oriented Arabic Morphological Analyzer (2009). Community based approaches to develop and evaluate morphological analysers for Arabic text namely: the MorphoChallenge competition and the ALECSO/KACST initiative were discussed. More detailed discussion of them is presented in Chapter 8 and Chapter 9.

Morphological analysers are designed to generate all possible analyses of the analysed words out of their context. Disambiguating the analysis suitable to the context is done by using part-of-speech taggers. Part-of-speech tagging technology was surveyed in this chapter. The survey listed state of the art part-of-speech taggers for English, the tagged corpora and the standards. Then, existing part-of-speech taggers for Arabic text were briefly listed focusing on their development approaches and their accuracy as reported by their developers.

Part II: Background Analysis and Design

Summary of Part II

Part II is an attempt to plan ahead for what is required for the full SALMA – Tagger in Chapter 8. Firstly, an analysis of the failings of morphological analyzers and stemmers is presented in Chapter 3. Secondly, development of a broad-coverage lexical resource, the SALMA – ABCLexicon, required by the development of the morphological analyzer is presented in Chapter 4. Finally, an analysis of existing tag sets as background to designing the SALMA –Tag Set, Chapters 3, 4 and 5 is a necessary prior step to develop the SALMA – Tagger.

Chapter 3
Comparative Evaluation of Arabic Morphological Analyzers and Stemmers

This chapter is based on the following sections of published papers:

Sections: 2, 3, 4, 5 and 6 are based on sections 1, 2, 3 and 4 in (Sawalha and Atwell 2008)

Section 7 is based on section 3.1 in (Sawalha and Atwell 2009a)

Chapter Summary

Arabic morphological analysers and stemming algorithms have become a popular area of research. Several computational linguists have designed and developed algorithms to tactile the problem of morphology and syntax; but each researcher proposed an evaluation methodology based on different text corpora. Therefore, we cannot make comparisons between these algorithms. This chapter discusses four different fair and precise evaluation experiments using a gold standard for evaluation consisting of two 1000-words text documents from the Holy Qur'an and the Corpus of Contemporary Arabic. Secondly, it discusses a combination of the results of these morphological analysers and stemming algorithms to allow "voting" on analysis of each word. The evaluation of the algorithms shows that Arabic morphology is still a challenge. Finally, it presents an analytical study of the triliteral Arabic roots based on the Qur'an as corpus roots, and the triliteral roots of a broad-coverage lexical resource of traditional Arabic lexicons. The study shows that more than 25% of Arabic triliteral roots are hard to analyze.

3.1 Introduction

Stemming is the process of assigning morphological variants of words to equivalent classes, such that each class corresponds to a single stem. It is also defined as reducing inflected words to their stem, base, or root form[23]. For example words such as *writing, write, writer* and *written* are reduced to the root *write*. Stemming has been widely used in several fields of natural language processing such as data mining, information retrieval, text analytics applications (e.g. compression, spell checking, text searching, and text analysis), and multivariate analysis.

A widely used simple stemming algorithm for English is the Porter Stemmer (Porter 1980). It is available as a freely distributed implementation written in several programming languages[24]. The stemmer is based on a series of simple cascaded rewrite rules which can be viewed as a lexicon-free finite state transducer FST stemmer. However, modern stemmers need to be more complicated than the Porter Stemmer. For instance the word *Illustrator* (*i.e.* a software package) does not share the stem *illustrate* with the word *illustrator* (*i.e.* one who gives or draws illustrations) (Jurafsky and Martin 2008). It also need to distinguish whether the part of the word is a suffix or looks like a suffix *e.g*. the *–ion* in *lion* looks like a suffix (Khoja 2003).

The Natural Language Toolkit[25] (NLTK) provides three stemmers for English namely: Porter Stemmer (**`nltk.stem.porter(PorterStemmer)`**), Lancaster Stemmer (**`nltk.stem.lancaster(LancasterStemmer)`**) and Regular Expression Stemmer (**`nltk.stem.regexp(RegexpStemmer)`**). The Porter and Lancaster stemmers are used as black boxes while the Regular Expression stemmer requires the user to provide the affixes that the stemmer should deal with.

Many stemming algorithms have been developed for many languages including Arabic; see section 2.3.4. They attempt to reduce morphological variants of words which have similar semantic interpretations to their common stem. Arabic has a complex morphological structure. So, it is difficult to deal with. Arabic is considered to be a root-based language: Arabic words are morphologically derived from roots following derivational templates called patterns, where many affixes (*i.e.* prefixes, infixes and suffixes) and clitics (*i.e.* proclitics and enclitics) can be attached to form surface words. These roots are made up of three, four or five consonants (Thabet 2004).

The motivation for comparing between different stemming algorithms and morphological analysers is that such systems are prerequisites for Part-of-Speech tagging and then parsing. It is also considered an essential step in many computational linguistic applications.

[23] Wikipedia definition, http://en.wikipedia.org/wiki/Stemming

[24] The Porter Stemmer implementation http://tartarus.org/~martin/PorterStemmer/

[25] The Natural Language Toolkit (NLTK) http://www.nltk.org

3.2 Three Stemming Algorithms

Many stemming algorithms for Arabic already exist (Al-Sughaiyer and Al-Kharashi 2002; Al-Shalabi et al. 2003; Thabet 2004; Al-Shalabi 2005; AlSerhan and Ayesh 2006; Yusof, Zainuddin and Baba 2010; Hijjawi et al. 2011), but few are open-source or readily accessible. The selection of the stemming algorithms to be studied is limited to three stemming algorithms namely: Khoja's stemmer (Khoja 2003), Buckwalter's morphological Analyzer (BAMA) (Buckwalter 2002) and Al-Shalabi et. al, triliteral root extraction algorithm (Al-Shalabi et al. 2003) for which a ready access to the implementation and/or results is available. These three stemmers are freely available online or through personal communication with the authors. A fact about the selected systems worth mentioning here is that these stemmers differ in the implementation methodology used in their development. This means that our comparative evaluation compares between three different stemming methodologies as well as three existing stemmers and morphological analyzers.

3.2.1 Shereen Khoja's Stemmer

We obtained a Java implementation of Shereen Khoja's stemmer[26]. Khoja's stemmer is the rule-based component of her Arabic part-of-speech tagger (APT). It removes the longest suffix and the longest prefix. Then, it matches the remaining word with verbal and noun patterns to extract the root. It deals with language specific variation to the general rules of the language to produce the correct root such as: weak letters (*'alif*, *wāw*, and *yā'*) and *hamza^h* that change their form during derivation, deleted root letters during derivation, and stop words (function words) that do not have roots. The stemming algorithm restores the weak root letter to *wāw* as default solution. It does not deal with the orthographic issues of writing the *hamza^h* and it always places the *hamza^h* on *'alif* (Khoja 2003). The stemmer makes use of several linguistic data files such as a list of all diacritic characters (7), punctuation characters (38), definite articles (5), stop words (168), prefixes (11), suffixes (28), triliteral roots (3,822), quadriliteral roots (926) and triliteral root patterns (46) (Larkey and Connell 2001). The purpose of constructing the stemmer was to identify the affixes and to find the pattern of the word, because the affixes and the pattern of the word provide linguistic information useful to guess the tag of the word.

Khoja's reported accuracy of her stemmer is 96% using newspaper text on the assumption it was evaluated on the developed corpus. The errors are mainly proper nouns and borrowings from foreign languages (Khoja 2003). However, there is not any detail of

[26] Java version of Khoja's stemmer is available to download from http://zeus.cs.pacificu.edu/shereen/research.htm

the evaluation methodology, text used in evaluation and accuracy metrics. Figures 3.4 and 3.6 in section 3.5, shows sample output of Khoja's stemmer.

3.2.2 Tim Buckwalter's Morphological Analyzer

Tim Buckwalter developed a morphological analyzer for Arabic (BAMA) (Buckwalter 2002). Buckwalter compiled three Arabic-English lexicon files; the prefixes file contains 299 entries, the suffixes file contains 618 entries, and the stems file contains 82,185 entries representing 38,600 lemmas. To control prefix-stem-suffix combinations, the analyzer is provided with three morphological compatibility tables which consist of 1,648 prefix-stem combinations, 1,285 stem-suffix combinations and 598 prefix-suffix combinations. Short vowels and diacritics were included in the lexicons[27] (Maamouri and Bies 2004; Maamouri et al. 2004).

BAMA was used to morphologically annotate the Penn Arabic Treebank distributed by the Linguistic Data Consortium (LDC). The results of the Arabic Treebank part 1 v 2.0, part 2 v 2.0 and part 3 v 1.0 were recycled through the system to modify the system and update the lexicon. With each cycle, the accuracy of the morphological analyzer and the coverage of the lexicon were improved from 90.63% for part 1 v 2.0 and 99.24% for part 2 v 2.0 to 99.25% for part 3 v 1.0. The most frequent accuracy problems were the absence of non-Arabic proper names (*i.e.* geographical and organizational names) which caused 38% of errors, false-positives (*i.e.* foreign names recognized as valid Arabic words), missing Arabic proper names (15% of errors), incorrect vocalization (21% of errors), plus the total cases where the analyzer failed to identify the passive voice or provide the proper verbal prefix or suffix (Maamouri and Bies 2004; Maamouri et al. 2004). Figures 3.4 and 3.6 in section 3.5, shows sample output of BAMA.

3.2.3 Triliteral Root Extraction Algorithm

Al-Shalabi, Kanaan and Al-Serhan developed a root extraction algorithm which does not use any dictionary. It depends on assigning weights for a word's letters multiplied by the letter's position, Consonants were assigned a weight of zero and different weights were assigned to the augmented letters of (أ *hamza*^{h}, ا *'alif*, ت *tā'*, س *sīn*, ل *lām*, م *mīm*, ن *nūn*, هـ *hā'*, و *wāw*, ي *yā'*) where all affixes are formed by combinations of these letters. The algorithm selects the letters with the lowest weights as root letters. The algorithm achieved an accuracy rate of about 93% texted on a sample of modern standard Arabic text comprising 242 non-vowelized Arabic abstracts chosen randomly from the proceedings of the Saudi Arabian National Computer Conference (Al-Shalabi et al. 2003). Figures 4 and 6 show a sample output of the triliteral root extraction algorithm.

[27] Tim Buckwalter's web site: http://www.qamus.org

3.3 Stemming by Ensemble or Voting

Natural language engineering aims to design systems that make as few errors as possible with as little effort and cost as possible. There are many ways to reduce errors. First, a better representation of the problem will reduce errors. Second, spending more time on encoding language knowledge of hand-crafted systems, or on finding more training data for data-driven systems, will reduce errors of the system as well. However, these solutions are not always available because of lack of resources (Chan and Stolfo 1995; Atwell et al. 2000; Borin 2000; Dˇzeroski, Erjavec and Zavrel 2000; Escudero, Mhrquez and Rigau 2000; Banko and Brill 2001; Halteren, Zavrel and Daelemans 2001; Marques and Lopes 2001; Hu and Atwell 2003; Banko and Moore 2004; Glass and Bangay 2005; Yonghui et al. 2006).

Rather than giving better representation of the problem or spending more time in encoding language knowledge and finding more training data; combining different systems of known representation will, hopefully, reduce errors of a system. The idea behind combining different systems is that systems designed differently in terms of using different formalism or containing different knowledge will produce different types of errors. Provided that these differences are (i) *complementary* (i.e. systems produce different types of errors, where a system's errors are not the same as the other system or not a subset of the other systems errors) and (ii) *systematic* (i.e. errors are not random). So, fixing some types of errors generated will reduce the errors of the combined system. By employing these disagreements of systems we might get better results and fewer errors of the combined system (Borin 2000; Halteren et al. 2001).

Much research has been done in the field of machine learning to find ways to improve the accuracy of supervised classifiers. An ensemble of classifiers that generate uncorrelated decisions can be more accurate than any of its component classifiers. There are many varieties of ensemble classifiers in terms of selecting individual classifiers or in the way they are combined (Halteren et al. 2001). If the classifiers are accurate and diverse, then the ensemble of classifiers will be more accurate than any of its individual members. An accurate classifier has an error rate of better than random guessing on new values. Diversity means that two classifiers make different errors on new data points (Dietterich, 2000).

A question raised is: Is it possible in practice to build an ensemble that outperforms any of its individual members? There are three sources of evidence for the possibility of building a good ensemble. The first is statistical. Suppose that $\mathcal{H}$ is the search space of hypotheses to identify the best hypothesis of a learning algorithm. If the amount of training data is too small, compared to the size of hypothesis space, then the learning algorithm can find many different hypotheses in $\mathcal{H}$. All of them give the same accuracy.

The ensemble that combines all of these accurate classifiers can “average” their votes, and reduces the risk of choosing the wrong classifiers. The second reason is computational; many learning algorithms get stuck in local optima while performing some form of local search. Constructing an ensemble that runs the search from different starting points may provide a better approximation to the true unknown function than any of the individual classifiers. The final reason is representational; the true function f in most machine learning applications cannot be represented by any hypothesis in $\mathcal{H}$. It may be possible to expand the space of representable functions by forming weighted sums of hypotheses drawn from $\mathcal{H}$. Figure 3.1 below depicts the three reasons (Dietterich 2000).

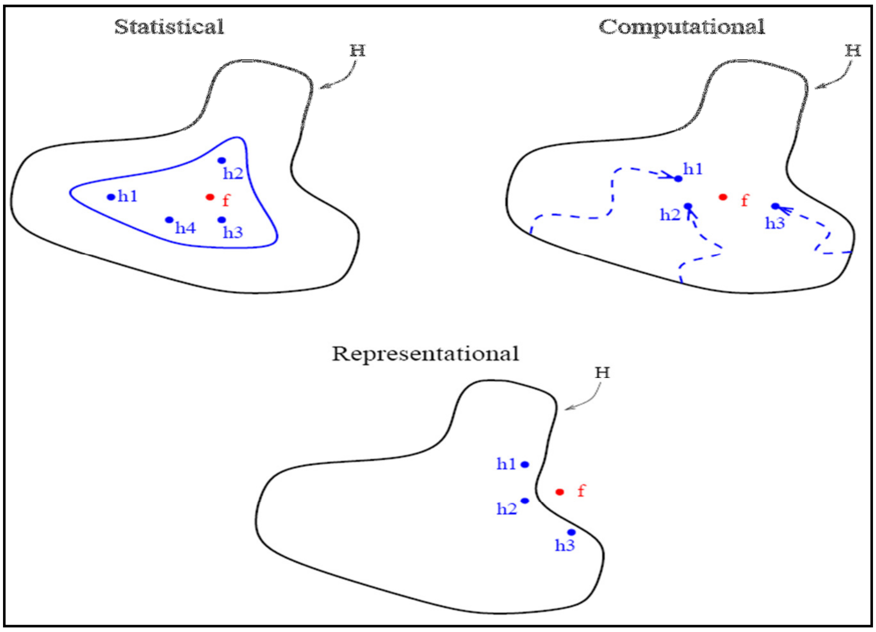


Figure 3.1 The statistical, computational and representational methods for better and more accurate ensemble (Dietterich 2000)

The reuse of existing components is an established principle in software engineering. A voting program is developed to allow “voting” on the analysis, of procured results from several candidate systems, of each word: for each word, examine the set of candidate analyses. Where all systems are in agreement, the common analysis is copied; but where contributing systems disagree on the analysis; take the “majority vote”, the analysis given by most systems. If there is a tie, take the result produced by the system with the highest accuracy (Atwell and Roberts 2007)

The output analysis of the stemming algorithms is considered as input for the “voting” program. The program reads in these files, tokenizes them, and stores the words and the roots extracted by each stemming algorithm in temporary lists to be used by the voting procedures.

The temporary lists work as a bag of words that contains all the result analysis of the stemming algorithms. These roots are ranked in best-first order according to accuracy results; see section 3.6. Khoja's stemmer results are inserted to the list first then the results from triliteral stemming algorithm and finally the results of BAMA.

After the construction of the lists of all words and their roots, a majority voting procedure is applied to it to select the most common root among the list. If the systems disagree on the analysis, the voting algorithm selects "Majority Vote" root as the root of the word. If there is a tie, where each stemming algorithm generates a different root analysis then the voting algorithm selects the root by two ways.

- In experiment 1, the algorithm simply selects the root randomly from the list using the `FreqDist()` Python function.
- In experiment 2, the algorithm selects the root generated from the highest accuracy stemming algorithm which is simply placed in the first position of the list as the candidate roots of the word are inserted to the list using the best-first in terms of accuracy strategy.

Figures 3.4 and 3.6 in section 3.5, show sample output of the voting algorithm for both experiments.

3.4 Gold standard for Evaluation

A gold standard for evaluating morphological analyzer and stemming algorithms for Arabic text was built using a randomly selected chapter of the Qur'an; chapter number 29 سُورَةُ العَنْكَبُوت *sūra^{tu} al-ankabūt* "The Spider", consisting of about 1000 words and representing classical Arabic text; see figure 3.2. Also, a modern standard Arabic (MSA) text sample of the Corpus of Contemporary Arabic[28] CCA (Al-Sulaiti and Atwell 2006) was used consisting of about 1000 words. The MSA text sample is selected from three genres; politics, sports and economics section, of newspaper and magazine articles; see figure 3.2. The gold standard is constructed by manually extracting the root of each word of the test documents. The manually extracted roots have been checked by Arabic language experts. Figures 3.4 and 3.6 in section 3.5, show samples of the gold standard's roots for both text types.

Table 3.1 shows number of word tokens, number of word types and detailed frequency of 4 texts: the gold standard's Qur'an text document, the full Qur'an as a corpus, the gold standard's CCA text document and a daily MSA newspaper article from Al-Rai daily newspaper[29] published in Jordan. The analysis also shows that function words such as في *fī* "in", من *min* "from", على *'alā* "on" and الله *'allāh* "GOD" are the most frequent words in any Arabic text. On the other hand, non-function words with high

[28] The Corpus of Contemporary Arabic http://www.comp.leeds.ac.uk/eric/latifa/research.htm
[29] Al-Rai daily newspaper http://www.alrai.com/

frequency such as الجامعات *al-ğāmiʿāt* “Universities” and الكويت*al-kuwayt* “Kuwait” give a general idea about the main topic or the theme of the article.

Simple tokenization is applied for the text of the gold standard documents. This will ensure that test documents can be used to test any stemming algorithm smoothly and correctly.

الم أَحَسِبَ النَّاسُ أَن يُتْرَكُوا أَن يَقُولُوا آمَنَّا وَهُمْ لَا يُفْتَنُونَ وَلَقَدْ
فَتَنَّا الَّذِينَ مِن قَبْلِهِمْ فَلَيَعْلَمَنَّ اللَّهُ الَّذِينَ صَدَقُوا وَلَيَعْلَمَنَّ الْكَاذِبِينَ
أَمْ حَسِبَ الَّذِينَ يَعْمَلُونَ السَّيِّئَاتِ أَن يَسْبِقُونَا سَاءَ مَا يَحْكُمُونَ مَن
كَانَ يَرْجُو لِقَاءَ اللَّهِ فَإِنَّ أَجَلَ اللَّهِ لَآتٍ وَهُوَ السَّمِيعُ الْعَلِيمُ وَمَن
جَاهَدَ فَإِنَّمَا يُجَاهِدُ لِنَفْسِهِ إِنَّ اللَّهَ لَغَنِيٌّ عَنِ الْعَالَمِينَ وَالَّذِينَ آمَنُوا
وَعَمِلُوا الصَّالِحَاتِ لَنُكَفِّرَنَّ عَنْهُمْ سَيِّئَاتِهِمْ وَلَنَجْزِيَنَّهُمْ أَحْسَنَ الَّذِي
كَانُوا يَعْمَلُونَ وَوَصَّيْنَا الْإِنسَانَ بِوَالِدَيْهِ حُسْنًا وَإِن جَاهَدَاكَ لِتُشْرِكَ بِي
مَا لَيْسَ لَكَ بِهِ عِلْمٌ فَلَا تُطِعْهُمَا إِلَيَّ مَرْجِعُكُمْ فَأُنَبِّئُكُم بِمَا كُنتُمْ
تَعْمَلُونَ وَالَّذِينَ آمَنُوا وَعَمِلُوا الصَّالِحَاتِ لَنُدْخِلَنَّهُمْ فِي الصَّالِحِينَ

ستبقى العولمة وإلى وقت ممتد مثيرة للأسئلة والأجوبة وفي هذا
المقال وقفة تأمل عميقة في بعض هذه الأسئلة بدأت منذ فترة
موجة جديدة من الكتابات تروج للعولمة باعتبارها الشكل
الجديد لحياة البشر في ظل القطب الأمريكي وهناك نمط من
هذه الكتابات يروج للنمط الأمريكي متعدد الأعراق والثقافات
بوصفه النمط الأمثل للحياة في القرية الكونية الجديدة التي
قاربت وسائل الاتصالات والمواصلات ونظم المعلومات ووسائل
الإعلام بين أجزائه المختلفة ويبشر أصحاب هذه النظرة ببشر
من نوع جديد بشر كوزموبوليتان

Figure 3.2 Sample from Gold Standard first document taken from Chapter 29 of the Qur’an (left) and the CCA (right).

Table 3.1 Summary of detailed analysis of the Arabic text documents used in the experiments

	Qur’an as Corpus		**Gold standard document 1 Chapter 29**		**Gold standard document 2 CCA Document**		**Al-Rai newspaper article**	
Tokens	77,787		987		1005		977	
Word Types	19,278		616		710		678	
	Token	Freq.	Token	Freq.	Token	Freq.	Token	Freq.
1	فِي	1179	فِي	21	في	35	في	39
2	مِن	872	اللَّهِ	17	من	21	من	16
3	مَا	832	مِن	14	على	12	على	13
4	الَّذِينَ	808	اللَّهُ	12	التي	12	التي	10
5	عَلَى	652	وَمَا	12	الكويت	11	إلى	9
6	وَمَا	640	إِلَّا	12	أن	10	المبنى	8
7	إِنَّ	605	الَّذِينَ	11	هذه	10	الجامعات	8
8	اللّهِ	464	مَا	8	إلى	8	أن	7
9	أَن	499	اللَّهَ	8	اماء	8	السلام	7
10	قَالَ	416	كَانُوا	8	عن	7	جلالته	7

3.5 Four Experiments and Results

In order to compare fairly between different stemming algorithms, four different experiments were applied to compute the accuracy of each algorithm. The accuracy of each experiment is measured using f-score; see formula 1. Each time the experiment is done, a comparison of the results with the gold standard is performed.

$$\text{Accuracy} = \frac{\text{Number of Correct Roots}}{\text{Number of Tokens/Types in the Sample}} * 100\% \quad \text{........ (1)}$$

The first experiment compares each token’s root output by the three stemming algorithms separately against the token’s roots in the gold standard. The second experiment excludes stop words (function words). The third experiment compares all word-type roots. Finally, word-type roots excluding the stop words (function words) are compared to the gold standard roots. The evaluation is done by comparing roots of the three algorithms according to the four experimental specifications against the manually extracted gold standard roots. Then the accuracy rate of each algorithm is computed using formula (1). Table 3.2 and figure 3.3 show the accuracy rates resulting from the four different experiments for the Qur’an test document. Table 3.3 and figure 3.5 show the accuracy rates resulting from the four different experiments for the CCA test document. Figure 3.4 and 3.6 show sample outputs of the stemming algorithms and the gold standard.

Table 3.2 Results of the four evaluation experiments of the 3 stemming algorithms tested using the Qur’an text sample

Algorithm	**Experiment 1: All Tokens (978 tokens)**			**Experiment 3: All Word Types (616 word types)**		
	Errors	**Fault Rate**	**Accuracy**	**Errors**	**Fault Rate**	**Accuracy**
Khoja’s Stemmer	311	31.8%	68.2%	224	36.36%	63.64%
BAMA	419	42.8%	57.16%	267	43.34%	56.66%
Triliteral	394	40.3%	59.71%	266	43.18%	56.82%
Voting Exp.1	434	44.4%	55.6%	242	39.3%	60.7%
Voting Exp.2	405	41.4%	58.6%	219	35.6%	64.4%
	Experiment 2: Tokens excluding Stop words (554 tokens)			**Experiment 4: Word Types excluding Stop words (451word types)**		
Khoja’s Stemmer	209	37.73%	62.27%	155	34.37%	65.63%
BAMA	325	58.66%	41.34%	251	55.65%	44.34%
Triliteral	279	50.36%	49.64%	214	47.45%	52.55%
Voting Exp.1	266	48.0%	52.0%	174	38.6%	61.4%
Voting Exp.2	229	41.3%	58.7%	151	33.5%	66.5%

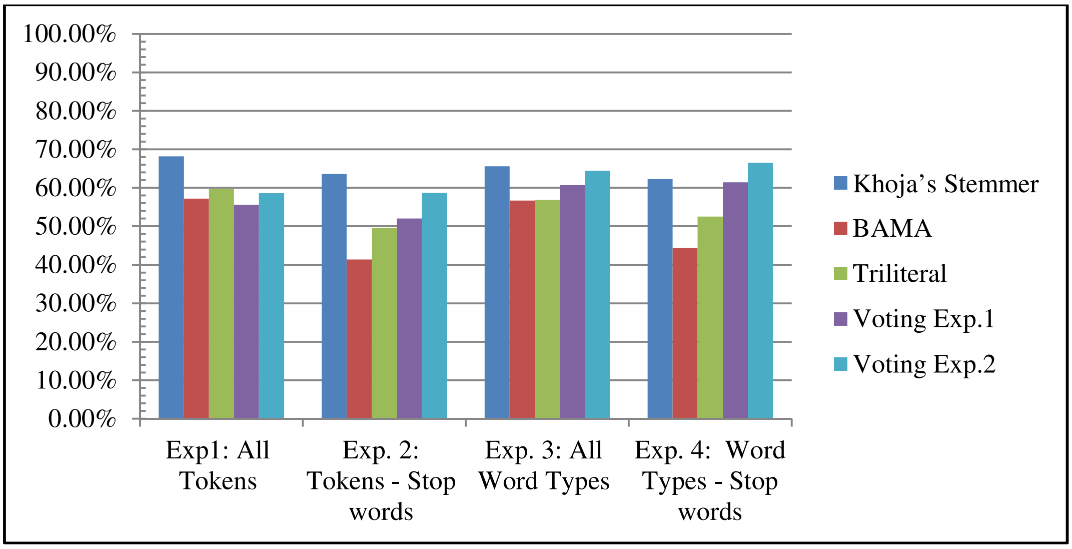


Figure 3.3 Accuracy rates resulting from the four different experiments for the Qur'an test document

Word	Khoja's stemmer	BAMA	Triliteral	Voting Exp. 1	Voting Exp. 2	Gold Standard	
الم	ألم	ألم	الم	ألم	ألم	ألم	Stop word
أَحَسِبَ	حسب	حسب	حسب	حسب	حسب	حسب	
النَّاسُ	نوس	ناس	ناس	ناس	ناس	ناس	
أَن	أن	إن	أن	أن	أن	أن	Stop word
يُتْرَكُوا	ترك	ترك	ركو	ترك	ترك	ترك	
أَن	أن	إن	أن	أن	أن	أن	Stop word
يَقُولُوا	قول	قال	يقولوا	يقولوا	قول	قول	
آمَنَّا	منا	آمن	آمن	آمن	آمن	آمن	
وَهُمْ	وهم	وهم	وهم	وهم	وهم	وهم	Stop word
لَا	لا	لا	لا	لا	لا	لا	Stop word
يُفْتَنُونَ	فتن	فتن	فنن	فتن	فتن	فتن	

Figure 3.4 Sample output of the three algorithms, the voting experiments and the gold standard of the Qur'an test document

The results shown in table 3.2 and figure 3.3 are computed by running the four experiments using the Qur'an text sample. The results of each stemming and voting algorithm in the four experiments are compared against the gold standard roots, and then accuracy rates are computed. In experiment 1 containing all word tokens, Khoja's stemmer achieved the highest accuracy of 68.2%. The triliteral root extraction algorithm and BAMA achieved quite similar results of 59.71% and 57.16% respectively. Neither voting experiment achieved better accuracy rates: 55.6% for voting experiment 1 and 58.6% for voting experiment 2.

In the second experiments excluding stop words, Khoja's stemmer scored the highest accuracy at 62.27%, then the triliteral root extraction algorithm at 49.64%, and finally BAMA at 41.34%. The voting algorithm scored 58.7% in voting experiment 1 and 55.6% in voting experiment 2.

The third experiment compares the results of each algorithm with respect to word-type roots. Khoja's stemmer achieved the highest accuracy at 63.64%. Triliteral root extraction algorithm and BAMA achieved similar accuracy rates of 56.82% and 56.66% respectively. The voting algorithm in this experiment performed better and achieved an accuracy of 64.40% for voting experiment 2 and 60.70% for voting experiment 1. Voting experiment 2 outperforms the best algorithm results by 0.76%.

The final experiment evaluates word-type accuracy excluding stop words. Khoja's stemmer achieved the highest accuracy rate at 65.63%. The triliteral root extraction algorithm achieved 52.55%, and finally BAMA achieved 44.34%. The voting algorithm achieved better results at 66.5% and 61.4% for voting experiment 2 and voting experiment 1 respectively. Voting experiment 2 outperforms the best algorithm results by 0.87%.

In summary, Khoja's stemmer achieved the highest accuracy rate at 68.2% in experiment 1. The rank of the stemming algorithms is Khoja's stemmer, then triliteral root extraction algorithm, and finally BAMA. The voting algorithm of the voting experiment 2 outperforms the best algorithm results by about 0.8% in experiments 3 and 4.

Table 3.3 Tokens and word types accuracy of the 3 stemming algorithms and voting algorithms tested on CCA sample

	Experiment 1: All Tokens (1005 tokens)			**Experiment 3: All Word Types (710 word types)**		
Algorithm	**Errors**	**Fault Rate**	**Accuracy**	**Errors**	**Fault Rate**	**Accuracy**
Khoja's Stemmer	231	22.99%	77.01%	232	32.68%	67.32%
BAMA	596	59.30%	40.70%	431	60.70%	39.30%
Triliteral	234	23.28%	76.72%	253	35.63%	64.37%
Voting Exp.1	303	30.15%	69.85%	248	34.93%	65.07%
Voting Exp.2	266	26.47%	73.53%	215	30.28%	69.71%
	Experiment 2: Tokens excluding Stop words (766 tokens)			**Experiment 4: Word Types excluding Stop words (640 word types)**		
Khoja's Stemmer	212	27.7%	72.3%	184	28.75%	71.25%
BAMA	431	60.70%	39.30%	423	66.09%	33.91%
Triliteral	253	35.63%	64.37%	224	35.00%	65.00%
Voting Exp.1	303	39.56%	60.44%	252	39.4%	60.6%
Voting Exp.2	266	34.73%	65.27%	195	30.5%	69.5%

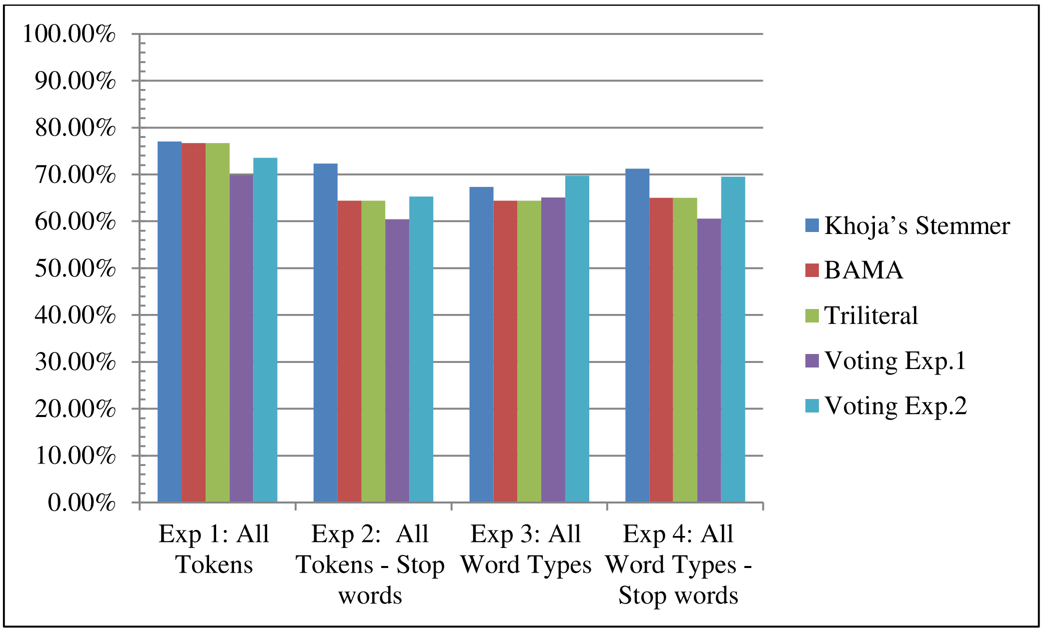


Figure 3.5 Accuracy rates results of the four different experiments for the CCA test document

Word	Khoja's stemmer	BAMA	Triliteral roots alg.	Voting Exper. 1	Voting Exper. 2	Gold Standard	
ستبقى	بقي	بقي	بقى	بقي	بقي	بقى	
العولمة	عولمة	عولمة	علم	عولمة	عولمة	علم	
وإلى	إلى	إلى	إلى	إلى	إلى	إلى	Stop Word
وقت	وقت	وقت	وقت	وقت	وقت	وقت	
ممتد	ممتد	ممتد	متد	ممتد	ممتد	مدَّ	
مثيرة	ثور	مثير	مثر	ثور	ثور	ثار	
للأسئلة	سول	سؤال	أسل	سؤال	سول	سأل	
والأجوبة	جوب	جواب	أجب	أجب	جوب	جوب	

Figure 3.6 Sample output of the three algorithms, the voting experiments and the gold standard of the CCA test document

The results shown in table 3.3 and figure 3.5 are computed by running the four experiments using the CCA text sample. The results of each stemming and voting algorithm in the four experiments are compared against the gold standard's roots, and then accuracy rates are computed.

In experiment 1 containing all tokens, Khoja's stemmer achieved the highest accuracy at 77.01%. The triliteral root extraction algorithm achieved 76.72%, and finally BAMA achieved 40.70%. Neither voting experiments achieved better accuracy rates: 69.85% for voting experiment 1 and 73.53% for voting experiment 2.

In the second experiment excluding stop words, Khoja's stemmer scored the highest accuracy at 72.30%, then the triliteral root extraction algorithm at 64.37%, and finally

BAMA at 39.30%. The voting algorithm scored 60.44% in voting experiment 1 and 65.27% in voting experiment 2.

The third experiment compares the results of each algorithm by word-type, Khoja's stemmer achieved the highest accuracy at 67.32%, then the triliteral root extraction algorithm at 64.37%, then BAMA at 39.30%. The voting algorithm in this experiment performed better and achieved 69.71% for voting experiment 2 and 65.07% for voting experiment 1. Voting experiment 2 outperforms the best algorithm results by 2.39%.

The final experiment excludes stop words when comparing word-type roots, Khoja's stemmer achieved the highest accuracy rate at 71.25%, then the triliteral root extraction algorithm at 65.00%, and finally BAMA at 33.91%. The voting algorithm achieved better accuracy rates, 69.50% and 60.60%, for voting experiment 2 and voting experiment 1 respectively.

In summary, Khoja's stemmer achieved the highest accuracy rate at 77.01% in experiment 1. The rank of the stemming algorithms is Khoja's stemmer, then triliteral root extraction algorithm, and finally BAMA. The voting algorithm of voting experiment 2 outperforms the best algorithm results by 2.39% in experiment 3.

3.6 Comparative Evaluation Conclusions

This study compared three existing stemming algorithms: Khoja's stemmer, BAMA and the Triliteral root extraction algorithm. Results of the stemming algorithms were compared with the gold standard of classical and MSA text samples of 1,000 words each. Four experiments were performed to fairly and accurately compare the outputs of the three different stemming algorithms and morphological analysis for Arabic text. The four experiments on both text samples show the same accuracy rank for the stemming algorithms: Khoja's stemmer achieved the highest accuracy then the triliteral root extraction algorithm and finally BAMA. Khoja's and the triliteral stemming algorithms generate only one result analysis for each input word, while BAMA generates one or more result analysis.

The voting algorithm achieves about 62% average accuracy for Qur'an text and about 70% average accuracy for newspaper text. The results show that the stemming algorithms used in the experiments work better on MSA text (*i.e.* newspaper text) than classical Arabic (*i.e.* Qur'an text), not unexpectedly as they were originally designed for stemming MSA text (*i.e.* newspaper text).

All stemming algorithms involved in the experiments agreed and generate correct analysis for simple roots that do not require detailed analysis. So, more detailed analysis and enhancements are recommended as future work.

Most stemming algorithms are designed for information retrieval systems where accuracy of the stemmers is not such an important issue. On the other hand, accuracy is vital for natural language processing. The accuracy rates show that even the best algorithm failed to achieve accuracy of more than 75%. This proves that more research is required, as Part-of-Speech tagging and then Parsing cannot rely on such stemming algorithms because errors from the stemming algorithms will propagate to such systems.

The experiments are limited to the three stemming algorithms. Other algorithms are not available freely on the web, and it is hard to acquire them from the authors. Open-source development of resources is important to advance research on Arabic NLP.

3.7 Analytical Study of Arabic Triliteral Roots

To understand the nature of Arabic roots, and the derivation process of words, triliteral roots are classified into 22 groups depending on the internal structure of the root itself; whether it contains only consonant letters, *hamza*^{h}, or defective letters (Dahdah 1987; Wright 1996; Al-Ghalayyni 2005; Ryding 2005). Section 6.2.21 discusses the classification of triliteral roots. Arabic triliteral root distribution is studied over the 22 categories by analyzing real text corpora: the Qur'an as corpus, which contains 45,534 triliteral-root words (*i.e.* not including function words which do not have triliteral roots such as demonstrative pronouns *e.g.* هٰذَا *hāḏā* "this", and words with quadriliteral roots such as دَرَاهِم *darāhim* "dirhams" from the root د-ر-هـ-م *d-r-h-m*, or quinquilitiral roots). This is an example of a natural corpus where words are repeated in different contexts; and 376,167 word types, derived from triliteral roots, an example of a dictionary of Arabic where each word of the test sample occurs once. Chapter 4 will discuss the processing steps, statistics and evaluation of the broad-coverage lexical resource the SALMA – ABCLexicon.

3.7.1 A Study of Triliteral Roots in the Qur'an

In general it is said that an Arabic word has a root of 3 consonants. However, there are many exceptions which cause problems for analysis. *hamza*^{h} is a special letter which is not a normal consonant but can appear in a root. Also, a few roots include vowels, and these are called "defective". Sometimes a consonant is doubled, and this also cause ambiguity in analysis.

The results show that 68% of the triliteral roots of Qur'an and 61% of the Qur'an words are derived from triliteral roots, mainly intact roots which are represented in categories 1 to 5 in table 3.4. 29% of the triliteral roots of Qur'an are defective roots (*i.e.* they contain one or two vowels in - their root) represented in categories 6-11 in table 3.4.The percentage of the words belonging to this category is 32% of the words of the Qur'an. The third category contains one or two vowels and *hamza*^{h} in its root, represented in categories 12-22 in table 3.4. The percentage of such triliteral roots of the Qur'an is 3%, and 7% of the words of the Qur'an belong to this category. Table 3.5 and figure 3.7 show the distribution of the Qur'an's words and roots into the three main root categories.

Table 3.4 Category distribution of Roots-Types and Word-Tokens extracted from the Qur'an

	Category				**Roots-Types**		**Word-Tokens**	
					count	**Percentage**	**count**	**Percentage**
1	Sound	C1	C2	C3	870	54.04%	20,007	43.94%
2	Doubled	C1	C2	C2	136	8.45%	3,814	8.38%
3	Initially-hamzated	H	C2	C3	44	2.73%	3,243	7.12%
4	Medially-hamzated	C1	H	C3	15	0.93%	281	0.62%
5	Finally-hamzated	C1	C2	H	32	1.99%	459	1.01%
6	Initially-defective	V	C2	C3	70	4.35%	1,252	2.75%
7	Medially-defective	C1	V	C3	198	12.30%	8,162	17.93%
8	Finally-defective	C1	C2	V	167	10.37%	3,584	7.87%
9	Separated doubly-weak	V	C2	V	12	0.12%	710	1.56%
10	Finally-adjacent doubly-weak	C1	V1	V2	19	1.18%	473	1.04%
11	Initially-adjacent doubly-weak	V1	V2	C3	2	0.12%	445	0.98%
12	Initially-hamzated and doubled	H	C2	C2	7	0.43%	175	0.38%
13	Initially-defective and Doubled	V	C2	C2	2	0.12%	40	0.09%
14	Initially-hamzated and finally-defective	H	C2	V	13	0.81%	958	2.10%
15	Initially-hamzated and medially-defective	H	V	C3	6	0.37%	153	0.34%
16	Adjacent doubly-weak and initially-hamzated	H	V1	V2	2	0.12%	418	0.92%
17	Finally-defective and medially-hamzated	C1	H	V	2	0.12%	330	0.72%
18	Separated doubly-weak and medially-hamzated	V1	H	V2	0	0.00%	0	0.00%
19	Initially-defective and medially-hamza	V	H	C3	3	0.19%	15	0.03%
20	Medially-defective and finally-hamzated	C1	V	H	8	0.50%	998	2.19%
21	Initially-defective and finally-hamzated	V	C2	H	2	0.12%	17	0.04%
22	Adjacent doubly-weak and finally-hamzated	V1	V2	H	0	0.00%	0	0.00%
Totals					1610	100.00%	45,534	100.00%

Table 3.5 Summary of category distribution of root and tokens of the Qur'an

Category	**Root**		**Tokens**	
	Total	**Percentage**	**Total**	**Percentage**
Intact	1097	68.14%	27,804	61.06%
Defective	468	29.07%	14,626	32.12%
Defective and hamzated	45	2.80%	3,104	6.82%
Totals	1610	100.00%	45,534	100.00%

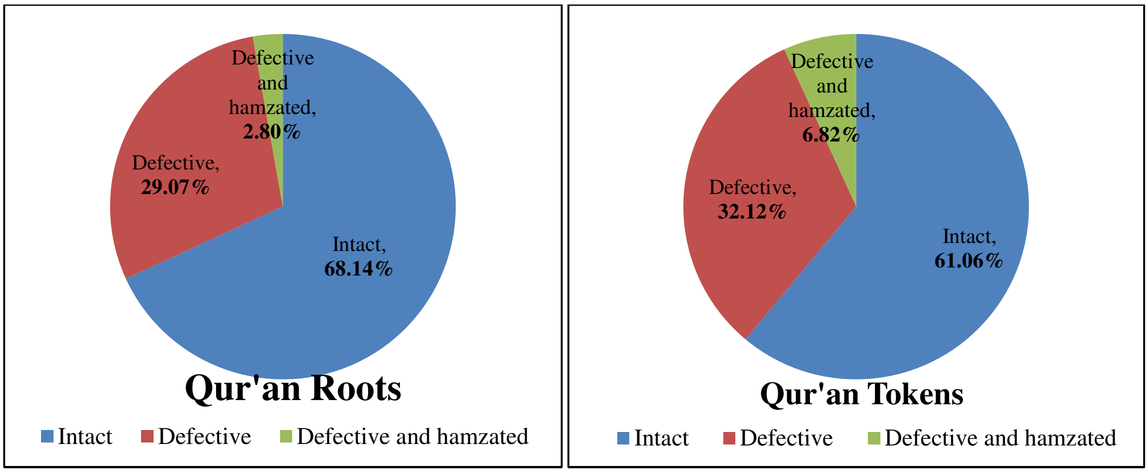


Figure 3.7 Root distribution (left) and word distribution (right) of the Qur'an

3.7.2. A Study of Triliteral Roots in Traditional Arabic Lexicons

Similar root and word distributions were obtained from the roots and the word types stored in the broad-coverage lexical resource. About 63% of the roots stored in the broad-coverage lexical resource are intact words, categories 1-5 in table 3.6, and slightly more than 68% of the word types belong to this category. Defective roots represented by categories 6-11 in table 3.6, form about 33% of the roots of the broad-coverage lexical resource and 29% of the word types belong to this category. Finally, defective and hamzated roots, represented by categories 12-22 in table 3.6, of the broad-coverage lexical resource are approximately 4% of roots, and about 2% of the word types belong to this category. Figure 3.8 and table 3.7 show the root and word types distribution after analyzing the broad-coverage lexical resource.

Table 3.6 Category distribution of Root and Word type extracted from the lexicon

	Category				Root		Word Type	
					Count	**Percentage**	**Types**	**Percentage**
1	Sound	C1	C2	C3	4147	48.78%	201,385	53.54%
2	Doubled	C1	C2	C2	446	5.25%	32,007	8.51%
3	Initially-hamzated	H	C2	C3	289	3.40%	10,449	2.78%
4	Medially-hamzated	C1	H	C3	216	2.54%	3,909	1.04%
5	Finally-hamzated	C1	C2	H	270	3.18%	8,985	2.39%
6	Initially-defective	V	C2	C3	386	4.54%	19,219	5.11%
7	Medially-defective	C1	V	C3	1115	13.11%	43,512	11.57%
8	Finally-defective	C1	C2	V	1151	13.54%	41,295	10.98%
9	Separated doubly-weak	V	C2	V	45	0.08%	2,372	0.63%
10	Finally-adjacent doubly-weak	C1	V1	V2	106	1.25%	4,057	1.08%
11	Initially-adjacent doubly-weak	V1	V2	C3	22	0.26%	211	0.06%
12	Initially-hamzated and doubled	H	C2	C2	30	0.35%	888	0.24%
13	Initially-defective and Doubled	V	C2	C2	29	0.34%	463	0.12%
14	Initially-hamzated and finally-defective	H	C2	V	74	0.87%	2,111	0.56%
15	Initially-hamzated and medially-defective	H	V	C3	47	0.55%	892	0.24%
16	Adjacent doubly-weak and initially-hamzated	H	V1	V2	7	0.08%	135	0.04%
17	Finally-defective and medially-hamzated	C1	H	V	42	0.49%	1,041	0.28%
18	Separated doubly-weak and medially-hamzated	V1	H	V2	2	0.02%	52	0.01%
19	Initially-defective and medially-hamza	V	H	C3	15	0.18%	292	0.08%
20	Medially-defective and finally-hamzated	C1	V	H	42	0.49%	1,590	0.42%
21	Initially-defective and finally-hamzated	V	C2	H	21	0.25%	1,302	0.35%
22	Adjacent doubly-weak and finally-hamzated	V1	V2	H	0	0.00%	0	0.00%
Totals					8502	100.00%	376,167	100.00%

Table 3.7 Summary of category distribution of root and word types of the lexicons

Category	Root		Word Types	
	Total	**Percentage**	**Total**	**Percentage**
Intact	5368	63.30%	256,735	68.25%
Defective	2803	33.05%	110,666	29.42%
Defective and hamzated	309	3.64%	8,766	2.33%
Totals	8480	100.00%	376,167	100.00%

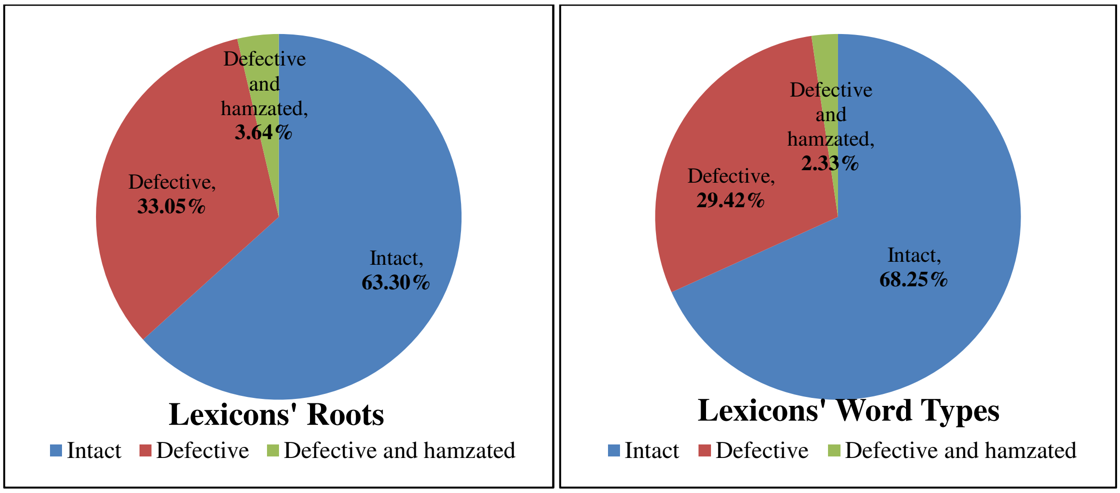


Figure 3.8 Root distribution (left) and Word type distribution (right) of the broad-lexical resource

3.7.3 Discussion of the Analytical Study of Arabic Triliteral Roots

The above analysis gives a clear picture of the distribution of the 22 categories and 3 broad categories of triliteral roots, words and word types. The study clearly shows that about a third of any Arabic text words have roots belonging to defective or defective and hamzated root categories. Words belonging to these two root categories are hard to analyze and the root extraction process for such words always has higher error rates than words belonging to the intact root category. Stemming and morphological analyzers are subject to mistakes when analyzing words belonging to these two broad categories.

Similar distribution results were obtained by analyzing the Qur'an's roots and words and the broad-coverage lexicon roots and word types. About 65% of roots, words and word types belong to intact triliteral roots. About 35% of the roots, words and word types are classified into the defective triliteral root category. Finally, 5% of the roots, words and word types belong to the defective and hamzated triliteral root category.

These figures prove that any successful stemming and morphological analysis system has to deal with issues specific to Arabic word derivation such as: incorporation, substitution and deletion of a weak vowel letter. Moreover, dealing with orthographic issues such as *hamza*^{h} in writing is critical for stemming and morphological analysis of Arabic text. Root extraction accuracy of any stemming or morphological analysis which does not deal with these special language specifications will not achieve an accuracy rate more than 65% in the best case.

A question raised in this context is: how to improve stemming and morphological analysis so the algorithm can deal successfully with the hard cases of the 35% of words belonging to defective and defective and hamzated triliteral root categories? Two methodologies can be followed; either building a sophisticated algorithm that deals with the hard cases or simply by providing the algorithm with a prior-knowledge broad-coverage lexical resource that contains most of the hard case words and their triliteral roots. Then the stemming algorithm will look up the word to be analyzed in the lexicon and get the correct analysis for that word. A look-up methodology is needed here.

Chapter 4 discusses the motivation and the processing steps in constructing the prior-knowledge broad-coverage lexical resource the SALMA-ABCLexicon[30]. The lexicon was constructed by analyzing the text of 23 traditional Arabic lexicons which are freely available open-source documents (PDF and MS-Word files). The main purpose of constructing the SALMA-ABCLexicon was to improve the morphological analysis of Arabic text. Constructing a broad-coverage lexical resource to improve the accuracy of Arabic morphological analysis has advantages over developing a sophisticated stemming algorithm. These advantages are discussed in detail in section 4.4. The constructed lexicon has about half a million different Arabic words which covers 85% or more of any Arabic text.

3.8 Summary and Conclusions

Arabic morphological analysers and stemming algorithms have become a popular area of research. Several computational linguists have designed and developed algorithms to solve the problems of morphology and syntax. Stemming algorithms have been developed for many languages including Arabic. Several stemming algorithms for Arabic already exist, but each researcher proposed an evaluation methodology based on different text corpora. Therefore, we cannot make direct comparisons between these evaluations. This chapter discussed four different fair and precise evaluation experiments using a gold standard for evaluation consisting of two 1000-word text documents from the Holy Qur'an and the Corpus of Contemporary Arabic. The selection of the stemming algorithms was limited to the algorithms where we have ready access to the implementation and/or results. The three selected algorithms are Khoja's stemmer (Khoja 2003), Buckwalter's morphological Analyzer (BAMA) (Buckwalter 2002) and Al-Shalabi et. al, triliteral root extraction algorithm (Al-Shalabi et al. 2003). A reuse of the results of the three algorithms in a voting program was developed to allow "voting" on the analysis of the three stemming algorithms.

The four experiments on both text samples show the same accuracy rank for the stemming algorithms: Khoja's stemmer achieved the highest accuracy then the triliteral root extraction algorithm and finally BAMA. The results show that the stemming algorithms used in the experiments work better on MSA text (*i.e.* newspaper text) than

[30] SALMA-ABCLexicon (Sawalha Atwell Leeds Morphological Analysis – Arabic Broad-Coverage Lexicon) http://www.comp.leeds.ac.uk/cgi-bin/scmss/arabic_roots.py

classical Arabic (*i.e.* Qur'an text), not unexpectedly as they were originally designed for stemming MSA text (*i.e.* newspaper text). All stemming algorithms involved in the experiments agreed and generated correct analyses for simple roots that do not require detailed analysis. So, more detailed analysis and enhancements are recommended as future work. Most stemming algorithms are designed for information retrieval systems where accuracy of the stemmers is not such an important issue. On the other hand, accuracy is vital for natural language processing. The accuracy rates show that even the best algorithm failed to achieve accuracy rate of more than 75%. This proves that more research is required, as Part-of-Speech tagging and then Parsing cannot rely on such stemming algorithms because errors from the stemming algorithms will propagate to such systems.

A clear image of the percentage of triliteral roots, words and word types distribution on 22 categories of triliteral roots was presented. The study clearly showed that about one third of Arabic text words have roots belonging to the defective or defective and hamzated root categories. Words belonging to these two root categories are hard to analyze and the root extraction process of such words always has higher error rates than for words belonging to the intact root category. Existing stemming and morphological analyzers are subject to mistakes when analyzing words belonging to these two categories.

The construction of a broad-coverage lexical resource to improve the accuracy of Arabic morphological analysis was proposed as a practical solution. Chapter 4 will discuss the motivation and the processing steps in constructing the prior-knowledge broad-coverage lexical resource, the SALMA-ABCLexicon. The lexicon is constructed by analyzing the text of 23 traditional Arabic lexicons which are freely available open-source documents. The main purpose of constructing the SALMA-ABCLexicon is to improve morphological analysis of Arabic text. The constructed lexicon has about half a million different Arabic words, which covers about 85% of any Arabic text.

Chapter 4
The SALMA-ABCLexicon: Prior-Knowledge Broad-Coverage Lexical Resource to Improve Morphological Analyses

This chapter is based on the following sections of published papers:

Sections 1, 2, 3, 4, 5 and 6 are based on section 1, 2, 3, 4, 5, 6, and 7 in (Sawalha and Atwell 2010a)

Chapter Summary

Broad-coverage language resources which provide prior linguistic knowledge must improve the accuracy and the performance of NLP applications. A broad-coverage lexical resource, the SALMA ABCLexicon (Sawalha Atwell Leeds Morphological Analysis Arabic Broad-Coverage Lexicon) was constructed to improve the accuracy of morphological analyzers and part-of-speech taggers of Arabic text. Over the past 1200 years, many different kinds of Arabic language lexicons have been constructed; these lexicons are different in ordering, size and aim of construction. 23 machine-readable lexicons, which are freely available on the web as portable document format (.pdf) or MS-Word (.doc) documents, were collected. Lexical resources were combined into one large broad-coverage lexical resource, the SALMA-ABCLexicon, by extracting information from disparate formats and merging traditional Arabic lexicons. The construction process followed agreed criteria for constructing morphological lexical resources from raw text.

To evaluate the broad-coverage lexical resource, coverage was computed over the Qur'an, the Corpus of Contemporary Arabic, and a sample from the Arabic Internet Corpus, using two methods. Counting exact word matches between test corpora and lexicon scored about 65-68%; Arabic has a rich morphology with many combinations of roots, affixes and clitics, so about a third of words in the corpora did not have an exact match in the lexicon. The second approach is to compute coverage in terms of use in a lemmatizer program, which strips clitics to look for a match for the underlying lexeme; this scored about 82-85%.

4.1 Introduction

Lexicography is the applied part of lexicology. It is concerned with collating, ordering of entries, derivations and their meaning depending on the aim of the lexicon to be constructed and its size. Lexicography is defined as *"…the branch of applied linguistics concerned with the design and construction of lexica for practical use.*" (Eynde and Gibbon 2000). On the other hand, lexicology is defined as *"…the branch of descriptive linguistics concerned with the linguistic theory and methodology for describing lexical information, often focusing specifically on issues of meaning.*" (Eynde and Gibbon 2000). Long-term efforts in lexicographic projects have greatly accelerated since the advent and use of computers: this is known as computational lexicography. However, constructing a large-scale broad-coverage lexicon involves time-consuming development of specifications, design, collection of lexical data, information structuring, and user-oriented presentation formatting (Eynde and Gibbon 2000).

A realistic and useful lexicon for NLP requires an efficiently stored machine-readable database with a large number of words with associated syntactic and semantic information (Russell et al. 1986). Morphological lexicons are based on the idea of generating all possible combinations of morphemes. But filtering out the non-established, yet theoretically possible combinations of morphemes is the major problem of lexicon generation (Tadi and Fulgosi 2003). Morphological lexicons are useful for many natural language applications such as: spelling and syntactic checkers integrated to word processing applications, development of morphological and syntactic analyzers, search engines, machine translation, information filtering and extraction systems, etc. (Petasis et al. 2001). Morphosyntactic lexicons are valuable resources for many NLP applications. However, these lexicons need to meet certain specifications such as high coverage; high level of quality; directly reusable in NLP tools; and freely-available to potential users (Sagot 2010).

4.1.1 Morphological Lexicons of Other Languages

Morphological lexicons exist for many languages. The Special Interest Group on the Lexicon of the Association for Computational Linguistics (ACL SIGLEX) maintains an online comprehensive list of lexical resources[31]. The lists and files with linguistic information include: Brown Corpus Lexicon of 52,000 words; the XTAG project with an associated 300,000 word English lexicalized grammar; COMLEX (COMmon LEXicon) a monolingual English Dictionary consisting of 38,000 head words; the Oxford Text Archive (OTA) of machine readable dictionaries for many languages; Adam Kilgarriff's list of 6,318 most frequent lemmas extracted from the British National Corpus; The Moby

[31] Online lexical resources by ACL SIGLEX http://www.clres.com/online.html

lexicon project consisting of sub-lexicons including Moby Hyphenator (185,000 entries), Moby Part-of-Speech (230,000 entries), Moby Thesaurus (30,000 entries) and Moby Words (610,000 words and phrases); Upper Cyc Ontology containing about 3,000 words capturing the most general concepts of human consensus reality.

Russell, Pulman et al. (1986) developed a dictionary and morphological analyzer for English. They assumed that correct syntactic analyses are built in to the lexical entries, but allowing adaptation by users to suit different analyses. The morphological lexicon itself consists of a sequence of entries, each in the form of a Lisp s-expression which consists of five elements: first, the head word in written form; second, the head word in phonological transcription; third, a syntactic field consisting of a syntactic category; fourth, a semantic field providing the facility for users and any Lisp s-expression to be inserted in it; and finally, a user field which allows users to include additional information they desire. The prototype lexicon contains about 3,500 entries.

MULTEXT lexicons[32] are part of the MULTEXT project, which aims to develop tools, corpora, and linguistic resources for a wide variety of languages. The MULTEXT lexicons include four developed lexicons for German, Italian, Spanish and French. The lexicons are stored in tab separated column files where the first column represents the word form, the second column represents the lemma and the last column represents the lexical tag.

MULTEXT-East[33] language resources are multilingual datasets for language engineering focused on the morphosyntactic level of linguistic description. These resources cover 16 languages of mainly central and eastern Europe and include the EAGLES-based morphosyntactic specifications and morphosyntactic lexica. MULTEXT-East followed the same lexicon format as the original MULTEXT lexicons. The size of MULTEXT-East lexicons ranges from 13,006 entries for Persian to 2,461,491 entries for Slovak (Erjavec 2010).

The Croatian Morphological Lexicon (CML) is a lexicon developed to make a model of the Croatian morphological system. The CML has two sub-lexicons: derivative/compositional (*i.e.* a list of lexical and a list of derivational morphemes with rules for combining) and inflectional (*i.e.* a list of generated stems and a list of inflectional morphemes with rules for combining) which are produced by two morphological generators according to morphotactic rules. The CML followed the same lexicon format as MUTEXT-East. The CML contains 36,000 lemmas extracted from the Croatian dictionary. Then the generation of word forms generated 171,308 nouns, 232,276 verbs, 1,207,786 adjectives and 11,706 adverbs (Tadi and Fulgosi 2003).

[32] MULTEXT Lexicons http://aune.lpl.univ-aix.fr/projects/multext/MUL5.html

[33] MULTEXT-East http://nl.ijs.si/ME/V4/

A large-scale Greek morphological lexicon was developed by the Software and Knowledge Engineering Laboratory (SKEL) to be used to develop a lemmatizer and morphological analyzer in a controlled language checker for Greek. The SKEL lexicon is organized into two components: the query component which aims to facilitate the query of the lexicon about specific form and retrieve the associated linguistic information; and the generation component responsible for generating all possible word forms for a given lemma. The generation component also utilizes language specific rules regarding syllabication and accentuation. The morphological database consists of a fixed number of pages, where each page contains a set of morphological entries. Each entry contains a fixed number of morphological features such as lemma, stem, suffix, syllabication, part-of-speech and other morphological features such as number, inflectional type, gender, case, inflection, tense, person, voice, mood, etc. The SKEL lexicon contains 60,000 unique lemmas which generate 710,000 word forms. The morphological database contains about 2,500,000 morphological entries (Petasis et al. 2001).

A Latvian lexicon was developed as part of a lexicon-based morphological analyzer for Latvian which is an implementation of word inflection based on a stem and its properties already stored in the lexicon. The lexicon’s core data are the dictionary’s lexical units, which contain word stems, their morphological types and any other linguistic information related to the stems. The lexicon contains about 27,000 stems. The coverage of the lexicon is scored at 85%-90% after analyzing an unrestricted text corpus. A heuristic, based on last letter of the analyzed word, is integrated with the morphological analyzer for guessing the part-of-speech of the remaining uncovered percentage of words. XML files are used to store the lexicon and other data files (Paikens 2007).

A freely-available and wide-coverage morphosyntactic lexicon for French Le*fff*[34] (Lexique des formes fléchies du français – Lexicon of French inflected forms) is used in many NLP tools including large-coverage parsers. The Le*fff* uses the Alexina framework to ensure reusability of the lexicon in many NLP tools. Alexina is a lexical modelling and acquisition framework for both the morphological and syntactic levels, which is a language and grammatical formalism independent and compatible with Lexical Markup Framework (LMF) standards. The Alexina lexicon consists of entries (*i.e.* lexemes) where each entry is associated with a lemma, a category and an inflectional class. The Le*fff* (3.0.1) contains 536,375 entries corresponding to 110,477 lemmas covering the grammatical categories of verbs, verbal idioms, nouns, adjectives, adverbs, prepositions, proper nouns and others. The Le*fff* is evaluated by a quantitative comparison with other existing lexical resources for French. It has also been evaluated in terms of its use in POS tagger and deep parser. Integrating *Lefff* in a maximum-entropy-based part-of-speech

[34] Le*fff* http://www.labri.fr/perso/clement/lefff/

tagger for French trained on the French Treebank increased the accuracy from 97.0% (86.1% for unknown words) to 97.7% (90.1% for unknown words) (Sagot et al. 2006; Nicolas et al. 2008; Sagot 2010).

Sagot (2005) developed a lexicon for Slovak from a raw corpus and a morphological description of the language. Both inflectional and derivational morphology are used to enhance the accuracy (recall and precision) and to acquire the derivational relations in the lexicon. A three-step procedure is followed for the acquisition of the lexicon. First, given the morphological description of the language, build all possible lemmas that can possibly explain the inflected forms in the lexicon. Second, rank the lemmas according to their likelihood in the corpus. Finally, best ranked lemmas are manually validated. A claim is stated that this methodology can be used for morphologically rich languages. The acquired lexicon following this methodology contains 2,000 lemmas generating more than 50,000 inflected forms (Sagot 2005).

A morphological analyzer and language specific web crawler (*i.e.* a tool used to collect a list of word types) have a potential to enhance lexical resources for morphologically rich but resource-poor languages such as Tigrinya. Tigrinya is an Ethio-Semitic language spoken by about 6 million people in the Tigray region of northern Ethiopia and in central Eritrea. The web crawler collected a list of 227,984 word types. Then, the list was filtered and passed to the morphological analyzer. 65,732 words succeed the lexical analysis, and 46,979 words have at least one analysis generated by the guesser analyzer (Gasser 2010).

In summary, many existing morphological lexicons were constructed from raw text (Sagot 2005). The general requirements for constructing a morphological lexicon from raw text are:

- A representative corpus.
- A generation program or a morphological description of the language.
- A Lexical Markup Framework (LMF) for providing compatible structure to store the lexical entries to ensure reusability of the lexicon in many NLP tools.
- A searching facility over the lexical entries (querying the constructed lexicon).
- An evaluation methodology for the morphological lexicons, by computing the coverage of the lexicon, and by measuring the accuracy gained after integrating the lexicon to a NLP application such as part-of-speech tagger or syntactic parser.

4.1.2 Morphological Lexicons for Arabic

A morphological analyzer for Arabic (BAMA) (Buckwalter 2002; Buckwalter 2004) contains three Arabic-English lexicon files: a prefixes file containing 299 entries, a suffixes file containing 618 entries, and a stems file containing 82,185 entries representing 38,600 lemmas; see section 3.2.2. The lexicon component of BAMA is reused in other Arabic NLP tools such as the large-scale lexeme-based Arabic morphological generation Aragen (Habash 2004), and spell checking lexicons such as Duali[35], Baghdad[36] and Arabic-spell[37].

The AyaSpell[38] project aims to develop open-source resources for Arabic NPL including Arabic spell checker. The shortage of existing Arabic spell checkers comes from the lexicon they depend on. A lexicon is developed to support the AyaSpell checker. The lexicon consists of two components: the vocabulary list built by analyzing 5 traditional Arabic lexicons; and the affixes and morphological rules list. Each entry in the vocabulary list has its morphological description associated with it. The vocabulary list contains more than 50,000 entries distributed on more than 10,000 verbs and more than 40,000 nouns, particles and residuals (Zarrouki and Kebdani 2009; Zerrouki and Balla 2009).

WordNet is a broad coverage lexical resource which is developed to support many information retrieval applications. The basic idea behind WordNet is that knowledge of words is represented by meanings and the context in which they occur. The desired conceptual information is provided by linking words to appropriate concepts. Concepts in the WordNet are the organizational units. They can be single words, compounds, collocations, idiomatic phrases and phrasal verbs. The foundation of the Global WordNet Association and the Global WordNet project coordinates the production and the linkage of wordnets for all languages of the world including Arabic (Elkateb, Black and Farwell 2006).

Arabic WordNet (AWN) is a lexical resource for MSA which is based on the design and the contents of the Princeton WordNet (PWN) for English. The AWN is constructed following the same methods developed for Euro WordNet, which is compatible with other wordnets and focuses on manual encoding of the most complicated and important concepts. The AWN structure consists of four principal structures. First, the *items* represent conceptual entities including synsets, ontology classes and instances. Second, a *word* entity represents a word sense. Third, a *form* entity contains lexical information.

[35] Duali Arabic spell-checker http://www.arabeyes.org/project.php?proj=Duali

[36] Bahghdad Arabic spell checker http://home.foolab.org/cgi-bin/viewcvs.cgi/projects/baghdad/

[37] Arabic-spell http://sourceforge.net/projects/arabic-spell/

[38] AyaSpell Arabic spell checker http://ayaspell.sourceforge.net/index.php

Fourth, a *link* connects in a relation two items. The AWN is stored using XML files and relational database implemented by MySQL. 1,000 terms and 4,000 definition statements are the contents of the large ontology which is built to provide the semantic background for the AWN (Elkateb and Black 2001; Black and El-Kateb 2004; Elkateb et al. 2006; Rodríguez et al. 2008).

Arabic Verbnet is a large coverage verb taxonomy for Arabic, a lexicon for Arabic verbs. Arabic Verbnet provides key element information about the syntax and semantics of Arabic verbs using the notion of verb-classes similar to the Verbnet for English. Arabic Verbnet contains verb entries where each entry is a third person masculine singular perfect verb. Each verb entry contains four child nodes of the verb, its root, verbal noun(s), and participle(s). It uses 23 thematic roles which have been already used in the English Verbnet. It has 173 classes which contain 4,392 verbs and 498 frames. These frames provide the four verb entry child nodes information besides information about subcategorization frames and syntactic and semantic description of each verb. The Arabic Verbnet uses XML fromat to store its frames (Mousser 2010).

In summary, the surveyed Arabic lexicons are common morphological and linguistic lists that are specific to a certain Arabic NLP application. They are not general purpose and they are small in size. Moreover, all of them only deal with modern standard Arabic (MSA). Arabic WordNet and Verbnet are based on models for English and Indo-European languages, rather than on Semitic templatic root-based lexical principles.

4.2 Traditional Arabic Lexicons and Lexicography

Traditional Arabic lexicons are not available in computerized lexicographic databases. Moreover, traditional Arabic lexicons have different arrangement methodologies than modern English dictionaries. Common English dictionaries list lexical entries, which are words (*i.e.* lexical entries in form of lemmas), arranged alphabetically; followed by the meaning of that word, while Arabic lexicons are mainly arranged by selecting the root as main lexical entry. The roots are followed by a definition part which may span several pages. The definition part is written as a unit or an article (*i.e.* encyclopaedia entry) which defines all the derived words of a certain root. These lexical entries are not arranged or distinguished with special formatting.

A study of a traditional Arabic lexicon called *al-qāmūs al-muḥīṭ* القاموس المحيط "The comprehensive lexicon" showed three major drawbacks of traditional Arabic lexicons. First, they do not represent language development periods in different times. Second, there are ambiguities in defining and explaining lexical meaning of the derived words. Third, the ordering methodology of the derived words is unorganized and lacks the reference of the origin of the derivations. Khalil (1998) highlighted the importance of ordering the derivations of each lexical entry to directly access the meaning of the derivations, and to show the origin of the Arabic word and its specifications.

Arabic lexicography is one of the original and deep-rooted arts of Arabic literature. The first lexicon constructed was *kitāb al-ʻayn* كتاب العين ‘al-‘ayn lexicon’ by *al-farāhīdī* (died in 791). Over the past 1300 years, many different kinds of Arabic language lexicons were constructed; these lexicons are different in ordering, size and goal of construction. Many Arabic language linguists and lexicographers studied the construction, development and the different methodologies used to construct these lexicons.

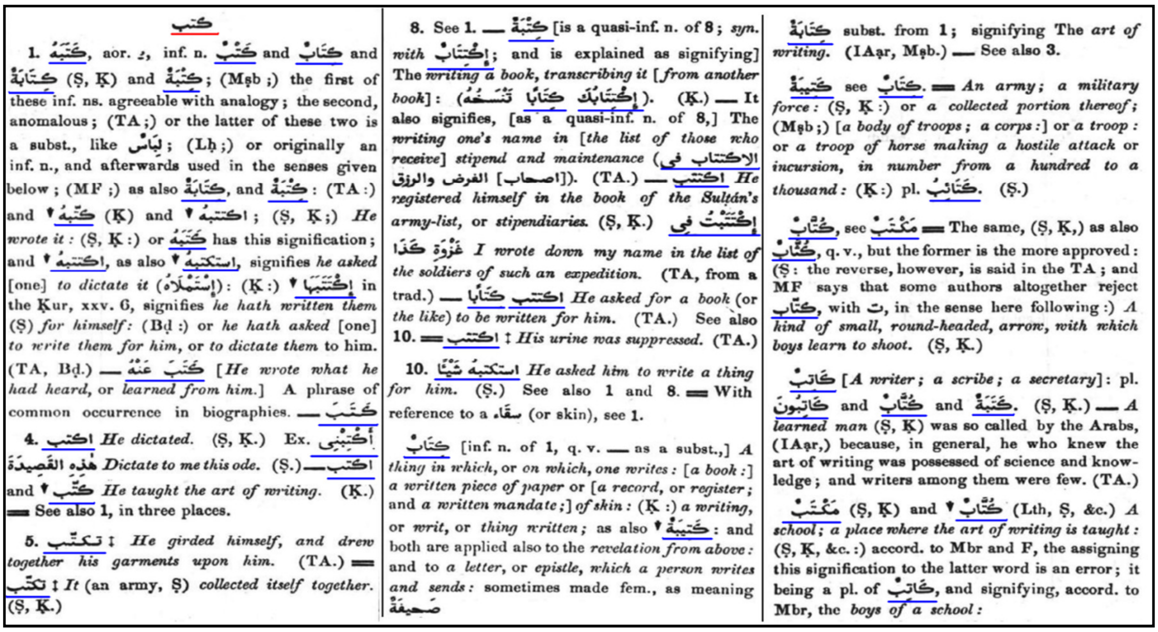
Several traditional Arabic lexicons have been scanned and put online as portable document format (*.pdf*) files. A few have been key-boarded and put online as MS-Word (*.doc*) or HTML text files. Figures 4.1 and 4.4 show samples of text taken from traditional Arabic lexicons; the target lexical entries are underlined and highlighted in blue. Figure 4.2 shows the human translation of the sample of figure 4.1, the target lexical entries are highlighted by square brackets. Figure 4.3 is a sample of the Arabic-English lexicon by Edward Lane (Lane 1968) volume 7, pages 117-119; the target lexical entries are underlined. Figure 4.5 shows a sample of the original manuscript of the traditional Arabic lexicon *aṣ-ṣiḥāḥ fī al-luğah* الصحاح في اللغة ‘The Correct Language’.

كتب: الكِتابُ: معروف، والجمع كُتُبٌ وكُتْبٌ. كَتَبَ الشيءَ يَكْتُبه كَتْباً وكِتاباً وكِتابةً، وكَتَّبَه: خَطَّه؛ قال أَبو النجم: أَقْبَلْتُ من عِنْدِ
زيادٍ كالخَرِفْ، تَخُطُّ رِجْلايَ بخَطٍّ مُخْتَلِفْ، تُكَتِّبانِ في الطَّريقِ لامَ أَلِفْ قال: ورأَيت في بعض النسخ تِكِتِّبانِ، بكسر التاء، وهي لغة بَهْراءَ،
يَكْسِرون التاء، فيقولون: تِعْلَمُونَ، ثم أَتْبَعَ الكافَ كسرةَ التاء.والكِتابُ أَيضاً: الاسمُ، عن اللحياني. الأَزهري: الكِتابُ اسم لما كُتب مَجْمُوعاً؛
والكِتابُ مصدر؛ والكِتابةُ لِمَنْ تكونُ له صِناعةً، مثل الصِّياغةِ والخِياطةِ. والكِتْبةُ: اكْتِتابُك كِتاباً تنسخه. ويقال: اكْتَتَبَ فلانٌ فلاناً أَي
سأَله أَن يَكْتُبَ له كِتاباً في حاجة. واستَكْتَبَه الشيءَ أَي سأَله أَن يَكْتُبَه له. ابن سيده: اكْتَتَبَه ككَتَبَه. وقيل: كَتَبَه خَطَّه؛ واكْتَتَبَه: اسْتَمْلاه،
وكذلك اسْتَكْتَبَه. واكْتتبَه: كَتَبه، واكْتَتَبْته: كَتَبْتُه. وفي التنزيل العزيز: اكْتَتَبَها فهي تُمْلى عليه بُكْرَةً وأَصِيلاً؛ أَي اسْتَكْتَبَها. ويقال: اكْتَتَبَ
الرجلُ إِذا كَتَبَ نفسَه في دِيوانِ السُّلْطان. وفي الحديث: قال له رجلٌ إِنَّ امرأَتي خَرَجَتْ حاجَّةً، وإِني اكْتُتِبْتُ في غزوة كذا وكذا؛ أَي كَتَبْتُ
اسْمِي في جملة الغُزاة. وتقول: أَكْتِبْنِي هذه القصيدةَ أَي أَمْلِها عليَّ. والكِتابُ: ما كُتِبَ فيه. وفي الحديث: مَن نَظَرَ في كِتابِ أَخيه بغير إِذنه،
فكأَنما يَنْظُرُ في النار؛ قال ابن الأَثير: هذا تمثيل، أَي كما يَحْذر النارَ، فَلْيَحْذَرْ هذا الصنيعَ، قال: وقيل معناه كأَنما يَنْظُرُ إِلى ما يوجِبُ عليه
النار؛ قال: ويحتمل أَنه أَرادَ عُقوبةَ البَصرِ لأَن الجناية منه، كما يُعاقَبُ السمعُ إِذا اسْتَمع إِلى قوم، وهم له كارِهُونَ؛ قال: وهذا الحديث محمولٌ
على الكِتابِ الذي فيه سِرٌّ وأَمانة، يَكْرَه صاحبُه أَن يُطَّلَع عليه؛ وقيل: هو عامٌّ في كل كتاب.

Figure 4.1 A sample of text from the traditional Arabic lexicons corpus “*lisān al-ʻarab*”, the target lexical entries are underlined and highlighted in blue.

k t b: [*Alkitab*] the book; is well known. The plural forms are [*kutubun*] and [*kutbun*]. [*kataba Alshay'*] He wrote something. [*yaktubuhu*] the action of writing something. [*katban*], [*kitaban*] and [*kitabatan*] means the art of writing. And [*kattabahu*] writing it means draw it up. Abu Al-Najim said: I returned back from Ziyad's house [after meeting him] and behaved demented, my legs drawn up differently (means walking in a different way). They wrote [*tukattibani*] on the road the letters of *Lam Alif* (describing how he was walking crazily and in a different way). He said: I saw in a different version, the word "they wrote" [*tikittibani*] using the short vowel *kasrah* on the first letter [taa], as it is used by Bahraa' [Arab tribe] dialect. They say: [ti'lamuwn] (you know). Then the short vowel kasrah is propagated to the following letter (kaf). Moreover, [*Alkitab*] the book is a noun. Al-lihyani Al-Azhari definition is: [*Alkitab*] The book is the name of a collection of what has been written (a collection of written materials or texts). And the book has gerund [*Alkitabatu*] writing (art of writing) for whoever has a profession, similar to drafting and sewing. And [*Alkitabatu*]: is copying a book [copying a book in several copies]. It is said: [*iktataba*] someone subscribed another means; he asked to write him a letter in something. [*istaktabahu*] He dictated someone something means to write him something. Ibn Sayyedah: [*Iktatabahu*] is similar to [*katabahu*]. It is said: [*katabahu*] write something down means draw up. And [*Iktatabahu*] writing something down means dictate someone something, which is the same meaning of [*Istaktabahu*]. [*Iktatabahu*] registering (masculine), and [*Iktatabathu*] registing (feminine). In the Qur'an: [*Iktatabaha*] He registered it, he has dictated it every sunrise and sunset, which means dictating it. It is said: [*Iktataba Al-rajul*] The man registered, if he registered himself in the Sultan's office. In Hadith: a man said to him (the prophet): my wife is pilgrimaging (to Mecca), and I have registered [*Oktutibtu*] in a conquest, which means that I have written my name among the conquerors. And you say: [*Aktibny*] let me copy this poem, means dictate me the poem. Also, [*Alkitab*] the book is something which has been written on. And in Hadith: who looks at his brother's book without permission is as looking to hell. Ibn Al-Atheer said: it is a similarity; which means as he avoids hell, he should avoid doing this. He said: the meaning (of the Hadith) is the punishment by hell will be applied if someone looks at a book without permission. He said: it might be the punishment of visual explorers as the crime is done by sight. Hearing explorer is punished if someone intentionally listened to other people who do not like anyone to listen to them. He said: this Hadith is specific for books of secrets and secure books, whose owners hate anybody to look at these books. It is also said: the Hadith is general; applied to any type of books.

Figure 4.2 A Human translation of the sample of text from the traditional Arabic lexicons "*lisān al-'arab*", the target lexical entries are highlighted in blue and square brackets.



كتب

1. كَتَبَهُ, aor. ـُ, inf. n. كَتْبٌ and كِتَابٌ and كِتَابَةٌ (Ṣ, Ḳ) and كِتْبَةٌ; (Mṣb ;) the first of these inf. ns. agreeable with analogy; the second, anomalous; (TA ;) or the latter of these two is a subst., like لِبَاسٌ; (Lḥ ;) or originally an inf. n., and afterwards used in the senses given below; (MF ;) as also كِتَابَةٌ, and كِتْبَةٌ: (TA :) and ↓ كَتَّبَهُ (Ḳ) and ↓ اكتتبهُ; (Ṣ, Ḳ ;) *He wrote it*: (Ṣ, Ḳ :) or كَتَّبَهُ has this signification; and ↓ اكتتبهُ, as also ↓ استكتبهُ, signifies *he asked* [one] *to dictate it* (إِسْتَمْلَاهُ): (Ḳ :) ↓ إِكْتَتَبَهَا in the Ḳur, xxv. 6, signifies *he hath written them* (Ṣ) *for himself*: (Bḍ :) or *he hath asked* [one] *to write them for him*, or *to dictate them* to him. (TA, Bḍ.) — كَتَبَ عَنْهُ [*He wrote what he had heard*, or *learned from him*.] A phrase of common occurrence in biographies. — كَتَّبَ

4. اكتب *He dictated.* (Ṣ, Ḳ.) Ex. أَكْتِبْنِى هٰذِهِ القَصِيدَةَ *Dictate to me this ode.* (Ṣ.) — اكتب and ↓ كَتَّبَ *He taught the art of writing.* (Ḳ.) ═ See also 1, in three places.

5. تكتّب ‡ *He girded himself, and drew together his garments upon him.* (TA.) ═ تكتّب ‡ *It* (an army, Ṣ) *collected itself together.* (Ṣ, Ḳ.)

8. See 1. — كِتْبَةٌ [is a quasi-inf. n. of 8; *syn. with* إِكْتِتَابٌ; and is explained as signifying] The *writing a book, transcribing it* [*from another book*]: (إِكْتِتَابُكَ كِتَابًا تَنْسَخُهُ). (Ḳ.) — It also signifies, [as a quasi-inf. n. of 8,] The *writing one's name in* [*the list of those who receive*] *stipend and maintenance* (الاكتتاب فى [اصحاب] الفرض والرزق). (TA.) — اكتتب *He registered himself in the book of the Sulṭán's army-list*, or *stipendiaries.* (Ṣ, Ḳ.) إِكْتَتَبْتُ فِى غَزْوَةِ كَذَا *I wrote down my name in the list of the soldiers of such an expedition.* (TA, from a trad.) — اكتتب كِتَابًا *He asked for a book* (or *the like*) *to be written for him.* (TA.) See also 10. ═ اكتتب ‡ *His urine was suppressed.* (TA.)

10. استكتبهُ شَيْئًا *He asked him to write a thing for him.* (Ṣ.) See also 1 and 8. ═ With reference to a سِقَاء (or skin), see 1.

كِتَابٌ [inf. n. of 1, q. v. — as a subst.,] *A thing in which*, or *on which, one writes*: [*a book* :] *a written piece of paper* or [*a record*, or *register*; and *a written mandate*;] *of skin*: (Ḳ :) *a writing*, or *writ*, or *thing written*; as also ↓ كَتِيبَةٌ: and both are applied also to the *revelation from above*: and to *a letter*, or *epistle, which a person writes and sends*: sometimes made fem., as meaning صَحِيفَةٌ

كِتَابَةٌ subst. from 1; signifying The *art of writing.* (IAạr, Mṣb.) — See also 3.

كَتِيبَةٌ see كِتَابٌ. ═ *An army; a military force*: (Ṣ, Ḳ :) or *a collected portion thereof*; (Mṣb ;) [*a body of troops*; *a corps* :] or *a troop*: or *a troop of horse making a hostile attack* or *incursion, in number from a hundred to a thousand*: (Ḳ :) pl. كَتَائِبُ. (Ṣ.)

كُتَّابٌ, see مَكْتَبٌ ═ The same, (Ṣ, Ḳ,) as also كُتَّابٌ, q. v., but the former is the more approved: (Ṣ: the reverse, however, is said in the TA; and MF says that some authors altogether reject كُتَّاب, with ت, in the sense here following :) *A kind of small, round-headed, arrow, with which boys learn to shoot.* (Ṣ, Ḳ.)

كَاتِبٌ [*A writer*; *a scribe*; *a secretary*]: pl. كَاتِبُونَ and كُتَّابٌ and كَتَبَةٌ. (Ṣ, Ḳ.) — *A learned man* (Ṣ, Ḳ) was so called by the Arabs, (IAạr,) because, in general, he who knew the art of writing was possessed of science and knowledge; and writers among them were few. (TA.)

مَكْتَبٌ (Ṣ, Ḳ) and ↓ كُتَّابٌ (Lth, Ṣ, &c.) *A school*; *a place where the art of writing is taught*: (Ṣ, Ḳ, &c. :) accord. to Mbr and F, the assigning this signification to the latter word is an error; it being a pl. of كَاتِبٌ, and signifying, accord. to Mbr, the *boys of a school*:

Figure 4.3 A Sample of the definition of the root *ktb* from an Arabic-English Lexicon by Edward Lane (Lane 1968), http://www.tyndalearchive.com/TABS/Lane/ , the target lexical entries are underlined.

(ك ت ب):

(كَتَبَهُ) كِتْبَةً وَكِتَابًا وَكِتَابَةً وَقَوْلُهُ وَإِذَا كَانَتْ السَّرِقَةُ صُحُفًا لَيْسَ فِيهَا كِتَابٌ أَيْ مَكْتُوبٌ (وَفِي حَدِيثِ أُنَيْسٍ) وَاحْكُمْ بِكِتَابِ اللَّهِ أَيْ بِمَا فَرَضَ اللَّهُ مِنْ كَتَبَ عَلَيْهِ كَذَا إذَا أَوْجَبَهُ وَفَرَضَهُ (وَمِنْهُ) الصَّلَاةُ الْمَكْتُوبَةُ وَأَمَّا قَوْلُهُ – صَلَّى اللَّهُ عَلَيْهِ وَآلِهِ وَسَلَّمَ - [مَا بَالُ أَقْوَامٍ يَشْتَرِطُونَ شُرُوطًا لَيْسَتْ فِي كِتَابِ اللَّهِ تَعَالَى] فَقِيلَ الْمُرَادُ قَوْله تَعَالَى (اُدْعُوهُمْ لِآبَائِهِمْ) إلَى أَنْ قَالَ وَمَوَالِيكُمْ فِيهِ أَنَّهُ نَسَبَهُمْ إلَى مَوَالِيهِمْ كَمَا نَسَبَهُمْ إلَى آبَائِهِمْ فَلَمَّا لَمْ يَجُزْ التَّحَوُّلُ عَنْ الْآبَاءِ لَمْ يَجُزْ عَنْ الْأَوْلِيَاءِ وَيَجُوزُ أَنْ يُرَادَ بِكِتَابِ اللَّهِ قَضَاؤُهُ وَحُكْمُهُ عَلَى لِسَانِ رَسُولِ اللَّهِ – صَلَّى اللَّهُ عَلَيْهِ وَآلِهِ وَسَلَّمَ – إنَّ الْوَلَاءَ لِمَنْ أَعْتَقَ (وَأَكْتَبَ الْغُلَامَ وَكَتَّبَهُ) عَلَّمَهُ الْكِتَابَ (وَمِنْهُ) سَلَّمَ غُلَامَهُ إلَى مُكَتِّبٍ أَيْ إلَى مُعَلِّمِ الْخَطِّ رُوِيَ بِالتَّخْفِيفِ وَالتَّشْدِيدِ (وَأَمَّا الْمَكْتَبُ) وَالْكُتَّابُ فَمَكَانُ التَّعْلِيمِ وَقِيلَ الْكُتَّابُ الصِّبْيَانُ (وَكَاتَبَ) عَبْدَهُ مُكَاتَبَةً وَكِتَابًا قَالَ لَهُ حَرَّرْتُكَ يَدًا فِي الْحَالِ وَرَقَبَةً عِنْدَ أَدَاءِ الْمَالِ (وَمِنْهُ) قَوْله تَعَالَى (وَالَّذِينَ يَبْتَغُونَ الْكِتَابَ) وَقَدْ يُسَمَّى بَدَلُ الْكِتَابَةِ مُكَاتَبَةً وَأَمَّا الْكِتَابَةُ فِي مَعْنَاهَا فَلَمْ أَجِدْهَا إلَّا فِي الْأَسَاسِ وَكَذَا تَكَاتَبَ الْعَبْدُ إذَا صَارَ مُكَاتَبًا وَمَدَارُ التَّرْكِيبِ عَلَى الْجَمْعِ (وَمِنْهُ كَتَبَ النَّعْلَ وَالْقِرْبَةَ) خَرَزَهَا (وَالْكُتَبُ الْخُرَزُ) الْوَاحِدَةُ كُتْبَةٌ (وَمِنْهُ كَتَبَ الْبَغْلَةَ) إذَا جَمَعَ بَيْنَ شُفْرَتَيْهَا بِحَلْقَةٍ (وَالْكَتِيبَةُ) الطَّائِفَةُ مِنْ الْجَيْشِ مُجْتَمِعَةً (وَبِهَا سُمِّيَ) أَحَدُ حُصُونِ خَيْبَرَ (وَقَوْلُهُمْ) سُمِّيَ هَذَا الْعَقْدُ مُكَاتَبَةً لِأَنَّهُ ضَمُّ حُرِّيَّةِ الْيَدِ إلَى حُرِّيَّةِ الرَّقَبَةِ أَوْ لِأَنَّهُ جَمْعٌ بَيْنَ نَجْمَيْنِ فَصَاعِدًا ضَعِيفٌ جِدًّا وَإِنَّمَا الصَّوَابُ أَنَّ كُلًّا مِنْهُمَا كَتَبَ عَلَى نَفْسِهِ أَمْرًا هَذَا الْوَفَاءَ وَهَذَا الْأَدَاءَ.

Figure 4.4 A sample of text from the traditional Arabic lexicon "*al-muğrib fī tartīb al-mu'rib*", the target lexical entries are underlined and highlighted in blue.

Figure 4.5 A sample of a traditional Arabic lexicon *aṣ-ṣiḥāḥ fī al-luğa^h* الصحاح في اللغة 'The Correct Language', the original manuscript.

4.3 Methodologies for Ordering Lexical Entries in the Traditional Arabic Lexicons

Traditional Arabic lexicons distinguish between four classes of ordering lexical entries in the lexicon. First, the *al-ẖalīl* methodology was developed by الخليل بن أحمد الفراهيدي *al-ẖalīl bin aḥmad al-farāhīdī* (died in 791). Second, the *abū ʻubayd* methodology was developed by *abū ʻubayd al-qāsim bin sallām* أبو عُبيدٍ القاسم بن سلاَّم(died in 838). Third, the *al-ğawharī* methodology was developed by *'ismā'īl bin ḥammād al-ğawharī* (died in 1002). Finally, the *al-barmakī* methodology was developed by *abū al-maʻālī moḥammad bin tamīm al-barmakī* أبو المعالي محمد بن تميم البرمكي, who lived in the same time period as *al-ğawharī*. *al-barmakī* did not construct a new lexicon; but he alphabetically re-arranged a lexicon called *aṣ-ṣiḥāḥ fī al-luğah* الصحاح في اللغة'The Correct Language' by *al-ğawharī*. He added little information to that lexicon.

4.3.1 The *al-ẖalīl* Methodology

The *al-ẖalīl* methodology was developed by الخليل بن أحمد الفراهيدي *al-ẖalīl bin aḥmad al-farāhīdī* (died in 791). His lexicon called كتاب العين *kitāb al-ʻayn* "*al-ʻayn* lexicon" was the first traditional Arabic lexicon. 'The *al-ʻayn'* lexicon lists the lexical entries phonologically according to places of articulation of phonemes from the mouth and throat, working forwards from glottal through to labial regions. He divided the lexicon into books, with one book for one letter. The books were then divided into 4 sections according to their internal structure: doubled biliteral roots; intact triliteral roots; doubly-defective roots; quadriliteral and quinquetiliteral roots. Many lexicons followed *al-ẖalīl's* methodology with slight changes in ordering. The following traditional Arabic lexicons followed this ordering methodology:

1. كتاب العين *kitābu al-ʻayn* "*al-ʻayn* Lexicon" by الخليل ابن أحمد الفراهيدي *al-ẖalīl bin aḥmad al-farāhīdī* died in 175H / 791AD.
2. مُعْجَمُ الْمُحِيطِ فِي اللغَةِ *mu'ğam al-muḥīṭ fī al- luğa*^{h} "The Comprehensive Language" by الصاحب بن عباد *aṣ-ṣāḥib bin ʻabbād* died in 385H / 995AD.
3. المحكم والمحيط الأعظم *al-muḥkam wa al-muḥīṭ al-'aʻaẓam* "The Greatest Verified and Comprehensive Lexicon" by (ابن سيدة) أبو الحسن علي بن اسماعيل النَّحوي اللغوي الأندلسي *'ibn sayyidah, abū al-ḥasan bin ''ismā ʻīl an-naḥawī al-lağawī al-'andalusī* died in 458H / 1065AD.
4. لسان العرب*lisān al-ʻrab* "Arab tongue" by جمال الدين محمد بن منظور *ğamāl ad-dīn moḥammed bin manẓūr* died in 629H / 1311AD.
5. معجم تهذيب اللغة *mu'ğam tahḏīb al-luğa*^{h} "The Lexicon of Refined Language" by أبو منصور الأزهري *abū manṣūr al-'azharī* died in 1205H / 1790AD.

4.3.2 The *abū ‘ubayd* Methodology

The *abū ‘ubayd* methodology was developed by *abū ‘ubayd al-qāsim bin sallām* أبو عُبيدٍ القاسم بن سلاَّم (died in 838). The first constructed lexicon which followed this methodology was الغريب المُصنّف في اللغة *al-ḡarīb al-muṣannaf fī al-luḡa*^{h} “The Irregular Classified Language”. This methodology arranges lexical entries according to their concepts or topics. The lexicon consists of many small books, each of which describes a topic or a concept, such as books describing horses, milk, honey, flies, insects, palms, and human creation. Then these small books are collated into one large lexicon. That lexicon consists of more than thirty small books. The following traditional Arabic lexicons followed *abī ‘ubayd* methodology:

6. الغريب المُصنّف في اللغة *al-ḡarīb al-muṣannaf fī al-luḡa*^{h} “The Irregular Classified Language” by أبي عُبيدٍ القاسم بن سلاَّم *’abi ‘ubayd al-qāsim bin sallām* died in 223H / 838AD.
7. المُنَجَّد في اللغة *al-munağğad fī al-luḡa*^{h} “The Decorated Language” by علي بن حسن الهنائي الأزدي*ali bin ḥasan al-hunā’ī al-’azdī* died in 310H / 922AD.
8. المخصص في اللغة *al-muẖaṣṣaṣ fī al-luḡa*^{h} “The Specified Language” by (ابن سيدة) أبو الحسن علي بن اسماعيل النَّحوي اللغوي الأندلسي *’ibn sayyidah, abū al-ḥasan bin ’ismā‘īl an-naḥawī al-laḡawī al-’andalusī* died in 458H / 1065AD.

4.3.3 The *al-ğawharī* Methodology

The *al-ğawharī* methodology was developed by *’ismā’īl bin ḥammād al-ğawharī* (died in 1002). The first lexicon which followed this methodology is called الصحاح في اللغة *aṣ-ṣiḥāḥ fī al-luḡa*^{h} ‘The Correct Language’. This methodology was based on the alphabetical order for ordering the lexical entries. However, the lexical entries were arranged in this lexicon depending on the last letter of the word, and then the first letter. The lexicon was organized into chapters where each chapter corresponds to the last letter of the word. Each chapter includes sections corresponding to the first letter of the word, then the second letter of triliteral roots, then the third letter of quadriliteral roots, then the fourth letter in quinquitiliteral roots. For example, the word بَسَطَ*baṣaṭ* “spread” is found in chapter ط*ṭ* representing the last letter of the word, then by looking to section ب*b* as it represents the first letter. The following lexicons followed this ordering methodology:

9. الصحاح في اللغة *aṣ-ṣiḥāḥ fī al-luḡa*^{h} “The correct language” by أبو نصر إسماعيل بن حماد الجوهري الفرابي*abū naṣr ’ismā‘īl bin ḥammād al-ğawharī al-farābī* died in 400H / 1009AD.
10. العباب الزاخر في اللغة *al-‘ibāb az-zāẖir fī al-luḡa*^{h} “The High Flood Water of Language” by الحسن بن محمد الصغاني *al-ḥasan bin muḥammad aṣ-ṣaḡānī* died in 650H / 1252AD.

11. تاج العروس من جواهر القاموس *tağ al-ʻarūs min ğawāhir al-qāmūs* “Bridal Crown Jewel of Dictionaries” by الزبيدي *az-zubaydī* died in 1205H / 1790AD.

12. القاموس المحيط *al-qāmūs al-muḥīṭ* “The Comprehensive Dictionary” by مجد الدين أبو طاهر محمد بن يعقوب الفيروزآبادى *mağd ad-dīn abū ṭāhir muḥammad bin yaʻqūb al-fayrūz’ābādī* died in 817H / 1414AD.

4.3.4 The *al-barmakī* Methodology

The *al-barmakī* methodology was developed by *abū al-maʻālī muḥammad bin tamīm al-barmakī* أبو المعالي محمد بن تميم البرمكي, who lived in the same time period as *al-ğawharī*. The *al-barmakī* methodology is based on arranging lexical entries alphabetically starting from the first root letter. *al-barmakī* did not construct a new lexicon. Rather, he re-arranged, using this ordering methodology, the lexical entries of الصحاح في اللغة *aṣ-ṣiḥāḥ fī al-luğa*^{h}, which was developed by *al-ğawharī* ordered using *al-ğawharī* methodology. Little information was added to this reordered version of the lexicon. After that, الزمخشري *az-zamaẖšarī* (died in 1143) followed the same methodology and constructing a lexicon called أساس البلاغة *asās al-balāğa*^{h} “Fundamentals of Rhetoric”. This methodology of ordering lexical entries in an Arabic lexicon become the most widely used ordering methodology. The following lexicons followed this ordering methodology:

13. معجم الجيم *muʻğam al-ğīm* “The *jīm* Lexicon” by أبو عمرو الشيباني *abū ʻamr aš-šībānī* died in 206H / 821AD.

14. جمهرة اللغة *ğamharat al-luğa*^{h} “The Gathering of the Language” by ابن دُرَيْد *’ibn durayd* died in 256H / 869AD.

15. معجم مقاييس اللغة *muʻğam maqāyīs al-luğa*^{h} “The Lexicon of the Standard Language” by أبي الحسين أحمد بن فارِس بن زكَرِيّا *’abī al-ḥusayn aḥmad bin fāris bin zakaryyiā* died in 395H / 1004AD.

16. معجم ما استعجم *muʻğam mā ’istaʻğam* “A Lexicon of Foreign Words” by البكري الاندلسي *al-bakrī al-’andalusī* died in 487H / 1094AD.

17. تهذيب الأفعال *tahḏīb al-afʻāl* “The Refined Verbs” by (ابن القطاع) أبو القاسم علي بن جعفر السعدي *(’ibn al-qiṭā’) abū al-qāsimʻalī bin ğaʻfar as-saʻdī* died in 515H/ 1121AD.

18. أساس البلاغة *asās al-balāğa*^{h} “Fundamentals of Rhetoric” by أبو القاسم محمود بن عمرو بن أحمد، الزمخشري جار الله *abū al-qāsim maḥmūd bin ʻamr bin aḥmad, az-zamaẖšarī ğār allā*^{h} died in 538H / 1143 AD.

19. الْمُغرِبِ فِي تَرْتِيبِ الْمُعْرِبِ *al-muğrib fī tartīb al-muʻrib* “Irregular Declinable Words” by أبو الفتح ناصر الدِّين المطرزي *’abū al-fatḥ nāṣir ad-dīn al-muṭrazī* died in 610H / 1213AD.

20. مختار الصحاح *muẖtār aṣ-ṣiḥāḥ* “The Selected of the Correct Language” by أبو بكر الرازي *abū bakr ar-rāzī* died in 666H / 1267AD.

21. المصباح المنير في غريب الشرح الكبير *al-muṣbāḥ al-munīr fī ğarīb aš-šarḥ al-kabīr* "The Illuminating Light on the Irregularity of the Great Explanations" by أحمد بن محمد بن علي الفيومي ثم الحموي، أبو العباس *aḥmad bin muḥammad 'alī al-fayyūmī ṯumma al-ḥamawī, abū al-'abbās* died in 538H / 1143AD.

22. المعجم الوسيط *al-mu'ğam al-wasīṭ* "The Intermediary Lexicon" by ابراهيم مصطفى . أحمد الزيات . حامد عبد القادر . محمد النجار *ibrāhīm muṣṭafā, aḥmad az-zayyāt, ḥāmid 'abdul-qādir, muḥammad an-nağğār* published in 1960.

23. معجم الأفعال المتعدية بحرف *mu'ğam al-'af'āl al-muta'adyya*^{h} *bi ḥarf* "The Lexicon of Transitive Verbs" by موسى بن محمد بن الملياني الأحمدي *mūsā bin muḥammad al-malyānī al-'aḥmadī* published in 1979.

4.4 Constructing the SALMA-ABCLexicon

Many existing morphological lexicons were constructed from raw text (Sagot 2005). The general requirements for constructing a morphological lexicon from raw text are: a corpus; a generation program or a morphological description of the language; a Lexical Markup Framework (LMF) for providing compatible structure to store the lexical entries; searching facility over the lexical entries (querying the constructed lexicon); and an evaluation methodology of the lexicon (Russell et al. 1986; Petasis et al. 2001; Tadi and Fulgosi 2003; Sagot 2005; Sagot et al. 2006; Paikens 2007; Nicolas et al. 2008; Erjavec 2010; Sagot 2010).

Broad-coverage language resources which provide prior linguistic knowledge must improve the accuracy and the performance of NLP applications. The main aim in constructing a broad-coverage lexical resource is to improve the accuracy of morphological analyzers and part-of-speech taggers of Arabic text. Chapter 3 discussed the shortcomings of the existing stemming algorithms for Arabic text. Constructing a broad-coverage lexical resource to improve the accuracy of Arabic morphological analysis has advantages over developing a sophisticated stemming algorithm. These advantages are:

- A prior-knowledge lexical resource will improve the Arabic morphological analysis.
- A lexical resource can be integrated to different stemming algorithms to give prior knowledge about the analyzed words.
- It can help in enhancing the performance of the morphological analyzers by reducing the complex analysis steps to a simpler look up procedure.

- The broad-coverage lexical resource can be a standalone resource which can be integrated in different Arabic natural language processing systems and benefits of integration can be gained.
- It is easier to update the lexical resource by adding new contents to it and correcting it than updating a sophisticated algorithm which needs specialized developers.
- It can also be used as a teaching material resource to help in assisting both teachers and students in a teaching-learning process.

The SALMA-ABCLexicon (Sawalha Atwell Leeds Morphological Analyses – Arabic Broad-Coverage Lexicon) was developed following the general requirements for constructing morphological lexicons from raw text. However, the absence of open-source Arabic corpora and the absence of a generation program led to the use of traditional Arabic lexicons as a corpus. The generation program for Arabic can generate verbs and derived nouns, but its major shortcomings are both over-generation and under-generation. The over-generation problem results in many lexical entries which are correctly structured but are not part of the real language vocabulary, while the under-generation problem happens when the generation cannot generate all possible vocabulary of the language.

In theory, any morphological generation program for Arabic will suffer from both over-generation and under-generation problems unless it has been provided with a comprehensive database that contains all the non-generated vocabulary (*i.e.* non-inflected words, primitive nouns and non-conjugated verbs) and comprehensive morphological descriptions of language encoded within the generation program. Both the dataset and the morphological descriptions of the language need huge amounts of manual work. As an alternative, the selection of traditional Arabic lexicons as a text corpus for constructing the SALMA-ABCLexicon will provide; first, a wide coverage of Arabic vocabulary (derived and non-derived words) where most of them appear in the lexicons in different forms as they are defined in the lexical entry. Second, the lexicons cover a range of the past 13 centuries (*i.e.* from 800 to 2000), a wide range of both classical and modern Arabic vocabulary and their development. Third, they provide a basic and comprehensive morphological dataset by mapping between the words and their roots; especially for words of hard cases where stemming algorithms and morphological analyzers fail to analyze them. This morphological dataset can be re-used by different text analytics applications.

This section discusses the construction steps for the SALMA-ABCLexicon following the three general requirements, mentioned above, for constructing morphological lexicons from raw text. Section 4.4.1 describes the text corpus used to construct the lexicon. Section 4.4.2 discusses the morphological knowledge used to extract the lexical entries and their basic morphological information. Section 4.4.3 describes the process of combining the lexical entries into one large lexical resource. Section 4.4.4 discusses the format of the lexicon. Section 4.4.5 explains the querying of the lexicon and the retrieval of its information.

4.4.1 The Text Corpus

As mentioned above, due to the absence of an open-source representative Arabic corpus and the absence of a generation program, the selection of a corpus to build the morphological lexicon was directed to select, as a corpus, the traditional Arabic lexicons. Twenty three freely available lexicons were collected from different resources from the web. These lexicons are listed in section 4.3. Meshkat Islamic Network[39] شبكة مشكاة الاسلامية *šabaka*^{t} *miškā*^{t} *al-'islāmiyya*^{h} provides most of these lexicons which are written in machine readable format using MS Word files or HTML web pages.

Common processing steps were applied to all lexicons. First, all lexicon files were converted from MS Word or HTML web pages into standard text files in Unicode 'utf-8' encoding. Second, a statistical analysis computed the word frequency and the vocabulary size for both vowelized and non-vowelized text of each lexicon. The complete corpus of 23 lexicon texts contains 14,369,570 words, 2,184,315 vowelized word types and 569,412 non-vowelized word types. Table 4.1 shows the summary of the statistical analyses of the lexicon texts used to construct the SALMA-ABCLexicon. Section 4.6 discusses the corpus of traditional Arabic lexicons.

Table 4.1 statistical analysis of the lexicon text used to construct the broad-coverage lexical resource

Number of files		**247**
Size		**178.32 MB**
Vowelized word analysis	Number of words	14,369,570
	Number of word types	2,184,315
Non-vowelized word analysis	Number of words	14,369,570
	Number of word types	569,412

4.4.2 Morphological Knowledge Used to Extract the Lexical Entries

Each lexicon was constructed following one of four ordering methodologies of their lexical entries, although most of them used the root as main lexical entry. Moreover, the 23 lexicons were typed into machine-readable files in different formats but without using any computerized lexicographic representations. These factors add more processing challenges. Therefore, each lexicon was processed separately using specialized programs. An important preprocessing step converts each lexicon text into a unified format by choosing the most common format for all the root entries in the lexicon. This step was

[39] شبكة مشكاة الإسلاميةMeshkat Islamic Network http://www.almeshkat.net

done manually, which involves going through all the text in the lexicon files and re-formatting the root entries that do not follow the selected format. The common basic structure of all lexicons is root-definition structure, where each root entry in the lexicon is followed by the definition part that groups all the derived words and their meanings. After that, a program was written to extract the roots and words derived from that root. The tokenizing module in the program must specify the root entries and their definition parts. Then, a bag of words was extracted from the definition text. The bag of words stores word-root pairs, where each word appearing in the definition part is associated with the root of that part.

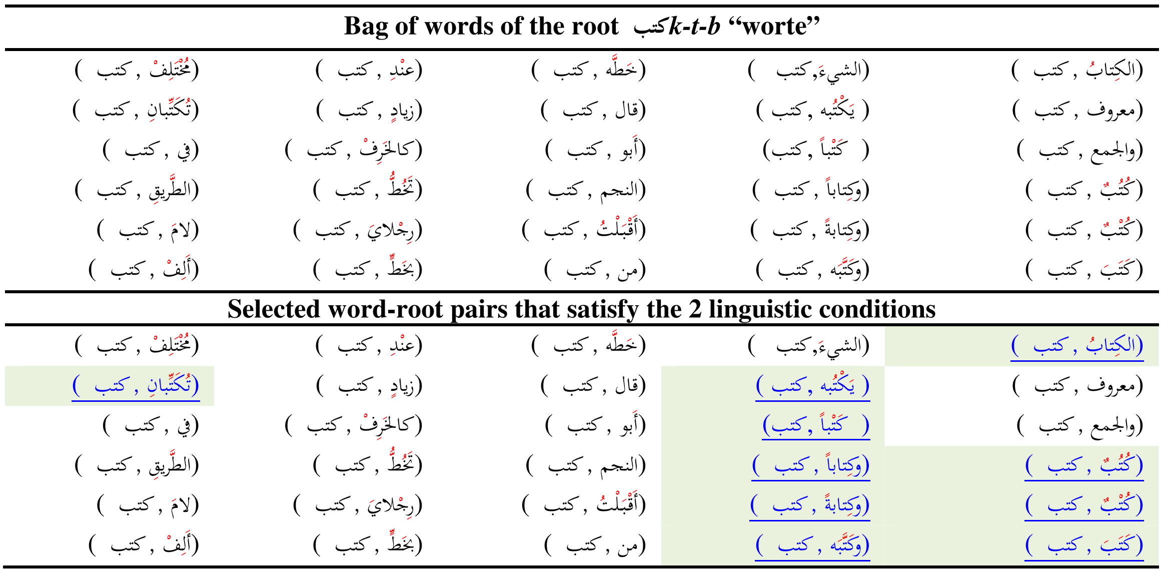
The definition parts of the roots are written as encyclopaedia articles that define each root and define the lexical entries derived from a certain root. The writing style of the definition part connects the lexical entries and their meanings together without following any structure or ordering methodology. The writing style of the definition parts show the lexical entries conjoined with all kinds of clitics and affixes. Clitics, such as conjunctions and pronouns, are used to connect the definitions of the lexical entries together as one unit.

Although the use of clitics and affixes adds a greater challenge to the construction of the broad-coverage lexical resource, they substitute and compensate for the generation program where derived words from a given root (*i.e.* lexical entry) appear in different shapes and formats. Moreover, the use of different lexicons, which share most of their lexical entries but differ in defining them, increases the potential for gathering a wider range of forms and shapes of the same derived words. Finally, because the definition part of the lexical entry is written as natural language text, the different forms of a derived word counted as a valid part of the language vocabulary, but excluded over-generated words; see figure 4.7. Non-derived words related to certain root lexical entries are also gathered and included in the lexicon.

Many words appearing in the definition part are not relevant to the root associated with that definition. Such words are found in the bag of words of that root. A normalization analysis that verifies the word-root pairs works by applying linguistic knowledge that governs the derivation process of words from their roots. These conditions are simply described as the following:

- **Condition 1 (check consonants):** If all consonant letters forming the root appear in the analyzed word, then check condition 2.
- **Condition 2 (consonants order):** If all root letters appear in the same order as the word's letters, then word-root combination is a candidate analysis, and can be inserted to the lexicon.

In the first condition (check consonants), we classified Arabic letters into four groups, letters that appear in clitics or affixes, vowels, *hamza*^{h} and letters that might be changed in derivation due to substitution إقلاب *'iqlāb* to simplify the pronunciation of the word. Then, a procedure is applied to verify each letter of the word. Another procedure is applied to match the order of the letters of both the analyzed word and its root. The analyses that meet the two conditions are candidate analyses and are stored in the lexicon database. The information about clitics, affixes and stem is also stored with the word-root combination. Figure 4.6 shows the process of selecting word-root pairs. Table 4.2 shows the number of words and the percentage of words extracted from the original text of the lexicons.



Bag of words of the root كتب*k-t-b* "worte"

(مُختَلِفْ , كتب)	(عِنْدِ , كتب)	(خَطَّه , كتب)	(الشيءَ,كتب)	(الكِتابُ , كتب)
(تُكتَبّانِ , كتب)	(زِيادٍ , كتب)	(قال , كتب)	(يَكْتُبه , كتب)	(معروف , كتب)
(في , كتب)	(كالخَرْفِ , كتب)	(أَبو , كتب)	(كَتْباً , كتب)	(والجمع , كتب)
(الطَّرِيقِ , كتب)	(تَخُطُّ , كتب)	(النجم , كتب)	(وكِتاباً , كتب)	(كُتُبٌ , كتب)
(لامَ , كتب)	(رِجْلايَ , كتب)	(أَقْبَلْتُ , كتب)	(وكِتابةً , كتب)	(كُتْبٌ , كتب)
(أَلِفْ , كتب)	(بِخَطٍّ , كتب)	(من , كتب)	(وكَتَّبه , كتب)	(كَتَبَ , كتب)

Selected word-root pairs that satisfy the 2 linguistic conditions

(مُختَلِفْ , كتب)	(عِنْدِ , كتب)	(خَطَّه , كتب)	(الشيءَ,كتب)	(الكِتابُ , كتب)
(تُكتَبّانِ , كتب)	(زِيادٍ , كتب)	(قال , كتب)	(يَكْتُبه , كتب)	(معروف , كتب)
(في , كتب)	(كالخَرْفِ , كتب)	(أَبو , كتب)	(كَتْباً , كتب)	(والجمع , كتب)
(الطَّرِيقِ , كتب)	(تَخُطُّ , كتب)	(النجم , كتب)	(وكِتاباً , كتب)	(كُتُبٌ , كتب)
(لامَ , كتب)	(رِجْلايَ , كتب)	(أَقْبَلْتُ , كتب)	(وكِتابةً , كتب)	(كُتْبٌ , كتب)
(أَلِفْ , كتب)	(بِخَطٍّ , كتب)	(من , كتب)	(وكَتَّبه , كتب)	(كَتَبَ , كتب)

Figure 4.6 Using linguistic knowledge to select word-root pairs from traditional Arabic lexicons. The selected word-root pairs are underlined and highlighted in blue

Table 4.2 Statistics of the traditional Arabic lexicons and morphological databases used to construct the SALMA-ABCLexicon

	Lexicon name	Word types	Words extracted		Roots extracted
1	*tağ al-'arūs min ğawāhir al-qāmūs*	831,504	474,351	57.05%	11,101
2	*lisān al-'rab*	507,860	274,305	54.01%	9,355
3	*mu'ğam al-muḥīṭ fī al-luğa*^{h}	168,870	66,763	39.54%	6,411
4	*kitābu al-'ayn*	141,098	54,970	38.96%	5,826
5	*al-mu'ğam al-wasīṭ*	112,164	45,614	40.67%	6,489
6	*al-muṣbāḥ al-munīr fī ğarīb aš-šarḥ al-kabīr*	61,422	29,742	48.42%	2,947
7	*muḫtār aṣ-ṣiḥāḥ*	40,295	17,636	43.77%	3,420
8	*al-muğrab fī tartīb al-mu'rab*	39,930	13,798	34.56%	2,322
9	*Arabic WordNet*	-	16,998	-	2,589
10	*Buckwalter's Lexicon*	-	82,158	-	-

4.4.3 Combining the Processed Lexicons into the SALMA-ABCLexicon

After manually converting each lexicon text into a unified format by choosing the most common format for all the root entries in the lexicon, information such as roots, words and meaning is automatically extracted using specialized programmes. The results are stored in separate dictionary files which include roots, words, and meanings. A combination algorithm combines the disparate lexicon information into one large broad-coverage lexical resource.

A combination algorithm is applied to construct the SALMA-ABCLexicon. The algorithm starts by selecting a large lexicon called لسان العرب *lisān al-‘rab* ‘Arab tongue’ as a seed to the SALMA-ABCLexicon. Then, the lexicons are combined one by one. Figure 4.7 shows the first 60 lexical entries of the root كتب *k-t-b* ‘wrote’ stored in the SALMA-ABCLexicon. After combining each lexicon the percentage of records added to the SALMA-ABCLexicon is computed. The percentage starts with 100% for the seed lexicon and decreases during the combination process. The percentage will tell us when the combination process should stop, and which lexicons are better to construct the SALMA-ABCLexicon. Table 4.3 shows the number of records extracted from 4.7 analyzed lexicons, and the number and percentage of records combined to form the SALMA-ABCLexicon.

The SALMA-ABCLexicon contains 2,774,866 word-root pairs, which represent 509,506 different words representing 261,125 different non-vowelized words. It contains 12 different biliteral roots; 8,585 different triliteral roots; 4,038 different quadriliteral roots; 63 different quinqueliteral roots; and 31 different sexiliteral roots. Word types of the lexicon are distributed into; 117 word types of biliteral roots; 483,356 word types of triliteral roots; 30,873 word types of quadriliteral roots; 615 word types of quinqueliteral; and 335 word types of sexiliteral roots.

Table 4.3 Number of records extracted from 7 analyzed lexicons, and the number and the percentage of records combined to the SALMA-ABCLexicon.

#	Lexicon	Word types [B]	Records inserted [A]	Percentage	
				(A/B)%	(A/C)%
1	*lisān al-‘rab*	207,992	207,992	100.00%	47.80%
2	*mu’ǧam al-muḥīṭ fī al- luġa*^{h}	74,507	61,113	82.02%	14.04%
3	*taǧ al-‘arūs min ǧawāhir al-qāmūs*	128,119	95,415	74.47%	21.93%
4	*muẖtār aṣ-ṣiḥāḥ*	19,540	16,573	84.82%	3.81%
5	*al-muġrib fī tartīb al-mu‘rib*	12,396	9,805	79.10%	2.25%
6	*kitāb*^{u} *al-‘ayn*	30,292	18,878	62.32%	4.34%
7	*al-mu’ǧam al-wasīṭ*	36,660	25,364	69.19%	5.83%
	Totals	***509,506***	***435,140*** [C]	**85.40%**	***100.00%***

أكتبه	*'aktabahu*	الكتاب	*al-kitāb*	الكُتْبةُ	*al-kutba*^{tu}
أَكْتَبَ	*'aktaba*	الكتابة	*al-kitāba*^{t}	الكُتْبَةُ	*al-kutba*^{tu}
أَكْتَبْتُ	*'aktabtu*	الكتابةَ	*al-kitāba*^{ta}	الكِتاب	*al-kitāb*
أَكْتِبْنِي	*'aktibnī*	الكتابَة	*al-kitāba*^{t}	الكِتابةُ	*al-kitāba*^{tu}
إِكْتاباً	*'iktāb*^{an}	الكتاتيب	*al-katātīb*	الكِتابَ	*al-kitāba*
استكتبه	*'istaktabahu*	الكتبة	*al-kitba*^{t}	الكِتابَةُ	*al-kitāba*^{tu}
اسْتَكْتَبَه	*'istaktabahu*	الكتيبة	*al-katība*^{t}	الكِتابُ	*al-kitābu*
اسْتَكْتَبَها	*'istaktabahā*	وكَتيبة	*wa katība*^{t}	الكِتابِ	*al-kitābi*
اكتتب	*'iktataba*	الكَتائِبَ	*al-katā'iba*	المكاتب	*al-mukātib*
اكْتَتَبَ	*'iktataba*	الكَتائِبُ	*al-katā'ibu*	المكاتبة	*al-mukātiba*^{t}
اكْتَتَبَه	*'iktatabahu*	الكَتيبةُ	*al-katība*^{ta}	المكتب	*al-maktab*
اكْتَتَبَها	*'iktatabahā*	الكَتَائبَ	*al-katā'iba*	المكتبة	*al-maktaba*^{t}
اكْتُبْ	*'uktub*	الكَتَبَة	*al-kataba*^{t}	المكتوبة	*al-maktūba*^{t}
اكْتُتِبْت	*'uktutibtu*	الكَتْبُ	*al-katbu*	الْكُتَّابُ	*al-kuttābu*
اكْتِتابُك	*'iktitābuk*	الكَتْبِ	*al-katbi*	الْكِتَابَ	*al-kitāba*
اكْتِتابُكَ	*'iktitābuka*	الكُتَبُ	*al-kutabu*	الْكِتَابَةُ	*al-kitāba*^{tu}
الاكْتِتابُ	*al-'iktitābu*	الكُتَيْبَةُ	*al-kutayba*^{tu}	الْكِتَابَةِ	*al-kitāba*^{ti}
التكاتب	*at-takātubu*	الكُتَّابَ	*al-kuttāba*	الْمَكْتَبُ	*al-maktabu*
الكاتب	*al-kātib*	الكُتَّابِ	*al-kuttābi*	الْمَكْتُوبَةُ	*al-maktūba*^{tu}
الكاتِبُ	*al-kātibu*	الكُتْبة	*al-kutba*^{t}	اِسْتَكْتَبَ	*'istaktaba*

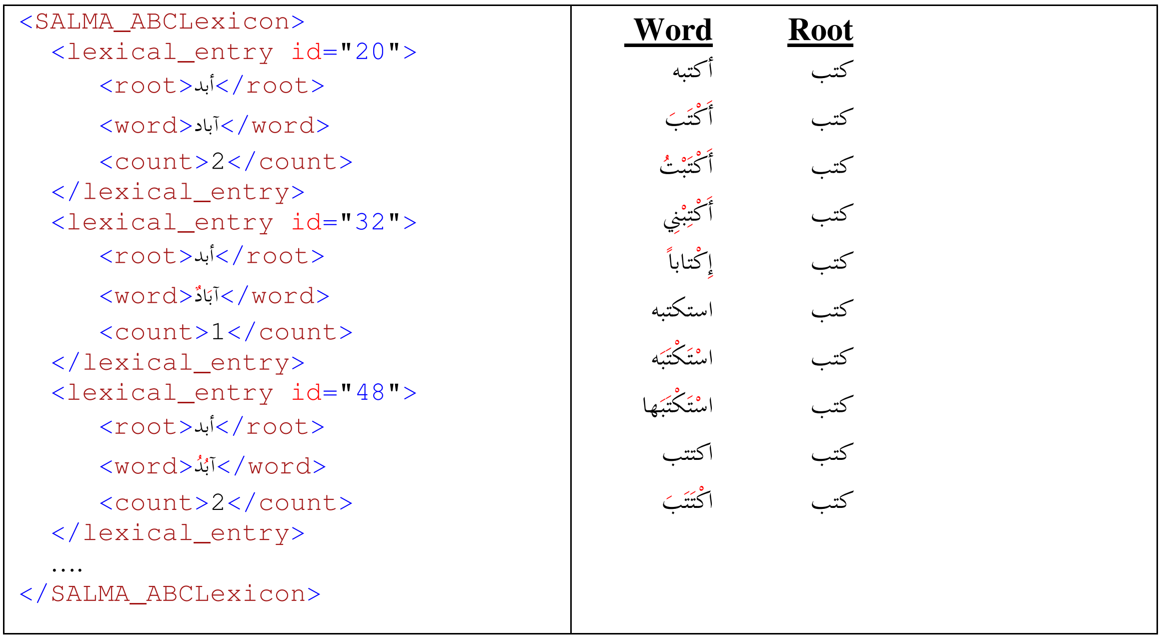
Figure 4.7 The first 60 lexical entries of the root كتب *k-t-b* 'wrote' stored in the SALMA – ABCLexicon

4.4.4 Format of the SALMA-ABCLexicon

Modern English dictionaries are stored using computerized lexicographic databases. The most widely accepted lexicographic database representation is lexical text markup using SGML (Standard Generalised Markup Language) such as XML. Other Database Management Systems (DBMS) can be used such as relational databases, object-oriented DBMS with inheritance mechanisms, and hybrid object-oriented/relational databases (Eynde and Gibbon 2000).

The Russell, Pulman et al. (1986) English morphological dictionary is stored as a sequence of entries, each in the form of a Lisp s-expression. MULTEXT, MULTEXT-East and CML is stored in tab separated column files (Erjavec 2010). SKEL lexicon is organized as a fixed number of pages, where each page contains a set of morphological entries (Petasis et al. 2001). The Latvian lexicon is stored in XML files (Paikens 2007). *Lefff* and the Slovak lexicons use Alexina framework (Sagot 2005; Sagot et al. 2006; Nicolas et al. 2008; Sagot 2010). Buckwalter's lexicon is stored as a relational database (Maamouri and Bies 2004; Maamouri et al. 2004).

Of these disparate formats, the SALAMA-ABCLexicon is stored as XML (Extensible Markup Language) files, as a relational database and tab separated column files. The three formats are used to ensure wider re-use of the lexicon in different text analytics applications for Arabic. Figure 4.8 shows the XML and tab separated column files. Figure 4.9 shows the entity diagram of the SALMA-ABCLexicon.



```
<SALMA_ABCLexicon>
  <lexical_entry id="20">
     <root>أبد</root>
     <word>آباد</word>
     <count>2</count>
  </lexical_entry>
  <lexical_entry id="32">
     <root>أبد</root>
     <word>آبَادٌ</word>
     <count>1</count>
  </lexical_entry>
  <lexical_entry id="48">
     <root>أبد</root>
     <word>آبُدُ</word>
     <count>2</count>
  </lexical_entry>
  ....
</SALMA_ABCLexicon>
```

Word	**Root**
أكتبه	كتب
أَكْتَبَ	كتب
أَكْتَبْتُ	كتب
أَكْتِبْنِي	كتب
إِكْتاباً	كتب
استكتبه	كتب
اسْتَكْتَبَه	كتب
اسْتَكْتَبَها	كتب
اكتتب	كتب
اكْتَتَبَ	كتب

Figure 4.8 XML and tab separated column files formats of the SALMA-ABCLexicon

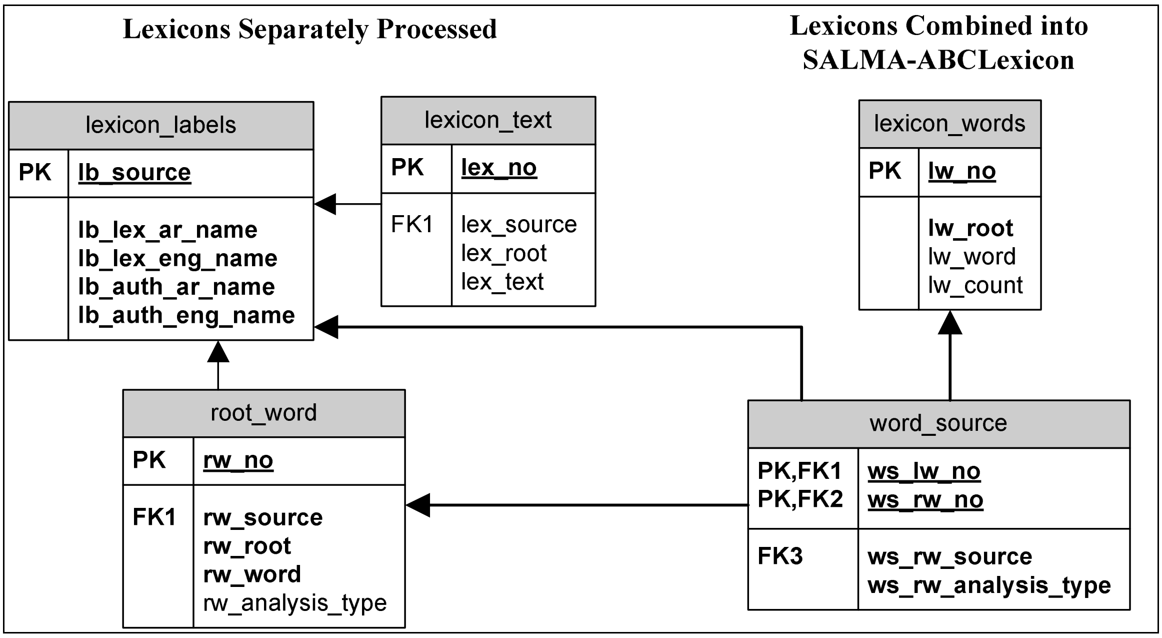


Figure 4.9 The entity relationship diagram of the SALMA-ABCLexicon

The first format uses XML to store the lexical entries of the SALMA-ABCLexicon. Each lexical entry has three pieces of information: Root, Word and Count. The Count is the number of times the word-root pair appeared in the lexicons text. The Count represents a verification criterion of the lexical entries. The second format uses a tab-separated column file where the first column represents the word and the second column represents the root. The last format uses relational databases to store the SALMA-ABCLexicon. The `lexicon_words` table represents the combined lexicon table. The `lexicon_words` table stores the Root, the Word and the Count. Simple SQLite3[40] was used to store and manage the lexicon database tables. SQLite is an open-source embedded SQL database engine which does not have a separate server process. SQLite reads and writes directly to ordinary disk files (*i.e.* is contained in a single disk file), which makes it a suitable choice for distributing the lexicon database file as a downloadable morphological database for Arabic.

4.4.5 Retrieval of the Lexical Entries

The lexicon has a searching facility that enables searching for a certain lexical entry in the lexicon, and returns back a Python object of type `LexiconEntry`. The `LexiconEntry` object represents an encapsulation of the word and its root as a unit of information; see figure 4.10. A specialized interface is provided to enable the morphological analyzer to communicate with the lexicon file; see section 8.3.2. This communication allows the morphological analyzer to retrieve the root(s) of the analyzed words. The `constructLexicon` function reads the tab separated column file and stores the lexicon in a dictionary data structure where the key of the dictionary is the non-vowelized word in string data type and the values of the dictionary are lists of LexiconEntry objects. The dictionary data structure of the lexicon is in this format

```
Lexicon = [nv_word:[LexiconEntry,...],...].
```

The `Lexicon` class interface represents the actual lexicon data and the communication facility between the lexicon and the morphological analyzer. Both `isLexiconEntry` and `getLexiconEntry` check whether the passed non-vowelized Arabic word is found in the lexicon and returns a list of `LexiconEntry` objects for the non-vowelized words found. Figure 4.10 shows the lexicon Python classes interface and the lexicon construction method – the implementation of the class methods is not included.

[40] SQLite http://www.sqlite.org/

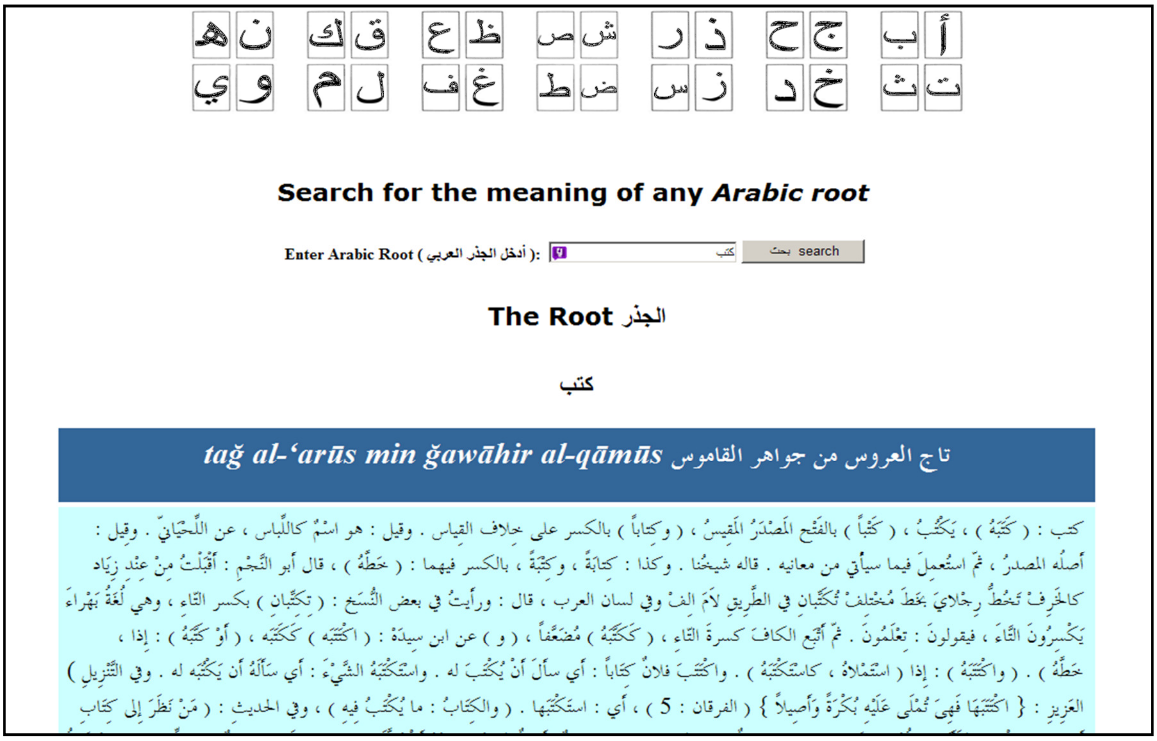
```
class LexiconEntry(object):
    def __init__(self, word, root):
        self.word = ArabicWord(word)
        self.root = ArabicWord(root)
    def __str__(self):
    def printLexEntry(self):

def constructLexicon():
    ''' This procedude reads the lexicon file and constructs the
        lexiocn dictionary of the following format
        {nv_word:[LexiconEntry,...],..., }'''
    return lexicon

class Lexicon(object):
    '''Lexicon class constructs the lexicon dictionary'''
    LexDict = constructLexicon()
    def printLexicon(cls):
    def isLexiconEntry(cls, nv_word): # return True or False
    def getLexiconEntry(cls, nv_word):
        return Lexicon.LexDict[nv_word]
```

Figure 4.10 Lexicon Python Classes interface – implementation of the methods is not included

A web interface[41] was developed to allow users to access the contents of the lexicon, to search for a given root. The interface searches the lexicon's relational database tables for the entered root and displays the definition parts from the analyzed lexicons. Figure 4.11 shows the web interface of the 7 analyzed traditional Arabic lexicons.



Search for the meaning of any *Arabic root*

Enter Arabic Root (أدخل الجذر العربي): كتب

The Root الجذر

كتب

tağ al-'arūs min ğawāhir al-qāmūs تاج العروس من جواهر القاموس

كتب : (كَتَبَهُ) ، يَكْتُبُ ، (كَتْباً) بالفَتْح المَصْدَرُ المَقِيسُ ، (وكِتاباً) بالكسر على خِلاف القِياس . وقيل : هو اسْمٌ كاللِّباس ، عن اللِّحْيانيّ . وقِيل :
أصلُه المصدرُ ، ثمّ استُعمِلَ فيما سيأْتي من معانيه . قاله شيخُنا . وكذا : كِتابَةً ، وكِتْبَةً ، بالكسر فيهما : (خَطَّهُ) ، قال أبو النَّجْمِ : أَقْبَلْتُ مِنْ عِنْدِ زِيَاد
كالخَرِفْ تَخُطُّ رِجْلايَ بِخَطٍّ مُخْتلِفْ تُكَتِّبانِ في الطَّرِيقِ لاَمَ الِفْ وفي لسان العرب ، قال : ورأَيتُ في بعض النُّسَخ : (تِكِتِّبانِ) بكسر التّاء ، وهي لُغَةُ بَهْرَاء
يَكْسِرُونَ التَّاءَ ، فيقولونَ : تِعْلَمُونَ . ثمّ أَتْبع الكافَ كسرةَ التّاءِ ، (كَكَتَّبَهُ) مُضَعَّفاً ، (و) عن ابن سِيدَهْ : (اكْتَتَبَه) كَكَتَبَه ، (أَوْ كَتَّبَهُ) : إذا ،
خَطَّهُ) . (واكْتَتَبَهُ) : إذا (استَمْلاهُ ، كاستَكْتَبَهُ) . واكْتَتَبَ فلانٌ كِتَاباً : أي سأَلَ أَنْ يُكْتَبَ له . واستَكْتَبَهُ الشَّيْءَ : أي سَأَلَهُ أن يَكْتُبَه له . وفي التَّنْزِيلِ)
العَزِيزِ : { اكْتَتَبَهَا فَهِيَ تُمْلَى عَلَيْهِ بُكْرَةً وَأَصِيلاً } (الفرقان : 5) ، أي : استَكْتَبَها . (والكِتَابُ : ما يُكْتَبُ فِيهِ) ، وفي الحديث : (مَنْ نَظَرَ إلى كِتَاب

Figure 4.11 Web interface for searching the traditional Arabic lexicons

[41] A web interface for searching the traditional Arabic lexicons for a certain root http://www.comp.leeds.ac.uk/cgi-bin/scmss/arabic_roots.py

4.5 Evaluation of the SALMA-ABCLexicon

The SALMA-ABCLexicon was evaluated by computing the coverage of the lexicon on different types of text corpora: the Qur'an; the Arabic Internet Corpus[42]; and the Corpus of Contemporary Arabic (CCA). Two experiments were carried out compute the coverage of the SALMA-ABCLexicon. First, exact match where each non-vowelized word in the test corpora is searched for in the lexicon. The results showed that the coverage of the three corpora is 65.5% - 67.5%. The highest coverage of 67.53% was achieved from the Qur'an. The coverage of both the Internet Arabic corpus and the CCA achieved 65.58% and 65.44% respectively. Table 4.4 and figure 4.12 show the coverage percentage of the SALMA-ABCLexicon using exact match. Table 4.4 shows the number of tokens and words in each corpus. Some tokens are not words (*i.e.* Arabic words) but numbers, dates, currency symbols, punctuations, HTML or XML tags and English words. Only Arabic words were selected to compute the coverage of the SALMA-ABCLexicon.

Table 4.4 The coverage of the lexicon using exact word-match method

Corpus	**Tokens**	**Words**	**Covered words**	**Coverage %**
Qur'an	77,800	77,799	52,536	67.53%
CCA	684,726	594,664	389,133	65.44%
Internet	1,128,114	833,916	546,880	65.58%

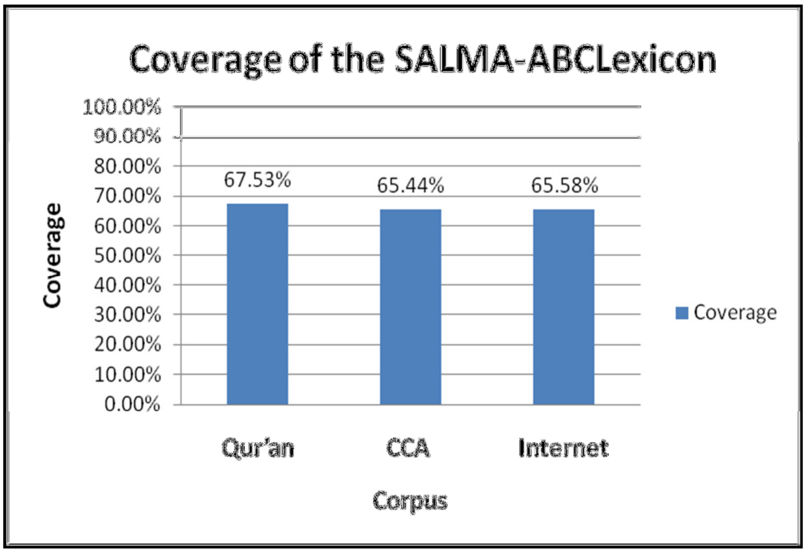


Figure 4.12 The coverage of the SALMA-ABCLexicon using exact match method

An Arabic word in any text may appear with many different forms of clitics attached to it, which makes the matching process of the word and the lexical entries not an easy task and decreases the coverage. The second experiment to compute the coverage of the SALMA-ABCLexicon is through an application that depends on it. The lemmatizer (Sawalha and Atwell 2011a) for Arabic text is used to process large-scale real data; the

[42] Leeds collection of Internet corpora: Arabic Internet Corpus http://corpus.leeds.ac.uk/internet.html

Arabic Internet Corpus which consists of 176 million words of Arabic collected from web pages. The lemmatizer depends on the SALMA-ABCLexicon to extract the root and generate the lemma of the word. Each word is tokenized into different forms consisting of proclitics, stem and enclitics, and then each stem is searched in the lexicon. If the stem is found in the lexicon then the root and the vowelized stems stored in the SALMA-ABCLexicon are retrieved. More details about the lemmatizer are given in chapters 8 and 10. When a correct analysis is retrieved from the lexicon then it is counted as a valid lexicon reference. The coverage of the SALMA-ABCLexicon is computed by the percentage of valid lexicon references to the number of words in the test sample. The lemmatizer uses three other linguistic lists; a list of function words (stop words) which have fixed syntactic analysis in any context (Diwan, 2004), a named entities list (Benajiba, Diab and Rosso 2008) and a list of broken plurals[43] (Elghamry 2010). The coverage of the SALMA-ABCLexicon was computed one time with the inclusion of these function word lists (*i.e.* function words list, named entities list and broken plurals), and another time without including the function word lists. Tables 4.5 and 4.6 show the coverage percentage of the lexicon computed using the lemmatizer program. Figure 4.13 shows a summary of the coverage of the SALMA-ABCLexicon using the lemmatizer.

Table 4.5 Coverage including function words

Corpus	Tokens	Words	Covered words	Coverage %
Qur'an	77,804	77,803	64,065	82.34%
CCA	685,161	595,099	507,943	85.35%
Internet	1,128,624	834,426	708,101	84.86%

Table 4.6 Coverage excluding function words

Corpus	Tokens	Words	Covered words	Coverage %
Qur'an	77,804	54,004	42,532	78.76%
CCA	685,161	411,482	338,790	82.33%
Internet	1,128,624	576,407	476,190	82.61%

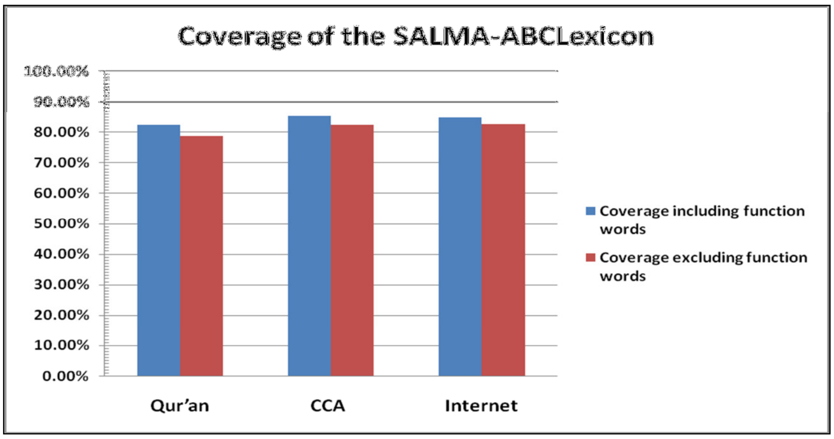


Figure 4.13 Coverage percentage of the SALMA-ABCLexicon using the lemmatizer

[43] Broken plural list source http://sites.google.com/site/elghamryk/arabiclanguageresources

The coverage is about 85% of the words, including function words, and about 82% of the words excluding function words. Both the CCA and the Arabic Internet Corpus achieved similar results when testing using the lemmatizer program and including function words. The coverage for them was 85.35% and 84.86% respectively. A coverage of 82.34% was achieved when analysing the Qur'an words. The second part of the experiment excluded the function words. Similar results were achieved. The Arabic Internet Corpus and the CCA scored 82.61% and 82.33% respectively. The coverage resulted from analyzing the Qur'an text was 78.76%.

Common words which are not covered by the SALMA-ABCLexicon include: function words (stop words); new Arabic terms; relative nouns; and borrowed words (Arabized words). Functional words (stop words)such as ذَلِكَ *ḏālika* "that"; وَإِلَى *wa-'ilā* "and to"; إِنَّهُمْ*'innahum* "they are"; and التي *allatī* "which", can be easily added to the lexicon along with their syntactical and morphological analysis by collecting them from traditional Arabic grammar books such as (Diwan 2004). New Arabic terms such as دردشة *dardaša*^{t} "chat"; انقر*'unqur* "click" and الانتخابات *al-'intiẖābāt* "elections" are not covered in the lexicon because such words have appeared recently due to modern technological development and the failure to add them to the traditional Arabic lexicons. Relative nouns الأسماء المنسوبة *al-'asmā' al-mansūba*^{h} are nouns that indicate affiliation of something to these nouns. See section 6.2.2. Relative nouns such as السياحية *as-siyāḥyya*^{t} "tourism"; الاجتماعية *al-iğtimāʿiyya*^{t} "social"; and الثقافية *aṯ-ṯaqāfiyya*^{t} "cultural" have become widely used in the media and modern standard Arabic. Borrowed words (Arabized words) such as الدكتور *ad-duktūr* "doctor"; الإيميل *al-'imayl* "e-mail"; التليفون *at-tilifūn* "telephone"; and الإنترنت *al-'intarnit* "Internet" are foreign words transliterated into Arabic by writing the word using Arabic letters. This is a common problem found in newspaper and web pages text due to the lack of the correct translation of the borrowed words which will increase the frequency of this type of word in contemporary Arabic text. Figure 14 shows a sample of common words not covered by the broad-coverage lexical resource.

ذَلِكَ	*ḏālika*	That	الاقتصادية	*al-’iqtiṣādiyya*^{t}	Economical
السَّمَاوَاتِ	*assamāwāti*	Skies	الإنسان	*al’insān*	The human
إِنَّهُمْ	*’innahum*	They are	الإيميل	*al-’īmayl*	E-mail
بِاللَّهِ	*billāhi*	Swear to God	التليفون	*at-tilifūn*	Telephone
عَنْهُمْ	*‘anhum*	After them	الفلسطيني	*al-filasṭīnī*	Palestinian
بِالْحَقِّ	*bilḥaqqi*	By the right	دردشة	*dardaša*^{t}	Chat
فَأُوْلَئِكَ	*fa’ulā’ika*	And those	انقر	*’unqur*	Click
فَبِأَيِّ	*fabi’ayyi*	In what	الأمريكية	*al-’amrīkiyya*^{t}	American
وَإِلَى	*wa-’ilā*	And to	الداخلية	*ad-dāẖiliyya*^{t}	Interior
فَسَوْفَ	*fasawfa*	It will	الانتخابات	*al-’intiẖābāt*	Elections
التي	*allatī*	which	الولايات	*al-wilāyāt*	States
المتحدة	*al-muttaḥida*^{t}	United	الاجتماعية	*al-iğtimā‘iyya*^{t}	Social
الدكتور	*ad-duktūr*	Doctor	الإنترنت	*al-’intarnit*	Internet
السياحية	*as-siyāḥiyya*^{t}	Tourism	التنمية	*at-tanmiya*^{t}	Developmental
الغربية	*al-ḡarbiyya*^{t}	Western	الثقافية	*aṯ-ṯaqāfiyya*^{t}	Cultural

Figure 4.14 A sample of common words which are not covered by the lexicon

4.6 The Corpus of Traditional Arabic Lexicons

Al-Sulaiti and Atwell (2006) developed the Corpus of Contemporary Arabic. This corpus contains 1 million words taken from different genres collected from newspapers and magazines. It contains the following domains; Autobiography, Short Stories, Children's Stories, Economics, Education, Health and Medicine, Interviews, Politics, Recipes, Religion, Sociology, Science, Sports, Tourist and Travel and Science. Like most Arabic corpora, the text of the Corpus Contemporary Arabic is taken from newspapers and magazines.

The Corpus of Traditional Arabic Lexicons consists of the text of 23 freely available traditional Arabic lexicons. This corpus has a different domain than existing corpora of contemporary Arabic. It covers a period of more than 1,300 years and consists of a large number of words (14,369,570) and word types (2,184,315). It also has both vowelized and non-vowelized text. Figure 4.15 shows the number of words and word types and the 25 words of highest frequency.

Partially-vowelized			Non-vowelized		
Word		**Frequency**	**Word**		**Frequency**
في	*fī* “in”	292,396	من	*min* “from”	322,239
من	*min* “from”	269,200	في	*fī* “in”	301,895
قال	*qāl* “he said”	172,631	قال	*qāl* “he said”	190,918
و	*wa* “and”	120,060	أي	*’ayy* “which”	132,635
على	*‘alā* “over”	108,252	و	*wa* “and”	130,809
ما	*mā* “what”	89,195	على	*‘alā* “over”	119,639
وقال	*wa qāl* “and he said”	88,233	إذا	*’iẖā* “if”	115,842
عن	*‘an* “about”	82,027	وقال	*wa qāl* “and he said”	99,601
إذا	*’iẖā* “if”	81,479	ابن	*’ibn* “son of”	94,980
أي	*’ay* “which”	78,622	ما	*mā* “what”	94,530
وهو	*wa huwa* “and he”	75,149	بن	*bin* “son of”	92,213
لا	*lā* “no”	69,737	عن	*‘an* “about”	87,064
ابن	*’ibn* “son of”	58,334	وهو	*wa huwa* “and he”	80,375
به	*bihi* “in it”	53,343	لا	*lā* “no”	73,066
وفي	*wa fī* “and in”	53,197	أبو	*abū* “father”	72,231
وقد	*wa qad* “and perhaps”	50,648	أن	*’an* “that”	65,419
أبو	*abū* “father”	47,915	أو	*’aw* “or”	62,298
بن	*bin* “son of”	46,880	الله	*allā*^{h} “Allah”	59,511
أَي	*’ay* “which”	46,788	به	*bihi* “in it”	58,941
هو	*huwa* “he”	45,916	يقال	*yuqāl* “it is said”	58,062
يقال	*yuqāl* “it is said”	45,794	وفي	*wa fī* “and in”	55,077
عليه	*‘alayhi* “about him”	44,786	وقد	*wa qad* “and perhaps”	53,992
ولا	*wa lā* “and not”	42,190	عليه	*‘alayhi* “about him”	50,906
الله	*allā*^{h} “Allah”	39,961	هو	*huwa* “he”	49,785
أو	*’aw* “or”	39,210	إلى	*’ilā* “to”	48,363

Figure 4.15 The Corpus of Traditional Arabic Lexicons frequency list

The Corpus of Traditional Arabic Lexicons is stored using 247 text files (178MB) using Unicode “utf-8” encoding. The text files contain the original lexicons’ text with the original ordering of the lexical entries. Another structured format for the corpus was created using XML technology. Seven lexicons which were analyzed to construct the SALMA-ABCLexicon, see section 4.4.2, were reformatted in alphabetical order of their lexical entries and stored in XML files. Figure 4.16 shows the XML structure used to store the corpus files. Note that XML version includes only seven lexicons.

```
<Lexicon id = "1" ar_name = "تاج العروس من جواهر القاموس" eng_name = "tağ al-‘arūs min ğawāhir al-qāmūs " author_ar = "الزبيدي" author_eng = "az-zubaydī">
…
<lexicon_entry id = "8391">
<root>كتب</root>
<text> وقيل: هو اسْمٌ كاللِّباس ، عن اللِّحْيَانِيّ. وقيل: بالكسر على خِلاف القِياس (وكِتاباً) بالفَتْح المَصْدَرُ المَقِيسُ ، (كَتْباً) ، يَكْتُبُ ، (كَتَبَهُ) :كتب
أَقْبَلْتُ مِنْ عِنْدِ زِيَاد كالخَرِفْ تَخُطُّ :، قال أبو النَّجْمِ (خَطَّهُ) :كِتابَةً ، وكِتْبَةً ، بالكسر فيهما :وكذا .قاله شيخُنا .أَصلُه المصدرُ ، ثمّ استُعمِلَ فيما سيأْتي من معانيه
بكسر التّاءِ ، وهي لُغَةُ بَهْراءَ ، يَكْسِرُونَ التَّاءَ ، فيقولونَ (تِكِتِّبانِ) :ورأَيتُ في بعض النُّسَخ :رِجْلايَ بِخَطّ مُخْتلِفْ تُكَتِّبانِ في الطَّرِيقِ لاَمَ الِفْ وفي لسان العرب ، قال
اسْتَمْلاهُ ، )إذا : (واكْتَتَبَهُ) . (خَطَّهُ )إذا : (أَوْ كَتَّبَهُ )كَكَتَبَه ، (اكْتَتَبَه) :عن ابن سِيدَهْ (و )مُضَعَّفاً ، (كَكَتَّبَهُ )ثمّ أَتْبَع الكافَ كسرةَ التّاءِ ، .تِعْلَمُونَ
( اكْتَتَبَهَا فَهِىَ تُمْلَى عَلَيْهِ بُكْرَةً وَأَصِيلاً } :وفي التَّنْزِيلِ العَزِيزِ .أَي سَأَلَهُ أَن يَكْتُبَه له :واسْتَكْتَبَهُ الشَّيْءَ .أَي سأَلَ أَنْ يُكْتَبَ له :واكْتَتَبَ فلانٌ كِتَاباً . (كاسْتَكْتَبَهُ
الفرقان: 5 ) أَي :، استَكْتَبَها... </text>
</lexicon_entry>

…
<lexicon_entry id = "9657">
<root>نجح</root>
<text> (وأَنْجَحَهَا اللَّهُ تعالى ) .وأَنْجَحْتُهَا لك (نَجَحَتِ الحاجَةُ ، كمَنَعَ ، وأَنْجَحَتْ )وقد .والفَوْزُ (الظَّفَرُ بالشَّيْءِ :النَّجَاح ، بالفتْح ، والنُّجْحُ بالضّمّ ) :نجح
وفي خطبة عائشة رضي الله عنها .وقد أَنْجَحْتُ حَاجتَه ، إِذا قضيْتها له . (مَنَاجِيحَ ومَنَاجِحَ )قومٍ (وهو مُنْجِحٌ ، مِنْ .صارَ ذا نُجْحٍ :وأَنْجَحَ زَيْدٌ ) .أَسْعَفَه بإِدراكِهَا :
والنَّجِيحُ ) .وبالله أَتَفتِحُ ، وإِيّاه أَستَنْجِحُ :ومن سَجَعَات الأَساس .، ونَجَحتْ هي (تَنَجَّزَهَا )، إِذا (وتَنَجَّجَ الحَاجَةَ واستَنْجَحَهَا ) . (وأَنْجَحَ إِذْ أَكْدَيْتُم ) :
رجُلٌ مُنْجِح : (الأَساس )نَجِيحٌ جَوَادٌ أَحو مَأْقِطٍ نِقَابٌ يُحَدِّثُ بالغائب وفي :، أَي مُنْجِحُ الحاجاتِ ، قال أَوْسٌ (المُنْجِحُ من النّاس ) :النَّجِيحُ (الصَّوابُ مِنَ الرَّأْيِ و
...وَشِيكٌ :، سَيْرٌ ناجحٌ ونَجِيحٌ (كالنَّاجِحِ )سارَ فلانٌ سَيراً نَجِيحاً ، أَي وَشيكاً ، :، يقال (الشَّدِيدُ مِن السَّيْرِ ) :النَّجيح :من المجاز (و ) .ذو نُجْحٍ :
</text>
</lexicon_entry>

…
</Lexicon>
```

Figure 4.16 XML structure of The Corpus of Traditional Arabic Lexicons

4.7 Discussion of the Results, Limitations and Improvement

The SALMA-ABCLexicon contains a large number of entries representing a wide coverage of Arabic words, word types and roots. The evaluation proved that the lexicon has wide coverage, where about 85% of the test corpora words have a valid reference to the lexicon entries. Despite the time span of 13 centuries of the traditional Arabic lexicons from which the SALMA-ABCLexicon has been derived, 15% of the test corpora words are not captured. The latest analyzed Arabic lexicon is المعجم الوسيط *al-mu‘ğam al-wasīṭ* which appeared in 1960s; so, new vocabulary items added to Arabic in the past 50 years is not included in the lexicon. Moreover, the use of borrowed words from foreign languages which do not have a proper translation in Arabic, but are written using Arabic letters (transliterated) has increased due to the technological advances. Advances in technology and communication means new products and their names have entered Arab countries, where these products keep their original names which have been widely used and become part of contemporary Arabic vocabulary. Moreover, the use of dialectical Arabic has increased in the written language due to open systems such as chat rooms, blogs and forums, which allow people to write text without restrictions on the web where they use dialectical words quite frequently.

The lexicon did not involve any manual correction due to the limitations of funding the correction process and voluntary work to correct the lexicon. However, the methodology followed to verify part of the lexicon was done by counting how many times the word-root pairs appear in the analyzed traditional Arabic lexicons. 976,427 word-root pairs representing 35.19% of the lexicon's word-root pairs scored a count of 2 or more. This means that these word-root pairs appeared in different lexicons and satisfied the linguistic knowledge of the two extraction conditions. Therefore, these word-root pairs have high potential to be valid and correct.

This is the first version of the SALMA-ABCLexicon. It can be extended to include the full morphological analyses of the lexical entries and other useful information that will enhance the accuracy of NLP applications. Special linguistic lists such as compounds, collocations, idiomatic phrases, phrasal verbs and named entities can be added to extend the lexicon. Moreover, morphological lists such as broken plurals, intransitive and transitive verbs, rational and irrational words and primitive nouns can be another extension to the lexicon. Chapter 8 will discuss the extension of the SALMA-ABCLexicon by adding special linguistic and morphological lists to enhance the guessing of the morphological features of the words by the developed morphological analyzer. The SALMA-ABCLexicon can also be extended by adding modern and dialect vocabulary from Corpus of Contemporary Arabic and Arabic Internet Corpus. But these corpora can only extend the vocabulary; the corpus does not provide a root for each word.

Manual correction of the word-roots pairs can be done in the future to make the SALMA-Lexicon an authenticated resource which can be used as a gold standard for stemming algorithms to be evaluated against a wide-coverage gold standard.

The SALMA-ABCLexicon is an open-source lexicon. There is also an online access method to its contents and searching facilities[44].

[44] SALMA-ABCLexicon http://www.comp.leeds.ac.uk/sawalha/SALMA-ABCLexicon.html

4.8 Chapter Summary

This chapter showed the process of constructing the SALMA-ABCLexicon to be used in Arabic text analytics applications such as lemmatizers, morphological analyzers and part-of-speech taggers. The motivations for constructing the SALMA-ABCLexicon are: the poor results achieved by comparing the outputs of existing morphological analyzers and stemmers discussed in chapter 3; the benefits gained by developing a morphological resource over developing a sophisticated stemming algorithm; the ability to reuse the SALMA-ABCLexicon in different Arabic text analytics applications; and the use of the text to construct the Corpus of Traditional Arabic Lexicons.

The chapter started by surveying morphological lexicons especially for Arabic and morphologically rich languages (mainly east European languages). The survey focused on the language of the lexicon, the construction methodology, the size and the evaluation of the lexicons. This was followed by the study of traditional Arabic lexicons focusing on the arrangement methodologies and the challenges and drawbacks of these lexicons. The focus of the survey was to investigate the agreed standard requirements for constructing morphological lexicons from raw text.

The development of constructing the SALMA-ABCLexicon followed the agreed standard for constructing a morphological lexicon from raw text. However, the absence of a large open-source representative Arabic corpus, the absence of an open-source generation programme and the generation programme problems of over-generation and under-generation, directed the selection of the raw text corpus to be the text of the traditional Arabic lexicons to substitute for the corpus and the generation program requirements. The major advantages of using the traditional Arabic lexicons text as a corpus are: the corpus contains a large number of words and word types and the possibility of finding the different forms of the derived words of a given root.

The SALMA-ABCLexicon is constructed by combining extracted information from disparate lexical resource formats and merging Arabic lexicons. The processing steps in constructing the SALMA-ABCLexicon involve; first, analyzing lexicon texts separately by manually converting each lexicon text into a unified format by choosing the most common format for all root entries. Then, for each lexicon a specialized program extracts the root and the words derived from that root depending on linguistic knowledge that governs the derivation of words from their roots. Second, a combination algorithm merges the information extracted from the previous step into one large broad-coverage lexical resource, the SALMA-ABCLexicon.

The evaluation of the SALMA-ABCLexicon was done by computing the coverage, using two methods: the first methodology computed the coverage by matching the words of the test corpora to the words in the lexicon, which scored about 67%. The second methodology used a lemmatizer program to compute the coverage, and scored about 82%.

The SALMA-ABCLexicon contains 2,781,796 vowelized word-root pairs which represent 509,506 different non-vowelized words. The lexicon is stored in three different formats: tab-separated column files; XML files; and relational database. It is also provided with access and searching facilities and a web interface that provide searching for a certain root and retrieving the original root definitions of the analyzed traditional Arabic lexicons. The different formats and the access and search facilities will increase the reusability of the lexicon in different Arabic text analytics applications. The SALMA-ABCLexicon is an open-source morphological resource.

The Corpus of Traditional Arabic Lexicons is a special corpus which is constructed from the text of 23 traditional Arabic lexicons. The corpus contains 14,369,570 words and 2,184,315 word types. The corpus is stored using three formats: text files encoded using Unicode utf-8; XML files; and a relational database. The corpus is an open-source resource for Arabic.

Chapter 5
Survey of Arabic Morphosyntactic Tag Sets and Standards; Background to Designing the SALMA Tag Set

This chapter is based on the following sections of published papers:

Sections 2, 3, 4, and 5 are based on sections 1.3, 1.4, 2 and 3 from (Sawalha and Atwell Under review)

Chapter Summary

A range of existing Arabic Part-of-Speech tag sets are illustrated and compared, and generic design criteria for corpus part-of-speech tag sets is reviewed in this chapter. Eight existing morphosyntactic annotation schemes for Arabic are compared in terms of the purpose of design, tag set characteristics, tag set size, and their applications. The main characteristics of the SALMA – Tag Set are to be: general purpose; reusable; and adhering to standards. The SALMA – Tag Set is not tied to a specific tagging algorithm or theory, and other tag sets could be mapped onto this standard, to simplify and promote comparisons between and reuse of Arabic taggers and tagged corpora. Sophisticated morphological and syntactic knowledge was extracted from traditional Arabic grammar books, then classified and used as a standard for the design of the SALMA – Tag Set. Tag set design criteria proposed by Atwell (2008) were applied and design decisions were investigated to handle each design dimension.

5.1 Introduction

The prerequisite for Part-of-speech annotation of corpora is a previously defined part-of-speech annotation scheme (Hardie 2004). The annotation scheme describes the morphosyntactic categories and enables annotators (human or computers) to label the corpus words by giving each word a label from the list of morphosyntactic categories according to its context; this is called a tag set.

Since the development of the Brown Corpus in 1963-1964, tag sets for English evolved. The Brown Corpus tagset has 87 tags. A smaller tagset for English is the 45-tag Penn Treebank tagset used to tag the Penn Treebank. A middle size of 61 tags for English is the C5 tagset used by the Lancaster UCREL project's CLAWS (The Constituent Likelihood Automatic Word Tagging System) to tag the British National Corpus (BNC). The current standard tagset for CLAWS is the 164-tag C7 tagset (Jurafsky and Martin 2008).

AMALGAM[45] (Automatic Mapping Among Lexico-Grammatical Annotation Models) multi-tagged corpus is pos-tagged according to a range of rival English corpus tagging schemes. These tagging schemes include: Brown corpus; ICE (International Corpus of English); LLC (London-Lund Corpus); LOB (Lancaster-Oslo/Bergen Corpus); PARTS (*i.e.* tag set used to tag the Spoken Corpus Recordings In British English SCRIBE); PoW (Polytechnic of Wales corpus); SEC (Spoken English Corpus); and UPenn (University of Pennsylvania corpus). Figure 5.1 shows an example of a sentence from the AMLGAM multi-tagged corpus illustrating the 8 tagging schemes used to tag the same sentence (Atwell 2007; Atwell 2008).

	Brown	**ICE**	**LLC**	**LOB**	**PARTS**	**PoW**	**SEC**	**UPenn**
select	VB	V(montr,imp)	VA+0	VB	adj	M	VB	VB
the	AT	ART(def)	TA	ATI	art	DD	ATI	DT
text	NN	N(com,sing)	NC	NN	noun	H	NN	NN
you	PPSS	PRON(pers)	RC	PP2	pron	HP	PP2	PRP
want	VB	V(montr,pres)	VA+0	VB	verb	M	VB	VBP
to	TO	PRTCL(to)	PD	TO	verb	I	TO	TO
protect	VB	V(montr,infin)	VA+0	VB	verb	M	VB	VB
.	.	PUNC(per)	.	.	.	.	.	.

Figure 5.1 Example sentence illustrating rival English part-of-speech tagging (from the ALMAGAM multi-tagged corpus)

Besides the evolution of the part-of-speech tag sets, standards and guidelines for morphosyntatic annotation of text corpora appeared. These standards and guidelines provide sophisticated knowledge of morphology and syntax where various heuristics are

[45] The AMALGAM project http://www.comp.leeds.ac.uk/amalgam/amalgam/amalghome.htm

given in the tagging manuals to help humans and computers to make decisions in pos-tagging the corpus (Jurafsky and Martin 2008). EAGLES (Expert Advisory Group on Language Engineering Standards) has become a widely used and most important recent standard for morphosyntactic annotation for Indo-European languages. The EAGLES guidelines were proposed in the interest of comparability, interchangeability and reusability of annotated corpora (Leech and Wilson 1996). Many morphosyntactic schemes for different languages applied the EAGLES guidelines. Example projects are: the MULTEXT project; the GRACE project; the CRATER project; and the morphosyntactic tag set of Urdu. The four projects and the tag set of Urdu are discussed in Hardie (2003 and 2004).

This chapter provides a background review of existing Arabic tag sets and discusses the design standards and guidelines applied in designing the morphological features tag set of Arabic, the SALMA Tag Set. The chapter starts by introducing traditional Arabic grammar in section 5.2. A survey and a comparative evaluation of existing Arabic part-of-speech tag sets are made in section 5.3. Section 5.4 discusses the design criteria proposed by Atwell (2008), which is applied in the design of the SALMA Tag Set. Finally, the complex morphology of Arabic is discussed in section 5.5.

5.2 Traditional Arabic Part-of-Speech Classification

Arabic, unlike English and modern European languages, has a long traditional of scholarly research into its grammatical description, spanning over a millennium. Most traditional Arabic grammar studies follow the order established by سِيبَوَيْه *Sībawayh*, about fourteen hundred years ago. It starts with syntax نحو *naḥw*, followed by morphology تصريف *taṣrīf*, and phonology علم الأصوات *'ilm al-'aṣwāt*. The grammarian's main preoccupation was the explanation of the case ending of the words in the sentence, called إعراب *'i'rāb*. The term originally meant the correct use of Arabic according to the language of the Bedouins but came to mean declension. Classical Arabic linguists classify words into three main parts of speech: Noun, name of a person, place, or object which does not have any tense; Verb, a word which indicates an action and has tense; and Particle, a word which cannot be understood without joining with a noun or a verb or both. However, there are also morphological criteria for this classification: a verb can be defined as a word derived from a specified morphological pattern, and has morphological features such as person and mood; while a noun can be definite or indefinite and has number and gender features. Derived nouns, which are derived from verbs, may have the same pattern with verbs. Particles are considered the most idiosyncratic words in Arabic, as these particles might span several grammatical categories. For example the particle *wa* وَ can indicate a conjunction between two adjectives قَضيتُ وَقْتاً سعيداً وَ مُمتعاً في الحَفْلَةِ *qaḍaytu waqt*^{an} *sa'īd*^{an} *wa* *mumti*$^{‘an}$ *fī al-ḥaflati* ‘I spent an interesting and happy time at the party’. While, in another case, the same particle *wa* وَ functions as locative preposition in the sentence مَشَيتُ وَ النَّهرَ *mašaytu* *wa* *an-nahra* ‘I walked along the river’(Al-Ghalayyni 2005).

Arabic is a highly inflectional language, and the traditional classification into nouns, verbs and particles does not say much about word structure. Arabic has many morphological and grammatical features, including sub-categories, person, number, gender, case, mood, etc. (Atwell 2008). A more fine-grained tag set is more appropriate for morphology research. The additional information may also help to disambiguate the base grammatical class (Schmid and Laws 2008). We aim to develop a part-of-speech tagger for annotating general-purpose Arabic corpus resources, in a wide range of text formats, domains and genres, including both vowelized and non-vowelized text; enriching the text with linguistic analysis will maximize the potential for corpus re-use in a wide range of applications. We foresee an advantage in enriching the text with part-of-speech tags showing very fine-grained grammatical distinctions, which reflect expert interest in syntax and morphology, rather than specific needs of end-users, because end-user applications are not known in advance.

Very fine-grain distinctions may cause problems for automatic tagging if some words can change grammatical tag depending on function and context (Atwell 2008); on the other hand, fine-grained distinctions may actually help to disambiguate other words in the local context. Practical experiments using a fine-grain morphological tag set were reported by (Schmid and Laws 2008). Their experiments were carried out using German and Czech as examples of highly inflectional languages. Their HMM part-of-speech tagger makes good use of the fine-grain tag set; it splits the part-of-speech into attribute vectors and estimates the conditional probabilities of the attribute with decision trees. This method achieved a higher tagging accuracy than two state-of-the-art general-purpose part-of-speech taggers (TnT and SVMTool). We believe that this kind of approach may yield better results for an Arabic part-of-speech tag set including fine-grained morphological features.

5.3 Existing Arabic Part-of-Speech Tag Sets

This section covers the most important Arabic tag sets and tag set design methodologies. These tag sets are; (1) Khoja’s Arabic tag set, (2) Penn Arabic Treebank tag set, (3) ARBTAGS, (4) The Quranic Arabic Corpus morphological tag set, (5) The MorphoChallenge 2009 Qur’an Gold Standard tag set and (6) CATiB part-of-speech tag set. The section describes each tag set and their characteristics, and a comparison table illustrates the differences between the different Arabic tag sets. The tag sets range from a small set of short tags analogous to BNC or LOB tag sets for English on one hand, to longer more detailed morphological tag sets (*e.g.* Penn Arabic Treebank (FULL) tag set) which are analogous to the ICE tag set for English.

5.3.1 Khoja’s Arabic Tag Set

During early research on developing a part-of-speech tagger for Arabic text, (Khoja, Garside and Knowles 2001; Khoja 2003) developed a tag set for Arabic which is based on traditional Arabic grammar categories rather than modern European EAGLES standards. The reasons for not following EAGLES morphosyntactic guidelines were: Arabic belongs to the Semitic language family while EAGLES guidelines were designed for European languages; and following EAGLES guidelines would not capture some Arabic morphosyntactic information such as imperative or jussive mood, dual number and inheritance. Inheritance is an important aspect of Arabic, where all subclasses of words inherit properties from the classes they are derived from. Khoja’s tag set contains 177 tags; 103 types of noun, 57 verbs, 9 particles, 7 residuals and 1 punctuation. Khoja’s tag set included the morphological features of gender, number, person, case, definiteness and mood. Figure 5.2 shows an example of a part-of-speech annotated sentence تنفيذاً لتوجيهات خادم الحرمين الشريفين *tanfīḏ*^{an} *li-tawjīhāt ẖādim al-ḥaramayn aš-šarīfayn* “Implementation of the directives of the Custodian of the Two Holy Mosques”, taken from the training corpus of the APT tagger (Khoja 2003).

Word			Khoja’s part-of-speech tag
تنفيذاً	*tanfīḏ*^{an}	Implementation	NCSgMI
لتوجيهات	*li-tawjīhāt*	directives	PPr’NCSgMI
خادم	*ẖādim*	Custodian	NCSgMI
الحرمين	*al-ḥaramayn*	Two Mosques	NCDuMD
الشريفين	*aš-šarīfayn*	Holy	NCDuMD

Figure 5.2 Example of tagged sentence using Khoja’s tag set

5.3.2 Penn Arabic Treebank (PATB) Part-of-Speech Tag Set

The most widely used tag set for Arabic is the Penn Arabic Treebank tag set used to annotate the Penn Arabic Treebank (PATB) with part-of-speech tags. Tim Buckwalter’s morphological analyser was used to compute a set of candidate solutions or analyses for each word, and then Arabic linguists selected the solution which best fitted the context. The Penn Arabic Treebank model postulates a FULL tag set which comprises over 2200 tag types (Diab 2007; Habash, Faraj and Roth 2009). This includes combinations of 114 basic tags listed in the Linguistic Data Consortium (LDC) Arabic part-of-speech/morphological tagging documentation[46] (Maamouri and Bies 2004; Maamouri et al. 2004; Habash 2010). Figure 5.3 shows these basic tags.

The FULL tag set exhibits a wider range of morphological features: case, gender, number, definiteness, mood, person, voice, tense and aspect. The LDC also introduced the reduced tag set (RTS) of 25 tags which is designed to maximize the performance of Arabic syntactic parsing. The RTS follows the tag set designed for the English Wall Street Journal. The morphological features marked by the RTS tag set are case, mood, gender, person and definiteness (Diab 2007).

```
ABBREV
ADJ
ADV
CONJ
DEM_PRON_F
DEM_PRON_FD
DEM_PRON_FS
DEM_PRON_MD
DEM_PRON_MP
DEM_PRON_MS
DET
EMPHATIC_PARTICLE
EXCEPT_PART
FUNC_WORD
FUT
INTERJ
INTERROG_PART
IV1P
IV1S
IV2D
IV2FS
IV2MP
IV2MS
IV3FD
IV3FP
IV3FS
IV3MD
IV3MP
IV3MS
IVSUFF_DO:1P
IVSUFF_DO:1S
IVSUFF_DO:2MP
IVSUFF_DO:2MS
IVSUFF_DO:3D
IVSUFF_DO:3FS
IVSUFF_DO:3MP
IVSUFF_DO:3MS
IVSUFF_SUBJ:2FS_MOOD:SJ
IVSUFF_SUBJ:D_MOOD:I
IVSUFF_SUBJ:D_MOOD:SJ
IVSUFF_SUBJ:FP
IVSUFF_SUBJ:MP_MOOD:I
IVSUFF_SUBJ:MP_MOOD:SJ
NEG_PART
NO_FUNC
NON_ALPHABETIC
NON_ARABIC
NOUN
NOUN_PROP
NSUFF_FEM_DU_ACCGEN
NSUFF_FEM_DU_ACCGEN_POSS
NSUFF_FEM_DU_NOM
NSUFF_FEM_DU_NOM_POSS
NSUFF_FEM_PL
NSUFF_FEM_SG
NSUFF_MASC_DU_ACCGEN
NSUFF_MASC_DU_ACCGEN_POSS
NSUFF_MASC_DU_NOM
NSUFF_MASC_DU_NOM_POSS
NSUFF_MASC_PL_ACCGEN
NSUFF_MASC_PL_ACCGEN_POSS
NSUFF_MASC_PL_NOM
NSUFF_MASC_PL_NOM_POSS
NSUFF_MASC_SG_ACC_INDEF
NUM
NUMERIC_COMMA
PART
POSS_PRON_1P
POSS_PRON_1S
POSS_PRON_2FS
POSS_PRON_2MP
POSS_PRON_2MS
POSS_PRON_3D
RESULT_CLAUSE_PARTICLE
POSS_PRON_3FP
POSS_PRON_3FS
POSS_PRON_3MP
POSS_PRON_3MS
PREP
PRON_1P
PRON_1S
PRON_2FS
PRON_2MP
PRON_2MS
PRON_3D
PRON_3FP
PRON_3FS
PRON_3MP
PRON_3MS
PUNC
PVSUFF_DO:1P
PVSUFF_DO:1S
PVSUFF_DO:3D
PVSUFF_DO:3FS
PVSUFF_DO:3MP
PVSUFF_DO:3MS
PVSUFF_SUBJ:1P
PVSUFF_SUBJ:1S
PVSUFF_SUBJ:2FS
PVSUFF_SUBJ:2MP
PVSUFF_SUBJ:3FD
PVSUFF_SUBJ:3FP
PVSUFF_SUBJ:3FS
PVSUFF_SUBJ:3MD
PVSUFF_SUBJ:3MP
PVSUFF_SUBJ:3MS
REL_PRON
REL_ADV
SUBJUNC
VERB_IMPERFECT
VERB_PERFECT
VERB PASSIVE
```

Figure 5.3 The Penn Arabic Treebank Tag Set; basic tags, which can be combined

[46] LDC Arabic POS tagging documentation http://www.ircs.upenn.edu/arabic/Jan03release/POS-info.txt

```
INPUT STRING: تم
LOOK-UP WORD: tm
     Comment:
* SOLUTION 1: (tam~) tam~/VERB_PERFECT
     (GLOSS):  + conclude/take place +
INPUT STRING: اعداد
LOOK-UP WORD: AEdAd
     Comment:
  SOLUTION 1: (>aEodAd) >aEodAd/NOUN
     (GLOSS):  + numbers/issues +
* SOLUTION 2: (<iEodAd) <iEodAd/NOUN
     (GLOSS):  + preparation +
INPUT STRING: الوثائق
LOOK-UP WORD: AlwvA}q
     Comment:
* SOLUTION 1: (AlwavA}iq) Al/DET+wavA}iq/NOUN
     (GLOSS): the + documents/charters +
INPUT STRING: المتوفرة
LOOK-UP WORD: Almtwfrp
     Comment:
* SOLUTION 1: (Almutawaf~irap) Al/DET+mutawaf~ir/ADJ+ap/NSUFF_FEM_SG
     (GLOSS): the + available/abundant + [fem.sg.]
INPUT STRING: ب
LOOK-UP WORD: b
     Comment: Separated
* SOLUTION 1: (bi-) bi-/PREP
     (GLOSS): by/with
INPUT STRING: كثرة
LOOK-UP WORD: kvrp
     Comment:
* SOLUTION 1: (-kavorap) -kavor/NOUN+ap/NSUFF_FEM_SG
     (GLOSS): abundance/frequency + [fem.sg.]
INPUT STRING: حول
LOOK-UP WORD: Hwl
     Comment:
* SOLUTION 1: (Hawola) Hawola/PREP
     (GLOSS):  + about/around +
  SOLUTION 2: (Haw~al) Haw~al/VERB_PERFECT
     (GLOSS):  + change/convert/switch +
  SOLUTION 3: (Hawol) Hawol/NOUN
     (GLOSS):  + power +
INPUT STRING: أول
LOOK-UP WORD: >wl
     Comment:
  SOLUTION 1: (>aw~al) >aw~al/VERB_PERFECT
     (GLOSS):  + explain/interpret +
* SOLUTION 2: (>aw~al) >aw~al/ADJ
     (GLOSS):  + first +
  SOLUTION 3: (>uwal) >uwal/ADJ
     (GLOSS):  + first +
INPUT STRING: رحلة
LOOK-UP WORD: rHlp
     Comment:
* SOLUTION 1: (riHolap) riHol/NOUN+ap/NSUFF_FEM_SG
     (GLOSS):  + journey/career + [fem.sg.]
INPUT STRING: طيران
LOOK-UP WORD: TyrAn
     Comment:
* SOLUTION 1: (TayarAn) TayarAn/NOUN
     (GLOSS):  + airline/aviation +
INPUT STRING: عثمانية
LOOK-UP WORD: EvmAnyp
     Comment:
  SOLUTION 1: (EuvomAniy~ap) EuvomAniy~/NOUN+ap/NSUFF_FEM_SG
     (GLOSS):  + Ottoman + [fem.sg.]
* SOLUTION 2: (EuvomAniy~ap) EuvomAniy~/ADJ+ap/NSUFF_FEM_SG
     (GLOSS):  + Ottoman + [fem.sg.]
INPUT STRING: فوق
LOOK-UP WORD: fwq
     Comment:
* SOLUTION 1: (fawoq) fawoq/PREP
     (GLOSS):  + above/over +
  SOLUTION 2: (fawoq) fawoq/NOUN
     (GLOSS):  + top/upper part +
INPUT STRING: البلاد
LOOK-UP WORD: AlblAd
     Comment:
* SOLUTION 1: (AlbilAd) Al/DET+bilAd/NOUN
     (GLOSS): the + (native) country/countries +
INPUT STRING: العربية
LOOK-UP WORD: AlErbyp
     Comment:
```

Figure 5.4 Buckwalter morphological analysis of a sentence from the Arabic Treebank

```
تم (tam~)                  tam~/VERB_PERFECT
اعداد (<iEodAd)            <iEodAd/NOUN
الوثائق (AlwavA}iq)        Al/DET+wavA}iq/NOUN
المتوفرة (Almutawaf~irap)   Al/DET+mutawaf~ir/ADJ+ap/NSUFF_FEM_SG
ب (bi-)                    bi-/PREP
كثرة (-kavorap)            -kavor/NOUN+ap/NSUFF_FEM_SG
حول (Hawola)               Hawola/PREP
أول (>aw~al)               >aw~al/ADJ
رحلة (riHolap)             riHol/NOUN+ap/NSUFF_FEM_SG
طيران (TayarAn)            TayarAn/NOUN
عثمانية (EuvomAniy~ap)      EuvomAniy~/ADJ+ap/NSUFF_FEM_SG
فوق (fawoq)                fawoq/PREP
البلاد (AlbilAd)           Al/DET+bilAd/NOUN
العربية (AlEarabiy~ap)      Al/DET+Earabiy~/ADJ+ap/NSUFF_FEM_SG
```

Figure 5.5 Disambiguated sentence from the Arabic Treebank using FULL tag set

```
INPUT STRING: وَوَصَّيْنَا
LOOK-UP WORD: wwSynA
* SOLUTION 1: (wawaS~ayonA) [waS~aY_1] wa/CONJ+waS~ay/VERB_PERFECT+nA/PVSUFF_SUBJ:1P
     (GLOSS): and + recommend/advise + we <verb>
  SOLUTION 2: (wawaSiy~nA) [waSiy~_1] wa/CONJ+waSiy~/NOUN+nA/POSS_PRON_1P
     (GLOSS): and + authorized agent/trustee + our

INPUT STRING: الإِنسَانَ
LOOK-UP WORD: Al<nsAn
* SOLUTION 1: (Al<inosAn) [<inosAn_1] Al/DET+<inosAn/NOUN
     (GLOSS): the + human being +

INPUT STRING: بِوَالِدَيْهِ
LOOK-UP WORD: bwAldyh
  SOLUTION 1: (biwAlidiy~h) [wAlidiy~_1] bi/PREP+wAlidiy~/ADJ+hu/POSS_PRON_3MS
     (GLOSS): by/with + parental + its/his
* SOLUTION 2: (biwAlidayohi) [wAlid_1]
               bi/PREP+wAlid/NOUN+ayo/NSUFF_MASC_DU_ACCGEN+hu/POSS_PRON_3MS
     (GLOSS): by/with + parents/father and mother + his/its two

INPUT STRING: حُسْنًا
LOOK-UP WORD: HsnA
  SOLUTION 1: (Hasun~A) [Hasun-u_1] Hasun/VERB_PERFECT+nA/PVSUFF_SUBJ:1P
     (GLOSS):  + be beautiful/be good + we <verb>
  SOLUTION 2: (HasunA) [Hasun-u_1] Hasun/VERB_PERFECT+A/PVSUFF_SUBJ:3MD
     (GLOSS):  + be beautiful/be good + they (both) <verb>
  SOLUTION 3: (Has~an~A) [Has~an_1] Has~an/VERB_PERFECT+nA/PVSUFF_SUBJ:1P
     (GLOSS):  + improve/decorate + we <verb>
  SOLUTION 4: (Has~anA) [Has~an_1] Has~an/VERB_PERFECT+A/PVSUFF_SUBJ:3MD
     (GLOSS):  + improve/decorate + they (both) <verb>
* SOLUTION 5: (HusonAF) [Huson_1] Huson/NOUN+AF/NSUFF_MASC_SG_ACC_INDEF
     (GLOSS):  + good/beauty + [acc.indef.]
  SOLUTION 6: (HasanAF) [Hasan_2] Hasan/NOUN+AF/NSUFF_MASC_SG_ACC_INDEF
     (GLOSS):  + good + [acc.indef.]
  SOLUTION 7: (HasanA) [Hasan_2] Hasan/NOUN+A/NSUFF_MASC_DU_NOM_POSS
     (GLOSS):  + good + two
  SOLUTION 8: (HasanAF) [Hasan_2] Hasan/ADV+AF/NSUFF_MASC_SG_ACC_INDEF
     (GLOSS):  + well + [acc.indef.]
  SOLUTION 9: (Has~anA) [Has~-i_1] Has~/VERB_PERFECT+a/PVSUFF_SUBJ:3MS+nA/PVSUFF_DO:1P
     (GLOSS):  + feel + he/it <verb> us
  SOLUTION 10: (Has~nA) [Has~_1] Has~/NOUN+nA/POSS_PRON_1P
     (GLOSS):  + perception/feeling + our
  SOLUTION 11: (His~nA) [His~_1] His~/NOUN+nA/POSS_PRON_1P
     (GLOSS):  + sensation/perception + our
```

Figure 5.6 Buckwalter morphological analysis of a sentence from the Quran

```
وَوَصَّيْنَا (wawaS~ayonA) wa/CONJ+waS~ay/VERB_PERFECT+nA/PVSUFF_SUBJ:1P
الإِنسَانَ (Al<inosAn)   Al/DET+<inosAn/NOUN
بِوَالِدَيْهِ(biwAlidayohi)bi/PREP
                        +wAlid/NOUN
                        +ayo/NSUFF_MASC_DU_ACCGEN+hu/POSS_PRON_3MS
حُسْنًا (HusonAF)         Huson/NOUN+AF/NSUFF_MASC_SG_ACC_INDEF
```

Figure 5.7 Disambiguated sentence from the Quran using FULL tag set

Figures 5.4-5.7 show examples of two sentences tagged by the FULL tag set. The first sentence is a newspaper text taken from the Arabic Treebank: تم اعداد الوثائق المتوفرة بكثرة حول أول رحلة طيران عثمانية فوق البلاد العربية *tamma 'i'dād al-waṯā'iqa al-mutawaffira*^{ti} *ḥawla 'awwali riḥla*^{ti} *ṭayyarān*^{in} *'uṯmāniyya*^{tin} *fawqa al-bilādi al-'arabiyya*^{ti} 'Many available documents relate to the first Ottoman's flight over the Arab countries'. The second sentence is taken from the Qur'an (chapter 29): وَوَصَّيْنَا الْإِنسَانَ بِوَالِدَيْهِ حُسْنًا *wa waṣṣaynā al-'insāna biwālidayhi ḥusn*^{an} 'We have enjoined on man kindness to parents'. Figures 5.4 and 5.6 show the full outputs of the Buckwalter morphological analyser including several possible solutions for some words; and Figures 5.5 and 5.7 show the correct disambiguated solution for each word in context.

Diab (2007) compared the FULL and RTS tag sets introduced by the LDC to PoS-tag the Arabic Treebank. The study is about designing the optimal part-of-speech tag set for Arabic. By analyzing the Arabic Treebank data, the RTS tag set is extended from 25 tags to 75 tags. Only morphological features, which are explicitly marked on the words, are added to the RTS. The new tag set is called the ERTS (extended reduced tag set). The ERTS has only the explicit or marked morphological features of gender, number and definiteness on nominals while maintaining the existing features from RTS. Figure 5.8 illustrates some differences between the three tag sets: FULL, RTS and ERTS from (Diab 2007).

Word			FULL	RTS	ERTS
حصيلة	HSylp	'result'	**NOUN+ NSUFF_FEM_SG+ CASE_IND_NOM**	**NN**	**NNF**
نهائية	nhA}yp	'final'	**ADJ+ NSUFF_FEM_SG+ CASE_IND_NOM**	**JJ**	**JJF**
حادث	HAdv	'accident'	**NOUN+ CASE_DEF_ACC**	**NN**	**NNM**
النار	AlnAr	'the-fire'	**DET+ NOUN+ CASE_DEF_GEN**	**NN**	**DNNM**
الجماعي	AlimAEy	'group'	**DET+ ADJ+ CASE_DEF_GEN**	**JJ**	**DJJM**
شخصين	$xSyn	'two-persons'	**NOUN+ NSUFF_MASC_DU_GEN**	**NN**	**NNMDu**

Figure 5.8 A sample of tagged sentence using the FULL, RTS and ERTS tag sets

5.3.3 ARBTAGS Tag Set

Alqrainy (2008) developed a new part-of-speech tag set called ARBTAGS to be used in the development of a part-of-speech tagger. The tag set design followed the criteria proposed by Atwell (2008). Like Khoja, Alqrainy built on traditional Arabic grammar books to design the tag set. Six morphological features of Arabic words were included: gender, number, case, mood, person and state. ARBTAGS contains 161 detailed tags and 28 general tags to cover the main part-of-speech classes and sub-classes. The 161 detailed tags are divided into 101 nouns, 50 verbs, 9 particles and 1 punctuation mark. Figure 5.9 shows the 28 general tags of the ARBTAGS tag set.

Tag	Description	Tag	Description
VePe	*Perfect verb*	**NuCd**	*Conditional noun*
VePi	*Imperfect verb*	**NuDe**	*Demonstrative noun*
VePm	*Imperative verb*	**NuIn**	*Interrogrative noun*
NuPo	*Proper noun*	**NuAd**	*Adverb*
NuCn	*Common noun*	**NuNn**	*Numeral noun*
NuAj	*Adjective noun*	**Fw**	*Foreign noun*
NuIf	*Infinitive noun*	**Pun**	*Punctuation mark*
NuRe	*Relative noun*	**PrPp**	*Preposition*
NuDm	*Diminutive noun*	**PrVo**	*Vocative Particle*
NuIs	*Instrument noun*	**PrCo**	*Conjunction Particle*
NuPn	*Noun of Place*	**PrEx**	*Exception Particle*
NuTn	*Noun of Time*	**PrAn**	*Annulment Particle*
NuPs	*Pronoun*	**PrSb**	*Subjunctive Particle*
NuCv	*Conjunctive noun*	**PrJs**	*Jussive Particle*

Figure 5.9 The 28 general tags of the ARBTAGS tag set

5.3.4 MorphoChallenge 2009 Qur'an Gold Standard Part-of-Speech Tag Set

MorphoChallenge2009[47] Qur'an gold standard was developed using the data of Morphological Tagging of the Qur'an database (Talmon and Wintner 2003; Dror et al. 2004). It was developed to be used to evaluate morphological analyzers in the Morphochallenge 2009 competition (Kurimo et al. 2009), which aimed to develop an unsupervised morphological analyzer to be used for different languages including Arabic. It contains the full morphological analysis for each word, according to the Tagged database of the Qur'an but reformatted to match other Morphochallenge test sets in other languages. The word's morphological analysis is shown after each word where the morphological features are separated by space and "+" sign. These features include the part-of-speech of the word, number, gender, person, case, definiteness, voice and others. Figure 5.10 shows a sample of the Qur'an gold standard.

This tag set was called a "gold standard" for the purpose of the MorphoChallenge 2009 contest, as it was the "target" or "solution" which the competitor system had to try to produce. The tagged text in other languages (*i.e.* English, German, French, Finish and Turkish) were also "gold standards" for the purposes of the MorphoChallenge contest. The term "gold standard" does not imply the tag set is better than others reviewed in the chapter.

[47] MorphoChallenge 2009 Qur'an Gold Standard http://www.cis.hut.fi/morphochallenge2009/datasets.shtml

```
وَوَصَّيْنَا          وصي  يُفَعِّلُ   وَ +Particle +Conjunction وَصَّينَا +Verb +Perf
                              +Act +1P +Pl +Masc/Fem
الإِنْسَانَ          ءنس  فِعلَان  ءِنسَان +Noun +Triptotic +Sg +Masc +Acc +Def
بِوَالِدَيْهِ          ولد  فَاعِل   ب +Prep وَالِد +Noun +Triptotic +Dual +Masc
                              +Obliquus +Pron +Dependent +3P +Sg +Masc
حُسْنًا             حسن   فُعل   حُسن +Noun +Triptotic +Sg +Masc +Acc +Tanwiin
```

```
wawaS~ayonaA wSy yufaE~ilu wa +Particle +Conjunction
                              waSSaynaA +Verb +Perf +Act +1P +Pl +Masc/Fem
Alo<insaAna 'ns fiElaAn 'insaAn +Noun +Triptotic +Sg +Masc +Acc +Def
biwaAlidayohi wld faAEil b +Prep waAlid +Noun +Triptotic +Dual +Masc
                                +Obliquus +Pron +Dependent +3P +Sg +Masc
HusonFA      Hsn fuEl    Husn +Noun +Triptotic +Sg +Masc +Acc +Tanwiin
```

Figure 5.10 Sample of tagged text taken from the MorphoChallenge 2009 Qur'an Gold Standard. The first part uses Arabic script and the second one uses romanized letters using Tim Buckwalter transliteration scheme.

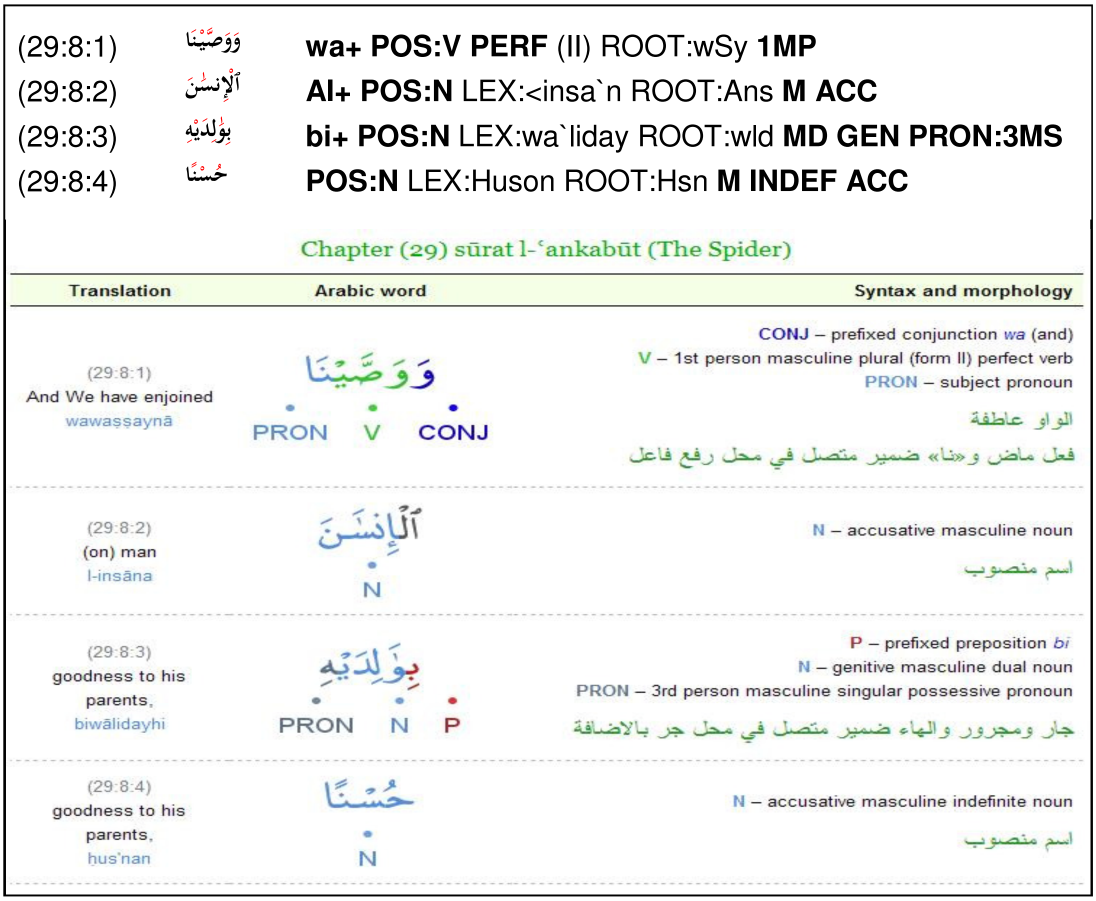
5.3.5 The Quranic Arabic Corpus Part-of-Speech Tag Set

The Quranic Arabic Corpus is a newly available resource enriched with multiple layers of annotation including morphological segmentation and part-of-speech tagging. The motivation behind this work is to produce a resource that enables further analysis of the Qur'an; a genre difficult to compare with other forms of Arabic, since the vocabulary and the spelling differs from modern standard Arabic (Dukes and Habash 2010).

Buckwalter's Arabic Morphological Analyzer (BAMA) was used to generate the initial tagging. The analyzer was adapted to work with the Quranic Arabic text. After that, the annotated corpus was then put online to allow for collaborative annotation (Dukes and Habash 2010), (Dukes et al., 2011).

A mapping was required to convert from the BAMA tag set to the Quranic Arabic Corpus tag set. Manual disambiguation was required for a few cases, where one-to-one mapping was not applicable such as particles. In order to adapt BAMA to process the Quranic Arabic Corpus text three modifications were made. First, spelling in the Qur'an differs from MSA. The differences involve orthographic variations of *hamza*^{h}, *'alif* and the long vowel *ā*. Second, the multiple diacritized analyses produced by BAMA for the processed words were ranked in terms of their edit-distance from the Qur'anic diacritization, with closer match ranked higher. Finally, filtering was done by choosing the highest rank analysis's part-of-speech as a solution (Dukes and Habash 2010).

The Quranic Arabic Corpus tag set adapts historical traditional Arabic grammar which leads to morphological annotation that uses terminology familiar to many readers of the Qur'an. This terminology enables people with Qur'anic syntax experience to participate in the online annotation to be verified against existing authenticated books on Quranic Grammar (Dukes and Habash 2010). Figure 5.11 shows a sample of the morphological and part-of-speech tags of the Quranic Arabic Corpus.



(29:8:1)	وَوَصَّيْنَا	**wa+ POS:V PERF** (II) ROOT:wSy **1MP**
(29:8:2)	ٱلْإِنسَٰنَ	**Al+ POS:N** LEX:<insa`n ROOT:Ans **M ACC**
(29:8:3)	بِوَٰلِدَيْهِ	**bi+ POS:N** LEX:wa`liday ROOT:wld **MD GEN PRON:3MS**
(29:8:4)	حُسْنًا	**POS:N** LEX:Huson ROOT:Hsn **M INDEF ACC**

Chapter (29) sūrat l-ʿankabūt (The Spider)

Translation	Arabic word	Syntax and morphology
(29:8:1) And We have enjoined wawaṣṣaynā	وَوَصَّيْنَا PRON V CONJ	CONJ – prefixed conjunction *wa* (and) V – 1st person masculine plural (form II) perfect verb PRON – subject pronoun الواو عاطفة فعل ماض و«نا» ضمير متصل في محل رفع فاعل
(29:8:2) (on) man l-insāna	ٱلْإِنسَٰنَ N	N – accusative masculine noun اسم منصوب
(29:8:3) goodness to his parents, biwālidayhi	بِوَٰلِدَيْهِ PRON N P	P – prefixed preposition *bi* N – genitive masculine dual noun PRON – 3rd person masculine singular possessive pronoun جار ومجرور والهاء ضمير متصل في محل جر بالاضافة
(29:8:4) goodness to his parents, hus'nan	حُسْنًا N	N – accusative masculine indefinite noun اسم منصوب

Figure 5.11 A sample of a tagged sentence taken from the Quranic Arabic Corpus

5.3.6 Columbia Arabic Treebank CATiB Part-of-Speech Tag Set

Another tag set was designed for the part-of-speech and syntactic annotation in the Columbia Arabic Treebank CATiB. A part-of-speech tag set consisting of only six tags is used for the part-of-speech annotation of CATiB. The main reason for using such a small tag set is a tradeoff between linguistic richness and Treebank size. The researchers' assumption for morpho-syntactically rich languages such as Arabic, is that the cost of fine grain annotation is a slower annotation process, a smaller Treebank and less data to train tools. CATiB is inspired by two ideas. First, it avoids annotation of redundant linguistic information. Second, it uses linguistic representation and terminology from traditional Arabic syntactic studies (Habash et al. 2009). The tag set is much smaller than the FULL tag set used by the Penn Arabic Treebank:

> *"... CATiB uses the same tokenization scheme used by PATB and PADT. However, unlike these resources, the CATiB POS tag set is much smaller. Whereas PATB uses 2,200 tags specifying every aspect of Arabic word morphology such as definiteness, gender, number, person, mood, voice and case; CATiB uses six POS tags: NOM (nominals such as nouns, pronouns, adjectives and adverbs), PROP (proper noun), VRB (verb), VRB-PASS (passive verb), PRT (particles such as prepositions or conjunctions) and PNX (punctuation). ..." (Habash and Roth 2009)*

Figure 5.12 shows an example of the sentence, خمسون ألف سائح زاروا لبنان وسوريا في أيلول الماضي *ẖamsūn 'alf sā'iḥ zārū lubnān wa sūriyyā fī 'aylūl al-māḍī* "50 thousand tourists visited Lebanon and Syria last September", tagged using part-of-speech tags used in the Columbia Arabic Treebank CATiB.

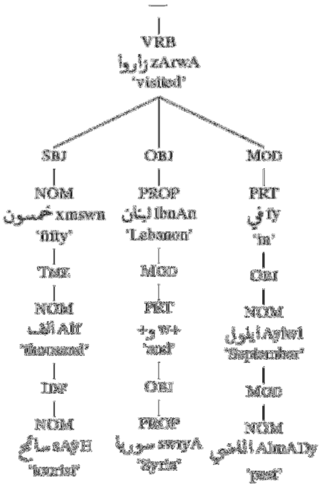
WORD			**CATIB PART-OF-SPEECH TAG**	**CATIB ANNOTATION**
خمسون	*ẖamsūn*	Fifty	**NOM**	 VRB زاروا zArwA 'visited' SBJ OBJ MOD NOM خمسون xmswn 'fifty' TMZ NOM ألف Alf 'thousand' IDF NOM سائح sA}H 'tourist' PROP لبنان lbnAn 'Lebanon' MOD PRT +و w+ 'and' OBJ PROP سوريا swryA 'Syria' PRT في fy 'in' OBJ NOM أيلول Aylwl 'September' MOD NOM الماضي AlmADy 'past'
ألف	*'alf*	Thousand	**NOM**	
سائح	*sā'iḥ*	Tourist	**NOM**	
زاروا	*zārū*	Visited	**VRB**	
لبنان	*lubnān*	Lebanon	**PROP**	
و	*wa*	And	**PRT**	
سوريا	*sūriyyā*	Syria	**PROP**	
في	*fī*	In	**PRT**	
أيلول	*'aylūl*	September	**NOM**	
الماضي	*al-māḍī*	Past	**NOM**	

Figure 5.12 Example of part-of-speech tagged sentence using CATiB tag set

5.3.7 Comparison of Arabic Part-of-Speech Tag Sets

Table 5.1 shows a comparison of the eight Arabic tag sets studied in this section. The comparison summarizes the characteristics of each tag set and helps to show the differences between them clearly. The drawbacks of the existing tag sets for Arabic were found to be:

- Existing Arabic tag sets vary in size from 6 tags to 2000 or more tags.
- Some of these tag sets follow standards for tag set design for English such as the PATB tag sets, and these may not always be appropriate for Arabic.
- The tag sets share common morphological features such as gender, number, person, case, mood and definiteness, but the attributes of the morphological feature categories are not standardized.

- These tag sets lack standardization in defining a suitable scheme for tokenizing Arabic words into their morphemes and they mix morpheme tagging with whole word tagging.
- They also lack suitable documentation that illustrates the decision made for each design dimension of the tag set.
- The tags assigned to words in a corpus are not consistent in either presentation of the tag itself or the morphological features which are encoded within the tag.

Moreover, the most widely used and important morphosyntactic annotation standards and guidelines, namely EAGLES, are designed for Indo-European languages. These guidelines are not entirely suitable for Arabic.

These drawbacks of existing tag sets are the motivation behind desining the SALMA (Sawalha Atwell Leeds Morphological Analysis) Tag Set for Arabic.

The comparison of the morphological features used in the different tag sets of Arabic shows shared common features such as gender, number, person, case, mood and definiteness. Features such as voice, tense and aspect are included in the PATB FULL tag set. State is included in the ARBTAGS tag set. Diptotic is a feature of the MorphoChallenge 2009 tag set, and verb form and derivation are features of the QAC tag set. Chapter 6 discusses the 22 morphological features of the SALMA Tag Set.

Table 5.1 Comparison of Arabic part-of-speech tag sets

1. Khoja's Tag set	
Purpose of design	Compiling a tag set as a standard tag set
Main characteristics	Based on traditional Arabic grammar rather than being based on an Indo-European one. Only the main classes and subclasses have been chosen.
Tag set size	177 tags (103 types of noun, 57 verbs, 9 particles, 7 residuals,1 punctuation)
Morphological features	Gender, Number, Case, Definiteness , Person, Mood
Applications	Used in the design of the APT tagger, and in the annotation of the training data of the APT tagger.
2. Penn Arabic Treebank (PATB) Part-of-Speech Tag Set (FULL)	
Purpose of design	Annotating the Arabic Treebank with part-of-speech tags
Main characteristics	Aims to cover detailed grammar features.
Tag set size	The FULL tag set comprises over 2000 tag types. This includes combinations of 114 basic tags.
Morphological features	Case, Gender, Number, Definiteness, Mood, Person, Voice, Tense, Aspect
Applications	Used in Tim Buckwalter's morphological analyser to annotate the Penn Arabic Treebank with part-of-speech tags.

3. Penn Arabic Treebank (PATB) Reduced Part-of-Speech Tag Set (RTS)	
Purpose of design	Maximizing the performance of Arabic syntactic parsing.
Main characteristics	Follows the tag set designed for the English Wall Street Journal.
Tag set size	25 tags
Morphological features	Case, Mood, Gender, Person, Definiteness
Applications	Used in the syntactic annotation of the Penn Arabic Treebank
4. Penn Arabic Treebank (PATB) Extended Reduced Part-of-Speech Tag Set (ERTS)	
Purpose of design	To be used for higher order processing of the language
Main characteristics	Is an extension of the RTS tag set which has only the explicit or marked morphological features of gender, number and definiteness on nominals.
Tag set size	75 tags
Morphological features	Gender, Number, Definiteness on nominals
Applications	To be used for parsing
5. ARBTAGS	
Purpose of design	Standardizing and building a comprehensive Arabic tag set.
Main characteristics	The tag set hierarchy follows the tradition of Arabic grammar.
Tag set size	161 detailed tags (101 nouns, 50 verbs, 9 particles, 1 punctuation mark including 28 different POS general tags to cover the main part-of-speech classes and sub-classes.
Morphological features	Gender, Number, Case, Mood, Person, State
Applications	Used in the Arabic Morphosyntactic Tagger AMT
6. MorphoChallenge 2009 Qur'an gold standard tag set	
Purpose of design	To annotate the Qur'an as a gold standard to be used to evaluate morphological analyzers in the MorphoChallenge 2009 competition.
Main characteristics	It was developed using the data for Morphological Tagging of the Qur'an database.
Tag set size	The tag set is combinations of the POS main and sub classes and the morphological features of the analysed words.
Morphological features	Gender, Number, Person, Case, Mood, Aspect, Voice, Definiteness, Diptotic
Applications	Used to construct the Qur'an gold standard for evaluating morphological analyzers in the MorphoChallenge 2009 competition.
7. Quranic Arabic Corpus POS tag set	
Purpose of design	To Annotate the Qur'an with morphological and part-of-speech tagging information.
Main characteristics	Used Tim Buckwalter's morphological analyzer as initial tagging, then a mapping from Buckwalter's tag set to the Quranic Arabic Corpus tag set. It adapts traditional Arabic grammar.
Tag set size	The tag set involves combinations of the POS main and sub classes and the morphological features of the analysed words.

Morphological features	Person, Gender, Number, Aspect, Mood, Voice, Verb form, Derivation, State
Applications	Used in the morphological and part-of-speech annotation of the Quranic Arabic Corpus
8. Columbia Arabic Treebank POS tag set	
Purpose of design	To be used for the part-of-speech annotation of Columbia Arabic Treebank CATiB.
Main characteristics	CATiB avoids the annotation of redundant linguistic information that is determinable automatically from syntax and morphological analysis, e.g., nominal case. CATiB uses linguistic representation and terminology inspired by the long tradition of Arabic syntactic studies.
Tag set size	6 part-of-speech tags (**VRB** – all verbs, **VRB-PASS** – passive-voice verbs, **NOM** – all nominals, **PROP** – proper nouns, **PRT** – particles, **PNX** – all punctuation marks)
Morphological features	No morphological features are encoded in the part-of-speech tag set of Columbia Arabic Treebank CATiB
Applications	Used in the part-of-speech annotation of Columbia Arabic Treebank CATiB.

5.4 Morphological Features in Tag Set Design Criteria

EAGLES[48] (Leech and Wilson 1996) proposed recommendations (guidelines) for morphosyntactic categories for European languages. The aim of the EAGLES guidelines is to propose standards in developing tag sets for morphosyntactic tagging, in the interest of comparability, interchangeability and reusability of annotated corpora. In addition to preferred standards, EAGLES guidelines also cater for extensibility, allowing specifications to extend to language-specific phenomena. The guidelines proposed standardisation in three important areas:

- Representation/Encoding: transparency, processability, brevity and unambiguity.
- Identifying categories/ subcategories/ structure: agreement on common categories and allowance for variation: obligatory, recommended and optional specification.
- Annotation schemes and their application to text: detailed annotation schemes should be made available to end-users and to annotators.

EAGLES recognizes four degrees of constraint in the description of word categories for morphosyntactic tags. First, *obligatory*; attributes have to be included in any morphosyntactic tag set: main categories of part-of-speech Noun, Verb, Adjective,

[48] EAGLES Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES document EAG-TCWG-MAC/R. http://www.ilc.cnr.it/EAGLES96/pub/eagles/corpora/annotate.ps.gz

Pronoun/Determiner, Article, Adverb, Adposition, Conjunction, Interjection, Unique/Unassigned, Residual, Punctuation. Second, *recommended*: attributes and values of widely-recognized grammatical categories which occur in conventional grammatical description (*e.g.* Gender, Number, Person). Third, *generic special extensions*: attributes and values which are not usually encoded, but might be included for particular purposes, for example semantic classes such as temporal nouns, manner adverbs, place names, etc. Finally, *language-specific special extensions*: additional attributes or values which may be important for a particular language.

Khoja et al (2001) compared their Arabic tag set against the EAGLES guidelines. The comparison showed: first, EAGLES tag set guidelines are based on Latin as a common ancestor, while Arabic has some novel features not found in Latin, for example certain categories and subcategories that inherit properties from the parent categories. Second, a Classical Arabic tag set has three main categories (nouns, verbs and particles), while EAGLES has eleven major part-of-speech categories. Third, apart from nouns and verbs, other major categories in EAGLES such as pronouns, numerals and adjectives are described as subcategories of major categories in a classical Arabic tag set. Fourth, Arabic, not only has singular and plural numbers, but it also has dual number. Moreover, Arabic verbs are classified as being perfect, imperfect and imperative, which differs from EAGLES classification of past, present and future tenses. Finally, the mood morphological feature is not covered by the EAGLES guidelines.

Atwell (2008) proposed criteria for tag set development, and stated that there are dimensions (choices) to be made by developers of a new part-of-speech tag set. Developers must decide on the set of grammatical tags or categories, and their definitions and boundaries. These criteria were applied to Arabic when the ARBTAGS tag set (Alqrainy 2008) was designed. We followed the same criteria as Atwell (2008) in designing the general-purpose morphological features tag set. Sections 5.4.1 - 5.4.12 explain the criteria and how they are applied in the SALMA – Tag set.

5.4.1 Mnemonic Tag Names

Generally, tag names for English PoS tag sets are chosen to help linguists to remember the grammatical categories such as CC for *Coordinating Conjunction* and VB for *VerB*. The SALMA Tag Set for Arabic has to encode much richer morphology: the tag is represented by a string of 22 characters. Each character represents a value or attribute which belongs to a morphological feature category. The position of the character in the tag string is important as it identifies the morphological feature category. The value of the feature is represented by one lowercase character, which is intended to remain readable, such as: **v** in the first position to indicate *verb*, **n** in the second position to indicate *name*, gender category values in the seventh position where *masculine* is represented by **m**, *feminine* is represented by **f** and *common gender* is represented by **x**. If the value of a certain feature is not applicable for the tagged word then dash "**-**" is used to indicate this. A question mark "**?**" indicates "unknown": a certain feature normally belongs to the word but at the moment is not available or the automatic tagger could not guess it.

The interpretation of the tag is handled by referring to the attribute value and its position in the tag string. The position of the attribute in the tag string identifies the morphological feature category, while the attribute value is identified by searching the morphological feature category for the specified symbol. Then, all these single interpretations of attributes are grouped together to represent the full tag of the word. The tag is still readable by linguists. Moreover, the tag is straightforwardly readable by software, for example by a search tool matching specified feature-value(s).

5.4.2 Underlying Linguistic Theory

Linguists who develop new tag sets will inevitably be swayed by the linguistic theories they espouse. In the case of English, there is disagreement between grammar theories on the range of grammatical categories and features to be tagged, and more complicated structural issues. It is difficult to have theory-neutral annotation, because every tagging scheme makes some theoretical assumptions (Atwell 2008).

Khoja's mophosyntactic tag set was derived from classical Arabic grammar (Khoja et al. 2001; Khoja 2003). ARBTAGS also tried to follow the Arabic grammatical system, which is based upon main three part-of-speech classes: verbs, nouns and particles, and enriched with inflectional features (Alqrainy 2008). The Arabic Penn Treebank tag set follows the same criteria used to develop the English Treebank (Maamouri and Bies 2004). ERTS (extended reduced tag set) extends the LDC reduced tag set (RTS) by adding morphological features namely (case, mood, definiteness, gender, number and person). This extends the 25 RTS tag set to 75 tag set of ERTS (Diab 2007).

The proposed SALMA – Tag Set adds more fine-grained details to the existing tag sets. The tag set follows traditional Arabic grammar theory (Dahdah 1987; Dahdah 1993; Wright 1996; Al-Ghalayyni 2005; Ryding 2005) in specifying 22 morphological features categories and their attributes or values. Section 6.2.1 justifies of the SALMA Tags in terms of this underlying theory.

5.4.3 Classification by Form or Function

For English an ambiguous word like '*open*' is tagged according to its function, and only its inflected forms are tagged by their form. Arabic words are highly inflected and hence word classification tends to be dependent on form. Classification by form is dependent on the word, while classification by function is dependent on the function of the word in context. For Arabic, the word class is heavily constrained by form, but if there is only one analysis, then it is determined by function. If there are two analyses, one needs to take context into account which means it is partially determined by function. In this case the function has to be taken into account for classification.

Arabic word-class is dependent on form. Traditional Arabic grammar groups words according to their inflexional behaviour. A challenging characteristic of Arabic is the treatment of short vowels, which are normally omitted in written Arabic. These short vowels can help in specifying some morphological feature information of grammatical categories. The Qur'an is fully vowelized to ensure it is pronounced correctly. This makes the Qur'an a potential "Gold Standard" corpus for Arabic tagging and NLP research (Atwell 2008).

Another challenge of Arabic words can appear when classifying words according to certain morphological feature such as gender. Classifying nouns into masculine or feminine can be viewed from two perspectives. First, according to the word's structure or morphologically; masculine nouns are not normally marked by any suffix, while feminine nouns have a suffix normally *–a*^{h} - added at the end of the noun. Second, semantically; nouns are arbitrarily classified into masculine or feminine, except when a noun refers to a human being or other creature having natural gender (sex), when it is normally conforms to natural gender (Ryding 2005). Therefore, a noun can have feminine suffix *–a*^{h}; which is classified as morphologically feminine, but it indicates a male such as حَمْزَة *ḥamza*^{h} 'Hamza (male proper name)', or vice versa, such as مَرْيَم *maryam* 'Mary (female proper name).

5.4.4 Idiosyncratic Words

Arabic has some words with special, idiosyncratic behaviour, such as particles which cannot be analyzed morphologically according to root and pattern. (Khoja et al. 2001) includes examples of this type in an "Exception" category, which covers group of particles that are equivalent to the English word "except" and the prefixes *non-*, *un-* , and *im-*.

5.4.5 Categorization Problems

A detailed categorisation scheme requires each tag to be defined clearly and unambiguously, by giving examples in a "case-law" document. This definition should include how to decide difficult, borderline cases, so that all examples in the corpus can be tagged consistently. Many words can belong to more than one grammatical category, depending on context of use. Tagging schemes should specify how to choose one tag as appropriate, if a word can have different part-of-speech tags in different contexts (Atwell 2008).

Vowelized Arabic text has less ambiguity than non-vowelized Arabic text. Short vowels and some affixes add linguistic information which reduces the ambiguity. In the SALMA Tag Set, each feature category is described, clearly documented and examples are provided. Moreover, tagging guidelines define the appropriate attribute for the morphological feature category.

5.4.6 Tokenisation: What Counts as a Word?

Arabic text tokenisation is not an easy task. Simple tokenisation of text can be carried out by dividing text into words by spaces, or punctuation. This tokenisation process is primitive and the first step in tokenising Arabic text. The majority of Arabic words are complex words; one or more clitics can be attached to the beginning and the end of the word [clitic(s) + word + clitic(s)]. These clitics are particles, pronouns or definite article. A tag is provided for each clitic attached to a word along with the tag of the word. For instance, the word وَبِحَسَنَاتِهِم *wabiḥasanātihim* ‘and with their good deeds’, consists of four parts, the conjunction letter و *wa* ‘and’, the preposition بِ *bi* ‘with’ the word حَسَنَاتِ *ḥasanāti* ‘good deeds’ and the pronoun هم *him* ‘ their’. The tag of this word will be the tags of the four morphemes and the whole word tag which is a combination of the morphemes tags. The clitics will help the tagging scheme in identifying some of the morphological features attributes; preposition بِ *bi* governs the genitive case of the noun.

5.4.7 Multi-Word Lexical Items

Multi-words lexical items are rare in Arabic (Alqrainy 2008). Such items might consist of two words; noun followed by adjective describing the proceeding noun, some compound proper names such as عَبْدُ الله *’abdu allāh* ‘Abdullah’, or compound particles such as فِيْمَا *fīmā* which consists of the preposition فِي *fī* and the non-human relative noun مَا *mā*. In the case of proper names; a single tag might be more appropriated. While, for the other cases a separate tags for each part of the lexical item will give more morphological detail about the multi-word lexical items.

The Penn Arabic Treebank guidelines ignore multi-word lexical items and tag each word of a compound word separately:

> *“....Divided/compound proper names in Arabic (Abdul Ahmed, e.g.): Label all parts of the name with the "Is a name" button.*
>
> *Idioms: (for example, in what in them = 'included'): Label each word independently for its own part of speech (ignore the idiomatic meaning)....”*[49]

[49] Penn Arabic Treebank annotation guidelines http://www.ircs.upenn.edu/arabic/pos.html

5.4.8 Target Users and/or Applications

Fitness for purpose and customer satisfaction are the most important practical criteria for a new tag set. One common use of part-of-speech tagged corpora is language teaching and research. A detailed tag set is required in teaching and learning to reflect fine distinctions of grammar, even though Machine Learning systems could cope better with a smaller tag set. General-purpose tag set developers should be more aware of potential re-use: detailed and more sophisticated part-of-speech tag schemes allow wider re-use of the corpus in future research (Atwell 2008).

The SALMA Tag Set is a general-purpose tag set. It encodes detailed information of morphological features embedded in any word. This morphological features information enables the tag set to be widely re-used.

5.4.9 Availability and/or Adaptability of Tagger Software

If a part-of-speech tag set is implemented in automatic tagger software, this has a clear advantage over a purely theoretical tag set (Atwell 2008). HMM taggers can be re-used for any language including Arabic. Experiments on highly inflectional languages such as German and Czech using an HMM tagger with a fine-grain tag set achieved higher tagging accuracy than two state-of-the-art general purpose part-of-speech taggers, The TnT tagger and SVMTool (Schmid and Laws 2008). Another experiment that uses a fine-grain tag set was done for Latin. Latin words require morphological analysis of nine features: part-of-speech, person, number, tense, mood, voice, gender, case and degree. The experiment used the TreeTagger analyzer which achieved an accuracy of 83% in correctly disambiguating the full morphological analysis (Bamman and Crane 2008).

5.4.10 Adherence to Standards

The EAGLES guidelines are designed for European languages. However, the Arabic language is different from Indo-European languages and has its own structure and morphological features. Instead, the standard adhered to in the SALMA Tag Set is that of traditional Arabic grammar books e.g. (Dahdah 1987; Dahdah 1993; Wright 1996; Al-Ghalayyni 2005; Ryding 2005).

5.4.11 Genre, Register or Type of Language

The SALMA Tag Set is intended to be general-purpose and to be used in part-of-speech tagging of different text types, formats and genres, of both vowelized and non-vowelized text. The tagging schemes and the tag set can be evaluated on a variety of text types, formats and genres. Corpora can include text in Classical Arabic such as; Qur’an, Classical Arabic dictionaries and poems from ancient Arabic literature, as well as Modern Standard Arabic text from newspapers, magazines, web pages, blogs, children’s books, and school text books, etc.

5.4.12 Degree of Delicacy of the Tag Set

The total number of tags is an indicator of the level of fine-grainedness of analysis. Existing Arabic corpus tag sets have degree of delicacy ranging from 6 for CATiB, 25 for the RTS tag set of the Penn Arabic Treebank, 75 tags for ERTS, 161 tags for ARABTAGS, 177 tags for Khoja's tag set, 2200 for PATB FULL tag set, and unspecified number of function combinations for QAC and MorphoChallenge 2009 tag sets. The SALMA Tag Set is a fine-grain tag set. It is unfeasible to enumerate all possible tags that can be generated from valid combinations of the 22 morphological feature categories; however, we can count the attributes of each feature category, and use these to estimate an upper bound or limit on the degree of delicacy of the SALMA Tag Set. Chapter 6 discusses the 22 morphological features of the SALMA – Tag Set and their attributes.

An upper limit of possible feature combinations is 4.07E+16, the total number of possible combinations of features in the SALMA Tag Set of Arabic, calculated by multiplying together the number of attributes of each of the 22 morphological features. But, of course, this includes many invalid tags that will never be used. A more realistic upper bound is given by counting the possible feature combinations for each major part of speech, and summing these. Table 2 shows the absolute upper limit of possible feature combinations for each major part of speech (Noun, Verb, Particle, Other (Residual), Punctuation); this gives an upper limit of 101,945,168 possible morphological feature combinations: about one hundred million possible SALMA tags.

Table 5.2 The upper limit of possible combinations of SALMA features

Feature		Number of attributes	Part of speech									
			Noun		Verb		Particle		Other		Punctuation	
			Template	Combinations	Template	Combinations	Template	Combinations	Template	Combinations	Template	Combinations
1	Main Part-of-Speech	5	n	1	v	1	p	1	r	1	u	1
2	Part-of-Speech: Noun	34	?	34	-	1	-	1	-	1	-	1
3	Part-of-Speech: Verb	3	-	1	?	3	-	1	-	1	-	1
4	Part-of-Speech: Particle	22	-	1	-	1	?	22	-	1	-	1
5	Part-of-Speech: Other	15	-	1	-	1	-	1	?	15	-	1
6	Punctuation marks	12	-	1	-	1	-	1	-	1	?	12
7	Gender	3	?	3	-	1	-	1	?	3	-	1
8	Number	9	?	9	-	1	-	1	?	3	-	1
9	Person	3	-	1	?	3	-	1	?	3	-	1
10	Inflectional morphology	4	?	3	?	2	?	1	?	1	-	1
11	Case or Mood	4	?	3	?	3	-	1	-	1	-	1
12	Case and Mood marks	10	?	7	?	6	?	4	?	4	-	1
13	Definiteness	2	?	2	-	1	-	1	-	1	-	1
14	Voice	2	-	1	?	2	-	1	-	1	-	1
15	Emphasized and non-emphasized	2	-	1	?	2	-	1	-	1	-	1
16	Transitivity	4	-	1	?	4	-	1	-	1	-	1
17	Rational	2	?	2	?	2	?	2	-	1	-	1
18	Declension and Conjugation	9	?	4	?	6	?	1	-	1	-	1
19	Unaugmented and Augmented	5	?	5	?	5	-	1	-	1	-	1
20	Number of root letters	3	?	3	?	2	-	1	-	1	-	1
21	Verb root	30	-	1	?	30	-	1	-	1	-	1
22	Nouns finals	6	?	6	-	1	-	1	-	1	-	1
Totals		**4.1E+16**	**83,280,960**		**18,662,400**		**176**		**1620**		**12**	
Upper limit of possible morphological feature combinations											**101,945,168**	

5.5 Complex Morphology of Arabic

Most Arabic words are derived from their roots following certain templates called patterns. The derivation process adds prefixes, suffixes and infixes to the root letters to generate a new word, which has a new function or meaning but preserves the main concept or meaning carried by the root. Moreover, using the derived word in a certain context will require clitics to be added to the beginning and the end of the word. Proclitics include prepositions, conjuctions and definite articles, and enclitics include relative pronouns. In addition, one or more affixes or clitics can be added to the derived word. In conclusion, most Arabic words are complex words consisting of multiple morphemes.

To specify a word's morphemes, tokenization is needed to analyse the word morphemes as clitics, affixes or stem. For example the tokenizer will specify the morphemes of the word وسيكتبونها *wasayaktubūnahā* 'and they will write it' as follows: preclitic و *wa* 'and' (conjunction), prefixes س *sa* 'will' and ي *ya* (imperfect prefix), the stem كتب *kataba* 'write', the suffix ون *ūn* 'they' and the enclitic ها *hā* 'it' (object suffixed pronoun). The word consists of 6 morphemes. Each morpheme carries morphological features and belongs to a specific part of speech category. The SALMA Tag Set assigns a tag to each morpheme of the word. Then in principle, the morphemes' tags are combined into one whole word tag. The word tag inherits its morphological feature attributes using an algorithm that establish agreements on morphological feature attributes. The description of the algorithm is beyond the scope of this chapter. This chapter is about the output of the tagger rather than describing the algorithm of tagging and combining morpheme tags into word tags. The following example in figure 5.13 shows the tokenization of the word into morphemes, the assignment of the part of speech tag for each morpheme and the result of combining the morpheme tags into one whole word tag.

Tokenization is a known problem even for English corpus tagging. The tagged LOB corpus defines the word or graphic word as a sequence of characters surrounded by spaces (or punctuation marks). Each word is assigned a tag. Differences in tagging occurred due to: first, variation in segmentation of compound terms, as in: *fancy free* given the tags **NN** (noun, singular, common) **JJ** (adjective), and *fancy-free* given the tag **JJ** (adjective). Second, hyphenated sequences, as in: *an above-the-rooftops position* given the tag **JJB** (adjective, attributive-only). Third, syntactic boundaries, as in: *Henry* **NP** (noun, singular, proper) *8's* **CD$** (numeral, cardinal, genitive) *hall*. In some cases, the LOB Corpus tagging guidelines have changed from 'one-word-one-tag-approach' to idiom tagging to handle the cases of recurrent multiword sequences functioning as units (Johansson et al. 1986).

On the other hand, contractions forming regular patterns such as, *I'll, she's, John's, let's, d'you*, etc. are split up in the tagged LOB corpus as the following: *I' ll*, *she' s, John'* *s, let' s, d' you*. Each part is treated as a separate word and assigned a single tag. Except where *'s* is possessive suffix, then the word gets a single tag entry **$** *e.g. John's* gets the tag **NP$** (Johansson et al. 1986).

Analyzed sentence: أقمت بمدينتي الجديدة لمدة عامين *'aqamtu bimadīnatī al-ğadīdat limuddat 'āmayn* "I have stayed in my new city for two years"

Analyzed word: بمدينتي *bimadīnatī* in my city

Step 1 : Tokenization of words into morphemes

Word	Proclitics	prefixes	Stem	Suffixes	enclitics
بمدينتي	ب *bi* in	-------	مدين ***madīna*** city	ت (ة) *t* feminine *tā'*	ي *ī* my

Step 2 : Assign morpheme tags

Morpheme	*Tag*	*Description*
ب *bi* in	`p--p------------------`	Particle; Preposition
مدين *madīna* city	`nl-------vg?i----tat-s`	Noun; Noun of place; Varied; Genitive; Indefinite; Primitive/ Concrete noun; Augmented by one letter; Triliteral root; Sound noun.
ت *t* feminine *tā'*	`r---f-fs-s-k----------`	Other (Residual); *tā'* of femininization; feminine; Singular; Invariable; *kasra*^{h};
ي *ī* my	`r---r-msfsgs----------`	Other (Residual); Connected pronoun; Common gender; Singular; First person; Invariable; Genitive; *sukūn* (Silence)

Step 3: Assign word tag

Word	*Tag*	*Description*
بمدينتي ***bimadīnatī***	`nl----fs-vgki----tat-s`	Noun; Noun of place; feminine; Singular; Declined; Genitive; *kasra*^{h}; Indefinite; Primitive/ Concrete noun; Augmented by one letter; Triliteral root; Sound noun.

Figure 5.13 Example of tokenization, the SALMA tag assignment for separate morphemes and the combination of the morphemes tags into the word tag

5.6 Chapter Summary

The release of the first Brown corpus in 1964 represented the start of tag set design as scheme for morphosyntactic annotation of corpora. Then, standards and guidelines for morphosyntactic annotation evolved. Eight Arabic tag sets are surveyed and compared in terms of purpose of design, characteristics, tag set size, and their applications. The most widely used and important morphosyntactic annotation standards and guidelines the EAGLES, are designed for Indo-European languages. These guidelines are not entirely suitable for Arabic. Therefore, the design of the SALMA Tag Set applied the standards of traditional Arabic grammar instead. Many Arabic grammar books have been written. A collection of comprehensive and widely used and referenced traditional Arabic grammar books was used as basic reference for morphosyntactic knowledge extraction. The

SALMA Tag Set adds more fine-grained details to the existing tag sets. It encodes 22 morphological feature categories of the word's morphemes where attributes or values are specified by referring to the widely-referenced traditional Arabic grammar books. Chapter 6 describes in detail the morphological feature categories and illustrates each feature and its possible values.

The SALMA Tag Set applied the tag set design criteria proposed by Atwell (2008). The design criteria are dimensions; in effect choices to be made by the designers of new part-of-speech tag sets. Through section 5.4, design decisions are investigated to handle each design dimension. Moreover, references to the existing Arabic tag sets showed the decisions made by these tag sets to handle each design dimension.

Part III: Proposed Standards for Arabic Morphological Analysis

Chapter 6
The SALMA – Tag Set

This chapter is based on the following sections of published papers:

Sections 1 and 2 are based on section 4 from (Sawalha and Atwell Under review)

Chapter Summary

The SALMA Morphological Features Tag Set (SALMA, Sawalha Atwell Leeds Morphological Analysis tag set for Arabic) captures long-established traditional morphological features of Arabic, in a compact yet transparent notation. For a morphologically-rich language like Arabic, the Part-of-Speech tag set should be defined in terms of morphological features characterizing word structure. A detailed description of the SALMA – Tag Set explains and illustrates each feature and its possible values. In our analysis, a tag consists of 22 characters; each position represents a feature and the letter at that location represents a value or attribute of the morphological feature; the dash "-" represents a feature not relevant to a given word. The first character shows the main Parts of Speech, from: noun, verb, particle, punctuation, and Other (residual); these last two are an extension to the traditional three classes to handle modern texts. The characters 2, 3, and 4 are used to represent subcategories; traditional Arabic grammar recognizes 34 subclasses of noun (letter 2), 3 subclasses of verb (letter 3), 22 subclasses of particle (letter 4). Others (residuals) and punctuation marks are represented in letters 5 and 6 respectively. The next letters represent traditional morphological features: gender (7), number (8), person (9), inflectional morphology (10) case or mood (11), case and mood marks (12), definiteness (13), voice (14), emphasized and non-emphasized (15), transitivity (16), rational (17), declension and conjugation (18). Finally there are four characters representing morphological information which is useful in Arabic text analysis, although not all linguists would count these as traditional features: unaugmented and augmented (19), number of root letters (20), verb root (21), types of nouns according to their final letters (22). The SALMA – Tag Set is not tied to a specific tagging algorithm or theory, and other tag sets could be mapped onto this standard, to simplify and promote comparisons between and reuse of Arabic taggers and tagged corpora.

6.1 The Theory Standard Tag Set Expounding Morphological Features

The SALMA – Tag Set is a general-purpose fine-grained tag set. The aim of this tag set is to be used by part-of-speech tagging software to annotate corpora with detailed morphological information for each word, and to enable direct comparisons between tagging algorithms and taggers using the same tag set. The tag set has been designed by grouping 22 morphological feature categories in one tag. Most of these morphological categories are described in any traditional Arabic language grammar book. In our study, all the morphological features are attested in five well known traditional Arabic grammar books (Dahdah 1987; Dahdah 1993; Wright 1996; Al-Ghalayyni 2005; Ryding 2005). Table 6.1 shows the 22 morphological feature categories.

The tag string consists of 22 characters. Each character represents a value or attribute which belongs to a morphological feature category. The position of the character in the tag string is important to identify the morphological feature category. Each morphological feature category attribute is represented by one lowercase letter, which is still human-readable, such as: **v** in the first position to indicate ***verb***, **n** in the second position to indicate ***name***, gender category values in the seventh position: ***masculine*** represented by **m**, ***feminine*** represented by **f** and *common gender* represented by **x.** If the value of a certain feature is not applicable for the word, then a dash '-' is used to indicate this; *e.g.* the mood morphological feature is not a noun feature. In contrast, a question mark '**?**' means a certain feature belongs to a word but, at the moment, the feature value is not available or the automatic tagger could not guess it.

The tag is intended to remain readable by linguists. Moreover, it can be rendered more readable if the interpretation of the tag string features is generated automatically: software can convert each position+letter to a human-readable English and/or Arabic grammar term. Figures 6.1 and 6.2 show examples of two sentences tagged by the SALMA Tag Set. The first sentence is a newspaper text taken from the Arabic Treebank: تم اعداد الوثائق المتوفرة بكثرة حول أول رحلة طيران عثمانية فوق البلاد العربية *tamma 'i'dād al-waṯāi'qa al-mutawaffira*^{ti} *ḥawla 'awwali riḥla*^{ti} *ṭayyarān*^{in} *'uṯmāniyya*^{tin} *fawqa al-bilādi al-'arabiyya*^{ti} 'Many available documents relate to the first Ottoman's flight over the Arab countries'. The second sentence is taken from the Qur'an (chapter 29): وَوَصَّيْنَا الْإِنسَانَ بِوَالِدَيْهِ حُسْنًا *wa waṣṣaynā al-'insāna biwālidayhi ḥusn*^{an} 'We have enjoined on man kindness to parents'.

Word	Morphemes				Tag
wa waaṣṣaynā And We have enjoined	وَوَصَّيْنَا	وَ	*wa*	*And*	p--c------------------
		وَصَّيْ	*waṣṣay*	*Have enjoined*	v-p---mpfs-s-amohvtt&-
		نَا	*nā*	*We*	r---r-xpfs-s----hn----
al-'insāna (on) man	الْإِنسَانَ	الْ	*al-*	*The*	r--d------------------
		إِنسَانَ	*'insāna*	*man*	nq----ms-pafd---htbt-s
bi- wālidayhi His parents	بِوَالِدَيْهِ	بِ	*bi*	*To*	p--p------------------
		وَالِدَ	*wālida*	*Parents*	nu----md-vgki---htot-s
		يْ	*y*	*Both*	r---r-xdts-s----------
		هِ	*hi*	*His*	r---r-msts-k----------
ḥusn^{an} Kindness	حُسْنًا	حُسْنً	*ḥusn*	*kindness*	ng----ms-vafi---ndst-s
		اً	*an*		r---k------f----------

Figure 6.1 Sample of Tagged vowelized Qur'an text using the SALMA Tag Set

Word	Morphemes			Tag
tamma Accomplished	تم	تم	*tamma* Accomplished	v-p---msts-f-amihdstb-
'i'dādu Preparing	اعداد	اعداد	*'i'dādu* Preparing	ng----ms-vndi---?db3-s
al-waṯā'iqa Documents	الوثائق	ال	*al* The	r---d-----------------
		وثائق	*waṯā'iqa* Documents	nq----fb-vafd---ndbt-s
al-mutawaffira^{ti} Available	المتوفرة	ال	*al* The	r---d-----------------
		متوفر	*mutawaffira* Available	nj----fs-vafd---ndtt-s
		ة	*ti*	r---t-fs--------------
bi kaṯra^{tin} In Many	بكثرة	ب	*bi* In	p--p------------------
		كثر	*kaṯra* Many	nj----fb-vgki----dat-s
		ة	*tin*	r---t-fs--------------
ḥawla About	حول	حول	*ḥawla* About	nv----m--s-fi----nst-s
'awwali First	أول	أول	*'awwali* First	n+----ms-vgki----dst-s
riḥla^{ti} Trip	رحلة	رحل	*riḥla* Trip	no----fs-vgki----dat-s
		ة	*ti*	r---t-fs--------------
tayyarān^{in} Flight	طيران	طيران	*tayyarān*^{in} Flight	ng----ms-vgki----dbt-s
uṯmāniyya^{t} Ottomani	عثمانية	عثمان	*uṯmān* Ottoman	n*----fs-pgki----daq-s
		ي	*iyya*	r---y-----------------
		ة	^{t} tā' marbūṭa^{h}	r---t-fs--------------
fawqa Over	فوق	فوق	*fawqa* Over	nv----m--s-fi----nst-s
al-bilādi Countries	البلاد	ال	*al* the	r---d-----------------
		بلاد	*bilād* countries	nl----mb-vgkd---ndat-s
al-'arabiyyati Arabian	العربية	ال	*al* the	r---d-----------------
		عرب	*'arab* Arab	n*----fb-vgkd---hdst-s
		ي	*iyya*	r---y-----------------
		ة	^{ti} tā' marbūṭa^{h}	r---t-fs--------------

Figure 6.2 Sample of Tagged non-vowelized newspaper text using the SALMA Tag Set

The categories and features are drawn from traditional Arabic grammar books (Dahdah 1987; Dahdah 1993; Wright 1996; Al-Ghalayyni 2005; Ryding 2005). In most cases there is agreement among them, but in some cases there are discrepancies. When there is agreement, the approach taken is simply a matter of presenting the agreed features. When there is a discrepancy in most cases the difference is that one text has more fine-grained subcategories which are merged in other texts; so the more fine-grained wider sub-classification is adopted. The only significant disagreement is in the number of nouns; see section 6.2.2, and in that case we adopted the widest most fine-grained sub-classification system.

Arabic grammar terms used to describe the attributes of the morphological feature categories in the SALMA - Tag Set are the same terms used by traditional Arabic grammar. The equivalent English translations of these grammar terms were extracted from 4 well-known traditional Arabic grammar reference books written in English. These books are: Wright, W. (1996), Ryding, K. C. (2005), Dahdah, A. (1993) and Cachia, P. (1973). These reference books agree on translating general Arabic grammar terms such as, noun, verb, adjective, person, number, case and mood. However, these reference books do not agree on translating some fine-grained attribute names such as الفعل السالم *al-fi'l as-sālim*, which is translated into 'the strong verb' by Wright, W. (1996), 'regular (sound) root' by Ryding, K. C. (2005), 'intact verb' by Dahdah, A. (1993), and 'sound verb; strong verb; verbum firmum' by Cachia, P. (1973). The agreed English translations of the grammar terms were directly used. For the non-agreed English translation, Professor James Dickins (head of Arabic and Middle Eastern Studies, University of Leeds, UK) was consulted to give advice on those English translations of Arabic grammar terms that would be clearest to English speaking linguists.

Appendix A lists the morphological features categories and their attribute values at each position of the 22 positions of the tag string.

6.2 The Morphological Features of the SALMA Tag Set

The SALMA Tag Set of Arabic consists of merging 22 morphological features of the Arabic into one compact morphological feature tag. The morphological features categories used to construct the SALMA Tags are listed in table 6.1 below. The following sub-sections 6.2.1 to 6.2.22 describe each morphological category and its attributes in more detail.

Table 6.1 Arabic Morphological Feature Categories

Position	Morphological Features Categories		
1	Main Part-of-Speech	أَقسام الكلام الرئيسيَّة	*’aqsām al-kalām ar-ra’īsiyya*^{t}
2	Part-of-Speech: Noun	أقسام الكلام الفرعيَّة (الاسم)	*’aqsām al-kalām al-far‘iyya*^{t} *(al-’ism)*
3	Part-of-Speech: Verb	أقسام الكلام الفرعيَّة (الفعل)	*’aqsām al-kalām al-far‘iyya*^{t} *(al-fi‘l)*
4	Part-of-Speech: Particle	أقسام الكلام الفرعيَّة (الحرف)	*’aqsām al-kalām al-far‘iyya*^{t} *(al-ḥarf)*
5	Part-of-Speech: Other (Residual)	أقسام الكلام الفرعيَّة (أخرى)	*’aqsām al-kalām al-far‘iyya*^{t} *(’uẖrā)*
6	Punctuation marks	أقسام الكلام الفرعيَّة (علامات الترقيم)	*’aqsām al-kalām al-far‘iyya*^{t} *(‘alāmāt at-tarqīm)*
7	Gender	المُذَكَّر والمُؤَنَّث	*al-muḏakkar wa al-mu’annaṯ*
8	Number	العدد	*al-‘adad*
9	Person	الاسناد	*al-’isnād*
10	Inflectional morphology	الصَّرف	*aṣ-ṣarf*
11	Case or Mood	الحالة الإعرابية للاسم أو الفعل	*al-ḥāla*^{tu} *al-’i‘rābiyya*^{tu} *lil-’ism ’aw al-fi‘l*
12	Case and Mood marks	علامة الإعراب أو البناء	*‘alāmāt al-’i‘rāb wa al-binā’*
13	Definiteness	المَعْرِفَة والنَّكِرَة	*al-ma‘rifa*^{ti} *wa an-nakira*^{ti}
14	Voice	المَبْني لِلمَعْلُوم و المَبْني لِلمَجْهُول	*al-mabnī lil-ma‘lūm wa al-mabnī lil-maǧhūl*
15	Emphasized and non-emphasized	المُؤكَّد وغيرُ المُؤكَّد	*al-mu’akkad wa ǧayr al-mu’akkad*
16	Transitivity	اللازم والمتعدي	*al-lāzim wa al-muta‘addi*
17	Rational	العاقل وغير العاقل	*al-‘āqil wa ǧayr al-‘āqil*
18	Declension and Conjugation	التَّصريف	*at-taṣrīf*
19	Unaugmented and Augmented	المجرَّد والمزيد	*al-muǧarrad wa al-mazīd*
20	Number of root letters	عَدَد أحْرُف الجَذْر	*‘adad ’aḥruf al-ǧaḏr*
21	Verb root	بُنية الفعل	*bunya*^{tu} *al-fi‘l*
22	Noun finals	أقسام الأسم تبعاً للفظ آخره	*’aqsām al-’ismi tib*$^{‘an}$ *li-lafẓi ’āẖirhi*

6.2.1 Main Part-of-Speech Categories

Generally, there is agreement among existing Arabic tag sets on the classification of main part-of-speech categories in traditional Arabic grammar books e.g. (Dahdah 1987; Dahdah 1993; Wright 1996; Al-Ghalayyni 2005; Ryding 2005; ALECSO 2008a) Arabic language scholars classify Arabic words into three main part-of-speech categories namely: nouns, verbs and particles. Khoja’s tag set added categories of punctuation marks and residuals. The punctuation marks used in Arabic are (! ؛ : ؟ - . ،). Others (residuals) include other non-Arabic words appearing in the text such as; currency, numbers or words in other languages. Figure 6.3 lists the attributes of the main part-of-speech category, which occupies the first character in the tag string.

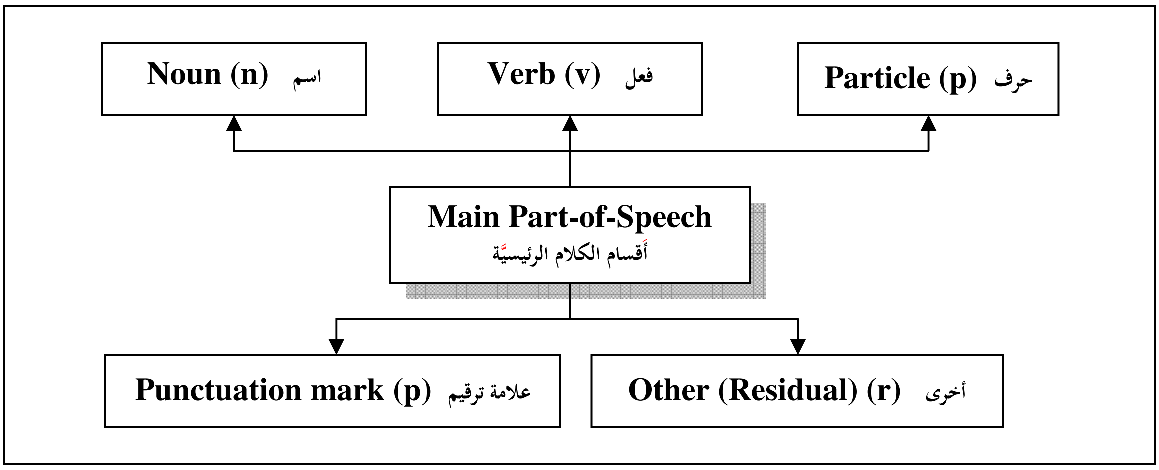


Figure 6.3 Main part-of-speech category attributes and letters used to represent them at position 1

6.2.2 Part-of-Speech Subcategories of Noun

A noun is defined as a word that has complete meaning and no tense associated with it. The Arabic concept of complete meaning corresponds approximately to content words except that it is also includes pronouns. Traditional Arabic grammar uses the concept of meaning to separate nouns and verbs from particles. This is roughly equivalent to content vs. function or lexical vs. grammatical in contemporary lexical terminology. This is not an exact correspondence since pronouns – a grammatical category – are a sub class of nouns. Arabic linguists distinguish many kinds of nouns. According to Dahdah (1987) nouns are classified into 21 kinds. Other classifications overlap. We classified nouns into 34 different types. Table 6.2 shows the 34 different types of nouns and examples of each type. Figure 6.4 shows the classification attributes of the noun part-of-speech category, which occupies the second character in the tag string.

Table 6.2 Noun types as classified in traditional Arabic grammar

	Noun types	**T**	**Meaning and Examples**
1	Gerund / verbal noun المصدر *al-maṣdar*	**g**	A noun which indicates a case or an action that is not related to time or tense. E.g. فَرَحٌ *faraḥ*^{un} ‘happiness’.
2	Gerund / verbal noun with initial *mīm* المصدر الميمي *al-maṣdar al-mīmī*	**m**	A noun which indicates a case or an action that is not related to time or tense. It has certain patterns which have the augmented letter (م) *mīm* at the beginning of the word. E.g. مُنْقَلِب ***munqalib*** ‘turned over’, مَوْعِد ***maw‘****id* ‘date’.
3	Gerund of instance مصدر المرَّة *maṣdar al-marra*^{h}	**o**	A noun that describes an action that has taken place once. It is formed by adding the feminine termination (ة) to the verbal noun. E.g. وَقْفَة *waqfa*^{h} ‘one stop’, زِيَارَة *ziyāra*^{h} ‘a visit’.

	Noun types	**T**	**Meaning and Examples**
4	Noun of state مصدر الهيئة/ مصدر النوع *maṣdar al-hay'a*^{h} / *maṣdar al-naw'*	**s**	A noun that describes an action. It indicates the manner (state, character and representation) of the action expressed by the verb. It always has the form فِعْلَة *fi'la*^{tun}. E.g. مَشى مِشْيَة الأسَد *mašā mišya*^{ta} *al-'asad* 'he walked like a lion'.
5	Gerund of emphasis مصدر التوكيد *maṣdar al-tawkīd*	**e**	A noun that emphasizes an action. E.g. صَوَّرَ اللهُ الخَلْقَ تَصْويراً *ṣawwara allāhu al-ẖalqa taṣwīr*^{an} 'God does shape the creatures'.
6	Gerund of profession المصدر الصناعي *al-maṣdar al-ṣinā'ī*	**i**	A noun which indicates an industry or profession. The gerund of industry ends with doubled yā' followed by feminine *tā' marbūṭa*^{h} (ة). E.g. إنسانيّة *'insāniyya*^{h} 'humanity', وطنيّة *waṭaniyya*^{h} 'nationality' and عالَمية *'ālamiyyah* 'internationality'.
7	Pronoun الضمير *al-ḍamīr*	**p**	Pronouns that belong to this category are the disconnected pronouns. A sentence can start with a pronoun. Pronouns can follow the word (إلّا) *'illā* 'except'. E.g أَنا مُجْتَهِدٌ *'anā muğtahid*^{un} 'I am a hard worker', and ما اجْتَهَدَ إِلّا أَنا *mā 'iğtahada 'illā 'anā* 'no one worked hard except me'. There are 24 pronouns classified into 12 nominative pronouns and 12 accusative pronouns. The nominative pronouns are: أَنَا *'anā* 'I', نَحْنُ *naḥnu* 'We', أَنْتَ *'anta* 'You', أَنْتِ *'anti* 'You', أَنْتُما *'antumā* 'You', أَنْتُم *'antum* 'You', أَنْتُنَّ *'antunna* 'You', هُوَ *huwa* 'He', هِيَ *hiya* 'She', هُما *humā* 'They', هُم *hum* 'They', and هُنَّ *hunna* 'They'. See table 11. The accusative pronouns are: إِيَّايَ *'iyyāya* 'Me', إِيَّانا *'iyyānā* 'us', إِيَّاكَ *'iyyāka* 'your', إِيَّاكِ *'iyyāki* 'your', إِيَّاكما *'iyyākumā* 'your', إِيَّاكُم *'iyyākum* 'your', إِيَّاكُنَّ *'iyyākunna* 'your', إِيَّاهُ *'iyyāhu* 'his', إِيَّاهَا *'iyyāhā* 'her', إِيَّاهما *'iyyāhumā* 'they', إِيَّاهم *'iyyāhum* 'they', إِيَّاهنَّ *'iyyāhunna* 'they'.
8	Demonstrative pronoun اسم الإشارة *'ism al-'išāra*^{h}	**d**	A noun that indicates by a tangible sign a person, an animal, a thing or a place such as; جَاءَ هذا الرجل *ğā' hāḏā ar-rağul* ' this man came', and رَأَيتُ تَينَ الفتاتين *ra'aytu tayna al-fatātayn* ' I saw these two girls'.

	Noun types	**T**	**Meaning and Examples**
9	Specific relative pronoun اسم الموصول الخاص *’ism al-mawṣūl al-ẖāṣ*	**r**	A group of nouns that connect two sentences to give a full meaning. The special relative pronouns are affected by three morphological feature categories, number, gender and humanness. E.g. الَّذي *al-laḏī* ‘who’ is a singular masculine human pronoun; التي *al-latī* ‘who’ is s singular feminine human pronoun; اللواتي *al-lawātī* ‘who’ is a plural feminine human pronoun.
10	Non-specific relative pronoun اسم الموصول المشترك *’ism al-mawṣūl al-muštarak*	**c**	A group of nouns that connect two sentences to give a full meaning. The common relative pronouns are not affected by gender and number, so they have invariable form. They are affected by the morphological feature of humanness. E.g. مَنْ *man* ‘who’ is used for human nouns, ما *mā* ‘who’ is used for non-human nouns, and ذا *ḏā* ‘what’ and أيّ *’ayyu* ‘which’ are used for non-human nouns.
11	Interrogative pronoun اسم الاستفهام *’ism al-’istfhām*	**b**	A pronoun used to make a query or question about a thing or an action, *e.g.* مَنْ هذا؟ *man haḏā?* ‘who is this?’. ما العمل؟ *mā al- ‘amal?* ‘what shall we do?’. The nouns مَنْ *man* ‘who’ and ما *mā* ‘what’ are interrogative nouns.
12	Conditional noun اسم الشرط *’ism al-šarṭ*	**h**	A noun which connects two sentences. It indicates that the action in the second sentence does not occur unless the action of the first sentence has occurred, e.g. أيُّ تِلْميذٍ يَجْتَهِدْ يَنْجَحْ *’ayyu tilmīḏ*^{in} *yağtahid yanğaḥ* ‘if any student studies hard, then he will succeed’. The noun أيُّ *’ayyu* ‘if any’, is a conditional noun.
13	Allusive noun الكناية *al-kināya*^{h}	**a**	A noun which indicates a specific intention by means of unclear terms. These nouns are: كَأيِّ *ka’ayyi* ‘Any’, كَذا *kaḏā* ‘So and so’, كَم *kam* ‘How …’, كَيْتَ *kayta* ‘So and so’**,** ذَيْتَ *ḏayta* ‘So and so’, بِضْعُ *biḍ‘u* ‘few’, فُلانُ *fulān* ‘someone’, e.g. كَأيِّ عصفوراً اصطدت *ka’ayyi ‘usfūr*^{an} *’isṭadta* ‘Like any bird you have hunted’. The word كَأيِّ *ka’ayyi* ‘As any’, is a generalization
14	Adverb الظَّرف *aẓ-ẓarf*	**v**	A noun which indicates the time or place of the action. It incorporates into its overall meaning a sence of relative locality on time or place, *e.g.* حينَ *ḥīna* ‘when’, مُدَّة *mudda*^{tu} ‘at a period of’, and أَمام *’amām* ‘straight forward (direction)’

	Noun types	**T**	**Meaning and Examples**
15	Active participle اسم الفاعل *’ism al-fā‘il*	**u**	A form that describes the doer of the action. This noun is derived from the action or the verb itself. E.g. كاتِبٌ *kātib*^{un} ‘writer’. This noun is derived from the action of *writing* or the verb *write* كَتَبَ *kataba.*
16	Intensive Active participle مُبَالَغة اسم الفاعِل *mubālağa*^{t} *’ism al-fā‘il*	**w**	A noun which has the same basic meaning as the present participle اسم الفاعل *’ism al-fā‘il* but indicates an augmentation of the meaning of the present participle. E.g. كَتَّابٌ *kattāb*^{un} ‘writer’, which indicates that the *writer* writes a lot. *kattāb*^{un} is derived from the verb ‘write’ كَتَبَ *kataba.*
17	Passive participle اسم المفعول *’ism al-maf‘ūl*	**k**	A derived noun which indicates an abstract meaning that describes something or someone affected by an action. E.g. مَكْسورٌ *maksūr*^{un} ‘broken’. This noun is derived from the verb break كَسَرَ *kasara.*
18	Adjective الصِّفة المشبَّهة *aṣ-ṣifa*^{h} *al-mušabbaha*^{h}	**j**	A derived noun which indicates a meaning of firmness. *i.e.* the absolute existence of the quality in its possessor. *E.g.* الجُنْدِيُّ شُجَاعٌ *al-ğundiyyu šuğā‘*^{un} ‘brave soldier’. The word شُجَاعٌ *šuğā‘*^{un} ‘brave’ describes the soldier. This word is an adjective.
19	Noun of place اسم المكان *’ism al-mkān*	**l**	A derived noun which indicates the place of an action. E.g. مَطْبَخٌ *maṭbaẖ*^{un} ‘kitchen’ indicates the place of cooking.
20	Noun of time اسم زمان *’ism zamān*	**t**	A derived noun which indicates the time of the action or a verb. E.g. مَغْرِبٌ *mağrib*^{un} ‘sunset’.
21	Instrumental noun اسم الآلة *’ism al-’āla*^{h}	**z**	A derived noun which indicates a tool used to some work. E.g. مِفْتاحٌ *miftāḥ*^{un} ‘key’, مِنشار *minšār* ‘saw’, and مِصباح *miṣbāḥ* ‘light’.
22	Proper noun اسم العلم *’ism al-‘alam*	**n**	The name of a dedicated or specific instance in a group or type. E.g. خالِدٌ *ẖālid*^{un} ‘Khalid’, عَبْدُاللّٰهِ *‘abdu allāhi* ‘Abdullah’, بَيْروتُ *bayrūt* ‘Beirut (the capital city of Lebanon)’.
23	Generic noun اسم الجنس *’ism al-ğins*	**q**	Indicates what is common to every element of the genus without being specific to any one of them. E.g. كِتابٌ *kitāb*^{un} ‘book’, رَجل *rağul* ‘man’, and بيت *bayt* ‘home’.

	Noun types	**T**	**Meaning and Examples**
24	Numeral اسم العدد *’ism al-‘adad*	+	A noun that indicates the quantity and order of countable nouns by transferring the numbers into the correct form of Arabic words. E.g. رَجُلٌ واحِدٌ *raǧul*^{un} ***wāḥid***^{un} ‘one man’. رَجُلانِ إثنانِ *raǧulāni* ***’iṯnāni*** ‘two men’. ثَلاثَةُ رِجالٍ *ṯalāṯatu riǧāl*^{in} ‘three men’. The words واحد ، اثنان و ثلاثة *wāḥid, ’iṯnāni and ṯalāṯa*^{h} ‘one’, ‘two’ and ‘three’, are ordinal numeral nouns.
25	Verb-like noun اسم الفعل *‘ism al-fi’il*	&	A noun which acts as a verb in its meaning. It indicates time of action, e.g. شَتَّانَ *šattāna* ‘how different they are!’, هَيهَاتَ *hayhāt* ‘but oh! far from the mark!’ and بَعُدَ *ba’uda* ‘far away’.
26	The five nouns الأسماء الخمسة *al-’asmā’ al-ẖamsa*^{h}	f	The five nouns are a group of five nouns belonging to the category of noun of genus. However, unlike standard nouns, which have three root letters, each of these nouns has only two root letters the third root letter being deemed to have been deleted. The five nouns are أَبٌ *’abun* ‘father’, أَخٌ *’aẖun* ‘brother’, حَمٌ *ḥamun* ‘father in law’, فو *fū* (فَم *fam*)‘mouth’, and ذو *ḏū* ‘owner’.
27	Relative noun اسم منسوب *’ism mansūb*	*	A declinable noun which has the suffix *–iyy*.. It indicates affiliation of something to this noun. E.g. أُرْدُنِيٌّ *’urduniyy*^{un} ‘Jordanian’ (*i.e.* affiliated to Jordan).
28	Diminutive اسم تصغير *’ism taṣġīr*	y	A declinable noun which has the sound *-ai-* after its second root letter. It indicates paucity, contempt or affection. E.g. دُرَيْهِمات *duraihimāt* ‘a few dirhams’, شُوَيْعِر *šuway‘ir* ‘poetaster’, and بُنَيَّ *bunayya* ‘my (little) son’.
29	Form of exaggeration صيغة مبالغة *ṣīġa*^{t} *al-mubālaġa*^{h}	x	It indicates exaggeration of the quality of the qualified noun and occurs as a derived noun with the basic meaning of the present participle. E.g. زَرّاع *zarrā‘* ‘a very good cultivator’.
30	Collective noun اسم جمع *’ism ǧam’*	$	A noun which indicates two or more. A singular form cannot be derived from this kind of noun. E.g. جَيْش *ǧayš* ‘army’, the corresponding singular being جندي *ǧundī* ‘a soldier’, or خَيْل *ẖayl* ‘horses’ the corresponding singular being فَرَس *faras* ‘a horse’.

	Noun types	**T**	**Meaning and Examples**
31	Plural collective noun اسم جنس جمعي *’ism ğins ğam‘ī*	#	A noun of genus where the singular and plural share the same basic form in meaning and pronunciation. The singular form is distinguished by adding the feminine *tā’ marbūtah* or the relative suffix *–ī*. E.g. زهر (زهرة) *zahr (zahra^h)* ‘flowers’ (‘a flower’), and عرب (عربي) *‘arab (‘arabī)* ‘Arabs’ (‘an Arab’).
32	Elative noun اسم تفضيل *’ism tafḍīl*	@	A derived noun used for the comparative and superlative when comparing persons or things. E.g. الأَسَدُ أقوى مِنَ الرَّجُلِ *al-’asadu ’aqwā mina ar-rağuli* ‘The lion is stronger than the man’. The noun أقوى *’aqwā* ‘stronger’ is used for comparing the strength of the lion and the man.
33	Blend noun اسم منحوت *’ism manḥūt*	%	This consists in composing a single word by the fusion of two or more words, so that some letters are dropped from each word on condition that the resultive form has an authentically acceptable pronunciation and meaning. E.g. جَعْفَلُ *ğa‘falu* ‘Could I but sacrifice myself for you’ composed from the words جَعِلْتُ فِداكَ *ğa‘altu fidāka* (same meaning).
34	Ideophonic interjection اسم صوت *‘ism ṣawt*	!	A noun improvised by human spontaneity and used initially as a verbal noun to talk to animals and small children, e.g. آه *āh* “Oh”, هَال *hāl* used for horses.

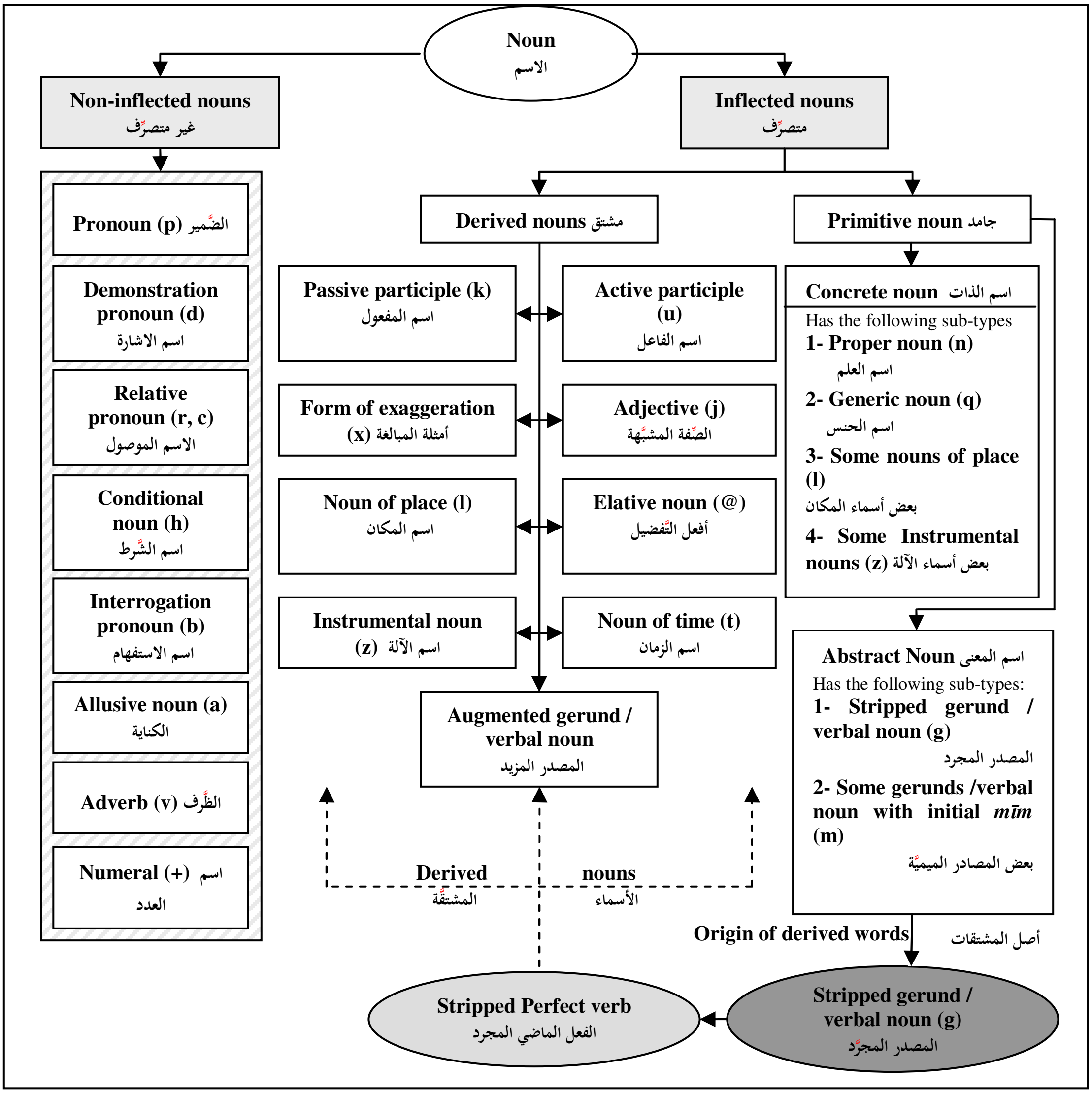


Figure 6.4 The classification attributes of noun part-of-speech subcategories with letter at position 2.

6.2.3 Part-of-Speech Subcategories of Verb

A verb is defined as a word that indicates a meaning by itself which is united with a tense or time; verbs takes words or affixes as indicators such as the particles قد *qad,* سوف *sawfa* , or suffixed pronouns or the prefixes س /s/, ت /t/, ن /n/ (Al-Ghalayyni 2005).

Verbs can be classified according to tense and morphological form into three groups. Table 6.3 shows the 3 attributes of the part-of-speech subcategories of verbs with their definition and examples of each attribute. Figure 6.5 below shows the subcategories of the verb, represented at position 3 of the tag string.

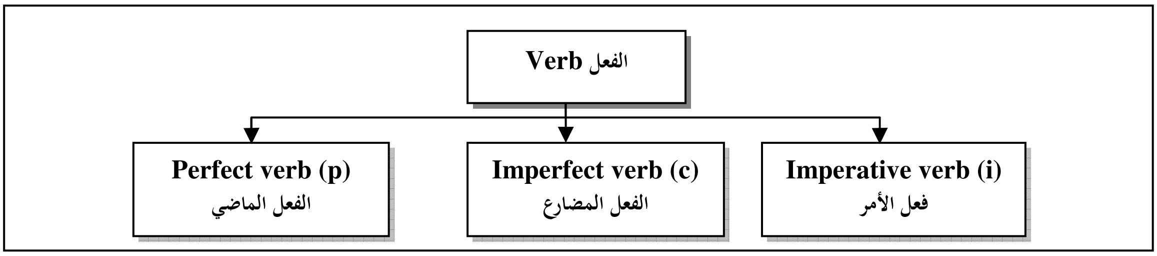


Figure 6.5 Part-of-Speech subcategories of verb, with letter at position 3

Table 6.3 Verb types as classified by Arab grammarians

Verb types	**T**	**Meaning and Examples**
Perfect verb الفعل الماضي *al-fi'l al-māḍī*	**p**	Indicates the occurrence of an action is in the past. E.g. كَتَبَ الطالبُ الدرسَ *kataba aṭ-ṭāilbu ad-darsa* 'The student wrote the lesson'. The verb كَتَبَ *kataba* 'wrote' is a perfect verb.
Imperfect verb الفعل المضارع *al-fi'l al-muḍāri'*	**c**	Indicates an action or case in the progressive tense or the action occurs at the time of speaking. E.g. يَتَكَلَّمُ *yatakallamu* 'someone is talking now'.
Imperative verb فعل الأمر *fi'l al-'amr*	**i**	Indicates a required action in the future, or a request (order) to do an action. E.g. أُكْتُبْ *'uktub* 'write' as a request or order.

6.2.4 Part-of-Speech Subcategories of Particles

Particles are classified in two broad categories. The first category is non-meaningful particles حروف المباني *ḥurūf al-mabānī* or alphabet letters. From these alphabet letters Arabic words are constructed. The second category is meaningful particles حروف المعاني *ḥurūf al-ma'ānī*. They are words which do not belong to noun or verb but they add specific meaning to the noun or verb in a sentence, or they connect two or more sentences. They are also classified according to their 'effect' on nouns or verbs into two groups; governing particles حروف عاملة *ḥurūf 'āmila^h* which affect the form of the following noun or verb; and non-governing particles غير عاملة *ḥurūf ġayr 'āmila^h* which do not affect the form of the following nouns or verbs (Dahdah 1987; Dahdah 1993).

Governing particles affect the following noun or verb by changing the mood of the verb or the case of the noun. They affect the verb by changing its mood to jussive, subjunctive or partially subjunctive. And they affect the case of noun in genitive, vocative or exception. Conjunctions حروف العطف *ḥurūf al-'aṭf* affect both nouns and verbs. Table 6.4 shows definitions and examples of the 22 subcategories of particles. Figure 6.6 shows the particles category attributes, represented at position 4 of the tag string.

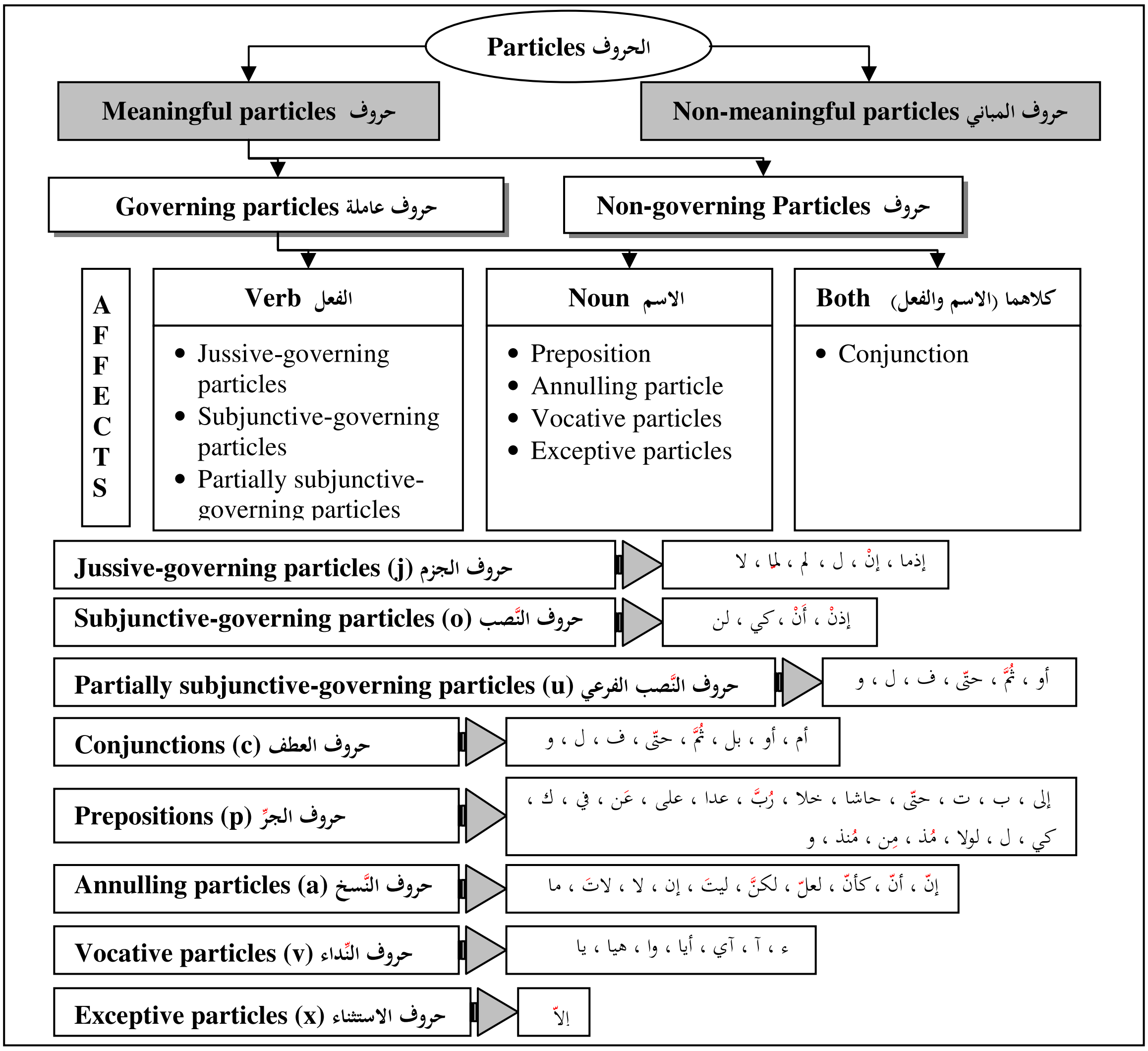

Particles الحروف
Meaningful particles حروف
Non-meaningful particles حروف المباني
Governing particles حروف عاملة
Non-governing Particles حروف
AFFECTS
Verb الفعل
Jussive-governing particles
Subjunctive-governing particles
Partially subjunctive-governing particles
Noun الاسم
Preposition
Annulling particle
Vocative particles
Exceptive particles
Both كلاهما (الاسم والفعل)
Conjunction
Jussive-governing particles (j) حروف الجزم
إذما ، إنْ ، ل ، لم ، لما ، لا
Subjunctive-governing particles (o) حروف النَّصب
إذنْ ، أَنْ ، كي ، لن
Partially subjunctive-governing particles (u) حروف النَّصب الفرعي
أو ، ثُمَّ ، حتّى ، ف ، ل ، و
Conjunctions (c) حروف العطف
أم ، أو ، بل ، ثُمَّ ، حتّى ، ف ، ل ، و
Prepositions (p) حروف الجرّ
إلى ، ب ، ت ، حتّى ، حاشا ، خلا ، رُبَّ ، عدا ، على ، عَن ، في ، ك ،
كي ، ل ، لولا ، مُذ ، مِن ، مُنذ ، و
Annulling particles (a) حروف النَّسخ
إنّ ، أنّ ، كأنّ ، لعلّ ، لكنَّ ، ليتَ ، إن ، لا ، لاتَ ، ما
Vocative particles (v) حروف النّداء
ء ، آ ، آي ، أيا ، وا ، هيا ، يا
Exceptive particles (x) حروف الاستثناء
إلاّ

Figure 6.6 Subcategories of Particle, with letter at position 4

Table 6.4 Examples of part-of-speech category attributes

	Particle Type	T	Meaning and Examples
1	Jussive-governing particle حرف جزم *ḥarf ğazim*	**j**	A group of particles that have the meaning of negation and prevention. They govern a following imperfect verb in the jussive mood. E.g. لا تيأسْ من رحمة الله *lā tay'as min raḥma*^{ti} *al-lā*^{h} 'Do not give up God's mercy'.
2	Subjunctive-governing particle حرف نصب *ḥarf naṣib*	**o**	A group of particles that govern a following imperfect verb in the subjunctive mood. Mainly used for conditions. E.g. جئتُ لكي أتعلَّمَ *ği'tu likay at'allama* 'I came to study'.
3	Partially Subjunctive-governing particle حرف نصب فرعي *ḥarf naṣib far'ī*	**u**	A group of particles that govern a following imperfect verb in the subjunctive mood through an implicit *'an* (أَنْ المضْمَرة). E.g. مُقاوَمَتُك العَدوَّ ثُمَّ تَنْتَصِرَ فَخْرٌ عَظِيمٌ *muqāwamatuka al-'aduwwa* ***ṯumma*** *tantaṣira faẖrun 'aẓīmun* 'your resistance to the enemy, then your victory, are the source of a great pride'.

	Particle Type	**T**	**Meaning and Examples**
4	Preposition حرف جرّ *ḥarf ǧarr*	**p**	A group of particles that govern a following noun in the genitive case. This group consists of true and fundamental markers of location and direction particles. E.g. دَرَستُ إلَى المِساءِ *darastu ’ilā almasā’i* ‘I studied up to the night’.
5	Annulling particle حرف ناسخ *ḥarf nāsiḫ*	**a**	A group of particles that ‘intervene’ in the nominal sentence and induce a change in the case of the following noun. These particles include إنَّ وأخَواتِها *’inna wa ’aḫawātihā* ‘indeed and its sisters’, لا النَّافية للجنس *lā an-nāfiyah lil-ǧins* ‘generic negative *lā* ’ and ما وَ أَخَواتِها *mā wa ’aḫawātihā* ‘*mā* and its sisters’. E.g. إنَّ الطَّقسَ جَمِيلٌ *’inna aṭ-ṭaqsa ǧamīlun* ‘Indeed, the weather is nice’
6	Conjunction حرف عطف *ḥarf ‘aṭf*	**c**	A group of particles used to connect elements of equal status in pronunciation or in meaning. This group includes ten conjunctions. E.g. جاءَ عليٌّ و خالدٌ *ǧā’a ‘aliyy*^{un} *wa ḫālid*^{un} ‘Ali and Khalid came’.
7	Vocative particle حرف نِداء *ḥarf nidā’*	**v**	A group of particles used to call or alert the person addressed. There are eight vocative particles. A noun preceded by a vocative particle is called a vocative noun. E.g. أَيَا طَالِبُ اسْتَمِعْ *’ayā ṭālibu ’istami‘* ‘Oh student, listen’.
8	Exceptive particle حرف استثناء *ḥarf ’istiṯnā’*	**x**	A group of particles used to exclude the following noun from the scope of the words before it. E.g. جَاءَ التَّلاميذُ إلّا سَميراً *ǧā’ at-talāmīḏu ’illā samīr*^{an} ‘The students came except Samir’.
9	Interrogative particle حرف استفهام *ḥarf ’istifhām*	**i**	A group of particles used to ask to elicit understanding, conception or approval. This group includes three interrogative particles. The noun which follows an interrogative particle is called an interrogative noun. E.g. هَلْ جَاءَ زَيدٌ؟ *hal ǧā’ zayd*^{un}*?* ‘Did Zaid come?
10	Particle of futurity حرف استقبال *ḥarf ’istiqbāl*	**f**	A group of particles which modifies the verb tense from the present tense to the future. The particles of futurity include the letter (س) *sīn* and the particle (سَوفَ) *sawfa*, both meaning ‘will’. E.g. سَوفَ أَعودُ *sawfa ’a‘ūdu* ‘I will come back’.
11	Causative particle حرف تعليل *ḥarf ta‘līl*	**s**	A group of particles used to express and confirm the logic of an argument. These eight particles are: إذْ *’iḏ* ‘since’, حَتّى *ḥattā* ‘in order to’, عَلى *‘alā* ‘on’, عَنْ *‘an* ‘About’, فِي *fī* ‘in’, كَيْ *kay* ‘so that’, اللّام *lām* ‘so that’, مِنْ *min* ‘from’. E.g. أُدْرُسْ حتى تنجح *’udrus **ḥattā** tanǧaḥ* ‘Study in order to succeed’.

	Particle Type	**T**	**Meaning and Examples**
12	Negative particle حرف نفي *ḥarf nafī*	**n**	A group of particles used to negate the proposition expressed after them, or to deny its affirmation. There are eight negative particles. These particles are: إِنْ *’in* ‘not’ (with more standard sense of ‘if’), كَلّا *kallā* ‘never’, لَمْ *lam* ‘not (in the past)’, لَمَّا *lammā* ‘ not yet’ , لَنْ *lan* ‘not (in the future)’, لا *lā* ‘not’, لاتَ *lāta* ‘not’, مَا *mā* ‘not’. E.g. لَمَّا يأْتِ القِطارُ *lammā ya’tī al-qiṭāru* ‘The train has not (yet) arrived’.
13	Jurative particles حرف قسم *ḥarf qasam*	**q**	A group of particles used to swear by the divine majesty or by another feature. There are four jurative particles. These are: ب *bā’*, ت *tā’*, ل *lām*, و *wāw*. E.g. بِاللهِ لأفعلنّ *bi-allāhi la-’af‘alanna* ‘By God I will surely do it’.
14	Yes/No response particle حرف جواب *ḥarf ğawāb*	**w**	A group of particles used to reply to an invocation, a question, a statement, a correspondence or an objection. There are eleven response particles. These particles are: أَجَلْ *’ağal* ‘yes’, إِذَنْ *’iḏan* ‘in that case’, إِذَاً *’iḏ^{an}* ‘ihen’, إي *’ī* ‘yes’, بَلى *balā* ‘yes’, جَلَلْ *ğalal* ‘yes’, جِيْرِ *ğayr* ‘yes’, الفاء *fā’* , الّلام *lām,* لا *lā* ‘no’, نَعَمْ *na‘am* ‘yes’. E.g. إِذَاً أنْتَ نَاجِحٌ ***’iḏ^{an}*** *anta nāğiḥ^{un}* ‘Then you have succeeded’.
15	Jussive-governing conditional particle حرف شرط جازم *ḥarf šarṭ ğāzim*	**k**	A group of particles used to express the occurrence of one event in connection with another one. There are two jussive-governing conditional particles. إِذ مَا *’iḏ mā* ‘whenever’ and وَإِنْ *wa ’in* ‘even if’ . E.g. إِذْ ما تَتَعَلَّم تَتَقَدَّم *’iḏ mā tata‘allam tataqaddam* ‘Whatever you learn you will progress’.
16	Incitement particle حرف تحضيض *ḥarf taḥḍīḍ*	**m**	A group of particles used to request something with force, incitement, and harassment. There are five incitement particles. These particles are: أَلا *’alā* ‘is it (etc.) not’, أَلَّا *’allā* ‘lest’, لولا *lalā* ‘were it (etc.) not’, لَوْما *lawmā* ‘if it were (etc.) not’, هَلَّا *hallā* ‘is it (etc.) not. E.g. هَلَّا تَقُومُ بِواجِبك *hallā taqūmu bi wāğibika* ‘Will not you carry out your duty’.
17	Gerund-equivalent particle حرف مصدري *ḥarf maṣdarī*	**g**	A group of particles used to ‘intervene’ in a sentence which can be replaced by gerund. These four particles are: الهمزة *hamza^{h}*, أَنْ *’an* ‘that’, كَيْ *kay* ‘so’, لَوْ *law* ‘if’. E.g. أُحِبُّ أَنْ أَخْدِمَ وَطَني *’uḥibbu ’an aḫdima waṭanī* ‘I like to serve my country’.

	Particle Type	T	Meaning and Examples
18	Particle of attention حرف تنبيه *ḥarf tanbī^{h}*	t	A group of particles used to clarify the matter for the orientation of the alert listener. There are two attention particles; أَلا *alā* ‘is it not’, and الهاء *hā’* ‘attention’. E.g. يَا أَيُّهَا الرَّجُلُ المعَلِّمُ غَيْرَهُ *yā’ayyuhā ar-rağulu al-mu‘allimu ḡayra^{hu}* ‘I call on you, man who teaches others’.
19	Emphatic particle حرف توكيد *ḥarf tawkīd*	z	A group of particles used to emphasise intention and to consolidate a pledge. There are eight emphatic particles. أمَّا *’ammā* ‘as for’, أَنْ *’an* ‘that’, إِنَّ *’inna* ‘indeed’, الباء *bā’*, عَلى *‘alā* ‘on’, الكاف *kāf*, النُّون *nūn*, نَّ *nna*. E.g. إِنَّ الطَّقسَ جَمِيلٌ *’inna aṭ–ṭaqsa ğamīlun* ‘Indeed, the weather is nice’
20	Explanatory particle حرف تفسير *ḥarf tafsīr*	d	A group of particles used to clarify the meaning of a word, to discover the purpose of a question and to interpret it. There are two explanatory particles. أَنْ *’an* ‘that’, and أَيْ *’ay* ‘That is’. E.g. هَذا عَسْجَدٌ أَيْ ذَهَبٌ *hadā ‘asğadun ’ay dahabun* ‘This is a precious metal, that is gold’.
21	Particle of comparison حرف تشبيه *ḥarf tašbī^{h}*	l	A group of particles used to liken one thing to another, but not in the same way as a metaphor. There are two particles of comparison; الكاف *kāf*, and كَأَنَّ *ka’anna* ‘As if’. E.g. كَأنَّكَ البَدْرُ *ka’annaka al-badru* ‘As if you are a full moon’.
22	Non-governing particle حرف غير عامل *ḥarf ḡayr ‘āmil*	b	A group of particles that do not affect the following word by changing its case or mood such as قَدْ *qad* ‘already/indeed’ or ‘perhaps’. E.g. قَدْ أَفْلَحَ مَنْ زَكَّاهَا *qad aflaḥa man zakkāhā* ‘Indeed, he has succeeded who has purified it’.

6.2.5 Part-of-Speech Subcategories of Others (Residuals)

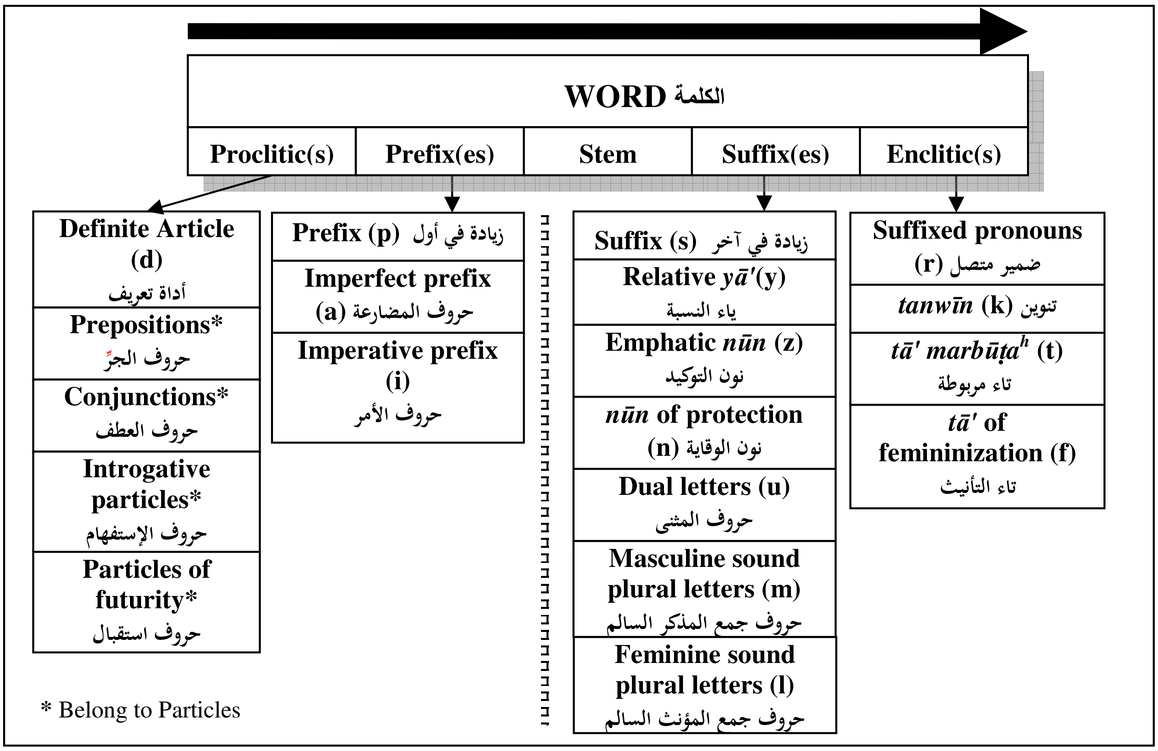
Most Arabic words consist of multiple parts. These parts are proclitic(s), prefix(es), stem, suffix(es) and enclitic(s). Clitics and affixes belong to nouns or particles. They affect some of the morphological features of the word. For example, prepositions change the case of nouns to genitive, while the letters ‘ون’ *wāw-nūn*, which are added to the end of the word (verb or noun), indicate plural number, masculine gender and nominative case when added to nouns. As these special particles or pronouns are attached to the word as affixes or clitics, we separated them in a morphological feature category of Others (residuals). Figure 6.7 shows the word structure and the residuals with part-of-speech Others (residuals) that belongs to each part of the word.

Table 6.5 lists the 15 subcategories of the part-of-speech Others (residuals), and explains the effects on verbs or nouns. The part-of-speech category of Others (residuals) is represented at the fifth position of the tag string.

Table 6.5 Examples of the part-of-speech category of Others (residuals)

	Others (Residuals)	**T**	**Explanation**
1	Prefix زيادة في أول الكلمة *ziyāda*^{h} *fī 'awwal al-kalima*^{h}	**p**	A morpheme added to the beginning of a basic word's pattern to derive another word. These letters will add more meanings to the word such as; emphasis, transitivity, etc.
2	Suffix زيادة في آخر الكلمة *ziyāda*^{h} *fī 'āḫir al-kalima*^{h}	**s**	A morpheme attached to the end of a basic word's pattern to derive another word. These letters will add more meanings to the verb such as; emphasis, transitivity, etc.
3	Suffixed pronoun ضمير متصل *ḍamīr muttaṣil*	**r**	A group of pronouns that are attached to the end of the verb or noun which represent the subject or the object of the verb.
4	*tā' marbūṭa*^{h} تاء مربوطة	**t**	A morpheme that is attached to the end of the noun or adjective to indicate feminine gender.
5	Relative *yā'* ياء النسبة *yā' an-nisba*^{h}	**y**	A morpheme that is attached to the end of the noun or adjective to mark relative nouns.
6	*tanwīn* تنوين	**k**	A morpheme (diacritic) attached to the end of the noun or adjective to mark indefiniteness morphological feature.
7	*tā'* of femininization تاء التأنيث *tā' al-ta'nīṯ*	**t**	A morphological letter that is attached to the end of the noun or verb to indicate feminine gender.
8	*Nūn* of protection نون الوقاية *nūn al-wiqāya*^{h}	**n**	A morphological letter that is attached to the end of the verb to separate between words ending with the ن *nūn* and other suffixes attached to the word starting with the letter ن *nūn*. E.g. عَلَّمَنِي *'allamanī* 'he taught me' *nūn* of protection appears between the perfect verb عَلَّمَ *'allama* and the object suffixed pronoun ي *–ī* 'me'.
9	Emphatic *nūn* نون التوكيد *nūn al-tawkīd*	**z**	A morpheme that is attached to the end of the verb to add emphasis to the word by adding the letter نْ *nūn* or doubled one نّ *nūn-nūn*.
10	Imperfect prefix *ḥarf muḍāra'a*^{h} حرف مضارعة	**a**	One of a group of morphemes attached at the beginning of the verb stem which mark the verb as being imperfect (or progressive) rather than perfect.

	Others (Residuals)	T	Explanation
11	Definite article أداة تعريف *’adā*^{t} *ta‘rīf*	d	A ‘definiteness particle’, added to the beginning of the nouns or adjectives and making them definite, rather than indefinite.
12	Masculine sound plural letters حروف جمع المذكر السالم *ḥurūf ğam‘ al-mudakkar as-sālim*	m	A morpheme that is attached to the end of singular nouns or adjectives to form sound plurals. They are used to derive masculine plural.
13	Feminine sound plural letters حروف جمع المؤنث السالم *ḥurūf ğam‘ al-mu’nnaṯ as-sālim*	l	A morpheme that is attached to the end of singular nouns or adjectives to form sound plurals. They are used to derive feminine plural.
14	Dual letters حروف المثنى *ḥurūf al-muṯannā*	u	A morpheme that is attached to the end of singular nouns or adjectives to derive dual noun or adjective. To derive feminine dual these letters must be preceded by the feminine letter *tā’* (ت) (تاء التأنيث).
15	Imperative prefix حرف الامر *ḥarf al-’amr*	i	A morpheme that is attached at the beginning of the verb stem and changes it from perfect to imperative verb.



WORD الكلمة
Proclitic(s)
Prefix(es)
Stem
Suffix(es)
Enclitic(s)
Definite Article (d) أداة تعريف
Prepositions* حروف الجرّ
Conjunctions* حروف العطف
Introgative particles* حروف الإستفهام
Particles of futurity* حروف استقبال
Prefix (p) زيادة في أول
Imperfect prefix (a) حروف المضارعة
Imperative prefix (i) حروف الأمر
Suffix (s) زيادة في آخر
Relative yā’(y) ياء النسبة
Emphatic nūn (z) نون التوكيد
nūn of protection (n) نون الوقاية
Dual letters (u) حروف المثنى
Masculine sound plural letters (m) حروف جمع المذكر السالم
Feminine sound plural letters (l) حروف جمع المؤنث السالم
Suffixed pronouns (r) ضمير متصل
tanwīn (k) تنوين
tā' marbūṭa^h (t) تاء مربوطة
tā' of femininization (f) تاء التأنيث
* Belong to Particles

Figure 6.7 The word structure and the residuals that belong to each part of the word, with letter at position 5

6.2.6 Part-of-Speech Subcategories of Punctuation Marks

Punctuation appears in most Arabic texts. Punctuation marks include: full stop, comma, colon, semi colon, parentheses, square brackets, quotation mark, dash, question mark, ellipsis and continuation mark. “Punctuation usage in original Arabic text is characterized by a great deal of fluidity” (Khafaji 2001) Figure 6.8 shows the punctuation marks that are used in Arabic text. Table 6.6 lists the 12 subcategories of punctuation marks and their use. The part-of-speech category of punctuation marks is represented at the sixth position of the tag string.

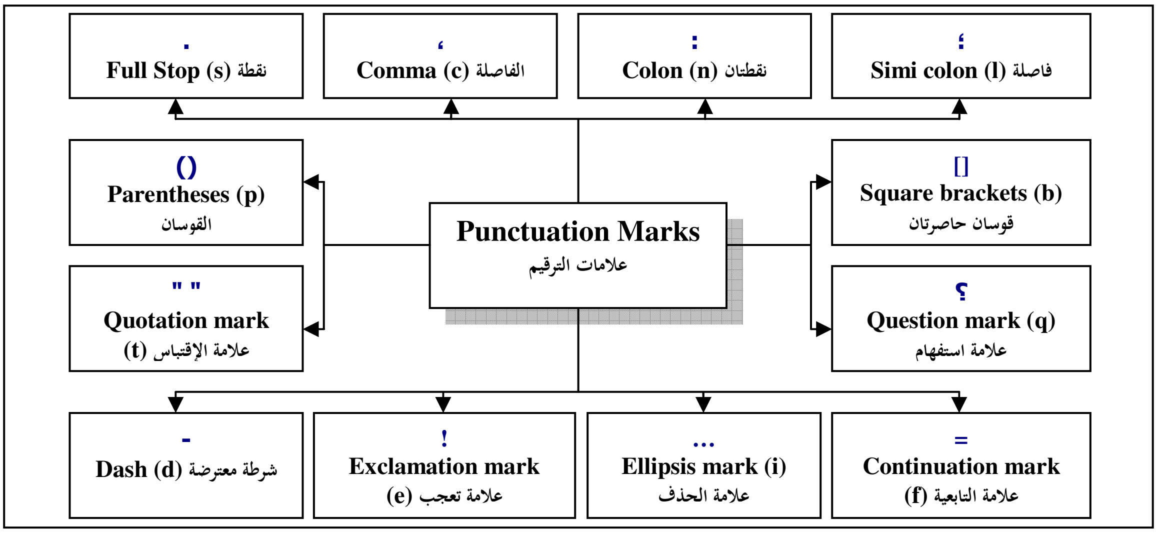


Figure 6.8 Punctuation marks used in Arabic, with letters at position 6

Table 6.6 Subcategories of punctuation and examples of their attributes

#	Punctuation marks	T	Example
1	Full stop نقطة (.) *nuqṭa*^{h}	**s**	A full stop is used at the end of paragraph, or after the meaning is completed. E.g. طَلَعتْ الشمسُ. *ṭalaʻat aš-šamsu* “the sun has risen.”
2	Comma فاصلة (،) *fāṣila*^{h}	**c**	A comma is used after the vocative and to separate phrases or clauses. E.g. يارجلُ، إنَّكَ مُهَدَّدٌ بِالخَطَرِ. *yā rağulu, ’innaka muhddad*^{un} *bilkhaṭar* “hey man, you are in danger.”
3	Colon نقطتان (:) *nuqṭatān*	**n**	A colon is used after reported speech. E.g. (قالَ: أنا ذَاهِبٌ.) *qāla: ‘anā ḏāhib*^{un}. “he said: I am leaving”
4	Semi-colon فاصلة منقوطة (؛) *fāṣila*^{h} *manqūṭa*^{h}	**l**	A semi-colon is used between two linked clauses, e.g. if one is the cause of the other. E.g. عَلِمْتُ أَنَّهُ قَادِمٌ؛ وَهَلْ يُعْقَلُ أَلاّ يَأْتِي؟ *‘alimtu ’annahu qadim*^{un}*; wahal yu’qalu ’allā ya’tī?* “I knew that he is coming; is it possible that he is not coming?”
5	Parentheses قوسان (( )) *qawsān*	**p**	Parentheses are used around numbers, and sometimes used for limitations. E.g. جاءَ ثَمانِي (8) نِساءٍ *ğā’ (8) nisā’* “8 women have come”.

#	Punctuation marks	T	Example
6	Square brackets ([ ]) قوسان حاصرتان *qawsān ḥāṣiratān*	**b**	Square brackets are used for limitation, and are also used around the sentence added to a quotations. E.g. قال المعريّ: "هَذَا جَنَاهُ أَبِي عَلَيَّ [مَعَ أَنَّ الجُنَاةَ عَلَيْهِ كُثُرٌ] ومَا جَنَيْتُ عَلَى أَحَدِ " . *qāl al-ma‘rrī: “haḏā ğanāhu ’abī ‘alayya [ma‘ ’anna al-ğunā^{ta} ‘alyhi kuṯur^{un}] wamā ğanaytu ‘lā ’aḥad”.* “al-ma’arry said: “This what my father did to me [although many people hurt him] and I have never hurt anybody”
7	Quotation mark (" ") علامة اقتباس *‘alāma^{tu} ‘iqtibās*	**t**	Quotation marks are used for quotations without changing the original text. E.g. قال جبران : " تَعَلَّمتُ الصَّمْتَ مِنَ الثَّرْثَارِ ... " *qāl ğubrān: ta‘almtu aṣ-ṣmta mina aṯ-ṯarṯār…”* (Jubran said: “I learnt how to be silent from a talkative person”.)
8	Dash (–) شرطة معترضة *šarṭa^{h} mu‘tariḍa^{h}*	**d**	A dash is used at the beginning and end of a parenthetical clause. It is also used when speaker is changed. E.g. ما اسمك؟ – اسمي سَمِيرٌ *mā ’ismuka? – ‘ismī samīr^{un}* “What’s your name? – My name is Samir”
9	Question mark (؟) علامة استفهام *‘alāma^{tu} ’istifhām*	**q**	A question mark is used after a question. E.g. ما اسمكَ؟ *mā ’ismuka?* “What’s your name?”
1	Exclamation mark (!) علامة تعجب *‘alāma^{tu} ta’ağğub*	**e**	An exclamation mark is used after an exclamation. E.g. ما أجملَ الرَّبيعَ! *mā ’ağmala ar-rabī ‘a!* “What a beautiful spring!”
1	Ellipsis mark (…) علامة حذف *‘alāma^{tu} ḥaḏf*	**i**	An ellipsis mark is used to mark an ellided word or phrase in a text. E.g. (جاءَ المعَلِّمُ وبَدَأَ …) *ğā’ al-mu‘alimu wa bada’a …* “ the teacher came and stared …”
1	Continuation mark (=) علامة التَّابعية *‘alāma^{tu} at-tabi‘yya^{h}*	**f**	A continuation mark is used in a footnote to indicate that the text has to be continued on another page.

6.2.7 Morphological Feature of Gender

Arabic classifies nouns according to gender into three classes[50]; nouns which are only masculine (مُذَكَّر) *muḏakkar*, nouns which are only feminine (مُؤَنَّث) *mu’annaṯ*, and nouns which are both masculine and feminine (common gender or neuter gender) (مُذَكَّر أَوْ مُؤَنَّث) *muḏakkar ’aw mu’annaṯ* such as; ملح *milḥ* ‘salt’, and روح *rūḥ* ‘spirit’ (Wright 1996). Figure 6.9 shows the morphological feature of gender subcategories. Table 6.7 lists the 3 subcategories, with examples of masculine, feminine and of common gender words. The morphological feature of gender is repsented at position 7 in the tag string.

[50] According to Wright’s (1986) classification. Ryding (2005) classifies nouns according to gender into two classes; *masculine* and *feminine*, and the “*dual gender noun*” is mentioned in a footnote on page 119.

Table 6.7 Examples of gender category attributes for nouns, verbs, adjectives and pronouns

#	Subcategories of gender	T	Examples			
			Noun	**Verb**	**Adjective**	**Pronoun**
1	Masculine مذكر *mudakkar*	**m**	كتاب *kitāb* book	يكتبون *yaktubūn* They are writing (Pl. / Masc.)	كاتب *kātib* writer (Sing. / Masc.)	هو *huwa* He
2	Feminine مؤنث *mu'annat*	**f**	مكتبة *maktaba*^{h} library	تكتبين *taktubīn* You are writing (sing. / Fem.)	كاتبة *kātiba*^{h} writer (Sing. / Fem.)	هي *hiya* She
3	Common gender مذكر أو مؤنث *mudakkar 'aw mu'annat*	**x**	ملح *milḥ* salt	نكتب *naktubu* We are writing (Pl. / Masc. or Fem)	[51]نائب *nā'ib* Parliament member (Sing./ Masc. or Fem.)	هما *humā* They (Dual)

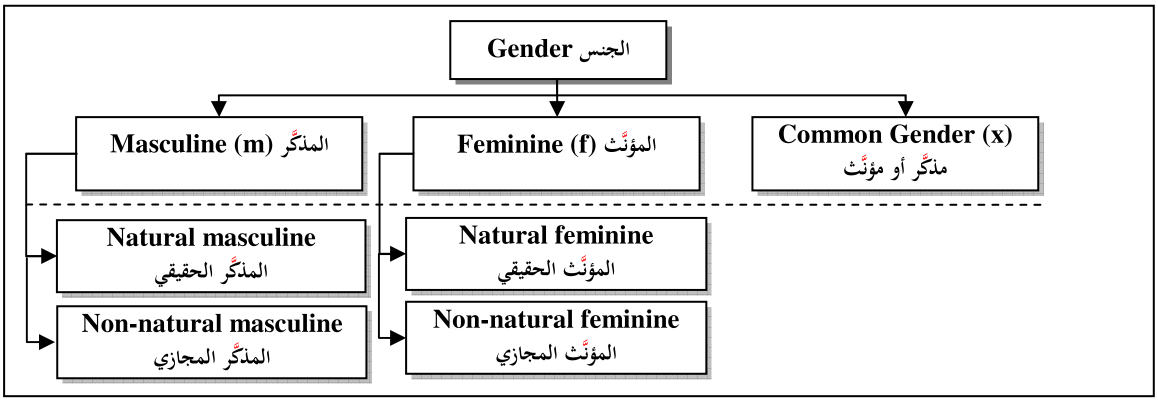


Figure 6.9 Arabic classification of nouns according to gender, with letter at position 7

Morphologically the masculine form is the simplest and most basic shape (word structure), whereas feminine nouns usually have a suffix that marks their gender. On the other hand, semantically, nouns are arbitrarily classified into masculine or feminine, except where a noun refers to a human being or other creature, when it is normally conforms to natural gender (Ryding 2005). Therefore, we can distinguish between two types of the morphological feature of gender that nouns can indicate: semantic gender where nouns indicate natural gender of humans, animals or things (male or female) whether the gender is a true characteristic of the human being or animal, or it is figurative for things that do not have natural gender. Morphological gender is defined if the noun is in its simplest form or if it contains a feminine suffix attached to it. Discussion of the detailed classifaction of the morphological feature of gender into morphological gender and semantic gender is beyond the scope of this thesis.

[51] Recently the word نائب *nā'ib* is being used for both masculine and feminine as the regular feminine form of this word نائبة *nā'iba*^{h} means *disaster*, which not suitable to indicate *feminine parliament member*.

6.2.8 Morphological Feature of Number

Singular, dual and plural are number morphological features identified in traditional Arabic grammar books. Singular applies for one entity of a category. Dual applies to "two" entities of a category, and plural applies to three or more entities. Number applies to nouns, adjectives, pronouns and verbs (*i.e.* the doer or the subject of verb). Other morphological categories, namely gender and rationality, affect the formation of the plural of nouns, particles or adjectives (Ryding 2005). Table 6.8 gives examples of singular, dual and plural words.

We distinguish between two types of plural: the sound plural جمع سالم *ğam' sālim* and the broken plural جمع تكسير *ğam' taksīr*. Sound plurals take specific suffixes to form the plural of certain masculine and feminine nouns. Broken plurals of nouns, by contrast do not follow regular rules but take one of a number of templatic patterns. For instance the word كِتَابٌ *kitāb* 'book', has the plural كُتُبٌ *kutub*^{un} 'books' following the templatic pattern فُعُلٌ *fu'ul*^{un}. Broken plurals are formed by adding letters to the singular form, by deleting letters from the singular form, or by changing the short vowels of the singular form. The plural of paucity جمع قلة *ğam' qilla*^{h} indicates few instances of a certain entity or type, while the Plural of Multitude جمع كثرة *ğam' kaṯra*^{h} indicates any number of instances more than three of a certain entity or type. The Ultimate plural منتهى الجموع *munthā al-ğumū'* is kind of Plural of Multitude but it follows only certain patterns. The Ultimate plural has an added infix *'alif* added to generate the broken plural from its corresponding singular noun followed by two consonants, or three consonants where the middle letter is silent (not followed by a vowel). Sometimes a broken plural can be further pluralized by a sound plural. If the broken plural is rational then the plural takes masculine plural suffixes, while, if it is an irrational broken plural, the feminine plural suffix is used to form the plural of the plural جمع الجمع *ğam' al- ğam'*, *e.g.* بُيوتات *buyūtāt* 'houses', which is formed by adding the feminine plural suffix ات *āt* to the broken plural بُيوت *buyūt* 'houses', which has the singular بيت *bayt* 'house'.

The category 'undefined' in the parser indicates cases where it is hard to guess the morphological feature of number of a particular word. For example, in the sentence كَتَبَ الطَّالِبُ الدَّرْسَ *katab aṭ-ṭālibu ad-darsa* 'the student wrote the lesson', the verb كَتَبَ *kataba* 'wrote' is singular and there is agreement between the verb and the subject of the sentence الطَّالِبُ *aṭ-ṭālibu* 'the student', which is also singular. On the other hand, in the sentence كَتَبَ الطَّالِبَانِ الدَّرْسَ *katab aṭ-ṭālibān ad-darsa* 'the two students wrote the lesson', the verb كَتَبَ *kataba* 'wrote' is singular while the subject الطَّالِبَانِ *aṭ-ṭālibān* 'the two students', is dual. The sentence كَتَبَ الطّلابُ الدَّرْسَ *kataba aṭ-ṭullābu ad-darsa* 'the students wrote the lesson', similarly has no agreement in gender between the singular form of the verb كَتَبَ

kataba ‘wrote’ and the plural form of the subject الطُّلاَّبُ *aṭ-ṭullābu* ‘the students’. The attribute ‘undefined’ is added to the number category of the verb to mark these cases.

Table 6.8 shows examples of the number category of nouns, verbs, adjectives and pronouns and illustrates the effects of the gender and humanness in the formation of the plural. Figure 6.10 shows the attributes of the morphological feature of number, represented at position 8 in the tag string.

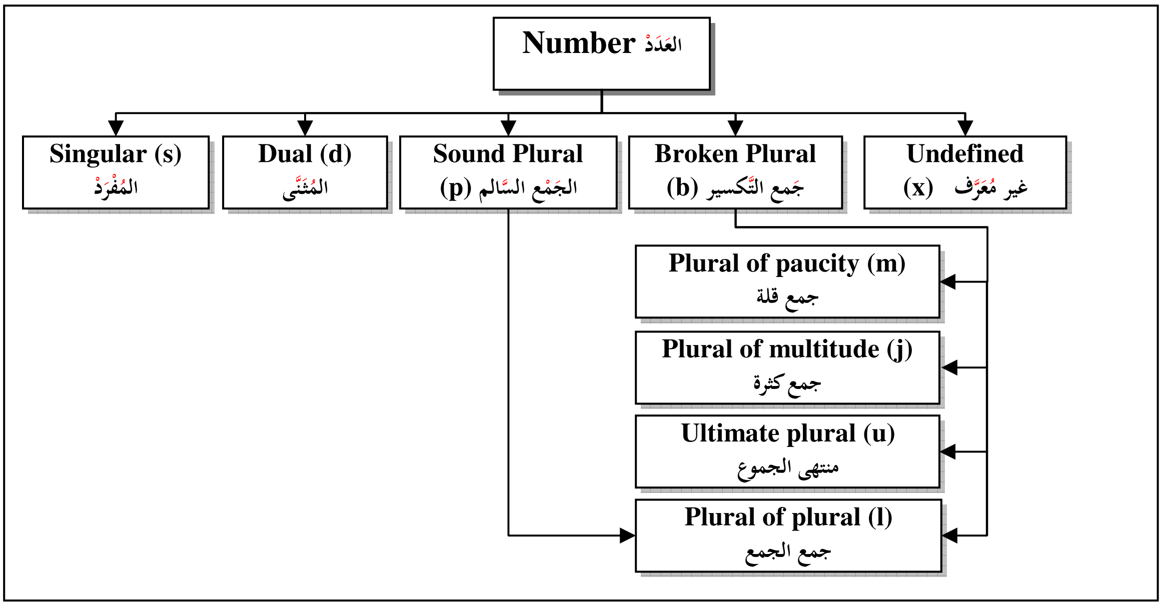

Number العَدَدْ
Singular (s) المُفْرَدْ
Dual (d) المُثَنَّى
Sound Plural (p) الجَمْع السَّالم
Broken Plural (b) جَمع التَّكسير
Undefined (x) غير مُعَرَّف
Plural of paucity (m) جمع قلة
Plural of multitude (j) جمع كثرة
Ultimate plural (u) منتهى الجموع
Plural of plural (l) جمع الجمع

Figure 6.10 Morphological feature of number category attributes, with letter at position 8

Table 6.8 Examples of the morphological feature category of Number

Category	**Noun**	**Verb**	**Adjective**	**Pronoun**[52]
Singular (s)	قَلَمٌ *qalam*^{un} pen (Masculine) وَرَقة *waraqa*^{h} paper (Feminine)	قَرَأَ *qara'a* he read قَرَأتْ *qara'at* she read	جَميل *ǧamīl* beautiful (masculine, singular) جَميلة *ǧamīla*^{h} beautiful (feminine, singular)	هو *huwa* he هي *hiya* she
Dual (d)	قَلَمانِ *qalamani* two pens(masculine) وَرَقتانِ *waraqatani* two papers (feminine)	يَقرآنِ *yaqra'āni* they (two) are reading (masculine) تقرآنِ *taqra'āni* they (two) are reading (feminine)	جَميلان *ǧamīlāni* beautiful (masculine, dual) جَميلتان *ǧamīlatān* beautiful (feminine, dual)	هما *humā* they (Common gender, dual)
Sound plural (p)	مراسلون *murāsilūn* agents (masculine) مُراسلات *murāsilāt* agents (feminine)	يَقرؤون *yaqra'ūn* they are reading (masculine) يقرأنَ *yaqra'na* they are reading (feminine)	جَميلون *ǧamīlūn* beautiful (masculine, plural) جَميلات *ǧamīlāt* beautiful (feminine, plural)	------------
Broken plural (b)	نِساء *nisā'* women عرب *'arab* Arabs	------------	كِبَار *kibār* senior (masculine, plural)	هم *hum* they (M) هُنَّ *hunna* they (F)
Plural of paucity (m)	أَبوابٌ *'abwāb*^{un} doors	------------	------------	------------
Plural of multitude (j)	كُتُبٌ *kutub*^{un} books	------------	رُكَّعٌ *rukka'*^{un} people who bow to the ground	------------
Ultimate plural (u)	مساجد *masāǧid* mosques	------------	------------	------------
Plural of plural (l)	رجالات *riǧālāt* men	------------	------------	------------
Undefined (x)	------------	كَتَبَ الطَّالِبُ الدَّرْسَ *katab aṭ-ṭālibu ad-darasa* 'the student wrote the lesson'; كَتَبَ الطَّالِبَانِ الدَّرْسَ *katab aṭ-ṭālibān ad-darsa* 'the two students wrote the lesson'; كَتَبَ الطَّلابُ الدَّرْسَ *kataba aṭ-ṭullābu ad-darsa* 'the students (plural) wrote the lesson'	------------	------------

[52] The number category applies to pronouns. They can be classified into singular, dual, and broken plural even though they are not templatic.

6.2.9 Morphological Feature of Person

Arabic has three main person attributes; first person المتَكَلِّم *al-mutakallim*, second person المخَاطَب a*l-muẖāṭab* and third person الغَائِب *al-ḡā’ib*. First person refers to the person or people speaking. The second person refers to the person or people who are present and sharing the talk or speech. The third person addresses the person or people who are absent and do not participate in the talk or speech (Ryding 2005).

The person category is affected by other morphological feature categories namely; gender and number. Thirteen personal pronouns and verb forms of person category, which are affected by gender and number, can be distinguished. There is no gender distinction in the first person but two forms of first person; singular and plural which is used as dual as well. There are five forms of second person; masculine singular, feminine singular, dual (masculine or feminine), masculine plural and feminine plural. The third person distinguishes between six forms of personal pronouns or verbs; masculine singular, feminine singular, masculine dual, feminine dual, masculine plural and feminine plural (Ryding 2005).

Table 6.9 shows the three main category attributes of person and how they are affected by gender and number categories with examples of both verbs and personal pronouns. Figure 6.11 shows the attributes of the morphological feature of person, represented at position 9 in the tag string.

Table 6.9 The three main attributes of person category with examples

	Person	**First Person (f)**		**Second Person (s)**		**Third person (t)**	
Number	**POS** **Gender**	**Personal pronoun**	**Verb**	**Personal pronoun**	**Verb**	**Personal pronoun**	**Verb**
Singular	**Masculine**	أنا *’anā* I	كَتَبتُ *katabtu* I wrote	أَنْتَ *’anta* you	كَتَبتَ *katabta* you wrote	هُوَ *huwa* he	كَتَبَ *kataba* he wrote
	Feminine			أَنْتِ *’anti* you	كَتَبتِ *katabti* you wrote	هِيَ *hiya* she	كَتَبَتْ *katabat* she wrote
Dual	**Masculine**	نَحْنُ *naḥnu* we	كَتَبنا *katabnā* we wrote	أَنْتُما *’antumā* you	كَتَبْتُما *katabtum ā* you wrote	هُما *humā* they	كَتَبا *katabā* they wrote
	Feminine						كَتَبَتا *katabatā* they wrote

	Person	First Person (f)		Second Person (s)		Third person (t)	
Number	POS Gender	Personal pronoun	Verb	Personal pronoun	Verb	Personal pronoun	Verb
Plural	Masculine	نَحْنُ *naḥnu* we	كَتَبنا *katabnā* we wrote	أَنْتُم *’antum* you	كَتَبتوا *katabtū* you wrote	هُم *hum* they	كَتَبُوا *katabū* they wrote
	Feminine			أَنْتُنَّ *’antunna* you	كَتَبتنَّ *katabtunna* you wrote	هُنَّ *hunna* they	كَتَبنَ *katabna* they wrote

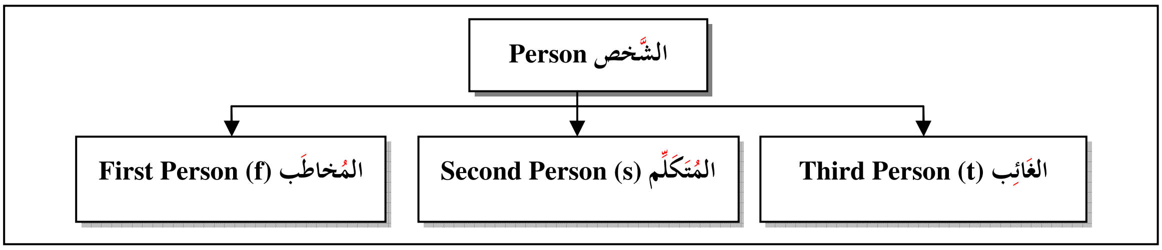

Person الشَّخص
First Person (f) المُخاطَب
Second Person (s) المُتَكَلِّم
Third Person (t) الغَائِب

Figure 6.11 Morphological feature of person category attributes, with letter at position 9

6.2.10 Morphological Feature Category of Inflectional Morphology

Inflectional morphology الصَّرف *aṣ-ṣarf* is an important feature of most Arabic word. Words are classified according to inflectional morphology into (i) invariable مبني *mabnī* or (ii) declined or conjugated معرب *mu‘rab*. Declined or conjugated words معرب *mu‘rab* are defined as these words which are affected by their preceeding word in context. The affect causes a change in case or mood of the word, changing its case or mood mark. By contrast, invariable words مبني *mabnī* are defined as words that do not change their case or mood marks in context, although they preceeded by words that otherwise have an effect on the following words in context (Dahdah 1987; Al-Ghalayyni 2005).

A declined or conjugated word can be an imperfect verb, e.g. يكتب *yaktubu* ‘he is writing’, and most nouns such as السَّماء *as-samā’* ‘the sky’, الأَرْض *al-‘arḍ* ‘the earth’ and الرَّجُل *ar-rağul* ‘the man’. An invariable word can be any particle, past and imperative verbs, and some nouns such as قَدْ *qad* ‘already or perhaps’, كَتَبَ *kataba* ‘he wrote’, اُكْتُبْ*’uktub* ‘write (order)’, هذه *hāḏihi* ‘this (fem.)’, أينَ*‘ayna* ‘where’, and مَنْ *man* ‘who’ (Dahdah 1987; Al-Ghalayyni 2005).

Most nouns are declined an exception being some nouns that are similar to particles. For example, pronouns are indeclinable nouns. Declined nouns are classified into (i) triptote or fully declined منصرف *munṣarif,* and (ii) diptote or non-declinable ممنوع من الصَّرف *mamnū’ min aṣ-ṣarf*. Triptote or fully declined nouns are regular nouns which change their case in context affected by the preceding word. The case mark can be any short vowel, *tanwīn* or a letter such as, *’alif* and *yā’*. Diptote or non-declinable nouns by contrast, cannot accept *tanwīn* or *kasra*^{h} as case mark; for example, أَحْمَدُ *'aḥmadu* 'Ahmad', يَعقوب *ya'qūba* 'Jacob', and عَطْشَانُ *'aṭšānu* 'thirsty' (Dahdah 1987; Al-Ghalayyni 2005).

Figure 6.12 shows the attributes of the morphological feature of Inflectional Morphology. Table 6.10 lists examples and definitions of the 4 attributes of the morphological feature category of Inflectional Morphology, represented at position 10 in the tag string.

Table 6.10 Examples of the morphological feature category of Inflectional Morphology

POS	**Morphology attributes**		**Examples**
Noun الأسم *al-'ism*	Invariable **(s)** مبني *mabnī*	An Invariable noun does not change its case marks in context. Although it is preceded by special words that have effects on the following words. E.g. Pronouns أَنْتُم *'antum* 'You (second person, plural)'.	
	Declined مُعرب *mu'rab*	Triptote / fully declined **(v)** مُنصَرِف *munṣarif*	Triptote or fully declined nouns are regular nouns which change their case in context due to the effect of the preceding word. E.g. السَّمَاء *as-samā'* 'the sky', الأَرْض *al-'arḍ* 'the earth', الرَّجُل *ar-rağul* 'the man'.
		Diptote / non-declined **(p)** ممنوع من الصَّرف *mamnū' min aṣ-ṣarf*	Diptote or non-declined nouns can not accept *tanwīn* or *kasra*^{h} as case mark , e.g. أَحْمَدُ*'aḥmadu* 'Ahmad', يَعقوب *ya'qūba* 'Jacob', عَطْشَانُ *'aṭšānu* 'thirsty'.
Verb الفعل *al-fi'l*	Invariable **(s)** مبني *mabnī*	An invariable مبني *mabnī* verb is defined as a word that does not change its mood marks in context. كَتَبَ *kataba* 'he wrote', and أُكْتُب *'uktub* 'write (order)'.	
	Conjugated **(d)** مُعرب *mu'rab*	A conjugated verb is affected by the preceding word in context. E.g. يَكتبُ *yaktubu* 'he is writing'. لَن يَكتُبَ *lan yaktuba* 'he will not write'. لَم يَكتُبْ *lam yaktub* 'he did not write'	

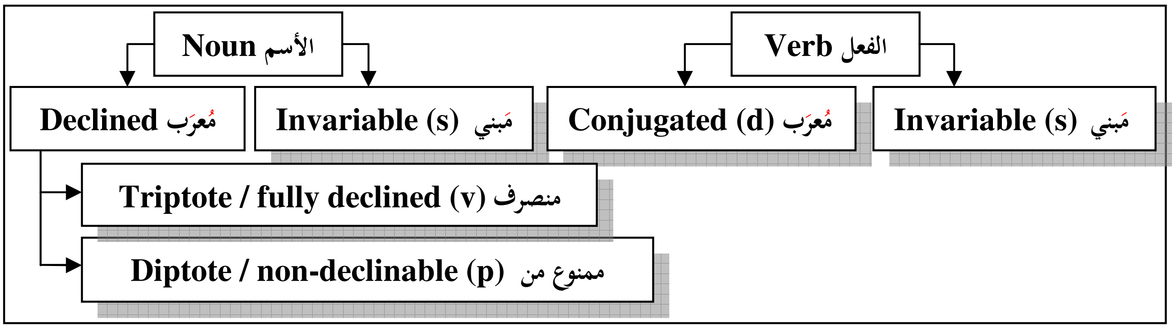


Figure 6.12 The morphological feature subcategories of Morphology attributes, with letter at position 10

6.2.11 Morphological Feature Category of Case or Mood

Case or mood is the morphological feature that determines the appropriate ending of a word, whether the word ends with a letter, short vowel or *tanwīn*. Case applies to nouns, and mood applies to verbs; since a word cannot be a noun and verb at the same time, no word can have both case and mood, they are mutually exclusive. So, we used position 11 to encode both case of noun and mood of verb. Case الحالة الإعرابيَّة للاسم *al-ḥāla*^{h} *al-'i'rābiyya*^{h} *lil'ism* is a morphological feature which applies to nouns and the subclasses of noun such as adjectives. There are three attributes of the case category: nominative مرفوع *marfū'*, genitive مجرور *mağrūr* and accusative منصوب *manṣūb*. Case marks are short vowel suffixes; *ḍamma*^{h} ضمَّة (ـُ) /u/ for nominative, *kasra*^{h} كسرة (ـِ) /i/ for genitive and *fatḥa*^{h} فتحة (ـَ) /a/ for accusative; with some exceptions to these general rules. Case is classified under morphology because it is part of word structure. Case is also classified under syntax because it is determined by the syntax of the sentence or clause. Subjects are marked by nominative case, direct objects of transitive verbs are marked by accusative case, and the object of a preposition and the possessor in a possessive structure are marked by genitive case (Ryding 2005).

Mood الحالة الإعرابيِّة للفعل *al-ḥāla*^{h} *al-'i'rābiyya*^{h} *lilfi'l* is a morphological feature which applies to verbs. There are three attributes of this category, namely indicative الرَّفِع *ar-raf'*, subjunctive النَّصِب *an-naṣb* and imperative or jussive الجَزِم *al-ğazm*. Straightforward statements or questions involve the indicative mood, whereas the subjunctive mood indicates an attitude toward the action (doubt, desire, wishing, necessity), and the imperative or jussive mood indicates an attribute of command or need (Ryding 2005). Imperative here describes the mood of the verb, while in section 6.2.3 imperative describes a verb category.

Like case, mood is classified under morphology because it is reflected in word structure. Mood is indicated by suffixes attached to the end of the verb stem. Mood is marked by *ḍamma*^{h} ضمَّة (ـُ) /u/ to indicate the indicative mood, marked by *fatḥa*^{h} فتحة (ـَ) /a/ to indicate the subjunctive mood, and by *sukūn* سُكون (ـْ) to indicate the imperative or jussive mood. Mood marking is determined by particular particles or by narrative context. This marking applies only to imperfect and imperative verbs. Perfect verbs do not have mood (Ryding 2005).

EAGLES guidelines for morphosyntatic annotation recommended putting attributes under part-of-speech headings. The standard requirement for these attributes/values is that it is advisable that the tag set of that language *should* encode them. The recommended attributes include type of noun, gender, number, case, person, definiteness, verb form / mood, tense, voice, status, degree, possessive, category of pronouns, and type for pronoun, determiner, article, adposition, conjunctions, numerals, and residuals. Case is a recommended attribute for nouns (N), adjectives (AJ), pronouns and determiners (PD), articles (AT) and numerals (NU). Table 6.11 shows the different attribute values of the case under each part-of-speech heading recommended by EAGLES. Mood or verb form is a recommended attribute specified for verbs. EAGLES guidlines distinguishes between eight attributes of mood for European languages. These values are indicative, subjunctive, imperative and conditional which are applicable to finite verbs, and infinitive, participle, gerund and supine which are applicable for non-finite verbs.

Table 6.11 The different attribute values of Case under each part-of-speech heading, as recommended by EAGLES

Part of Speech	**Attributes of Case**
Nouns (N)	1. Nominative 2. Genitive 3. Dative 4. Accusative 5. Vocative
Adjectives (AJ)	1. Nominative 2. Genitive 3. Dative 4. Accusative
Pronouns and Determiners (PD)	1. Nominative 2. Genitive 3. Dative 4. Accusative 5. Non-genitive 6. Oblique
Articles (AT)	1. Nominative 2. Genitive 3. Dative 4. Accusative
Numerals (NU)	1. Nominative 2. Genitive 3. Dative 4. Accusative

Case and mood are also important morphological features of an Arabic word. A good morphosyntatic annotation of Arabic text *should* include the case or mood of the word and the two main attributes associated with it, namely, the morphological feature of Inflectional Morphology and the morphological feature of Case and Mood Marks. For morphosyntatic annotation of Arabic text, these three morphological feature categories are obligatory attributes. Specifying the attributes of these morphological feature categories is a major topic of linguistic and grammatical studies of morphology and syntax of Arabic.

" ... الصرف والاعراب

للكلمات العربية حالتان: حالةُ إفرادٍ وحالة تركيب.

فالبحثُ عنها، وهي مُفردةٌ، لتكون على وزن خاصٍّ وهيئة خاصة هو من موضوع "علم الصرف".

والبحثُ عنها وهي مُركبةٌ، ليكونَ آخرُها على ما يَقتضيه مَنهجُ العرب في كلامهم – من رفعٍ، أو نصبٍ، أو جرّ، أو جزمٍ، أو بقاءٍ على حالةٍ واحدة، من تَغيُّر – هو من موضوع "علم الإعراب". ... " (Al-Ghalayyni, 2005 p.8)

" ... *Morphology and Syntax*

Arabic words have two states: stand alone words (out of context words) and in-context words.

Searching for an out-of-context word to specify its pattern and form is the subject of morphology علم الصرف *'ilm aṣ-ṣarf. And searching for a word in a contex to specify its case or mood according to the methods of Arabic grammar by determining the attribute of case or mood of the word such as nominative, accusative, genitive or jussive mood, or determing whether the word has only one state wherever it appears in context, is the subject of syntax, which is called* علم الإعراب *'ilm al- 'i'rāb ..."* (Al-Ghalayyni 2005 p.8)

Table 6.12 shows examples of Case or Mood attributes within sentences. Figure 6.13 shows the 6 attributes of the morphological feature of Case or Mood category, represented at position 11 in the tag string.

Table 6.12 Examples of morphological feature category of Case or Mood

Case or mood	**T**	**Example**
Case of noun الحالة الإعرابيَّة للاسم *al-ḥāla*^{tu} *al-'i'rābiyya*^{tu} *lil-'ism*		
Nominative مرفوع *marfū'*	**n**	Marked by *ḍamma*^{h} ضمَّة (ـُ) /u/. ذَهَبَ الطَّالِبُ الى المدرَسةِ *ḏahaba aṭ-ṭālibu 'ilā al-madrasati* 'The student went to the school'. The word الطَّالِبُ *aṭ-ṭālibu* 'The student' is the subject of the sentence and is in the nominative case.
Accusative منصوب *manṣūb*	**a**	Marked by *fatḥa*^{h} فتحة (ـَ) /a/. قَرَأَ الطَّالِبُ الدَّرسَ *qara'a at-talibu ad-darsa* 'The student read the lesson'. The word الدَّرسَ *ad-darsa* 'the lesson' is the direct object of the transitive verb قَرَأَ *qara'a* 'read', and is in the accusative case.
Genitive مجرور *maǧrūr*	**g**	Marked by *kasra*^{h} كسرة (ـِ) /i/. ذَهَبَ الطَّالِبُ الى المدرَسةِ *ḏahaba aṭ-ṭālibu 'ilā al-madrasati* 'The student went to the school'. The word المدرَسةِ *al-madrasati* 'the school' is the object of the preposition الى *'ilā* 'to' and is in the genitive case.
Mood of verb الحالة الإعرابيِّة للفعل *al-ḥāla*^{tu} *al-'i'rābiyya*^{tu} *lil-fi'l*		
Indicative **(n)** الرَّفِع *ar-raf'*	**n**	Marked by *ḍamma*^{h} ضمَّة (ـُ) /u/. يعمَلُ في الإدارَةِ *ya'malu fi al-'idarati* 'He works in administration'. The verb يعمَلُ *ya'malu* 'he works' is in the indicative mood.
Subjunctive النَّصِب *an-naṣb*	**a**	Marked by *fatḥa*^{h} فتحة (ـَ) /a/. يَجِبُ أنْ نَقومَ بِزيارَةٍ *yağibu 'an naqūma bi ziyārat*^{in} 'It is necessary that we undertake a visit'. The verb نَقومَ *naqwma* 'we undertake' is in the subjunctive mood because it is preceded by the subjunctive particle أنْ *'an*.
Imperative or jussive الجَزِم *al-ğazm*	**j**	Marked by *sukūn* سُكون (ـْ) or shortening of the final vowel of the verb if this vowel is otherwise long. إصلاحاتٌ لَمْ تَكمِلْ مُنذُ عامَينِ *'iṣlāḥāt lam taktamil mundu 'āmayni* renovations that haven't been completed for two years. لا تَنسَ! *lā tansa!* 'Don't forget!'. The verb تَكمِلْ *taktamil* 'completed' is in the jussive mood because it is been preceeded by the negative particle لَمْ *lam*. The verb تَنسَ *tansa* 'forget' is in the jussive mood, and is marked by shortening of the final vowel letter ى *'alif* of the original verb تَنْسى *tansā*.

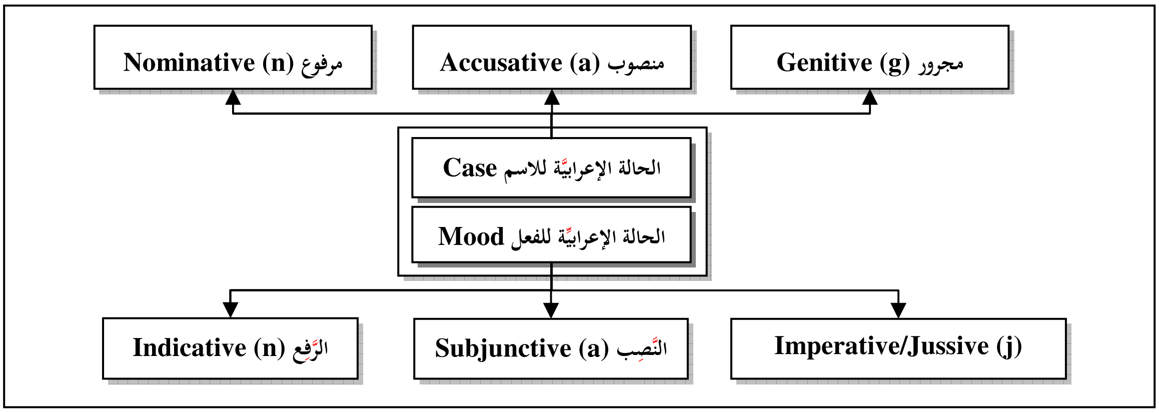


Figure 6.13 The morphological feature of Case or Mood, with letter at position 11

6.2.12 The Morphological Feature of Case and Mood Marks

The case or mood is an important morphological feature of the word. The case or mood of a word changes in context, and it is affected by the preceding words. The change of case or mood of the word affects the end of the word, by either change or omission of the word's last letter or the short vowel which appears on it. There are three kinds of case or mood marks; short vowel, letter or omission. The short vowels are *ḍamma*^{h} ضمّة (ـُ), *fatḥa*^{h} فتحة (ـَ) /a/ and *kasra*^{h} كسرة (ـِ) /i/. The letters are *'alif* (ا) /ā/, *nūn* (ن) /n/, *wāw* (و) /w/ and *yā'* (ي) /y/. Finally, omission is of three kinds; the deletion of the short vowel which is called *sukūn* سُكون (ـْ), the deletion of the vowel letter (*'alif*, *wāw*, *yā'*) and the deletion of the letter *nūn* (Al-Ghalayyni 2005).

The nominative case or indicative mood has four marks, *ḍamma*^{h} ضمّة, *wāw* (و), *'alif* (ا) and *nūn* (ن). The default mark for nominative case or indicative mood is *ḍamma*^{h} ضمّة. The accusative case or subjunctive mood has five marks; *fatḥa*^{h} فتحة, *'alif* (ا), *yā'* (ي), *kasra*^{h} كسرة and the deletion of letter *nūn*. The default mark is *fatḥa*^{h} فتحة. The genitive case has three marks; *kasra*^{h} كسرة, *'alif* (ا) and *yā'* (ي). The default mark is *kasra*^{h} كسرة. Finally, the imperative or jussive mood has three marks; *sukūn* سُكون, the deletion of the vowel letter (*'alif*, *wāw*, *yā'*) and the deletion of the letter *nūn* . The default mark is *sukūn* سُكون (Al-Ghalayyni 2005).

Table 6.13 shows examples of the 10 attributes of the Case and Mood Marks category. Figure 6.14 shows the 10 attributes of the morphological feature category of Case and Mood Marks, represented in position 12 of the tag string.

Table 6.13 Examples of each attribute of the Case and Mood Marks category

Case and Mood Mark			**T**	**Example**
Case (Noun)	Nominative مرفوع *marfū‘*	*ḍamma*^{h} ضمَّة	**d**	الصادقُ يُحَبُّ *yuḥabbu* ***aṣ-ṣādiqu*** ‘The honest (man) is loved’.
		wāw (و)	**w**	أفلحَ المؤمنونَ *aflaḥa* ***al-mu’minūna*** ‘The believers won’.
		’alif (ا)	**a**	يُكرَمُ التلميذان المجتهدان *yukramu* ***al-tilmīḏāni al-mujtahidāni*** ‘Both of the hardworking students are rewarded’.
	Accusative منصوب *manṣūb*	*fatḥa*^{h} فتحة	**f**	جانب الشَّرَّ فتسلَم *ğānib* ***aš-šarra*** *fa-taslam* ‘If you avoid evil, then you will be fine’
		’alif (ا)	**a**	أعطِ ذا الحقِّ حقهُ *’a‘ṭi* ***ḏā*** *al-ḥaqqi ḥaqqahu* “give the rightful man his right”
		yā’ (ي)	**y**	يحبُ الله المتقين *yuḥibbu ’allāhu al-muttaqīna* “God likes righteous people”
		kasra^{h} كسرة	**k**	أكرم الفتياتِ المجتهداتِ *’akrim* ***al-fatayāti*** *al-mujtahidāti* ‘reward the hardworking girls’
	Genitive مجرور *mağrūr*	*kasra*^{h} كسرة	**k**	تمسك بالفضائلِ *tamassak* ***bil-faḍā’ili*** ‘keep doing good deeds’
		yā’ (ي)	**y**	أطع أمر أبيك *’aṭi‘ ’amra* ***’abīka*** ‘obey your father’s order’.
		fatḥa^{h} فتحة	**f**	ليس فاعلُ الخيرِ بأفضلَ من الساعي فيه *laysa fā‘ilu al-ẖayri* ***bi-’afḍala*** *mina as-sā‘ī fīhi* “the one who does good deeds is not better that the one who help in them”
Mood (Verb)	Indicative الرَّفِع *ar-raf’*	*ḍamma*^{h} ضمَّة	**d**	يُحَبُّ الصادقُ ***yuḥabu*** *aṣ-ṣadiqu* ‘The honest (man) is loved’
		Inflectional *nūn* (ن)	**n**	تنطقون بالصدق ***tanṭiqūna*** *biṣ-ṣidqi* ‘You speak the truth’
	Subjunctive النَّصِب *an-naṣb*	*fatḥa*^{h} فتحة	**f**	يَجبُ أنْ نَقومَ بزيارَةٍ *yağibu ’an* ***naqūma*** *bi ziyāra*^{tin} ‘It is necessary that we undertake a visit’.
		deletion of *nūn*	**o**	لن تنالوا البرَّ حتى تُنفقُوا ممَّا تُحِبُّون *lan* ***tanālū*** *al-birra ḥattā* ***tunfiqū*** *mimmā tuḥibbūn* ‘You will not earn profit unless you spend what you like’
	Imperative or jussive الجَزِم *al-ğazm*	*sukūn* سُكون	**s**	إصلاحاتٌ لَمْ تَكمِلْ مُنْذُ عامَين *’iṣlāḥāt*^{un} *lam* ***taktamil*** *mundu ‘āmayni* ‘renovations that haven’t been completed for two years’.
		deletion of vowel letter حذف حرف العلّة	**v**	لا تَنْسَ! *lā* ***tansa!*** ‘Don’t forget!’.
		deletion of *nūn* حذف النون	**o**	قولوا خيراً تغنموا *qūlū ẖayr*^{an} ***tağnamū*** ‘If you speak well, you will get benefit’.

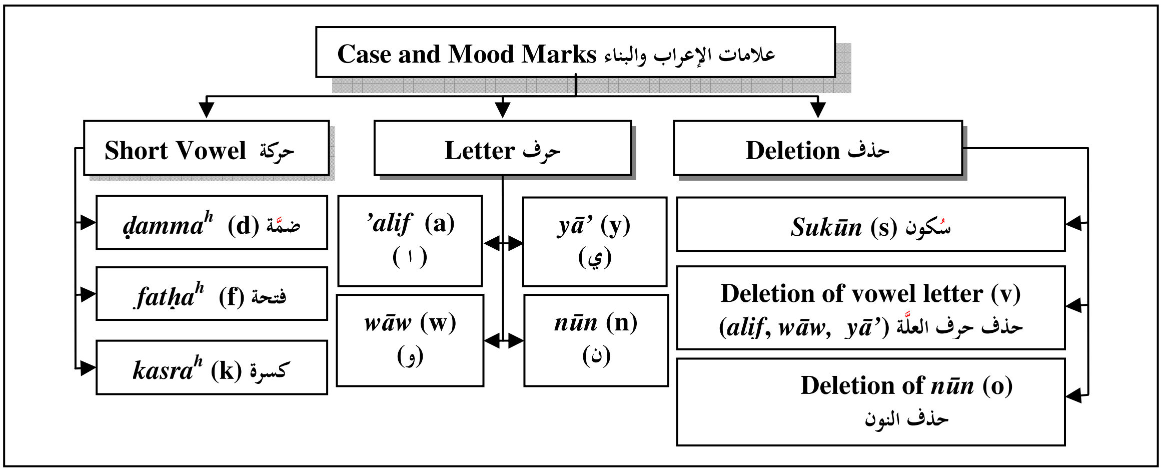


Figure 6.14 The morphological feature Case and Mood Marks, with letter at position 12

6.2.13 The Morphological Feature of Definiteness

Definiteness in Arabic has two attributes (markers); definiteness مَعْرِفَة *ma‘rifa*^{h} and indefiniteness نَكِرَة *nakira*^{h}. The prefix (ال) *alif-lām* (ال التعريف) is the definiteness prefix for nouns or adjectives; while the diacritical suffix (تنوين) *tanwīn* (ـٍ ، ـٌ ، ـً) /$^{-n}$/ is the indefiniteness suffix. The *tanwīn* is a diacritic mark which does not appear in non-vowelized text, while the definiteness mark, the definite article, (ال) *alif-lām* appears on definite nouns or adjectives in non-vowelized text (Ryding 2005).

Table 6.14 shows examples of the morphological feature of Definiteness. Figure 6.15 shows the 2 attributes of the morphological feature of Definiteness, represented at position 13 in the tag string.

Table 6.14 Examples of the morphological feature of Definiteness

	Definiteness	**T**	**Example**
1	Definiteness مَعْرِفَة *ma‘rifa*^{h}	**d**	البَيْت *al-bayt* ‘the home’. Is a definite noun marked with prefix (ال) *’alif-lām.*
2	Indefiniteness نَكِرَة *nakira*^{h}	**i**	بَيْتٌ *bayt*^{un} ‘home’. Is an indefinite noun marked with the diacritical suffix tanween (ـٌ)/^{un}/.

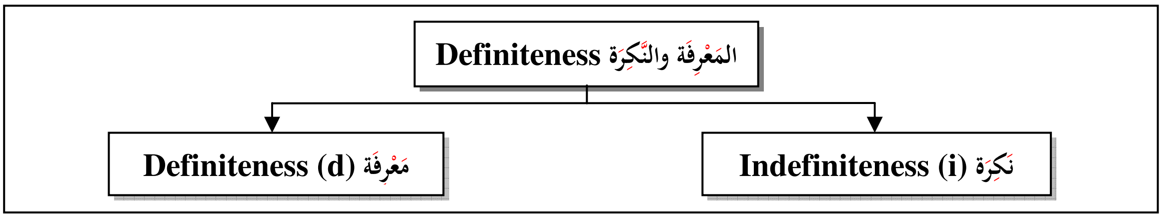

Definiteness المَعْرِفَة والنَّكِرَة
Definiteness (d) مَعْرِفَة
Indefiniteness (i) نَكِرَة

Figure 6.15 The morphological feature of Definiteness, with letter at position 13

6.2.14 Morphological Feature of Voice

Verbs in Arabic are either in the active voice مَبْنِي للمَعْلُوم *mabnī lil-ma'lūm* or the passive voice مَبْنِي للمَجْهُول *mabnī lil-maǧhūl*. The active voice standardly indicates that the doer of the action is the subject of the verb, while in the passive voice the subject of the verb is the direct object of the corresponding active, and the doer of the action (the active-voice subject) is unknown or not mentioned (Ryding 2005).

Table 6.15 shows examples of the 2 Voice category attributes in sentences. Figure 6.16 shows the 2 attributes of the morphological feature of Voice, represented at position 14 in the tag string.

Table 6.15 Examples of Voice category attributes in sentences

Voice	T	Example
Active مَبْنِي للمَعْلُوم *mabnī lil-ma'lūm*	**a**	كَتَبَ الطَّالِبُ الدَّرسَ ***kataba*** *aṭ-ṭālibu ad-darsa* 'The student **wrote** the lesson'. The verb كَتَبَ *kataba* 'wrote' is an active verb. The subject الطَّالِبُ *aṭ-ṭālibu* 'The student' appears in the sentence.
Passive مَبْنِي للمَجْهُول *mabnī lil-maǧhūl*	**p**	كُتِبَ الدَّرسُ ***kutiba*** *ad-darsu* 'The lesson **was written**'. The verb كُتِبَ *kutiba* 'was written' is a passive verb. The subject of the verb is the direct object الدَّرسُ *ad-darsu* 'The lesson'.

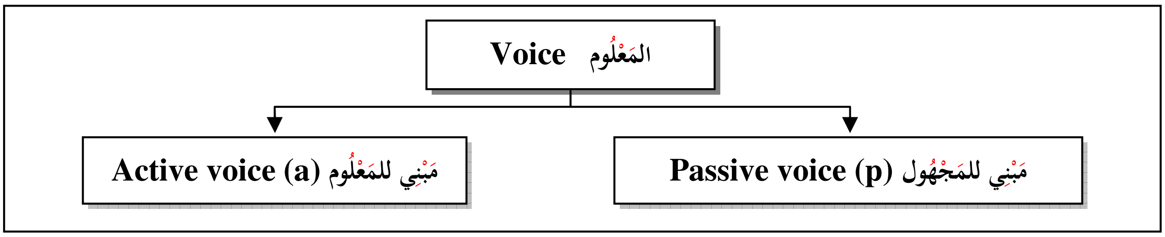

Voice المَعْلُوم
Active voice (a) مَبْنِي للمَعْلُوم
Passive voice (p) مَبْنِي للمَجْهُول

Figure 6.16 The morphological feature of Voice, with letter at position 14

6.2.15 Morphological Feature of Emphasized and Non-emphasized

The morphological feature of Emphasized and Non-emphasized المؤَكَّد وغيرُ المؤَكَّد *al-mu'akkad wa ḡayr al-mu'akkad* applies to verbs only. It has three attributes: non-emphasized غَيْرُ مُؤَكَّد *ḡayr mu'akkad* which applies to past or perfect verbs, obligatorily emphasized يَجب التأكيد *yağibu at-ta'kīd* and optionally emphasized مسموح التأكيد *masmūḥ at-ta'kīd*. Imperfect verbs must be emphasized in some circumstances when some conditions have been met such as: interrogation, wish, demand, encouragement, prevention, negation, and swearing. Emphasized verbs are marked by the suffix letter نْ /n/ added to the end of the verb stem; see table 6.5. There are two types of emphatic نْ /n/; one is the intensive *nūn* نّ /nn/ نون ثقيلة *nūn ṯaqīla*^{h} and the other is the non-intensive *nūn* نْ /n/ نون خفيفية *nūn ẖafīfa*^{h} (Dahdah 1987; Dahdah 1993).

Table 6.16 shows examples of Emphasized and Non-emphasized category attributes in sentences. Figure 6.17 shows the 2 attributes of the morphological feature of Emphasized and Non-emphasized, represented at position 15 in the tag string.

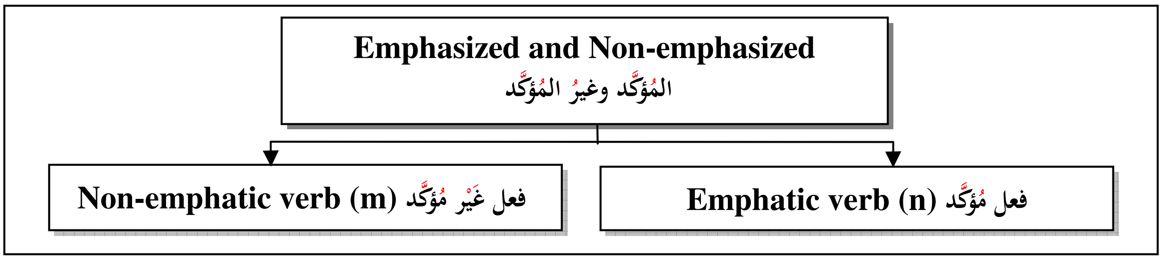

Emphasized and Non-emphasized
المُؤَكَّد وغيرُ المُؤكَّد
Non-emphatic verb (m) فعل غَيْر مُؤَكَّد
Emphatic verb (n) فعل مُؤَكَّد

Figure 6.17 The morphological feature of Emphasized and Non-emphasized, with letter at position 15

Table 6.16 Examples of the morphological feature Emphasized and Non-emphasized

Emphasized or Non-Emphasized	**T**	**Example**
Non-emphatic verb **فعل غَيْر مُؤَكَّد** *fi‘l g̅ayr mu’akkad*	**m**	ذَهَبَ الطَّالِبُ الى المِدرَسةِ ***d̲ahaba*** *aṭ-ṭalibu ‘ilā al-madrasati* ‘The student **went** to the school’. The perfect verb ذَهَبَ ***d̲ahaba*** ‘went’ is not emphasized.
Emphatic verb **فعل مُؤَكَّد** *fi‘l mu’akkad*	**n**	هَلْ تَذْهَبَنَّ؟ *hal tad̲habanna*? ‘Would you *go*?’ The verb تَذْهَبَنَّ *tad̲habanna* ‘go’ is emphasized. The suffix letter نّ /nn/ (النون الثقيلة) is added to the original verb تَذْهَبُ *tad̲habu* ‘go’.
		اذهَبْنَّ! *’id̲habnna* ‘Go!.’ The imperative verb اذْهَبْنَّ *’id̲habnna* ‘Go!’ is emphasized. The suffix letter نّ /nn/ (النون الثقيلة) is added to the original verb اذْهَبْ *’id̲hab* ‘go’.

6.2.16 The Morphological Feature of Transitivity

Verbs in Arabic are either transitive مُتعدِّي *muta‘addī* or intransitive لازِم *lāzim*. Intransitive verbs are verbs which give full meaning in a sentence without the need for an object. On the other hand, transitive verbs require an object to complete the meaning of the sentence. There are three types of transitive verbs. First, singly transitive مُتَعدِّي الى مَفعُول واحِد *muta‘addī ’ilā maf‘ūlin wāḥid* where there is only one object in the sentence. Second, doubly transitive verb مُتَعدِّي الى مَفْعُولَين *muta’addī ’ilā maf‘ūlayn* which requires two objects to complete the meaning in a sentence. Third, triply transitive verb مُتَعَدِّي الى ثَلاثَة مَفاعِيل *muta‘addī ’ilā t̲alāt̲ati mafā‘īl*, which require three objects to complete the meaning of a sentence; there are only seven of these verbs: أرى *’arā* ‘showed’, أَعلَمَ *’a‘lama* ‘notified’, حَدَّث *ḥaddat̲a* ‘narrated’, خَبَّرَ *h̲abbara* ‘informed’, أَخبَرَ *’ah̲bara* ‘gave information’, أَنْبَأَ

’anba’a, and نَبَّأَ *nabba’a* ‘advised’ ‘announced’ which share the meaning of telling or informing (Dahdah 1987; Dahdah 1993).

Table 6.17 shows examples of the 4 Transitivity category attributes in sentences. Figure 6.18 shows the 4 attributes of the morphological feature of Transitivity, represented at position 16 in the tag string.

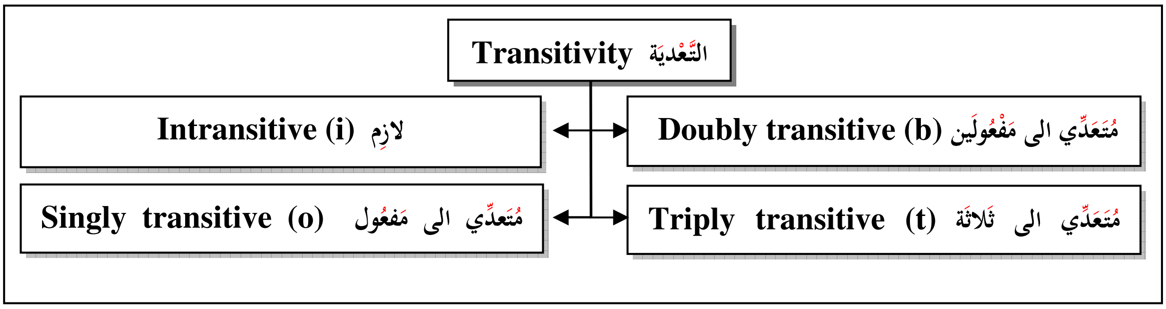

Transitivity التَّعْدِيَة
Intransitive (i) لازِم
Doubly transitive (b) مُتَعَدِّي الى مَفْعُولَين
Singly transitive (o) مُتعدِّي الى مَفعُول
Triply transitive (t) مُتَعَدِّي الى ثَلاثَة

Figure 6.18 The morphological feature of Transitivity, with letter at position 16

Table 6.17 shows examples of the Transitivity category attributes in sentences

Transitivity	**T**	**Example**
Intransitive verb لازِم *lāzim*	**i**	مَاتَ القَائِدُ ***māta*** *al-qā’idu* ‘The commander **has died**’. The verb مَاتَ *māta* ‘has died’ is an intransitive verb. The sentence is meaningful without the need for an object.
Singly transitive verb مُتَعدِّي الى مَفعُول واحِد *muta‘addī ’ilā maf‘ūlin wāḥid*	**o**	يَطْلُبُ البَاحِثُ المعْرِفَةَ ***yaṭlubu*** *al-bāḥṯu al-ma‘rifati* ‘The researcher **asks** for knowledge’. The verb يَطْلُبُ *yatlubu* ‘asks’ is a singly transitive verb. The sentence is not meaningful without the object المعْرِفَة *al-ma‘rifati* ‘knowledge’.
Doubly transitive verb مُتَعَدِّي الى مَفْعُولَين *muta’addī ’ilā maf‘ūlayn*	**b**	تَأْمُرُونَ النَّاسَ خَيْراً ***ta’murūna*** *an-nāsa ẖair*^{an} ‘You *order* people [to do] good’. The verb تَأْمُرُونَ *ta’muruuna* ‘order’ is a doubly transitive verb. The sentence is not meaningful without the first object النَّاس *an-nāsa* ‘people’ and the second object خَيْراً *ẖair*^{an} ‘for good’.
Triply transitive verb مُتَعَدِّي الى ثَلاثَة مَفاعِيل *muta‘addī ’ilā ṯalāṯati mafā‘īl*	**t**	أَرَى اللهُ المذنِبينَ أَعمَالَهم حَسَراتٍ ***’arā*** *allāhu al-mudnibīna ’a‘mālahum hasarāt*^{in} ‘God *shows* sinners what they did as repentances’. The verb أَرَى *’arā* ‘shows’ is a triply transitive verb. The sentence is not meaningful if any of the three objects are missing. المذنِبينَ *al-mudnibīna* ‘sinners’, أَعمَالَهم *’a’mālahum* ‘what they did’, and حَسَراتٍ *ḥasarāt*^{in} ‘repentances’.

6.2.17 The Morphological Feature of Rational

The morphological feature of rational describes the ability to be endowed with reason and comprehension, like human beings, angels and demons. The opposite is irrational. The morphological feature of “rational” or “rationality” differs from the linguistic concept of animacy because the latter divides nouns/entities into two categories: animate versus inanimate, while the former is used to denote human or human-like entities (*e.g.* djinn) at the top of the person hierarchy (Zaenen et al. 2004) and endowed with the faculty of reason as distinct from all other entities, whether animate or inanimate. Rational is a morphological feature which is applicable to some types of nouns such as singular proper nouns (names) اسم العلم المفرد *’ism al-‘alam al-mufrad*, demonstrative pronouns أسماء الإشارة *’asmā’ al-’išāra*^{h}, conditional nouns اسماء الشرط *’asmā’ aš-šarṭ* relative pronouns الأسماء الموصولة *al-’asmā’ al-mawṣūla*^{h}, interrogative pronouns أسماء الإستفهام *’asmā’ al-’istifhām* and allusive nouns الكناية *al-kināya*^{h} (Dahdah 1987; Dahdah 1993).

Table 6.18 shows the 2 attributes of the morphological feature Rational, with rational and irrational examples for these noun types. Figure 6.18 shows the noun types that have the Rational morphological feature, represented at position 17 in the tag string.

Table 6.18 Examples of the morphological feature category of Rational

Noun	**Rational**	**Irrational**
Singular proper name اسم العلم المفرد *’ism al-‘alam al-mufrad*	سمير *samīr* ‘Samir’, جبريل *ǧibrīl* ‘Gabriel’, إبليس *‘iblīs* ‘Satan’.	Irrational compound proper name such as; بَيْت لَحم *bayt laḥm* ‘Bethlehem’, بَعْلَبَك *ba’lbak* ‘Baalbak’.
Demonstrative pronouns أسماء الإشارة *’asmā’ al-’išāra*^{h}	أولئك *’ulā’ika* ‘hese’.	تلك *tilka* ‘that’.
Interrogation pronouns أسماء الإستفهام *’asmā’ al-’istifhām*	مَنْ *man* ‘who’, مَنْ ذا *man ḏā* ‘who is he’.	ما *mā* ‘that which’, ماذا *māḏā* ‘what’.
Conditional nouns اسماء الشرط *’asmā’ aš-šarṭ*	مَنْ *man* ‘who’.	ما *mā* that ‘which’. مهما *mahmā* ‘whatever’.
Relative pronouns الأسماء الموصولة *al-’asmā’ al-mawṣūla*^{h}	مَنْ *man* ‘who’.	ما *mā* ‘that which’.
Allusive nouns الكناية *al-kināya*^{h}	فُلان *fulān* (used to refer to rational singular masculine proper name)	------------------------

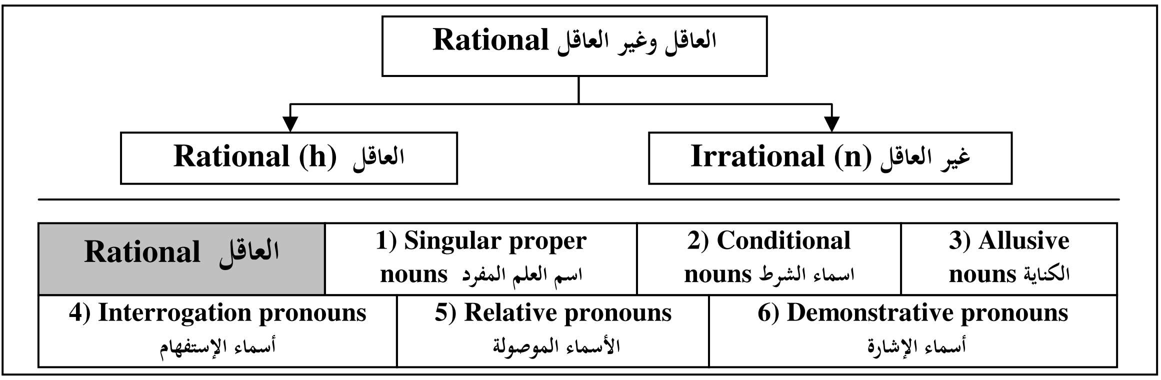

Rational العاقل وغير العاقل
Rational (h) العاقل
Irrational (n) غير العاقل
Rational العاقل
1) Singular proper nouns اسم العلم المفرد
2) Conditional nouns اسماء الشرط
3) Allusive nouns الكناية
4) Interrogation pronouns أسماء الإستفهام
5) Relative pronouns الأسماء الموصولة
6) Demonstrative pronouns أسماء الإشارة

Figure 6.19 Morphological feature category of Rational, with letter at position 17

6.2.18 The Morphological Feature of Declension and Conjugation

Declension means a class of nouns or adjectives having the same type of inflectional forms, and conjugation is the schematic arrangement of the inflectional forms of a verb[53]. In Arabic, both of the terms mean subject to change too. In Arabic grammatical terminology, declension and conjugation is put under the 'science' (area of enquiry) that describes the rules of word structure. It identifies the underlying letters of the word, the word's consonant letters and vowels. It also identifies which of the word's letters are changed during derivation. In addition, the meaning includes changing the word into different forms of different meanings, such as deriving the perfect verb الفعل الماضي *al-fiʿl al-maḍī*, imperfect verb الفعل المضارع *al-fiʿl al-muḍāriʿ*, imperative verb فعل الأمر *fiʿl al-'amr*, active participle اسم الفاعل *'ism al-fāʿil*, passive participle اسم المفعول *'ism al-mafʿūl*, relative noun الاسم المنسوب *al-'ism al-mansūb*, diminutive اسم التصغير *'ism at-taṣġīr* and others from the gerund المصدر *al-maṣdar* (Al-Ghalayyni 2005).

Nouns are classified into inflected nouns اسماء متصرِّفة *'asmā' mutaṣarrifa*^{h} and non-inflected nouns اسماء غير متصرِّفة *'asmā' ġayr mutaṣarrifa*^{h}. The inflected noun has number, *i.e.* it can be dual or plural as well as singular. It can be a relative noun اسم منسوب *'ism mansūb* or diminutive اسم مصغَّر *'ism muṣaġġar*. The non-inflected noun الاسم غير المتصرِّف *al-'ism ġayr al-mutaṣarrif*, by contrast has only one form which does not change in context. Non-inflected nouns include pronouns الضمائر *al-ḍamā'ir*, demonstrative pronouns أسماء الإشارة *'asmā' al-'išāra*^{h}, relative pronouns الأسماء الموصولة *al-'asmā' al-mawṣūla*^{h}, conditional nouns اسماء الشرط *'asmā' aš-šarṭ*, interrogative pronouns أسماء الإستفهام *'asmā' al-'istifhām*, allusive nouns الكناية *al-kināya*^{h}, adverbs الظَّروف *al-ẓurūf* and numerals اسماء الأعداد *'asmā' al-'aʿdād.*

Inflected nouns الاسماء متصرِّفة *al-'asmā' mutaṣarrifa*^{h} are classified into the derived nouns اسم مشتقّ *'ism muštaqq* and the primitive nouns اسم جامد *'ism ǧāmid*. The derived noun is derived from its verb; for example عالِم *ʿālim* 'scientist' and مُتعلِّم *mutaʿallim* 'learner' are derived from the verb عَلِمَ *ʿalima* 'knew' and تَعَلَّمَ *taʿallama* 'he learnt' respectively. Derived nouns includes 10 types of nouns; active participle اسم فاعل *'ism fāʿil* , passive

[53] Merriam Webester Dictionarry

participle اسم مفعول *’ism maf‘ūl*, adjective صفه مشبهة *ṣifa^h mušabbaha^h*, intensive active participle مبالغة اسم الفاعل *mubālaǧat ’ism al- fā‘il*, elative noun اسم تفضيل *’ism tafḍīl*, noun of time اسم زمان *’ism zamān*, noun of place اسم مكان *’ism makān*, gerund with initial *mīm* المصدر الميمي *al-maṣdar al-mīmī*, instrumental noun اسم آله *’ism al-’āla^h* and the gerund of the unaugmented verb consisting of more than three letters مصدر الفعل فوق الثلاثي المجرِّد *maṣdar al-fi‘l fawq al-ṯulāṯī al-muǧarrad* (Al-Ghalayyni 2005).

The primitive noun الاسم الجامد *al-’ism al-ǧāmid* cannot be derived from a verb. Examples are حجر *ḥaǧar* ‘stone’, سقف *saqf* ‘ceiling’ and دِرهَم *dirham* ‘Dirham (currency)’. They also include, the gerund of unaugmented triliteral verbs مصادر الأفعال الثلاثية المجرَّدة *maṣādir al-af‘āl al-ṯulāṯiyya^h al-muǧarrada^h* such as عِلْم *‘ilm* ‘science’ and قِراءة *qirā’a^h* ‘reading’ (Al-Ghalayyni 2005).

Verbs are classified into conjugated verbs أفعال متصرِّفة *af‘āl mutaṣarrifa^h* and non-conjugated verbs أفعال جامدة *af‘āl ǧāmida^h* according to whether the verb has a tense or not. Verb forms are changed to indicate the tense of an action; past tense, present tense and future tense. But if a verb does not indicate any tense or an action, then there is no need to change the verb form, because its meaning does not change when the tense or action changes. Only a change of tense or action requires changing the form of the verb to indicate different meanings in different tenses.

The non-conjugated verb الفعل الجامد *al-fi‘l al-ǧāmid* is similar to particles. It indicates an abstract meaning that has no tense or action. Therefore, the non-conjugated verb has only one form which does not change in any context. Non-conjugated verbs are either restricted to the perfect ملازم للماضي *mulāzim lil-maḍī* such as عسى *‘asā* ‘might’ and لَيْسَ *laysa* ‘not (negation)’, or restricted to the imperfect ملازم للمضارع *mulāzim lil-muḍāri‘* as in يَهِيطُ *yahīṭu* ‘scream’, or restricted to the imperative as in هَبْ *hab* ‘suppose’.

Finally, the conjugated verb الفعل المتصرَّف *al-fi‘l al-mutaṣarrif* indicates an action or tense. So, it accepts the changes of form which reflect the different meanings of different tenses. The majority of verbs belong to the class of fully conjugated verbs فعل تام التَّصريف *fi‘l tām at-taṣrīf* where the three types of signification are found as in كتب *katab* ‘he wrote’ (perfect), يَكْتُبُ *yaktunu* ‘he is writing’ (imperfect) and اكتُبْ *‘uktub* ‘write (imperative)’. The partially conjugated verb فعل ناقص التَّصريف *fi‘l nāqiṣ at-taṣrīf* has only two types of signification, *i.e.* either perfect and imperfect but not imperative as in كادَ *kāda* يَكادُ *yakādu* ‘[be] close near [to] or almost [to]’ and أوشكَ *’awšaka* يوشِكُ *yūšiku* ‘[be] about [to]’, or imperfect and imperative but not perfect as in يَدَعُ *yada‘u* ‘he leaves’, دَع *da‘* ‘leave’ and يَذَرُ *yaḏaru* ‘he leaves’ ذَرْ *ḏar* ‘leave’ (Al-Ghalayyni 2005).

Table 6.19 shows examples of the 9 attributes of the Declension and Conjugation morphological feature. Figure 6.20 shows the the classifications of nouns and verbs according to the Declension and Conjugation morphological feature, represented at position 18 in the tag string.

Table 6.19 Examples of the Declension and Conjugation morphological feature

Declension and Conjugation		**T**	**Examples**
Noun	Non-inflected غير مُتصَرِّف *ğayr mutaṣarrif*	**n**	The pronoun هُوَ *huwa* ‘he’
	Primitive / Concrete noun مُتَصَرِّف – جامِد– اسم ذات *mutaṣarrif – ğāmid – ’ism dāt*	**t**	The concrete noun is perceptible by one or more of the five senses and includes the generic noun إمرأة *‘imra’a^h* ‘woman’, the proper noun مِصْرَ *miṣra* ‘Egypt’, and some nouns of place and instrument: مِزْمَار *mizmār* ‘pipe’
	Primitive / Abstract noun مُتَصَرِّف – جامِد– اسم معنى *mutaṣarrif – ğāmid – ’ism ma‘nā*	**a**	The abstract noun is not preciptible by the five senses and includes the unaugmented gerund: شُرْبٌ *šurb^{un}* drinking, and some gerunds with initial ‘*mīm*’: مَطْلَبٌ *maṭlabun* ‘claim’
	Inflected / Derived noun مُتَصَرِّف – اسم مُشْتَقٌّ *mutaṣarrif - ’ism muštaqq*	**d**	عالِم *‘ālim* ‘scientist’ derived from the verb عَلِمَ *‘alima* ‘knew’ and مُتَعلِّم *muta’allim* ‘learner’ derived from the verb تَعَلَّمَ *ta’allama* ‘he learn’
Verb	Non-conjugated / restricted to the perfect فعل جامِد– ملازم للماضي *fi‘l ğāmid- mulāzim lil-māḍī*	**p**	عسى *‘asā* ‘might’ لَيْسَ *laysa* ‘not (negation)’
	Non-conjugated / restricted to the imperfect فعل جامِد– ملازم للمضارع *fi‘l ğāmid- mulāzim lil-muḍāri‘*	**c**	يَهِيطُ *yahīṭu* ‘scream’
	Non-conjugated / restricted to the imperative فعل جامِد– ملازم للأمر *fi‘l ğāmid- mulāzim lil-’amr*	**i**	هَبْ *hab* ‘suppose’
	Conjugated / fully conjugated verb مُتَصَرِّف – فعل تام التَّصريف *mutaṣarrif – fi‘l tāmm at-taṣrīf*	**v**	كتب *katab* ‘he wrote’, يَكْتُبُ *yaktubu* ‘he writes’ and اكتُبْ *‘uktub* ‘write’
	Conjugated / partially conjugated verb مُتَصَرِّف – فعل ناقص التَّصريف *mutaṣarrif –fi‘l nāqiṣ at-taṣrīf*	**m**	كادَ *kāda* يَكادُ *yakādu* ‘[be] close near [to] or almost [to]’ أوشكَ *‘awšaka* يوشِكُ *yūšiku* ‘[be] about [to]’, يَدَعُ *yada’u* ‘he leaves’ دَعْ *da’* ‘leave’ يَذَرُ *yadaru* ‘he leaves’ ذَرْ *dar* ‘leave’

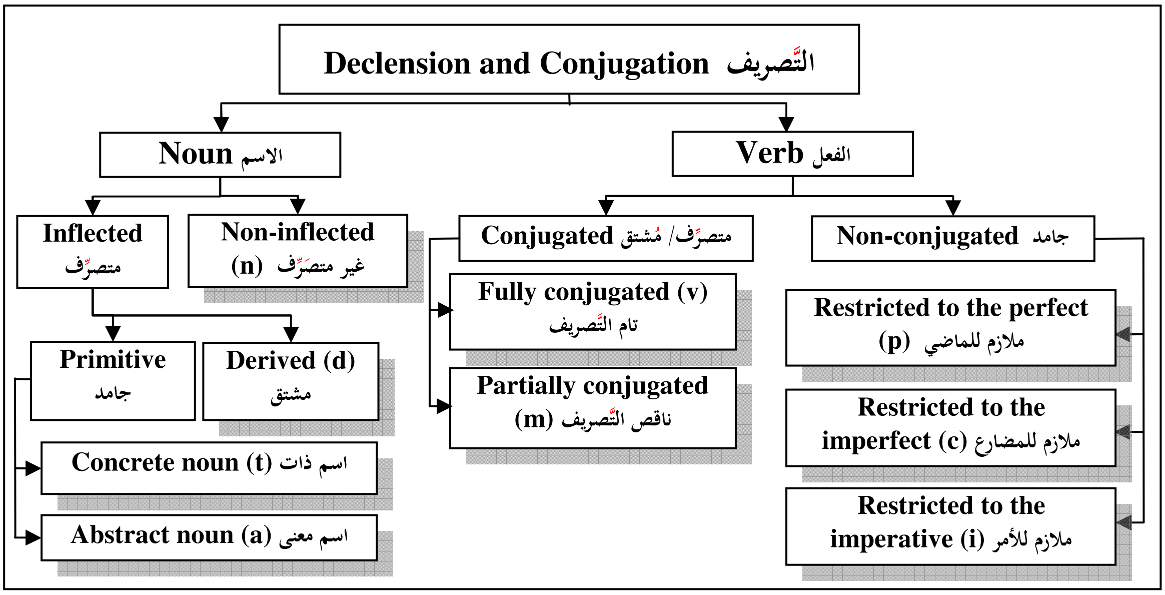

Declension and Conjugation التَّصريف
Noun الاسم
Verb الفعل
Inflected متصرِّف
Non-inflected (n) غير متصرِّف
Conjugated متصرِّف/ مُشتق
Non-conjugated جامد
Fully conjugated (v) تام التَّصريف
Partially conjugated (m) ناقص التَّصريف
Restricted to the perfect (p) ملازم للماضي
Restricted to the imperfect (c) ملازم للمضارع
Restricted to the imperative (i) ملازم للأمر
Primitive جامد
Derived (d) مشتق
Concrete noun (t) اسم ذات
Abstract noun (a) اسم معنى

Figure 6.20 The the classification of nouns and verbs according to the morphological feature of Declension and Conjugation, with letter at position 18

6.2.19 The Morphological Feature of Unaugmented and Augmented

Arabic verbs have roots consisting of three or four letters. From these roots many verbs can be derived by following certain patterns. There are many patterns for Arabic verbs. The standard way of determining the pattern of a verb is to refer to an Arabic lexicon or dictionary. Nonetheless, Arabic linguists have constructed general rules to extract these patterns. Verbs have two basic patterns consisting of three or four letters فَعَلَ *fa'ala* and فَعْلَلَ *fa'lala* respectively. Any verb derived following these two patterns is called an unaugmented verb (فعل مُجَرَّد) *fi'l muğarrad*. From فَعَلَ *fa'ala*; the basic triliteral pattern, 10 more patterns can be derived, and from فَعْلَلَ *fa'lala;* the basic quadriliteral pattern, 3 more patterns can be derived. These new patterns are derived by adding one, two or three letters to the basic patterns or by duplicating the second letter ع *'ayn* of the basic pattern. The group of letters that are added to the basic patterns to produce the other 13 patterns are; ا ، أ ، ت ، س ، ل ، م ، ن ، هـ ، و ، ي *(ā, ', t, s, l, m, n , h, w, y)* that combine with the word سألتمونيها *sa'altumūnīhā* 'you (second person, plural) asked me it (feminine, singular)' (Dahdah 1987; Dahdah 1993; Al-Ghalayyni 2005)**.**

Unagmented declineable nouns are either triliteral ثُلاثي *ṯulāṯī* such as حجر *ḥağr* 'stone', quadriliteral رُباعي *rubā'ī* such as جعفر *ğa'far* 'male proper name', or quinquiliteral خُماسي *ḫumāsī* such as سَفرجل *safarğal* 'quince [kind of fruit]'. A noun which consists of more than five letters is an augmented noun. A noun can be augmented by one letter مزيد بحرف *mazīd bi ḥarf* such as حصان *ḥiṣān* 'horse' (augmented by *ā* ا) and قنديل *qindīl* 'light' (augmented by *ī* ي), augmented by two letters مزيد بحرفين *mazīd bi ḥarfayn* such as مصباح *miṣbāḥ* 'lamp' (augmented by *m* م and *ā* ا), augmented by three letters مزيد بثلاثة أحرف *mazīd* *bi talāta*^{ti} *'aḥruf* such as انطلاق *'inṭilāq* 'starting' (augmented by ' ا, *n* ن and *ā* ا) and احرنجام *'iḥranğām* 'crowded' (augmented by ' ا, *n* ن and *ā* ا), or augmented by four letters مزيد بأربعة أحرف *mazīd bi 'arba'a*^{ti} *'aḥruf* such as استغفار *'istiğfār* 'asking for forgiveness' (augmented by ' ا, *s* س, *t* ت and *ā* ا).

Table 6.20 shows examples of the 5 Unaugmented and Augmented category attributes. Figure 6.21 shows the 5 attributes of the Unaugmented and Augmented category, represented at position 19 in the tag string.

Table 6.20 Examples of Unaugmented and Augmented category attributes

Unaugmented and Augmented	**T**	**Examples**		
		Triliteral verbs	**Quadriliteral verbs**	**Nouns**
Unaugmented **المُجَرَّد** *al-muğarrad*	**s**	فَتَحَ *fataḥa* 'he opened'.	دَحْرَجَ *daḥrağa* 'rolled'.	حجر *ḥağr* 'stone'. جعفر *ğa'far* 'a name'. سَفرجل *safarğal* 'quince, [kind of fruits]'
Augmented by one letter **مَزِيْد بِحَرف** *mazīd bi ḥarf*	**a**	يَفْتَحُ *yaftaḥu* 'he is opening. The letter (يَ) *yā* is added to the beginning of the verb stem فَتَحَ *fataḥa*	يُدَحْرِجُ *yudaḥriğu* 'he is rolling'. The letter (يَ) *yā* is added to the beginning of the verb stem دَحْرَجَ *daḥrağa.*	حصان *ḥiṣān* 'horse'. قنديل *qindīl* 'light'.
Augmented by two letters **مَزِيْد بِحَرفَيْن** *mazīd bi ḥarfayn*	**b**	انْكَسَرَ *'inkasara* ' has broken'. The letters ا *'alif* and نْ *nūn* are added to the beginning of the verb stem كَسَرَ *kasara* 'broke'.	يَتَدَحْرَجُ *yatadaḥrağu* ' is rolling'. The letters (يَ) *yā'* and تَ *tā'* are added to the verb stem دَحْرَجَ *daḥrağa* 'rolled'.	مصباح *miṣbāḥ* 'lamp'. احرنجام *'iḥranğām* 'crowded'
Augmented by three letters **مَزِيْد بِثَلاثَةِ حُرُوف** *mazīd bi talāta*^{ti} *ḥurūf*	**t**	اسْتَخْرَجَ *'istaẖrağa* has extracted. The letters ا *'alif,* س *sīn and* تَ *tā'* are added to the beginning of the verb stem خَرَجَ *ẖarağa* 'extracted'.	---------------------	انطلاق *'inṭilāq* 'starting'
Augmented by four letters **مزيد بأربعة أحرف** *mazīd bi 'arba'ati 'aḥruf*	**q**	---------------------	---------------------	استغفار *'istiğfār* 'asking for forgiveness'

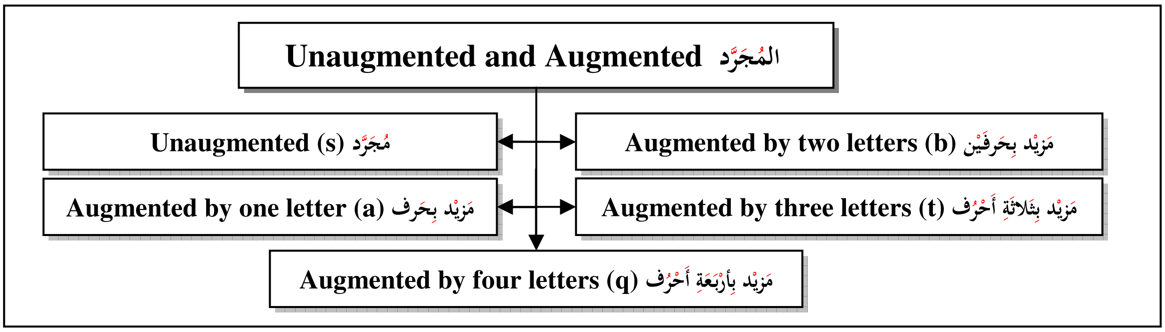

Unaugmented and Augmented المُجَرَّد
Unaugmented (s) مُجَرَّد
Augmented by two letters (b) مَزِيْد بِحَرفَيْن
Augmented by one letter (a) مَزِيْد بِحَرف
Augmented by three letters (t) مَزِيْد بِثَلاثَةِ أَحْرُف
Augmented by four letters (q) مَزِيْد بِأَرْبَعَةِ أَحْرُف

Figure 6.21 The Unaugmented and Augmented category attributes, with letter at position 19

6.2.20 The Morphological Feature of Number of Root Letters

> "*Root is a relatively invariable discontinuous bound morpheme, represented by two to five phonemes, typically three consonants in same order, which interlocks with a pattern to form a stem and which has lexical meaning*" (Ryding 2005)

Discontinuous means vowels can be interspersed between the root consonants *e.g* دَرَس *d-r-s* study. These consonants must always be present in the same sequence in the derived words first د /d/ then ر /r/ then س /s/ (Ryding 2005). Verbs, as mentioned in the previous section, have triliteral ثُلاثي *ṯulāṯī* or quadriliteral رُباعي *rubāʿī* roots. The general Arabic rule is that any noun with less than three letters or more than five letters then either has letters deleted from it or added on (Dahdah 1987). According to this rule, Arabic nouns are either triliteral ثُلاثي *ṯulāṯī* such as حجر *ḥağr* 'stone', quadriliteral رُباعي*rubāʿī* such as جعفر *ğaʿfar* 'a name', or quinquiliteral خُماسي *ẖumāsī* such as سَفرجل *safarğal* 'quince'.

Table 6.21 shows examples of the 3 attributes of the Number of Root Letters category. Figure 6.22 shows the 3 attributes of the Number of Root Letters category, represented at position 20 in the tag string.

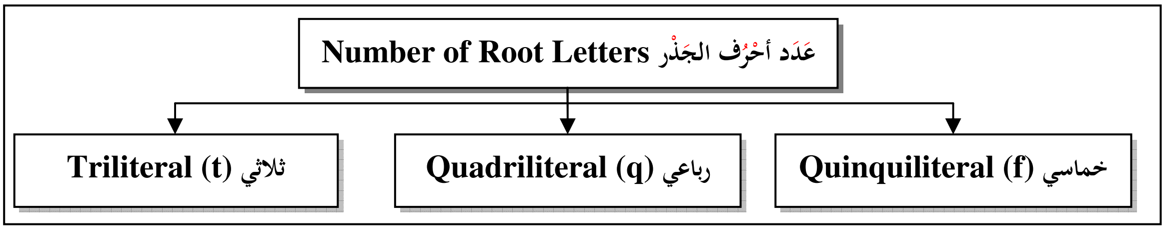


Figure 6.22 The Number of Root Letters category, with letter at position 20

Table 6.21 Examples of Number of Root Letters category attributes

Number of root letters	T	Examples
Triliteral ثُلاثي *ṯulāṯī*	t	ك ت ب *k t b* 'wrote'
Quadriliteral رُباعِي *rubāʿī*	q	د ح ر ج *d ḥ r ğ* 'rolled'
Quinquiliteral خُماسِي *ẖumāsī*	f	س ف ر ج ل *s f r ğ l* 'quince'

6.2.21 The Morphological Feature of Verb Root

Arabic linguists classify Arabic triliteral verbs (roots) into two main categories according to the groups of letters which construct the verb. These categories are the intact verb الفعل الصَّحيح *al-fi'l aṣ-ṣaḥīḥ* and the defective verb الفعل المعتل a*l-fi'l al-mu'tall*. Intact verbs are classified into three subcategories; sound verb الفعل السالم *al-fi'l as-sālim*, verb containing *hamza*^{h} الفعل المهموز *al-fi'l al-mahmūz*, and doubled verb الفعل المضعَّف *al-fi'l al-muḍa''af*. All the underlying (original) letters of the sound verb belong to the consonant letter group only; *i.e.* all letters except for the vowels and *hamza*^{h}. The second verb subcategory containing *hamza*^{h} has *hamza*^{h} (أ , إ , ؤ , ئ , ء) as one of its underlying (original) letters either as first, second or third letter. The doubled subcategory has the same letter as its second and third radicals (Al-Ghalayyni 2005).

The second category is the defective verb الأفعال المعتلة *al-'f'āl al-mu'talla*^{h} , where one or two of the the underlying (original) letters belong to the set of vowels ا , و , ي (*'alif, wāw, yā'*). This category has four subcategories. The first contains a vowel as the first letter of its root. This is called an initial-weak verb الفعل المثال *al-fi'l* al-*mithāl*. The second subcategory contains a vowel as the second letter of the root. This is called a hollow verb الفعل الأجوف *al-fi'l al-ağwaf*. The third subcategory contains a vowel as the third letter of its root. This is called a final-weak verb الفعل الناقص *al-fi'l an-nāqiṣ*. The last subcategory contains two vowels in its root. If these vowels are adjacent, as the first and second letters of the root, or as the second and third letters of the root, this is called an adjacent doubly-weak verb لفيف مقرون *lafīf maqrūn*. If it contains two vowels as the first and third root letters, it is called a separated doubly-weak verb لفيف مفروق *lafīf mafrūq* (Al-Ghalayyni 2005).

Figure 6.23 shows part of this classification of 30 Verb Root attributes. More detailed subclassification of triliteral verbs can be derived by combining the subcategories of verbs containing *hamza*^{h}, doubled letters and defective letters. Table 6.22 shows the 23 Verb Root attributes with an example of each attribute. The Verb Root category is represented at position 21 of the tag string.

Table 6.22 Verb Root category attributes and their tags at position 21

#	**Category attributes**		**Tag**	**Examples**
1	Sound verb	صحيح *saḥīḥ*	**a**	حسب *ḥasaba* 'calculated'
2	Doubled verb	مضعف *muḍa''af*	**b**	حبّ *ḥabba* 'loved'
3	Initially-hamzated verb	مهموز الفاء *mahmūz al-fā'*	**c**	أكل *'akala* 'ate'
4	Initially-hamzated and doubled verb	مهموز الفاء مضعَّف *mahmūz al-fā' muḍa''af*	**d**	أنّ *'anna* 'moan'
5	Initially- and finally-hamzated verb	مهموز الفاء ومهموز اللام *mahmūz al-fā' wa mahmūz al-lām*	**e**	أثأ *'aṯa'a* 'hit'
6	Medially-hamzated verb	مهموز العين *mahmūz al-'ayn*	**f**	سأل *sa'ala* 'asked'
7	Finally-hamzated verb	مهموز اللام *mahmūz al-lām*	**g**	بدأ *bada'a* 'started'

#	**Category attributes**		**Tag**	**Examples**
8	*wāw*-initial verb	مثال واوي *miṯāl wāwī*	**h**	وعد *wa‘ada* ‘promised’
9	*wāw*-initial and doubled verb	مثال واوي مضعف *miṯāl wāwī muḍa’’af*	**i**	ودّ *wadda* ‘wished’
10	*wāw*- initial and medially-hamzated verb	مثال واوي مهموز العين *miṯāl wāwī mahmūz al-‘ayn*	**j**	وئب *wa'iba* 'be angry'
11	*wāw*-initial and finally-hamzated verb	مثال واوي مهموز اللام *miṯāl wāwī mahmūz al-lām*	**k**	وطئ *waṭi’a* ‘trampled’
12	*yā'*-initial verb	مثال يائي *miṯāl yā’ī*	**l**	يقن *yaqina* ‘certained’
13	*yā'*-initial and doubled verb	مثال يائي مضعف *miṯāl yā’ī muḍa’’af*	**m**	يمّ *yamma* ‘to betake’
14	*yā'*- initial and medially-hamzated verb	مثال يائي مهموز العين *miṯāl yā’ī mahmūz al-‘ayn*	**n**	يئس *ya’isa* ‘to despair’
15	Hollow with *wāw*	أجوف واوي *’ağwaf wāwī*	**o**	قام *qāma* ‘to stand up’
16	Hollow with *wāw* and initially-hamzated verb	أجوف واوي مهموز الفاء *’ağwaf wāwī mahmūz al-fā’*	**p**	آب *āba* ‘to return’
17	Hollow with *wāw* and finally-hamzated verb	أجوف واوي مهموز اللام *’ağwaf wāwī mahmūz al-lām*	**q**	ناء *nā’a* ‘to fall down’
18	Hollow with *yā'*	أجوف يائي *’ağwaf yā’ī*	**r**	باع *bā‘a* ‘to sell’
19	Hollow with *yā'* and initially-hamzated verb	أجوف يائي مهموز الفاء *’ağwaf yā’ī mahmūz al-fā’*	**s**	أيس *’ayisa* ‘to despair’
20	Hollow with *yā'* and finally-hamzated verb	أجوف يائي مهموز اللام *’ağwaf yā’ī mahmūz al-lām*	**t**	شاء šā’ ‘to want’
21	Defective with *wāw* verb	ناقص واوي *nāqiṣ wāwī*	**u**	سرو *saraw* ‘to rid s.o’s worries’
22	Defective with *wāw* and initially-hamzated verb	ناقص واوي مهموز الفاء *nāqiṣ wāwī mahmūz al-fā’*	**v**	أسا *’asā* ‘to nurse’
23	Defective with *wāw* and medially-hamzated verb	ناقص واوي مهموز العين *nāqiṣ wāwī mahmūz al-‘ayn*	**w**	مأى *ma’ā* ‘to extend’
24	Defective with *yā'* verb	ناقص يائي *nāqiṣ yā’ī*	**x**	خشي *ẖašiya* ‘to fear’
25	Defective with *yā'* and initially-hamzated verb	ناقص يائي مهموز الفاء *nāqiṣ yā’ī mahmūz al-fā’*	**y**	أذي *’aḏiya* ‘to suffer damage’
26	Defective with *yā'* and medially-hamzated verb	ناقص يائي مهموز العين *nāqiṣ yā’ī mahmūz al-‘ayn*	**z**	رأى *ra'ā* ‘saw’
27	Adjacent doubly-weak verb	لفيف مقرون *lafīf maqrūn*	*****	قوي *qawiya* ‘to become strong’
28	Adjacent doubly-weak and initially-hamzated verb	لفيف مقرون مهموز الفاء *lafīf maqrūn mahmūz al-fā’*	**$**	أوى *’awā* ‘to seek refuge’
29	Separated doubly-weak verb	لفيف مفروق *lafīf mafrūq*	**&**	وقى *waqā* ‘to guard’
30	Separated doubly-weak and medially-hamzated verb	لفيف مفروق مهموز العين *lafīf mafrūq mahmūz al-‘ayn*	**@**	وأى *wa’ā* ‘to garantee’

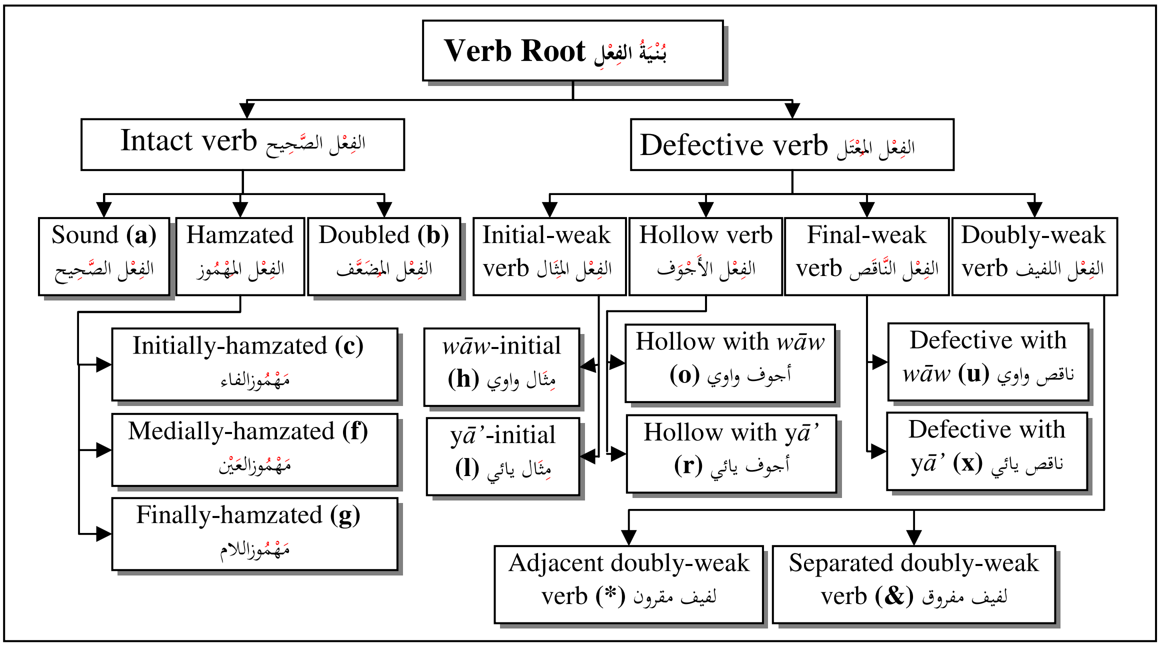

Verb Root بُنْيَةُ الفِعْلِ
Intact verb الفِعْل الصَّحِيح
Defective verb الفِعْل المعْتَل
Sound (a) الفِعْل الصَّحِيح
Hamzated الفِعْل المَهْمُوز
Doubled (b) الفِعْل المضَعَّف
Initial-weak verb الفِعْل المِثَال
Hollow verb الفِعْل الأَجْوَف
Final-weak verb الفِعْل النَّاقص
Doubly-weak verb الفِعْل اللفيف
Initially-hamzated (c) مَهْمُوزالفاء
Medially-hamzated (f) مَهْمُوزالعَيْن
Finally-hamzated (g) مَهْمُوزاللام
wāw-initial (h) مِثَال واوي
yā'-initial (l) مِثَال يائي
Hollow with wāw (o) أجوف واوي
Hollow with yā' (r) أجوف يائي
Defective with wāw (u) ناقص واوي
Defective with yā' (x) ناقص يائي
Adjacent doubly-weak verb (*) لفيف مقرون
Separated doubly-weak verb (&) لفيف مفروق

Figure 6.23 Verb Root attributes, with letter at position 21

6.2.22 The Morphological Feature of Types of Noun Finals

Nouns are classified according to their final letters into six categories.

1. The sound noun الاسم صحيح الآخر *al-'ism ṣaḥīḥ al-'āir* is a noun which ends with a consonant rather than a vowel or extended *'alif* ألف ممدودة *'alif mamdūda*^{h} which is an *'alif* followed by *hamza*^{h}. Case and mood marks appear at the end of sound nouns. Examples of sound nouns are; الرَّجُل *ar-rağul* 'the man', المَرْأة *al-mar'a*^{h} 'the woman', الكِتَاب *al-kitāb* 'the book', and القَلَم *al-qalam* 'the pen' (Al-Ghalayyni 2005).

2. The semi-sound noun الاسم شبه الصحيح *al-'ism šibh aṣ-ṣaḥīḥ* is a noun which ends with a vowel preceded by a silent consonant. Examples are دَلْو *dalw* 'bucket', ظَبْي *ẓaby* 'oryx', هَدْي *hady* 'guidance' and سَعْي *sa'y* 'striving'. Case and mood marks appear on the end of semi-sound nouns; for example the genitive case of the word دَلْوٍ *dalw* 'bucket' is marked by *tanwīn kasr* and the nominative case of the word ظَبْيٌ *ẓaby* 'oryx' is marked by *tanwīn ḍamm* as in the following sentence يَشْرَبُ ظَبْيٌ مِن دَلْوٍ *yašrabu ẓaby*^{un} *min dalw*^{in} 'an oryx is drinking from a bucket'. Similarly, the accusative case of the word ظَبْي *ẓaby* 'oryx' is marked by *tanwīn fatiḥ* in the following رَأَيْتُ ظَبْيًا *ra'aytu ẓaby*^{an} 'I saw an oryx' (Al-Ghalayyni 2005).

3. The noun with shortened ending الاسم المقصور *al-'ism al-maqṣūr* is a declinable noun ending with *'alif* of either *'alif* or *yā'* shapes. The final *'alif* is the underlying (original) letter, but it is either changed or augmented. The underlying (original) letter of the changed *'alif* is the vowel *wāw* or the vowel *yā'*. The underlying (original) vowel of the changed *'alif* appears in the dual form of the noun. The

noun final is affected by other morphological features such as number, root letters, and case and mood marks. For example, the underlying (original) vowel of the final *'alif* of the noun عصا *'aṣā* 'stick' is *wāw*, which appears in the dual form عَصَوَان *'aṣawān* 'two sticks', and the underlying (original) vowel of the final *'alif* of the noun فَتَى *fatā* 'boy' is *yā'*, which appears in the dual form فَتَيَان *fatayān* 'two boys'. The augmented *'alif* is added to the noun to make it similar to other nouns or to match a certain pattern such as أَرْطى *'arṭā* 'kind of trees' and ذِفْرَى *ḏifrā* 'bone behind the ear'. The final *'alif* is written either as *'alif* or *yā'*. If the word consists of four or more letters such as مُسْتَشفَى *mustašfā* 'hospital', or if it is derived from *yā'*, which is its third underlying radical, as in فَتَى *fatā* 'boy', it is as *yā'*. It is written as an *'alif,* if it is derived from the vowel letter *wāw* which is its third underlying radical. An example is نَدَى *nadā* 'dew', where the root is ندو *n-d-w* (Al-Ghalayyni 2005).

4. The noun with extended ending الاسم الممدود *al-'ism al-mamdūd* is a declinable noun ending with *hamza*^{h} preceded by augmented *'alif* such as سَمَاء *samā'* 'sky' and صَحْرَاء *ṣaḥrā'* 'desert'. The *hamza*^{h} at the end of the noun is either underlying (original) as in قُرَّاء *qurrā'* 'readers' or derived from *yā'* or *wāw* as in, سَمَاء *samā'* 'sky' and بِنَاء *binā'* 'building' where the former is derived from *yā'* and the later is drived from *wāw*. The *hamza*^{h} might be an added letter indicating feminine nouns as in حَسْنَاء *ḥasnā'* 'beautiful', or might be added to make it similar to certain patterns as in حِرْبَاء *ḥirbā'* 'chameleon' (Al-Ghalayyni 2005).

5. The noun with curtailed ending الاسم المنقوص *al-'ism al-manqūṣ* is a declinable noun ending with *yā'* and preceded by a letter with the short vowel *kasra*^{h} such as القَاضِي *al-qāḍī* 'the judge' and الرَّاعِي *ar-rā'ī* 'shepherd'. The final *yā'* is deleted if the noun is an indefinite noun, where the definite article *'alif-lām* (ال) is not attached to the beginnig of the word, and the noun is in nominative or genitive case as in حَكَمَ قاضٍ على جانٍ *ḥakama qāḍ*^{in} *'alā ğān*^{in} 'A judge judged a criminal'. However, the final *yā'* appears if the definite article is attached to the noun or if it is added to another noun which defines it as in حَكَمَ القاضِي على الجَانِي *ḥakama al-qāḍī 'alā al-ğānī* 'The judge judged the criminal' and جَاءَ قَاضِي القُضَاة *ğā' qāḍī al-quḍāt* 'A chief justice came' (Al-Ghalayyni 2005).

6. The noun with deleted ending الاسم محذوف الآخر *al-'ism maḥḏūf al-'āḫir* is a noun where its final underlying vowel is deleted. This kind of noun may consist of two letters such as يَدٌ *yad* 'hand', where the final underlying vowel *yā'* is deleted يدي *y-d-y*. Other examples are; سَنَة *sana*^{h} 'year', where the final underlying vowel *wāw* is deleted سنو *s-n-w*, and لُغَة *luga*^{h} 'language', where the underlying vowel *wāw* is deleted لغو *l-ğ-w* (Al-Ghalayyni 2005).

Figure 6.24 shows this classification of Noun Finals. Table 6.23 shows examples of the 6 attributes of the morphological feature of Noun Finals, represented at position 22 of the tag string.

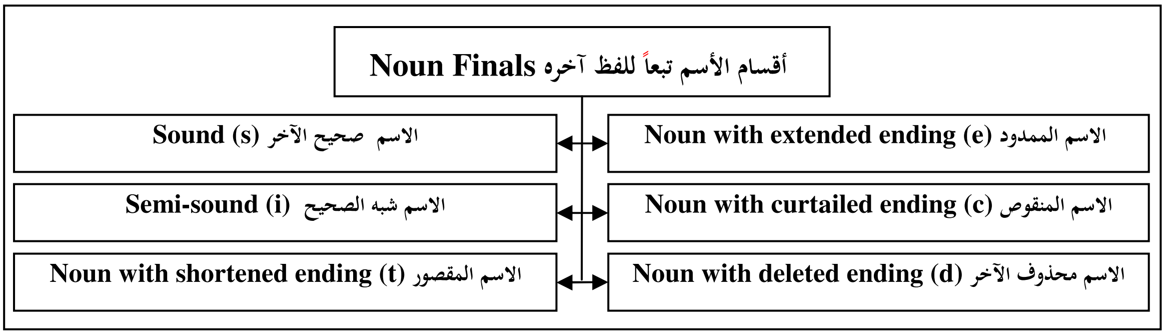

Noun Finals أقسام الأسم تبعاً للفظ آخره
Sound (s) الاسم صحيح الآخر
Semi-sound (i) الاسم شبه الصحيح
Noun with shortened ending (t) الاسم المقصور
Noun with extended ending (e) الاسم الممدود
Noun with curtailed ending (c) الاسم المنقوص
Noun with deleted ending (d) الاسم محذوف الآخر

Figure 6.24 The classification of nouns according to their final letters, for the morphological feature of Noun Finals, with letter at position 22

Table 6.23 Examples of the attributes of the morphological feature of Noun Finals

Attributes of noun final letters category	**T**	**Examples**
Sound noun الاسم صحيح الآخر *al-’ism ṣahīh al-’āir*	**s**	الرَّجُل *ar-rağul* ‘the man’, المَرْأَة *al-mar’a*^{h} ‘the woman’, الكِتَاب *al-kitāb* ‘the book’, and القَلَم *al-qalam* ‘the pen’.
Semi-sound noun الاسم شبه الصحيح *al-’ism šibh aṣ-ṣaḥīḥ*	**i**	دَلْو *dalw* ‘bucket’, ظَبْي *ẓaby* ‘oryx’, هَدْي *hady* ‘guide’ and سَعْي *sa’y* ‘striving’.
Noun with shortened ending الاسم المقصور *al-’ism al-maqṣūr*	**t**	عَصا *‘aṣā* ‘stick’, فَتَى *fatā* ‘boy’, مُسْتَشفَى *mustašfā* ‘hospital’, أَرْطى *‘arṭā* ‘kind of trees’, ذِفْرَى *ḏifrā* ‘A bone behind the ear’ and نَدَى *nadā* ‘dew’.
Noun with extended ending الاسم الممدود *al-’ism al-mamdūd*	**e**	سَمَاء *samā’* ‘sky’, صَحْرَاء *ṣaḥrā’* ‘desert’, بِنَاء *binā’* ‘building’, حَسْنَاء *ḥasnā’* ‘beautiful’ and حِرْبَاء *ḥirbā’* ‘chameleon’.
Noun with curtailed ending الاسم المنقوص *al-’ism al-manqūṣ*	**c**	القَاضِي *al-qāḍī* ‘the judge’ and الرَّاعِي *ar-rā‘ī* ‘shepherd’, حَكَمَ قَاضٍ على جانٍ *ḥakama qāḍ*^{in} *‘alā ğānin* ‘A judge judged a criminal’ and جَاءَ قَاضِي القُضَاة *ğā’ qāḍī al-quḍāt* ‘A chief justice came’.
Noun with deleted ending الاسم محذوف الآخر *al-’ism maḥḏūf al-’āẖir*	**d**	يَدْ *yad* ‘hand’, سَنَة *sana*^{h} ‘year’, and لُغَة *luğa*^{h} ‘language’.

6.3 Chapter Summary

This chapter discussed the SALMA Tag Set morphological feature categories and their attribute values. The SALMA Tag Set captures long-established traditional morphological features of Arabic, in a compact yet transparent notation. For a morphologically-rich language like Arabic, the Part-of-Speech tag set should be defined in terms of morphological features characterizing word structure. A detailed description of the SALMA Tag Set explains and illustrates each feature and its possible values. In our analysis, a tag consists of 22 characters; each position represents a feature and the letter at that location represents a value or attribute of the morphological feature; the dash "-" represents a feature not relevant to a given word. The SALMA Tag Set is not tied to a specific tagging algorithm or theory, and other tag sets could be mapped onto this standard, to simplify and promote comparisons between and reuse of Arabic taggers and tagged corpora.

The SALMA Tag Set has been applied to a sample from the Quranic Arabic Corpus (QAC) to prove its applicability to morphologically annotate Arabic text with very fine-grained morphological analysis of each morpheme of the corpus words. The next chapter (chapter 7) discusses the steps in applying the SALMA Tag Set to annotate a sample of 1000 words from the Quranic Arabic Corpus.

Chapter 7
Applying the SALMA – Tag Set

This chapter is based on the following sections of published papers:

Section 3 depends on section 5 from (Sawalha and Atwell Under review)

Sections 4 and 5 are based on sections 3 and 4 from (Sawalha and Atwell 2011c)

Chapter Summary

Morphosyntactic tag sets are evaluated by studying external and internal design criteria. The external design criterion involves measuring the capability of making the linguistic distinctions required by higher level NLP applications. The internal design criterion evaluates the application of the tag set in tagging of a corpus.

The SALMA – Tag Set has been validated in two ways. First, it was validated by proposing it as a standard to the Arabic language computing community, and it has been adopted in several Arabic language processing systems. Second, an empirical approach to evaluating the SALMA – Tag Set of Arabic showed that it can be applied to an Arabic text corpus, by mapping from an existing tag set to the more detailed SALMA Tag Set. The morphological tags of a 1000-word test text, chapter 29 of the Quranic Arabic Corpus, were automatically mapped to SALMA tags.

The SALMA – Tag Set and the SALMA – Gold Standard tagged corpus are open-source resources and standard to promote comparability and interoperability of Arabic morphological analyzers and Part-of-Speech taggers.

7.1 Introduction

The evaluation of morphosyntactic tag sets has been less studied in the literature than the evaluation of the morphosyntactic tools (Dejean 2000). Evaluating the external and internal design criteria of tag sets are two types of evaluation methodology. The external criterion for evaluation checks if the tag set is capable of making the linguistic distinctions required by higher level NLP applications such as part-of-speech taggers and parsers. The internal criterion evaluates the applicability in accurately tagging corpus (Elworthy 1995; Dejean 2000; Melamed and Resnik 2000; Sharoff et al. 2008; Zeman 2008). Modifying the tag set (e.g. decreasing the cardinality of the tag set by omitting certain attributes) and comparing the tagging accuracy of the modified tag set with the accuracy gained using the original tag set is an evaluation approach for tag sets (Dejean 2000; Dzeroski, Erjavec and Zavrel 2000; Melamed and Resnik 2000; Diab 2007). Another evaluation methodology involves mapping from an existing coarse tag set to a fine-grained tag set and enriching the corpus by linguistically informed knowledge, then measuring the increment in accuracy gained by using the mapped tag set to train part-of-speech tagging systems (Melamed and Resnik 2000; MacKinlay 2005). (Dickinson and Jochim 2010) evaluated different tag set mappings and their distributional properties depending on the external and internal design criteria. Theoretical comparison of tag sets depending on certain specifications and requirements of application or tagging scheme of a corpus is also seen as evaluation methodology for tag sets (Gopal, Mishra and Singh 2010). However, evaluating the tag set by measuring whether the tag set is useful for certain application depends on how much information the application needs (Jurafsky and Martin 2008).

Moreover, tag sets are always associated with a certain annotated corpus or annotation system. For instance, the Brown tag set is used in the part-of-speech tagging of the Brown corpus; the C5 tag set is associated with both the CLAWS part-of-speech tagger and the BNC; the Penn Arabic Treebank tag set is used by the Buckwalter morphological analyzer and to part-of-speech tag the Penn Arabic Treebank; and the QAC tag set is used in the morphosyntcatic annotation layer of the Quranic Arabic Corpus. Applying the tag set in real-life data or applications, represented by text corpora and part-of-speech taggers, is the validation methodology of the tag sets.

Section 7.3 discusses two proposed evaluation methodologies for evaluating the SALMA Tag Set. First, evaluating the tag set by proposing the morphosyntactic annotation scheme to be used by wider the NLP community. Second, by tagging a test corpus, by mapping from an existing tag set to the SALMA Tag Set.

7.2 Why was Manual Annotation not Applied?

An essential prerequisite to implementing an automatic morphosyntactic analyzer is to try out the tag set manually. Two benefits are gained by trying the tag set manually. First, tag sets which are designed depending of the published grammar of the language rather than direct reference to data, need to be applied to reflect valid distinctions of their categories in the language, and to identify phenomena which are difficult to categorize or intrinsically ambiguous. Second, the manually tagged text represents training data for tagging systems that apply machine learning algorithms, and it represents a gold standard for evaluating morphosyntactic analyzers in general (Hardie 2004).

Due to the limitations of time, funds to hire annotators, and the lack of availability of professional annotators especially in a non-Arabic speaking country such as the UK where the project is taking place, purely manual annotation for an Arabic corpus was not practical. However, samples of both Classical Quranic Arabic and Modern Standard Arabic (MSA) were morphologically annotated using the SALMA – Tag Set. Section 7.4 and Chapter 9 discuss the construction of the SALMA – Gold Standard.

Moreover, fine-grained distinctions might affect inter-annotator agreement. Hence, measuring inter-annotator agreements and defining clear decision criteria for suitable tags, are time-consuming and require major effort.

On balance, it was more practical to adapt an existing tagged text. The mapping from the Quranic Arabic Corpus morphological tags to SALMA tags allowed the construction of a gold standard and verified that the SALMA Tag Set is applicable and can be used to enrich Arabic text corpora with fine-grained morphosyntactic information.

As a future work project, applying the SALMA Tag Set to a larger representative Arabic corpus will be of high priority. Chapter 11 discusses this future work project.

7.3 Methodologies for Evaluating the SALMA Tag Set

Two ways to validate the SALMA Tag Set of Arabic are: first, to propose it as a standard to the Arabic language computing community and have the standard adopted by others. Second, another empirical evaluation is to see how readily it can be applied to a sample of Arabic text, for example by mapping from an existing tagged corpus to the SALMA tag set.

The SALMA Tag Set has been used in the SALMA Tagger (Sawalha Atwell Leeds Morphological Analysis Tagger). It is used as the standard for specifying the word's morphemes and for encoding the morphological features of each morpheme (Sawalha and Atwell 2009b; Sawalha and Atwell 2009a). The SALMA Tag Set has been published online (http://www.comp.leeds.ac.uk/sawalha/tagset.html) and has been adopted as a standard by other Arabic language computing researchers. For instance, part of the tag set is also used in the Arabic morphological analyzer and part-of-speech tagger Qutuf (Altabbaa, Al-Zaraee and Shukairy 2010). Qutuf uses the main part-of-speech, the subcategories of nouns, the subcategories of verbs named as verb aspects, the subcategories of particles and the morphological features of gender, number, person, case or mood, definiteness, voice, transitivity, and part of the declension and conjugation category named as perfectness. Qutuf does not use the SALMA tag format. Rather it uses a tag consisting of slots for each feature separated by a comma. Another re-use of the SALMA – Tag Set has been reported as a standard for evaluating Arabic morphological analyzers, and for building a Gold Standard for evaluating Arabic morphological analyzers and part-of-speech taggers (Hamada 2010).

The second method for evaluating the SALMA Tag Set is to apply it to a sample of Arabic text, by mapping from an existing broad tag set to the more fine-grained SALMA Tag Set. Morphologically annotated sample text from the Quranic Arabic Corpus (QAC), chapter 29, consisting of about 1000 words, was selected. Then, an automated mapping algorithm was developed to map the QAC morphological tags to the SALMA tags. After that, the automatically mapped morphological features tags were manually verified and corrected, to provide a new fine-grain Gold Standard for evaluating Arabic morphological analyzers and part-of-speech taggers.

The mapping from the QAC morphological tag set to the SALMA Tag Set was done by the following six-step procedure.

1. **Mapping classical to modern character-set:** the QAC uses the classical Othmani script of the Qur'an (77,430 words) which was mapped to Modern Standard Arabic (MSA) script (77,797 words).

2. **Splitting whole-word tags into morpheme-tags:** the morphological tag in the QAC is a whole-word tag, composed by combining the prefix with the stem and suffix morphological tags, while the SALMA Tag Set is designed for word morpheme tagging.

3. **Mapping of feature-labels:** the mnemonics of the Quranic Arabic Corpus tags were mapped to their equivalent in the SALMA Tag Set.

4. **Adjustments to morpheme tokenization:** due to differences between the underlying word tokenization model used in the QAC and the one required for the SALMA Tag Set, the mapped tags of the prefixes and suffixes were replaced with SALMA tags by matching them to the clitics and affixes lists used by the SALMA Tagger (Sawalha and Atwell 2009a; Sawalha and Atwell 2010b).

5. **Extrapolation of missing fine-grain features:** for the morphological features which are not included in the QAC tag set, automatic "feature-guessing" procedures applied linguistic knowledge extracted from traditional Arabic grammar textbooks, encoded as a computational rule-based system, to automatically predict the values of the missing morphological features of the word.
6. **Manually proofread and corrected the mapped SALMA tags:** proofreading and correction is done by an Arabic language expert. The result is a sample Gold Standard annotated corpus for evaluating morphological analyzers and part-of-speech taggers for Arabic text.

Section 7.4 explains the mapping procedures followed to map the QAC morphological tags to the SALMA tags.

7.4 Mapping the Quranic Arabic Corpus (QAC) Morphological Tags to SALMA Tags

The reuse of existing components is an established principle in software engineering. The Quranic Arabic Corpus (QAC) is a newly available resource enriched with multiple layers of annotation including morphological segmentation and part-of-speech tagging (Dukes and Habash 2010). A morphologically annotated test text sample from the QAC, chapter 29, consisting of about 1000 words, was selected. Then, an automated mapping methodology mapped the QAC morphological tags to SALMA morphological features tags.

The mapping from the QAC morphological tags to the SALMA morphological features tags is done by following a six-step procedure. The following sub-sections describe in detail the mapping steps, highlight the challenges of mapping and show examples of mapping the QAC morphological tags to the SALMA morphological features tags.

7.4.1 Mapping Classical to Modern Character-Set

The QAC uses the Othmani script of the Qur'an. Most Arabic NLP applications deal with MSA script. These programs need some modifications to deal with the Othmani script. However, the Qur'an script is also available in MSA script. One-to-one mapping, between the Qur'anic words written in Othmani script and the Qur'an written in MAS script, can be applied to the QAC except for a few special cases. Such cases exist due to the spelling variations between the Othmani script and the MSA script. For instance the vocative particle يا *yā* is written connected to the next word in Othmani script, and it is written as standalone token in MSA script e.g. the word يَٰمُوسَىٰٓ *yāmūsā* 'O Musa "Moses"!'in Othmani script is one token but it is written as two tokens in MSA script as يَا

مُوسَى *yā mūsā* ‘O Musa “Moses”!’. Therefore, The QAC has 77,430 words while the Quran in written MSA has 77,797 tokens. Figure 7.1 gives some examples of the spelling variations between the Othmani script and MSA script.

Othmani		Standard Arabic		Meaning
يَٰمُوسَىٰ	*yāmūsā*	يَا مُوسَى	*yā mūsā*	O Musa (Moses)!
يَٰٓأَهْلَ	*yā’ahla*	يَا أَهْلَ	*yā ’ahla*	O people of
يَٰلَيْتَنِى	*yālaytanī*	يَا لَيْتَنِي	*yā laytanī*	I wish if I had
وَأَلَّوِ	*wa’allaw*	وَأَنْ لَوِ	*wa’n law*	And if not
يَٰعِيسَى	*yā‘isā*	يَا عِيسَى	*yā ‘isā*	O Issa (Jesus)!
يَٰقَوْمِ	*yāqawm*	يَا قَوْمِ	*yā qawm*	O people

Figure 7.1 Examples of spelling / tokenization variations between the Othmani script and MSA script

The one-to-one mapping was done automatically. The difference of 375 tokens between the two writing schemes was manually corrected, by grouping two tokens of MSA that match one token of the Othmani script. This grouping is done to preserve the morphological tag of the words. From the previous example the word يَٰمُوسَىٰ *yāmūsā* ‘O Musa “Moses”!’ has the QAC morphological tag **ya+ POS:PN LEM:muwsaY` M NOM**, which is mapped to the two tokens يَا and مُوسَى *yā mūsā* ‘O Musa “Moses”!’ and these two tokens are given the same morphological tag as illustrated in figure 7.2.

Othmani	QAC morphological tag	MSA	QAC morphological tag
يَٰمُوسَىٰ	ya+ POS:PN LEM:muwsaY` M NOM	يَا	ya+
		مُوسَى	POS:PN LEM:muwsaY` M NOM

Figure 7.2 mapping example, preserving the part-of-speech tag

7.4.2 Splitting Whole-Word Tags into Morpheme-Tags

Tokenizing the word into its morphemes is not an easy task for Arabic words. The tokenization of QAC words into morphemes was done automatically using BAMA. However, there is no resource provided by the QAC that tokenizes the words into their morphemes and assigns the morphological tags for each morpheme. The given morphological tags are whole word tags, combining the prefix with the stem and the suffix morphological components separated by a + sign. So, for our mapping process, the words and their morphological tags were automatically tokenized into morphemes and morphemes tags. Figure 7.3 shows an example of tokenizing a word and its morphological tag into morphemes and morpheme tags.

Word no.	Othmani word	MSA word	QAC morphological tag
(16:72:16)	أَفَبِٱلْبَٰطِلِ	أَفَبِالْبَاطِلِ	A:INTG+ f:REM+ bi+ Al+ POS:N ACT PCPL LEM:ba`Til ROOT:bTl M GEN
Morpheme [1]	أَ	أَ	A:INTG
Morpheme [2]	فَ	فَ	f:REM
Morpheme [3]	بِ	بِ	Bi
Morpheme [4]	ٱلْ	الْ	Al
Morpheme [5]	بَٰطِلِ	بَاطِلِ	POS:N ACT PCPL LEM:ba`Til ROOT:bTl M GEN

Figure 7.3 Example of tokenizing Quranic Arabic Corpus words and their morphological tags into morphemes and their morpheme tags

The QAC has 18,994 word types (Othmani script) and 18,123 different morphological tags. This large number of different morphological tags can be reduced to 1,067 different morpheme tags after dividing the morphological tag of the whole word into morpheme tags and removing the ROOT: and LEM: parts of the QAC morphological tags.

7.4.3 Mapping of Feature-Labels

The third mapping step starts by mapping the mnemonics of the QAC to their equivalent in the SALMA – Tag Set, followed by application of the morphological feature templates that determine the applicable and non-applicable morphological features of the analyzed morphemes.

A mapping dictionary was constructed to map the mnemonics of the QAC that captures the morphological features of the analyzed morphemes, to their SALMA Tag Set equivalent attribute values and the attributes' positions in the SALMA tag string. Figure 7.4 shows part of the dictionary data structure used to map between QAC and SALMA tags. The dictionary consisting of 158 entries was used via a specialized program that matches the QAC morphemes tags after tokenization, and returns the attributes' values and the positions for the mapped features. Then, the attributes are placed in their specified positions in the SALMA tag string.

```
{"1FP"     :[(7,'f'),(8,'p'),(9,'f')],   # 1st person / Feminine /Plural
"1FS"      :[(7,'f'),(8,'s'),(9,'f')],   # 1st person / Feminine /Singular
"1MP"      :[(7,'m'),(8,'p'),(9,'f')],   # 1st person / Masculine / Plural
"1P"       :[(8,'p'),(9,'f')],           # 1st person / Plural
"1S"       :[(8,'s'),(9,'f')],           # 1st person / Singular
"2D"       :[(8,'d'),(9,'s')],           # 2nd person / Dual
"2FD"      :[(7,'f'),(8,'d'),(9,'s')],   # 2nd person / Feminine / Dual
"2MS"      :[(7,'m'),(8,'s'),(9,'s')],   # 2nd person / Masculine / Singular
"POS:ACC"  :[(1,'p'),(4,'o')],           # Accusative particle
"POS:ADJ"  :[(1,'n'),(2,'j')],           # Adjective
"POS:N"    :[(1,'n')],                   # Noun
"POS:P"    :[(1,'p'),(4,'p')],           # Preposition
"POS:V"    :[(1,'v')],                   # Verb
```

Figure 7.4 Part of the dictionary data structure used to map the Quranic Arabic Corpus tag set to the morphological features tag set

The SALMA tag string consists of 22 features. Not all these features are applicable for a given part-of-speech. For instance, *number* and *gender* at positions 7 and 8 respectively, are noun features, while *person* and *voice* at positions 9 and 14 respectively are verb features. The SALMA Tag Set uses ‘-’ to show that the feature in that position is not applicable, and it uses ‘?’ to show that the feature is applicable but its attribute value is not known yet.

A matrix of the main and sub parts of speech and their applicable features (or possible attributes) has been constructed and used by the mapping program and the SALMA – Tagger (Sawalha and Atwell 2009b; Sawalha and Atwell 2009a; Sawalha and Atwell 2010b). Chapter 8 discusses in detail the SALMA – Tagger algorithm. The matrix is used as SALMA tag string templates. For each main or sub part-of-speech there is a template that shows the applicable and non-applicable morphological features. The main part of speech and some of the sub part of speech categories are already marked in the initially mapped tag. A string, formed by grouping the attributes of the first 6 positions of the initial SALMA tag string representing the main and the sub part of speech categories, is used as a key to search the templates dictionary that stores the SALMA tag templates. These templates are used to add ‘-’, ‘?’ or any other specified attributes to the initially mapped tag string. Figure 7.5 shows a sample of SALMA tag templates.

```
{`n?----`   :  `n?----??-????---????-?'    #  Noun
 `v-?---`   :  `v-?-----????-????????-`    #  Verb
 `p--?--`   :  `p--?-----????---?-----`    #  Particle
 `r---?-`   :  `r---?-??????????------`    #  Residual
 `u----?'   :  `u----?----------------`    #  Punctuation
 `ng----`   :  `ng----??-v???---?d??-?'    #  Gerund
 `np----`   :  `np----???s-??---?ns---`    #  Pronoun
 `v-p---`   :  `v-p-----?s-?-?m??????-`    #  Past verb
 `v-c---`   :  `v-c-----?d??-????????-`    #  Present verb
 `v-i---`   :  `v-i-----?s-?-a???????-`    #  Imperative verb
 `p--p--`   :  `p--p-----s-?-----n----`    #  Preposition
 `p--a--`   :  `p--a-----s-?-----n----`    #  Annular
 `p--c--`   :  `p--c-----s-?-----n----`    #  Conjunction
 `r---r-`   :  `r---r-???s-?----------`    #  Connected pronoun
 `r---t-`   :  `r---t-fs-s-?----------`    #  tā' Marbouta
 `r---d-`   :  `r---d-------d---------`    #  Definite article
 `u----s'   :  `u----s----------------`    #  Full stop
 `u----c'   :  `u----c----------------`    #  Comma
 `u----n'   :  `u----n----------------`    #  Colon
```

Figure 7.5 A sample of the morphological features tag templates

7.4.4 Adjustments to Morpheme Tokenization

Due to the differences between the underlying word’s morpheme tokenization models used in the QAC and the one required for the SALMA – Tag Set, adjustment to morpheme tokenization is required. The fine-grained SALMA – Tagger divides the word into five parts: proclitic(s), prefix(es), stem, suffix(es) and enclitic(s). Clitics and affixes can be multiple clitics or affixes. The underlying word’s morpheme tokenization model used by the QAC is inherited from BAMA. So, the SALMA-Tagger is used to tokenize the words into morphemes and to assign the morpheme tag by matching the clitics and affixes morphemes of the analyzed words with the clitics and affixes from the clitics and affixes dictionaries of the SALMA-Tagger.

The clitics and affixes dictionaries contain detailed information about proclitic and prefix combinations and suffix and enclitic combinations. This information includes suitable SALMA tags and three information labels that help in matching the correct combination of proclitics and prefixes from one side with the suffixes and enclitics from the other side. The first label [proc, perf, suf, enc] indicates whether the clitic or affix is a proclitic, prefix, suffix or enclitic respectively. The second label [n, v, x] represents the main part-of-speech of the stem morpheme which indicates whether the clitic or affix belongs to noun, verb or both. The final information is [y, n]. This indicates whether the clitic or affix is part of the pattern or not. This information is useful for pattern generator and lemmatizer programs. The construction and the properties of clitics and affixes dictionaries are discussed in more detail in chapter 8. The SALMA – Tagger selects the clitic and affix combinations that match this information and match the main part of speech of the stem. Figure 7.6 shows examples from the clitics and affixes lists. Figure 7.7 shows a sample of the mapped morphological features tags after applying step 4.

Proclitics and prefixes list								
وَلَيَ		وَلَيَعْلَمَنَّ *walaya'lamanna* "And he will surely make evident"						
1	وَ	*wa*	**p--c------------------**	proc	x	n	حرف عطف	Conjunction
2	لَ	*la*	**p--z-----s-f----------**	proc	v	n	حرف توكيد	Emphatic particle
3	يَ	*ya*	**r---a-----------------**	pref	v	y	حرف مضارعة	Imperfect prefix
Suffixes and enclitics list								
اتِهَا		وَتَطْبِيْقَاتِهَا *wataṭbīqātihā* "And its applications"						
1	اتِ	*āti*	**r---l-fp--------------**	suf	n	y	حروف جمع المؤنث السا لم	Feminine sound plural letters
2	هَا	*hā*	**r---r-fsts-s----------**	enc	x	n	ضمير متصل	Suffixed pronoun

Figure 7.6 Examples of the clitics and affixes lists

Morpheme	QAC morpheme tag	SALMA tags after the 4th step
الم	POS:INL	p--?-----????---?-----
أَ	A:INTG+	p--i-----s------------
حَسِبَ	POS:V PERF 3MS	v-p---mst--?-?-??????-
ال	Al+	r---d-----------------
نَاسُ	POS:N MP NOM	n?----mp-?n??---????-?
أَنْ	POS:SUB	p--g-------?----------
يُ	NULL	r---a-----------------
تْرَكُ	POS:V IMPF PASS 3MP MOOD:SUBJ	v-c---mptda?-p???????-
وا	PRON:3MP	r---r-mptsnw----------
أَنْ	POS:SUB	p--g-------?----------
يَ	NULL	r---a-----------------
قُولُ	POS:V IMPF 3MP MOOD:SUBJ	v-c---mptda?-????????-
وا	PRON:3MP	r---r-mptsnw----------
آمَنَّ	POS:V PERF (IV) 1MP	v-p---mpf--?-?-??????-
نَا	PRON:1MP	r---r-xpfs??----------
وَ	wa+	p--c------------------
هُمْ	POS:PRON 3MP	np----mpt--??---?-----
لَا	POS:NEG	p--n-------?----------
يُ	NULL	r---a-----------------
فْتَنُ	POS:V IMPF PASS 3MP	v-c---mpt-??-p???????-
ونَ	PRON:3MP	r---r-mp?snn----------

Figure 7.7 A sample of the mapped SALMA tags after applying mapping steps 1 to 4

After applying the four-step mapping procedure to a sample of 1000 words, chapter 29 of the Qur'an, the success rate in mapping each morphological features category was computed by comparing with the final version after proof reading. Table 7.1 shows how successful the mapping was for each individual target feature. Full mapping was done for the main part-of-speech and sub part of speech categories, with a success rate of nearly 100% except for noun sub-categories of which only about 50% were mapped successfully. The morphological categories of gender, number, person, inflectional morphology and case or mood were mapped with a success rate of 68% to 89%. Case and mood marks, definiteness, voice, emphasized and non-emphasized, and declension and conjugation were poorly mapped with a success-rate of 5% to 17%. Transitivity, rational, unaugmented and augmented, number of root letters, verb root and noun finals were not mapped at all, because these morphological features do not exist in the QAC tag set.

Table 7.1 The mapping success rate after applying the first four mapping steps

Category		?	-	Applicable	Not mapped	mapped
1	Main Part-of-Speech	16	0	1935	0.83%	99.17%
2	Part-of-Speech: Noun	247	1435	500	49.40%	50.60%
3	Part-of-Speech: Verb	0	1675	260	0.00%	100.00%
4	Part-of-Speech: Particle	31	1424	511	6.07%	93.93%
5	Part-of-Speech: Other	0	1287	648	0.00%	100.00%
6	Punctuation marks	0	1935	0	0.00%	100.00%
7	Gender	125	785	1150	10.87%	89.13%
8	Number	244	847	1088	22.43%	77.57%
9	Person	103	1267	668	15.42%	84.58%
10	Inflectional morphology	85	1141	794	10.71%	89.29%
11	Case and Mood	280	1043	892	31.39%	68.61%
12	Case and Mood marks	1120	581	1354	82.72%	17.28%
13	Definiteness	402	1467	468	85.90%	14.10%
14	Voice	220	1698	237	92.83%	7.17%
15	Emphasized and non-emphasized	114	1805	130	87.69%	12.31%
16	Transitivity	260	1675	260	100.00%	0.00%
17	Rational	712	1223	712	100.00%	0.00%
18	Declension and Conjugation	482	1428	507	95.07%	4.93%
19	Unaugmented and Augmented	603	1332	603	100.00%	0.00%
20	Number of root letters	654	1281	654	100.00%	0.00%
21	Verb root	260	1675	260	100.00%	0.00%
22	Nouns finals	394	1541	394	100.00%	0.00%

7.4.5 Extrapolation of Missing Fine-Grain Features

As previously discussed, The SALMA – Tag Set is a fine-grained tag set that captures 22 morphological features in the tag string. As shown in table 7.1 above, some of these morphological features are poorly mapped such as case and mood marks; definiteness; voice; emphasized and non-emphasized; and declension and conjugation; while others are not mapped because they are not represented by the QAC morphological tag set. The non-mapped features are: transitivity; rational; unaugmented and augmented; number of root letters; verb root; and types of nouns according to their final letters.

The morphological features which are not included in the QAC tag set are automatically guessed using the SALMA – Tagger. The SALMA – Tagger has specialized procedures that apply the linguistic knowledge extracted from traditional Arabic grammar books as a computational rule-based system to automatically guess the value of the remaining morphological features of the word's morphemes. Chapter 8 discusses in detail these procedures.

A rule-based approach was used for morphological analysis of the 22 morphological features. Rules were extracted from traditional Arabic grammar books. Then, these rules were programmed and integrated to the SALMA – Tagger to predict the morphological feature values of each morpheme of the analyzed word. The rules depend on the structure of the analyzed words and their morphemes to predict the value of a given category. For instance, if the analyzed word has a prefix يَ *yā* and suffixed pronoun ونَ *ūna* then the appropriate tag of the person category is ‘***t***’ representing third person and the subject’s number and gender guessed values are ‘***p***’ and ‘***m***’ representing plural and masculine respectively. The rules also depend on linguistic lists for the features that are hard to predict depending on the structure of the analyzed words. The SALMA – Tagger has linguistic lists such as a broken plural list to predict the number feature of nouns; list of doubly transitive verbs and list of triply transitive verbs to predict the values of the transitivity feature; lists of restricted to perfect, restricted to imperfect, restricted to imperative, and partially conjugated verbs which are used to guess the values of the declension and conjugation morphological feature.

Table 7.1 showed that the mapping percentage after applying the first four mapping steps for these morphological features is less than 20% and most of them have 0% mapping. These procedures are also used to verify the already mapped morphological features such as number, gender, person and case or mood. After applying these rule-based procedures the mapping success rate increased and reached 83% to 100% for most of the morphological features. Table 7.2 shows the mapping success-rate after applying the fifth mapping step of applying the rule-based system to morphological analysis.

Table 7.2 The mapping success rate after applying the fifth mapping step

Category		**?**	**-**	**Applicable**	**Not Mapped**	**Mapped %**
1	Main Part-of-Speech	0	0	1935	0.00%	100.00%
2	Part-of-Speech: Noun	247	478	1457	16.95%	83.05%
3	Part-of-Speech: Verb	0	716	1219	0.00%	100.00%
4	Part-of-Speech: Particle	26	758	1177	2.21%	97.79%
5	Part-of-Speech: Other	0	976	959	0.00%	100.00%
6	Punctuation marks	0	976	959	0.00%	100.00%
7	Gender	123	219	1716	7.17%	92.83%
8	Number	305	218	1717	17.76%	82.24%
9	Person	0	673	1262	0.00%	100.00%
10	Inflectional morphology	0	0	1935	0.00%	100.00%
11	Case and Mood	250	241	1694	14.76%	85.24%
12	Case and Mood marks	262	0	1935	13.54%	86.46%
13	Definiteness	0	478	1457	0.00%	100.00%
14	Voice	0	716	1219	0.00%	100.00%
15	Emphasized and non-emphasized	0	716	1219	0.00%	100.00%
16	Transitivity	0	716	1219	0.00%	100.00%
17	Rational	0	218	1717	0.00%	100.00%
18	Declension and Conjugation	0	218	1717	0.00%	100.00%
19	Unaugmented and Augmented	0	346	1589	0.00%	100.00%
20	Number of root letters	0	336	1599	0.00%	100.00%
21	Verb root	0	721	1214	0.00%	100.00%
22	Nouns finals	121	478	1457	8.30%	91.70%

7.4.6 Manual proofreading and correction of the mapped SALMA tags

I manually proofread and corrected the mapped morphological features tags. The result of correcting the automatically mapped morphological features tags is a sample gold standard for evaluating morphological analyzers and part-of-speech taggers for Arabic text. Constructing the gold standard for evaluating morphological analyzers is one of the objectives of evaluating the SALMA – Tag Set. The gold standard is stored in different formats and published online[54] to allow the wider Arabic NLP community to use it in evaluating morphosyntactic systems for Arabic. Chapter 9 discusses in detail the construction and the specifications of the SALMA – Gold Standard. Figure 7.8 shows an example of mapping from the QAC into SALMA tags, the results after applying steps 1 to 4, the results after applying step 5 and the results after manually correcting the tags.

[54] The SALMA Gold Standard http://www.comp.leeds.ac.uk/sawalha/goldstandard.html

	QAC morpheme tag	SALMA tags after mapping steps 1-4	SALMA tags after mapping step 5	Corrected SALMA tags
الم	POS:INL	**p--?-----????---?-----**	**p--?-----s-s----------**	**p--b-----s-s----------**
أَ	A:INTG+	**p--i-----s------------**	**p--i-----s------------**	**p--i-----s------------**
حَسِبَ	POS:V PERF 3MS	**v-p---mst--?-?-??????-**	**v-p---msts-f-ambhvsta-**	**v-p---msts-f-amohvsta-**
ال	Al+	**r---d-----------------**	**r---d-----------------**	**r---d-----------------**
نَاسُ	POS:N MP NOM	**n?----mp-?n??---????-?**	**n?----mp-vndd---ndst-s**	**n#----mj-vndd---hdst-s**
أَنْ	POS:SUB	**p--g-------?----------**	**p--g-----s-s----------**	**p--g-----s-s----------**
يُ	NULL	**r---a-----------------**	**r---a-----------------**	**r---a-----------------**
تْرَكُ	POS:V IMPF PASS 3MP MOOD:SUBJ	**v-c---mptda?-p???????-**	**v-c---mptdao-pmbhvtta-**	**v-c---mptdao-pmohvtta-**
وا	PRON:3MP	**r---r-mptsnw----------**	**r---r-mptsnw----------**	**r---r-mpts-s----------**
أَنْ	POS:SUB	**p--g-------?----------**	**p--g-----s-s----------**	**p--g-----s-s----------**
يَ	NULL	**r---a-----------------**	**r---a-----------------**	**r---a-----------------**
قُولُ	POS:V IMPF 3MP MOOD:SUBJ	**v-c---mptda?-????????-**	**v-c---mptdao-amohvtto-**	**v-c---mptdao-amohvtto-**
وا	PRON:3MP	**r---r-mptsnw----------**	**r---r-mptsnw----------**	**r---r-mpts-s----------**
آمَنَّ	POS:V PERF (IV) 1MP	**v-p---mpf--?-?-??????-**	**v-p---mpfs-s-amohvttc-**	**v-p---mpfs-s-amohvttc-**
نَا	PRON:1MP	**r---r-xpfs??----------**	**r---r-xpfs??----------**	**r---r-xpfs-s----------**
وَ	wa+	**p--c------------------**	**p--c------------------**	**p--c-----s-f----------**
هُمْ	POS:PRON 3MP	**np----mpt--??---?-----**	**np----mpts-si---hn---?**	**np----mpts-si---hn----**
لَا	POS:NEG	**p--n-------?----------**	**p--n-----s-s----------**	**p--n-----s-s----------**
يُ	NULL	**r---a-----------------**	**r---a-----------------**	**r---a-----------------**
فْتَنُ	POS:V IMPF PASS 3MP	**v-c---mpt-??-p???????-**	**v-c---mptdnn-pmohvtta-**	**v-c---mptdnn-pmohvtta-**
ونَ	PRON:3MP	**r---r-mp?snn----------**	**r---r-mp?snn----------**	**r---r-mpts-f----------**

Figure 7.8 A Sample of the QAC tags and their mapped SALMA tags after applying the mapping procedure's steps 1-4, step 5 and manually correcting the tags.

7.5 Evaluation of the Mapping Process

The correction process of the automatically mapped tags involves correcting the individual morphological feature categories tags of each morpheme. This process specifies whether a morphological feature category is applicable or not. If it is applicable, the automatically mapped attribute is checked and corrected. Otherwise, if it is not applicable and the mapped tag is not "-", the correction will replace any attribute by "-". During the correction process, the following types of correction were observed.

- Changing the automatic tag from "-", to the correct tag of a certain morphological feature attribute.
- Changing the automatic tag from "?", to the correct tag of a certain morphological feature attribute.

- Changing an automatic tag which is not "-" or "?", to the correct tag of a certain morphological feature attribute.
- Changing the automatic tag from "?", to "-" where a given morphological feature is not applicable to a given morpheme.
- Changing an automatic tag which is not "-" or "?", to "-" where a given morphological feature is not applicable to a given morpheme.

Depending on the above observed correction types and the standard definitions of accuracy metrics[55], the rules for measuring the accuracy of the mapping process were inferred. The following classifications of the different cases of the corrected SALMA tags are used as bases to measure the accuracy of the mapping process.

- **TN**: True and not applicable; case was not applicable and predicted not applicable.
- **TP**: True and applicable; case was applicable and predicted correctly.
- **FN**: False and not applicable; case was not applicable and predicted applicable.
- **FP**: False and applicable; case was applicable and predicted not applicable.

The accuracy metrics of the automatically mapped tags are based on the above observations to calculate the recall, precision and accuracy. Accuracy is the percent of predictions where were correct. Formula [2] illustrates the computation of accuracy.

$$\text{Accuracy} = \frac{\text{TN+TP}}{\text{Total number of morphemes}} \;\ldots\ldots\ldots\; (2)$$

Recall is defined as the percentage of applicable cases that are correctly mapped from the mapped cases. Formula [3] illustrates the computation of recall.

$$Recall = \frac{\text{TP}}{\text{TP+FN}} \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\; (3)$$

Precision is defined as the percentage of the applicable cases which are correctly predicted from the total number of the applicable cases. Formula [4] illustrates the computation of precision.

$$\text{Precision} = \frac{\text{TP}}{\text{number of applicable cases}} \;\ldots\ldots\ldots\ldots\; (4)$$

Table 7.3 shows accuracy, recall and precision after applying the first four mapping steps and after applying the fifth mapping step. Figures 7.9, 7.10 and 7.11 show the increase in accuracy, recall and precision after using the procedures of linguistic rules, for mapping the QAC morphological tags to the SALMA tags.

[55] Standard definition of Recall and Precision http://en.wikipedia.org/wiki/Recall_and_precision

Table 7.3 Accuracy, recall and precision of the mapping procedure after steps 4 and 5

Category	Mapping steps 1-4			Mapping steps 1-5		
	Accuracy	Recall	Precision	Accuracy	Recall	Precision
Main part-of-speech	72.30%	100.00%	72.30%	97.99%	99.43%	97.99%
Part-of-speech: Noun	58.96%	99.16%	46.81%	86.15%	99.16%	46.81%
Part-of-speech: Verb	87.18%	99.62%	99.62%	99.95%	99.62%	99.62%
Part-of-speech: Particle	83.73%	100.00%	88.37%	96.24%	98.03%	86.63%
Part-of-speech: Other	72.45%	30.84%	19.31%	94.90%	95.50%	86.43%
Punctuation marks	100.00%	-	-	100.00%	-	-
Gender	71.11%	100.00%	79.11%	89.03%	97.66%	88.72%
Number	63.13%	100.00%	64.82%	79.09%	97.09%	70.91%
Person	79.40%	100.00%	96.23%	94.28%	96.11%	89.02%
Inflection	15.65%	100.00%	22.04%	88.47%	95.30%	86.73%
Case and Mood	18.54%	100.00%	75.31%	79.71%	99.56%	94.98%
Case and Mood marks	0.41%	100.00%	0.58%	74.25%	94.20%	66.11%
Definiteness	16.68%	100.00%	12.96%	96.40%	100%	88.46%
Voice	67.97%	100.00%	5.38%	98.61%	100%	89.62%
Emphasis	68.07%	100.00%	6.15%	99.95%	100%	99.62%
Transitivity	67.25%	0.00%	0.00%	99.69%	100%	98.45%
Rationality	6.59%	0.00%	0.00%	94.34%	100%	86.68%
Declension and conjugation	34.65%	95.65%	2.89%	90.11%	99.83%	75.03%
Unaugmented and augmented	33.37%	0.00%	0.00%	95.21%	98.56%	86.19%
Number of root letters	33.42%	0.00%	0.00%	99.74%	100%	100%
Verb root	73.84%	0.00%	0.00%	100.00%	100%	100%
Noun finals	46.96%	0.00%	0.00%	93.31%	100%	97.64%

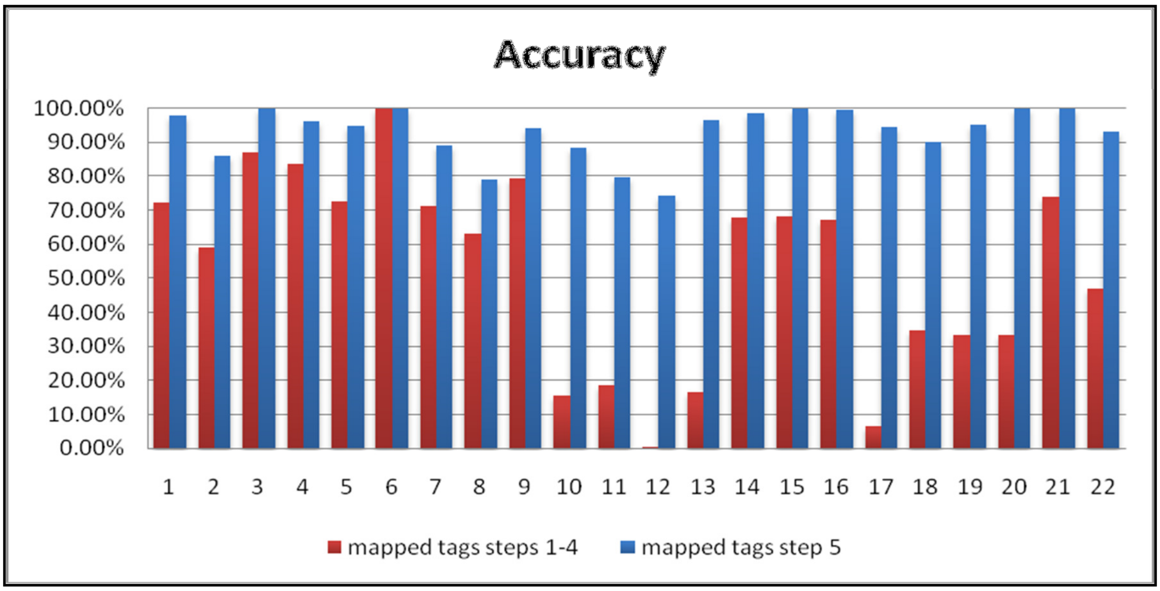


Figure 7.9 Accuracy of mapping after steps 4 and step 5 of mapping QAC to SALMA tags

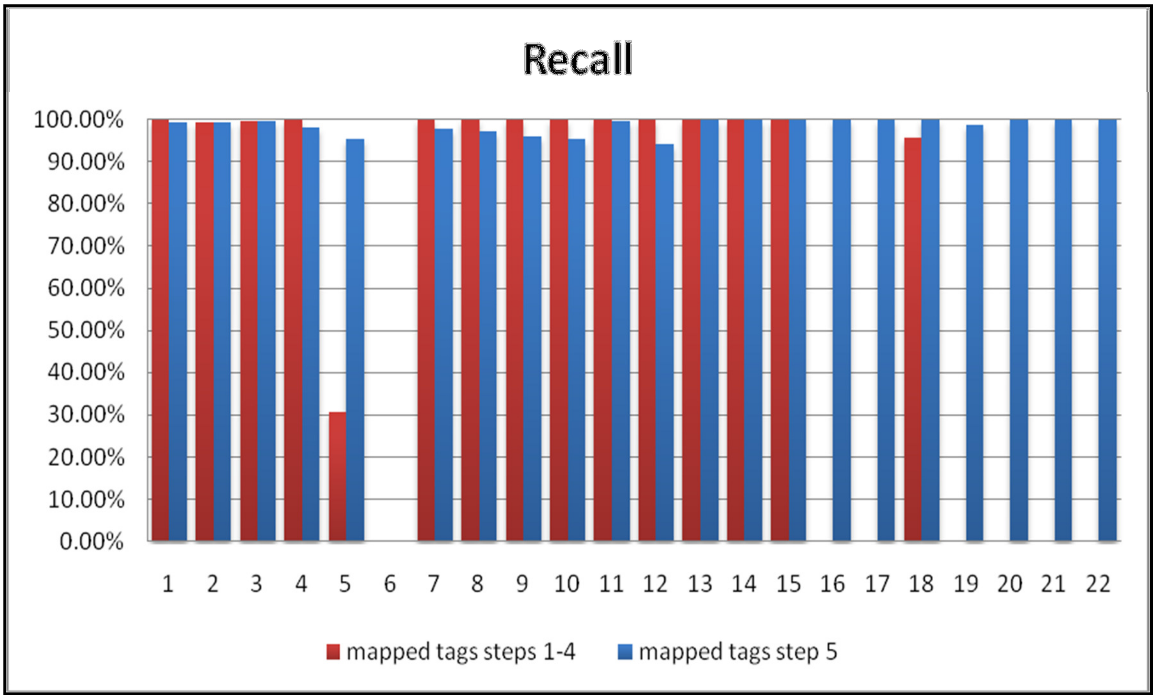


Figure 7.10 Recall of mapping after steps 4 and step 5 of mapping QAC to SALMA tags

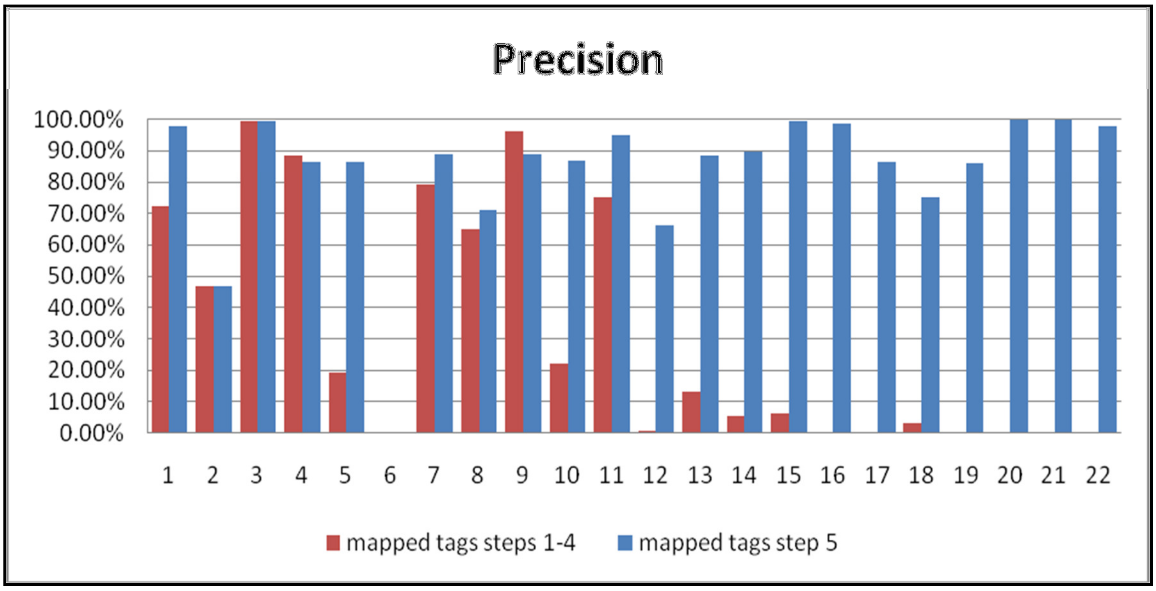


Figure 7.11 Precision of mapping after steps 4 and step 5 of mapping QAC to SALMA tags.

7.6 Discussion of Evaluation of the SALMA Tag Set

Arabic has a complex morphology and fine-grain tag assignment is significantly challenging. Arabic words should be decomposed into five parts: proclitics, prefixes, stem or root, suffixes and enclitics. The morphological analyzer should add appropriate linguistic information to each of these parts of the word. Instead of a tag for the whole word, sub-tags are required for each part. More detailed morphological feature information that describes each part of the word is generally more useful and appreciated.

The software engineering principle of reuse was applied to build a morphologically tagged corpus enriched with detailed analysis of each word's morphemes, by recycling an existing morphologically tagged corpus, the Quranic Arabic Corpus (QAC). This chapter demonstrated that this resource can be reused and enriched with detailed analysis by mapping the existing morphological analysis of a sample chapter of the QAC to the detailed morphological analysis using the SALMA – Tag Set and the SALMA – Tagger. This empirical study was achieved by following a 6-step procedure which involves direct mapping of the existing features and building a rule-based system which depends on the linguistic knowledge extracted from traditional Arabic grammar books.

A measure of accuracy is "exact match". The exact match of the prediction of all 22 features for a morpheme whole tags for the test sample is 53.5%, but some of the errors were very minor such as replacing one '?' by '-'. The error-rate of individual features scored 2.01% for main part of speech, between 3% and 15% for morphological features coded in the QAC tags, and between 2% and 24% for features which do not exist in the QAC tags but can be automatically guessed. Due to the use of 22 morphological features categories for each morpheme, which increase the potential for making annotation mistakes, this result demonstrates that the reuse and enriching of existing resource with more detailed morphological features information is applicable and can provide tagged Arabic corpora with fine grain analysis.

7.7 Conclusions and Summary

A range of Arabic Part-of-Speech taggers exist, each with a different tag set. The existing tag sets for Arabic were illustrated and compared, and this suggests the need for a common standard to simplify and promote comparisons and sharing of resources. Generic design criteria for corpus tag sets were reviewed in chapter 5. Some of these principles have been applied in existing tag sets; but there is still room for improvement, in the design of a theory-neutral standard tag set for Arabic Part-of-Speech taggers and tagged corpora. The SALMA – Tag Set captures long-established traditional morphological features of Arabic, in a compact yet transparent notation. A tag consists of 22 characters; each position represents a feature and the letter at that location represents a value or attribute of the morphological feature; the dash '-' represents a feature not relevant to a given word. The SALMA – Tag Set is not tied to a specific tagging algorithm or theory, and other tag sets could be mapped onto this standard, to simplify and promote comparisons between and reuse of Arabic taggers and tagged corpora. The SALMA – Tag Set design decisions were made through chapter 6.

The SALMA – Tag Set has been validated in two ways. First, it was validated by proposing it as a standard to the Arabic language computing community, and has been adopted in Arabic language processing systems. The SALMA – Tag Set has been used in the SALMA – Tagger to encode the morphological features of each morpheme (Sawalha and Atwell 2009a; Sawalha and Atwell 2010b). Parts of The SALMA – Tag Set were also used in the Arabic morphological analyzer and part-of-speech tagger Qutuf (Altabbaa et al. 2010). Moreover, the SALMA – Tag Set has been reported as a standard for evaluating morphological analyzers for Arabic text and for building a gold standard for evaluating morphological analyzers and part of speech taggers for Arabic text (Hamada 2010).

Second, an empirical approach to evaluating the SALMA – Tag Set of Arabic showed that it can be applied to an Arabic text corpus, by mapping from an existing tag set to the more detailed SALMA – Tag Set. The morphological tags of a 1000-word test text, chapter 29 of the Quranic Arabic Corpus, were automatically mapped to SALMA tags. Then, the mapped tags were proofread and corrected. The result of mapping and correction of the SALMA tagging of this corpus is a new Gold Standard for evaluating Arabic morphological analyzers and part-of-speech taggers with a detailed fine-grain description of the morphological features of each morpheme, encoded using SALMA tags.

We invite other Arabic language computing researchers to take up the SALMA – Tag Set and the SALMA – Gold Standard tagged corpus, to promote comparability and interoperability of Arabic morphological analyzers and Part-of-Speech taggers.

Part IV: Tools and Applications for Arabic Morphological Analysis

Chapter 8
The SALMA Tagger for Arabic Text

This chapter is based on the following sections of published papers:

Section 3 is expanded from section 2 in (Sawalha and Atwell 2009b) and section 3.2 in (Sawalha and Atwell 2009a)

Section 5 is based on section 3 in (Sawalha and Atwell 2010b)

Chapter summary

Morphological analyzers and part-of-speech taggers are key technologies for most text analysis applications. The main aim of this thesis is to develop a morphosyntactic tagger for annotating a wide range of Arabic text formats, domains and genres including both vowelized and non-vowelized text. Enriching the text with linguistic analysis will maximize the potential for corpus re-use in a wide range of applications. We foresee the advantage of enriching the text with part-of-speech tags of very fine-grained grammatical distinctions, which reflect expert interest in syntax and morphology, but not specific needs of end-users, because end-user applications are not known in advance.

This chapter describes the fine-grained Arabic morphological analyzer algorithm, the SALMA – Tagger. The SALMA – Tagger is adherent to an agreed standard of the ALECSO/KACST initiative for designing and evaluating morphological analyzers for Arabic text. The SALMA Tagger is enriched with dictionaries: SALMA – ABCLexicon, pre-stored lists of clitics and affixes, roots, patterns dictionary, function words list, and other linguistic lists such as broken plural list and proper noun list.

The SALMA – Tagger combines sophisticated modules that break down complex morphological analysis problem into achievable tasks which each address a particular problem and also constitute stand-alone units. These modules are: the SALMA – Tokenizer, the SALMA – Lemmatizer and Stemmer, the SALMA – Pattern Generator, the SALMA – Vowelizer and the SALMA – Tagger module. These modules are useful as stand-alone tools which users can select and/or customise to their own applications.

8.1 Introduction

A morphological analyzer is a program which analyzes words. It extracts the root from the derived word and/or generates all possible words from a certain root. It analyzes the word into morphemes by dividing the word into proclitics, prefixes, stem or root, suffixes and enclitics. Moreover, it identifies the word's part of speech and generates the correct derivation pattern of the analyzed word.

Morphological analysis is defined as the process of analysing a word in its orthographic form, and generates all possible analyses of the analysed word. The morphological analyser, a program that does the morphological analysis of the word, must generate all possible analyses and identify the morphological features for each morpheme of the analysed word. The morphological features should be encoded using a specified scheme- morphological features tags, which can be used by higher level text analytics applications such as part-of-speech tagging and parsing. Moreover, morphological analysis involves extracting the root and matching the pattern of the word. Morphological analysers can be used to add the correct vowelization (diacritics) for each letter of the analysed word.

Section 2.3 in chapter 2 has more background on morphological analysis for Arabic text.

8.2 Specifications and Standards of Arabic Morphological Analyses

A robust and well-designed morphological analyser for Arabic text has to meet agreed design standards for Arabic morphological analyses. Many researchers have investigated the morphology of Arabic, and they built their morphological analysers according to specific application requirements. For instance, stemming involves morphological analyses for Arabic words where the outputs of the stemmers are the roots of the analysed words (Al-Sughaiyer and Al-Kharashi 2004). However, the complex morphology of Arabic requires more detailed analyses. Therefore, the morphological analyser for Arabic text should meet the following requirements (Al-Bawaab 2009; Hamada 2009b; Hamada 2010).

1. It can correctly divide the analysed word into morphemes such as proclitics, prefixes, stem or root, suffixes and enclitics.
2. It can generate the correct pattern of the word and specify whether the generated pattern is a noun pattern, verb pattern or both.
3. It can correctly specify the morphological features for each morpheme.
4. It can extract the correct root of the word whether it is triliteral or quadriliteral.

5. It can deal with unambiguous words (inert or stop words), irregular words, rare words and borrowed words.
6. If an orthographic form is ambiguous, it should generate a set of plausible/possible analyses to be disambiguated at a subsequent processing stage taking context into account.
7. It allows the rules of transitive and intransitive verbs to be specified.
8. It allows the derivation rules of perfect verbs, imperfect verbs and imperative verbs to be specified.
9. It can deal with the orthographic features of words such as vowelizing, incorporation, substitution and the writing of *hamza*^{h}. This helps in correcting spelling mistakes.

The most widely-agreed and recent specification and standard is the ALECSO/KACST initiative on morphological analysers for Arabic text; see section 2.3.4.7. The organization and the institution invited specialized researchers on morphological analysers for Arabic text to present their morphological analysers, to agree on the design and development specifications and standards, and to agree on an evaluation methodology for the different morphological analysers. This section will discuss the ALECSO/KACST initiative. The ALECSO/KACST design specifications and standards will be followed in the design of the SALMA – Tagger.

8.2.1 ALECSO/KACST Initiative on Morphological Analyzers for Arabic Text

This section discusses our experience in developing and evaluating morphological analysers for Arabic text. The section analyses an exemplar of how the community should work together to advance the field. The exemplar is The Arab League Educational, Cultural and Scientific Organization (ALECSO) and the King Abdul-Aziz City of Science and Technology (KACST) initiative on morphological analysers of Arabic text[56] which aims to encourage research on developing open-source morphological analysers for Arabic text, which are of high accuracy, easy to use and can be integrated into higher levels of applications for processing Arabic text.

The ALECSO/KACST initiative contains recommendations and standards for designing morphological analysers. These recommendations are written as papers appearing in the workshop proceedings (Al-Bawaab 2009; Hamada 2009b; Zaied 2009). It also includes agreed specifications for developing morphological analysers represented by the participants' papers and presentations. Moreover, the initiative includes an evaluation methodology and criteria for evaluating the outputs of the morphological

[56] ALECSO/KACT initiative on morphological analyzers for Arabic text http://www.alecso.org.tn/index.php?option=com_content&task=view&id=1234&Itemid=1002&lang=ar

analysers. ALECSO/KACST organized a competition between the participants' analyzers. *AlKhalil* morphological analyzer (Boudlal et al. 2010) was announced as the winner of the competition. However, these design specifications and standards, evaluation methodology and the results of the competition have not been widely publicized. Hamada (2010) reported the evaluation methodology in Arabic only. Another aim of this section is to publicize these important specifications, standards, methodology and the competition to the English-speaking Arabic NLP community.

8.2.2 ALECSO/KACST Prerequisites for a Good Morphological Analyser for Arabic Text

The ALECSO/KACST design specifications and standards stated some essential prerequisites of robust morphological analysers for Arabic text. These prerequisites involve dealing with clitics, affixes, roots, patterns, non-inflected words, non-conjugated verbs and primitive nouns (Hamada 2009a). This requires the morphological analyser to have comprehensive lists that cover the information. Having these morphological lists previously stored within the morphological analyser will meet the first five general requirements of the Arabic morphological analyser. These prerequisites as described by (Hamada 2009a) are:

- A list of all prefixes, such as definite article, subject prefix, etc.
- A list of all suffixes, such as feminine *nūn*, masculine sound plural letters, etc.
- A list of all patterns, such as فَعَلَ *fa‘ala*, فَعُول *fa‘ūl*, مَفَاعِيْل *mafa‘īl,* etc.
- A list of all triliteral and quadriliteral roots.
- A list of non-inflected words, non-conjugated verbs and primitive nouns.

Moreover, the lists of prefixes and suffixes need to be classified into noun affixes, verb affixes and affixes which are common between nouns and verbs.

8.2.3 ALECSO/KACST: Design Recommendations

The ALECSO/KACST initiative for morphological analysis for Arabic text has specified the general design specifications and standards as recommendations for the developers of morphological analyzers for Arabic text. These recommendations include recommendations for the inputs of the morphological analyzer, the analysis process, and the outputs of the morphological analyzer. The following subsections discuss these design recommendations as described by Al-Bawaab (2009).

8.2.3.1 ALECSO/KACST: Design Recommendations of Inputs

A well-designed morphological analyzer for Arabic text can accept a single word, a sentence, or a text as inputs. The morphological analyser should provide analyses for each word of an input sentence or text.

Moreover, the morphological analyser should accept the input word(s) to be fully vowelized, partially vowelized or non-vowelized. In order to deal with the different word vowelization variations, the morphological analyzer should contain special functions that can generate the non-vowelized form of the input word(s), preserve the vowelization, and deal with the specific orthographic challenges of the Arabic word such as *šadda*^{h}.

8.2.3.2 ALECSO/KACST: Design Recommendations of Analysis

An Arabic word form may be assigned several analyses due to the absence of vowelization and the treatment of the word out of its context. Then the number of analyses differs from word to word. Because the morphological analyser analyzes the words out of their context, it should produce all possible analyses of each word form.

Arabic words are classified into nouns, verbs and particles. Due to the absence of vowelization words can share noun or verb properties. Thus ورد *wrd* can be وَرْدٌ *ward*^{un} "roses" representing a noun or وَرَدَ *warada* "to come" representing a verb. The word can be a noun or particle. An example is رب *rb* where رَبٌّ *rubb*^{un} "God" is a noun, while رُبَّ *rubba* "many" is a particle. The word can be a verb and particle as in عدا *'dā*; عَدَا *'adā* "ran" is a verb, while عَدَا *'adā* "except" is a particle. The word can also be a noun, verb and particle as in بل *bl*; بَلٌّ *ball*^{un} "moistering" is a noun; بَلَّ *balla* "to moisten, wet, make wet" is a verb; بَلْ *bal* "nay, -rather …, (and) even, but, however, yet" is a particle.

Therefore, the analyser assumes that the analyzed word is noun, verb and particle then follows certain procedures to analyze verbs, nouns and particles, to extract morphological features specified below.

A- Analyzing verbs

The morphological analyzer must extract the following information assuming the analyzed word is a verb.

1- **Verb prefixes:** a one-letter or two-letter prefix can be attached to the beginning of the verb. Thus in وَكَتَبَ *wakataba* "and he wrote" وَ+كَتَبَ *wa+kataba* has a one letter prefix وَ *wa* "and" representing a conjunction particle; and in وَسَيَكْتُبُ *wasayakubu* "and he will write" وَسَ+يَكْتُبُ *wasa+yaktubu* has a two letter prefix consisting of وَ *wa* "and" representing a conjunction particle and سَ *sa* "will" representing a particle of futurity. The equivalent feature-numbers in the SALMA – Tag Set are 4 and 5.

2- **Verb suffixes:** These are the subject-suffix pronouns and the object-suffix pronouns. The verb suffix can be one of the suffixed pronouns or a combination of both types of pronouns. For example, the verb قَرَأْتُ *qara'tu* "I have read" has تُ *tu* as a subject-suffix pronoun. The verb عَلَّمَهَا *'allamahā* "he taught her" has هَا *hā* "her" as an object-suffix pronoun, and the word زَوَّجْنَاكَهَا *zawwağnākahā* "we have let you marry her" has نَا *nā* "we" as a subject-suffix pronoun, كَ *ka* "you" as a first object-suffix pronoun, and هَا *hā* "her" as a second object-suffix pronoun. The equivalent feature-number in the SALMA – Tag Set is 5.

3- **Verb subcategory:** the morphological analyser should specify the subcategory of the analyzed verb. The analyzed verb can be a perfect verb, imperfect verb or imperative verb. The analyzed verb can share properties of two or three verb subcategories as in أكرم *'akrm*. Here أَكْرَمَ *'akrama* "treated reverentially with hospitably" is a perfect verb; أُكْرِمُ *'ukrimu* "I treat reverentially with hospitably" is an imperfect verb; and أَكْرِمْ *'akrim* "You! Treat reverentially with hospitably" is an imperative verb. The equivalent feature-number in the SALMA – Tag Set is 3.

4- **The pattern of the verb:** the morphological analyser extracts the correct pattern of the verb. For example the verb اسْتَقَامَ *'istaqāma* "straighten" is an augmented triliteral verb which has the pattern اسْتَفْعَلَ *'istaf'ala*. Some verbs can have more than one pattern. Thus يُقَال *yuqāl* has the pattern يَفْعُلُ *yaf'ulu* then it means "said", and the pattern يُفْعِلْ *yuf'il* when it means "been sacked".

5- **The root of the verb:** the morphological analyzer specifies the correct root for the analyzed verb. For example, يَرِثُ *yariṯu* "he inherits" has the root و ر ث *w-r-ṯ*, the imperative verb قُلْ *qul* "You! Say" has the root ق و ل *q-w-l*, and the imperative verb قِ *qi* "You! Protect" has the root و ق ي *w-q-y*.

6- **Verb augmentation:** the morphological analyser specifies whether the verb is unaugmented, augmented by one letter, augmented by two letters or augmented by three letters. It also specifies whether the verb has a triliteral root or quadriliteral root. For instance, the verb عَلَّمَ *'allama* "he taught" is a triliteral verb augmented by one letter. The verb اطْمَأَنَّ *'iṭma'anna* "he reassured" is quadriliteral verb augmented by two letters. The equivalent feature-number in the SALMA – Tag Set for verb augmentation is 20, and for number of root letters 21.

7- **Person morphological feature:** the morphological analyser determines whether the analyzed verb is first person, second person or third person depending on the subject-suffix pronouns and whether the short vowels appear on the analyzed verb. The verb لاحظْت *lāḥaẓtu* "I have noticed" is a first person verb. The verb لاحَظْتَ *lāḥaẓta* "You have noticed" is a second person verb. And the verb لاحَظَتْ *lāḥaẓat*

"She has noticed" is a third person verb. The equivalent feature-number in the SALMA – Tag Set is 10.

8- **Voice morphological feature:** the morphological analyser determines whether the analyzed verb is active voice or passive voice. For example, the verb يُصَارُ *yuṣāru* "has become" is an imperfect passive verb. The equivalent feature-number in the SALMA – Tag Set is 15.

9- **The mood marks:** the morphological analyser determines the mood marks of the analyzed verb. The mood marks of the verb can be a short vowel (*i.e. fatḥa*^{h}*, ḍamma*^{h}*, sukūn*), a letter (*i.e. nūn*), or omission (*i.e.* omission of vowel letter). The equivalent feature-number in the SALMA – Tag Set is 13.

10- **Full vowelization:** the morphological analyser adds the correct full vowelization to the analyzed verb whatever the original vowelization of the input verb.

B) Analyzing nouns

The morphological analyser should extract the following morphosyntactic information assuming the analyzed word is a noun.

1- **Noun prefixes:** the noun prefix consists of one to five letters. The prefix letters can be homographic with the noun original letters (*i.e.* the root radicals of the noun). *E.g.* بِطَاقَات *biṭāqāt*; can be analyzed بِ+طَاقَات *bi+ṭāqāt* "with the abilities" where the first letter the preposition بِ *bi* "with" is a prefix, or بِطَاقَات *biṭāqāt* "cards" without any prefix. The equivalent feature-number in the SALMA – Tag Set is 4.

2- **Noun suffixes:** genitive suffixed pronouns are the most common suffixes of nouns. The suffix letters can be a suffix on the noun or on underlying letter of the noun. *E.g.* the word فكه *fkh* can be analyzed فَكُّ+هُ *fakkuhu* "his jaw" where هُ *hu* is a suffix, or as فَكِهٌ *fakih*^{un} "humorous" which has the root ف ك ه *f-k-h* and lacks any suffix. The equivalent feature-number in the SALMA – Tag Set is 5.

3- **The pattern of the noun:** the morphological analyser specifies the pattern of the analyzed noun. *E.g.* the pattern of the noun بِنَاء *binā'* "building" is فِعَال *fi'āl*, the pattern of the noun سَيِّد *sayyid* "master" is فَيْعِل *fay'il*, and the pattern of the word أَكُفٌّ *akuff*^{un} "hands" is أَفْعُلٌ *'af'ul*^{un}.

4- **The root of the noun:** the morphological analyzer extracts the root of the analyzed noun. E.g. اسْم *'ism* "name" has the root س م و *s-m-w*, حَيْوَان *ḥaywān* "animal" has the root ح ي ي *ḥ-y-y*, and مِيْنَاء *mīnā'* "port" has the root و ن ي *w-n-y*.

5- **Noun sub-category:** Arabic language scholars classified Arabic words into three main categories, namely noun, verb and particle. This classification is coarse-grained. More details are needed to distinguish the sub-categories of nouns, verbs and particles. The sub-categories of nouns include: common nouns, proper nouns, relative pronouns, demonstrative pronouns, nouns of time and place, adjectives, adverbs, etc. There is no agreement between part-of-speech tag sets of Arabic text on the sub-categories of nouns. The CATiB tag set groups nominals such as nouns, pronouns, adjectives and adverbs into one tag NOM, and gives proper nouns a specific tag PROP. The PATB Full tag set distinguishes between NOUN (common noun), ADJ (adjective), ADV (adverb) and NOUN_PROP (proper noun). The QAC tag set has four categories to tag nouns. These are nouns (N noun, PN proper noun, IMPN imperative verbal noun), pronouns (PRON personal pronoun, DEM demonstrative pronoun, REL relative pronoun), nominals (ADJ adjective, NUM number) and adverbs (T time adverb, LOC location adverb). (See section 5.3 for more details about part-of-speech tag sets of Arabic text). The SALMA Tag Set classifies nouns into 34 sub categories at position 2 which include more descriptions of inflected and non-inflected noun categories. See section 6.2.2 for the details of the part-of-speech subcategories of noun. ALECSO/KACST design recommendations for morphological analysis for Arabic text distinguish between 18 noun subcategories. Table 8.1 shows the subcategories of nouns with examples.

Table 8.1 The 18 subcategories of nouns with examples

	Noun subcategory			**Example**
1	Primitive noun	اسْم جَامِد	’ism ğāmid	كِتَاب *kitāb* “book”
2	Active participle	اسم الفاعل	*’ism al-fā’il*	ضارِب *ḍārib* ‘hitter’
3	Passive participle	اسم المفعول	*’ism al-maf’ūl*	مَضْرُوب *maḍrūb* ‘Struck’
4	Noun of place	اسم المكان	*’ism al-makān*	مَكْتَب *maktab* ‘office’
5	Noun of time	اسم زمان	*’ism zamān*	مَطْلِع *maṭla‘* start time
6	Adjective	الصِّفة المشبَّهة	*aṣ-ṣifa*^{h} *al-mušabbaha*^{h}	طويل *ṭawīl* ‘tall’
7	Instrumental noun	اسم الآلة	*’ism al-‘āla*^{h}	مِنْشار *minšār* ‘saw’
8	Gerund / Verbal noun	المصدر الأصْلِي	*al-maṣdar al-aṣlī*	ضَرْب *ḍarb* ‘hitting’
9	Gerund of profession	المصدر الصناعي	*al-maṣdar al-ṣinā‘ī*	فُروسيَّة *furūsiyya*^{h} ‘horsemanship’
10	Gerund of instance	مصدر المرَّة	*maṣdar al-marra*^{h}	نَظْرَة *naẓra*^{h} ‘one look’
11	Gerund of state	مصدر الهيئة	*maṣdar al-hay’a*^{h}	جِلْسَة *ğilsa*^{h} ‘sitting position’
12	Proper noun	اسم العلم	*’ism al-‘alam*	فاطِمَة *fāṭima*^{h} ‘Fatima’
13	Gerund/ verbal noun with initial *mīm*	المصدر الميمي	*al-maṣdar al-mīmī*	مَوعِد *maw ‘id* ‘date’
14	Elative noun	اسم تفضيل	*’ism tafḍīl*	أفضل *’afḍal* ‘better’
15	Intensive Active participle	مبالغة اسم الفاعل	*mubālağa*^{t} *’ism al-fā’il*	جَرَّاح *ğarraḥ* ‘surgeon’
16	Generic noun	اسم الجنس	*’ism al-ğins*	حِصان *hiṣān* ‘horse’
17	Plural generic noun	اسم جنس جمعي	*’ism ğins ğam’ī*	تفاح tuffāḥ ‘apple’
18	Collective noun	اسم جمع	*’ism ğam’*	قوم *qawm* ‘folk’

6- **The Morphological Features of Inflectional Morphology:** Most Arabic nouns are declined nouns. However, some nouns are non-declined because they are generated from certain patterns, or they satisfy certain conditions. For example, the noun مَدَارِس *madāris* "schools" is non-declined because it has the pattern مَفَاعِل *mafā'il*. And the noun إِبْرَاهِيْم *'ibrāhīm* "Abraham" is non-declined because it is not an Arabic proper name. The equivalent feature-number in the SALMA – Tag Set is 11.

7- **The Morphological Feature of Gender:** the morphological analyser specifies the gender of the analyzed noun; for example قَمَر *qamar* "moon" is masculine; شَمْس *šams* "sun" is feminine; and طَرِيْق *ṭarīq* "road" is of common gender. The equivalent feature-number in the SALMA – Tag Set is 7.

8- **The Morphological Feature of Number:** the morphological analyser recognizes the number of the analyzed noun whether it is singular, dual or plural. For example, the noun عَصَوَان *'aṣawān* "two sticks" is dual and its singular is عَصَا *'aṣā* "one stick"; the noun أَرْضُون *'arḍūn* "earths" is the plural form of the noun أَرْض *'arḍ* "earth"; and the noun صَحْرَاوَات *ṣaḥrāwāt* "deserts" is the plural of the noun صَحْرَاء *ṣaḥrā'* "desert". The equivalent feature-number in the SALMA – Tag Set is 8.

9- **The Relative and Diminutive Nouns:** the morphological analyser specifies the noun sub-categories of relative and diminutive nouns. For example, the noun خَلَوِيّ *ḫalawyy* "cellular" is a relative noun of خَلِيَّة *ḫalyya*^{h} "cell"; and the noun عُصَيَّة *'uṣayya*^{h} "small stick" is a diminutive of عَصَا *'aṣā* "stick". The equivalent feature-number in the SALMA – Tag Set is 2.

10- **The Case Mark:** the morphological analyzer specifies the case of the analyzed noun and the correct case mark. The case mark can be a short vowel (*i.e. fatḥa*^{h}*, ḍamma*^{h}*, kasra*^{h}*, sukūn*) or a letter (*i.e. 'alif, wāw, yā'*). For example, أَبَا *'abā* "father" is an accusative noun which has *'alif* as case mark; فَلاَّحُوْنَ *fallāḥūna* "peasants" is a nominative noun which has *wāw* as case mark because it is a masculine sound plural; حَذَارِ *ḥaḏāri* "beware" is an invariable verb-like noun marked by *kasra*^{h}. The equivalent feature-number in the SALMA – Tag Set is 13.

11- **Vowelization of nouns:** the morphological analyser adds the full vowelization to the analyzed noun regardless of the original vowelization of the input noun. For example, some of the vowelized variations of the non-vowelized noun المدرسة *al-mdrs*^{t} are; الْمَدْرَسَة *al-madrasa*^{t} "the school"; الْمُدَرِّسَة *al-mudarrisa*^{t} "the female-teacher"; الْمُدَرَّسَة *al-mudarrasa*^{t} "the female-student", etc.

C) Analyzing Particles

The morphological analyser assumes that the analyzed word is a particle and extracts the following information:

1- **The Prefix of the Particle**: the particle's prefix consists of one letter such as وإِذَا *wa'iḏā* "and if" where وَ *wa* is a prefixed conjunction, or two letters such as فَلَرُبَّمَا *falarubbamā* "and perhaps" where the two letters فَلَ *fala* at the beginning of the particle represent the prefix.

2- **The suffix of the particle:** the suffixes are the genitive suffixed pronouns such as عَنْكُمَا *'ankumā* "about both of you".

3- **The Inflectional Morphology Mark:** particles are always invariable. The result of analyzing particles shows the inflectional morphology mark of particles. For example, حَيْثُ *ḥayṯu* "where (*adv.*)" has the mark *ḍamma*^{h}; بَلْ *bal* "nay, -rather …, (and) even, but, however, yet" has the mark *sukūn*; and سَوْفَ *sawfa* "will" has the mark *fatḥa*^{h}.

8.2.3.3 ALECSO/KACST: Design Recommendations of Outputs

The output should include all possible analyses of the analyzed word, assuming the analyzed word is verb, noun and particle. The recommended morphosyntactic information, discussed above, represents the core information that is displayed in the outputs of the morphological analyzer. As described by the ALCSO/KACST initiative, figure 8.1 shows examples of the output verb analyses; figure 8.2 shows examples of the output noun analyses; and figure 8.3 shows examples of the output particle analyses.

***w'dt = wa'adtu = wa'ad+tu* "I promissed"**	**وعدت = وَعَدْتُ = وَعَدْ + تُ**
Perfect verb with active voice Unaugmented, has the pattern *fa'ala yaf'ul* and has the root (*w-'-d*) Invariable verb has *sukūn* as inflectional morphology mark Third person verb which has a singular subject of common gender The suffix is subject suffixed pronoun *tā'*	فعل ماض، معلوم مجرد، على وزن (فَعَلَ يَفْعُلْ) من الجذر (و ع د) مبني على السكون مسند إلى المتكلِّم المفرد متصل بضمير الرفع (ت)
w'dt = *wa'adta* = *wa'ad+ta* "You (*masc.*) promissed"	وعدت = وَعَدْتَ = وَعَدْ + تَ
w'dt = *wa'adti* = *wa'ad+ti* "You (*fem.*) promissed"	وعدت = وَعَدْتِ = وَعَدْ + تِ
w'dt = *wa'adat* = *wa'ada+t* "She promissed"	وعدت = وَعَدَتْ = وَعَدَ + تْ
w'dt = *wu'idtu* = *wu'id+tu* "I have been promissed"	وعدت = وُعِدْتُ = وُعِدْ + تُ
w'dt = *wa'udtu* = *wa+'ud+tu* "And I have returned back"	وعدت = وَعُدْتُ = وَ+عُدْ + تُ
w'dt = *wa'addat* = *wa+'adda+t* "she counted"	وعدت = وَعَدَّتْ = وَ+عَدَّ + تْ

Figure 8.1 Examples of the output verb analyses

***wmfṣlk = wamafṣiluka = wa+mafṣilu+ka* “And your joint”**	**ومفصلك = وَمَفْصِلُكَ = وَ + مَفْصِلُ + كَ**
Prefix وَ *wa* “And”	السابقة (و)
mafṣilu, is a masculine noun has the pattern (maf‘il) and the root (*f-ṣ-l*)	مَفْصِلُ، اسم مذكر على وزن (مَفْعِل) من الجذر (ف ص ل)
Is in nominative case and has the *ḍamma*^{h} case mark	مرفوع وعلامة رفعه الضمة
Is connected to the genitive suffixed pronoun *kāf*	متصل بضمير الجر (ك)
wmfṣlk = *wamafṣiluki* = *wa+mafṣilu+ki* “And your (*fem.*) joint”	ومفصلك = وَمَفْصِلُكِ = وَ + مَفْصِلُ + كِ
wmfṣlk = *wamifṣiluka* = *wa+mifṣilu+ka* “And your (*masc.*) tongue”	ومفصلك = وَمِفْصِلُكَ = وَ + مِفْصِلُ + كَ
wmfṣlk = *wamufṣiluka* = *wa+mufṣilu+ka* “And your (*masc.*) separator”	ومفصلك = وَمُفْصِلُكَ = وَ + مُفْصِلُ + كَ
wmfṣlk = *wamufṣṣiluka* = *wa+mufṣṣilu+ka* “And your interpreter”	ومفصلك = وَمُفْصِّلُكَ = وَ + مُفْصِّلُ + كَ

Figure 8.2 Examples of the output noun analyses

***fmnkm = faminkum = fa+min+kum* “and among you”**	**فمنكم= فَمِنْكُمْ = فَ + مِنْ + كُمْ**
The prefix is فَ fa “and”	السابقة (ف)
مِنْ *min* “among” is a preposition, Invariable particle, and *sukūn* is its inflectional morphology mark	(مِنْ) حرف جر ، مبني على السكون
It is connected to the genitive suffix pronoun كُمْ *kum* “you”	متصل بضمير الجر (كُمْ)

Figure 8.3 Examples of the output particle analyses

8.2.4 Discussion of ALECSO/KACST Recommendations

The ALECSO/KACST recommendations for designing an Arabic morphological analyzer are morphological descriptions of the analyzed words. These linguistic descriptions involve variant analyses of the analyzed word, such as assuming the word is a noun, verb and particle, then analyzing the word according to that assumption. The descriptions clarify the tokenization of the analyzed word into morphemes, where the prefix letters or suffix letters can be homographic with the original letters of the analyzed word. Therefore, different analyses can be produced by tokenizing the word into different morphemes. The recommendations provide information about the morphological features of the analyzed words. They provide 11 morphological features for nouns and 10 morphological features for verbs. They also provide information about the root, pattern, prefixes, suffixes and vowelization of the analyzed words.

On the other hand, the ALECSO/KACST recommendations lack the description of how to encode the morphological features of the analyzed words in a machine-readable way. The recommendations are not specific to a morphosyntactic tag set, and they do not provide intermediate coding to enable mapping of different morphosyntactic tagging schemes. The classification by linguists of morphological features of nouns, verbs and other information such as root, pattern and affixes does not prioritise these features, so that order of presentation can be exploited as procedural steps in the development of the morphological analyzer.

8.3 The SALMA – Tagger Algorithm

The SALMA – Tagger algorithm involves several processing steps for Arabic text. These steps, described below, are executed sequentially where each step depends on the previous one. Intermediate results can be obtained from each processing step. Figure 8.4 shows the steps and module components of the SALMA – Tagger.

The SALMA – Tagger was developed according to the long-established Arabic grammar knowledge extracted from traditional Arabic grammar books. It also has the SALMA – ABCLexicon as a main component for extracting the root of the word, and for finding the different vowelization variations of the analyzed words. The SALMA – Tagger depends on the SALMA – Tag Set as a design standard. The SALMA design standard for morphological analysis of Arabic includes the ALCESO/KACST design recommendations and standards.

However, the SALMA standards for designing fine-grained morphological analysis for Arabic text are more detailed, and adherent to standards of global computational linguistic knowledge and traditional Arabic grammar. The SALMA standards are not tied to a specific application, as user needs are not known yet. The standards are designed to be general purpose, can be integrated into different levels of applications, and different tag sets can be mapped to this standard to allow reusability and comparability between these different morphosyntactic annotation schemes.

Following the ALECSO/KACST recommendations convention, inputs, analysis process and outputs are described in this section. The morphological analyzer accepts a single Arabic word, a sentence or an Arabic text document, whether they are vowelized, partially vowelized, or non-vowelized, as inputs to the system.

The SALMA – Tagger is a morphological analyser that consists of five components. Each component can be a standalone text analytics application that performs a specific task, and they work together to process the input text and provide all morphological information of each analysis of the analyzed words. Sections 8.3.1 to 8.3.5 will discuss the component modules of the SALMA – Tagger.

The outputs of morphological analyser are the full analyses of the words from the analyzed text. Full analysis means all possible analyses of the word such as all possible roots, clitics, affixes, stems, lemmas, patterns, different forms of vowelization, and the morphological features of each analysis represented by a morphological tag using the SALMA – Tag Set. The subsections of section 8.3 will discuss the outputs of each tagger's components. Section 8.6 discusses the output formats of the SALMA Tagger.

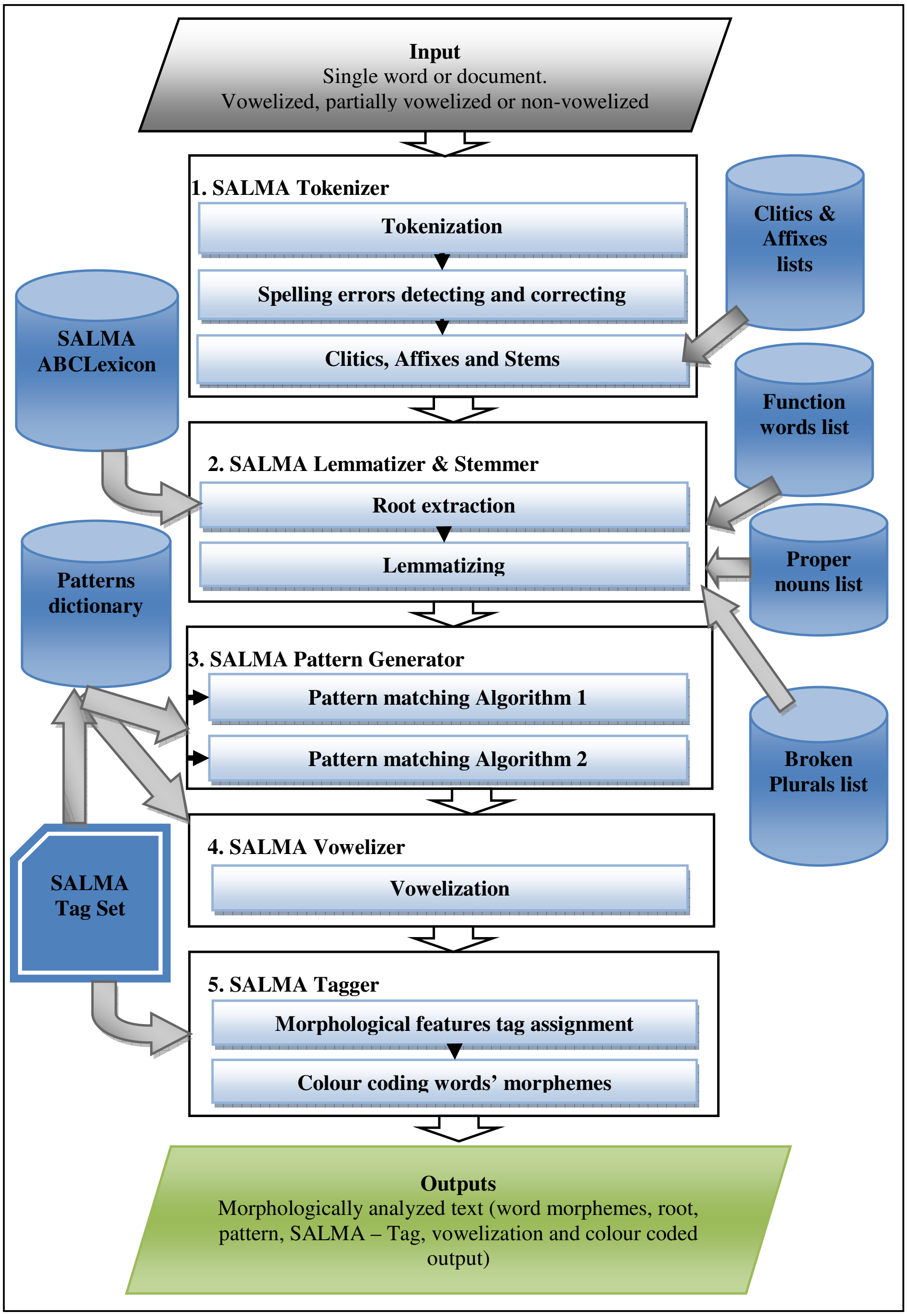

Input
Single word or document.
Vowelized, partially vowelized or non-vowelized
1. SALMA Tokenizer
Tokenization
Spelling errors detecting and correcting
Clitics, Affixes and Stems
Clitics & Affixes lists
SALMA ABCLexicon
Function words list
2. SALMA Lemmatizer & Stemmer
Root extraction
Lemmatizing
Proper nouns list
Patterns dictionary
3. SALMA Pattern Generator
Pattern matching Algorithm 1
Pattern matching Algorithm 2
Broken Plurals list
4. SALMA Vowelizer
Vowelization
SALMA Tag Set
5. SALMA Tagger
Morphological features tag assignment
Colour coding words' morphemes
Outputs
Morphologically analyzed text (word morphemes, root, pattern, SALMA – Tag, vowelization and colour coded output)

Figure 8.4 The SALMA Tagger algorithm

8.3.1 Module 1: SALMA – Tokenizer

The first module of the SALMA – Tagger is the SALMA – Tokenizer. The main task of this module is to split the input running text into tokens. Then, the tokens are decomposed into morphemes (Attia 2007; Attia 2008). The SALMA – Tokenizer has three main parts. Each part is important for analyzing Arabic text. The Tokenization part deals with the input text files, determines what is considered an Arabic word, and stores the Arabic word in a unified format that enables the other components to deal with the word whether the word is fully vowelized, partially vowelized or non-vowelized. The Spelling Errors Detection and Correction part checks the spelling of the tokenized words and corrects the spelling of the words if the word letters do not match certain patterns. The Word Segmentation part is responsible for generating all possible variant morpheme tokenizations of the analyzed word. This part mainly depends on matching the affixes and clitics of the analyzed word and comprehensive lists of affixes and clitics. The following sections discuss these parts in detail.

8.3.1.1 Step 1, Tokenization

In this section; Buckwalter's transliteration scheme is used in the example as it illustrates 1-to-1 mapping between Arabic letters and diacratics and their equivelant in Roman letters. The tokenizer program uses the NLTK regular expression tokenizer to tokenize the input text into Arabic words, punctuation marks, currency tokens, numbers, words written in Latin letters, and HTML/XML tags. The regular expression tokenizer uses regular expression patterns that suit the Arabic text. Then the tokenizer processes the extracted Arabic words, by resolving the doubled letters الحروف المضعَّفة *al-ḥurūf al-muḍa‘‘afa*^{h} and the extensions المدّ *al-madd*. The doubled letter marked by *šadda*^{h} الشَّدَّة is replaced by two letters similar to the original letter; the first is silent marked by *sukūn*, and the second is vowelized by the same short vowel as appears on the original letter. For example the word وَصَّى *waṣṣā waS~aY* has the doubled letter ص *ṣ S* and after processing it will be in the form وَصْصَى *waṣṣā waSoSaY* “He enjoined”. The extension المدّ *al-madd* (آ) is replaced by (*hamza*^{h}) and *’alif*, as in the word آمَنُوا *’āmanū |manuwA* “They believed” which will be in the form ءامَنُوا *’āmanū ’AmanuwA*.

Only one short vowel can be associated with any letter of the word. Based on this fact, a unified data structure to store Arabic words was designed. This data structure consists of a list of tuples of size two, where each tuple stores the letter in the first position and the short vowel (if it is present) at the second position. And so on for all letters and short vowels of the word. The data structure is represented as [(**C**,**V**), (**C**,**V**),…,(**C**,**V**)], where **C** represents a consonant and **V** represents a short vowel. Figure 8.5 shows the data structure storing the words وَصْصَى *waSoSaY* and ءامَنُوا *’āmanū ‘AmanuwA*. This data structure is also used to match the word and the patterns.

Position	**0**		**1**		**2**		**3**		**4**		**5**	
وَصْصَى	و	◌َ	ص	◌ْ	ص	◌َ	ى	–				
waSoSaY	**w**	**a**	**S**	**o**	**S**	**a**	**Y**	**-**				
ءامَنُوا	ء	–	ا	–	م	◌َ	ن	◌ُ	و	–	ا	–
‘AmanuwA	**‘**	**-**	**A**	**-**	**m**	**a**	**n**	**u**	**w**	**-**	**A**	**-**

Figure 8.5 The word data structure

Figure 8.6 shows a tokenized sentence of chapter 29 of Qur'an. It shows the original fully vowelized word. Then the tokenizer module produces three variations of the analyzed word; the non-vowelized word, the processed word extracted from the unified word's data structure, and the processed non-vowelized word.

Word			Non-vowelized	Processed vowelized word	Processed non-vowelized word
أَمْ	*'am*	Or	أم >m	أَمْ >amo	أم >m
حَسِبَ	*ḥasiba*	Think	حسب Hsb	حَسِبَ Hasiba	حسب Hsb
الَّذِينَ	*al-lḏīna*	those who	الذين Al*yn	الْلَذِينَ Alola*iyna	اللذين All*yn
يَعْمَلُونَ	*ya 'malūna*	do	يعملون yEmlwn	يَعْمَلُونَ yaEomaluwna	يعملون yEmlwn
السَّيِّئَاتِ	*as-sayyi'āt*	evil deeds	السيئات Alsy}At	السَيْيِئَاتِ Alsayoyi}aAti	السييئات Alsyy}At
أَن	*'an*	that	أن >n	أَن >an	أن >n
يَسْبِقُونَا	*yasbiqūnā*	they can outrun us	يسبقونا ysbqwnA	يَسْبِقُونَا yasobiquwnaA	يسبقونا ysbqwnA
سَاء	*Sā'a*	Evil is	ساء sA'	سَاء saA'	ساء sA'
مَا	*mā*	what	ما mA	مَا maA	ما mA
يَحْكُمُونَ	*yaḥkumūn*	they judge	يحكمون yHkmwn	يَحْكُمُونَ yaHkumuwna	يحكمون yHkmwn

Figure 8.6 A sample output of the tokenization module component after processing the Qur'an , chapter 29

8.3.1.2 Step 2, Spelling Errors Detection and Correction

A large number of potential spelling errors are to be expected because of a variety of word processing tools with different spelling conventions that are used to generate Arabic text. Most word processing tools that support Arabic are not aware of what letter and diacritic combinations can appear on a letter in a given position of the word. Therefore, it is the responsibility of the editor (person) who should check the word's spelling while writing a document or a authoring a web page.

The absence of such a special module in the word processing tools that support Arabic increases the potential for mis-spelling Arabic words. Such spelling errors include adding more than one short vowel to the same letter; starting the word with *taṭwīl,* a special character that is used to extend the Arabic word; adding a diacritic to *taṭwīl* (also considered a spelling error). Another type of constraint that the word processing tools should deal with is whether a certain diacritic can appear on a letter in a given position in the word. This constraint has many rules such as; a word cannot start with a 'silent' letter, (*i.e. sukūn* cannot appear on the first letter of the word). A Similar rule is *tanwīn*, which appears only on the last letter of the word.

The algorithm divides the Arabic word into three parts; the front part consisting of the first letter and any diacritics appearing on it; the middle part consisting of the letters starting from the second letter till the letter before the last and their diacritics; and the rear part which consists of the last letter and its diacritics. Each part has its own valid letter-diacritics combinations. The front part is checked if it matches the following 3 valid letter-diacritic combinations [(*letter* + *šadda*^{h} + *a short vowel*[57]*), (letter + a short vowel), (letter)]*. Each letter-diacritic combination from the middle part is checked if it matches the following 5 valid letter-diacritic combinations; [(*letter* + *šadda*^{h} + *a short vowel), (letter + a short vowel), (letter + sukūn), (letter), (taṭwīl)]*. The rear part is checked if it matches one of the following letter-diacritic valid combinations [(*letter* + *šadda*^{h} + *a short vowel),* (*letter* + *šadda*^{h} + *tanwīn), (letter + a short vowel), (letter + sukūn), (letter + tanwīn), (letter)]*. Figure 8.7 shows an example of applying the letter-vowelization templates to the analyzed word. The matching templates are highlighted in bold.

Word	Rear	Middle part				Front
سَيَّارَةٌ *sayyāra*^{tun} “Car” Letter vowelization templates	ةٌ	رَ	ا	ـ	يَّ	سَ
	1) Letter + *tanwīn*	1) Letter + Short vowel	2) Letter	*4) taṭwīl*	5a) Letter + *šadda*^{h} (O) + short vowel	1) Letter + Short vowel
	2) Letter + *sukūn* 3) Letter 4) Letter + *šadda*^{h} + *tanwīn* 5) Letter + *šadda*^{h} + a short vowel		3) Letter + *sukūn*		5b) letter + *šadda*^{h} (ph) + short vowel	2) Letter 3) Letter + *šadda*^{h} (ph) + short vowel

Figure 8.7 Example of applying letter-vowelization templates to a word. The matching templates are highlighted in bold.

8.3.1.3 Step 3, Word Segmentation (Clitics, Affixes and Stems)

For each tokenized Arabic word, a special module divides the word into three parts: proclitics and prefixes, stem/root, and suffixes and enclitics. The first part is matched against a list of proclitics and prefixes consisting of 220 entries, and the third part is matched with a list of suffixes and enclitics consisting of 474 entries. Only the analyses that match both of the lists of clitics and affixes are taken as candidate analyses.

8.3.1.4 Which Segmentation to Use?

Several morphological systems exist for Arabic text. These systems apply tokenization to the input text because tokenization is an essential prerequisite. However,

[57] Short vowels are *fatḥa*^{h}, *ḍamma*^{h} and *kasra*^{h} [(◌َ) (◌ُ), (◌ِ)]

these systems do not describe the tokenization decisions. Only Attia (2007); also Attia (2008) described the tokenization of Arabic as a challenge which needs more investigation.

The SALMA Standard decomposes the tokens (word) into five parts: proclitics; prefixes; stem; suffixes; and enclitics. Each part can be a single part or multiple of more than one clitic or affix, except there is only one stem in a word. This fine-grain decomposition is required by the SALMA – Tag Set. Then, a SALMA – Tag is assigned to each morpheme.

The distinction between affixes and clitics can be confusing. Clitics and affixes are defined as follows:

> *"…affixes carry morpho-syntactic features (such as tense, person, gender or number), while clitics serve syntactic functions (such as negation, definition, conjunction or preposition) that would otherwise be served by an independent lexical item.*" (Attia, 2008 p. 59)

This definition distinguishes between the morphosyntactic features of affixes and the syntactic functions of the clitics. The SALMA standard bases the definition of the clitics and affixes on the patterns of the words where the morphosyntactic features of affixes and the syntactic functions of the clitics are preserved as defined by Attia (2008). Affixes are the morphemes shared between the word and its pattern, and clitics are the word's morphemes that do not match morphemes of the pattern. Therefore, suffixed pronouns can be classified as suffixes if they are subject pronouns. On the other hand, they are classified as enclitics if they are object-suffix pronouns or genitive-suffix pronouns. This classification is based on patterns, where subject-suffix pronouns are part of the pattern. Subject-suffix pronouns carry morphosyntactic features (*i.e.* gender, number and person) of the verb, while object-suffix pronouns and genitive-suffix pronouns serve syntactic functions (*e.g.* object of the verb) that can be expressed by an independent lexical item. Figure 8.8 shows an example of tokenization of some words.

فرمت *frmt*	فرمت	*farmata* "he formatted"	وهم *whm*	وهم	*wahm* "delusive imagination"
	فرم + ت	*faram+ti* "you (2*SF*) chopped"		و+هم	*wa+hum* "and they"
	ف + رم + ت	*fa+ ram+t* "you (2*SF*) throwed "	أمس *'ms*	أمس	*'ams* "yesterday"
حسب *ḥsb*	حسب	*ḥasaba* "he computed"		أ + مس	*'a+ massa* "did he touched?"
تسربل *tsrbl*	ت + سربل	*ta+sarbala* "he dressed"	يسر *ysr*	يسر	*yasir* "ease, prosperity"
وراثة *wirāṯa^t*	وراث + ة	*wirāṯa* + ^{t} "inheretance"		ي + سر	*ya+sirru* "he telld a secret"
زوجناكها *zwǧnākhā*	زوج + نا + ك + ها	*zawwaǧ+nā+ka+hā* "we allowed you to marry her"			

Figure 8.8 Example of tokenization of some words

8.3.1.5 Constructing the Clitics and Affixes Dictionaries

Using traditional Arabic language grammar books (Dahdah 1987; Dahdah 1993; Wright 1996; Al-Ghalayyni 2005; Ryding 2005), lists of **proclitics** (*e.g.* conjunctions, prepositions, vocative particles, interrogative particles, particle of futurity, definite article[58]), **prefixes** (*e.g.* imperfect prefix, imperative prefix), **suffixes** (*e.g.* relative *yā'*, emphatic *nūn*, *nūn* of protection, dual letters, masculine sound plural letters, feminine sound plural letters), and **enclitics** (*e.g.* suffixed pronouns, *tā' marbūṭa*^{h}, *tā'* of feminization, *tanwīn*) were constructed. These lists were provided to a generating program which generates all the possible combinations of proclitics and prefixes together, and suffixes with enclitics. The generated lists of these combinations were extremely large because the generation process produced all possible combinations of proclitics and prefixes; and suffixes and enclitics. These generated lists were checked by analyzing words in four corpora; the Qur'an text corpus, the Corpus of Contemporary Arabic, the Penn Arabic Treebank, and the Corpus of Traditional Arabic Dictionaries. Then, two lists were constructed; first, a list of proclitics and prefixes containing 220 entries, and second, a list of suffixes and enclitics containing 474 entries.

Khoja's stemmer contains 11 prefixes and 28 suffixes (Khoja 2003). BAMA has a prefixes file containing 299 prefixes and a suffixes file containing 618 suffixes. BAMA provides a morphological compatibility table containing 598 prefix-suffix combinations (Maamouri and Bies 2004; Maamouri et al. 2004). The *Alkhalil* morphological analyzer has 65 prefixes and 65 suffixes. The prefixes and suffixes are stored in separate XML files (Boudlal et al. 2010).

The clitics and affixes dictionaries add more morphosyntactic features to each entry. The entry is compound (*i.e.* consists of one or multiple clitics or affixes representing distinct morphemes). Instead of one tag for the clitic and affix entry, multiple tags were added. Each part (morpheme) is assigned a SALMA – Tag where the morphological features of that part are encoded. The nature of that part whether it is a proclitic (*proc*), a prefix (*pref*), a suffix (*suf*) or an enclitic (*enc*) is distinguished. Whether that part is part of a pattern or not is also determined. This information is useful for tokenization and pattern matching. The prefix-stem-suffix agreement is illustrated by adding the main part-of-speech information for each part. ***n*** indicates that part of clitic and affix entry can be used on a *noun* stem and other noun clitics and affixes parts. ***v*** indicates verb part. And ***x*** indicates the part is either noun or verb.

[58] The definite article *al-* is classified as proclitic because it does not appear in the patterns and it is not part of the underlying letters of the word. The definite article *al-* is also different than other proclitics such as prepositions and conjunctions because *al-* cannot appear as a stand-alone morpheme.

Figures 8.9 and 8.10 show samples of these lists with the morphosyntactic information added to each entry in the list.

Prefix	**Example**	**Morphemes**	**SALMA – Tag**	**Morpheme type**	**Stem POS**	**Part of pattern**	**Description**
من **mn**	منقلبة **mnqlibp**	من **mn**	r---p-----------------	pref	n	y	زيادة في أول الكلمة Prefix
فاست **fAst**	فاستبقوا **fAstbqwA**	ف **f**	p--c------------------	proc	x	n	حرف عطف Conjunction
		است **Ast**	r---p-----------------	pref	v	y	زيادة في أول الكلمة prefix
كالمت **kAl**	كالمتعجب **kAlmtEjb**	ك **k**	p--l------------------	proc	n	n	حرف تشبيه Simile particle
		ال **Al**	r---d-----------------	proc	n	n	أداة تعريف Definite article
		مت **mt**	r---p-----------------	pref	n	y	زيادة في أول الكلمة Prefix
أفبال **>fbAl**	أفبالباطل **>fbAlbATl**	أ **>**	p--i-----s------------	proc	x	n	حرف استفهام Interrogative particle
		ف **f**	p--c------------------	proc	x	n	حرف عطف Conjunction
		ب **b**	p--p------------------	proc	n	n	حرف جر Preposition
		ال Al	r---d-----------------	proc	n	n	أداة تعريف Definite article

Figure 8.9 Sample of the proclitics and prefixes with their morphological tags, attributes and descriptions

Suffix	Example	Morphemes	SALMA Tag	Morpheme type	Stem POS	Part of pattern	Description
هم **hm**	كتابهم **ktAbhm**	هم	r---r-mpts-s----------	enc	x	n	ضمير متصل (مذكر، جمع، غائب) في محل (نصب أو جر)
		hm					Suffixed pronoun (MP3)
ني **ny**	عَلَّمَنِي **Eallamany**	ن	r---n----s-s----------	enc	v	n	نون الوقاية
		n					*Nūn* of protection
		ي	r---r-xsfs-s----------	enc	x	n	ضمير متصل(مفرد ، متكلم) في محل نصب
		y					Suffixed pronoun (XS2)
تماناها **tmAnAhA**	أَعْطَيْتُمَانَاهَا **>ETytmAnAhA**	تما	r---r-xdss-s----------	suf	v	y	ضمير متصل (مثنى، مخاطب)في محل رفع
		tmA					Suffixed pronoun (XD1)
		نا	r---r-x?fs-s----------	suf	v	y	ضمير متصل(جمع، متكلم) في محل رفع
		nA					Suffixed Pronoun (XP1)
		ها	r---r-fsts-s----------	enc	x	n	ضمير متصل (مذكر، مفرد ، غائب) في محل جر
		hA					Suffixed pronoun (MS3)
انيتك **Anytk**	إنسانيتك	ان	r---s-----------------	suf	n	y	زيادة في آخر الكلمة
	>nsAnytk	**An**					Suffix
		ي	r---y-----------------	enc	n	n	ياء النسبة
		y					Relative *yā'*
		ت	r---f-fs-s-s----------	suf	n	y	تاء التأنيث (منقلبة عن تاء مربوطة)
		t					*tā'* of femininization
		ك	r---r-xsss------------	enc	x	n	ضمير متصل (مفرد، مخاطب) في محل نصب أو جر
		k					Suffixed pronoun (XS2)
اً **F**	ذَهَباً ***hbAF**	اً	r---k-------i---------	suf	n	y	تنوين
		F					*tanwīn*

Figure 8.10 Sample of the suffixes and enclitics with their morphological tags, attributes and descriptions

8.3.1.6 Matching the Affixes and Clitics with the Word's Segments

The analyser divides the word into three parts of different sizes. Then it searches the proclitics and prefixes list for the first part, and the suffixes and enclitics list for the third part. If the first or the third parts are found in the lists, the morphosyntactic information associated to the prefix or suffix is assigned to these parts. Then the analyzer selects the analyses of the word where the first part matches one of the proclitics and prefixes from the list, and the third part matches one of the suffixes and clitics from the list. Table 8.2 shows the process of matching prefixes and suffixes and the process of selecting the candidate analyses.

The selection of the candidate analyses maintains the prefix-stem-suffix agreement. At this stage, the main part of speech of the stem is still unavailable. However, agreement is maintained between the part of speech information of the proclitics, prefixes, suffixes and enclitics. For example, the analysis ي **y** + عمل **Eml** + ون **wn** is accepted because the first part ي **y** is found in the proclitics and prefixes list, and the third part ون **wn** is found in the suffixes and enclitics list. However, the analysis يع **yE** + م **m** + لون **lwn** is not accepted because the first part يع **yE** and the third part لون **lwn** are not found in the clitics and affixes lists. The main part of speech of the stem can be predicted at this stage.

Table 8.2 Example of the process of selecting the matched clitics and affixes

Word		First Part		Second Part		Third Part		Possible analyses
يَعْمَلُونَ	*yaEomaluwna*			**يعملون**	***yEmlwn***			**Candidate analysis**
يَعْمَلُونَ	*yaEomaluwna*			**يعملو**	***yEmlw***	**ن**	***n***	**Candidate analysis**
يَعْمَلُونَ	*yaEomaluwna*			**يعمل**	*yEml*	**ون**	***wn***	**Candidate analysis**
يَعْمَلُونَ	*yaEomaluwna*			يعم	*yEl*	لون	*lwn*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*			يع	*yE*	ملون	*mlwn*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*			ي	*y*	عملون	*Emlwn*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*	**ي**	***y***	**عملون**	***Emlwn***			**Candidate analysis**
يَعْمَلُونَ	*yaEomaluwna*	**ي**	***y***	**عملو**	***Emlw***	**ن**	***n***	**Candidate analysis**
يَعْمَلُونَ	*yaEomaluwna*	**ي**	***y***	**عمل**	***Eml***	**ون**	***wn***	**Candidate analysis**
يَعْمَلُونَ	*yaEomaluwna*	ي	*y*	عم	*Em*	لون	*lwn*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*	ي	*y*	ع	*E*	ملون	*mlwn*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*	يع	*yE*	ملون	*mlwn*			Not accepted
يَعْمَلُونَ	*yaEomaluwna*	يع	*yE*	ملو	*mlw*	ن	*n*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*	يع	*yE*	مل	*ml*	ون	*wn*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*	يع	*yE*	م	*m*	لون	*lwn*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*	يعم	*yEm*	لون	*lwn*			Not accepted
يَعْمَلُونَ	*yaEomaluwna*	يعم	*yEm*	لو	*lw*	ن	*n*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*	يعم	*yEm*	ل	*l*	ون	*wn*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*	يعمل	*yEml*	ون	*wn*			Not accepted

Figure 8.11 shows an example of prefix-stem-suffix agreement between parts of the analyzed word. The suffix ون *wn* has two entries in the suffixes and enclitics dictionary. The first entry represents subject a suffixed pronoun which is a verb suffix. The second is the masculine plural suffix, which is a noun suffix. The prefix-stem-suffix agreement is valid between the the imperative prefix ي *y* and the subject suffixed pronoun where both

are verb affixes. On the other hand, agreement is not satisfied between the imperative prefix and the masculine plural suffix. The prefix-stem-suffix agreement can distinguish the main part-of-speech of the stem عمل *Eml* as a verb.

Analyzed word	يَعْمَلُونَ **yaEomaluwna** *ya‘malūna* “They work”								
	Prefix				*Stem*	*Suffix*			
Possible tokenization	ي **y**				**عمل** **Eml**	ون **wn**			
Affixes information	r---a----------------	pref	**v**	y	Match	r---r-mp?s-f----------	suf	**v**	y
					No match	r---m-mp-s-f----------	enc	**n**	n

Figure 8.11 Example of prefix-stem-suffix agreement between a word’s morphemes

8.3.2 Module 2: SALMA- Lemmatizer and Stemmer

Stemming and lemmatizing have been widely used in several fields of natural language processing. Stemming is the process of assigning morphological variants of words to equivalence classes, such that each class corresponds to a single stem. It is also defined as reducing inflected words to their stem, base, or root form. Lemmatizing is the process of grouping a set of words into the canonical form, dictionary form, or citation form which is also called the *lemma*. *E.g.*, in English, *run, runs, ran* and *running* are forms of the same lexeme, with *run* as the lemma[59].

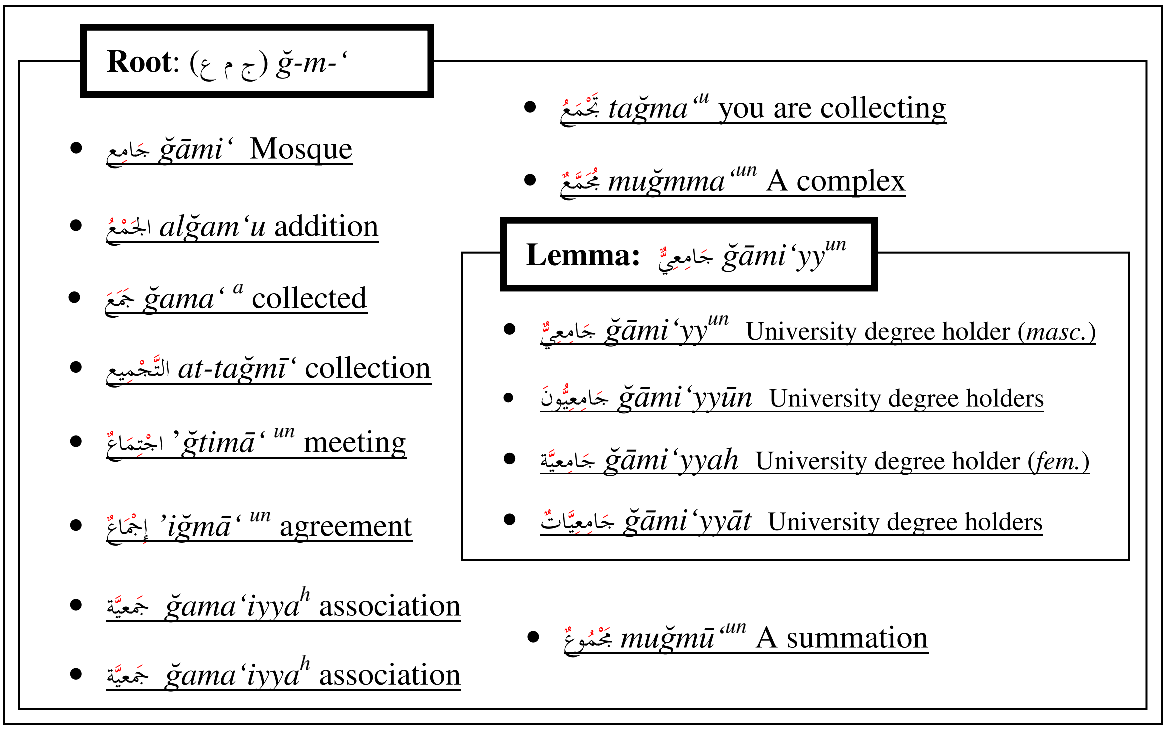
Chapter 3 discusses the comparative evaluation of three existing stemming algorithms and morphological analyzers: Khoja’s stemmer (Khoja 2003); Buckwalter’s morphological Analyzer (BAMA) (Buckwalter 2002); and Al-Shalabi et. al’s, triliteral root extraction algorithm (Al-Shalabi et al. 2003). The comparative evaluation shows that all stemming algorithms involved in the experiments agreed and generate correct analysis for simple roots that do not require detailed analysis. But they make mistakes in analysis of complex cases. So, more detailed analysis and enhancements are recommended. Most stemming algorithms are designed for information retrieval systems where accuracy of the stemmers is not an important issue. On the other hand, accuracy is vital for natural language processing. The accuracy rates show that the best algorithm failed to achieve an accuracy rate of more than 75%. This proves that more research is required.

A breakdown of the percentage of triliteral roots, words and word types’ distribution on 22 categories of triliteral roots was depicted. The study clearly showed that about 35% of any Arabic text words have roots which belonging to the defective or defective and hamzated root categories. Words which belong to these two root categories are hard to analyze and the root extraction process of such words always has higher error rates than

[59] Definition of Lemma from Wikipedia http://en.wikipedia.org/wiki/Lemma_(linguistics)

words which belong to the intact root category. Section 3.7 discusses the details of the analytical study of Arabic triliteral roots.

A lemma in Arabic is different from the root. The root represents the 3 to 5 letter underlying form of the word, while the lemma is the canonical form that can be used as a head word in a dictionary. Lemmatizing an Arabic word produces the singular form of nouns and the third person masculine perfect form of verbs. This requires removing the clitics attached to the beginning and the end of the word; recognizing the number of nouns and dealing with both sound and broken plural; and feminine sound plural nouns require replacing the feminine sound plural letters ات *āt* with ة *tā’ marbūta*^{h} to extract the lemma. Figure 8.12 shows a set of words sharing the same root and lemma.



Root: (ج م ع) *ǧ-m-‘*

- جامِع *ǧāmi‘* Mosque
- الجَمْعُ *alǧam‘u* addition
- جَمَعَ *ǧama‘*^{a} collected
- التَّجْميع *at-taǧmī‘* collection
- اجْتِمَاعٌ *’ǧtimā‘*^{un} meeting
- إِجْمَاعٌ *’iǧmā‘*^{un} agreement
- جَمعيَّة *ǧama‘iyya*^{h} association
- جَمعيَّة *ǧama‘iyya*^{h} association
- تَجْمَعُ *taǧma‘*^{u} you are collecting
- مُجَمَّعٌ *muǧmma‘*^{un} A complex

Lemma: جَامِعِيٌّ *ǧāmi‘yy*^{un}

- جَامِعِيٌّ *ǧāmi‘yy*^{un} University degree holder (*masc.*)
- جَامِعِيُّونَ *ǧāmi‘yyūn* University degree holders
- جَامِعِيَّة *ǧāmi‘yyah* University degree holder (*fem.*)
- جَامِعِيَّاتٌ *ǧāmi‘yyāt* University degree holders

- مَجْمُوعٌ *muǧmū‘*^{un} A summation

Figure 8.12 Example set of words grouped to root and lemma

8.3.2.1 The Use of the SALMA ABCLexicon

The SALMA – ABCLexicon, as discussed in chapter 4, is a broad-coverage lexical resource which provides prior knowledge to support the development and to improve the accuracy of morphological analysis. The SALMA – ABCLexicon is constructed by extracting information from disparate formats and merging 23 traditional Arabic lexicons by following agreed criteria for constructing morphological lexical resources from raw text. The SALMA – ABCLexicon contains 2,774,866 word-root pairs representing 509,506 different vowelized words and 261,125 different non-vowelized words.

The SALMA – ABCLexicon is stored in three alternative formats: XML files, a relational database; and tab-separated column files. The lexicon is provided with a search facility that enables searching for a certain lexical entry in the lexicon, to return an object `LexiconEntry` representing an encapsulation of the word and its root. A specialized interface is provided to enable the morphological analyzer to communicate with the lexicon file. The dictionary data structure of the lexicon is in this format:

```
Lexicon = [nv_word:[LexiconEntry,...],...]
```

The `Lexicon` class interface represents the actual lexicon data and the communication facility between the lexicon and the morphological analyzer. It has procedures that check whether the passed non-vowelized Arabic word is found in the lexicon and returns a list of `LexiconEntry` objects for the found non-vowelized words. Section 4.4.5 discussed the lexicon data structure and how the lexicon is searched to retrieve the lexicon objects.

8.3.2.2 Step 1, Root extraction

The system mainly depends on the SALMA – ABCLexicon to extract the root of the analyzed word. The SALMA – ABCLexicon contains 12 different biliteral roots, 8,585 different triliteral roots, 4,038 different quadriliteral roots, 63 different quinquiliteral roots, and 31 different sextiliteral roots. After selecting the candidate analyses that match the first part of the word with the proclitics and prefixes list, and the third part of the word with the suffixes and enclitics list, the analyzer searches the second part in the SALMA – ABCLexicon and retrieves all the `LexiconEntry` objects representing word-root pairs.

For each candidate analysis from the word segmentation step in the previous module the SALMA – Tokenizer, the second part of the segmented word, stem/root, is searched in the SALMA – ABCLexicon. If the non-vowelized stem/root is found in the lexicon then all vowelized word-root combinations are retrieved and attached to that analysis, which is accepted as a candidate analysis. The common (*i.e.* highly frequent) root for each analysis is specified. Also, the common root of the word's analyses is specified. Figure 8.13 shows examples of extracting the root of the different segmentation candidate analyses. The common root of the word and the common root of each analysis are shown in the figure.

	Word	**يَعْمَلُونَ**		**Common Root**		**عمل *E-m-l***			
	Word	**First part**		**Second part**		**Third Part**		**Root**	**Long stem**
يَعْمَلُونَ	*yaEomaluwna*			يعملون	*yEmlwn*			عمل *E-m-l*	يَعْمَلُونَ
يَعْمَلُونَ	*yaEomaluwna*			يعمل	*yEml*	ون	*wn*	عمل *E-m-l*	يَعْمَلُونَ
يَعْمَلُونَ	*yaEomaluwna*	ي	*y*	عملون	*Emlwn*			***Root is not found***	
يَعْمَلُونَ	***yaEomaluwna***	ي	***y***	**عمل**	***Eml***	ون	***wn***	**عمل *E-m-l***	يَعْمَلُونَ

Figure 8.13 Example of root extraction module

8.3.2.3 Step 2, Function Words

Function words are words with little semantic content. They serve as important clues to the structure of sentences. They define the grammatical relationships with other words within a sentence. They also signal the structural relationships that words have to one another[60]. Function words include pronouns, prepositions, determiners, conjunctions, auxilliary and modal verbs (Baker et al. 2006). A function word has a special morphological analysis wherever it appears in the text. The percentage of function words in any typical Arabic text is around 40%.

The system contains a list of 523 function words collected from a traditional Arabic grammar book (Diwan 2004). The morphological analyzer searches for the word in the function words list, and if it is founded, the analyzer adds the morphological analysis associated with it to the set of analyses generated by the morphological analyzer. Then the analyzer processes the next word. Figure 8.14 shows a sample of function words.

أنا	_>nA_	me	الذي	_Al*y_	who	حول	_Hwl_	about	عن	_En_	about
نحن	_nHn_	we	على	_ElY_	on	في	_fy_	in	بضع	_bDE_	few
هي	_hy_	she	عند	_End_	next to	بما	_bmA_	Although	بلى	_blY_	yes
هؤلاء	_h&lA'_	they	ذلك	_*lk_	that	بين	_byn_	between	مع	_mE_	with

Figure 8.14 Sample of the function words list

8.3.2.4 Step 3, Lemmatizing

In this step, the second part of each analysis, which represents the stem or root, is searched for in three other linguistic lists: a list of function words; a named entities list (Benajiba et al. 2008); and a list of broken plurals[61]. If the stem/root of any analysis matches one of these lists, then a new analysis entry along with its morphological analysis is added to the candidate analyses of the word.

The function word list, as discussed in the previous section, consists of 523 function words. The named entity list is the ANERGazet (Benajiba et al. 2008), which consists of three gazetteers: Locations gazetteer containing names of continents, countries, cities, etc; People gazetteer containing names of people collected manually from different Arabic websites; and Organizations gazetteer containing names of organizations like companies, football teams, etc. The Locations gazetteer contains 1,543 names; the People gazetteer contains 2,099 names; and the Organizations gazetteer contains 316 names. Figure 8.15 shows examples of the three gazetteers.

[60] Wikipedia: Function words http://en.wikipedia.org/wiki/Function_words

[61] Khaled Elghamry (2007) Broken Plural List http://sites.google.com/site/elghamryk/arabiclanguageresources

		Locations gazetteer			
اثيوبيا	*'iṯyūbiyā*	Ethiopia	ابو حماد	*'abū hammād*	Abu Hammad
القاهرة	*Al-qāhira*^{h}	Cairo	اكسفورد	*'uksfurd*	Oxford
جمهورية الكونغو الديمقراطية	*ğomhūryyat al-kongū ad-dīmoqrātiyyah*			Democratic Republic of the Congo	
		People gazetteer			
ابراهيم	*'ibrāhīm*	Abraham	زهرة	*zahra*^{h}	Zahra
عبدالله	*'abdullā*^{h}	Abdullah	غراهام	*ḡrāhām*	Graham
		Organizations gazetteer			
اخبار الخليج	*'aẖbār al-ẖalīğ*	Gulf News	ريال مدريد	*riyāl madrīd*	Real Madrid F.C
وكالة انباء البتراء	*wikala*^{t} *'anbā' al-batrā'*		Petra News Agency		

Figure 8.15 Examples of the three named entities gazetteers

The third list used is the broken plural list. The list is compiled using the broken plural lists of Elghamry (2007). These lists were automatically extracted from three Arabic Dictionaries: المتقن *al-mutqan* "The professional", الوسيط *al-wasīṭ* "The median", and الغني *al-ḡanī* "The rich". As a singular form is hard to guess from the broken plural form of the word, the lemmatizer is provided with a list of broken plural words of Arabic consisting of 11,367 broken plurals. Each broken plural entry in the list is provided with the root and the singular form of the broken plural which represents the lemma. Figure 8.16 shows examples from the broken plural list.

Broken plural			**Singular**		
أبواق	*'abwāq*	Horns	بوق	*būq*	Horn
حفظة	*ḥafaẓa*^{h}	Ones who know Qur'an by heart	حافظ	*ḥāfaẓ*	One who knows Qur'an by heart
حَيارَى	*ḥayārā*	Confused people	حيران	*ḥayrān*	To become confused
خياشيم	*ẖayāšīm*	Noses; gills	خيشوم	*ẖayšūm*	Nose
نسخ	*nusaẖ*	Copies	نسخة	*nusẖa*^{h}	Copy

Figure 8.16 Examples of broken plurals

The SALMA – Lemmatizer and Stemmer has been applied to lemmatize a large and varied Arabic Internet Corpus consisting of 176 million words of documents collected from the web (Sawalha and Atwell 2010b). Chapter 10 discusses the application of the SALMA – Lemmatizer and Stemmer used to lemmatize the Arabic Internet Corpus. See section 2.3.4.2 for the definition of lemma, lemmatizing and stem. For further distinctions between concatenative morphology and templatic morphology see Habash (2010).

8.3.3 Module 3: SALMA – Pattern Generator

The templatic morphology of Arabic words is based on three elements: root, pattern and vowelization (vocalisim). Roots are the three, four or five underlying letters of words. Roots are classified according to the number of their radicals into: triliteral, quadriliteral or quinquitiliteral (Habash 2010). The previous section 8.3.2 defines roots and explains the methodology followed to extract the roots of the analyzed words.

Patterns are the templates of combinations of consonants and vowels. The consonants represent slots for the root radicals to be inserted and the vowels represent the vocalism. The pattern is represented by sequences of Cs representing the consonants and Vs representing the vocalism. For instance, the pattern **mVC1C2VC3** where the vocalisim **V=a**. Using this pattern and the root كتب (*k-t-b*) "to write", the word *maktab* مَكتَب "office" is derived. The CV approach for representing patterns is widely used a cross languages (McCarthy and Prince 1990b; McCarthy and Prince 1990a; Smrz 2007; Attia 2008; Habash 2010).

Hundreds of years ago, patterns were defined by Arabic grammarians as الميزان الصرفي *al-mīzān aṣ-ṣarfī* "the morphological scale". The root letters of the patterns are represented by three letters ف *fā'* **f**, ع *'ain* **E** and ل *lām* **l** representing the first, second and third radicals of the word respectively. The purpose of using the patterns is to standardize the morphological description including the root letters and the vocalism of the derived words. The patterns group derivations of different roots into a template that describes the derivation process, the vocalism and the changes that might happen to the word during derivation (Ali 1987; al-Saydawi 2006).

The patterns are templates that enable root letters to be slotted in. Therefore, there are patterns that have three slots to suit triliteral roots (*e.g.* the word لَهَب *lahab* "flame" has the pattern فَعَل *fa'al* **faEal**, the word جِسْم *ğism* "body" has the pattern فِعْل *fi'l* **fiEl**, and the word كُسُوف *kusūf* "eclips" has the pattern فُعُول *fu'ūl* **fuEuwl**). If the root is quadrilateral - having four radicals - then the fourth radical is represented by (ل *lām* **l**), which is a repetition of the third radical. For example, the word صُعْلُوك *ṣu'lūk* "robber" has the quadriliteral root ص-ع-ل-ك (*ṣ-'-l-k*) and the pattern فُعْلُول *fu'lūl* **fuEluwl**). Second, if one of the triliteral root letters is doubled, then the symbol that represents that letter in the pattern is also doubled. For example the word رَسَّام *rassām* "painter" which is derived from the triliteral root ر-س-م *r-s-m* "to paint", has the pattern فَعَّال *fa''āl* **faEEaAl**). In general, if a letter is added or doubled in the word, then the same letter is added or the corresponding letter is doubled in the pattern (Ali 1987; al-Saydawi 2006).

The pattern not only has slots for root letters and vocalism to be inserted, it also captures morphosyntactic and semantic characteristics of the derived words. These characteristics are the basis for grouping Arabic words into families of formally and semantically related forms (Ali 1987). These morphosyntactic features are inherited by the derived word of that pattern. The next section 8.3.3.1 describes the construction of the pattern dictionary. The pattern dictionary depends on the SALMA morphosyntactic standards to describe the morphosyntactic attributes of the patterns which are propagated to the derived words. Therefore, knowing the analyzed word's pattern results in knowing most of the morphological feature values. Two pattern matching algorithms are used to extract the correct pattern of the analyzed word. These algorithms depend on the pattern dictionary to match the word with its possible patterns. Sections 8.3.3.2 and 8.3.3.3 discuss the pattern matching algorithms.

Pattern matching has been investigated by many researchers and several pattern matching algorithms have been proposed to match the word with possible patterns. The Xerox Arabic morphological analyzer depends only on finite-state operations (Beesley 1996; Beesley 1998). *Alkhalil* depends on large morphophonemic patterns (Mazroui et al. 2009; Boudlal et al. 2010). ElixirFM uses the morphophonemic patterns pertaining to the morphological stem and reflects its phonological qualities (Smrz 2007).

The choice of using morphosyntactic patterns or morphophonemic patterns depends on the ability of the pattern matching algorithm to deal with the three types of changes that might happen to the word during the derivation. Matching the morphophonemic pattern with the word can be easier than matching with morphosyntactic patterns. However, the number of patterns in the patterns dictionary will be very large, and it is hard to collect, encode and describe the features of each pattern. On the other hand, morphosyntactic patterns are easier to collect, encode and describe the features of each pattern entry. However, the pattern matching algorithm must deal with the three types of changes: incorporation or assimilation, substitution and deletion of vowel letters. Thus, a more sophisticated pattern matching algorithm needs developing.

Incorporation is a common phonological process by which the sound of one letter blends with the sound of the following letter. For example, the word آمَنَّا *'āmannā* "we believe" has two incorporations: *madda*^{h} which represents incorporation of the letter *hamza*^{h} and the following *'alif*, and the doubled ن *nūn*, which involves incorporation of the *nūn* (*i.e.* the last letter of آمَنْ *'āman*) and the following letter *nūn* (*i.e.* the first letter of the subject suffixed pronoun نَا *nā*). The word آمَنَّا *'āmannā* **|Aman~aA** will match the pattern فَاعَلْنَا *fā'alnā* **fAElnaA**. After resolving the two incorporations, the word will be ءامَنْنَا *'āmannā* **>AmanonaA**. Incorporation appears in the written script of the word and it is marked by *šadda*^{h}.

Substitution is the process of changing one of the root radicals into another letter during the derivation process. Substitution happens to weak root letters; و *wāw* and ي *yā'* are changed into *'alif* or *hamza*^{h}. The *'alif* in the word صَلَاةٌ *ṣalā*^{tun} "a prayer" is underlyingly و *wāw* in its root ص-ل-و *ṣ-l-w*. Substitution happens to other letters of the pattern such as ت *tā'* in the pattern إِفْتَعَلَ *'ifta'ala* **>ifotaEala**. Where the first radical is ز *zāy* or ص *ṣād* the ت *tā'* is changed into د *dāl* or ط *ṭah* respectively. This kind of substitution happens because it is hard to pronounce the /t/ sound after /z/ or /sˤ/. The word إِزْدِهَار

’izdihār **>izodihaAr** “prosperity” has the root (ز-هـ-ر) *z-h-r* and the pattern إِفْتِعَال *’ifti‘āl* **>ifotiEaAl.** Here the third letter of the word د *dāl* has changed from the letter ت *tā’* in the pattern. إِصْطَدَمَ *’iṣṭadama* **>iSoTdama** “clashed” has the root (ص-د-م) *ṣ-d-m* and the pattern إِفْتَعَلَ *’ifta‘ala* **>ifotaEala**. Here the third letter of the word ط *ṭah* has changed from the letter ت *tā’* in the pattern.

Deletion of vowel letters or *nūn* is a mood mark; section 6.2.12 discussed the case and mood marks including deletion. A vowel letter at the end of an indicative verb is deleted if the verb is in the imperative or jussive mood. For example, لا تَنْسَ! *lā* *tansa!* ‘Don’t forget!’, The verb تَنْسَ *tansa* ‘forget’ is in the jussive mood marked by deleting the vowel letter ى *’alif* from the end of the original verb تَنْسى *tansā*. The *nūn* at the end of indicative verbs which follow one of the five common verb patterns الأَفْعَال الخَمْسَة *al-’af‘āl al-ẖamsa*^{h}, is deleted in subjunctive or jussive mood. For example, قولوا خيراً تَغْنَمُوا *qūlū ẖayr*^{an} *taġnamū* ‘If you speak well, you will get benefits’, the verb تغنموا *taġnamū* “you will get benefits” is in the jussive mood. Therefore, the final letter *nūn* is deleted from the verb to indicate the jussive mood. The same verb in the indicative mood is تَغْنَمُونَ *taġnamūna*.

8.3.3.1 Constructing the Patterns Dictionary

The construction of the pattern dictionary started by collecting the morphosyntactic patterns from traditional Arabic grammar books (Ya‘qūb 1996) which provided the vowelized patterns and the morphosyntactic description in Arabic for each pattern. The morphosyntactic attributes of each pattern were determined and encoded using the SALMA – Tag Set standards. Also, the full vowelization (vocalism) of each pattern was added. The dictionary of morphosyntactic patterns contains 2,730 verb patterns and 985 noun patterns. Figure 8.17 shows sample entries of the patterns dictionary.

We chose to construct a pattern dictionary that contains morphosyntactic patterns, rather than morphophonemic patterns or CV patterns and vocalisms, because the morphosyntactic patterns are easier to collect, encode and describe the features of each pattern entry. The two words تَدَحْرَجَ *tadaḥraǧ* **tadaHraj** “rolled” and تَدَحْرُجَ *tadaḥruǧ* **tadaHruja** “rolling” have the same CV pattern CVCVCCVC. It ia thus impossible by this means to distinguish between the third person singular perfect verb تَدَحْرَجَ *tadaḥraǧ* **tadaHraj** “rolled” and the gerund تَدَحْرُجَ *tadaḥruǧ* **tadaHruja** “rolling”. However, the two words have the morphosyntactic patterns تَفَعْلَلَ *tafa‘lal* **tafaElal** and تَفَعْلُلَ *tafa‘lul* **tafaElul** respectively. The two patterns match the previous words and distinguish between the morphosyntactic features of each word. Unaugmented triliteral perfect verbs have the morphosyntactic pattern فَعَلَ *fa‘ala* **faEala** which also indicates a third person masculine singular subject as in: the verbs قَالَ *qāla* **qaAla** “he said”, and كَتَبَ *kataba* **kataba** “he wrote”. However, they have two morphophonemic patterns فَالَ *fāla* **faAla** and فَعَلَ *fa‘ala* **faEala** respectively.

A pattern matching algorithm matches the analyzed words with their morphosyntactic patterns in the pattern dictionary. The morphosyntactic attributes are represented as a SALMA – Tag and the vowelization of the matched patterns are propagated to the analyzed words. Two pattern matching algorithms were developed. Both of them mainly depend on the pattern dictionary. The next sub-sections discuss the pattern matching algorithms.

A syllabified version of the pattern was stored alongside the pattern to be used in a future Arabic prosody project, (see chapter 11 for future work). Dashes were used to separate the syllables of the patterns.

Verb Patterns		Syllabification	SALMA Tag
فَعَلْتُ	**faEalotu**	فَ–عَلْ–تُ	`v-p---nsfs-s-an??dst?-`
فَعَلْنَا	**faEalonaA**	فَ–عَلْ–نَا	`v-p---npfs-s-an??dst?-`
فَعَلْتَ	**faEalota**	فَ–عَلْ–تَ	`v-p---msss-s-an??dst?-`
فَعَلْتِ	**faEaloti**	فَ–عَلْ–تِ	`v-p---fsss-s-an??dst?-`
فَعَلْتُمَا	**faEalotumaA**	فَ–عَلْ–تُ–مَا	`v-p---xdss-s-an??dst?-`
Noun Patterns		**Syllabification**	**SALMA Tag**
أُفْعُلاوَى	**>ufoEulAwaY**	أُفْ–عُ–لا–وَى	`n?----??-v???---?dqt-?`
اِفْعِيلال	**AifoEiylAl**	اِفْ–عِي–لال	`ng----??-v???---?dtt-?`
فاعُولاء	**fAEuwlA’**	فا–عُو–لاء	`n?----??-v???---?dqt-?`
فُعُلْعُلان	**fuEuloEulAn**	فُ–عُلْ–عُ–لان	`n?----??-v???---?dqt-?`
فُعَّيْلاء	**fuE~ayolA’**	فُعْ–عَيْ–لاء	`n?----??-v???---?dqt-?`

Figure 8.17 Sample of the patterns dictionary

8.3.3.2 Pattern Matching Algorithm 1

The first pattern matching algorithm depends on the word itself and its root as inputs. The algorithm replaces the root letters in the word with the pattern letters ف *fa’* **f**, ع *‘ain* **E**, and ل *lām* **l**. Then it searches in the patterns dictionary for the generated pattern and returns the morphosyntactic attributes and the vowelization of the analyzed word.

However, the process of replacing the root letters with the letters ف *fa’* **f**, ع *‘ain* **E**, and ل *lām* **l** is not easy, as some root letters might be changed. The changes include incorporation, turnover, defection and replacement. The algorithm must deal with these changes and extract the correct pattern of the word. The algorithm follows these steps to match the pattern which deals with the changes that happen to the word during derivation:

1. Determine the root letters in the word:

   a) Find the index or indices of each root letter in the word. If the root letter is *’alif*, *wāw*, *yā’* or *hamza*^{h} then add -1 to the indices list of that root letter. The -1 value indicates that the root radical has changed. See figure 8.18 step 1a.

b) Construct the candidate root indices lists by generating all possible permutations of the indices of the root radicals (step 1a), by selecting an index from each indices list of the root radicals into one combined list. See figure 8.18 step 1b.

c) Select the candidate root indices lists that satisfy the linguistic rule of derivation where root letters must appear in the same order in the derived words. This means that the index of the first root radical must be less than the index of the second root radical, and they must be less than the index of the third root radical. The -1 value in the list does not violate the rule. See figure 8.18 step 1c.

2. Replace the root letters in the words with the pattern letters ف *fa'* **f**, ع *'ain* **E**, and ل *lām* **l**. The indices of the the root letters in the words are determined from the previous step (1c). See figure 8.18 step 2.

3. Search for the candidate pattern in the patterns dictionary. If the pattern is found in the list, the SALMA – Tag associated with the pattern in the list is assigned to the analyzed word.

4. If the word is fully vowelized or partially vowelized, then match the vowelization of the word with the vowelization of the pattern. Select only the vowelization of the patterns which best match the vowelization of the word.

The algorithm is repeated for each analysis of the candidate analyses produced by the previous analyzer module. The patterns and the morphosyntactic attributes are added to each analysis.

8.3.3.3 Pattern Matching Algorithm 2

The second method of extracting the pattern of the word is based on the Pattern Matching Algorithm (PMA) (Alqrainy, 2008). This algorithm matches partially vowelized word, with the last diacritic mark only, with a pattern lexicon without doing any analyses for the clitics and affixes of the word.

Pattern matching algorithm 2 searches the patterns list for patterns of similar size as the analyzed word after removing the clitics of the word. For example, a form كتب *ktb* has a size of 3 according to the data structure we used, whether the word is fully-vowelized, partially-vowelized or non-vowelized. It matches the following patterns (فَعْل *FaEol*, فَعَل *faEal*, فَعُل *faEul*, فَعِل *faEil*, فُعْل *fuEol*, فُعَل *fuEal*, فُعُل *fuEul*, فُعِل *fuEil*, فِعْل *fiEol*). In the second step, the algorithm replaces the letters of the word corresponding to the letters ف *fa'* **f**, ع *'ain* **E**, and ل *lām* **l** of the pattern. Then these generated patterns are searched in the pattern list. If the pattern is found in the pattern list, then it is a candidate pattern of the word, and the morphological tag associated with the pattern in the list is assigned to the analyzed word. Figure 8.19 shows example of extracting the pattern of the word using this method. Figure 8.20 shows examples of matches pattern and their SALMA Tags. The pattern matching algorithm 2 steps are the following:

1. Get the patterns, from the patterns list, which have a similar size to the analyzed word after removing the clitics of the word.
2. Choose the patterns that share the maximum number of letters with the analyzed words. This will reduce the number of patterns to be processed.
3. Replace the letters of the word corresponding to the letters ف *fa'* **f**, ع *'ain* **E**, and ل *lām* **l** of the pattern.
4. Search the candidate generated patterns in the pattern list. If the pattern is found in the pattern list, then the SALMA – Tag associated with the pattern in the list is assigned to the analyzed word.
5. If the word is fully vowelized or partially vowelized, then match the vowelization of the word with the vowelization of the pattern. Select only the vowelization of the patterns that best match the vowelization of the word.

Both pattern matching algorithms are used by the SALMA – Pattern generator to match the analyzed with its pattern from the patterns dictionary. The pattern matching algorithm 1 requires the root information to be available, while the pattern matching algorithm 2 depends only on the patterns dictionary. The pattern matching algorithm 1 was developed mainly to solve the problems of the incorporation, deletion, and substitution of the root radicals during the derivation process. The pattern matching algorithm is an improved version of the PMA of Alqrainy (2008). The original PMA matches the word with the patterns of provided with a dictionary containing 8,718 patterns most of them verb patterns. The PMA does not deal with clitics and affixes. This requires providing the algorithm with a large pattern dictionary of all possible combinations of clitics and affixes attached to the pattern types. The SALMA – Pattern generator uses only the matching steps of the PMA to match the word with patterns stored in our patterns dictionary after removing the clitics and affixes that are marked as they are not part of the pattern; see section 8.3.1.5 for the details of the clitics and affixes dictionaries. The removal of the unwanted clitics and affixes generalize the pattern matching algorithm to a finite set of patterns represented by the patterns dictionary that we have constructed.

Step 1 **Determine the root letters in the word**

Word أَحْسَنَ *'aḥsana* **>aHosana** "better"

Root ح-س-ن *ḥ-s-n* **H-s-n**

Step 1a Find the index or indices of each root letter in the word

Word	[(أ **>**)$_0$, (ح **H**)$_1$, (س **s**)$_2$, (ن **n**)$_3$] (short vowels are not shown)
Indices of 1st Root radical **(ح H)**	[1]
Indices of 2nd Root radical **(س s)**	[2]
Indices of 3rd Root radical **(ن n)**	[3]

Step 1b **Construct the candidate root indices**

Candidate indices list [1, 2, 3]

Step 1c Select the candidate root indices lists that satisfy the linguistic rule

Indices list [1, 2, 3]

Step 2 Replace the root letters in the words by the with the pattern letters

Word [(أ **>**)$_0$, (ح **H**)$_1$, (س **s**)$_2$, (ن **n**)$_3$]

Pattern [(أ **>**)$_0$, (ف **f**)$_1$, (ع **E**)$_2$, (ل **L**)$_3$] أفعل **>fEl** *'f'l*

Step 3 Search for the candidate pattern in the patterns dictionary

Matched patterns

أَفْعَل	*>afoEal*	`n@----m?-v???---?dat-?`	أُفْعِلَ	*>ufoEila*	`v-c---xsfdaf-an??dat?-`
أَفْعَل	*>afoEal*	`nj----m?-v???---?dat-?`	أُفْعِلْ	*>ufoEilo*	`v-c---xsfdjs-an??dat?-`
أَفْعُلُ	*>afoEulu*	`v-c---xsfdnd-an??dst?-`	أَفْعِلْ	*>ufoEilo*	`v-i---msss-s-an??dat?-`
أَفْعُلْ	*>afoEulo*	`v-c---xsfdjs-an??dst?-`	أُفْعَلُ	*>ufoEalu*	`v-c---xsfdnd-pn??dtt?-`
أَفْعِلُ	*>afoEilu*	`v-c---xsfdnd-an??dst?-`	أُفْعُلَ	*>ufoEula*	`v-c---xsfdaf-pn??dtt?-`
أَفْعِلَ	*>afoEila*	`v-c---xsfdaf-an??dst?-`	أُفْعُلْ	*>ufoEula*	`v-c---xsfdjs-pn??dtt?-`
أَفْعِلْ	*>afoEilo*	`v-c---xsfdjs-an??dst?-`	أُفْعَلُ		`v-c---xsfdnd-pn??dat?-`
أَفْعَلُ	*>afoEalu*	`v-c---xsfdnd-an??dst?-`	أُفْعَلَ		`v-c---xsfdaf-pn??dat?-`
أَفْعَلَ	*>afoEala*	`v-c---xsfdaf-an??dst?-`	أُفْعَلْ		`v-c---xsfdjs-pn??dat?-`
أَفْعَلْ	*>afoEalo*	`v-c---xsfdjs-an??dst?-`			

Step 4 **Match the vowelization of the word with the vowelization of the pattern**

أَفْعَل	`n@----m?-v???---?dat-?`	أَفْعَلَ	`v-c---xsfdaf-an??dst?-`
أَفْعَل	`nj----m?-v???---?dat-?`		

Figure 8.18 Example of extracting the pattern of the words using the first method (the word and its root)

Step 1	**Get the patterns, from the patterns list, which have similar size as the analyzed word**
Word	يَعْمَلُونَ *ya'malūna* **yaEomaluwna** "They work" word length = 6
Patterns	يَفْعَلُونَ *yaf 'alūna* **yafoEaluwna**, يَفْعَلاَنِ *yaf'alāni* **yafoEalaAni**, تَفْعَلِينَ *taf'alīn* **tafoEaliyna**, تَفْعَلاَنِ *tafo'alāni* **tafoEalaAni**, يَفْعُلاَنِ *yaf'ulān* **yafoEulaAn**,…etc.
Step 2	Choose the patterns that share the maximum number of letters with the analyzed words
Patterns	يَفْعَلُونَ = 4, يَفْعَلاَنِ = 3, يَفْعُلاَنِ = 3, تَفْعَلِينَ = 2, تَفْعَلاَنِ = 2
Step3	Replace the letters of the word corresponding to the letters (ف *fa'* **f**, ع *'ain* **E**, and ل *lām* **l**) of the pattern.

Word	يَعْمَلُونَ	ي **y**$_0$	ع **E**$_1$	م **m**$_2$	ل **l**$_3$	و **w**$_4$	ن **n**$_5$	**yaEmlwn**
Pattern	يَفْعَلُونَ	ي **y**$_0$	ف **f**$_1$	ع **E**$_2$	ل **l**$_3$	و **w**$_4$	ن **n**$_5$	**yfElwn**
Generated pattern	**يفعلون**	ي **y**$_0$	ف **f**$_1$	ع **E**$_2$	ل **l**$_3$	و **w**$_4$	ن **n**$_5$	**yfElwn**

Step 4	Search the candidate generated patterns in the pattern list

يَفْعُلُونَ	*yafoEuluwna*	`v-c---mptdnn-an??dst?-`
يَفْعِلُونَ	*yafoEiluwna*	`v-c---mptdnn-an??dst?-`
يَفْعَلُونَ	*yafoEaluwna*	`v-c---mptdnn-an??dst?-`
يُفْعِلُونَ	*yufoEiluwna*	`v-c---mptdnn-an??dat?-`
يُفْعَلُونَ	*yufoEaluwna*	`v-c---mptdnn-pn??dtt?-`

Step 5	**Match the vowelization of the word with the vowelization of the pattern**		
Pattern	يَفْعَلُونَ	*yafoEaluwna*	`v-c---mpt--ian?-st?`

Figure 8.19 Example on Pattern Matching Algorithm 2 processing steps

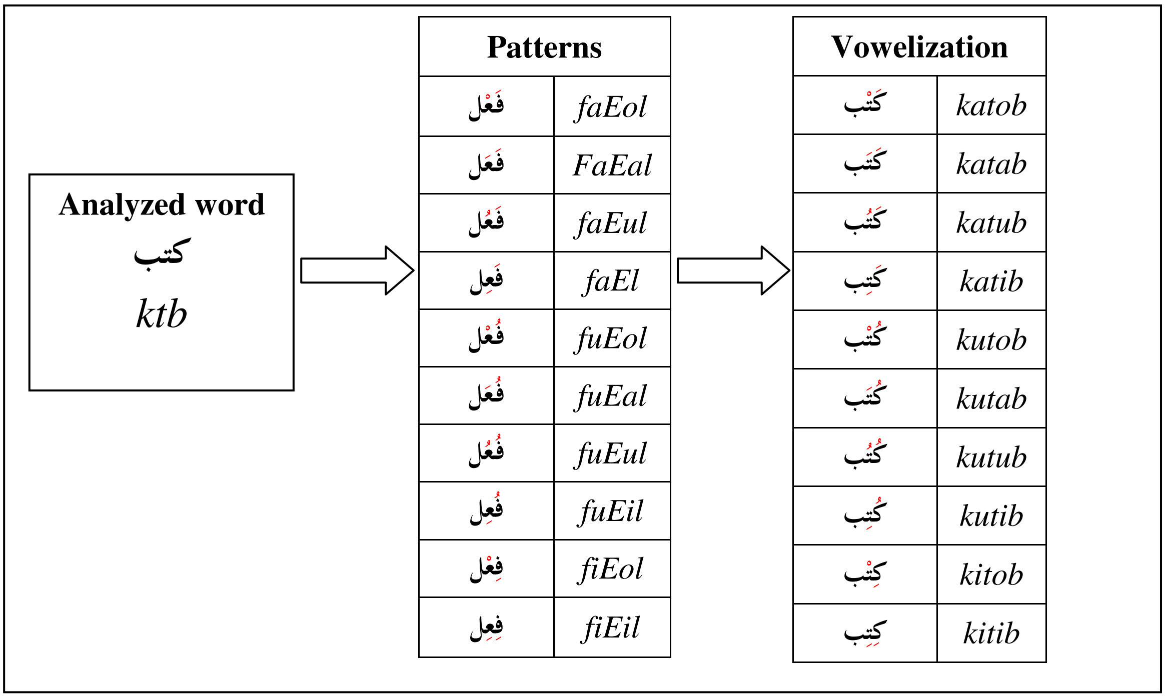
Word		Pattern		SALMA Tag
كتب	*ktb*	فَعَلَ	*faEala*	`v-p---msts-a-an??dst?-`
كتب	*ktb*	فَعِلَ	*faEila*	`v-p---msts-f-an??dst?-`
كتب	*ktb*	فَعُلَ	*faEula*	`v-p---msts-f-an??dst?-`
كتب	*ktb*	فُعِلَ	*fuEila*	`v-p---msts-f-pn??dtt?-`
كتب	*ktb*	فَعْل	*faEol*	`nj----m?-v???---?dst-?`
كتب	*ktb*	فَعَل	*FaEal*	`ng----m?-v???---?dst-?`
كتب	*ktb*	فَعُل	*faEul*	`n?----??-v???---?dst-?`
كتب	*ktb*	فَعِل	*faEil*	`nx----??-v???---?dst-?`
كتب	*ktb*	فُعْل	*fuEol*	`ng----??-v???---?dst-?`
كتب	*ktb*	فُعَل	*fuEal*	`n?----??-v???---?dst-?`
كتب	*ktb*	فُعُل	*fuEul*	`n?----??-v???---?dst-?`
كتب	*ktb*	فُعِل	*fuEil*	`n?----??-v???---?dst-?`

Figure 8.20 Example of using the Pattern Matching Algorithm 2

8.3.4 Module 4: SALMA – Vowelizer

Vowelization is an important characteristic of the Arabic word. Vowelization helps in determining some morphological features of the words. The presence of the short vowel on the last letter helps in determining the case or mood of the word. The presence of the vowels on the first letter determines whether the verb is active or passive. The presence of other diacritics such as *šadda*^{h} and *madda*^{h} (extension) solve some ambiguities of words.

After matching the patterns and the analyzed word, in the previous step, taking into account that the patterns are fully vowelized, the analyzer adds the short vowels which appear on the patterns to the analyzed word, whether it is partially-vowelized or non-vowelized. The result is a correctly fully vowelized list of words with the possible analyses. Figure 8.21 shows the process of adding vowels to the non-vowelized words.



Patterns	
فَعْل	*faEol*
فَعَل	*FaEal*
فَعُل	*faEul*
فَعِل	*faEl*
فُعْل	*fuEol*
فُعَل	*fuEal*
فُعُل	*fuEul*
فُعِل	*fuEil*
فِعْل	*fiEol*
فِعِل	*fiEil*

Vowelization	
كَتْب	*katob*
كَتَب	*katab*
كَتُب	*katub*
كَتِب	*katib*
كُتْب	*kutob*
كُتَب	*kutab*
كُتُب	*kutub*
كُتِب	*kutib*
كِتْب	*kitob*
كِتِب	*kitib*

Figure 8.21 Vowelization process example

8.3.5 Module 5: SALMA – Tagger

The SALMA – Tagger is built on top of the previous modules: the SALMA-Tokenizer, the SALMA – Lemmatizer and Stemmer, the SALMA – Pattern Generator and the SALMA – Vowelizer. Each module processes input words and produces direct results such as: root, lemma and pattern, and intermediate results which are passed to the next module. The previous intermediate results are necessary to perform the specified tasks of that module. For instance, the SALMA – Pattern Generator accepts the root from the SALMA – Stemmer and the input word's tokenization resulting from the SALMA – Tokenizer, as inputs and uses the patterns dictionary to provide the necessary morphosyntactic information to find the pattern of the word. Figure 8.4 shows the complete SALMA – Tagger algorithm and the relations of its component modules.

The SALMA – Tagger module is the last module which is responsible for adding the SALMA Tags to the analyzed word morphemes. Each morpheme is assigned a single SALMA Tag. The initially-assigned SALMA – Tags were given to the word's morphemes by matching the morpheme with its equivalent from the morphosyntactic dictionaries included in the system. The initial morphological features tag assignment is discussed in the next sub-section 8.3.5.1. A rule-based system was developed and integrated to the SALMA – Tagger to predict the value of the morphological features which are not assigned in the initial tag assignment process. Sub-section 8.3.5.2 discusses the different kinds of rules that were used to predict the morphological features of the analyzed word. It gives examples of the rules used to predict the morphological features. Section 8.4 gives two examples of the complete set of linguistic rules used to predict the morphological features of person and rationality. Section 8.3.5.3 shows the colour-coded tags for the word's morphemes.

8.3.5.1 Initially-assigned SALMA Tags

Most Arabic words are complex words consisting of multiple morphemes. Each morpheme carries morphological features and belongs to a specific part of speech category. The SALMA-Tagger assigns a tag for each morpheme of the word; given that the linguistic lists used by the morphological analyzer all have the morphological feature tags assigned to each entry in these lists. The previous SALMA – Tokenizer and SALMA – Pattern Generator modules assign an initial SALMA – Tag for each morpheme of the analyzed words.

As discussed before, words should be decomposed into five parts: proclitics, prefixes, stem or root, suffixes and postclitics. The morphological analyser should then add the appropriate linguistic information to each of these parts of the word; in effect, instead of a tag for a word, we need a subtag for each part (and possibly multiple subtags if there are multiple proclitics, prefixes, suffixes and enclitics) (Sawalha and Atwell 2009a).

The SALMA – Tokenizer implements the above definition and segments the analyzed word into five parts. It assigns a SALMA – Tag for each clitic or affix by searching in the clitics and affixes dictionaries. Once the clitic or affix is found in the clitics and affixes dictionaries, the SALMA Tag associated with that dictionary entry is assigned to the clitic or affix of the word. See section 8.3.1.6 for more details about matching the word segments with the clitics and affixes dictionary entries. The SALMA Tags assigned to the clitics and affixes of the analyzed words represent the initial tag assignment.

The SALMA – Pattern Generator extracts the pattern of the word by applying two pattern matching algorithms that depend on a pattern dictionary. The pattern dictionary associates a SALMA – Tag with each pattern entry. This tag will be assigned to the analyzed word as an initial tag, which will represent the tag of the stem of the word. The initially-assigned SALMA – Tags specify whether a morphological feature category is applicable to the morpheme or not applicable represented by "-" in the tag string. If the feature is applicable, then the value of that feature is either determined and represented by a single letter, or cannot be initially-predicted and represented by "?". Figure 8.22 shows an example of assigning the initial tags to a word. The example shows that morphological features of Transitivity, Rational and Verb Root cannot be predicted at this stage of analysis.

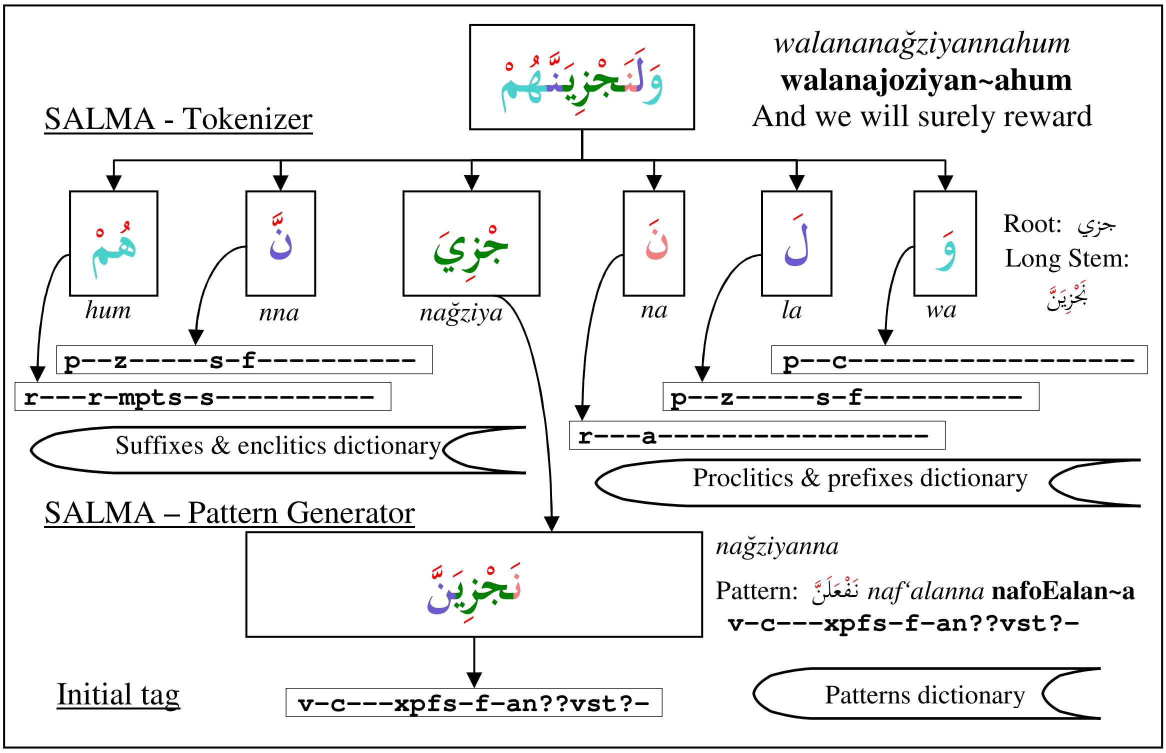


Figure 8.22 Example of assigning initial SALMA Tags to all word's morphemes

8.3.5.2 Rule-Based System to Predict the Morphological Feature Values of the Word's Morphemes

A rule-based system was developed to predict the values of the morphological features of the analyzed word. A set of rules was extracted from traditional Arabic grammar books that predict the value of each morphological feature category. The SALMA – Tagger validates the initially-predicted values of the morphological features and predicts the value of the morphological features which were not assigned in the previous step. Figure 8.23 shows examples of the linguistic rules applied to validate and predict the values of the morphological features which were assigned for these particular words in context. The example shows how other morphological feature values help in distinguishing a given morphological feature. Different rules will apply to different words in context.

Section 8.4 gives examples of two sets of rules used to predict the morphological features of Person, Rational and Noun Finals.

Analyzed word	نَجْزِيَنَّ *nağziyanna* **najoziyan~a** “surely reward”	
Initial SALMA Tag	**`v-c---xpfs-f-an??vst?-`**	
Categorey	***Tag***	***Linguistic Rule Applied***
Inflectional Morphology Case or Mood Case and Mood Marks	**s** **-** **f**	If the imperfect verb (1, “v”), (3, “c”) is emphasized (15, “n”), has the suffix نْ *n* or نَّ *nna* the emphasis *nūn* as one of the word’s morphemes
Transitivity	**o**	If the verb (1, “v”) has an object suffixed-pronoun in its suffixes then it is transitive to one object.
Rational	**h**	Rational is set as default value for verbs (1, “v”).
Verb Roots	**x**	The root is جزي *ğ-z-y* has the template C1-C2-Y

The analyzed word نَجْزِيَنَّ is assigned the following SALMA Tag:

`v-c---xpfs-f-anohvstx-`

Analyzed word	نَصْرٌ *naṣr*^{un} “victory”	
Initial SALMA Tag	**`ng----??-v???---?dst-?`**	
Categorey	*Tag*	*Linguistic Rule Applied*
Gender	**m**	Masculine is a default value, if the word does not include femeinine suffixes ة *tā’ marbūṭa*^{h}, ى *’alif maqṣūrā* or اء *madd* extension.
Number	**s**	If the word is declined noun (1, “n”), (10, “v or p”) and the word does not have any of dual or plural suffixes and it is not found in the broken plural list.
Inflectional Morphology	**v**	If the word ends with *tanwīn*, then the word is a Triptote.
Case and Mood Case and Mood Marks Definitness	**n** **d** **i**	If the word ends with *tanwīn al-ḍamm*
Rational	**n**	Irrational is the default value for Gerund (1, “n”), (2, “g”)
Noun Finals	s	If the last letter of the word is a consonant and it is not a *hamza*^{h}, then the word is sound noun.

The analyzed word نَصْرٌ is assigned the following SALMA Tag:

`ng----ms-vndi---ndst-s`

Figure 8.23 Examples of the linguistic rules applied to validate and predict the values of the morphological features

8.3.5.3 Colour Coding the Analyzed Words

To visualize the analysis, the word morphemes can be colour-coded. The colour-coding scheme depends on the morphological information of the analyzed word. The SALMA – Tokenizer and the SALMA – Tagger modules specify each of the word's morphemes, its class (*i.e.* proclitic, prefix, stem, suffix and enclictic) and the part-of-speech category for each morpheme. The part of speech category of the stem was used to colour the stem. If the part-of-speech of the stem is a verb, noun, particle, other (residual) or punctuation mark, then it is coloured in green, purple, blue, dark grey or black respectively. Morpheme class is used to colour-code the word's morphemes of type proclitic, prefix, suffix and enclitic. Each part was coded in a different colour (and possibly multiple colours if there are multiple proclitics, prefixes, suffixes and enclitics). Four colours are used to colour prefixes and suffixes: SlateBlue, LightCoral, Violet and Gold. And four colours are used to colour proclitics and enclitics: MediumTurquoise, SteelBlue, PowderBlue and MediumAquaMarine. Figure 8.24 shows the different colours used to colour-code the word's morphemes. Figure 8.25 shows an example of a colour-coded word from the Qur'an Gold Standard. Figure 8.29 shows colour-coded visualization of a full text - Qur'an Chapter 29 and a MSA sample from CCA, showing just the morphemes, without full SALMA – Tags; this illustrates morpheme boundaries.

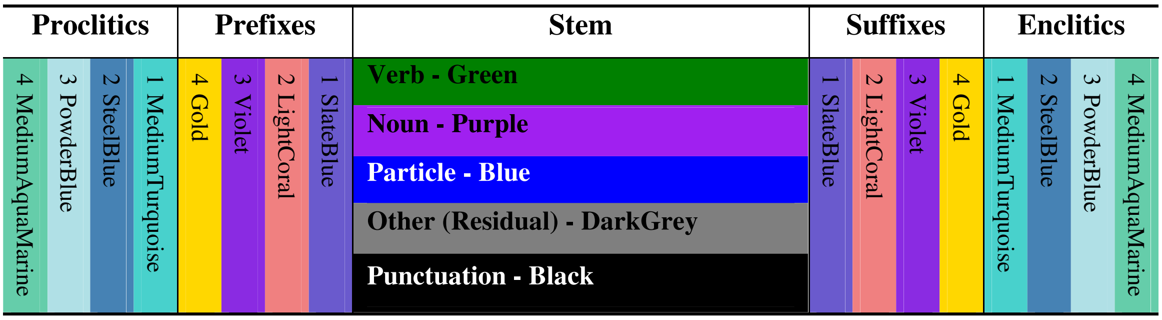


Figure 8.24 Colour codes used to colour code the morphemes of the analyzed words

فَلَيَعْلَمَنَّ

Root	*Stem*	*Long stem*	*Pattern*	*Word-by-word translation*
علم	عَلِم	يَعْلَمَنَّ	يَفْعَلَنَّ	and-allah-will-surely-make-

فَ	`p--c------------------`	حرف | حرف عطف| Particle |Conjunction |
لَ	`p--z-----s-f----------`	حرف | حرف توكيد | مبني | الفتحة /الفتح| Particle |Emphatic particle | Invariable (v, n) |*fatḥa*^{h} |
يَ	`r---a-----------------`	أخرى | حرف مضارعة | Other (Residual) |Imperfect prefix |
عْلَمَ	`v-c---msts-f-anohvtta-`	فعل | فعل مضارع | مذكر | مفرد |الغَائِب | مبني | الفتحة / الفتح | مَبْنِي للمَعْلُوم | فعل مُؤَكَّد | مُتَعدٍّ إلى مَفعُول واحِد | عاقِل | مُتَصَرِّف – فعل تام التَّصريف | مَزِيْد بِثَلاثَةِ أحرف |ثُلاثِي | صحيح| Verb |Imperfect verb |Masculine |Singular |Third Person | Invariable (v, n) |*fatḥa*^{h} |Active voice |Emphatic verb |Singly transitive |Rational |Conjugated / fully conjugated verb |Augmented by three letters |Triliteral |Intact verb |
نَّ	`r---z----s-f----------`	أخرى | نون التوكيد | مبني |الفتحة / الفتح| Other (Residual) |Emphatic nūn | Invariable (v, n) |*fatḥa*^{h} |

Figure 8.25 Colour-coded example of a word from the Qur'an gold standard

8.4 Rules for Predicting the Morphological features of Arabic Word Morphemes

A rule-based system was designed to predict the morphological features of the analyzed word's morphemes. It depends on linguistic knowledge extracted from traditional Arabic grammar books (Dahdah 1987; Wright 1996; Al-Ghalayyni 2005; Ryding 2005). For each morphological feature category of the SALMA – Tag Set, a set of rules were extracted and encoded in the SALMA – Tagger. The SALMA – Tagger executes these rules to predict and validate the values of the morphological features of the initial tags assigned to the word's morphemes. Sophisticated linguistic knowledge was encoded as a rule-based system within the SALMA – Tagger. The encoded rules represent a variety of linguistic knowledge types. In the following, SALMA – Tagger features are cross-referenced to subsections defining them.

First come, rules that depend on data lists or dictionaries. These rules search the analyzed word in the data dictionaries to predict the value of a given feature. The rule-based system includes several data lists: the broken plural list contains 9,513 entries used in predicting the morphological feature of Number (section 6.2.8); the named entities list includes personal names list which contains 2,099 entries, the location names list which contains 1,715 entries, and the organization names list which contains 384 entries. This is used to predict the morphological feature attribute of proper name and the morphological feature of Rational (section 6.2.17). The transitive verbs lists (*i.e.* the doubly transitive verb list contains 2,889 verbs and the triply transitive verbs list contains 1,065 verbs) are used to predict the values of the morphological feature of Transitivity (section 6.2.16). The five nouns list contains 21 entries including all the variations of the five nouns that can be found in a text. The list is used to predict the morphological feature attribute of the five nouns and some attributes of the morphological features of Case or Mood (section 6.2.11) and Case and Mood Marks (section 6.2.12). The non-conjugated and partially-conjugated verbs lists are used to predict some values of the morphological features category of Declension and Conjugated (section 6.2.18). These lists include: a partially-conjugated verb list which contains 13 entries; a non-conjugated/restricted to the perfect verb list containing 42 verbs, a non-conjugated/restricted to the imperfect verb list containing 4 verbs, and a non-conjugated/restricted to the imperative verb list containing 13 verbs.

Second come, rules that depend on the affixes and clitics of the words. Rules for predicting the morphological features of Gender (section 6.2.7), Number (section 6.2.8) and Person (section 6.2.9) of verbs check the combinations of prefixes and suffixes in the analyzed word. The number of nouns is predicted depending on both the suffixes of the analyzed word and on searching the analyzed word in the broken plural list. The morphological feature of emphasized and non-emphasized (section 6.2.15) depends on the presence and absence of the emphatic *nūn* suffix in the analyzed word. An emphasized verb which has emphatic *nūn* as a suffix, is an invariable verb, the morphological feature of Case or Mood (section 6.2.11) is not applicable and the Case and Mood Mark (section 6.2.12) is always *fatḥa*^{h}. A definite noun has a definite article as a proclitic.

Third come, rules which depend on the pattern of the analyzed word. Some rules of predicting intransitive verbs (section 6.2.16) depend on patterns such as افْتَعَلَ *'ifta'ala* **AfotaEala**, تَفاعَلَ *tafā'ala* **tafaAEala** and تَفَعَّلَ *tafa''ala* **tafaEEala**. Determining whether the verb has one of the five-verb patterns الأَفْعَالُ الخَمْسَة *al-'af'āl al-ẖamsa*^{h} is essential to predict the values of the morphological features of Gender (section 6.2.7), Number (section 6.2.8), Person (section 6.2.9), Inflectional Morphology (section 6.2.10), Case or Mood (section 6.2.11) and Case and Mood Mark (section 6.2.12). The SALMA – Pattern Generator is used to extract the pattern of the analyzed word.

Fourth come, rules depend on the root and stem of the analyzed word. The SALMA – Stemmer and Lemmatizer is used to extract the root of the analyzed word. The root is essential to predict the values of the morphological features of Number of Root Letters (section 6.2.20) and Verb Roots (section 6.2.21). The SALMA – Tokenizer defines the analyzed word's morphemes including the stem and the long stem of the word. The stem is the middle part of the analyzed words after removing both the clitics and affixes morphemes, while the long stem is the middle part of the analyzed word after removing the clitics only. Long stem is used to predict the value of the morphological feature of Noun Finals (section 6.2.22). It is also used with the root to predict the morphological feature of Unaugmented and Augmented (section 6.2.19).

Finally come, rules which depend on the vowelization of the word. The main Case and Mood Marks (section 6.2.12) attributes are specified by the final short vowel appearing on the final letter of the word. A noun that has *tanwīn* on its final letter is an indefinite noun. A passive voice verb has *ḍamma*^{h} on its first letter.

A default value was selected for each morphological feature category. The default value is used when the rules of predicting the attribute value of a certain morphological feature are not applicable. The selection of the default value was determined by the linguistic knowledge of the attribute values of the morphological features, rather than statistical analysis of the most frequent attribute values in a tagged corpus. A corpus analysis approach is not applicable because of the absence of a tagged Arabic corpus using the full SALMA – Tag Set. Examples of default values are: the default value of the verb mood (section 6.2.11) is set to be indicative; the default value for the Rational (section 6.2.17) is rational for verbs and irrational for nous; and the default value of the

Number of Root Letters (section 6.2.20) is triliteral as most roots of Arabic words are triliteral.

In this section, three examples are represented to show the complexity of designing and implementing the rule-based system to predict the values of the morphological features of the word's morphemes. Section 8.4.1 shows the rules for predicting the values of the morphological feature of Person (section 6.2.9). It also shows other morphological features where their value can be predicted using these rules: the Gender (section 6.2.7) and Number (section 6.2.8) of verbs. Section 8.4.2 shows an example of hard-to-predict morphological features, Rational (section 6.2.17). This example focuses on the need to construct comprehensive dictionaries and linguistic lists. It also gives a good example of selecting the default value for Rational. Section 8.4.3 discusses the rules of the morphological feature of Noun Finals (section 6.2.22). These rules depend on the long stem of the analyzed word.

8.4.1 Rules for Predicting the Morphological Feature of Person

An Arabic verb has three main person attribute values; first person المُتَكَلِّم *al-mutakallim*, second person المُخاطَب a*l-muḫāṭab* and third person الغَائِب *al-ḡā'ib*. First person refers to the person or people speaking. Second person refers the person or people who are present and sharing the talk or speech. Third person refers to the person or people who are absent and do not participate in the talk or speech (Ryding 2005).

The rules for predicting the morphological feature of person mainly depend on the combinations of prefixes and suffixed pronouns attached to the end of the verbs. Subject suffixed-pronouns and genitive suffixed pronouns describe the reference person of the verb and agree with the number and gender of the doer of the verb.

The subject suffix-pronouns are part of the circumfix (long stem), as the subject suffix-pronouns are part of the verb pattern, while the genitive suffix-pronouns are treated as enclitics. The values of the morphological features of Gender, Number and Person of the subject suffix-pronouns agree with their equivalent of the doer of the verb (the subject), while genitive suffixed-pronouns agree with the object of the sentence (*i.e.* the person or thing who received the action done by the subject of the verb) in the values of the morphological features of Gender, Number and Person. Subject suffix-pronouns and genitive suffix-pronouns can appear together in the same verb, and the agreement is maintained with the subject and the object of the sentence. For instance, the word يَقْرَؤُوْنَهَا *yaqra'ūnahā* 'they read it' has the prefix (يَ) *yā'* and the subject suffixed-pronoun (ون) *ūn*. The combination of prefix and suffix pronouns indicates third person, masculine gender and plural number of the verb, while the genitive suffix-pronoun ها *hā* indicates third person, feminine and singular object (*it*).

Tables 8.3-8.5 list the rules for predicting the values of the morphological feature of Person, and the values of the other related morphological features: Gender and Number of perfect, imperfect and imperative verbs respectively.

Table 8.3 Rules for predicting the values of the morphological features of Person, Number and Gender for perfect verbs

Position 9	**Person** الاسناد *al-’isnād*					
	Person Category	**Subject suffixed-pronoun**	**Genitive suffixed-pronoun**	**Person (9)**	**Number (8)**	**Gender (7)**
Perfect verb **(1, “v”)** **(3, “p”)**	First Person المتَكَلِّم *al-mutakallim*	تُ *tu*	نِي *nī*	**f**	**s**	**x**
		نَا *nā*	نَا *nā*	**f**	**p**	**x**
	Second Person المخاطَب *al-muẖāṭab*	تَ *ta*	كَ *ka*	**s**	**s**	**m**
		تُمَا *tumā*	كُما *kumā*	**s**	**d**	**x**
		تُم *tum*	كُم *kum*	**s**	**p**	**m**
		تِ *ti*	كِ *ki*	**s**	**s**	**f**
		تُنَّ *tunna*	كُنَّ *kunna*	**s**	**p**	**f**
	Third Person الغَائِب *al-ḡā’ib*	-	هُ *hu*	**t**	**s**	**m**
		ا *ā*	هُما *humā*	**t**	**d**	**x**
		وا *ū*	هُم *hum*	**t**	**p**	**m**
		-	ها *hā*	**t**	**s**	**f**
		نَ *na*	هُنَّ *hunna*	**t**	**p**	**f**

Table 8.4 Rules for predicting the values of the morphological features of Person, Number and Gender for imperfect verbs

	Person Category	**Prefix Aoristic letter**	**Subject suffixed-pronoun**	**Person (9)**	**Number (8)**	**Gender (7)**
Imperfect verb **(1, “v”)** **(3, “c”)**	First Person المتَكَلِّم *al-mutakallim*	أَ *’a*	-	**f**	**s**	**x**
		نَ *na*	-	**f**	**p**	**x**
	Second Person المخاطَب *al-muẖāṭab*	تَ *ta*	-	**s**	**s**	**m**
		تَ *ta*	انِ *āni*	**s**	**d**	**x**
		تَ *ta*	ونَ *ūna*	**s**	**p**	**m**
		تَ *ta*	يْنَ *īna*	**s**	**s**	**f**
		تَ *ta*	نَ *na*	**s**	**p**	**f**
	Third Person الغَائِب *al-ḡā’ib*	يَ *ya*	-	**t**	**s**	**m**
		يَ *ya*	انِ *āni*	**t**	**d**	**m**
		يَ *ya*	ونَ *ūna*	**t**	**p**	**m**
		تَ *ta*	يْنَ *īna*	**t**	**s**	**f**
		تَ *ta*	انِ *āni*	**t**	**d**	**f**
		يَ *ya*	نَ *na*	**t**	**p**	**f**

Table 8.5 Rules for predicting the values of the morphological features of Person, Number and Gender for imperative verbs

	Person Category	Prefix Imperative letter	Subject suffixed-pronoun	Person (9)	Number (8)	Gender (7)
Imperative verb **(1, “v”)** **(3, “i”)**	Second Person المُخاطَب *al-muḫāṭab*	ا ’	-	**s**	**s**	**m**
		ا ’	ا *ā*	**s**	**d**	**x**
		ا ’	وا *ū*	**s**	**p**	**m**
		ا ’	يْ *ī*	**s**	**s**	**f**
		ا ’	نَ *na*	**s**	**p**	**f**

8.4.2 Rules for Predicting the Morphological Feature of Rational

The Morphological feature of Rational (see section 6.2.17) is important in deriving the sound plural from rational or irrational nouns (*i.e.* an adjective describing an irrational masculine word, may forme its feminine sound plural by adding ات *āt* to the end of the adjective, as in جَبَلٌ شَاهِقٌ *ğabal*^{un} *šāhiq*^{un} “high mountain” has the plural of جِبَالٌ شَاهِقَاتٌ *ğibāl*^{un} *šāhiqāt*^{un} high mountains).

Rules for predicting the morphological feature of Rational depend on the main and sub part-of-speech categories of the analyzed word. Table 8.6 lists the set of rules used to predict the value of the morphological feature of Rational.

The morphological feature of Rational is hard to predict automatically depending on the rules of the main and sub part-of-speech of the word. Classifying words into rational or irrational depends on the semantics of the word itself and its context. For example, an adjective should agree in terms of rationality with the person or thing being described. If the adjective describes a person as in رَجُلٌ طَوِيلٌ *rağul*^{un} *ṭawīl*^{un} “a tall man”, then the adjective طَوِيلٌ *ṭawīl*^{un} “tall” is rational. But if the adjective describes a thing such as طَرِيْقٌ طَوِيلٌ *ṭarīqun ṭawīl*^{un} “a long road”, then the adjective طَوِيلٌ *ṭawīl*^{un} “long” is irrational. Therefore, a comprehensive dictionary which includes Rational information for each dictionary entry is needed to determine the correct attribute value of rational for the described nouns. An agreement algorithm is also needed to match Rational attributes of the adjective and the described nouns. Other types of agreement such as verb-subject agreement are also applicable to predict the value of Rational.

The set of rules designed to predict the value of the morphological feature of Rational depends on assigning a default value of rational or irrational to words depending on their sub part of speech, especially for words that need dictionary lookup to find their morphological features. Some words which belong to sub part-of-speech category such as demonstrative pronouns can be gathered and classified into rational and irrational. Table 8.6 shows some of these rules. If these rules are not applied then a default value is assigned depending on the sub part of speech of the analyzed word. Table 8.7 shows the types of nouns that accept rational as a default value, and the types of nouns that accept irrational as a default value. The default value of Qur’an verbs is rational.

Table 8.6 Rules for predicting the values of the morphological features of Rational

Position 17	**Rational** العاقل وغير العاقل *al-‘āqil wa ġayir al-‘āqil*		
Category	***Rule***		
Rational عاقِل ***‘āqil*** **(h)**	Singular proper nouns (personal names)	**n**	Personal nouns list
	Some demonstrative pronouns	**d**	أولئك *’ulā’ika* “Those”
	Some conditional nouns	**n**	من *man* “who?”
	Some relative pronouns	**r, c**	من، man “who”
	Some interrogative pronouns	**b**	من، من ذا *man, man ḏā* “who?, who is?”
	Allusive nouns	**a**	
Irrational غَيْر عَاقِل ***ġayr ‘āqil*** **(n)**	Singular proper nouns (organization and location names)	**n**	Organizations list and Locations list
	Some demonstrative pronouns	**d**	تلك *tilka* “that”
	Some conditional nouns	**h**	ما ، مهما *mā, mahmā* “what, whatever”
	Some relative pronouns	**r, c**	ما *mā* “what”
	Some interrogative pronoun	**b**	ما، ماذا *māḏā,mā* “what”
	Allusive nouns	**a**	

Table 8.7 Default value of Rational and Irrational for sub part-of-speech categories of nouns, with a tag symbol at position 2

Category	**Noun types**	
Rational	• Pronoun (**p**)	• Five nouns (**f**)
	• Active participle (**u**)	• Relative noun (*****)
	• Intensive Active participle (**w**)	• Diminutive (**y**)
	• Passive participle (**k**)	
Irrational	• Gerund / Verbal noun (**g**)	• Instrumental noun (**z**)
	• Gerund with initial *mīm* (**m**)	• Generic noun (**q**)
	• Gerund of instance (**o**)	• Numeral (**+**)
	• Gerund of state (**s**)	• Verb-like noun (**&**)
	• Gerund of emphasis (**e**)	• Form of exaggeration (**x**)
	• Gerund of profession (**i**)	• Collective noun (**$**)
	• Allusive noun (**a**)	• Plural generic noun (**#**)
	• Adverb (**v**)	• Elative noun (**@**)
	• Adjective (**j**)	• Blend noun (**%**)
	• Noun of place (**l**)	• Ideophonic interjection (**!**)
	• Noun of time (**t**)	

8.4.3 Rules for Predicting the Morphological Feature of Noun Finals

Nouns are classified into six categories according to their final letters. Nouns that end with a consonant letter are called sound nouns. Semi-sound nouns end with a vowel letter proceeded by a silent letter. A noun with a shortened ending ends with *'alif* or *'alif maqṣūrā,* if the last letter of the root is *wāw* or *yā'*. If the noun ends with an added*'alif* and *hamza*^{h} then it is called a noun with extended ending. A Noun with a curtailed ending ends with *yā'* proceeded by a letter that has the short vowel of *kasra*^{h}. Finally, a noun with a deleted ending has fewer letters than its root. See section 6.2.22. Table 8.8 shows the rules for predicting the morphological feature of Noun Finals and the related features.

The rules for predicting the value of the morphological feature of Noun Finals mainly depends on the long stem and the root of the analyzed word. The rules check the final letters of the long stem against a set of conditions that classify nouns into 6 categories. Knowing the value of the Noun Finals feature helps in specifying other features such as morphological features of Inflectional Morphology and Case and Mood Marks. Case marks cannot appear on the last letter of the nouns with shortened ending, and only *fatḥa*^{h}, the mark of the accusative case appears on the last letter of nouns with curtailed ending.

Table 8.8 Rules for predicting the values of the morphological features of Noun Finals

Category	Rule	Tag	Other features
Sound noun **الاسم صحيح الآخر** ***al-'ism ṣahīh al-'āir***	The last letter of the long stem is a consonants and not *hamza*^{h}.	**s**	• Inflectional Morphology: noun is triptote / fully declined (10, '**v**'). • Case marks appear on the last letter of the long stem.
Semi-sound noun **الاسم شبه الصحيح** ***al-'ism šibh aṣ-ṣaḥīḥ***	The last letter of the stem is a vowel and the previous letter is silent (*i.e.* has *sukūn* as short vowel).	**i**	• Inflectional Morphology: noun is triptote / fully declined (10, '**v**'). • Case marks appear on the last letter of the long stem.
Noun with shortened ending **الاسم المقصور** ***al-'ism al-maqṣūr***	The last letter of the stem is either *'alif* or *'alif maqṣūrā*, and the last letter of the root is *wāw* or *yā'*.	**t**	• Inflectional Morphology: noun is triptote / fully declined (10, '**v**'). • Case markers do not appear on the last letter of the stem.
Noun with extended ending **الاسم الممدود** ***al-'ism al-mamdūd***	The last letter of the stem is either added *'alif*, or the last two letters of the stem are added *'alif* followed by *hamza*^{h} or added *'alif* followed by *wāw*, and the last letter of the root is not *wāw* or *yā'*.	**e**	• Inflectional Morphology: noun is triptote / fully declined (10, '**v**'). Except, if the root is quadriliteral or quinquiliteral, then the noun is non-declinable (10, '**p**'). • Case markers appear on the last letter of the stem.
Noun with curtailed ending **الاسم المنقوص** ***al-'ism al-manqūṣ***	The last letter of the stem is *yā'* proceeded by a letter that has the short vowel *kasra*^{h}, and the last letter of the root is *yā'*.	**c**	• Inflectional Morphology: noun is triptote / fully declined (10, '**v**'). Except, if the word is a broken plural (8, '**b**'), then the noun is non-declinable (10, '**p**'). • Only accusative case marker appears on the last letter of the stem. Nominative and genitive case markers do not appear.
Noun with deleted ending **الاسم محذوف الآخر** ***al-'ism maḥḏūf al-'āẖir***	The stem consists of two letters, or the stem consists of three letter where the third letter is *tā' marbūta*^{h}, and the word has a triliteral root where the last root letter is a vowel.	**d**	• Inflectional Morphology: noun is triptote / fully declined (10, '**v**'). • Case marks appear on the last letter of the long stem.

8.5 Output Format

The final outputs of the SALMA – Tagger include the input word and all possible analyses. Each analysis includes information about the root, the lemma, the pattern, the full vowelized form, the tokenization of the word into morphemes, and the detailed description of the morphosyntactic information of each morpheme using SALMA – Tag. The output of the SALMA – Tagger covers all types of information recommended by the ALCCSO/KACST standards. Moreover, the SALMA – Tagger assigns a SALMA – Tag to each morpheme which captures the detailed and fine-grained morphosyntactic information of that morpheme whether it is a proclitic, prefix, stem, suffix or enclitic. The ALECSO/KACST standards recommend the description of the morphosyntactic information of the whole word or main stem only. Intermediate results can also be obtained from the different modules of the SALMA – Tagger such as root, lemma, pattern and possible vowelized forms of the word.

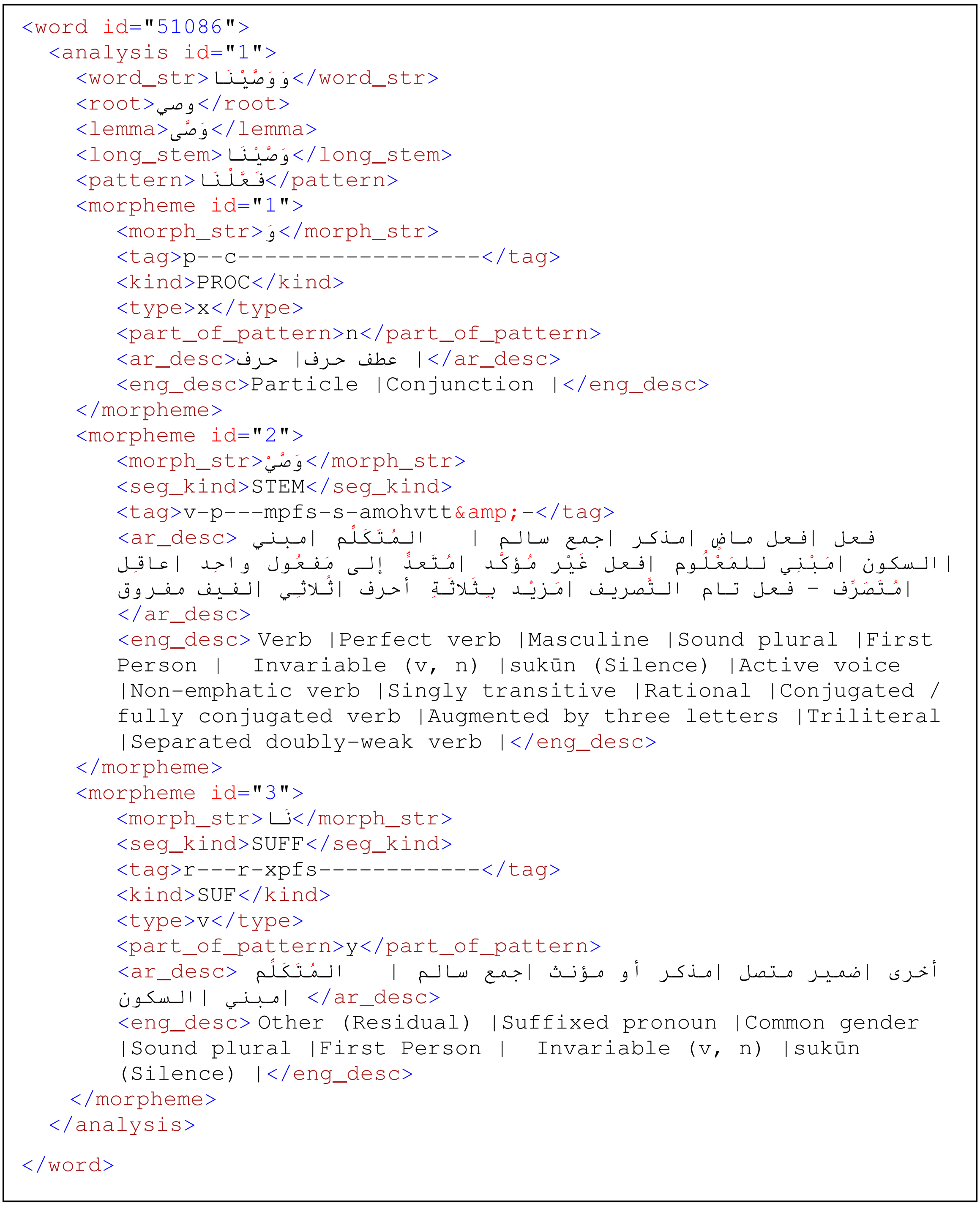
Several formats are available to format the analyses resulted by SALMA – Tagger. The results are output as a tab-separated file, as XML file and/or HTML page. The alternative formats and file types are provided to ensure wider re-use of the results of the SALMA – Tagger in different text analytics applications for Arabic. We want to tag an Arabic Corpora with fine-grained morphosyntactic information. Therefore, these formats were selected to be compatible with accepted standards for storing text corpora. These standard formats also allow the results to be easily integrated with corpus analysis software where simple tokenization, concordancing and corpus query language can be used to investigate the results of the SALMA – Tagger.

A widely-used format to store text corpora is the tab-separated column text-file. This format has been used since the first version of Brown and LOB corpus. The SALMA – Tagger formats its outputs in a tab-separated column file which represents a compatible result format with the widely-used corpus format. The SALMA – Tagger follows the same format as the MorphoChallenge 2009 Qur'an gold standard, see chapter 9. This format stores a word and its analyses per line. The first column contains the input word, and then the analysis is broken down into three columns: the root, the pattern, and the morphemes. A SALMA – Tag is assigned to each morpheme separated by a single space. The morphemes are comma separated. Figure 8.26 shows sample of the SALMA – Tagger results formatted in a tab separated column file.

```
وَوَصَّيْنَا    وصي    فَعَّلْنَا    وَ p--c------------------, وَصَّيْ v-p---mpfs-s-amohvtt&-, نَا r---r-xpfs-s----------
الْإِنْسَانَ    أنس    فِعْلَان    الْ r---d-----------------, إِنْسَانَ nq----ms-pafd---hdbt-s
بِوَالِدَيْهِ    ولد    فَاعِل    بِ p--p------------------, وَالِدَ nq----ms-pafd---hdbt-s, يْ r---r-xdts-s----------,
                        هِ r---r-msts-k----------
حُسْنًا    حسن    فُعْل    حُسْنَ ng----ms-vafi---ndst-s, ًا r---k------f----------
```

Figure 8.26 SALMA – Tagger output formatted in a tab separated column file

The second format uses XML files to store the results of the SALMA – Tagger. XML technology has become a widely-used and accepted standard to store text corpora when adding structures to the stored corpus. XML tags are used to provide the appropriate structure to the data stored in XML files. The format has a hierarchical structure where the word is at the top of the XML document object model. Several analyses are provided by the SALMA – Tagger to each word of the input text. Each analysis contains the root, the lemma, the long stem, the pattern and the morphemes of the word. For each morpheme the morphosyntactic information is stored. This is: the morpheme string, the SALMA – Tag, and the Arabic and English descriptions of the morphological features encoded in the tag. If the morpheme is a clitic or affix, then information such as morpheme kind, part of pattern and type are stored with the morpheme structure. Figure 8.27 shows the format of a word's analysis stored using XML file.



```
<word id="51086">
  <analysis id="1">
    <word_str>وَوَصَّيْنَا</word_str>
    <root>وصي</root>
    <lemma>وَصَّى</lemma>
    <long_stem>وَصَّيْنَا</long_stem>
    <pattern>فَعَّلْنَا</pattern>
    <morpheme id="1">
       <morph_str>وَ</morph_str>
       <tag>p--c-------------------</tag>
       <kind>PROC</kind>
       <type>x</type>
       <part_of_pattern>n</part_of_pattern>
       <ar_desc>حرف |حرف عطف |</ar_desc>
       <eng_desc>Particle |Conjunction |</eng_desc>
    </morpheme>
    <morpheme id="2">
       <morph_str>وَصَّيْ</morph_str>
       <seg_kind>STEM</seg_kind>
       <tag>v-p---mpfs-s-amohvtt&amp;-</tag>
       <ar_desc> فعل |فعل ماضٍ |مذكر |جمع سالم |   المُتَكَلِّم |مبني
       السكون |مَبْنِي لـلمَعْلُوم |فعل غَيْر مُؤكَّد |مُتَعدٍّ إلى مَفعُول واحِد |عاقِل
       مُتَصَرِّف – فعل تام التَّصريف |مَزيْد بِثَلاثَةِ أحرف |ثُلاثِي |لفيف مفروق|
       </ar_desc>
       <eng_desc> Verb |Perfect verb |Masculine |Sound plural |First
       Person |  Invariable (v, n) |sukūn (Silence) |Active voice
       |Non-emphatic verb |Singly transitive |Rational |Conjugated /
       fully conjugated verb |Augmented by three letters |Triliteral
       |Separated doubly-weak verb |</eng_desc>
    </morpheme>
    <morpheme id="3">
       <morph_str>نَا</morph_str>
       <seg_kind>SUFF</seg_kind>
       <tag>r---r-xpfs-------------</tag>
       <kind>SUF</kind>
       <type>v</type>
       <part_of_pattern>y</part_of_pattern>
       <ar_desc> أخرى |ضمير متصل |مذكر أو مؤنث |جمع سالم |   المُتَكَلِّم
       السكون| |مبني| </ar_desc>
       <eng_desc> Other (Residual) |Suffixed pronoun |Common gender
       |Sound plural |First Person |  Invariable (v, n) |sukūn
       (Silence) |</eng_desc>
   </morpheme>
  </analysis>

</word>
```

Figure 8.27 SALMA – Tagger outputs format stored in XML file

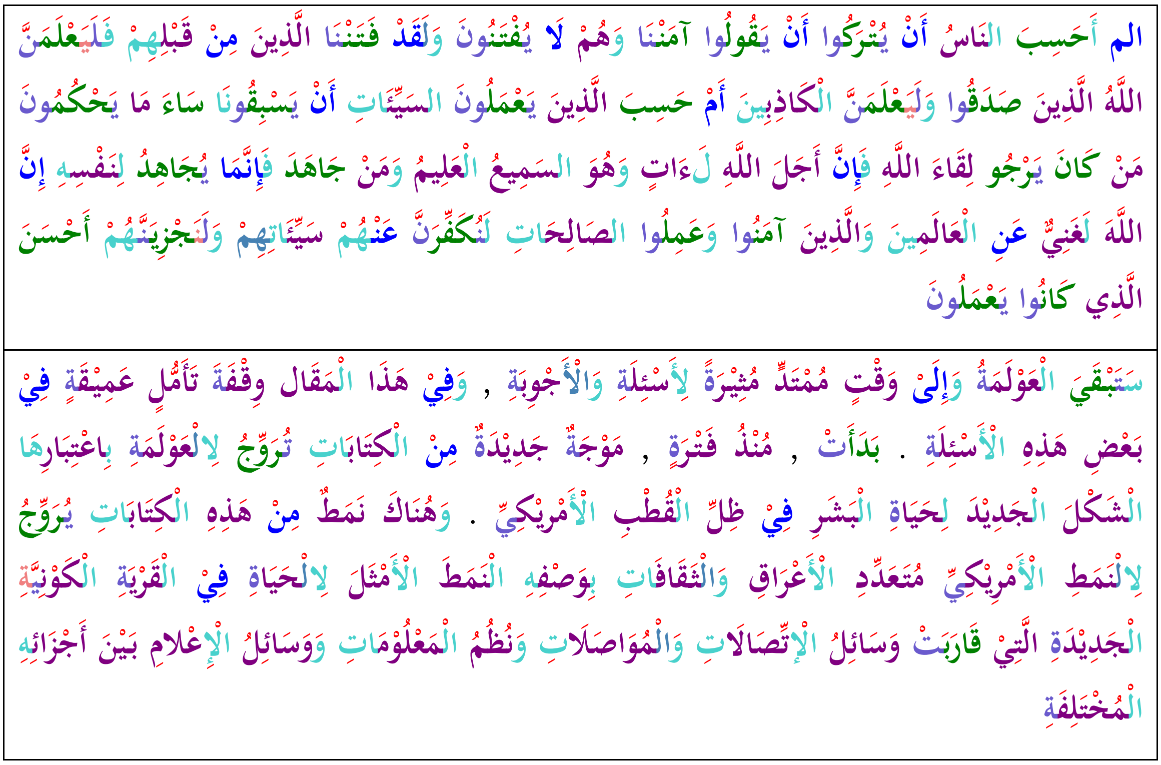
The third format uses HTML files to store and display the results of the SALMA – Tagger. HTML technology is used to display the results in a visualized way that shows the analyses of the words directly to the end user. This type of formatting is needed when an online interface is used to run the SALMA – Tagger by end users. However, the end-user has still got the choice to store the results in a tab-separated column file or XML file, to be downloaded directly after the user finishes the execution of the analyzer. The HTML format also allows the hyper-linking of the results with other online applications. For instance, the root of the analyzed word is linked with the web interface of the SALMA-ABCLexicon.The HTML output file contains the morphosyntactic information of the analyzed words such as: the root, the lemma, the long stem, the pattern, the word type and the word's morphemes. The morpheme type, the SALMA Tag and the Arabic and English descriptions are shown for each morpheme. Figure 8.28 shows a sample HTML page displaying some results of the SALMA – Tagger.

Word	**Root**	**Lemma**	**Long stem**	**Pattern**	**Word type**
وَوَصَّيْنَا	وصي	وَصَّى	وَصَّيْنَا	فَعَّلْنَا	
#	***Morpheme***		***Type***	***SALMA Tag***	
1	وَ		PROC	`p--c------------------`	
	Arabic description		حرف | حرف عطف|		
	English description		Particle |Conjunction |		
2	وَصَّيْ	(وَصَّيْنَا)	STEM	`v-p---mpfs-s-amohvtt&-`	
	Arabic description		فعل | فعل ماضٍ | مذكر | جمع سا لم | المتَكلِّم | مبني | السكون | مَبْنِي للمَعْلُوم | فعل غَيْر مُؤَكَّد | مُتَعدٍّ إلى مَفعُول واحِد | عاقِل | مُتَصَرِّف – فعل تام التَّصريف | مَزِيْد بِثَلاثَةِ أحرف | ثُلاثِي | لفيف مفروق |		
	English description		Verb |Perfect verb |Masculine |Sound plural |First Person | Invariable (v, n) |*sukūn* (Silence) |Active voice |Non-emphatic verb |Singly transitive |Rational |Conjugated / fully conjugated verb |Augmented by three letters |Triliteral |Separated doubly-weak verb |		
3	نَا		SUF	`r---r-xpfs-s----------`	
	Arabic description		أخرى | ضمير متصل | مذكر أو مؤنث | جمع سا لم | المتَكلِّم | مبني | السكون		
	English description		Other (Residual) |Suffixed pronoun |Common gender |Sound plural |First Person | Invariable (v, n) |*sukūn* (Silence) |		
Word	**Root**	**Lemma**	**Long stem**	**Pattern**	**Word type**
الْإِنْسَانَ	أنس	إِنْسَانَ	إِنْسَانَ	فِعْلَان	
#	***Morpheme***		***Type***	***SALMA Tag***	
1	الْ		PROC	`r---d-----------------`	
	Arabic description		أخرى | أداة تعريف |		
	English description		Other (Residual) |Definite article |		
2	إِنْسَانَ	(إِنْسَانَ)	STEM	`nq----ms-pafd---hdbt-s`	
	Arabic description		اسم | اسم الجنس | مذكر | مفرد | مُعرب – ممنوع من الصرف | منصوب | الفتحة / الفتح | مَعْرِفَة | عاقِل | مُتَصَرِّف – اسم مُشْتَقّ | مَزِيْد بِحَرفَيْن | ثُلاثِي | الاسم صحيح الآخر|		
	English description		Noun |Generic noun |Masculine |Singular |Non-declinable |Accusative (n), Subjunctive (v) |*fatḥa*^{h} |Definiteness |Rational |Inflected / Derived noun |Augmented by two letters |Triliteral |Sound noun |		

Word	Root	Lemma	Long stem	Pattern	Word type
بِوَالِدَيْهِ	ولد	وَالِدَ	وَالِدَيْ	فَاعِل	
#	***Morpheme***		***Type***	***SALMA Tag***	
1	بِ		PROC	`p--p------------------`	
	Arabic description			حرف | حرف جر |	
	English description			Particle |Preposition |	
2	وَالِدَ	(وَالِدَيْ)	STEM	`nu----md-vgki---ndbt-s`	
	Arabic description			اسم | اسم الفاعل | مذكر | مثنى | مُعرب – منصرف | مجرور | الكسرة | نَكِرَة | غَيْر عَاقِل | مُتَصَرِّف – اسم مُشْتَقٌّ | مَزِيْد بِحَرفَيْن | ثُلاثِي | الاسم صحيح الآخر	
	English description			Noun |Active participle |Masculine |Dual |Triptote / fully declined |Genitive (n) |*kasra*^{h} |Indefiniteness |Irrational |Inflected / Derived noun |Augmented by two letters |Triliteral |Sound noun |	
3	يْ		SUF	`r---r-xdts-s----------`	
	Arabic description			أخرى | ضمير متصل | مذكر أو مؤنث | مثنى | الغَائِب | مبني | السكون|	
	English description			Other (Residual) |Suffixed pronoun |Common gender |Dual |Third Person | Invariable (v, n) |*sukūn* (Silence) |	
4	هِ		ENC	`r---r-msts-k----------`	
	Arabic description			أخرى | ضمير متصل | مذكر | مفرد | الغَائِب | مبني | الكسرة|	
	English description			Other (Residual) |Suffixed pronoun |Masculine |Singular |Third Person | Invariable (v, n) |*kasra*^{h} |	
Word	**Root**	**Lemma**	**Long stem**	**Pattern**	**Word type**
حُسْنًا	حسن	حُسْنَ	حُسْنَاً	فُعْل	
#	***Morpheme***		***Type***	***SALMA Tag***	
1	حُسْنَ	(حُسْنَاً)	STEM	`ng----ms-vafi---ndst-s`	
	Arabic description			اسم | المصدر | مذكر | مفرد | مُعرب – منصرف | منصوب | الفتحة / الفتح | نَكِرَة | غَيْر عَاقِل | مُتَصَرِّف – اسم مُشْتَقٌّ | مُجَرَّد | ثُلاثِي | الاسم صحيح الآخر|	
	English description			Noun |Gerund |Masculine |Singular |Varied (n) |Accusative (n), Subjunctive (v) | *fatḥa*^{h} |Indefinite |Non-human |Derivable – Derived noun (n) |Unaugmented |Tri-literal |Sound noun |	
2	اً		SUF	`r---k------f----------`	
	Arabic description			أخرى | تنوين | الفتحة / الفتح |	
	English description			Other (Residual) |*tanwīn* |*fatḥa*^{h} |	

Figure 8.28 SALMA – Tagger outputs formatted in HTML file

Finally, the colour-coding module is used to visualize the morphosyntactic information such as the word's morphemes and its part of speech coded in colours. This colour-coding output format visualizes the complexity of the Arabic words, and the number and types of morphemes that forms a single word. Each morpheme is coloured depending on its type and part of speech. The details of the colouring scheme were discussed in section 8.3.5.3. The coloured outputs are displayed to the end-user through a web interface as coloured-coded text. The hyper-linking properties of web applications allow us to show the detailed analyses of each word of the displayed text by following the link assigned to each word. Figure 8.25 in section 8.3.5.3 shows an example of detailed analysis of the colour-coded word. Figure 8.29 shows two samples of colour-coded text, the top text is a Qur’an text – chapter 29, and the second sample is a MSA text taken from the CCA.



الم أَحَسِبَ النَاسُ أَنْ يُتْرَكُوا أَنْ يَقُولُوا آمَنْنَا وَهُمْ لَا يُفْتَنُونَ وَلَقَدْ فَتَنْنَا الَّذِينَ مِنْ قَبْلِهِمْ فَلَيَعْلَمَنَّ
اللَّهُ الَّذِينَ صَدَقُوا وَلَيَعْلَمَنَّ الْكَاذِبِينَ أَمْ حَسِبَ الَّذِينَ يَعْمَلُونَ السَيِّئَاتِ أَنْ يَسْبِقُونَا سَاءَ مَا يَحْكُمُونَ
مَنْ كَانَ يَرْجُو لِقَاءَ اللَّهِ فَإِنَّ أَجَلَ اللَّهِ لَءَاتٍ وَهُوَ السَمِيعُ الْعَلِيمُ وَمَنْ جَاهَدَ فَإِنَّمَا يُجَاهِدُ لِنَفْسِهِ إِنَّ
اللَّهَ لَغَنِيٌّ عَنِ الْعَالَمِينَ وَالَّذِينَ آمَنُوا وَعَمِلُوا الصَالِحَاتِ لَنُكَفِّرَنَّ عَنْهُمْ سَيِّئَاتِهِمْ وَلَنَجْزِيَنَّهُمْ أَحْسَنَ
الَّذِي كَانُوا يَعْمَلُونَ

سَتَبْقَي الْعَوْلَمَةُ وَإِلَىْ وَقْتٍ مُمْتَدٍّ مُثِيْرَةً لِأَسْئِلَةِ وَالْأَجْوِبَةِ , وَفِيْ هَذَا الْمَقَال وِقْفَةَ تَأَمُّلٍ عَمِيْقَةٍ فِيْ
بَعْضِ هَذِهِ الْأَسْئِلَةِ . بَدَأَتْ , مُنْذُ فَتْرَةٍ , مَوْجَةٌ جَدِيْدَةٌ مِنْ الْكِتَابَاتِ تُرَوِّجُ لِالْعَوْلَمَةِ بِاعْتِبَارِهَا
الْشَكْلَ الْجَدِيْدَ لِحَيَاةِ الْبَشَرِ فِيْ ظِلِّ الْقُطْبِ الْأَمْرِيْكِيِّ . وَهُنَاكَ نَمَطٌ مِنْ هَذِهِ الْكِتَابَاتِ يُرَوِّجُ
لِالْنَمَطِ الْأَمْرِيْكِيِّ مُتَعَدِّدِ الْأَعْرَاقِ وَالْثَقَافَاتِ بِوَصْفِهِ الْنَمَطَ الْأَمْثَلَ لِالْحَيَاةِ فِيْ الْقَرْيَةِ الْكَوْنِيَّةِ
الْجَدِيْدَةِ الَّتِي قَارَبَتْ وَسَائِلُ الْإِتِّصَالَاتِ وَالْمُوَاصَلَاتِ وَنُظُمُ الْمَعْلُوْمَاتِ وَوَسَائِلِ الْإِعْلَامِ بَيْنَ أَجْزَائِهِ
الْمُخْتَلِفَةِ

Figure 8.29 Colour coded output of the analyzed text samples of the Qur’an and MSA.

8.6 Chapter Summary

Morphological analyses and part of speech (PoS) tagging are very important and basic applications of Natural Language Processing. In this chapter we highlighted the importance of morphosyntactic analyses in a wide range of NLP applications. Arabic has many morphological and grammatical features, including sub-categories, person, number, gender, case, mood, etc. More fine-grained tag sets are often considered more appropriate. The additional information may also help to disambiguate the (base) part of speech.

The SALMA – Tagger is a morphological analyzer for Arabic text which depends on pre-stored lists of prefixes, suffixes, roots, patterns, function words, etc. These lists were extracted by referring to traditional grammar books. The affixes lists were verified by analyzing the Qur’an, the Corpus of Contemporary Arabic, the Penn Arabic Tree bank and the text of the 23 traditional Arabic lexicons as a fourth corpus. The prefixes list contains 220 prefixes. The suffixes list contains 474 suffixes and the patterns list contains 2,730 verb patterns and 985 nouns patterns.

The morphological analyzer was developed to analyze the word and specify its morphological features. The SALMA – Tag Set is used as standard for the development of the morphological analyzers. The morphological analyzer uses the tokenization scheme of Arabic words that distinguishes between five parts of word's morphemes (*i.e.* proclitics, prefixes, stem, suffixes and enclitics). Each part is given a fine-grained SALMA Tag that encodes 22 morphosyntactic categories of the morpheme (or possibly multiple tags if the part has multiple clitic or affix).

The morphological analyzer uses linguistic lists of functional words, named entities and broken plural lists. It also used the broad-coverage lexical resource constructed by analyzing 23 traditional Arabic lexicons. The coverage of the constructed broad-coverage lexical resource showed that about 85% of the words processed using the lemmatizer referenced the broad-coverage lexicon and retrieved correct analyses for the analyzed words.

The SALMA – Tagger algorithm involves a pipeline of processing stages, as shown in figure 8.4: Tokenization, Spelling error detecting and correcting, Clitics and affixes matching, Root extraction, lemmatizing, Pattern matching, Vowelization, Morphological features tag assignment and Colour-coding word's morphemes. These processing stages are useful on their own, such that users can choose the tool that suits their applications.

The SALMA – Tagger is an open-source fine-grain morphological analyzer for Arabic text. It only depends on open-source materials: lexicons, word lists and linguistic knowledge. The SALMA – Tagger consists of several modules which can be used independently to perform a specific task such as root extraction, lemmatizing and pattern extraction. Or, they can be used together to produce full detailed analyses of the words.

Chapter 9
Evaluation for the SALMA – Tagger

This chapter is based on the following sections of published papers:

Section 4 is based on section 5 in Sawalha and Atwell (2009a) and section 5 in Sawalha and Atwell (2009)

Section 5.1 is based on section 3 in Sawalha and Atwell (2011) and section 5 in Sawalha and Atwell (Under review)

Chapter Summary

The evaluation for the SALMA - Tagger depends on developing proposed standards for evaluating morphological analyzers for Arabic text, based on our experiences and participation in two evaluation contests: the ALECSO/KACST initiative for developing and evaluating morphological analyzers; and the MorphoChallenge 2009 competition. A reusable general purpose gold standard (the SALMA – Gold Standard) was constructed for evaluating the SALMA – Tagger. It can be reused to evaluate other morphological analyzers for Arabic text and to allow comparisons between the different analyzers. The SALMA – Gold Standard is adherent to standards, enriched with fine-grained morphosyntactic information of each morpheme of the gold standard text samples, contains two text samples of about 1000-word each representing two different text domains and genres of both vowelized and non-vowelized text taken from the Qur'an – chapter 29 and the CCA, and it is stored in several standard formats to allow wider reusability.

The SALMA – Gold Standard was used to evaluate the SALMA-Tagger. The evaluation focused on measuring the prediction accuracy of the 22 morphological features encoded in the SALMA – Tags for each of the gold standard's text sample morphemes. The results show that 53.50% of the Qur'an text sample morphemes and 71.21% of the CCA text sample morphemes were correctly tagged using "exact match" with the gold standard's morpheme tags. The evaluation reported the accuracy, recall, precision, f1-score and the confusion matrix for each morphological feature category to report for users who will use/reuse the SALMA – Tagger or parts of it, the prediction accuracy of the attributes of each morphological feature category. The prediction accuracy scored highly for 15 morphological feature categories at 98.53% -100% for the CCA test sample and 90.11% - 100% for the Qur'an test sample, while slightly lower accuracy was scored by the other 7 morphological feature categories at 81.35% - 97.51% for the CCA test sample and 74.25% - 89.03% for the Qur'an test sample.

9.1 Introduction

Several morphological analyzers for different languages and especially for English are available online, such as: *EMERGE*, *SProUT*, *FLEMM*, *FreeLing, POSTAG, ROSANA, TWOL,* and *XeLDA,* see section 2.3. The high accuracy results achieved by the morphological analyzers is due to: the availability of standard tag sets used to encode the morphosyntactic features of the analyzed words; the availability of morphosyntactically annotated corpora for free use by the research community; and the availability of the evaluation methodologies and standards for evaluating the results of the morphological analyzers and allowing comparative evaluations between them (Hamada 2010).

However, there are no evaluation prerequisites (*i.e.* standards and resources) available for Arabic whether automatic or manual. Therefore, the evaluation of morphological analyzers for Arabic text is not an easy task, and needs more investigation of the specific morphosyntactic features of Arabic, development of a morphosyntactically tagged representative corpus and the proposal of agreed standards to encode the results of the morphosyntactic features of the output analyses.

Two community-based experiences for evaluating morphological analyzers for Arabic text and proposed guidelines for evaluation are the ALECSO/KACST initiative[62] (Hamada 2010) and the MorphoChallenge[63] competition (Kurimo et al. 2009). The ALECSO/KACST initiative aimed to encourage the development of open-source morphological analyzers for Arabic text which are high-accuracy, and easy to develop, can be integrated into higher-level text analytics applications, and adhere to agreed standard guidelines. The MorphoChallenge competition aims to develop unsupervised morphological analyzers to be used for different languages including English, French, German, Finish, Turkish and Arabic. The competition evaluates the participant systems against previously prepared gold standards for each language. The unsupervised morphological analyzer that achieves the highest accuracy results in its outputs applied to the 6 languages wins the competitions. The two experiences are discussed in sections 9.2 and 9.3 respectively.

This chapter focuses on evaluation techniques for morphological analyzers for Arabic text. The chapter reflects our experiences on evaluating morphological analyzers as participants in the ALECSO/KACST initiative and the MorphoChallenge 2009 competition. The chapter develops and proposes applicable standard guidelines for evaluating morphological analyzers for Arabic text. These guidelines were applied to

[62] The workshop of morphological analyzers experts for Arabic language (اجتماع خبراء المحلّلات الحاسوبية الصرفيّة للغة العربية) 26 -28 April 2009, Damascus, Syria http://www.alecso.org.tn/index.php?option=com_content&task=view&id=1234&Itemid=1002&lang=ar

[63] MorphoChallenge 2009 http://research.ics.tkk.fi/events/morphochallenge2009/

evaluate the SALMA – Tagger. The evaluation procedure and results are discussed in the chapter.

9.2 ALECSO/KACST Initiative Guidelines for Evaluating Morphological Analyzers for Arabic Text

The ALECSO/KACST initiative aimed to encourage the development of open-source morphological analyzers for Arabic text which are high-accuracy, and easy to develop, can be integrated into higher-level text analytics applications, and adhere to agreed standard guidelines. The organizers invited world-wide Arabic morphological analyzer experts from universities, research institutions, software companies, a private legal institution and a non-governmental research funding organization along with Arabic language scholars to a workshop held in the Arabic Language Academy of Damascus, Syria in April 2009.

The participants presented the specifications of their morphological analyzers, the development methodologies, the initial results of evaluation, and demos of the developed systems. The ALECSO/KACST initiative evaluation committee presented the specifications of the required morphological analyzer for Arabic text (Al-Bawaab 2009; Hamada 2009a); see section 8.2. The evaluation committee also presented the evaluation methodology. Then the participants discussed the proposed evaluation methodology and agreed on the evaluation guidelines and procedures that would be followed to fairly evaluate and compare the different morphological analyzers. The discussions were based on the proposed evaluation methodologies presented by the participants (Dichy 2009; Hamada 2009b; Sawalha and Atwell 2009b).

The ALECSO/KACST initiative agreed to organize a competition between the participants' analyzers. The evaluation committee provided the output format of the morphological analyzer and a test dataset consisting of selected words to represent most morphological and inflectional cases of Arabic words. A period of two months was given to the researchers to format the output of their analyzers to match the recommended format. On the day of the competition, the evaluation committee provided the participants with the test dataset containing 15 words. The participants ran their morphological analyzers on this test list and they returned the results of their systems one day after receiving the test list. Then the evaluation committee evaluated the results received and announced the winner of the competition. However, the procedure they followed to evaluate the morphological analyzer was not reported, and the comparative evaluation results from participants' analyzers in respect to the agreed evaluation guidelines were not revealed. This section describes in detail the ALECSO/KACST initiative standards and guidelines for evaluating morphological analyzers for Arabic text.

The evaluation process involves analyzing the outputs of the analyzers given a test dataset consisting of selected words which represent most morphological and inflectional cases of Arabic words. The outputs of the morphological analyzers are evaluated according to two criteria: linguistic analyses and technical specifications (*i.e.* the approach to implementation, the extent to which it is user-friendly, the database management, the copyright and licensing issues and the accuracy metrics of recall and precision) (Hamada 2009b).

9.2.1 Evaluation of the Linguistic Specifications

The evaluation according to linguistic specifications checks the ability of the morphological analyzer to specify the morphosyntactic features of the analyzed words. The evaluation criteria are mainly based on the recommended morphosyntactic requirements for developing robust morphological analyzers for Arabic text (Al-Bawaad 2009; Hamada 2009b, Zaied 2009) and the development standards agreed by the participants, see section 8.2. The evaluation criteria include (Hamada 2009b):

- The ability to analyze all forms of words (*i.e.* fully vowelized, partially vowelized and non-vowelized).
- The ability to tokenize the analyzed word and to specify the word's morphemes (*i.e.* proclitics, prefixes, stem, suffixes and enclitics).
- The ability to extract all correct roots and patterns of the analyzed word.
- The ability to specify the main part of speech of the analyzed word.
- The ability to add the correct vowelization to the analyzed word.
- The ability to identify the morphological features of verbs such as: transitivity, augmented or unaugmented, number of root letters, person, voice and mood.
- The ability to identify the morphological features of nouns such as: gender, number, relative noun or noun of diminution, and variability and conjugation.

9.2.2 Evaluation of the Technical Specifications

The guidelines for evaluating the technical specifications contain five evaluation criteria. These criteria are: the approach to implementation, user friendliness, database management, copyright and licensing, and the accuracy metrics of recall and precision:

9.2.2.1 The Approach to Implementation

- The clarity and simplicity of the morphological analyzer algorithm and development approach.
- The novelty of the algorithm.
- The ability to integrate the morphological analyzer or parts of it into other Arabic text analytics applications.

- The availability of complete documentation that describes the morphological analyzer development approach and usage.

9.2.2.2 User Friendliness

- The user interface of morphological analyzer.
- The speed performance when analyzing words (word/second).
- The programming language used to develop the morphological analyzer.

9.2.2.3 Database Management

- The independence of the database (dictionaries) from the actual programs of the morphological analyzer.
- The ability to update the database (insert/delete/update) by the user, without running the morphological analyzer, or during the execution.

9.2.2.4 Copyright and licensing

This criterion checks whether the morphological analyzer depends on open-source resources or closed-source resources developed by others.

9.2.2.5 Evaluation Metrics of Recall and Precision

Recall and precision can be used to compute the accuracy of the results for each morphological analyzer. Then, the accuracy results can be ranked for comparative evaluation of morphological analyzers. Recall and precision are defined in the following formulas 9.1 and 9.2.

$$\text{Recall} = \frac{\text{Number of correct analyses}}{\text{Number of input words (test words)}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(9.1)$$

$$\text{Precision} = \frac{\text{Number of correct analyses}}{\text{Number of analyzed words}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(9.2)$$

9.3 MorphoChallenge Guidelines for Evaluating Morphological Analyzers for Arabic Text

The Morpho Challenge task is to develop an unsupervised learning algorithm which can return the morpheme analyses of each word given lists of words of in a number of target languages. In 2009, these were Arabic, English, Finish, German and Turkish. The algorithm should be as language-independent as possible. All words in the training corpus occur in sentences, so the algorithm might utilize information about word context (Kurimo et al. 2009).

The training corpora were 3 million sentences for English, Finnish and German, and 1 million sentences for Turkish in plain unannotated text files. The training corpus for Arabic was the Qur'an, which is a small corpus consisting of only 78K words. The text of the Qur'an corpus is available in both vowelized and non-vowelized formats. For Arabic, the participants could test their algorithms using the vowelized words or the unvowelized, or both. The algorithms were separately evaluated against the vowelized and the non-vowelized gold standard analyses. For all Arabic data, the Arabic writing scripts were provided as well as the Roman script (Buckwalter transliteration[64]). However, only morpheme analyses submitted in Roman script were evaluated (Kurimo et al. 2009).

MorphoChallenge 2009 established three competitions for evaluating the morpheme analyses. Competition 1 evaluated the proposed morpheme analyses against a linguistic gold standard. It included all five test languages. The winners were selected separately for each language according to the highest F-measure of accuracy. Competition 2 evaluated the proposed morpheme analyses against information retrieval (IR) experiments, where the search was based on morphemes instead of words. The words in the documents and queries were replaced by their proposed morpheme representations. This competition included three of the test languages (Finish, German and English). Competition 3 evaluated the proposed morpheme analyses using a machine translation (MT) model where the translation was based on morphemes instead of words. The words in the source language document were replaced by their morpheme representation. This competition included two of the test languages (Finish and German). Translation was done from the test language to English. The performance was measured with BLEU scores (Kurimo et al. 2009).

9.3.1 MorphoChallenge 2009 Competition 1: Evaluation using Gold Standard

In Competition 1 the proposed unsupervised morpheme analyses were compared to the correct grammatical morpheme analyses of the linguistic gold standard. The gold standard morpheme analyses were prepared in the same format as the result file the participants were asked to submit, alternative analyses being separated by commas. The Qur'an gold standard included each word in a separate line. Each line contains the word, the root, the pattern and then the morphological and part-of-speech analysis (Kurimo et al. 2009).

[64] Buckwalter transliteration http://www.qamus.org/transliteration.htm

Unsupervised learning algorithms for analyzing Arabic text were only evaluated in competition 1.

> *"... The basis of the evaluation is, thus, to compare whether any two word forms that contain the same morpheme according to the participants' algorithm also has a morpheme in common according to the gold standard and vice versa. In practice, the evaluation is performed by randomly sampling a large number of morpheme sharing word pairs from the compared analyses. Then the precision is calculated as the proportion of morpheme sharing word pairs in the participant's sample that really has a morpheme in common according to the gold standard. Correspondingly, the recall is calculated as the proportion of morpheme sharing word pairs in the gold standard sample that also exist in the participant's submission ..."*
>
> (Kurimo et al. 2009)

The F-measure, which is the harmonic mean of precision and recall, was selected as the final evaluation measure:

$$F - measure = \frac{1}{\frac{1}{Precision}+\frac{1}{Recall}} \text{......................................}(9.3)$$

9.3.2 MorphoChallenge 2009 Qur'an Gold Standard

We developed the gold standard of the Qur'an to be used to evaluate morphological analyzers in Morphochallenge 2009 competition 1[65], which aimed to develop an unsupervised morphological analyzer to be used for different languages including Arabic. The gold standard size is 78,004 words. The Qur'an gold standard contains the full morphological analysis for each word, according to the morphological analysis of the Qur'an in the Tagged database of the Qur'an developed at the University of Haifa (Dror et al. 2004). Figure 9.1 shows a sample of the Qur'an gold standard.

[65] Qur'an dataset http://www.cis.hut.fi/morphochallenge2009/datasets.shtml

Vowelized Arabic script			
بِسْمِ	سم	None	ب+Prep , سم+Noun+Triptotic+Sg+Masc+Gen ,
اللّٰهِ	None	None	لَلاه+Noun+ProperName+Gen+Def ,
الرَّحْمَنِ	رحم	فَعلَان	رَحمَان+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def ,
الرَّحِيمِ	رحم	فَعِيل	رَحِيم+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def ,
Non-Vowelized Arabic script			
بسم	سم	None	ب+Prep , سم+Noun+Triptotic+Sg+Masc+Gen ,
الله	None	None	للاه+Noun+ProperName+Gen+Def ,
الرحمن	رحم	فعلان	رحمان+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def ,
الرحيم	رحم	فعيل	رحيم+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def ,
Vowelized Romanized script using Buckwalter transliteration scheme			
bisomi	sm	None	b+Prep , sm+Noun+Triptotic+Sg+Masc+Gen ,
All~hi	None	None	llaah+Noun+ProperName+Gen+Def ,
Alr~aHomani	rHm	faElaAn	raHmaan+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def ,
Alr~aHiymi	rHm	faEiyl	raHiim+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def ,
Von-vowelized Romanized script using Buckwalter transliteration scheme			
bsm	sm	None	b+Prep , sm+Noun+Triptotic+Sg+Masc+Gen ,
Allh	None	None	llAh+Noun+ProperName+Gen+Def ,
AlrHmn	rHm	fElAn	rHmAn+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def
AlrHym	rHm	fEyl	rHym+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def ,

Figure 9.1 A sample of the MorphoChallenge2009 Qur'an gold standard, in 4 alternate formats

9.4 Gold Standard for Evaluation

As with other NLP tasks, it is customary to use gold standards for evaluating morphological analyzers. This is discussed in section 2.3.2 of this thesis, along with construction of gold standard data sets for the Qur'an and MSA in section 3.4. This section proposes guidelines for constructing and using a gold standard for evaluation of a fine-grained morphological analyzer for Arabic text.

Gold standards are used to evaluate and measure the accuracy of automatic systems. The evaluation can be used to compare between different systems or algorithms on the same problem domain. It shows the successes and failings of an algorithm. Gold standards can be used to compute similarity between systems by highlighting the cases of agreed analyses and the cases when a tie resulted.

Moreover, a gold standard can be used to determine the specifications of the morphological analyzers by specifying which morphological features it can or cannot handle. This is another way to evaluate morphological analyzers, by describing their specifications.

To construct a gold standard for evaluation, we need to determine the problem domain of the algorithms to be evaluated, the corpus to be used as gold standard, the format of the gold standard, its size, the script used and transliteration scheme, and the phases of constructing the gold standard.

9.4.1 Problem domain

The gold standard will be used to evaluate morphological analyzers and part-of-speech taggers for Arabic text. The gold standard should have morphological information and part-of-speech tags for each word of the selected corpus.

9.4.2 The Corpora

Corpora are used to build gold standards. Many Arabic language corpora have been developed. But to build a widely used general purpose gold standard, corpora of different text domains, formats and genres of both vowelized and non-vowelized Arabic text are needed. Two open-source corpora are recommended to be used. First, the Qur'an corpus can be used in the construction of the gold standard. The Qur'an text is Classical Arabic, representing a genre-specific corpus which is morphologically different from Modern Standard Arabic. It represents a challenge to morphological analyzers for Arabic text because of its complex morphosyntactic features. The Qur'an sample is fully vowelized text. Second, the Corpus of Contemporary Arabic (CCA) is an open-source Arabic corpus representing Modern Standard Arabic (Al-Sulaiti and Atwell 2004; Al-Sulaiti and Atwell 2005; Al-Sulaiti and Atwell 2006).This corpus contains 1 million words taken from different genres collected from newspapers and magazines. It contains the following domains; Autobiography, Short Stories, Children's Stories, Economics, Education, Health and Medicine, Interviews, Politics, Recipes, Religion, Sociology, Science, Sports, Tourist and Travel and Science. The text in the CCA is non-vowelized.

9.4.3 Gold Standard Format

The gold standard will include detailed morphosyntactic information for each word of the gold standard. The analysis divides the words into their morphemes: proclitics, prefixes, stem, suffixes and enclitics. For each morpheme fine-grain morphological features information will be provided. The SALMA – Tag Set is recommended to be used to encode the morphological features of the word's morphemes (Hamada 2010). Moreover, the gold standard will contain the basic morphological information such as: the root, the lemma and the pattern of the words. The gold standard will be stored using different file formats to meet the wider-user specifications. Both tab-separated column files and XML files are recommended. A visual representation of the gold standard such as HTML tables is recommended. The visual representation allows the end-user to view the morphosyntactic information of the gold standard. Unicode utf-8 encoding is recommended to be used in all files (Bird et al. 2009 p.93) to enable a unified representation for Arabic letters on different platforms.

9.4.4 Gold Standard Size

The gold standard should be large enough to cover most cases that morphological analyzers have to handle. The gold standard size is measured by the number of words it contains.

9.5 Building the SALMA – Gold Standard

This section discusses the process of building the SALMA - Gold Standard for evaluating morphological analyzers for Arabic text. The proposed standards are based on the agreed standards and guidelines and our experiences and contributions to the ALECSO/KACST initiative and MorphoChallenge 2009 competition for developing and evaluating morphological analyzers for Arabic text.

The SALMA – Gold Standard is aimed at the wider research community for evaluating morphological analyzers for Arabic text, and comparisons between their outputs. Therefore, it includes detailed morphosyntactic information that can be produced by morphological analyzers such as: the input word, its root, lemma, pattern, word type and the word's morphemes. For each of the word's morphemes, the standard shows the morpheme type classified into proclitic, prefix, stem, suffix and enclitic, and a fine-grained SALMA – Tag which encodes 22 morphological feature categories of each morpheme. These morphological features are described in Arabic and English.

The format of the gold standard is an important issue. The proposed gold standard is formatted in different formats to meet a range of user needs. XML technology allows storage of the gold standard in a machine-readable structured format that increases its reusability. Tab separated column files are widely used by researchers. They are used following the Morphochallenge 2009 recommendations for constructing gold standards. Other formats are used to display the information of the gold standard for the end users. These formats include HTML files and the visual display of the gold standard in colour-coded format. The SALMA – Gold Standard for evaluating Arabic morphological analyzers is an open-source resource that is available to download.

Two text samples were selected to construct the SALMA – Gold Standard. The first text sample is Chapter 29 of the Qur'an representing classical Arabic. Section 9.5.1 discusses the construction of the Qur'an gold standard. The second text sample is taken from the CCA representing Modern Standard Arabic. Section 9.5.2 discusses the construction of the CCA gold standard. Both samples were selected to represent a wider range of text types, formats and genres.

9.5.1 The Qur'an Gold Standard

The SALMA Gold Standard Qur'an text sample was constructed by mapping from an existing specific format and broad tag set to the standardized format and fine-grained SALMA – Tag Set see section 7.2.

The Quranic Arabic Corpus sample text chosen was chapter 29, consisting of about 1000 words. An automated mapping algorithm was developed to map the Quranic Arabic Corpus script, morpheme tokenization and morphological tags to meet our proposed standards and guidelines. After that, the automatically mapped results including the morphological feature tags were manually verified and corrected, to provide a new fine-grain Gold Standard for evaluating Arabic morphological analyzers and part-of-speech taggers.

The mapping from the Quranic Arabic Corpus format and morphological tag set to the proposed standards and guidelines for constructing gold standards and the SALMA – Tag Set was done by the following six-step procedure:

1. **Mapping classical to modern character-set:** the Quranic Arabic Corpus uses the classical Othmani script of the Qur'an (77,430 words) which was mapped to Modern Standard Arabic (MSA) script (77,797 words). This was achieved by applying one-to-one mapping except for some cases where one word in Othmani script is mapped to two words in MSA such as the word يٰمُوسَىٰ *yāmūsā* 'O Musa "Moses"!' - in Othmani script this is one word but it is written as two words in MSA script: يَا مُوسَى *yā mūsā*.
2. **Splitting whole-word tags into morpheme tags:** the morphological tag in the Quranic Arabic Corpus is a whole-word tag, composed by combining the prefix with the stem and suffix morphological tags, separated by (+) signs. The words and their morphological tags were automatically divided into morphemes and morpheme tags.
3. **Mapping of feature-labels:** the mnemonics of the Quranic Arabic Corpus tags were mapped to their equivalent in the SALMA Tag Set. Then, SALMA Tag Set templates were applied to specify the applicable and non-applicable morphological features of the analyzed morpheme.
4. **Adjustments to morpheme tokenization:** due the differences between the underlying word tokenization model used in the Quranic Arabic Corpus and the one required for the SALMA Tag Set, we replaced the mapped tags of the prefixes and suffixes with SALMA tags by matching them to the clitics and affixes lists used by the SALMA Tagger.
5. **Extrapolation of missing fine-grain features:** for morphological features which are not included in the Quranic Arabic Corpus tag set, automatic "feature-

prediction” procedures applied linguistic knowledge extracted from traditional Arabic grammar textbooks, encoded as a computational rule-based system, to automatically predict the values of the missing morphological features of the word.

6. **Proofreading and correction:** the mapped SALMA tags were manually proofread and corrected by an Arabic language expert. The result is a sample Gold Standard annotated corpus for evaluating morphological analyzers and part-of-speech taggers for Arabic text. Sections 7.3 and 7.4 discuss the mapping process in detail.

The exact match of the prediction of all 22 features for a morpheme whole tags for the test sample is 53.5%, but some of the errors were very minor such as replacing one ‘?’ by ‘-’. The error-rate of individual features scored 2.01% for main part of speech, between 3% and 15% for morphological features coded in the QAC tags, and between 2% and 24% for features which do not exist in the QAC tags but can be automatically predicted.

9.5.1.1 Specifications of the Qur’an part of the SALMA Gold Standard

The construction of the SALMA – Gold Standard applied the proposed guidelines and standards for constructing gold standards for evaluating morphological analyzers of Arabic text. This section shows their application on the Qur’an sample of the SALMA – Gold Standard.

1- Problem domain

The Qur’an part of the SALMA – Gold Standard was constructed to evaluate morphological analyzers and part-of-speech taggers on Classical Arabic. This information includes the input word, root, lemma, pattern, and the appropriate segmentation of the word into its morphemes. The morphological features for each of the word’s morphemes were encoded using SALMA – Tags. The detailed and fine-grain morphosyntactic information was provided to enable the wider research community to evaluate their morphological systems using a unified standard that enables comparisons between the various evaluated systems.

2- The Corpus

This is text sample of the Qur’an, chapter 29 سورة العنكبوت *sūra*^{t} *al-‘ankabūt*. The Qur’an text represents a genre specific corpus which is morphologically different from Modern Standard Arabic. It represents a challenge to morphological analyzers for Arabic text because of its complex morphosyntactic features. The Qur’an sample is fully vowelized text. A non-vowelized copy is provided to evaluate morphological analyzers which do not accept vowelization for their input text. Morphological analyzers of Arabic text are expected to perform better on Modern Standard Arabic text than the Qur'an text.

3- Gold Standard Format

The SALMA – Gold Standard is stored using a variety of file formats. Firstly, XML files were used for storage. Suitable xml-tags were added to describe the detailed information of the analyses for words and their morphemes. Figure 9.3 shows an example of the SALMA – Gold Standard, Qur'an part, stored using XML files.

Secondly, widely used tab separated column files were used to store the gold standard following the Morphochallenge 2009 recommendations for constructing gold standards. Each word and its analysis were stored in a line where the word occupied the first column, followed by the root, the pattern and the morphemes on separate columns. The last column contains each morpheme which is followed by its SALMA Tag separated by a comma. Figure 9.2 shows an example of the SALMA – Gold Standard, Qur'an part, stored using a tab separated column file.

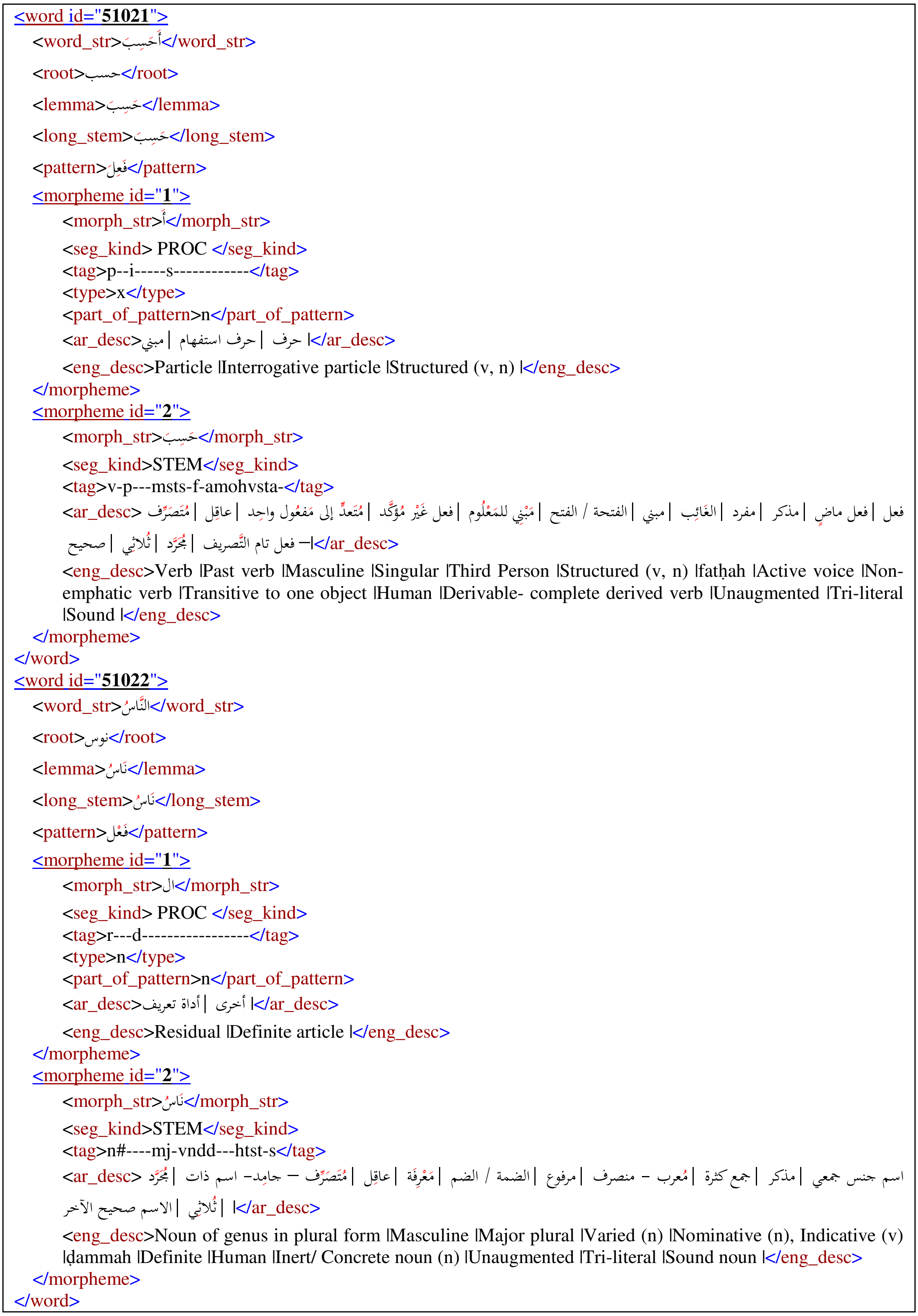
Other formats are used to display the information of the gold standard for end users. These formats include HTML files and the visual display of the gold standard in colour-coded format. The SALMA – Gold Standard for evaluating Arabic morphological analyzers is an open-source resource that is available to download. See section 8.5 output format of the SALMA – Tagger.

4- Gold Standard Size

The size of the gold standard is measured by the number of words it contains. The SALMA – Gold Standard, Qur'an part contains 976 words, of 603 word types. These words were generated from 243 different roots, 367 different lemmas and 175 different patterns. The number of morphemes in this part is 1,942 having 471 different SALMA – Tags.

أَحَسِبَ	حسب	حَسِبَ	فَعِلَ	أَ p--i-----s------------, حَسِبَ v-p---msts-f-amohvsta-
النَّاسُ	نوس	نَاسُ	فَعْل	ال r---d-----------------, نَاسُ n#----mj-vndd---htst-s
أَنْ				أَنْ p--g-----s-s----------
يُتْرَكُوا	ترك	تَرَكَ	يُفْعَلُوا	يُ r---a-----------------, تْرَكُ v-c---mptdao-pmohvtta-, وا r---r-mpts-s----------
أَنْ				أَنْ p--g-----s-s----------
يَقُولُوا	قول	قَالَ	يَفْعُلُوا	يَ r---a-----------------, قُولُ v-c---mptdao-amohvtto-, وا r---r-mpts-s----------
آمَنَّا	أمن	آمَنَ	فَاعَلْنَا	آمَنْ v-p---mpfs-s-amohvttc-, نَا r---r-xpfs-s----------
وَهُمْ				وَ p--c-----s-f----------, هُمْ np----mpts-si---hn----
لَا				لَا p--n-----s-s----------
يُفْتَنُونَ	فتن	فَتَنَ	يُفْعَلُونَ	يُ r---a-----------------, فْتَنُ v-c---mptdnn-pmohvtta-, ونَ r---r-mpts-f----------

Figure 9.2 A sample of the SALMA – Gold Standard, Qur'an part, stored using text file



```
<word id="51021">
   <word_str>أَحَسِبَ</word_str>
   <root>حسب</root>
   <lemma>حَسِبَ</lemma>
   <long_stem>حَسِبَ</long_stem>
   <pattern>فَعِلَ</pattern>
   <morpheme id="1">
      <morph_str>أَ</morph_str>
      <seg_kind> PROC </seg_kind>
      <tag>p--i-----s------------</tag>
      <type>x</type>
      <part_of_pattern>n</part_of_pattern>
      <ar_desc>حرف | حرف استفهام | مبني |</ar_desc>
      <eng_desc>Particle |Interrogative particle |Structured (v, n) |</eng_desc>
   </morpheme>
   <morpheme id="2">
      <morph_str>حَسِبَ</morph_str>
      <seg_kind>STEM</seg_kind>
      <tag>v-p---msts-f-amohvsta-</tag>
      <ar_desc> فعل | فعل ماضٍ | مذكر | مفرد | الغَائِب | مبني | الفتحة / الفتح | مَبْنِي للمَعْلُوم | فعل غَيْر مُؤَكَّد | مُتَعدٍّ إلى مَفعُول واحِد | عاقِل | مُتَصَرِّف
      صحيح | ثُلاثِي | مُجَرَّد | فعل تام التَّصريف –|</ar_desc>
      <eng_desc>Verb |Past verb |Masculine |Singular |Third Person |Structured (v, n) |fatḥah |Active voice |Non-emphatic verb |Transitive to one object |Human |Derivable- complete derived verb |Unaugmented |Tri-literal |Sound |</eng_desc>
   </morpheme>
</word>
<word id="51022">
   <word_str>النَّاسُ</word_str>
   <root>نوس</root>
   <lemma>نَاسُ</lemma>
   <long_stem>نَاسُ</long_stem>
   <pattern>فَعْل</pattern>
   <morpheme id="1">
      <morph_str>ال</morph_str>
      <seg_kind> PROC </seg_kind>
      <tag>r---d-----------------</tag>
      <type>n</type>
      <part_of_pattern>n</part_of_pattern>
      <ar_desc>أداة تعريف | أخرى |</ar_desc>
      <eng_desc>Residual |Definite article |</eng_desc>
   </morpheme>
   <morpheme id="2">
      <morph_str>نَاسُ</morph_str>
      <seg_kind>STEM</seg_kind>
      <tag>n#----mj-vndd---htst-s</tag>
      <ar_desc> اسم جنس جمعي | مذكر | جمع كثرة | مُعرب - منصرف | مرفوع | الضمة / الضم | مَعْرِفَة | عاقِل | مُتَصَرِّف – جامِد- اسم ذات | مُجَرَّد
      الاسم صحيح الآخر | ثُلاثِي | |</ar_desc>
      <eng_desc>Noun of genus in plural form |Masculine |Major plural |Varied (n) |Nominative (n), Indicative (v) |ḍammah |Definite |Human |Inert/ Concrete noun (n) |Unaugmented |Tri-literal |Sound noun |</eng_desc>
   </morpheme>
</word>
```

Figure 9.3 A sample of the SALMA – Gold Standard, Qur'an part, stored using XML file

9.5.2 The Corpus of Contemporary Arabic Gold Standard

The SALMA – Gold Standard CCA text sample was constructed by using the SALMA – Tagger, then manually selecting and correcting the analysis of each word according to its context. This semi-automatic approach was followed because of limitations of time, funds and availability of professional annotators. Therefore, manual annotation was not practical. On balance, it was more practical to run the SALMA – Tagger which produced the initial analyses necessary to construct the gold standard. Mapping from non-open-source part-of-speech tagged corpora such as the PATB was avoided because it contradicted the aim of constructing the SALMA – Gold Standard as an open-source resource available for the wider research community.

A 1000-word text sample was selected from the CCA. This MSA text sample was selected from three genres of the CCA: politics, sport and economics, the main three genres of newspaper articles. The selected text sample is non-vowelized. The construction of the SALMA – Gold Standard from the CCA text sample was done by selecting and correcting the outputs of the SALMA – Tagger run on this text sample. The SALMA – Tagger provided the detailed morphosyntactic information required by the gold standard such as: root, lemma, long stem, pattern, vowelized word and the word's morphemes. A SALMA Tag was provided for each morpheme as well.

The manual selection and correction was done because the SALMA – Tagger generates all possible analyses for each word. Therefore, one analysis suitable for the context was selected as a candidate analysis. Then, manual correction was carried out. The correction process involves verifying and correcting the detailed information about root, lemma, pattern, fully vowelized form of the word and the word's morphemes. The SALMA – Tag for each morpheme was then proofread and corrected.

The exact match of the prediction of all 22 features for a morpheme whole tags for the test sample is 71.12%, but some of the errors were very minor such as replacing one '?' by '-'.

9.5.2.1 Specifications of the CCA part of the SALMA Gold Standard

A similar methodology was followed to construct the SALMA – Gold Standard CCA part that applied the proposed guidelines and standards for constructing gold standards for evaluating morphological analyzers of Arabic text. This section shows their application on the CCA sample of the SALMA – Gold Standard.

1- Problem domain

The CCA part of the SALMA – Gold Standard was constructed to evaluate morphological analyzers and part-of-speech taggers on MSA text. The SALMA – Gold Standard contains detailed analysis of each word of the gold standard. This information includes the input word, root, lemma, pattern, fully vowelized form of the word, and the appropriate segmentation of the word into its morphemes. The morphological features for each of the word's morphemes were encoded using SALMA – Tags. The detailed and fine-grain morphosyntactic information was provided to satisfy a wider research community to evaluate their morphological systems using a unified proposed standard that enables comparisons between the various evaluated systems.

2- The Corpora

A text sample of the CCA consisting of about 1,000 words was selected. The CCA is a 1-million word open-source MSA corpus collected from newspapers and magazines which contains 14 genres. The selected sample was selected from politics, sport and economics, the main three genres of newspaper articles. The words of the CCA are morphologically simpler that the Qur'an text. However, this still represents a challenge to morphological analyzers for Arabic text. Possible challenges of the CCA text to morphological analyzers are borrowed word, named entities, new vocabulary, transliterated words and relative nouns. The CCA sample is non-vowelized text. Fully-vowelized forms of the words are provided in the gold standard. The morphological analyzers for Arabic text are required to produce the fully-vowelized form of the analyzed words.

3- Gold Standard Format

The SALMA – Gold Standard, CCA part used the unified file format which is used to store the Qur'an part of the gold standard. Both XML files provided with the appropriate xml-tags that describe the information stored in the gold standard, and tab separated column files where each column contains a piece of information stored in the gold standard, were used to format the detailed information of the gold standard. Figure 9.4 shows example of the SALMA – Gold Standard, CCA part, stored using XML files. Figure 9.5 shows example of the SALMA – Gold Standard, CCA part, stored using a tab separated column file.

Other formats are used to display the information of the gold standard for the end users. These formats include HTML files and the visual display of the gold standard in colour-coded format.

4- Gold Standard Size

The size of the gold standard is measured by the number of words it contains. The SALMA – Gold Standard, CCA part contains 1,122 tokens distributed into 1,015 Arabic words, 99 punctuation marks and 8 numbers. The sample contains 775 token types distributed into 756 Arabic word types, 13 punctuation marks and 6 numbers.

The Arabic words in the sample were generated from 421 different roots, 594 different lemmas and 215 different patterns. The number of morphemes in this part is 2,172 having 452 different SALMA – Tags.

```
<word id="11">
    <word_str>هذا</word_str>
    <v_word>هَذَا</v_word>
    <root>هذا</root>
    <lemma>هَذَا
    </lemma>
    <long_stem>هَذَا</long_stem>
    <word_type>Arabic Word</word_type>
    <word_kind>Stop Word</word_kind>
    <morpheme id="1">
        <morph_str>هَذَا</morph_str>
        <seg_kind>STEM</seg_kind>
        <tag>nd----ms-s-si---nns---</tag>
        <ar_desc>اسم | اسم الإشارة | مذكر | مفرد | مبني | السكون | نَكِرَة | غَيْر عَاقِل | غير مُتصرِّف | مُجَرَّد |</ar_desc>
        <eng_desc>Noun |Demonstrative pronoun  |Masculine |Singular |  Invariable (v, n) |sukūn (Silence)
        |Indefiniteness |Irrational |Non-Inflected (n, v) |Unaugmented  |</eng_desc>
    </morpheme>
</word>
<word id="12">
    <word_str>المقال</word_str>
    <v_word>المقَال</v_word>
    <root>قول</root>
    <lemma>مَقَال
    </lemma>
    <long_stem>مَقَال</long_stem>
    <pattern>مَفْعَل</pattern>
    <word_type>Arabic Word</word_type>
    <morpheme id="1">
        <morph_str>الْ</morph_str>
        <seg_kind>PRE</seg_kind>
        <tag>r---d-----------------</tag>
        <kind>proc</kind>
        <type>n</type>
        <part_of_pattern>n</part_of_pattern>
        <ar_desc>أخرى | أداة تعريف |</ar_desc>
        <eng_desc>Other (Residual) |Definite article |</eng_desc>
    </morpheme>
    <morpheme id="2">
        <morph_str>مَقَال</morph_str>
        <seg_kind>STEM</seg_kind>
        <tag>nq----fb-v??d---ntat-s</tag>
        <ar_desc> اسم | اسم الجنس | مؤنث | جمع تكسير | مُعرب - منصرف | مَعْرِفَة | غَيْر عَاقِل | مُتصَرِّف – جامِد- اسم ذات | مَزِيْد بِحَرف | ثُلاثِي |
        الاسم  صحيح الآخر |</ar_desc>
        <eng_desc>Noun |Generic noun |Feminine |Broken plural |Triptote / fully declined |Definiteness
        |Irrational |Primitive / Concrete noun  |Augmented by one letter |Triliteral |Sound noun |</eng_desc>
    </morpheme>
</word>
```

Figure 9.4 A sample of the SALMA – Gold Standard, CCA part, stored using XML file

وفي	وَفِيْ	في	فِيْ		Word	وَ p--c-----------------, فِيْ p--p-----s-?-----n----
هذا	هَذَا	هذا	هَذَا		Word	هَذَا nd----ms-s-si---nns---
المقال	المِقَال	قول	مَقَال	مَفْعَل	Word	الْ r---d----------------, مَقَال nq----fb-v??d---ntat-s
وقفة	وِقْفَةَ	وقف	وِقْفَةَ	فِعْلَة	Word	وِقْفَ ns----fs-vafi---ndat-s, ةَ r---t-fs-------------
تأمل	تَأَمُّلٍ	أمل	تَأَمُّلٍ	تَفَعُّل	Word	تَأَمُّلٍ ne----ms-vgki---ndbt-s
عميقة	عَمِيْقَةٍ	عمق	عَمِيْقَةٍ	فَعِيْلَة	Word	عَمِيْقَ nj----fs-v??i---hdbt-s, ةٍ r---t-fs-------------
في	فِيْ	في	فِيْ		Word	فِيْ p--p-----s-?-----n----
بعض	بَعْضِ	بعض	بَعْضِ	فَعْل	Word	بَعْضِ n+----ms-vgki---nnst-s
هذه	هَذِهِ	هذه	هَذِهِ		Word	هَذِهِ nd----mb-s-si---nns---
الأسئلة	الأَسْئِلَةِ	سأل	سُؤَال	أَفْعِلَة	Word	الْ r---d----------------, أَسْئِلَ nq----mb-vgkd---ntbt-s, ةِ r---t-fs-------------
.	.				Punct.	. u----s---------------

Figure 9.5 A sample of the SALMA – Gold Standard, CCA part, stored using text file

9.6 Deciding on Accuracy Measurements

The ALECSO/KACST initiative evaluated morphological analyzers for Arabic text according to both linguistic and technical specifications of the morphological analyzer and its outputs. However, no gold standard for evaluation was provided. They relied on linguists to assess the linguistic information produced by the morphological analyzers for examples of challenging words. The technical specifications were assessed by a computational linguist. Even though no evaluation results were reported by the ALECSO/KACST initiative for evaluation of morphological analyzers, they recommended to use recall and precision metrics to compute the accuracy of the morphological analyzers according to formulas 9.1 and 9.2. Section 9.2 discusses the ALECSO/KACST initiative for evaluating morphological analyzers.

The MorphoChallenge 2009 competition 1 evaluates the proposed morpheme analysis against a linguistic gold standard. The results of the participants' algorithms were compared with the gold standard by checking whether any two words have a morpheme in common. The best morphological analyzer was selected according to the highest F-measure of accuracy calculated using formula 9.3. The F-measure score is the harmonic mean of recall and precision. Precision was defined as the proportion of word pairs that share the same morpheme and that have a morpheme in common in the gold standard. Recall was defined as the proportion of morphemes sharing word pairs in the gold standard also found in the participants' results.

In general, morphological analyzers of Arabic text are required to produce all possible analyses of the word form out of context. The SALMA – Tagger produces all possible analyses of the analyzed word form. The absence of a gold standard for evaluating morphological analyzers that contains all possible and correct analyses and their morphosyntactic information (*i.e.* root, lemma, pattern, vowelization, word's morphemes and their morphological feature descriptions) makes such an evaluation of an Arabic morphological analyzer impractical.

On the other hand, the SALMA – Gold Standard contains one correct analysis for each word suitable to its context. The evaluation of a morphological analyzer using the SALMA – Gold Standard, will check whether the correct analysis of the gold standard is among the possible analyses of the morphological analyzer. One analysis produced by the morphological analyzer that matches the correct word segmentation into morphemes and possibly the SALMA – Tags of each morpheme is selected. Then the tags for each morpheme of the selected analysis are compared with their equivalents in the gold standard. The percentage of the correct whole morpheme tags is computed and reported. In the following evaluation, scores are for the "best" analysis, chosen by hand from the set of possible analyses output by the SALMA – Tagger.

Accuracy, precision, recall and F-measure are applicable to measure the accuracy of the individual morphological categories of the morpheme tags. The computed accuracy metrics measure the capacity of a morphological analyzer to predict the detailed morphological features information encapsulated within the analyzed word. They also show the interdependency and the interrelationships between the different morphological categories of the morphemes. The next section discusses the evaluation of the SALMA – Tagger using the gold standard concentrating on the application of evaluation metrics to measure the accuracy of the individual morphological feature categories. Chapter 10 discusses the evaluation of the SALMA – Lemmatizer and Stemmer on the Qur'an and the Arabic Internet Corpus.

9.7 Evaluating the SALMA – Tagger Using Gold Standards

The focus in evaluating the outputs of the SALMA – Tagger is to evaluate the prediction accuracy of the 22 morphological feature categories of each morpheme using the SALMA – Gold Standard. Other intermediate outputs can be evaluated separately e.g. the evaluation of the SALMA – Lemmatizer and Stemmer; see section 10.2.

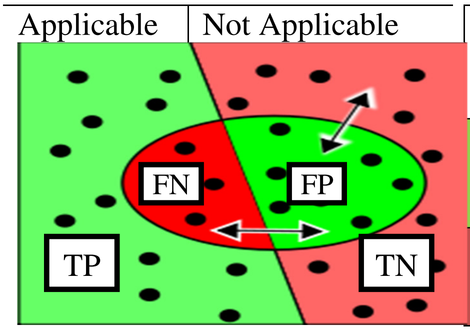
Therefore, for each word of the test samples (*i.e.* the Qur'an text sample and the CCA text sample) the analysis that best matches its equivalent in the SALMA – Gold Standard was chosen as a candidate analysis for evaluation. Then the evaluation metrics of accuracy, recall, precision and F-measure were computed. Two aspects for measuring the accuracy of the SALMA – Tagger were investigated:

- **Applicability:** equates to whether or not a value is entered at the expected position in the tag string.

- **Correctness:** equates to the correct value for a feature, mapped to the correct position in the tag string.

These aspects were used to define the elements of the confusion matrix. One advantage of a confusion matrix is counting and visualizing when the system is confusing two classes (*i.e.* commonly giving one tag as another). Another advantage of a confusion matrix is to compute the values of accuracy, recall, precision and f-measure of the SALMA – Tagger outputs. The confusion matrix elements are TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative), see figure 9.6. These elements were defined according to the observations of the outputs as follows:

- **TP (True Positive)**: True and applicable; the case was applicable and predicted correctly. Two conditions of applicability and correctness are needed to classify the prediction as TP. First, the morphological feature is applicable. Second, the prediction of the attribute value of that morphological feature is correctly predicted.
- **TN (True Negative)**: True and not applicable cases; the case was not applicable and predicted as not applicable.
- **FN (False Negative)**: False prediction of applicable cases; the case was applicable but predicted as not applicable.
- **FP (False Positive)**: False prediction of not applicable cases; the case was not applicable but predicted as applicable by giving a tag in the expected position.



Confusion Matrix	Predicted positive	Predicted Negative
Positive cases	TP	FN
Negative cases	FP	TN

Figure 9.6 The confusion matrix aspects and elements

The definition of the confusion matrix elements depends on two conditions: applicability and correctness. These conditions overlap in some cases where the positive cases are given a wrong tag other than "-". Using a confusion matrix, the analyses are classified into four categories but the observations made from analysing the output data distinguish between 5 categories:

1- Applicable case and predicted correctly, which represents the TP category. E.g. predicting the gender of a noun as singular '**s**' which matches the gender feature of the same noun in the gold standard, which is tagged as singular '**s**'.

2- Not applicable case and predicted not applicable, which represents the TN category. E.g. the morphological feature category of person is not a feature for proper nouns. Hence, proper nouns have "-" in the ninth position of their tags. A case is classified as TN, if the morphological analyzer predicts the value of the person feature as non-applicable and gives a "-" tag.

3- Applicable case and predicted not applicable tagged by "-", which can fit into the FN category. This case happens if the morphological analyzer gives a "-" tag for the morphological feature of number which is an applicable feature for proper nouns. The gold standard has a valid tag for the number feature of proper nouns either '***s***' (singular), '***d***' (dual), '***p***' (sound plural), '***b***' (broken plural).

4- Not applicable cases tagged in the gold standard by "-" and predicted as applicable, which can fit into the FP category. Theoretically, this case should not occur. However, some morphological features such as Inflectional Morphology, Case or Mood, and Case and Mood Marks depend on each other. Predicting the value of inflectional morphology for a perfect verb as '***d***' (conjugated) will affect the prediction of Case or Mood by giving a tag for a non-applicable morphological feature.

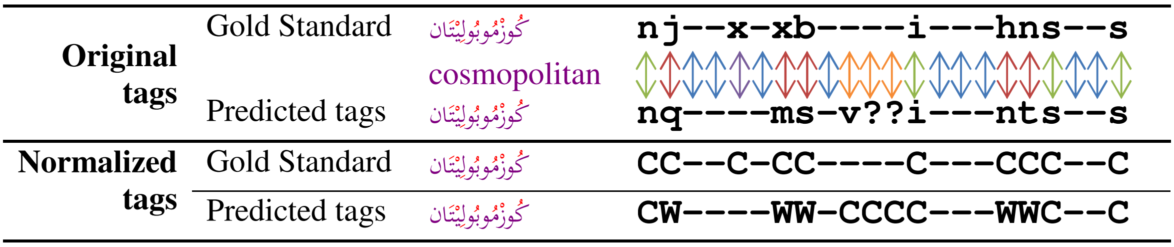
5- Applicable cases and predicted wrongly by tagging with a tag other than "-". E.g. predicting the number of a proper noun as singular by giving the tag '***s***' while that proper noun is broken plural which is tagged by '***b***' in the gold standard.

The last observation (O_5) can fit into the FP category because it is part of the positive predictions made by the analyzer, and the FN category because it is summed with the number of positive cases in the gold standard. According to the definition of precision and recall, see formula 9.5 and 9.6, the fifth observation will affect both the recall and the precision of the system.

However, the confusion matrix will only allow data to be classified into one of its four categories. An extended version of the confusion matrix where the data of the five observations fit only into one category was developed. The development of the extended confusion matrix required normalizing the tags of the gold standard and the outputs of the analyzer were normalized to three symbols ('**C**' (character), '**W'** (wrong), '-' (not applicable)). According to the above observations new tags for the gold standard and the outputs of the analyzer were generated by mapping the original tag into the three tags used for evaluation. These three evaluation tags are not shown in the outputs of the analyzer. They are only used to extend the confusion matrix that can fit 5 categories instead of the ordinary four categories. Figure 9.7 illustrates the mapping rules of the original tags into the three tags for evaluation depending on the above five observations. Figure 9.8 gives an example of the mapping process and the normalized evaluation tags for the word كُوزْمُوبُولِيتَان *kuzmūbūlītān* ‘cosmopolitan’ a borrowed word which represent a challenging example for the morphological analyzer to predict its morphological features. However, it is good example because it contains all the five observations and demonstrates the mapping process.

Observations		Original tags		Normalized tags	
		Gold	Predicted	Gold	Predicted
Applicable case and predicted correctly	O_1	**a**	**a**	**C**	**C**
Not applicable case and predicted not applicable	O_2	**-**	**-**	**-**	**-**
Applicable case and predicted not applicable	O_3	**b**	**-**	**C**	**-**
Not applicable cases and predicted as applicable	O_4	**-**	**c**	**-**	**C**
Applicable cases and predicted wrongly	O_5	**d**	**e**	**C**	**W**

Figure 9.7 Normalizing the gold standard and predicted tags into (-, C, W) evaluation tags



Original tags	Gold Standard	كُوزْمُوبُولِيتَان	`nj--x-xb----i---hns--s`
		cosmopolitan	
	Predicted tags	كُوزْمُوبُولِيتَان	`nq----ms-v??i---nts--s`
Normalized tags	Gold Standard	كُوزْمُوبُولِيتَان	`CC--C-CC----C---CCC--C`
	Predicted tags	كُوزْمُوبُولِيتَان	`CW----WW-CCCC---WWC--C`

Figure 9.8 Example of normalizing the gold standard and predicted tags into (-, C, W) evaluation tags

The new extended confusion matrix will contain three rows and three columns marked by (-, **C**, **W**). Then the confusion matrix is filled by the values by comparing the normalized tags. The 5 observations will fit directly in the confusion matrix. Figure 9.9 shows the skeleton of the confusion matrix and where the five observation values fit in the matrix. The five observations are marked by O_1-O_5 where the numbers 1-5 represent the observation number as listed above. The other entries in the confusion matrix are always zero marked by ‘**.**’ because the output tags of the analyzer cannot be classified into these entries. The figure shows the entries of the confusion matrix that are used to compute the values of the accuracy, precision and recall. Figures 9.10 and 9.11 show the confusion matrices generated for each morphological feature category of the morphemes SALMA – Tags.

Confusion Matrix

	–	C	W
–	<O$_2$>	O$_4$	.
C	O$_3$	<O$_1$>	O$_5$
W	.	.	<.>

(row = reference; col = test)

Entries used to compute 'Accuracy'

	–	C	W
–	<O$_2$>	O$_4$	.
C	O$_3$	<O$_1$>	O$_5$
W	.	.	<.>

(row = reference; col = test)

Entries used to compute 'Precision'

	–	C	W
–	<O$_2$>	O$_4$	.
C	O$_3$	<O$_1$>	O$_5$
W	.	.	<.>

(row = reference; col = test)

Entries used to compute 'Recall'

	–	C	W
–	<O$_2$>	O$_4$	.
C	O$_3$	<O$_1$>	O$_5$
W	.	.	<.>

(row = reference; col = test)

Figure 9.9 The confusion matrix and the entries used to compute the evaluation metrics

Using the extended confusion matrix, the values of the accuracy metrics were computed and reported. The first accuracy metric computed is Accuracy. The accuracy is defined as the percentage of correct predictions made for a certain morphological feature category. Formula 9.4 is used for the computation of accuracy.

$$\text{Accuarcy} = \frac{\text{TP+TN}}{\text{Total Number of morphemes}} = \frac{\text{O}_1 + \text{O}_2}{\text{Total Number of morphemes}} \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots(9.4)$$

Recall is defined as the percentage of applicable cases that are correctly predicted from the total number of actual positive cases in the gold standard. Formula 9.5 illustrates the computation of recall.

$$Recall = \frac{\text{Number of applicable cases correctly predicted}}{\text{Number of actual positive cases in the gold standard}} = \frac{\text{TP}}{\text{TP+FN}} = \frac{\text{O}_1}{\text{O}_1\text{+(O}_3\text{+ O}_5\text{)}} \;\ldots.(9.5)$$

Precision is defined as the percentage of applicable cases which are correctly predicted from the total number of positive predictions. Formula 9.6 illustrates the computation of precision.

$$\text{Precision} = \frac{\text{Number of applicable cases correctly predicted}}{\text{Total number of positive predictions}} = \frac{\text{TP}}{\text{TP+FP}} = \frac{\text{O}_1}{\text{O}_1\text{+(O}_4\text{+ O}_5\text{)}} \;\ldots\ldots\;(9.6)$$

F-measure (F$_1$ score) is the harmonic mean that combines precision and recall. It is interpreted as a weighted average of the precision and recall. F$_1$ score reaches its best value at 1 (100%) and worst score at 0 (0%). Formula 9.7 illustrates the computation of F$_1$ score.

$$F_1 \text{ score } = 2.\frac{\text{Precision .Recall}}{\text{Precision+Recall}} \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(9.7)$$

Results reported err on the side of caution by adding the number of cases of O$_5$ to both recall and precision equations.

(1) Main Part-of-Speech

```
  |    -     C     W |
--+-----------------+
- |   <.>    .     . |
C |     . <2170>   1 |
W |     .     .  <.>|
--+-----------------+
```

(2) Part-of-Speech: Noun

```
  |    -     C     W |
--+-----------------+
- |<1454>    1     . |
C |     .  <708>   8 |
W |     .     .  <.>|
--+-----------------+
```

(3) Part-of-Speech: Verb

```
  |    -     C     W |
--+-----------------+
- |<2057>    .     . |
C |     .  <112>   2 |
W |     .     .  <.>|
--+-----------------+
```

(4) Part-of-Speech: Particle

```
  |    -     C     W |
--+-----------------+
- |<1798>    .     . |
C |     1  <372>   . |
W |     .     .  <.>|
--+-----------------+
```

(5) Part-of-Speech: Other

```
  |    -     C     W |
--+-----------------+
- |<1301>    .     . |
C |     1  <861>   8 |
W |     .     .  <.>|
--+-----------------+
```

(6) Punctuation marks

```
  |    -     C     W |
--+-----------------+
- |<2072>    .     . |
C |     .   <93>   6 |
W |     .     .  <.>|
--+-----------------+
```

(7) Gender

```
  |    -     C     W |
--+-----------------+
- | <970>   10     . |
C |     .<1137>   54 |
W |     .     .  <.>|
--+-----------------+
```

(8) Number

```
  |    -     C     W |
--+-----------------+
- | <970>   10     . |
C |     .<1122>   69 |
W |     .     .  <.>|
--+-----------------+
```

(9) Person

```
  |    -     C     W |
--+-----------------+
- |<1940>    8     . |
C |     4  <177>  42 |
W |     .     .  <.>|
--+-----------------+
```

(10) Inflectional Morphology

```
  |    -     C     W |
--+-----------------+
- | <942>    9     . |
C |     1<1205>   14 |
W |     .     .  <.>|
--+-----------------+
```

(11) Case or Mood

```
  |    -     C     W |
--+-----------------+
- |<1457>   12     . |
C |     8  <602>  92 |
W |     .     .  <.>|
--+-----------------+
```

(12) Case and Mood Marks

```
  |   -      C     W |
--+-----------------+
- |<987>     9     . |
C |   1   <779>  395 |
W |   .      .   <.>|
--+-----------------+
```

(13) Definiteness

```
  |    -     C     W |
--+-----------------+
- |<1425>   18     . |
C |     .  <725>   3 |
W |     .     .  <.>|
--+-----------------+
```

(14) Voice

```
  |    -     C     W |
--+-----------------+
- |<2049>    8     . |
C |     .  <105>   9 |
W |     .     .  <.>|
--+-----------------+
```

(15) Emphasized and Non-emphasized

```
  |    -     C     W |
--+-----------------+
- |<2049>    8     . |
C |     .  <114>   . |
W |     .     .  <.>|
--+-----------------+
```

(16) Transitivity

```
  |    -     C     W |
--+-----------------+
- |<2049>    8     . |
C |     .  <114>   . |
W |     .     .  <.>|
--+-----------------+
```

(17) Rational

```
  |    -     C     W |
--+-----------------+
- |<1340>    5     . |
C |     .  <695> 131 |
W |     .     .  <.>|
--+-----------------+
```

(18) Declension and Conjugation

```
  |    -     C     W |
--+-----------------+
- |<1085>    1     . |
C |     1<1080>    4 |
W |     .     .  <.>|
--+-----------------+
```

(19) Unaugmented and Augmented

```
  |    -     C     W |
--+-----------------+
- |<1344>    8     . |
C |     3  <795>  21 |
W |     .     .  <.>|
--+-----------------+
```

(20) Number of Root Letters

```
  |    -     C     W |
--+-----------------+
- |<1398>    3     . |
C |     4  <765>   1 |
W |     .     .  <.>|
--+-----------------+
```

(21) Verb Root

```
  |    -     C     W |
--+-----------------+
- |<2058>    .     . |
C |     .  <112>   1 |
W |     .     .  <.>|
--+-----------------+
```

(22) Noun Finals

```
  |    -     C     W |
--+-----------------+
- |<1500>    6     . |
C |     .  <656>   9 |
W |     .     .  <.>|
--+-----------------+
```

For all confusion matrices in this figure
`(row = reference; col = test)`

Figure 9.10 Confusion matrices for the CCA test sample

(1) Main Part-of-Speech

```
  |     -     C     W |
--+-----------------+
- |    <.>    .     . |
C |    11<1903>   28 |
W |     .     .    <.>|
--+-----------------+
```

(2) Part-of-Speech: Noun

```
  |     -     C     W |
--+-----------------+
- |<1438>    2     . |
C |     2 <235> 265 |
W |     .     .    <.>|
--+-----------------+
```

(3) Part-of-Speech: Verb

```
  |     -     C     W |
--+-----------------+
- |<1681>    .     . |
C |     1 <260>    . |
W |     .     .    <.>|
--+-----------------+
```

(4) Part-of-Speech: Particle

```
  |     -     C     W |
--+-----------------+
- |<1422>    4     . |
C |     9 <447>   60 |
W |     .     .    <.>|
--+-----------------+
```

(5) Part-of-Speech: Other

```
  |     -     C     W |
--+-----------------+
- |<1270>    9     . |
C |    27 <573>   63 |
W |     .     .    <.>|
--+-----------------+
```

(6) Punctuation marks

```
  |     -     C     W |
--+-----------------+
- |<1942>    .     . |
C |     .   <.>    . |
W |     .     .    <.>|
--+-----------------+
```

(7) Gender

```
  |     -     C     W |
--+-----------------+
- |<769>    91     . |
C |   23  <960>   99 |
W |     .     .    <.>|
--+-----------------+
```

(8) Number

```
  |   -   C   W |
--+--------------+
- |<768> 91    . |
C |  23<768>292 |
W |   .   .  <.>|
--+--------------+
```

(9) Person

```
  |     -     C     W |
--+-----------------+
- |<1312>   47     . |
C |    21 <519>   43 |
W |     .     .    <.>|
--+-----------------+
```

(10) Inflectional Morphology

```
  |     -     C     W |
--+-----------------+
- | <522>   41     . |
C |    59<1196> 124 |
W |     .     .    <.>|
--+-----------------+
```

(11) Case or Mood

```
  |     -     C     W |
--+-----------------+
- |<1094> 370     . |
C |     2 <454>   22 |
W |     .     .    <.>|
--+-----------------+
```

(12) Case and Mood Marks

```
  |    -      C     W |
--+-----------------+
- |<533>    34     . |
C |  56   <909> 410 |
W |     .     .    <.>|
--+-----------------+
```

(13) Definiteness

```
  |     -     C     W |
--+-----------------+
- |<1435>   13     . |
C |     . <437>   57 |
W |     .     .    <.>|
--+-----------------+
```

(14) Voice

```
  |     -     C     W |
--+-----------------+
- |<1682>    .     . |
C |     . <233>   27 |
W |     .     .    <.>|
--+-----------------+
```

(15) Emphasized and Non-emphasized

```
  |     -     C     W |
--+-----------------+
- |<1682>    .     . |
C |     . <259>    1 |
W |     .     .    <.>|
--+-----------------+
```

(16) Transitivity

```
  |     -     C     W |
--+-----------------+
- |<1682>    2     . |
C |     . <254>    4 |
W |     .     .    <.>|
--+-----------------+
```

(17) Rational

```
  |     -     C     W |
--+-----------------+
- |<1175>    9     . |
C |     . <657> 101 |
W |     .     .    <.>|
--+-----------------+
```

(18) Declension and Conjugation

```
  |     -     C     W |
--+-----------------+
- |<1179>    2     . |
C |     1 <571> 189 |
W |     .     .    <.>|
--+-----------------+
```

(19) Unaugmented and Augmented

```
  |     -     C     W |
--+-----------------+
- |<1300>    5     . |
C |     8 <549>   80 |
W |     .     .    <.>|
--+-----------------+
```

(20) Number of Root Letters

```
  |     -     C     W |
--+-----------------+
- |<1298>    5     . |
C |     . <639>    . |
W |     .     .    <.>|
--+-----------------+
```

(21) Verb Root

```
  |     -      C    W |
--+-----------------+
- |<1687>     .    . |
C |     .   <255>  . |
W |     .      .  <.>|
--+-----------------+
```

(22) Noun Finals

```
  |     -     C     W |
--+-----------------+
- |<1440> 121     . |
C |     . <372>    9 |
W |     .     .    <.>|
--+-----------------+
```

For all confusion matrices in this figure
`(row = reference; col = test)`

Figure 9.11 Confusion matrices for the Qur'an – chapter 29 test sample

The SALMA – Tagger was evaluated using two samples of text documents: chapter 29 of the Qur'an and a sample from the CCA. The outputs of analysing the two samples were evaluated using the SALMA – Gold Standard. The confusion matrix of each morphological feature category was generated. Then the four accuracy metrics were computed. The confusion matrices of the morphological feature categories of the two test texts are shown in figures 9.10 and 9.11. The accuracy metrics are shown in tables 9.1 and 9.2. The figures of the evaluation metrics are shown in figures 9.12 and 9.13. The results are discussed in the next section 9.8.

Found P represents the positive predictions made by the SALMA – Tagger where it gave a tag other than '-' at the expected position. Actual P represents the positive cases in the gold standard. Found N represents the non-applicable predictions made by the SALMA – Tagger where it gave the tag '-'. Actual N represents the non-applicable cases in the gold standard tagged by '-'.

Table 9.1 Accuracy metrics for evaluating the CCA test sample

#	Category	Found (P)	Actual (P)	Found (N)	Actual (N)	Accuracy	Recall	Precision	F1-score
1	Main Part-of-Speech	2170	2171	0	0	99.95%	99.95%	99.95%	99.95%
2	Noun	708	717	1454	1455	99.59%	98.88%	98.74%	98.81%
3	Verb	112	114	2057	2057	99.91%	98.25%	98.25%	98.25%
4	Particle	372	372	1798	1798	99.95%	99.73%	100.00%	99.87%
5	Other	861	869	1301	1301	99.59%	98.97%	99.08%	99.02%
6	Punctuations	93	99	2072	2072	99.72%	93.94%	93.94%	93.94%
7	Gender	1137	1201	970	980	97.05%	95.47%	94.67%	95.07%
8	Number	1122	1201	970	980	96.36%	94.21%	93.42%	93.81%
9	Person	177	227	1940	1948	97.51%	79.37%	77.97%	78.67%
10	Inflectional Morphology	1205	1228	942	951	98.89%	98.77%	98.13%	98.45%
11	Case or Mood	602	706	1457	1469	94.84%	85.76%	85.27%	85.51%
12	Case and Mood Marks	779	1183	987	996	81.35%	66.30%	65.85%	66.07%
13	Definiteness	725	746	1425	1443	99.03%	99.59%	97.19%	98.37%
14	Voice	105	122	2049	2057	99.22%	92.11%	86.07%	88.98%
15	Emphasis	114	122	2049	2057	99.63%	100.00%	93.44%	96.61%
16	Transitivity	114	122	2049	2057	99.63%	100.00%	93.44%	96.61%
17	Rational	695	831	1340	1345	93.74%	84.14%	83.63%	83.89%
18	Declension and Conjugation	1080	1085	1085	1086	99.72%	99.54%	99.54%	99.54%
19	Unaugmented and Augmented	795	824	1344	1352	98.53%	97.07%	96.48%	96.77%
20	Number of Root Letters	765	769	1398	1401	99.63%	99.35%	99.48%	99.42%
21	Verb Root	112	113	2058	2058	99.95%	99.12%	99.12%	99.12%
22	Noun Finals	656	671	1500	1506	99.31%	98.65%	97.76%	98.20%

Table 9.2 Accuracy metrics for evaluating the Qur'an – Chapter 29 test sample

#	Category	Found (P)	Actual (P)	Found (N)	Actual (N)	Accuracy	Recall	Precision	F1-score
1	Main Part-of-Speech	1903	1931	0	0	97.99%	97.99%	98.55%	98.27%
2	Noun	235	502	1438	1440	86.15%	46.81%	46.81%	46.81%
3	Verb	260	260	1681	1681	99.95%	99.62%	100.00%	99.81%
4	Particle	447	511	1422	1426	96.24%	86.63%	87.48%	87.05%
5	Other	573	645	1270	1279	94.90%	86.43%	88.84%	87.61%
6	Punctuations	0	0	1942	1942	100.00%	0.00%	0.00%	0.00%
7	Gender	960	1150	769	860	89.03%	88.72%	83.48%	86.02%
8	Number	768	1151	768	859	79.09%	70.91%	66.72%	68.76%
9	Person	519	609	1312	1359	94.28%	89.02%	85.22%	87.08%
10	Inflectional Morphology	1196	1361	522	563	88.47%	86.73%	87.88%	87.30%
11	Case or Mood	454	846	1094	1464	79.71%	94.98%	53.66%	68.58%
12	Case and Mood Marks	909	1353	533	567	74.25%	66.11%	67.18%	66.64%
13	Definiteness	437	507	1435	1448	96.40%	88.46%	86.19%	87.31%
14	Voice	233	260	1682	1682	98.61%	89.62%	89.62%	89.62%
15	Emphasis	259	260	1682	1682	99.95%	99.62%	99.62%	99.62%
16	Transitivity	254	260	1682	1684	99.69%	98.45%	97.69%	98.07%
17	Rational	657	767	1175	1184	94.34%	86.68%	85.66%	86.16%
18	Declension and Conjugation	571	762	1179	1181	90.11%	75.03%	74.93%	74.98%
19	Unaugmented and Augmented	549	634	1300	1305	95.21%	86.19%	86.59%	86.39%
20	Number of Root Letters	639	644	1298	1303	99.74%	100.00%	99.22%	99.61%
21	Verb Root	255	255	1687	1687	100.00%	100.00%	100.00%	100.00%
22	Noun Finals	372	502	1440	1561	93.31%	97.64%	74.10%	84.26%

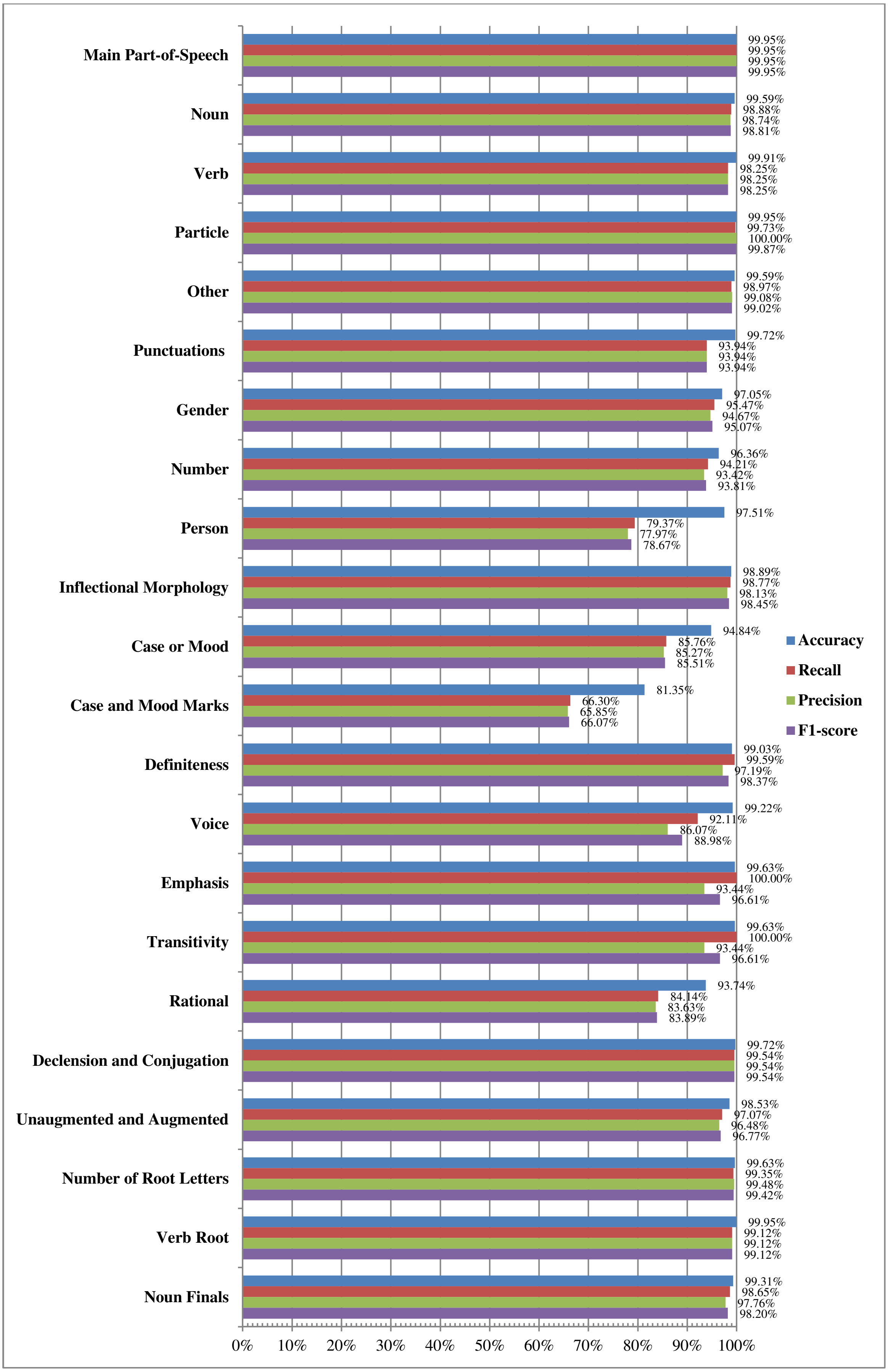


Figure 9.12 Accuracy metrics for evaluating the CCA test sample

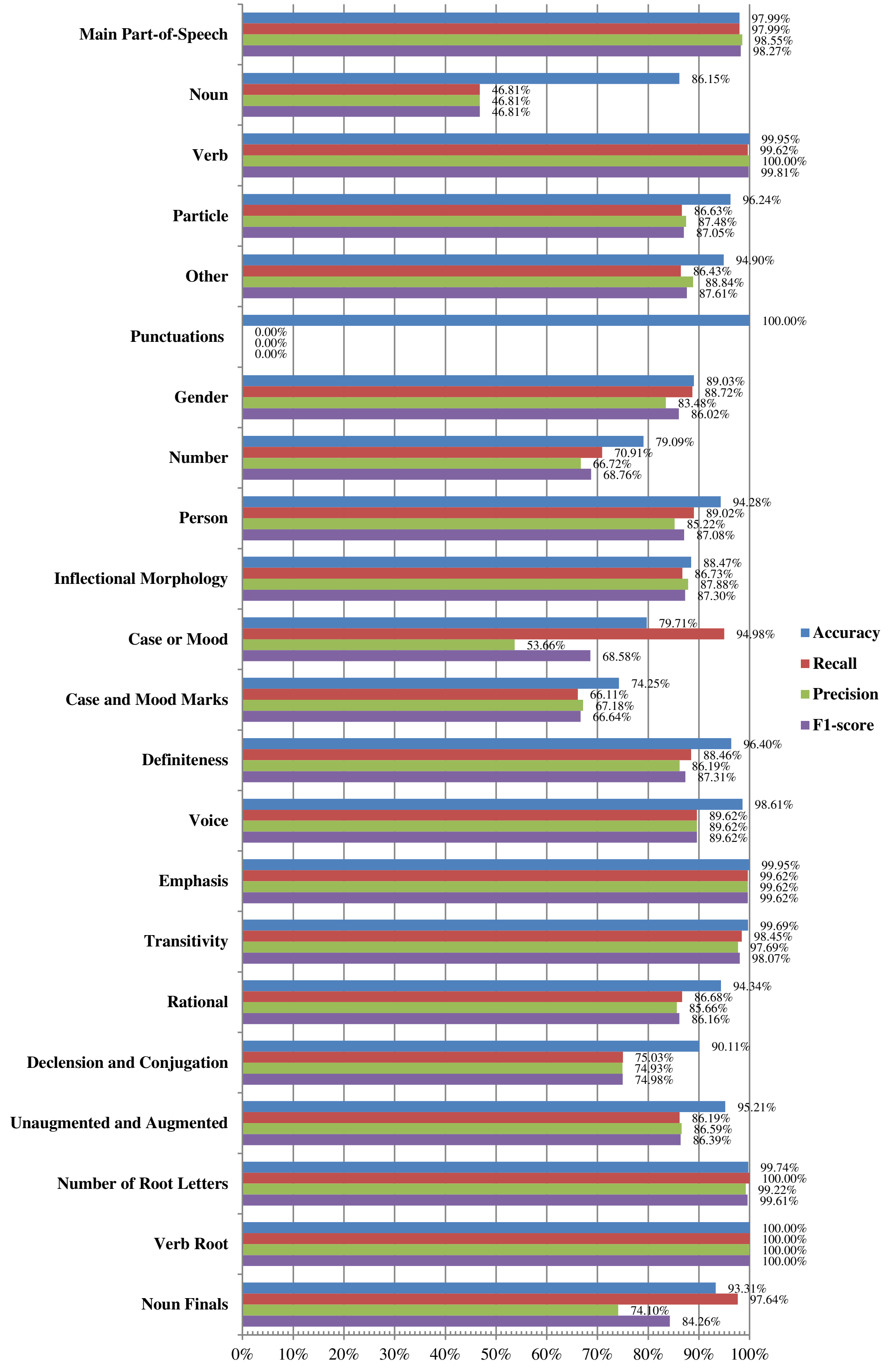


Figure 9.13 Accuracy metrics for evaluating the Qur'an – Chapter 29 test sample

9.8 Discussion of Results

The results of evaluating the SALMA – Tagger for two different text genres: the MSA text from the CCA and the Classical Arabic text from the Qur'an, showed the applicability of the SALMA – Tagger to process different types of text types, domains and genres of both vowelized and non-vowelized Arabic text. The SALMA – Tagger can be used to POS-tag Arabic text corpora and to provide detailed fine-grained analysis for each morpheme of the corpus words. The SALMA – Tagger divides the analyzed word into 5 parts (*i.e.* proclitics, prefixes, stem, suffixes and enclitics) and gives each part a detailed morphological feature tag (SALMA - Tag) or possibly multiple tags if the parts have multiple clitics or affixes. Each SALMA – Tag consists of 22 morphological feature categories that encode fine-grain morphological information about each morpheme of the analyzed words.

The evaluation of the SALMA – Tagger using MSA text showed better overall results than the evaluation using the Qur'an text. The measure of accuracy is "exact match". The exact match of the prediction of all 22 features for a morpheme whole tags for the CCA test sample is 71.21% and for the Qur'an – chapter 29 test sample is at 53.5%, but some of the errors were very minor such as replacing one '?' by '-'. This shows that the Qur'an text has a more complex morphological structure than the MSA text. These complex morphological structures need more future work that investigates the differences between the two genres.

As long as, there is no disambiguation facility of the SALMA – Tagger, and the best match analyses were selected manually for the purpose of evaluation. The achieved accuracy results of evaluation represent the highest accuracy scores that can be achieve by the SALMA – Tagger to predict the values of the morphological feature categories attributes. The accuracy scores for part of speech tagging system as surveyed in section 2.4.1 and reported by their developers, range from 91% for the AMT tagger by Alqrainy (2008) to 97% for the HMM part-of-speech tagger for Arabic developed by Al-Shamsi and Guessoum (2006). Errors of a disambiguation tool, that will be added to the SALMA – Tagger as future work, will decrease the overall accuracy results between 3% and 9%.

The focus of this evaluation is to show the applicability of the SALMA – Tagger in distinguishing the fine-grain morphological features of the Arabic text corpus words. The evaluation shows which morphological feature the SALMA – Tagger can distinguish. It also shows the accuracy rate for each morphological feature category. The purpose of this evaluation is to report for users who will use the SALMA – Tagger or parts of it on the SALMA – Tagger capability in distinguishing the fine-grain morphological features of the words. For instance, anaphora resolution applications can benefit from the morphological features of main part of speech, gender, number, person and rational outputs of the SALMA – Tagger to maintain agreement of these features between verbs and pronouns in sentences. Limitations, examples of hard cases and methods for improvements are discussed for each morphological feature category.

9.8.1 Results of Predicting the Value of Main Part of Speech

The results show high accuracy in predicting the main part of speech of the analyzed morphemes. 99.05% of the Qur'an sample morphemes and 97.99% of the CCA sample were correctly predicted. The prediction of the main part of speech of the morphemes depends on both: (i) maintaining agreement between the word's affixes and clitics where the clitics and affixes dictionaries contain the part-of-speech information that matches them, see section 8.3.1.5; and (ii) the patterns dictionaries where the main part of speech information is encoded within the SALMA – Tag given to each pattern; see section 8.3.3.1. The clitics and affixes dictionaries are used in the prediction of the main part of speech for all morphemes of the analyzed word, while the patterns dictionary is mainly used to predict the main part of speech of the stem morpheme.

9.8.2 Results of Predicting the Value of the Part-of-Speech Subcategory of Noun

The prediction of the part-of-speech subcategory of Noun scored an accuracy of 99.59% for the CCA text, while it scored a lower accuracy of 86.15% for the Qur'an test sample. The prediction of the part-of-speech subcategory of noun was not easy for the Qur'an text sample due to the nature of Quranic Arabic. The Qur'an text sample involves repeated use of old personal names such as فِرْعَوْنَ *fir'awn* 'firaun' and places such as ثَمُودَ *ṯamūd* 'thamud', while the list of the proper nouns used by the SALMA – Tagger was constructed from MSA newswire corpus; see section 8.3.2.4. The MSA text sample contains many relative nouns such as الثَّقَافِيِّ *aṯ-ṯaqāfī* 'cultural' and gerunds of profession such as الْوَطَنِيَّةَ *al-waṭaniyya*^{h} 'nationality', which are repeated frequently in the CCA text sample. These two types of repeated nouns are frequently used in MSA text. They are formed by adding the relative *yā'* and *tā' marbūta*^{h} as suffixes. Therefore, the rule for predicting these attributes is simple. The Qur’an sample does not contain any examples of these two noun types.

9.8.3 Results of Predicting the Value of the Part-of-Speech Subcategories of Verb and Particle

High accuracy for predicting the part-of-speech sub category of verbs was scored about 99.95% accuracy for both the Qur’an and the CCA text samples. The prediction of verbs depends on the analysis of the prefixes and suffixes and the matching of the stem morpheme with a patterns dictionary entry. High accuracy was scored for the part-of-speech subcategory of particle as well. An accuracy of 99.95% was scored for the CCA text sample and 96.24% for the Qur’an text sample. Most particles are stored in the function words list; see section 8.3.2.3. However, some particles in the Qur’an text sample are complex particles which consist of more than one morpheme such as أَوَلَمْ *’a-wa-lam* ‘and not’ which consists of three morphemes. Such complex particles need to be included in the function words list to improve the accuracy of the predicting particles.

9.8.4 Results of Predicting the Value of the Part-of-Speech Subcategory of Others (Residuals)

The accuracy of predicting the part-of-speech subcategory of others (residuals) scored 99.59% for the CCA test sample and 94.24% for the Qur’an test sample. The residuals are part of the clitics and affixes. The prediction of these affixes depends on matching the morphemes of the analyzed word with the entries of the clitics and affixes dictionaries. The errors made in the Qur’an sample are due to the use of ambiguous enclitics which can be classified into different categories such نَّ *nna* and نْ *n* which can be feminine suffixed pronoun or emphatic *nūn*. The CCA text sample contains numbers, currency and Arabized words which belong to the ‘others’ category but the SALMA – Tag Set does not include them yet. Section 9.10 (below) discusses the extension of the SALMA – Tag Set to include these attributes.

9.8.5 Results of Predicting the Value of Punctuations

The Qur’an test sample has no punctuation; therefore predicting that the punctuation category is not applicable for the analyzed words morphemes scored 100% accuracy. The CCA test sample contains punctuation. The accuracy of prediction was 99.72%. The prediction of punctuation is done in the tokenization step; see section 8.3.1. Special characters are used in the MSA text which cannot be classified as a word or a morpheme and not part of the standard punctuation described in section 6.2.6. These special characters such as ‘/’ slash are given a new tag ‘o’ which represents other punctuation marks.

9.8.6 Results of Predicting the Value of the Morphological Features of Gender, Number and Person

The prediction of the morphological features of gender, number and person scored 97.05%, 96.36% and 97.51% for the CCA test sample respectively, and 89.03%, 79.09%, 94.28% for the Qur’an test sample, respectively. The three morphological features are related to each other and share the same prediction methodology. Nouns have the morphological features of gender and number but not person, except for pronouns. Verbs have all three features. The prediction of the morphological features of gender and number for nouns depends on suffix analysis. Feminine and singular words have the suffix *ta’ marbuta*^{h}. Dual words are marked by ان *ān* or ين *ayn*. Masculine sound plural words have the suffix ون *wn* or ين *ayn*, while feminine sound plural words have the suffix ات *āt*. Broken plural words are searched in the broken plural list and the investigation of the gender feature is done on the retrieved singular form of the matched words. For example, the gender for أَنْحَاء *’anḥā’* “directions; regions” which is a broken plural of the singular نَاحِيَة *nāḥiya*^{t} “directions; regions”, is feminine because the singular feminine suffix *ta’ marbuta*^{h} appears on the singular form of the analyzed word. However, if the word is a broken and not found in the broken plural list, then the assigned tags ‘***ms-***’ (masculine, singular and not applicable) are wrong.

The prediction of the three morphological features for verbs depends on the combinations of prefixes and suffixed pronouns attached to the end of the verbs. Subject suffix-pronouns and genitive suffix-pronouns describe the reference person of the verb and agree with the number and gender of the doer of the verb; see section 8.4.1. False predictions of the morphological features of gender, number and person of verbs occur because some verbs are ambiguous. These verbs such as تَرْبِطُ *tarbiṭu* “you are tying / she is tying” can be masculine, singular and second person, or feminine, singular and third person. The SALMA – Tagger predicts/assigns the tags ‘***xs?***’ (of common gender, singular, applicable feature) to these kind of verbs. The difference comes by comparing against the gold standard where these features match the context of the words. These wrong predictions can be solved by applying contextual rules that define the agreement between the verb and its doer (the subject of the sentence). Contextual rules are also needed to disambiguate the number of verbs where singular verb forms have following plural subjects such as the phrase وَيُرَوِّجُ هَؤُلَاءِ *wa yurawwiğu hā’ulā’i* “and those who are spreading”, the verb يُرَوِّجُ *yurawwiğu* “spreading” is in singular form while the subject هَؤُلَاءِ *hā’ulā’I* “those” is a plural demonstrative pronoun.

9.8.7 Results of Predicting the Value of the Morphological Features of Inflectional Morphology, Case or Mood, and Case and Mood Marks

The prediction accuracy of the morphological features of inflectional morphology, case or mood, and case and mood marks scored 98.89%, 94.84% and 81.35% for the CCA test sample and 88.47%, 74.71% and 74.25% for the Qur’an test sample respectively. The prediction of morphological feature of inflectional morphology for verbs depends on the part-of-speech subcategory of verbs and analysis of suffixes for imperfect verbs to determine whether the verb is conjugated or invariable. The disambiguation of nouns into declined or invariable depends on applying many rules that deal with the part-of-speech subcategory of nouns, noun finals and patterns. These rules classify the declined nouns into fully declined or non-declined. The prediction of the morphological feature of case and mood depends on the result of the prediction of the morphological feature of inflectional morphology, where a declined noun has case (*i.e.* nominative, accusative or genitive) and a conjugated verb has mood (*i.e.* indicative, subjunctive, or imperative/jussive), while case and mood are not applicable to invariable nouns and verbs. The prediction of a noun’s case investigates the proclitics attached to the beginning of the noun which might affect the case and its syntactic mark such as prepositions and jurative particles. Prediction rules also investigate the dual and plural suffixes which change according to the case of the noun. For example, ون *wn* is a masculine plural suffix of nominative case, while ين *ayn* is a masculine plural or dual suffix of accusative or genitive case. The five nouns أَبٌ *’ab*^{un} ‘father’, أَخٌ *’aẖ*^{un} ‘brother’, حَمٌ *ḥam*^{un} ‘father-in-law’, فُو *fū* (فَم *fam*) ‘mouth’, and ذُو *ḏū* ‘possessor; owner’ change their suffix according to the context, the suffix و *waw* indicates nominative case, ا *’alif* indicates accusative case and ي *yā’* indicates genitive case. Rules for predicting the case or mood, and case and mood marks for singular and broken plural nouns depend on the short vowel (*i.e.* the syntactic mark) that appears on the end of the word. The absence of short vowels and the contextual rules that deal with the nouns according to their context (*i.e.* subject or object) increases the potential of wrong prediction especially for singular and broken plural nouns. Moreover, determining the morpheme that carries the syntactic mark of the word is not an easy task. For example the word بِأَجْنِحَتِهِ *bi-’ağniḥati*^{hi} ‘by its wings’ has four morphemes: preposition بِ *bi*, stem morpheme أَجْنِحَ *’ağniḥa*, feminine suffix تِ *ti*, and the suffixed pronoun هِ *hi*. The case mark, which is always considered by traditional Arabic grammar to be at the end of the word, is carried by the third morpheme the feminine suffix تِ *ti* in this example, rather than the final morpheme the suffixed pronoun هِ *hi*.

The prediction of the morphological features of case or mood, and case and mood marks for verbs depends on the previous prediction made for the morphological feature of inflectional morphology that classifies verbs into conjugated or invariable. Only a conjugated verb has mood. The prediction rules for mood depend on the part-of-speech subcategory of verb where mood is applicable to imperfect verbs and not applicable to perfect and imperative verbs. The rules also analyze the suffixes of the imperfect verb to determine the applicability of mood. Imperfect verbs that contain the third person feminine suffix pronoun ن *nūn* are invariable verbs which are marked by *sukūn* such as يَكْتُبْنَ *yaktubna* ‘they (*fem.*) write’. Those containing the emphatic *nūn* suffix are invariable verbs which are marked by *fatḥa*^{h} such as فَلَيَعْلَمَنَّ *falaya‘lamanna* ‘and *allā*^{h} will surely make evident’. The final rule of prediction depends on the short vowel which appears on the morpheme that carries the mood mark, where *ḍamma*^{h} indicates indicative mood, *fatḥa*^{h} indicates subjunctive mood, and *sukūn* indicates imperative or jussive mood. The absence of short vowels and the contextual rules that deal with nouns according to their context (*i.e.* subject or object) increases the potential for wrong prediction especially for subjunctive, and imperative or jussive verbs which are always preceded by subjunctive-governing particles and jussive-governing particles respectively.

The results show the interdependency of these three morphological feature categories. The morphological feature category of case and mood marks depends on both case or mood, and inflectional morphology. Case or mood depends on inflectional morphology. The prediction errors for inflectional morphology are propagated to the case or mood category, and then to case and mood markers. Therefore, accuracy rates were decreased in the direction of error propagation.

9.8.8 Results of Predicting the Value of the Morphological Feature of Definiteness

The accuracy of predicting the morphological feature of definiteness was high at 99.03% and 96.40% for the CCA test sample and the Qur'an test sample respectively. The prediction of the morphological feature of definiteness depends on the availability of the definite article ال as a proclitic for the analyzed noun. If the noun contains the definite article in its proclitics then the noun is definite; otherwise it is an indefinite noun. The morphological feature of definiteness is not applicable to verbs. Errors in classifying the word into noun or verb will be propagated to this category especially for indefinite prediction.

9.8.9 Results of Predicting the Value of the Morphological Feature of Voice

The prediction of the morphological feature of voice achieved a high accuracy score of 99.22% and 98.61% for the CCA test sample and the Qur'an test sample respectively. The morphological feature of voice is only applicable to verbs. The prediction rules classify verbs into active verbs or passive verbs depending on the short vowel appearing on the first letter of the verb after removing proclitics. If a *fatḥa*^{h} appears on the verb's first letter, then it is classified as an active voice verb. If *ḍamma*^{h} appears on the verb's first letter, then it is classified as a passive voice verb. Errors can happen in some cases where *ḍamma*^{h} appears on the first letter of active voice verbs such as يُرِيْدُونَ *yurīdūna* 'they want' which matches the pattern يُفْعِلُونَ *yuf'ilūn*. The passive verb form of this example is يُرَادُونَ *yurādūna* 'they are wanted to be' which matches the pattern يُفْعَلُونَ *yuf'alūn*. The difference between the two patterns is the short vowel that appears on the second root radical. The short vowel on the second root radical is *kasra*^{h} for active voice and *fatḥa*^{h} for all verbs generated from these patterns. The patterns dictionary used by the SALMA – Tagger distinguishes between active voice and passive voice patterns. Applying prediction rules for the morphological feature of voice that depend on patterns rather than the short vowel of the first letter of the verb will increase the prediction accuracy.

9.8.10 Results of Predicting the Value of the Morphological Feature of Emphasized and Non-Emphasized

The prediction accuracy of the morphological feature of emphasized and non-emphasized was high at 99.63% and 99.95% for the CCA test sample and the Qur'an test sample respectively. The morphological feature of emphasized and non-emphasized is applicable only to verbs. Prediction rules for classifying verbs into emphasized or non-emphasized depends on the part-of-speech subcategory of the verb. Perfect verbs are always non-emphasized while imperfect and imperative verbs can be emphasized. The prediction rules also investigate the suffixes of the verb. Emphasized verbs contain the emphatic *nūn* as a suffix.

9.8.11 Results of Predicting the Value of the Morphological Feature of Transitivity

The prediction accuracy of the morphological feature of transitivity was high at 99.63% and 99.69% for the CCA test sample and the Qur'an test sample respectively. The morphological feature of transitivity is applicable only to verbs. The prediction rules of the morphological feature of transitivity classify verbs into: intransitive verbs which complete their meaning without the need for an object; singly transitive verbs which need one object to complete their meaning; doubly transitive verbs, which need two objects to complete their meaning; or triply transitive verbs, which need three objects to complete their meaning. The prediction rules of the morphological feature of transitivity depend on matching the analyzed verb with one verb stored in the lists of doubly transitive and triply transitive verbs. The singly transitive verb attribute is the default value of the morphological feature of transitivity. The absence of contextual rules for predicting the attributes of the morphological feature of transitivity increases the potential for making prediction mistakes. On the other hand, suffix pronouns analysis can capture some attributes of this morphological feature.

9.8.12 Results of Predicting the Value of the Morphological Feature of Rational

The prediction of the morphological feature of rational scored an accuracy of 93.74% for the CCA test sample and an accuracy of 94.34% for the Qur'an test sample. The morphological feature of rational is applicable to both nouns and verbs. The rationality of the subject (or the doer) of the verb determines the rationality attribute of the analyzed verb. The prediction rules for the morphological feature of rational assign default values to the analyzed words depending on their part-of-speech subcategory; see section 8.4.2. Proper nouns are classified as rational if the proper noun is found in the personal proper nouns list, and as irrational if they are found in the locations or organizations proper nouns lists. Demonstrative pronouns are classified according their use as rational or irrational. Qur'an verbs are assigned a default value of rational as most of the Qur'an verbs represent dialogue between God and people. Classifying words into rational or irrational depends on the semantics of the word itself and its context, such that agreement is maintained between sentence parts such as verb-subject agreement and adjective-descriptive noun agreement. A comprehensive dictionary which includes Rational information for each dictionary entry is needed to determine the correct attribute value of rational for nouns.

9.8.13 Results of Predicting the Value of the Morphological Feature of Declension and Conjugation

The prediction of the morphological feature of declension and conjugation was highly accurate at 99.72% for the CCA test sample and slightly less accurate at 90.11% for the Qur'an test sample. The morphological feature of declension and conjugation is applicable to nouns, verbs and particles. The prediction rules of the values of declension and conjugation of nouns depend on the part-of-speech subcategories. The rules for predicting the values of declension and conjugation of verbs depend on searching four lists of verbs: the non-conjugated/restricted-to-the-perfect verb list; the non-conjugated/restricted-to-the-imperfect verb list; the non-conjugated/restricted-to-the-imperative verb list; and the partially conjugated verb list. The default value of the morphological feature of declension and conjugation for verbs is fully conjugated verb. Including the declension and conjugation information in the Arabic dictionary will increase the correct prediction of attributes for this morphological feature.

9.8.14 Results of Predicting the Value of the Morphological Features of Unaugmented and Augmented, Number of Root Letters, and Verb Roots

The prediction accuracy of the morphological features of unaugmented and augmented, number of root letters, and verb roots was 98.53%, 99.63% and 99.95% for the CCA test sample and 95.21%, 99.74% and 100% for the Qur'an test sample respectively. The morphological features of unaugmented and augmented, and number of root letters are applicable to both nouns and verbs, while the morphological feature of verb roots only applies to verbs. The rules for predicting the three morphological features mainly depend on the root of the analyzed word. The prediction rule of unaugmented and augmented attributes subtracts the length of the root from the length of the analyzed word. The prediction rule of the attributes of the number of root letters depends on the length of the root. The prediction rules of the morphological feature of verb roots depend on the nature of the root letters - whether they are consonants, containing *hamza*^{h}, or containing one or two vowels. The prediction errors are higher for the morphological feature of unaugmented and augmented due to the ambiguous word boundaries. In some cases of non-vowelized text *tanwīn fatiḥ* (اً) appears as *'alif* which will be counted as an augmented letter. In other cases, vowels might be deleted from the word. Therefore, the rules for counting the added letters to the word need to know whether a vowel is deleted or not. For example, the verb يَجِدُ *yağidu* 'he finds' has the root و-ج-د *w-ğ-d* and is augmented by one letter يَ *yā'* representing the imperfect prefix. The first root letter و *wāw* is a vowel and is deleted from the word.

9.8.15 Results of Predicting the Value of the Morphological Feature of Noun Finals

The prediction of the morphological feature of noun finals was a highly accurate at 99.31% for the CCA test sample and slightly lower at 93.31% for the Qur'an test sample. The rules for predicting the value of the morphological feature of Noun Finals mainly depend on the long stem and the root of the analyzed word. The rules check the final letters of the long stem against a set of conditions that classify nouns into 6 categories; see section 8.4.3. Knowing the value of the Noun Finals feature helps in specifying other features such as the morphological features of Inflectional Morphology and Case and Mood Marks. Case marks cannot appear on the last letter of nouns with shortened ending, and only *fatḥa*^{h}, the mark of accusative case, appears on the last letter of nouns with curtailed ending.

9.8.16 More Conclusions

In conclusion, the SALMA – Tagger was evaluated on two text samples from different genres: chapter 29 of the Qur'an representing classical Arabic, and a sample from the CCA represents Modern Standard Arabic. The focus of this evaluation was to report on the applicability of the SALMA – Tagger in distinguishing the fine-grained morphological features of the Arabic text corpus, by measuring the accuracy of each of the 22 morphological feature categories represented by the SALMA – Tag for each morpheme in the two samples. The evaluation used the SALMA – Gold Standard. One advantage of carrying out this type of evaluation is to report for users who will use/reuse the SALMA – Tagger or parts of it the accuracy of predicting the attributes of the fine-grained morphological features. Users can depend on this evaluation to decide which parts of the SALMA – Tagger can be used directly. Another advantage directly addresses our interest in developing an Arabic morphological analyzer that is able to analyze Arabic text corpora by providing fine-grain analysis for each word. Fine-grain analysis of the Arabic word involves dividing the word into five parts and giving each part a detailed morphological features tag or possibly multiple tags if the part has multiple clitics or affixes.

The prediction accuracy was high for 15 morphological features: the morphological features of main part-of-speech; part-of-speech subcategory of verb; part-of-speech subcategory of particle; part-of-speech subcategory of other (residual); part-of-speech subcategory of punctuation; morphological feature of definiteness; morphological feature of voice; morphological feature of emphasized and non-emphasized; morphological feature of transitivity; morphological feature of declension and conjugation; morphological feature of unaugmented and augmented; morphological feature of number of root letters; morphological feature of verb roots; and morphological feature of noun finals. The accuracy for predicting the attributes of these 15 morphological features was between 98.53% and 100% for the CCA test sample and 90.11% and 100%for the Qur'an test sample. The morphological features of part-of-speech subcategory of noun, gender, number, person, inflectional morphology, case or mood, case and mood marks, and rational, scored slightly lower accuracy of prediction at 81.35% - 97.51%for the CCA test sample and 74.25% - 89.03%for the Qur'an test sample.

The next section (9.9) discusses the limitations, and the factors that affected the prediction accuracy of the morphological features, and suggests solutions that might improve this accuracy.

9.9 Limitations and improvements

The SALMA – Tagger achieved high prediction accuracy for 15 morphological features, and lower accuracy for 7 morphological features. The high prediction accuracy was due to the factors of the detailed analysis of words into morpheme and classifying these morphemes into distinctive classes that helped in predicting the attributes of these morphological feature categories. The reuse of the predicted attributes of some categories helped in predicting the correct attribute value of other categories. Providing the SALMA – Tagger with lists of (function words, broken plurals, named entities, doubly transitive verbs and triply transitive verbs, and conjugated and non-conjugated verbs) was the basis for predicting the attributes of many morphological feature categories. The SALMA – ABCLexicon is mainly used to extract the correct root of the analyzed words. The root information represents the basis for predicting the correct attribute of some morphological features. Finally, the patterns dictionary and the pattern matching algorithms were used in the prediction rules of most of the morphological feature categories.

The lower accuracy achieved with the other 7 morphological feature categories was due to an absence of contextual rules in the SALMA – Tagger, such that it treats words out of their context. The absence of short vowels on text especially for MSA text makes the prediction of the attributes of some morphological features difficult. Moreover, the interdependency between some morphological features such as the morphological features of inflectional morphology, case or mood, and case or mood marks decreases the accuracy of the dependent features by propagating errors from one feature to another. Finally, prediction errors increase, if the number of attributes of a certain morphological feature increases.

To improve the accuracy of predicting the attributes of the morphological feature categories, contextual rules can be implemented as a second pass. The contextual rules will also help in reducing the number of candidate analyses of the analyzed words by excluding those analyses that do not satisfy certain contextual rules. Some morphological feature categories such as rational depend on the semantic nature of the analyzed word itself. Providing rationality information for Arabic dictionary entries and reusing this information in morphological analyzers will increase the accuracy of prediction. Moreover, updating the dictionaries which are used by the SALMA – Tagger by increasing their coverage will increase the prediction accuracy.

9.10 Extension of the SALMA – Tag Set

The SALMA – Tag Set is a general-purpose fine-grain tag set. The aim of developing this tag set is that it should be used as the standard for part-of-speech tagging software to annotate corpora with more detailed morphological information for each word. The SALMA – Tag Set was evaluated by applying it to two text samples of different genres: chapter 29 of the Qur'an representing classical Arabic, and a sample of the CCA representing modern standard Arabic. Both samples and their annotations were used in the SALMA – Gold Standard.

The application of the SALMA – Tag Set to the Qur'an text sample did not introduce any reason for extending the tag set. However, the CCA text sample introduced some examples of tokens that appear in MSA text. These examples include numbers (digits), currency, non-Arabic words, borrowed (foreign) words, dates and special characters.

Extensions of the SALMA – Tag Set were made to two morphological feature categories: others (residual) and punctuation. The morphological feature of others (residuals) was extended to include new attributes for numbers (digits), currency, non-Arabic words, borrowed (foreign) words and dates. Table 9.3 shows the new attributes added to the part-of-speech subcategory of others (residuals). The part-of-speech subcategory of punctuation marks was extended by adding an attribute for special characters that are used as punctuation marks. These special characters appear on the MSA text due to the use of word-editing software that enables typing of special characters within text easily, and because of the lack of knowledge about using standard punctuation in Arabic text. Table 9.4 shows the attribute added to the part-of-speech subcategory of punctuation marks.

Borrowed (foreign) words are words borrowed from other languages which have become part of the language because they have become used widely by Arabic speakers. They also appear in text in transliteration format using Arabic letters. These words are used within the sentence like normal Arabic words. They accept inflectional affixes and change their form according to the context. Therefore, the SALMA – Tag Set treats them as Arabic words by classifying them within the main part-of-speech category attributes and assigning the morphological feature attributes that are applicable to them. They are given the tag '***x***' in the fifth position of the tag string to distinguish them as borrowed (foreign) words. Figure 9.14 shows an example of tagging a borrowed (foreign) word.

Table 9.3 Extended attributes of the Part-of-speech subcategories of Other (Residuals) and their tags at position 5

Position	Feature Name				Tag
5	**Part-of-Speech: Other** أقسام الكلام الفرعيَّة (أُخْرَى) *’aqsām al-kalām al-far’iyya^{t} (’uḫrā)*				
	Number (digits)	رَقَم	*raqam*	(+325461) (-897,653) (0.986) (13x10^{-3}) (-1.2E2) (1.2e-2)	**g**
	Currency	عُمْلَة	*‘umla^{t}*	(1,500د.أ) (2,927ر.س) ($250) (£430)	**c**
	Date	تَارِيخ	*tārīḫ*	(27/09/2011) (2011 أيلول 27) (27.09.11) (11 سبتمبر 27)	**e**
	Non-Arabic word	كَلِمَة غَيْر عَرَبِيَّة	*kalima^{t} ġayr ‘arabiyya^{h}*	windows, photoshop, games, download	**w**
	Borrowed (foreign) word	كَلِمَة مُعَرَّبَة	*kalima^{t} mu‘arraba^{h}*	كُوزْمُوبُولِيْتَان *kuzmūbūlītān* ‘cosmopolitan’ ستَاد *stād* ‘stadium’	**x**

Table 9.4 Extended attributes of the Part-of-speech subcategories of Punctuation Marks and their tags at position 6

Position	Feature Name				Tag
6	**Punctuation Marks** أقسام الكلام الفرعية (علامات الترقيم) *’aqsām al-kalām al-far’iyya^{t} (‘alāmāt at-tarqīm)*				
	Other punctuations	عَلامَات أُخْرَى	*‘alāmāt ’uḫrā*	/	**o**

Word	**SALMA – Tag**
كُوزْمُوبُولِيْتَان *kuzmūbūlītān* ‘cosmopolitan’	`nj--x-xb----i---hns--s`

Figure 9.14 Example of tagging a borrowed (foreign) word

9.11 Chapter Summary

This chapter discussed the evaluation of the SALMA – Tagger. The evaluation methodologies for morphological analyzers are not standardized yet. The first part of the chapter discussed the development of agreed standards for evaluating morphological analyzers for Arabic text, based on our experiences and participation in two community-based evaluation contests: the ALECSO/KACST initiative for developing and evaluating morphological analyzers, and the MorphoChallenge 2009 competition. The guideline recommendations, evaluation specifications and procedures, and evaluation metrics were reused to generate a global standard for evaluating morphological analyzers for Arabic text. The developed standards were applied for evaluating the SALMA – Tagger.

The developed evaluation standards depend on using gold standards for evaluating morphological analyzers for Arabic text. A reusable general purpose gold standard (the SALMA – Gold Standard) was constructed to evaluate various morphological analyzers for Arabic text and to allow comparisons between the different analyzers. The SALMA – Gold Standard is adherent to standards, and enriched with fine-grained morphological information for each morpheme of the gold standard text samples. The detailed information is: the input word, its root, lemma, pattern, word type and the word's morphemes. For each of the word's morphemes, the morpheme type is classified into proclitic, prefix, stem, suffix and enclitic, and a fine-grain SALMA – Tag which encodes 22 morphological feature categories of each morpheme, was included.

The SALMA – Gold Standard contains two text samples of about 1000-words each representing two different text domains and genres of both vowelized and non-vowelized text taken from the Qur'an – chapter 29 representing Classical Arabic, and from the CCA representing Modern Standard Arabic. The SALMA – Gold Standard is stored using different standard formats to allow wider reusability. XML technology allows storage of the gold standard in a machine-readable structured format. Tab-separated column files are widely used by researchers. They are used to store the gold standard following the Morphochallenge 2009 recommendations for constructing gold standards. Other formats are used to display the information of the gold standard for end users. These formats include HTML files and the visual display of the gold standard in colour-coded format.

The SALMA – Gold Standard was used to evaluate the SALMA – Tagger. The evaluation focused on measuring the prediction accuracy of the 22 morphological features encoded in the SALMA – Tags for each of the gold standard's text sample morphemes. The results show that 53.50% of the Qur'an text sample morphemes and 71.21% of the CCA text sample were correctly tagged using "exact match" of the gold standard's morpheme tags, but some of the errors were very minor such as replacing '**?**' by '-'.

The evaluation reported accuracy, recall, precision, f1-score and the confusion matrix for each morphological feature category. The individual category accuracy results are useful for users who will use/reuse the SALMA – Tagger or parts of it, to know in advance the prediction accuracy of the attributes of each morphological feature category. Accuracy scores are high for 15 morphological feature categories at about 98.53%-100% for the CCA test sample and 90.11% -100% for the Qur'an test sample. These categories are: the morphological feature of main part-of-speech; part-of-speech subcategory of verb; part-of-speech subcategory of particle; part-of-speech subcategory of other (residual); part-of-speech subcategory of punctuation; definiteness; voice; emphasized and non-emphasized; transitivity; declension and conjugation; unaugmented and augmented; number of root letters; verb roots; and noun finals.

The other 7 morphological feature categories: part-of-speech subcategory of noun; gender; number; person; inflectional morphology; case or mood; case and mood marks; and rational, were less accurately predicted: 81.35% - 97.51% for the CCA test sample and 74.25%-89.03% for the Qur'an test sample.

The absence of contextual rules, the absence of short vowels, the interdependency between some morphological features, and the number of attributes of a certain morphological category increase the potential for prediction errors of some morphological feature categories. To improve the accuracy of predicting the attributes of the morphological feature categories, contextual rules can be implemented as a second pass. Some morphological feature categories such as rational depend on the semantic nature of the analyzed word itself. Providing rationality information for Arabic dictionary entries and reusing this information in morphological analyzers will increase the accuracy of prediction. Moreover, updating the dictionaries which are used by the SALMA – Tagger by increasing their coverage will increase the prediction accuracy.

The SALMA – Gold Standard for evaluating Arabic morphological analyzers is an open-source resource that is available to download, for reuse in evaluation of other Arabic morphological analyzers.

Chapter 10
Practical Applications of the SALMA – Tagger

This chapter is based on the following sections of published papers:

Section 2 is based on section 4 in (Sawalha and Atwell 2010b) and section 1 in (Sawalha and Atwell 2011a)

Section 3 is based on section 1 in (Sawalha and Atwell 2011b)

Chapter Summary

The SALMA Tagger has been used in two important applications of Arabic text analytics: first, lemmatizing the 176-million words Arabic Internet Corpus, and second, as corpus linguistic resources and tools for Arabic lexicography. This chapter will illustrate how the tools- the SALMA – Tagger and SALMA – Lemmatizer and Stemmer, the resources - the SALMA – ABCLexicon and the Corpus of Traditional Arabic Lexicons, and the proposed standards - the SALMA – Tag Set - have been useful tools, resources and standards to advance Arabic computational linguistic technologies.

10.1 Introduction

In this research, resources (the SALMA – ABCLexicon, Chapter 4), Standards (the SALMA – Tag Set, Chapters 5, 6 and 7), and tools (the SALMA – Tagger, Chapters 8 and 9) were developed and evaluated. The main purpose in developing the resources, standards and tools is for morphosyntactic annotation of Arabic text with fine-grain morphosyntactic information. This chapter will investigate two applications of these resources, standards and tools: lemmatizing the 176-million word Arabic Internet Corpus[66] (AIC) (Sawalha and Atwell 2011a), and as language engineering resources to construct the Arabic dictionary (Sawalha and Atwell 2011b).

The resources, standards and tools were evaluated on samples of Arabic text to measure their accuracy and applicability to text analytics tasks. However, the performance aspects of the SALMA – Tagger such as speed, memory and ability to perform the desired analysis tasks were not evaluated previously. Applying the SALMA – Lemmatizer and Stemmer to lemmatize the 176-million word Arabic Internet Corpus is a practical application through which to evaluate performance and investigate the challenges of applying the resources, standards and tools on real, large-scale data.

The second application is a proposal about how these resources, standards and tools can be used as a language engineering toolkit for Arabic lexicography. This study reviews the resources and tools which are used in modern lexicography, and shows that the developed resources, and standards constitute a toolkit for constructing Arabic bi-lingual and monolingual dictionaries.

Section 10.2 discusses the application of lemmatizing the 176-million word AIC. Section 10.3 discusses the resources and tools for Arabic lexicography.

10.2 Lemmatizing the 176-million words Arabic Internet Corpus

The Arabic Internet Corpus is one of several large corpora collected for Translation Studies research at http://corpus.leeds.ac.uk/internet.html alongside Internet corpora for English, Chinese, French, German, Greek, Italian, Japanese, Polish, Portuguese, Russian and Spanish (Sharoff 2006). The Arabic Internet Corpus consists of about 176 million words[67]. Initially it consisted of raw text, with no further processing such as lemmatization or part-of-speech tagging. This section shows how the lemma and root were added for each word of the AIC.

[66] Querying Arabic Corpora http://smlc09.leeds.ac.uk/query-ar.html

[67] The frequency list of the Arabic Internet Corpus http://corpus.leeds.ac.uk/frqc/i-ar-forms.num

Arabic is a morphologically rich and highly inflectional language. Hundreds of words can be derived from the same root; and a lemma can appear in the text in many different forms due to the glutination of clitics at the front and end of the word. Therefore, lemmatization and root extraction is necessary for search applications, to enable inflected forms of a word to be grouped together. We used the lemmatizing part of the SALMA – Tagger (see section 8.3.2) to annotate the Arabic Internet Corpus words at two levels; the lemma and the root, as shown in Figure 10.1. The SALMA – Lemmatizer and Stemmer is relatively slow. In initial tests it processed 7 words per second, because it deals with orthographic issues, spell checking of the word's letters, short vowels and diacritics and the large dictionaries provided to perform its task. The estimated execution time for lemmatizing the full Arabic Internet Corpus was roughly 300 days using an ordinary uni-processor machine.

To reduce the processing time of the whole task, we used the power of HPC (High Performance Computing). NGS[68] (National Grid Services) aims to enable coherent electronic access for UK researchers to all computational and data-based resources and facilities required to carry out their research, independent of resource or researcher location. The huge computational power of NGS was used to lemmatize the Arabic internet corpus. As a result, a massive reduction in execution time was gained.

The Arabic Internet Corpus was divided the into half-million-word files. Then a specialized program distributed copies of the SALMA – Lemmatizer and Stemmer to multiple CPUs and assigned different input files to run the lemmatizer for the partitioned corpus files in parallel. The output files were combined in one lemmatized Arabic Internet Corpus, comprising 176 million word-tokens, 2,412,983 word-types, 322,464 lemma-types, and 87,068 root-types.

By using the NGS, a massive reduction was gained in execution time for processing the 176-million words corpus to only 5 days. It might have been a few hours, if enough CPUs had been allocated to process all files strictly in parallel; NGS provides virtual parallel processing on a reduced set of CPUs. Therefore, the half-million-word files were divided into three groups containing 100, 150 and 80 files respectively depending on the number of CPUs they were allocated. The average CPU time used to lemmatize a file of average 584,599 words was 91,102 seconds (25 hours, 18 minutes and 22 seconds) at an average of 6.4 words per second. The total CPU time used to lemmatize all the corpus files was 30,245,965 seconds (8401 hours, 39 minutes and 25 second – approximately one year). However, five days were enough to lemmatize the 176-million word Arabic Internet Corpus via parallel processing.

[68] NGS (National Grid Services) http://www.ngs.ac.uk
NGS case study: Accelerating the Processing of Large Corpora, http://www.ngs.ac.uk/accelerating-the-processing-of-large-corpora-using-grid-computing-technologies-for-lemmatizing-176

After lemmatizing the three groups of corpus files, the lemmatized output files were combined into one lemmatized Arabic Internet Corpus. The lemmatized corpus was stored in one large tab-separated column file where the words occupy the first column, the lemmas occupy the second column, the roots occupy the third column, and special tags were added in the fourth column. These tags are: **STOP_WORD** to mark function words; **N_BP** to mark broken plural nouns; **NE_PERS** to mark personal named entities; **NE_LOC** to mark locational named entities and **NE_ORG** to mark organizational named entities.

Figure 10.1 shows a one-sentence example of the lemmatized Arabic Internet Corpus. The sentence is:

> لعله أن يكون كابوسا ويستفيق منه على الأشياء الأليفة والطيبة والحبيبة. وامتد الشارع الضيق طويلا.. طويلا، وجلست
> البيوت ساكنة، مطرقة، والمصابيح الصفراء المقرورة تنزف ضوءا.
>
> *la‘alla^{hu} ’an yakūna kābūs^{an} wa yastafīqu minhu ‘alā al-’ašyā’i al-’alīfa^{ti} wa aṭ-ṭayyiba^{ti} wa al-ḥabība^{ti}. wa imtadda aš-šāri‘u al-ḍayyiqu ṭawīl^{an}.. ṭawīl^{an} wa ğalasat al-buyūtu sākinat^{an}, muṭriqat^{an}, wa al-maṣābīḥu aṣ-ṣafrā’u al-maqrūra^{tu} tanzifu ḍaw’^{an}*
>
> ‘Perhaps it is a nightmare and he will wake up to the usual, good and beloved things. The narrow road is extend long. long. The homes sat silent, listening, speechless, and the yellow bubbled lamps bled light.’

لعله	عل	علل	
أن	أن	أن	STOP_WORD
يكون	كان	كون	STOP_WORD
كابوسا	كابوس	كبس	
ويستفيق	يستفيق	فوق	
منه	منه	منه	STOP_WORD
على	على	على	STOP_WORD
الأشياء	أشياء	شيأ	
الأليفة	أليف	ءلف	
والطيبة	طيب	طيب	
والحبيبة	حبيب	حبب	
.	.	.	
وامتد	امتد	مدد	
الشارع	شارع	شرع	
الضيق	ضيق	ضيق	
طويلا	طويل	طول	
.	.	.	
.	.	.	
طويلا	طويل	طول	
،	،	،	
وجلست	جلس	جلس	
البيوت	بيت	بيت	N_BP
ساكنة	ساكن	سكن	
،	،	،	
مطرقة	مطرق	طرق	
،	،	،	
والمصابيح	مصابيح	صبح	
الصفراء	صفراء	صفر	
المقرورة	مقرور	قرر	
تنزف	نزف	زفف	
ضوءا	ضوء	ضوأ	

Figure 10.1 Sample of lemmatized sentence from the Arabic Internet Corpus

The main challenge of lemmatizing the 176-million words Arabic Internet Corpus was the long execution time that might take several months. This challenge was solved by using the high performance computational power provided by the NGS. The lemmatization of the AIC was significantly reduced to 5 days.

The other challenge that appeared during lemmatizing the AIC was the many cases of spelling errors. The AIC was collected automatically from web pages (Sharoff 2006). These web pages were constructed using different web authoring tools which have integrated word processing modules. Most of these word processing tools that support Arabic are not aware of what letter and diacritic combinations can appear on a letter in a given position in the word. The absence of such a module in word processing tools that support Arabic increases the potential for mis-spelling Arabic words. Many spelling-errors are found in the AIC. Such errors are: adding more than one short vowel to the same letter; starting or ending the word with *taṭwīl*; adding a diacritic to *taṭwīl*; starting the Arabic word with a silent letter by adding *sukūn* to the first letter; and adding *tanwīn* to any of the word's letters other than the last letter.

The SALMA – Tokenizer has a specialized procedure that checks whether the letter and diacritic combinations are correct or not; see section 8.3.1. The first step in lemmatization is the tokenization of the corpus words that classifies words into Arabic words or other words (*i.e.* number, currency, non-Arabic word and date). The Arabic words are passed to the spell-checking procedure that discovers the spelling errors and corrects them. The mis-spelled words are replaced by the correct words.

10.2.1 Evaluation of the Lemmatizer Accuracy

There was not a gold standard for evaluating the accuracy of the AIC lemmas and roots accuracy. Therefore, small random samples were selected and the accuracy was computed for each sample. To evaluate the accuracy of the lemmatizer, in terms of lemma and root accuracies, 10 samples of 100-words each from the lemmatized AIC were randomly selected. For each word in the sample the lemma and root accuracies were computed by counting the percentage of correct lemma and root analyses in the samples. Tables 10.1 and 10.2 show the accuracy results for each sample. Accumulative averages of both the lemma and root accuracies were computed to track the accuracy changes from one sample to another. The accumulative average accuracy showed steady accuracy rates among the selected samples. So, the evaluation stopped adding more samples. The accumulative accuracy averages were reported as the lemma and root accuracies of the AIC. Figure 10.2 shows the lemma accuracy and root accuracy for each sample, the accumulative average of the lemma accuracy, and the accumulative average of the root accuracy.

The results show that the accumulative average root accuracy is 81.20% and the average lemma accuracy is 80.80%.

Table 10.1 Lemma accuracy

Sample	Sample name	Start line	Tokens	Correct lemmas	Accuracy %	Average %
1	newdp_out.txt	111,435	100	81	81.00%	81.00%
2	newfo_out.txt	384,384	100	76	76.00%	78.50%
3	newih_out.txt	113691	100	78	78.00%	78.33%
4	newca_out.txt	13,076	100	80	80.00%	78.75%
5	newfc_out.txt	59,313	100	78	78.00%	78.60%
6	newlg_out.txt	234,254	100	85	85.00%	79.67%
7	newdr_out.txt	570,807	100	77	77.00%	79.29%
8	newmi_out.txt	507,492	100	80	80.00%	79.38%
9	newir_out.txt	355,144	100	82	82.00%	79.67%
10	neweu_out.txt	149,057	100	91	91.00%	80.80%
			1000	**808**	**80.80%**	**80.80%**

Table 10.2 Root accuracy

Sample	Sample name	Start line	Tokens	Correct roots	Accuracy %	Average %
1	newdp_out.txt	111,435	100	85	85.00%	85.00%
2	newfo_out.txt	384,384	100	72	72.00%	78.50%
3	newih_out.txt	113691	100	80	80.00%	79.00%
4	newca_out.txt	13,076	100	82	82.00%	79.75%
5	newfc_out.txt	59,313	100	79	79.00%	79.60%
6	newlg_out.txt	234,254	100	85	85.00%	80.50%
7	newdr_out.txt	570,807	100	71	71.00%	79.14%
8	newmi_out.txt	507,492	100	85	85.00%	79.88%
9	newir_out.txt	355,144	100	84	84.00%	80.33%
10	neweu_out.txt	149,057	100	89	89.00%	81.20%
			1000	**812**	**81.20%**	**81.20%**

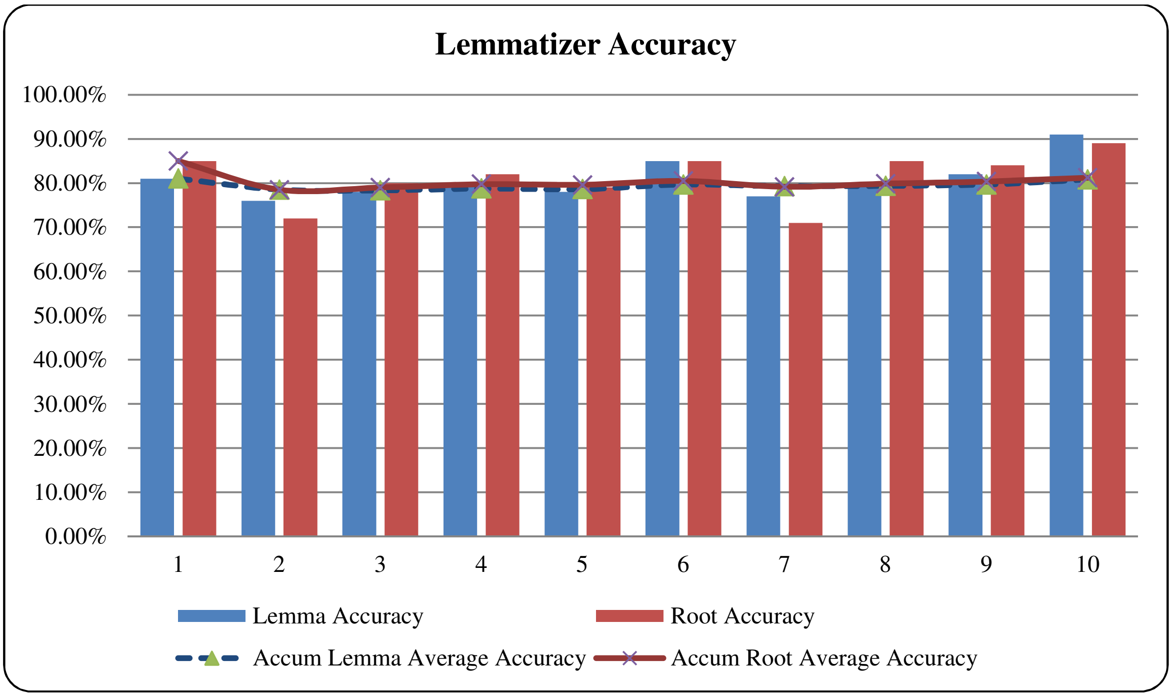


Figure 10.2 Lemma and root accuracy of the lemmatized Arabic internet corpus

10.3 Corpus Linguistics Resources and Tools for Arabic Lexicography

Corpora have been used to construct dictionaries since the release of the Collins-Birmingham University International Database COBUILD. Computer technology was used in the four stages of constructing COBUILD: data-collection, entry-selection, entry construction and entry-arrangement (Ooi 1998).

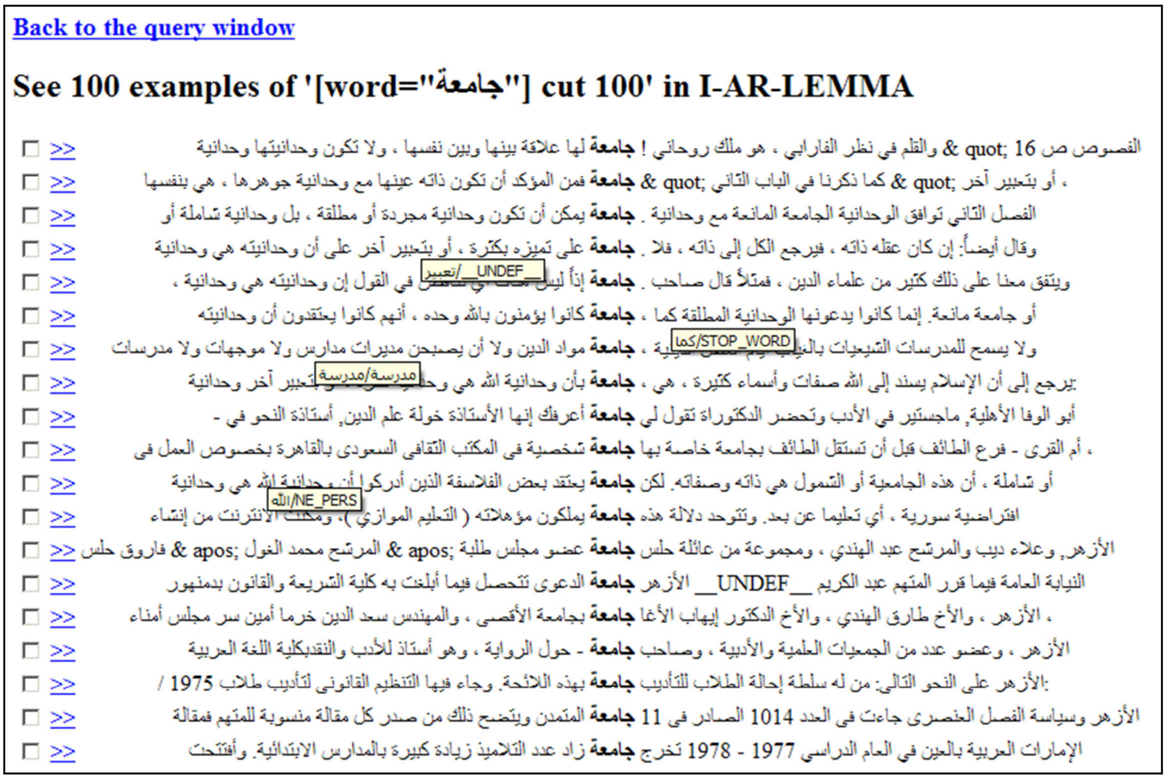
A Large and representative corpus which is made up of texts of many different domains, formats and genres provides detailed information about all aspects of written language that can be studied. Corpus and corpus analysis tools *e.g.* Sketch Engine[69], have brought about a revolution in dictionary building. Corpus analysis tools are used to build a detailed statistical profile of every word in the corpus, which enables lexicographers to understand the words, their collocations, their behaviors, usages and the connotations they may carry. Ways of producing new words and expressions and the popularity of coinages can be identified with the help of the corpus. Oxford dictionaries[70] represent an exemplar of the use of corpus in constructing dictionaries.

The second and traditional source of information which is used to construct dictionaries is citations. Citations represent the objective evidence of language in use. They are a prerequisite for a reliable dictionary but they have their limitations (Atkins and Rundell 2008).

[69] Corpus analysis tools such as Sketch Engine (www.sketchengine.co.uk)

[70] Oxford dictionaries http://www.oxforddictionaries.com

Arabic corpora have not been used to construct Arabic dictionaries[71]. Advances in corpora construction technologies, corpora analysis tools and the availability of large quantities of Arabic text of different domains, formats and genres on the web can allow us to build a large and representative lexicographic corpus of Arabic to be used in constructing new Arabic dictionaries. A lemmatizing tool is needed to group words that share the same lemma. It also helps in finding the collocations of the word. Figures 10.3 and 10.4 show examples of the word جامِعَة *ğāmi'a^t* "University" and its collocations.



Back to the query window

See 100 examples of '[word="جامعة"] cut 100' in I-AR-LEMMA

□ >> الفصوص ص 16 ;quot & والقلم في نظر الفارابي ، هو ملك روحاني ! **جامعة** لها علاقة بينها وبين نفسها ، ولا تكون وحدانيتها وحدانية

□ >> ، أو بتعبير آخر ;quot & كما ذكرنا في الباب الثاني ;quot & **جامعة** فمن المؤكد أن تكون ذاته عينها مع وحدانية جوهرها ، هي بنفسها

□ >> الفصل الثاني توافق الوحدانية الجامعة المانعة مع وحدانية . **جامعة** يمكن أن تكون وحدانية مجردة أو مطلقة ، بل وحدانية شاملة أو

□ >> وقال أيضاً: إن كان عقله ذاته ، فيرجع الكل إلى ذاته ، فلا . **جامعة** على تميزه بكثرة ، أو بتعبير آخر على أن وحدانيته هي وحدانية

□ >> ويتفق معنا على ذلك كثير من علماء الدين ، فمثلاً قال صاحب . **جامعة** إذاً ليس هناك أي تناقض في القول إن وحدانيته هي وحدانية ،

□ >> أو جامعة مانعة. إنما كانوا يدعونها الوحدانية المطلقة كما ، **جامعة** كانوا يؤمنون بالله وحده ، أنهم كانوا يعتقدون أن وحدانيته

□ >> ولا يسمح للمدرسات الشيعيات [illegible] ، **جامعة** مواد الدين ولا أن يصبحن مديرات مدارس ولا موجهات ولا مدرسات

□ >> يرجع إلى أن الإسلام يسند إلى الله صفات وأسماء كثيرة ، هي ، **جامعة** بأن وحدانية الله هي [illegible] وبتعبير آخر وحدانية:

□ >> أبو الوفا الأهلية, ماجستير في الأدب وتحضر الدكتوراة تقول لي **جامعة** أعرفك إنها الأستاذة خولة علم الدين, أستاذة النحو في -

□ >> ، أم القرى - فرع الطائف قبل أن تستقل الطائف بجامعة خاصة بها **جامعة** شخصية فى المكتب الثقافى السعودى بالقاهرة بخصوص العمل فى

□ >> أو شاملة ، أن هذه الجامعية أو الشمول هي ذاته وصفاته. لكن **جامعة** يعتقد بعض الفلاسفة الذين أدركوا أن وحدانية الله هي وحدانية

□ >> افتراضية سورية ، أي تعليما عن بعد. وتتوحد دلالة هذه **جامعة** يملكون مؤهلاته (التعليم الموازي)، ومكنت الانترنت من إنشاء

□ >> الأزهر, وعلاء ديب والمرشح عبد الهندي ، ومجموعة من عائلة حلس **جامعة** عضو مجلس طلبة ;apos & المرشح محمد الغول ;apos & فاروق حلس

□ >> النيابة العامة فيما قرر المتهم عبد الكريم __UNDEF__ الأزهر **جامعة** الدعوى تتحصل فيما أبلغت به كلية الشريعة والقانون بدمنهور

□ >> ، الأزهر ، والأخ طارق الهندي ، والأخ الدكتور إيهاب الأغا **جامعة** بجامعة الأقصى ، والمهندس سعد الدين خرما أمين سر مجلس أمناء

□ >> الأزهر ، وعضو عدد من الجمعيات العلمية والأدبية ، وصاحب **جامعة** - حول الرواية ، وهو أستاذ للأدب والنقدبكلية اللغة العربية

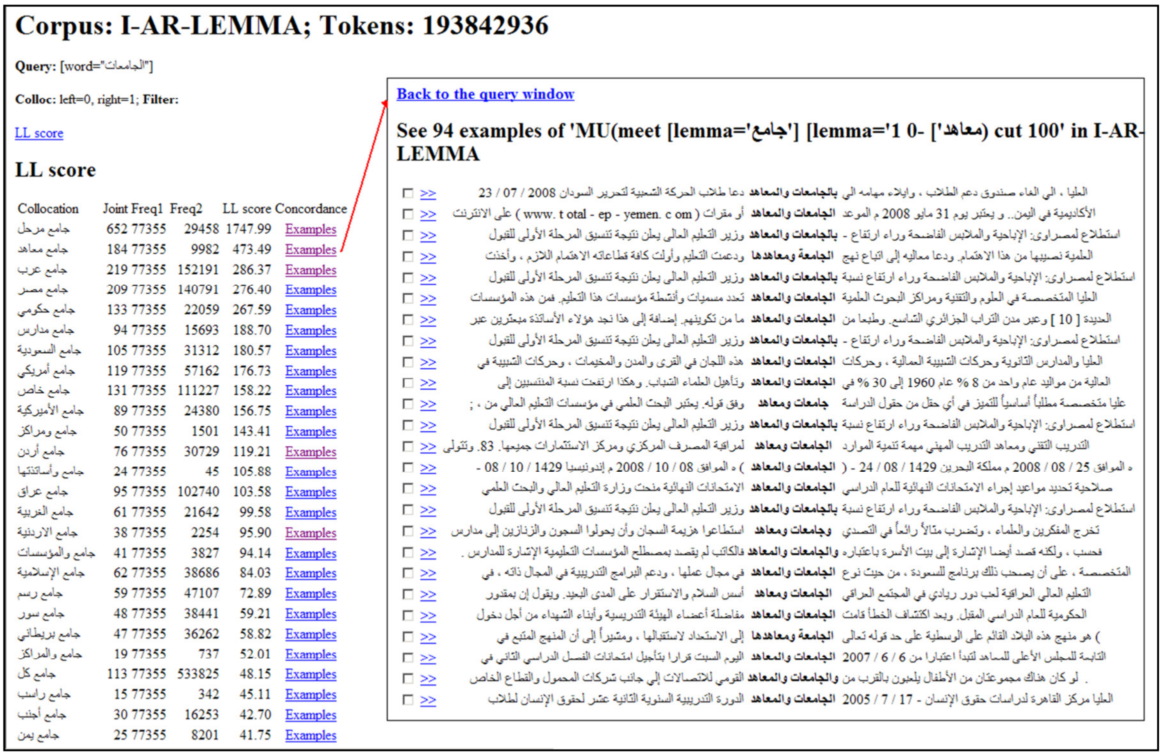
□ >> الأزهر على النحو التالي: من له سلطة إحالة الطلاب للتأديب **جامعة** بهذه اللائحة. وجاء فيها التنظيم القانونى لتأديب طلاب 1975 /:

□ >> الأزهر وسياسة الفصل العنصرى جاءت فى العدد 1014 الصادر فى 11 **جامعة** المتمدن ويتضح ذلك من صدر كل مقالة منسوبة للمتهم فمقالة

□ >> الإمارات العربية بالعين في العام الدراسي 1977 - 1978 تخرج **جامعة** زاد عدد التلاميذ زيادة كبيرة بالمدارس الابتدائية. وأفتتحت

Figure 10.3 Example of the concordance line of the word جامعة *ğāmi'a^t* "University" from the Arabic Internet Corpus

[71] The last Arabic dictionary المُعْجَمُ الوَسِيْطُ *mu'jam al-wasīṭ* "Al-Waseet Lexicon" appeared in 1960's by the Arabic language academy in Cairo.



Corpus: I-AR-LEMMA; Tokens: 193842936

Query: [word="الجامعات"]

Colloc: left=0, right=1; **Filter:**

LL score

LL score

Collocation	Joint	Freq1	Freq2	LL score	Concordance
جامع مرحل	652	77355	29458	1747.99	Examples
جامع معاهد	184	77355	9982	473.49	Examples
جامع عرب	219	77355	152191	286.37	Examples
جامع مصر	209	77355	140791	276.40	Examples
جامع حكومي	133	77355	22059	267.59	Examples
جامع مدارس	94	77355	15693	188.70	Examples
جامع السعودية	105	77355	31312	180.57	Examples
جامع أمريكي	119	77355	57162	176.73	Examples
جامع خاص	131	77355	111227	158.22	Examples
جامع الأميركية	89	77355	24380	156.75	Examples
جامع ومراكز	50	77355	1501	143.41	Examples
جامع أردن	76	77355	30729	119.21	Examples
جامع وأساتذتها	24	77355	45	105.88	Examples
جامع عراق	95	77355	102740	103.58	Examples
جامع الغربية	61	77355	21642	99.58	Examples
جامع الاردنية	38	77355	2254	95.90	Examples
جامع والمؤسسات	41	77355	3827	94.14	Examples
جامع الإسلامية	62	77355	38686	84.03	Examples
جامع رسم	59	77355	47107	72.89	Examples
جامع سور	48	77355	38441	59.21	Examples
جامع بريطاني	47	77355	36262	58.82	Examples
جامع والمراكز	19	77355	737	52.01	Examples
جامع كل	113	77355	533825	48.15	Examples
جامع راسب	15	77355	342	45.11	Examples
جامع أجنب	30	77355	16253	42.70	Examples
جامع يمن	25	77355	8201	41.75	Examples

Back to the query window

See 94 examples of 'MU(meet [lemma='جامع'] [lemma='1 0- ['معاهد) cut 100' in I-AR-LEMMA

```
□ >> العليا ، الي الغاء صندوق دعم الطلاب ، وايلاء مهامه الي بالجامعات والمعاهد دعا طلاب الحركة الشعبية لتحرير السودان 23 / 07 / 2008
□ >> الأكاديمية في اليمن.. و يعتبر يوم 31 مايو 2008 م الموعد الجامعات والمعاهد أو مقرات ( www. t otal - ep - yemen. c om ) على الانترنت
□ >> استطلاع لمصراوى: الإباحية والملابس الفاضحة وراء ارتفاع - بالجامعات والمعاهد وزير التعليم العالى يعلن نتيجة تنسيق المرحلة الأولى للقبول
□ >> العلمية نصيبها من هذا الاهتمام. ودعا معاليه إلى اتباع نهج الجامعة ومعاهدها ودعمت التعليم وأولت كافة قطاعاته الاهتمام اللازم ، وأخذت
□ >> استطلاع لمصراوى: الإباحية والملابس الفاضحة وراء ارتفاع نسبة بالجامعات والمعاهد وزير التعليم العالى يعلن نتيجة تنسيق المرحلة الأولى للقبول
□ >> العليا المتخصصة في العلوم والتقنية ومراكز البحوث العلمية الجامعات والمعاهد تعدد مسميات وأنشطة مؤسسات هذا التعليم. فمن هذه المؤسسات
□ >> الجديدة [ 10 ] وعبر مدن التراب الجزائري الشاسع. وطبعا من الجامعات والمعاهد ما من تكوينهم. إضافة إلى هذا نجد هؤلاء الأساتذة مبعثرين عبر
□ >> استطلاع لمصراوى: الإباحية والملابس الفاضحة وراء ارتفاع - بالجامعات والمعاهد وزير التعليم العالى يعلن نتيجة تنسيق المرحلة الأولى للقبول
□ >> العليا والمدارس الثانوية وحركات الشبيبة العمالية ، وحركات الجامعات والمعاهد هذه اللجان في القرى والمدن والمخيمات ، وحركات الشبيبة في
□ >> العالية من مواليد عام واحد من 8 % عام 1960 إلى 30 % في الجامعات والمعاهد وتأهيل العلماء الشباب. وهكذا ارتفعت نسبة المنتسبين إلى
□ >> عليا متخصصة مطلباً أساسياً للتميز في أي حقل من حقول الدراسة جامعات ومعاهد وفق قوله. يعتبر البحث العلمي في مؤسسات التعليم العالي من ، ;
□ >> استطلاع لمصراوى: الإباحية والملابس الفاضحة وراء ارتفاع نسبة بالجامعات والمعاهد وزير التعليم العالى يعلن نتيجة تنسيق المرحلة الأولى للقبول
□ >> التدريب التقني ومعاهد التدريب المهني مهمة تنمية الموارد الجامعات ومعاهد لمراقبة المصرف المركزي ومركز الاستثمارات جميعها. 83. وتتولى
□ >> ه الموافق 25 / 08 / 2008 م مملكة البحرين 1429 / 08 / 24 - ( الجامعات والمعاهد ) ه الموافق 08 / 10 / 2008 م إندونيسيا 1429 / 10 / 08 -
□ >> صلاحية تحديد مواعيد إجراء الامتحانات النهائية للعام الدراسي الجامعات والمعاهد الامتحانات النهائية منحت وزارة التعليم العالي والبحث العلمي
□ >> استطلاع لمصراوى: الإباحية والملابس الفاضحة وراء ارتفاع نسبة بالجامعات والمعاهد وزير التعليم العالى يعلن نتيجة تنسيق المرحلة الأولى للقبول
□ >> تخرج المفكرين والعلماء ، وتضرب مثالاً رائعاً في التصدي وجامعات ومعاهد استطاعوا هزيمة السجان وأن يحولوا السجون والزنازين إلى مدارس
□ >> فحسب ، ولكنه قصد أيضا الإشارة إلى بيت الأسرة باعتباره والجامعات والمعاهد فالكاتب لم يقصد بمصطلح المؤسسات التعليمية الإشارة للمدارس .
□ >> المتخصصة ، على أن يصحب ذلك برنامج للسعودة ، من حيث نوع الجامعات والمعاهد في مجال عملها ، ودعم البرامج التدريبية في المجال ذاته ، في
□ >> التعليم العالي العراقية لعب دور ريادي في المجتمع العراقي الجامعات ومعاهد أسس السلام والاستقرار على المدى البعيد. ويقول إن بمقدور
□ >> الحكومية للعام الدراسي المقبل. وبعد اكتشاف الخطأ قامت الجامعات والمعاهد مفاضلة أعضاء الهيئة التدريسية وأبناء الشهداء من أجل دخول
□ >> ) هو منهج هذه البلاد القائم على الوسطية على حد قوله تعالى الجامعة ومعاهدها إلى الاستعداد لاستقبالها ، ومشيراً إلى أن المنهج المتبع في
□ >> التابعة للمجلس الأعلى للمعاهد لتبدأ اعتبارا من 6 / 6 / 2007 الجامعات والمعاهد اليوم السبت قرارا بتأجيل امتحانات الفصل الدراسي الثاني في
□ >> . لو كان هناك مجموعتان من الأطفال يلعبون بالقرب من والجامعات والمعاهد القومي للاتصالات إلي جانب شركات المحمول والقطاع الخاص
□ >> العليا مركز القاهرة لدراسات حقوق الإنسان - 17 / 7 / 2005 الجامعات والمعاهد الدورة التدريبية السنوية الثانية عشر لحقوق الإنسان لطلاب
```

Figure 10.4 Example of the collocations of the word جامعة *ğāmi‘a^t* “University” from the Arabic Internet Corpus

The second important resource of information needed to construct new Arabic dictionaries is the long established traditional Arabic lexicons. Over the past 1200 years, many different kinds of Arabic lexicons were constructed; these lexicons are different in ordering, size and goal of construction. The traditional Arabic lexicons followed four main methodologies for ordering their lexical entries. These methodologies use the root as lexical entry. The main disadvantage of these methodologies is that the words derived from the root are not arranged methodically within the lexical entry. Ordering of dictionary entries is the main challenge in constructing Arabic dictionaries.

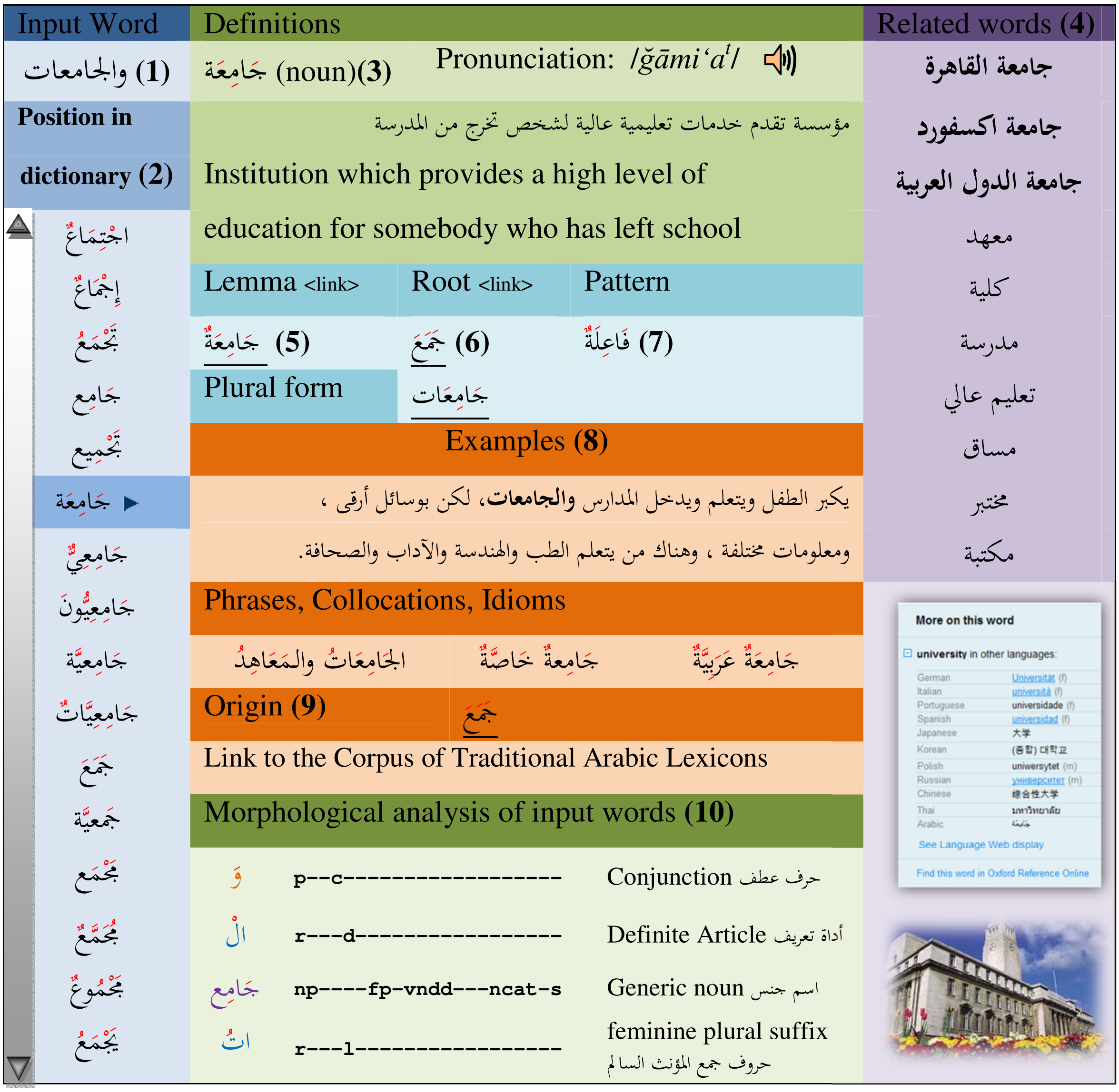
Traditional Arabic lexicons represent a citation bank to be used in the construction of modern Arabic dictionaries. They include citations for each lexical entry from the Qur’an and authentic poetry that represents the proper use of keywords. They provide information about the origin of words. They also include phrases, collocations, idioms, and well-known personal names and places derived from that root (lexical entry).

The corpus of traditional Arabic lexicons is a collection of 23 lexicons. It represents a different domain than existing Arabic corpora. It covers a period of more than 1200 years. It consists of a large number of words, about 14,369,570 and about 2,184,315 word types. The corpus of traditional Arabic lexicons has both types of Arabic text; vowelized and non-vowelized. Figure 10.5 shows the most frequent words of the Corpus of Traditional Arabic Lexicons, see section 4.6.

Partially-vowelized		Non-vowelized	
Word	**Frequency**	**Word**	**Frequency**
في *fī* “in”	292,396	من *min* “from”	322,239
من *min* “from”	269,200	في *fī* “in”	301,895
قال *qāl* “he said”	172,631	قال *qāl* “he said”	190,918
و *wa* “and”	120,060	أي *’ay* “which”	132,635
على *‘alā* “over”	108,252	و *wa* “and”	130,809
ما *mā* “what”	89,195	على *‘alā* “over”	119,639
وقال *wa qāl* “and he said”	88,233	إذا *’iẖā* “if”	115,842
عن *‘an* “about”	82,027	وقال *wa qāl* “and he said”	99,601
إذا *’iẖā* “if”	81,479	ابن *’ibn* “son of”	94,980
أي *’ay* “which”	78,622	ما *mā* “what”	94,530
وهو *wa huwa* “and he”	75,149	بن *bin* “son of”	92,213
لا *lā* “no”	69,737	عن *‘an* “about”	87,064
ابن *’ibn* “son of”	58,334	وهو *wa huwa* “and he”	80,375
به *bihi* “in it”	53,343	لا *lā* “no”	73,066
وفي *wa fī* “and in”	53,197	أبو *abū* “father”	72,231
وقد *wa qad* “and perhaps”	50,648	أن *’an* “that”	65,419
أبو *abū* “father”	47,915	أو *’aw* “or”	62,298
بن *bin* “son of”	46,880	الله *allā*^{h} “Allah”	59,511
أَي *’ay* “which”	46,788	به *bihi* “in it”	58,941
هو *huwa* “he”	45,916	يقال *yuqāl* “it is said”	58,062
يقال *yuqāl* “it is said”	45,794	وفي *wa fī* “and in”	55,077
عليه *‘alayhi* “about him”	44,786	وقد *wa qad* “and perhaps”	53,992
ولا *wa lā* “and not”	42,190	عليه *‘alayhi* “about him”	50,906
الله *allā*^{h} “Allah”	39,961	هو *huwa* “he”	49,785
أو *’aw* “or”	39,210	إلى *’ilā* “to”	48,363

Figure 10.5 The Corpus of Traditional Arabic Lexicons frequency lists

Figure 10.6 shows a proposed web interface for an Arabic dictionary that illustrates the adaptation of the resources, standards and tools developed in this research as language-engineering tools to construct Arabic dictionaries.



Input Word

والجامعات (1)

Position in

dictionary (2)

اجْتِمَاعٌ
إِجْمَاعٌ
تَجَمُّعٌ
جَامِع
تَجْمِيع
جَامِعَة
جَامِعِيٌّ
جَامِعِيُّون
جَامِعِيَّة
جَامِعِيَّاتٌ
جَمَعَ
جَمْعِيَّة
مُجْمَع
مُجَمَّعٌ
مَجْمُوعٌ
يَجْمَعُ

Definitions

جَامِعَة (noun)(3) Pronunciation: */ğāmiʻa^t/*

مؤسسة تقدم خدمات تعليمية عالية لشخص تخرج من المدرسة

Institution which provides a high level of education for somebody who has left school

Lemma <link>	Root <link>	Pattern
جَامِعَةٌ (5)	جَمَعَ (6)	فَاعِلَةٌ (7)
Plural form	جَامِعَات	

Examples (8)

يكبر الطفل ويتعلم ويدخل المدارس **والجامعات**، لكن بوسائل أرقى ،

ومعلومات مختلفة ، وهناك من يتعلم الطب والهندسة والآداب والصحافة.

Phrases, Collocations, Idioms

جَامِعَةٌ عَرَبِيَّةٌ | جَامِعَةٌ خَاصَّةٌ | الجَامِعَاتُ والمَعَاهِدُ

Origin (9) جَمَعَ

Link to the Corpus of Traditional Arabic Lexicons

Morphological analysis of input words (10)

وَ	`p--c------------------`	Conjunction حرف عطف
الْ	`r---d-----------------`	Definite Article أداة تعريف
جَامِع	`np----fp-vndd---ncat-s`	Generic noun اسم جنس
اتُ	`r---l-----------------`	feminine plural suffix حروف جمع المؤنث السا لم

Related words (4)

جامعة القاهرة
جامعة اكسفورد
جامعة الدول العربية
معهد
كلية
مدرسة
تعليم عالي
مساق
مختبر
مكتبة

More on this word

⊟ **university** in other languages:

German	Universität (f)
Italian	università (f)
Portuguese	universidade (f)
Spanish	universidad (f)
Japanese	大学
Korean	(종합) 대학교
Polish	uniwersytet (m)
Russian	университет (m)
Chinese	综合性大学
Thai	มหาวิทยาลัย
Arabic	جَامِعَة

See Language Web display

Find this word in Oxford Reference Online

Figure 10.6 A proposed web interface for Arabic dictionary

The number label on the figure is mapped to one of the resources, standards and tools:

- **Label number 1:** This allows users to search for any word. The SALMA – Lemmatizer and Stemmer can be used to extract the lemma (lexical entry) related to the input word and retrieve the definitions stored in the dictionary.
- **Label number 2:** The SALMA – ABCLexicon can be used to retrieve a list of alphabetically ordered lexical entries that share the same root.
- **Label number 3:** The SALMA – Tagger can provide the main part-of-speech of the lexical entry.

- **Label number 4:** The lemmatized AIC can be used to retrieve related words by measuring the Loglikelihood, T-score and Mutual Information to extract the collocation of the searched word
- **Labels number 5 and 6:** The SALMA-Lemmatizer can be used to extract the lemma and the root of the entered word.
- **Label number 7:** The pattern information can be produced using the SALMA – Pattern Generator.
- **Label number 8:** Examples are selected from the lemmatized AIC concordance lines of the input word and its lemma.
- **Label number 9:** The origin of this word and the time line of the semantic development of the lexical entries can be investigated via a link to the Corpus of Traditional Arabic Lexicons.
- **Label number 10:** The morphological analysis of the input word, its morphemes and the morphological features of each morpheme are described using both the SALMA – Tag Set and the SALMA – Tagger.

10.4 Chapter Summary

Resources, standards and tools developed in this research have many potential applications as they work as fundamental prerequisites for most Arabic text analytics applications. The main purpose in developing the resources, standards and tools is to annotate an Arabic text corpus with fine-grain morphosyntactic information. This chapter investigated two applications of these resources, standards and tools: lemmatizing the 176-million word Arabic Internet Corpus (AIC), and as language engineering resources to construct an Arabic dictionary.

The developed resources, standards and tools were evaluated on a sample of Arabic text to measure their accuracy and applicability for use to perform text analytics tasks. However, the performance aspects of the SALMA – Tagger such as speed, memory and ability to perform the desired analysis tasks were not evaluated previously. Applying the SALMA – Lemmatizer and Stemmer to lemmatize the 176-million word Arabic Internet Corpus is a practical application that evaluated its performance and investigated the challenges of applying the resources, standards and tools on real and large-scale data. Two main challenges arose during the lemmatizing of the AIC: the speed and the spelling errors. NGS was used to lemmatize the divided parts of the AIC in parallel. A massive reduction in execution time was gained. The SALMA – Tokenizer was used to detect and correct the spelling errors that appear in the AIC due to poor word processing tools used in authoring web pages.

The second application is a proposal about how these resources, standards and tools can be used as a language engineering toolkit for Arabic lexicography. This study reviews the resources and tools which are used in modern lexicography, and shows that the developed resources, and standards constitute a toolkit for constructing Arabic monolingual and bi-lingual dictionaries.

Part V: Conclusions and Future Work

Chapter 11
Conclusions and Future Work

11.1 Overview

Arabic morphological analyzers and stemming algorithms have become a popular area of research. This chapter reviews the main contributions of this thesis to this area. It discusses the conclusions drawn from experimental work, and connects these findings with related future work. Finally, the chapter summarises PhD impact, originality and contributions to Arabic NLP.

Several computational linguists have designed and developed algorithms to address problems in automatic morphosyntactic annotation of Arabic text. This thesis has surveyed current Arabic morphological analyzers, and conducted experiments to discover the theoretical and practical challenges of morphological analysis for Arabic. Practical work includes the development of resources to enhance the accuracy of such systems, where these resources can also be reused in diverse Arabic text analytics applications. It also includes the proposal of linguistically informed standards for Arabic morphological analysis which draw on the long-established traditions of Arabic grammar. Finally, resources and proposed standards are brought together in the development of the SALMA – Tagger: a fine-grained morphological analyzer for Arabic text of different domains, formats and genres.

Resources, proposed standards and tools are intended to be open-source. The development of the SALMA – Tagger used the open source programming language Python because it is intended for integration into the Natural Language Toolkit (NLTK[72]), a set of open source Python modules, linguistic data and documentation for research and development in natural language processing and text analytics.

11.2 Thesis Achievements and Conclusions

This section summarises the main achievements of this thesis and the conclusions drawn from experimental work. It starts by discussing the practical challenges of Arabic morphological analysis. The second section discusses the motivations and benefits of creating the SALMA – ABCLexicon as a lexical resource for improving Arabic

[72] Natural Language Toolkit (NLTK) http://www.nltk.org

morphological analyzers. Section 11.2.3 discusses standardization of morphosyntactic annotation for Arabic corpora. Section 11.2.4 covers the application of proposed standards and resources developed in the SALMA – Tagger, a tool for fine-grain morphological analysis of Arabic text. Finally, section 11.2.5 discusses the evaluation of the SALMA – Tagger, focusing on the fine-grained morphological feature categories, and draws conclusions from this evaluation that suggest opportunities for future work to enhance the performance and accuracy of the SALMA – Tagger as a language-engineering toolkit for morphosyntactic analysis for Arabic text.

11.2.1 The Practical Challenge of Morphological Analysis for Arabic Text

Several stemming algorithms for Arabic already exist, but each researcher proposes an evaluation methodology based on different text corpora. Therefore, direct comparisons between these evaluations cannot be made. At the time of the experiment, only three stemming algorithms and morphological analyzers for Arabic text were readily accessible to assess their implementation and/or performance results. The three selected algorithms are Khoja's stemmer (Khoja 2003), Buckwalter's morphological Analyzer (BAMA) (Buckwalter 2002) and the triliteral root extraction algorithm (Al-Shalabi et al. 2003).

A range of four fair and precise evaluation experiments was conducted using a gold standard for evaluation consisting of two 1000-word text documents from the Holy Qur'an and the Corpus of Contemporary Arabic. The four experiments on both text samples show the same accuracy rank for the stemming algorithms: Khoja's stemmer achieved the highest accuracy, then the triliteral root extraction algorithm, and finally BAMA. The results show that:

- The stemming algorithms used in the experiments work better on MSA text (*i.e.* newspaper text) than Classical Arabic (*i.e.* Qur'an text), not unexpectedly as they were originally designed for stemming MSA text (*i.e.* newspaper text). The SALMA – Tagger is designed for wide coverage and so can deal with both genres.
- All stemming algorithms involved in the experiments agree and generate correct analysis for simple roots that do not require detailed analysis. So, more detailed analysis and enhancements are recommended as future work.
- Most stemming algorithms are designed for information retrieval systems where accuracy of the stemmers is not such an important issue. On the other hand, accuracy is vital for natural language processing, and this what the SALMA – Tagger is designed for.

- Accuracy rates surveyed show that even the best algorithm failed to achieve an accuracy rate of more than 75%. This proves that more research is required: part-of-speech tagging and then parsing cannot rely on such stemming algorithms because errors from the stemming algorithms will propagate to such systems.

To give a clear picture of the stemming problem, an analytical study was conducted to compute the percentage of triliteral roots, words, and word type distribution on 22 categories of triliteral roots, as classified in sections 3.7 and 6.2.21. The roots, words and word types of the Qur'an and the SALMA-ABCLexicon were analysed. The study clearly showed that about one third of Arabic text words have roots belonging to the defective or defective and hamzated root categories (*i.e.* one or two root radicals belong to vowels or *hamza*^{h}). Words belonging to these two root categories are hard to analyze and the root extraction process of such words always has higher error rates than for words belonging to the intact root category. Existing stemming and morphological analyzers are subject to mistakes when analysing words belonging to these two categories.

The evaluation methodology used in this thesis for stemming algorithms and morphological analyzers for Arabic text based on the gold standard has since been reused and referenced by Alotaiby, Alkharashi et al. (2009), Kurimo, Virpioja et al. (2009), Harrag, Hamdi-Cherif et al. (2010), Yusof, Zainuddin et al. (2010), Al-Jumaily, Martínez et al. (2011), and Hijjawi, Bandar et al. (2011)..

11.2.2 Resources for improving Arabic Morphological Analysis

The previous section raises the following question: How can we improve stemming and morphological analysis for Arabic so the algorithm can deal successfully with the hard cases of the 35% of words belonging to defective and defective and hamzated triliteral root categories? Two methodologies can be adopted: either to build a sophisticated algorithm that deals with the hard cases or simply to provide the algorithm with a prior-knowledge broad-coverage lexical resource that contains most of the hard case words and their triliteral roots and enables direct access to its contents. The stemming algorithm then looks up the word to be analysed in the lexicon and gets the correct analysis for that word.

We chose to construct a broad-coverage lexical resource, the SALMA - ABCLexicon to improve the accuracy of Arabic morphological analysis rather than developing a sophisticated stemming algorithm. Our choice was influenced by our interest in Arabic lexicon development and the advantages to be gained from developing the SALMA – ABCLexicon such as:

- Improving Arabic morphological analysis by providing a broad-coverage lexical resource that can be integrated to different stemming algorithms and can reduce the series of complex analysis steps to a simpler look-up procedure.
- The broad-coverage lexical resource can be a stand-alone resource which can be integrated in different Arabic natural language processing systems and benefits from such integration can be gained.
- It is easier to update the lexical resource by adding new content to it and correcting it than updating a sophisticated algorithm which needs specialized developers.
- It can also be used as a material resource to assist in the teaching-learning process.

The SALMA-ABCLexicon was constructed by analysing the text of 23 traditional Arabic lexicons, all of which are freely available open-source documents, and by following an agreed standard for constructing a morphological lexicon from raw text. However, three factors directed the selection of traditional Arabic lexicons as our raw text corpus: (i) the absence of an open-source, large, representative Arabic corpus; (ii) the absence of an open-source generation program; and (iii) the generation programme problems of over-generation and under-generation. The major advantages of using the traditional Arabic lexicons text as a corpus are: the corpus contains a large number of words (14,369,570) and word types (2,184,315), and the possibility of finding the different forms of the derived words of a given root.

The SALMA-ABCLexicon is constructed by combining information extracted from disparate lexical resource formats and merging Arabic lexicons. The coverage of the SALMA – ABCLexicon was computed via two methods. The first was to match the words of the test corpora to the words in the lexicon, which scored about 67%. The second was to use a lemmatizer to compute the coverage, which scored about 82% for the Qur'an, the CCA, and a million-word sample of the AIC.

The SALMA-ABCLexicon contains 2,781,796 vowelized word-root pairs which represent 509,506 different non-vowelized words. The lexicon is stored in three different formats: tab-separated column files, XML files, and a relational database. It is also provided with access and searching facilities and a web interface that provides a facility for searching a certain root and retrieving the original root definitions of the analyzed traditional Arabic lexicons.

In addition, the Corpus of Traditional Arabic Lexicons (14,369,570 words, and 2,184,315 word types) was created as a special corpus constructed from the text of 23 traditional Arabic lexicons.

11.2.3 Standards for Arabic Morphosyntactic Analysis

The initial evaluation of morphological analyzers and stemmers for Arabic text pointed out the lack of standardization and guidelines for morphosyntactic annotation for Arabic text. These standards and guidelines are the prerequisites for morphosyntactic annotation of corpora. Therefore, eight existing Arabic tag sets were surveyed and compared in terms of purpose of design, characteristics, tag-set size, and their applications (section 5.3.7). The drawbacks of the existing tag sets for Arabic were found to be:

- Existing Arabic tag sets vary in size from 6 tags to 2000 or more tags.
- Some of these tag sets follow standards for tag set design for English such as the PATB tag sets, and these may not always be appropriate for Arabic.
- The tag sets share common morphological features such as gender, number, person, case, mood and definiteness, but the attributes of the morphological feature categories are not standardized.
- These tag sets lack standardization in defining a suitable scheme for tokenizing Arabic words into their morphemes and they mix morpheme tagging with whole word tagging.
- They also lack suitable documentation that illustrates the decision made for each design dimension of the tag set.
- The tags assigned to words in a corpus are not consistent in either presentation of the tag itself or the morphological features which are encoded within the tag.

Moreover, the most widely used and important morphosyntactic annotation standards and guidelines, namely EAGLES, are designed for Indo-European languages. These guidelines are not entirely suitable for Arabic.

The previous comparative evaluation of Arabic tag sets and the opportunity for making an original contribution motivated the development of the SALMA – Tag Set as proposed standard for morphological annotation for Arabic text corpora. This constitutes a common standard to simplify and promote comparisons and sharing of resources. For a morphologically rich language like Arabic, the Part-of-Speech tag set should be defined in terms of morphological features characterizing word structure. The SALMA – Tag Set has the following characteristics:

- The SALMA – Tag Set captures long-established traditional morphological features of Arabic, in a notation format intended to be compact yet transparent.
- A detailed description of the SALMA – Tag Set explains and illustrates each feature and its possible values.
- A tag consists of 22 characters; each position represents a feature and the letter at that location represents a value or attribute of the morphological feature; the dash "-" represents a feature not relevant to a given word.
- The SALMA – Tag Set is not tied to a specific tagging algorithm or theory, and other tag sets could be mapped onto this standard, to simplify and promote comparisons between and reuse of Arabic taggers and tagged corpora.

The SALMA – Tag Set has been validated in two ways. First, it was validated by proposing it as a standard for the Arabic language computing community, and it has been adopted in Arabic language processing systems.

- It has been used in the SALMA – Tagger to encode the morphological features of each morpheme (Sawalha and Atwell 2009a; Sawalha and Atwell 2010b).
- Parts of The SALMA Tag Set were also used in the Arabic morphological analyzer and part-of-speech tagger Qutuf (Altabbaa et al. 2010).
- It has been reported as a standard for evaluating morphological analyzers for Arabic text and for building a gold standard for evaluating morphological analyzers and part-of-speech taggers for Arabic text (Hamada 2010).

Second, an empirical approach to evaluating the SALMA Tag Set of Arabic showed that it can be applied to an Arabic text corpus, by mapping from an existing tag set to the more detailed SALMA Tag Set. The morphological tags of a 1000-word test text, chapter 29 of the Quranic Arabic Corpus, were automatically mapped to SALMA tags. Then, the mapped tags were proofread and corrected. The result of mapping and correction of the SALMA tagging of this corpus is a new Gold Standard for evaluating Arabic morphological analyzers and part-of-speech taggers with a detailed fine-grain description of the morphological features of each morpheme, encoded using SALMA tags.

11.2.4 Applications and Implementations

Morphosyntactic analysis is a very important and basic application of Natural Language Processing which can be integrated into a wide range of NLP applications. Arabic has many morphological and grammatical features, including sub-categories, person, number, gender, case, mood, etc. More fine-grained tag sets are often considered more appropriate. The additional information may also help to disambiguate the (base) part of speech.

The SALMA – Tagger is an open-source fine-grain morphological analyzer for Arabic text which puts together the developed resources (*i.e.* mainly the SALMA – ABCLexicon) and standards (the SALMA – Tag Set). It also depends on pre-stored lists (*i.e.* prefixes, suffixes, roots, patterns, function words, broken plurals, named entities, etc.) which were extracted from traditional grammar books. The morphological analyzer was developed to analyze the word and specify its morphological features. It uses a tokenization scheme for Arabic words that distinguishes between five parts of a word's morphemes as defined by the SALMA – Tag Set. Each part is given a fine-grained SALMA Tag that encodes 22 morphosyntactic categories of the morpheme (or possibly multiple tags if the part has multiple clitics or affixes). The SALMA – Tagger consists of several modules which can be used independently to perform a specific task such as root extraction, lemmatizing and pattern extraction. Or, they can be used together to produce full detailed analyses of the words.

The SALMA – Tagger was evaluated on a sample of Arabic text to measure its accuracy and applicability for use in text analytics tasks. It was also practically evaluated by applying the SALMA – Lemmatizer and Stemmer to lemmatize the 176-million word Arabic Internet Corpus (AIC) (section 10.2). This application measured the performance aspects of the SALMA - Tagger such as speed, memory and ability to perform the desired analysis tasks. Two main challenges arose during the lemmatizing of the AIC:

- **Speed:** which is solved by using the NGS to lemmatize the divided parts of the AIC in parallel giving a massive reduction in execution time.
- **Spelling errors:** which are solved by using the SALMA-Tokenizer to detect and correct the spelling errors that appear in the AIC due to poor word processing tools used in authoring web pages.

The second application is a proposal about how these resources, standards and tools can be used as a language engineering toolkit for Arabic lexicography. We reviewed the resources and tools which are used in modern lexicography, and we showed that the resources, proposed standards, and tools developed constitute a toolkit for constructing Arabic monolingual and bi-lingual dictionaries (section 10.3).

11.2.5 Evaluation

The evaluation for the SALMA – Tagger showed that evaluation methodologies for morphological analyzers are not standardized yet. Therefore, we developed agreed standards for evaluating morphological analyzers for Arabic text, based on our experiences and participation in two community-based evaluation contests: the ALECSO/KACST initiative for developing and evaluating morphological analyzers; and the MorphoChallenge 2009 competition. The guideline recommendations, evaluation specifications and procedures, and evaluation metrics were reused to generate a global standard for evaluating morphological analyzers for Arabic text. The developed standards were applied when evaluating the SALMA – Tagger.

The developed evaluation standards depend on using gold standards for evaluating morphological analyzers for Arabic text. A reusable general purpose gold standard (the SALMA – Gold Standard) was constructed to evaluate various morphological analyzers for Arabic text and to allow comparisons between the different analyzers. The SALMA – Gold Standard is adherent to standards, and enriched with fine-grained morphological information for each morpheme of the gold standard text samples. The detailed information is: the input word, its root, lemma, pattern, word type and the word's morphemes. For each of the word's morphemes, the morpheme type is classified into proclitic, prefix, stem, suffix and enclitic, and a fine-grain SALMA Tag which encodes 22 morphological feature categories of each morpheme, is also included.

The SALMA – Gold Standard contains two text samples of about 1000-words each representing two different text domains and genres of both vowelized and non-vowelized text taken from the Qur'an – chapter 29 representing Classical Arabic, and from the CCA representing Modern Standard Arabic. The SALMA – Gold Standard is stored using different standard formats (*i.e.* XML files, tab-separated column files, HTML and colour-coded format) to allow wider reusability.

The evaluation using the SALMA – Gold Standard focused on measuring the prediction accuracy of the 22 morphological features encoded in the SALMA – Tags for each of the gold standard's text samples morphemes. The evaluation aimed to answer the following questions:

- Is fine-grained morphological analysis for Arabic text practical?
- Can traditional Arabic grammar be leveraged to inform the knowledge-base for predicting the attribute values of the morphological feature categories?
- How can accuracy metrics report usefully for potential users who will use/reuse the SALMA – Tagger or parts of it?
- How are morphological feature categories related to each other (*i.e.* what interdependencies exist between the morphological features categories)?

The results show that 53.50% of the Qur'an text sample morphemes and 71.21% of the CCA text sample were correctly tagged using "exact match" of the gold standard's morpheme tags, but some of the errors were very minor such as replacing '**?**' by '-'. These results of applying the SALMA – Tagger answer the first question and show that fine-grained morphological analysis for Arabic text is practical. The results show the applicability of the SALMA – Tagger to process different types of text types, domains and genres of both vowelized and non-vowelized Arabic text. The SALMA – Tagger can be used to POS-tag Arabic text corpora and to provide detailed fine-grained analysis for each morpheme of the corpus words.

Moreover, these general results and the individual accuracy rates reported for each morphological feature show that the linguistically-informed knowledge-based system for predicting the values of the morphological feature categories is applicable to Arabic morphological analysis. The traditional Arabic grammar rules are leveraged to inform and construct the knowledge-based system for predicting the attribute values of the morphological feature categories.

The evaluation reported the accuracy, recall, precision, f1-score and the confusion matrix for each morphological feature category. The individual category accuracy results are useful for users who will use/reuse the SALMA – Tagger or parts of it, to know in advance the prediction accuracy of the attributes of each morphological feature category. Prediction accuracy was high for 15 morphological feature categories: namely, 98.53%-100%for the CCA test sample and 90.11%-100% for the Qur'an test sample. These categories are: main part-of-speech; subcategory of verb; subcategory of particle; subcategory of other (residual); punctuation; definiteness; voice; emphasized and non-emphasized; transitivity; declension and conjugation; unaugmented and augmented; number of root letters; verb roots; and noun finals.

The remaining 7 morphological feature categories, namely: the subcategory of noun; gender; number; person; inflectional morphology; case or mood; case and mood marks; and the morphological feature of rational, achieved slightly lower prediction accuracy: 81.35%-97.51%for the CCA test sample and 74.25%-89.03% for the Qur'an test sample.

Insights gained from this evaluation process for the morphological feature categories of Arabic words have been investigated in terms of the main background knowledge used for prediction and are as follows:

- The prediction of the main part-of-speech of a word's morphemes depends on both maintaining agreement between the word's affixes and clitics and the patterns dictionaries. Main part-of-speech information is provided in the clitics and affixes dictionaries and the patterns dictionary.
- The prediction of the part-of-speech subcategory of noun was not easy for the Qur'an text sample due to the nature of Quranic Arabic. The Qur'an text sample has repeated examples of proper nouns of historical persons and places. One characteristic of MSA text is the frequent use of relative nouns such as الْثَّقَافِيّ *aṯ-ṯaqāfī* 'cultural' and gerunds of profession such as الْوَطَنِيَّة *al-waṭaniyya*^{h} 'nationalism' where the rule for predicting these attributes is simple.
- The prediction of verbs depends on the analysis of the prefixes and suffixes and the matching of the stem morpheme with a patterns dictionary entry.
- Most particles are stored in the function words list. However, some of the particles of the Qur'an text sample are complex particles which consist of more than one morpheme such as أَوَلَمْ *'a-wa-lam* 'and not', which consists of three morphemes.
- The prediction of these affixes depends on matching the morphemes of the analyzed word with the entries of the clitics and affixes dictionaries. Ambiguous clitics can be classified into different categories.
- The prediction of punctuation is done in the tokenization step. Special characters used in the MSA text which are not standard punctuation marks are given a special tag '**o**' at position 6 of the tag string.
- The morphological features of gender, number and person are related to each other and share the same prediction methodology which depends on suffix analysis. Contextual rules that define agreement between the verb and its doer (the subject of

the sentence) are needed to support the prediction of these features when the affixes are ambiguous and cannot provide enough prediction information.

- The prediction of the morphological feature of inflectional morphology for verbs depends on the part-of-speech subcategory of verbs and analysis of suffixes for imperfect verbs to determine whether the verb is conjugated or invariable.
- The disambiguation of nouns into declined and invariable depends on applying many rules that deal with the part-of-speech subcategory of nouns, noun finals and patterns. These rules classify nouns into fully-declined or non-declined.
- The prediction of the morphological feature of case and mood depends on the result of the prediction of the morphological feature of inflectional morphology, such that a declined noun has case (*i.e.* nominative, accusative and genitive) and a conjugated verb has mood (*i.e.* indicative, subjunctive, and imperative or jussive), while case or mood is not applicable to invariable nouns and verbs.
- The prediction of a noun's case investigates the proclitics attached to the beginning of the noun which might affect the case and its syntactic mark such as prepositions and jurative particles. Prediction rules also investigate the dual and plural suffixes which change according to the case of the noun.
- Rules for predicting the case or mood, and case and mood marks for singular and broken plural nouns depend on the short vowel (*i.e.* the syntactic mark) that appears on the end of the word. The absence of short vowels and contextual rules that deal with nouns according to their context (*i.e.* subject or object) increases the potential of wrong prediction especially for singular and broken plural nouns.
- Determining the morpheme that carries the syntactic mark of the word is not an easy task and needs more investigation and standardization. Defining the morpheme that carries the syntactic mark has an impact on the development of the syntactic parsers for Arabic text.
- Only a conjugated verb has mood. The prediction rules of mood depend on the part-of-speech subcategory of verb, such that mood is applicable to imperfect verbs and not applicable to perfect and imperative verbs. The rules also analyze the suffixes of the imperfect verb to determine the applicability of mood. The final rule of prediction depends on the short vowel.
- Interdependency is clear between the three morphological feature categories: inflectional morphology, case or mood, and case and mood marks.

- The prediction of the morphological feature of definiteness depends on the availability of the definite article ال as a proclitic for the analyzed noun.
- The prediction rules classify verbs into active verbs or passive verbs depending on the short vowel appearing on the first letter of the verb after removing proclitics. If a *ḍamma*^{h} does not appear on the verb's first letter, then it is classified as an active voice verb. Errors can happen in some cases where *ḍamma*^{h} appears on the first letter of active voice verbs. Applying prediction rules for the morphological feature of voice that depend on the patterns rather than the short vowel of the first letter of the verb will increase the prediction accuracy.
- Prediction rules for classifying verbs into emphasized or non-emphasized depend on the part-of-speech subcategory of the verb. Perfect verbs are always non-emphasized while imperfect and imperative verbs can be emphasized. The prediction rules also investigate the suffixes of the verb. Emphasized verbs contain the emphatic *nūn* as a suffix.
- The prediction rules for the morphological feature of transitivity depend on matching the analyzed verb with one verb stored in the lists of doubly transitive and triply transitive verb lists. The singly transitive verb attribute is the default value for the morphological feature of transitivity. The absence of contextual rules for predicting the attributes of the morphological feature of transitivity increases the potential for making prediction mistakes. On the other hand, suffix pronoun analysis can capture some attributes of this morphological feature.
- Classifying words into rational or irrational depends on the semantics of the word itself and its context, which determines agreements between sentence parts such as verb-subject agreement and adjective-noun agreement. A comprehensive dictionary which includes Rational information for each dictionary entry is needed to determine the correct attribute value of rational for nouns.
- The morphological feature of declension and conjugation is applied to nouns, verbs and particles. The prediction rules of the values of declension and conjugation of nouns depend on the part-of-speech subcategories. Including declension and conjugation information in the Arabic dictionary will increase the correct prediction of attributes for this morphological feature.
- The prediction rule of unaugmented and augmented attributes subtracts the length of the root from the length of the analyzed word. The prediction rule of the

attributes of the number of root letters depends on the length of the root. The prediction rules of the morphological feature of verb roots depend on the nature of the root letters: whether they are consonants, containing *hamza*^{h}, or whether they contain one vowel or two.

- The rules for predicting the value of the morphological feature of Noun Finals mainly depends on the long stem and the root of the analysed word which checks the final letters of the long stem against a set of conditions that classify nouns into 6 subcategories. Knowing the value of the Noun Finals feature helps in specifying other features such as the morphological features of Inflectional Morphology and Case and Mood Marks.

To summarize, the absence of contextual rules, the absence of short vowels, the interdependency between some morphological features, and the number of attributes of a certain morphological feature increase the potential of prediction errors for some morphological feature categories. To improve the accuracy of predicting the attributes of the morphological feature categories, contextual rules can be implemented as a second pass. Some morphological feature categories such as rational depend on the semantic nature of the analyzed word itself. Providing rationality information for Arabic dictionary entries and reusing this information in morphological analyzers will increase prediction accuracy. Moreover, updating the dictionaries which are used by the SALMA – Tagger by increasing their coverage will increase prediction accuracy.

11.3 Future work

This section explores four possible applications of the SALMA – Tagger, and the resources developed in this thesis to future work projects: improving the SALMA – Tagger; a syntactic parser; the international corpus of Arabic ICA; and as a tool for annotating phrase-breaks and other prosodic features in a corpus. The Tagger can also be integrated with similar level applications that combine two systems together to maximise the capabilities of both systems.

11.3.1 Improving the SALMA – Tagger

The evaluation of the SALMA – Tagger showed that the prediction rules for 7 morphological feature categories (namely: the subcategories of noun, gender, number, person, inflectional morphology, case or mood, case and mood marks, and the morphological feature of rational) achieved a slightly lower than expected prediction accuracy: 81.35%-97.51% for the CCA test sample and 74.25%-89.03% for the Qur'an test sample. The lower accuracy achieved with the 7 morphological feature categories was due to:

- The absence of contextual rules in the SALMA – Tagger, which treats words out of their context.
- The absence of short vowels in text, and especially MSA text. This makes the prediction of the attributes of some morphological features difficult.
- The interdependency between some morphological features such as the morphological features of inflectional morphology, case and mood, and case and mood marks. The decreases the accuracy of the dependent features by propagating errors from one feature to another.
- Prediction errors. These increase, if the number of attributes of a certain morphological feature increases.

To improve the accuracy of predicting the attributes of the morphological feature categories, three practical solutions can be implemented as a second phase of the development of the SALMA – Tagger. These solutions are:

- Contextual rules, which can be implemented as a second pass. The contextual rules will also help in reducing the number of candidate analyses of the analyzed words by excluding the analyses that do not satisfy certain contextual rules.
- Enriching Arabic dictionary entries with fine-grain morphological information such as gender, number, inflectional morphology, rationality, and transitivity and reusing this information in morphological analyzers. This will increase the accuracy of prediction.
- Updating the dictionaries and the linguistic lists which are used by the SALMA – Tagger by increasing their coverage. This will increase prediction accuracy.

The morphological feature categories such as rational depend on the semantic nature of the analyzed word itself. Therefore, the development of the morphological analyzer of Arabic text is an ongoing project that will be integrated in different levels of applications (*i.e.* phonology, syntax and semantics) into these application levels on an information sharing basis. The morphological analyzer which is integrated to these levels will provide detailed morphological information about words and at the same time will benefit from feedback from these levels of analysis.

11.3.2 A Syntactic Analyzer (parser) for Arabic Text

The SALMA - Tagger generates all possible analyses for the analyzed words out of their context. A disambiguation tool that selects a suitable analysis within a certain context is needed. A syntactic analyzer (parser) is required as a tool for automatically annotating the Arabic corpus with the correct syntactic information. It is also required to build the syntactic parse trees for Arabic corpus sentences. The aim of this project is to build a syntactic analyzer (parser) to annotate the Arabic corpus with the syntactic information for each word in the corpus. The aim of this corpus annotation is to create a Treebank corpus and a dependency Treebank of Arabic. These tools and standards will be tied into a specific corpus, but they can be reused to annotate any Arabic corpus to meet the needs of updating the contents of any Arabic corpus or building new Arabic corpora for specific purposes.

The syntactic analyzer for Arabic text will depend on both the linguistic information extracted from traditional Arabic grammar books and the use of machine leaning algorithms such as HMM and decision trees, to build the disambiguation tool that selects the appropriate morphosyntactic analysis of the word in its context.

The following resources and tools are needed to develop a syntactic analyzer (parser) for Arabic text:

- **Morphological analysis tool and standard**: The SALMA – Tagger and the SALMA – Tag Set are essential prerequisites for the syntactic parser, providing a detailed morphological analysis of all morphemes of words in the Arabic corpus.
- **Linguistic model of Arabic sentence structure and the syntactic tag set:** The methodology used to develop the fine-grain morphological features tag set, the SALMA – Tag Set, can be reused to develop a syntactic tag set that is based on traditional Arabic grammar. The syntactic tag set of Arabic will specify the types of Arabic sentences and phrases (*i.e.* verbal sentences, nominal sentences and phrases); the components of Arabic sentences and phrases (*i.e.* verb, subject, object and complement); the linguistic attributes (*i.e.* syntactic features) of each sentence component; and the forms of agreement between the sentence components.
- **Representative Open Source Arabic Corpus:** Very few open source Arabic corpora are available which can be used as seeds for the new representative open source Arabic corpus. Such available open source corpora are the Corpus of Contemporary Arabic (Al-Sulaiti and Atwell 2006), the Corpus of Traditional Arabic Dictionaries (Sawalha and Atwell 2010a), and the Quranic Arabic Corpus (Dukes et al. 2010). The first two corpora do not have any morphosyntactic annotation, but the Quranic Arabic Corpus is annotated with morphosyntactic analyses which can be reused by mapping the annotation to our standards.

- **Evaluation Standards:** The standard development methodology of the SALMA – Tagger can be reused to develop standards and guidelines to evaluate the syntactic parser. The evaluation standards will mainly depend on developing a gold standard for evaluation. The gold standard aims to be widely used by the Arabic NLP community and to be general purpose. It will be used as a standard for comparing different Arabic syntactic parsers. Therefore, the construction of the gold standard should follow specific guidelines for size, the corpora used in constructing it and its format. The gold standard should be large enough to cover most of the morphosyntactic phenomena that morphosyntactic analyzers have to handle. The corpus used to construct the gold standard should be representative, including text of different text domains, formats and genres, with both vowelized and non-vowelized Arabic text. The format of the gold standard will specify what information it has to include and in which format it has to be stored.
- **The Project Collaborators:** this project is part of a future project that meets our interest in morphosyntactic analysis for Arabic text. Initial agreements have already been made between the project collaborators: Majdi Sawalha and Dr. Eric Atwell (Arabic Language Engineering team at the University of Leeds, UK); Professor Azzeddine Mazroui (Natural Language Processing team at the University of Mohammed I, Morocco); and Dr. AlMoutaz Bi-Allah Al-Sa'eed (Cairo University, Egypt).

11.3.3 Open Source Morphosyntactically Annotated Arabic Corpus

The main objective in developing the SALMA – Tagger and the syntactic parser (previous section) is to annotate the Arabic corpus with detailed morphosyntactic analyses of each word in the corpus. There is as yet no open source Arabic Corpus with full morphosyntactic annotation. The construction of such a corpus aims to advance Arabic NLP studies. The survey of Arabic corpora in section 2.2 showed that there are only two open source Arabic corpora eligible for morphosyntactic annotation. These existing corpora are the Corpus of Contemporary Arabic (Al-Sulaiti and Atwell 2006) and the Quranic Arabic Corpus (Dukes et al. 2010). The CCA is an MSA corpus of raw text, while the QAC represents Classical Arabic which has morphological and syntactic annotations. The Corpus of Traditional Arabic Dictionaries (Sawalha and Atwell 2010a) developed in this thesis is a special corpus of raw text which represents text from a period of 1,300 years.

A representative open-source Arabic corpus will be constructed by selecting the text from different genres and formats including both vowelized and non-vowelized Arabic text. The previously mentioned open-source corpora can represent a seed for our corpus. Each document of the corpus will be described by adding information of date, author, country, topic/genre, vowelization information, source, etc. These descriptions can be used to train text classifiers.

An annotation tool and annotation guidelines are needed to achieve our objective. The design of the annotation program should take into account the choices for the annotator to manually annotate the corpus and to correct the automatically tagged text by selecting the appropriate morphological analysis resulting from the morphological analyzer and the ability to correct the syntactic analysis generated automatically using the syntactic parser. The annotation program should have capabilities for searching for morphosyntactic patterns in the annotated text, and for visualizing the sentences and the syntactic annotations as parse trees in a readable and representative way, with the added capacity to access parts of the parse tree and make corrections if necessary. The annotation program should also have an intelligent design that facilitates the annotation process.

Some open source annotation tools already exist such as GATE (http://gate.co.uk). Our annotation tools and analyzers can be integrated into GATE, which can help widen usage of the tools and standards that will be produced in this project.

The Morphosyntactic Analyses Training Corpus of Arabic is useful for developing machine learning algorithms. The latter requires a training corpus of Arabic text annotated with the appropriate morphosyntactic analyses. Parts of the open source Arabic corpus can be manually/semi-automatically annotated using the developed tools to train the machine learning algorithms that will be used to build statistical models for morphosyntactic analyses of Arabic text corpora.

The project collaborators are: Majdi Sawalha and Dr. Eric Atwell (Arabic Language Engineering team at the University of Leeds, UK); Professor Azzeddine Mazroui (Natural Language Processing team at the University of Mohammed I, Morocco); and Dr. Al-Moutaz Bi-Allah Al-Sa'eed (Cairo University, Egypt).

11.3.4 Arabic Phonetics and Phonology for Text Analytics and Natural Language Processing Applications

This research applies Text Analytics techniques honed on English for resource creation and corpus-based exploration of Arabic speech and language for Arabic Natural Language Processing (NLP) applications. Such techniques depend on a *corpus* or *sample* of naturally occurring language texts capturing empirical data on the phenomena being studied, for example prosodic-syntactic patterns in the vicinity of phrase breaks or perceived pauses in the speech stream. Computational analysis of text also requires gold-standard (human) annotation of target phenomena and other linguistic knowledge inherent in text, such as part-of-speech (POS) categories. The approach is then to mine the annotations as well as plain text.

Collaborators on this project have research interests and expertise in Corpus Linguistics, Artificial Intelligence, Text Analytics, and Lexicography for English and Arabic (Brierley and Atwell 2008; Dukes et al. 2010; Sawalha and Atwell 2010b). One area to focus on is the prosody-syntax interface: this approach builds on previous work on English prosody and Text Analytics (Brierley and Eric 2010) and involves mining rhythmic junctures to derive boundary templates and phrasing strategies from Arabic texts as diverse as transcribed speech recordings (*e.g.* Modern Standard Arabic newsreel), Classical Arabic poetry and Quranic Arabic. Some editions of the Quran have fine-grained prosodic-boundary annotations, inviting comparison with conventions for British and American English (*e.g.* ToBI (Beckman and Hirschberg 1994)). Collaborators will report on an essential pre-requisite for this approach: an Arabic pronunciation lexicon and automatic text annotation tool modelled on a similar tool for English (Brierley and Atwell 2008). The SALMA patterns dictionary enriched with syllable and primary stress information, and the SALMA Tagger and Vowelizer are required as part of the language-engineering toolkit for this project.

The project plans to represent significant boundary and phrasing patterns thus derived as categorical features for machine learning and to test these in phrase break models for Arabic Text-to-Speech Synthesis (TTS). Enhanced performance in TTS relates to the longer-term goal of achieving more realistic speech in virtual characters for both English and Arabic HCI (Human-Computer Interaction), with diverse applications in education, therapy and entertainment.

The collaborators on this project are: Majdi Sawalha, Claire Brierley and Eric Atwell (Arabic Language Engineering team at the University of Leeds, UK).

11.4 Summary: PhD impact, originality, and contributions to research field

Our research into morphosyntactic analysis of Arabic text corpora involves original scientific research, and focuses on the question of how to widen the scope of Arabic morphosyntactic analyses, to develop an NLP toolkit that can process Arabic text in a wide range of formats, domains, and genres, of both vowelized and non-vowelized Arabic text. This final section presents a brief summary of research contributions and achievements of this PhD.

11.4.1 Utilizing the Linguistic Wisdom and Knowledge in Arabic NLP

The inspiration behind this research is centuries-old linguistic wisdom and knowledge captured and readily available in traditional Arabic grammars and lexicons. The knowledge can be utilized in an Arabic NLP toolkit which can be accessed, standardized, reused and implemented in Arabic natural language processing. The detailed knowledge is applicable to both Classical and Modern Standard Arabic and can be used to restore orthographic (*e.g.* short vowels) and morphosyntactic features which signify important linguistic distinctions. Fine-grained morphosyntactic analysis is possible, achievable and advantageous in processing Arabic text. Enriching the text with linguistic analysis will maximize the potential for corpus re-use in a wide range of applications. We foresee the advantage of enriching the text with part-of-speech tags of very fine-grained grammatical distinctions, which reflect expert interest in syntax and morphology, but not specific needs of end-users, because end-user applications are not known in advance.

The objective of the thesis has been achieved through developing a novel language-engineering toolkit for morphosyntactic analysis of Arabic text, the SALMA – Tagger. The SALMA – Tagger combines sophisticated modules that break down the complex morphological analysis problem into achievable tasks which each address a particular problem and also constitute stand-alone units. The novel language-engineering tool depends on two novel and original resources and standards (i) the SALMA – Tag Set and (ii) the SALMA – ABCLexicon.

11.4.2 Dimensions of Contributions to Arabic NLP

This research has contributed to Arabic NLP in three dimensions: Resources, standards and tools (*i.e.* practical software). The following is a list of the contributions classified into the three dimensions:

D. Resources

1. The SALMA – ABCLexicon: a novel broad-coverage lexical resource constructed by extracting information from many traditional Arabic lexicons, constructed over 1,300 years, of disparate formats.
2. The Corpus of Traditional Arabic Lexicons: a special corpus of Arabic which is compiled from the text of 23 traditional Arabic lexicons that cover a period of 1,300 years and shows the evolution of Arabic vocabulary. It contains about 14 million word tokens and about 2 million word types.
3. The morphological lists of the SALMA – Patterns Dictionary and the SALMA – Clitics and Affixes lists.

4. The several linguistic lists that are used by the SALMA – Tagger such as: function words list, named entities lists, broken plural list, conjugated and non-conjugated verbs list, and transitive verbs lists.
5. The Lemmatized version of the Arabic Internet Corpus.

E. Proposed standards

16. The SALMA – Tag Set: a morphological features tag set for Arabic text which captures long-established traditional morphological features of Arabic, in a compact yet transparent notation.
17. The SALMA – Gold Standard for evaluating morphological analyzers for Arabic text.
18. The MorphoChallenge 2009 Qur'an Gold Standard.
19. Proposed standards for developing morphological analyzers for Arabic text.
20. Proposed standards for evaluating morphological analyzers for Arabic text.

F. Tools (practical software)

1. The SALMA – Tokenizer, which tokenizes the input text files and identifies the Arabic words, spell-checks and corrects the words, and identifies the words' parts or morphemes.
2. The SALMA – Lemmatizer and Stemmer, which extracts the lemma and the root of the analysed word.
3. The SALMA – Pattern Generator, which is responsible for matching the word with its pattern.
4. The SALMA – Vowelizer, which is responsible for adding the short vowels to the analysed words.
5. The SALMA – Tagger module, which predicts the fine-grained morphological features for each of the analysed word's morphemes.

Finally, a potential future application of these contributions is as a language-engineering toolkit for Arabic lexicography to construct Arabic monolingual and bi-lingual dictionaries (Section 10.3).

11.4.3 Impact

Journal and conference papers resulting from this thesis have addressed a range of research communities: Computational linguistics, Arabic Natural language processing, Language Resources and Evaluation, Linguistic studies (word structure analysis), and Lexicography. These publications have already been cited by other researcher such as Alotaiby, Alkharashi et al. (2009), Kurimo, Virpioja et al. (2009), Altabbaa, Al-Zaraee et al. (2010), Hamada 2010; Harrag, Hamdi-Cherif et al. (2010), Yusof, Zainuddin et al. (2010), Al-Jumaily, Martínez et al. (2011), and Hijjawi, Bandar et al. (2011).

References

Al-Bawaab, M. 2009. مواصفات نظام التحليل الصرفي في اللغة العربية Specifications of Arabic Morphological Analyzer. *Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Science and Technology (KACST) and Arabic Language Academy.*, Damascus, Syria.

Al-Ghalayyni. 2005. جامع الدروس العربية *"Jami' Al-Duroos Al-Arabia"*. Saida - Lebanon: Al-Maktaba Al-Asriyiah "المكتبة العصرية".

Al-Jumaily, H., Martínez, P., Martínez-Fernández, J., and Goot, E.v.d. 2011. *A real time Named Entity Recognition system for Arabic text mining*. Language Resources and Evaluation.1-21.

al-Saydawi, Y. 2006. الكفاف: كتاب يعيد صوغ القواعد العربية *Sufficiency: A Book Reformulating Arabic Grammar*. Damascus, Syria: Dar Al-Fikr.

Al-Shalabi, R. 2005. Pattern-based Stemmer for Finding Arabic Roots. *Information Technology Journal* 4(1): 38-43.

Al-Shalabi, R., Kanaan, G. and Al-Serhan, H. 2003. New approach for extracting Arabic roots. *in ACIT '2003: Proceedings of The 2003 Arab conference on Information Technology*, Alexandria, Egypt.

Al-Shammari, E. and Lin, J. 2008. A novel Arabic lemmatization algorithm. *AND '08: Proceedings of the second workshop on Analytics for noisy unstructured text data*, pp. 113--118. Singapore: ACM.

Al-Shamsi, F. and Guessoum, A. 2006. A Hidden Markov Model-Based POS Tagger for Arabic. *8es Journees internationales d'Analyse statistique des Donnees Textuelles*.

Al-Sughaiyer, I. A. and Al-Kharashi, I. A. 2002. Rule Parser for Arabic Stemmer *Text, Speech and Dialogue*, pp. 11-18. Springer Berlin / Heidelberg.

Al-Sughaiyer, I. A. and Al-Kharashi, I. A. 2004. Arabic morphological analysis techniques: A comprehensive survey. *Journal of the American Society for Information Science and Technology* 55(3): 189-213.

Al-Sulaiti, L. and Atwell, E. 2004. Designing and developing a corpus of contemporary Arabic *TALC 2004: Proceedings of the sixth Teaching And Language Corpora conference*, pp. 92-93.

Al-Sulaiti, L. and Atwell, E. 2005. Extending the corpus of contemporary Arabic. *Proceedings of Corpus Linguistics 2005*.

Al-Sulaiti, L. and Atwell, E. 2006. The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics* 11: 135-171.

ALECSO. 2008a. Arabic Derivation System.

ALECSO. 2008b. Sarf - Arabic Morphology System The Arab League Educational, Cultural and Scientific Organization (ALECSO).

Ali, A. S. M. 1987. *A Linguistic Study of the development of Scientific Vocabulary in Standard Arabic*. London and New York: Kegan Paul International.

Alotaiby, F., Alkharashi, I. A. and Foda, S. G. 2009. Processing Large Arabic Text Corpora: Preliminary Analysis and Results. *Paper presented to the Proceedings of the Second International Conference on Arabic Language Resources and Tools,* Cairo, Egypt, 2009.

Alqrainy, S. 2008. A Morphological-Syntactical Analysis Approach For Arabic Textual Tagging. *2008*, pp. 197. Leicester, UK: De Montfort University.

AlSerhan, H. and Ayesh, A. 2006. A Triliteral Word Roots Extraction Using Neural Network For Arabic. *IEEE International Conference on Computer Engineering and Systems (ICCES06)*, pp. 436-440. Cairo, Egypt.

Altabbaa, M., Al-Zaraee, A. and Shukairy, M. A. 2010. An Arabic Morphological Analyzer and Part-Of-Speech Tagger Qutuf 'قُطُوف'. *Faculty of Informatics Engineering*, pp. 100. Damascus: Arab International University.

Atkins, B. T. S. and Rundell, M. 2008. *The Oxford guide to practical lexicography* Oxford ; New York Oxford University Press.

Attia, M. A. 2007. Arabic Tokenization System. *ACL-Workshop on Computational Approaches to Semitic Languages*, Prague.

Attia, M. A. 2008. Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation. *Faculty of Humanities*, pp. 279. Manchester: University of Manchester.

Atwell, E. 2007. A cross-language methodology for corpus Part-of-Speech tag-set development *Proceedings of Corpus Linguistics 2007.*

Atwell, E. 2008. Development of tag sets for part-of-speech tagging. In A. Ludeling and M. Kyto (eds.). *Corpus Linguistics: An International Handbook, Volume 1*, pp. 501-526 Mouton de Gruyter.

Atwell, E., Demetriou, G., Hughes, J., Schiffrin, A., Souter, C. and Wilcock, S. 2000. A comparative evaluation of modern English corpus grammatical annotation schemes. *ICAME Journal, International Computer Archive of Modern and medieval English, Bergen* 24: 7-23.

Atwell, E. and Roberts, A. 2007. CHEAT: combinatory hybrid elementary analysis of text *Proceedings of CL'2007 Corpus Linguistics Conference*.

Baayen, R. H., Piepenbrock, R. and Rijn, H. v. 1995. The CELEX Lexical Database. Release 2.

Baker, P., Hardie, A. and McEnery, T. 2006. *A Glossary of Corpus Linguistics*. Edinburgh, UK: Edinburgh University Press.

Bamman, D. and Crane, G. 2008. Building a Dynamic Lexicon from a Digital Library. *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008)*, Pittsburgh.

Banko, M. and Brill, E. 2001. Scaling to Very Very Large Corpora for Natural Language Disambiguation. *39th annual meeting & 10th conference of the European Chapter :* , Toulouse, 9-11 July 2001 Morgan Kaufman Publishers, [S. l.], INCONNU (2001) (Monographie).

Banko, M. and Moore, R. C. 2004. Part of Speech Tagging in Context. *20th International Conference on Computational Linguistics (Coling 2004), pp. 556-561*, Geneva, Switzerland: International Conference on Computational Linguistics.

Beckman, M. E. and Hirschberg, J. 1994. The ToBI Annotation Conventions.

Beesley, K. R. 1996. Arabic finite-state morphological analysis and generation. *Proceedings of the 16th conference on Computational linguistics - Volume 1*, Copenhagen, Denmark: Association for Computational Linguistics.

Beesley, K. R. 1998. Arabic morphology using only finite-state operations. *Proceedings of the Workshop on Computational Approaches to Semitic Languages*, Montreal, Quebec, Canada: Association for Computational Linguistics.

Benajiba, Y., Diab, M. T. and Rosso, P. 2008. Arabic named entity recognition using optimized feature sets. *Proceedings of the Conference on Empirical Methods in Natural language Processing, EMNLP'08*, pp. 248-293. Honolulu, Hawaii: Association for Computational Linguistics.

Benmamoun, E. 1999. Arabic morphology: The central role of the imperfective. *Lingua* 108.175-201.

Bird, S., Klein, E., and Loper, E. 2009. Natural Language Processing with Python (1st edition edn.: O’Reilly Media, Inc.).

Black, W. J. and El-Kateb, S. 2004. A Prototype English-Arabic Dictionary Based on WordNet. *The Second Global Wordnet Conference 2004* Brno, Czech Republic, January 20-23, 2004, pp. 67-74.

Borin, L. 2000. Something Borrowed, Something Blue: Rule-Based Combination of POS Taggers. *Proceedings of Second International Conference on Language Resources and Evaluation (LREC)*, pp. 21-26. Athens, Greece.

Boudlal, A., Belahbib, R., Lakhouaja, A., Mazroui, A., Meziane, A. and Bebah, M. O. A. O. 2011. A Markovian Approach for Arabic Root Extraction. *The International Arab Journal of Information Technology* 8(1): 91-98.

Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., Bebah, M. O. A. O. and M.Shoul. 2010. Alkhalil Morpho Sys: A Morphosyntactic analysis system for Arabic texts. *IJCSI International Journal of Computer Science Issues*.

Brierley, C. and Atwell, E. 2008. ProPOSEL: a human-oriented prosody and PoS English lexicon for machine learning and NLP. Proceedings of COLING 2008, CogALex Workshop on Cognitive Aspects of the Lexicon.

Brierley, C. and Eric, A. 2010. Holy smoke: vocalic precursors of phrase breaks in Milton's Paradise Lost. *Literary and Linguistic Computing Journal* 25(2).

Buckwalter, T. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, catalog number LDC2002L49 and ISBN 1-58563-257-0.

Buckwalter, T. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. Linguistic Data Consortium, catalog number LDC2004L02 and ISBN 1-58563-324-0.

Cachia, P. 1973. *The monitor : a dictionary of Arabic grammatical terms : Arabic-English, English-Arabic* / compiled by Pierre Cachia. Beirut, Librairie du Liban.

Chan, P. K. and Stolfo, S. J. 1995. A Comparative Evaluation of Voting and Meta-learning on Partitioned Data. *Proceedings of International Conference on Machine Learning*, pp. 90-98.

Clark, A. 2007. Supervised and Unsupervised Learning of Arabic Morphology. In A. Soudi, A. v. Bosch and G. Neuman (eds.). *Arabic Computational Morphology*, pp. 181-200. Springer.

Dˇzeroski, S. s., Erjavec, T. z. and Zavrel, J. 2000. Morphosyntactic Tagging of Slovene: Evaluating Taggers and Tagsets. *Proceedings of the Second International Conference on Language Resources and Evaluation. ELRA*, pp. 1099-1104. Paris-Athens.

Dahdah, A. 1987. *A Dictionary of Arabic Grammer in Charts and Tables* " معجم قواعد اللغة العربيه – في جداول ولوحات ". Beirut, Lebanon: Librairie du Liban publisher.

Dahdah, A. 1993. *A dictionary of Arabic Grammatical nomenclature Arabic – English* " معجم لغة النحو العربي عربي-انكليزي". Beirut, Lebanon: Librairie du Liban publishers.

Dejean, H. 2000. How to Evaluate and Compare Tagsets? A Proposal. *Proceedings of the second international conference on Language Resources and Evaluation LREC 2000*, Ahens, Greece: European Language Resources Association (ELRA).

Diab, M. T., Hacioglu, K., and Jurafsky, D. 2004. Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks. *Paper presented to the Proceedings of HLT-NAACL 2004.*

Diab, M. T. 2007. Towards an Optimal POS Tag Set for Arabic Processing. *Proc RANLP.*

Dichy, J. 2001. On lemmatization in Arabic, A formal definition of the Arabic entries of multilingual lexical databases. *ACL/EACL 2001 Workshop on Arabic NLP*, Toulouse, France, Friday 6 July 2001.

Dichy, J. 2009. A basic method for assessing arabic morphological analysers : some crucial criteria. *Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Science and Technology (KACST) and Arabic Language Academy.*, Damascus, Syria.

Dichy, J. and Farghaly, A. 2003. Roots & patterns vs. stems plus grammar-lexis specifications: on what basis should a multilingual database centred on Arabic be built? *MT Summit IX -- workshop: Machine translation for semitic languages*, New Orleans, USA.

Dickinson, M. and Jochim, C. 2010. Evaluating Distributional Properties of Tagsets. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pp. 2522-2529. Valletta, Malta: European Language Resources Association (ELRA).

Dietterich, T. G. 2000. Ensemble Methods in Machine Learning. *Lecture Notes in Computer Science*, pp. 1-15.

Diwan, A.-H. 2004. المعجم النحوي لمفردات اللغة العربية *The Syntactic Lexicon of Arabic Words*. Aleppo, Syria: Fusselat Publishers.

Dror, J., Shaharabani, D., Talmon, R. and Wintner, S. 2004. Morphological Analysis of the Qur'an. *Literary and Linguistic Computing* 19(4): 431-452.

Duh, K. and Kirchhoff, K. 2005. POS Tagging of Dialectal Arabic: A Minimally Approach. *ACL-05, Computational Approaches to Semitic Languages Workshop Proceedings*, pp. 55-62. University of Michigan Ann Arbor, Michigan, USA.

Dukes, K., Atwell, E. and Sharaf, A.-B. M. 2010. Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank. *Language Resources and Evaluation Conference (LREC 2010)*, Valletta, Malta.

Dukes, K. and Habash, N. 2010. Morphological Annotation of Quranic Arabic. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta,19-21 May 2010.: European Language Resources Association (ELRA).

Dzeroski, S., Erjavec, T. and Zavrel, J. 2000. Morphosyntactic Tagging of Slovene: Evaluating Taggers and Tagsets. *Proceedings of Second International Conference on Language Resources and Evaluation (LREC)*, pp. 1099-1104.

Elghamry, K. 2010. Broken Plurals. http://sites.google.com/site/elghamryk/arabiclanguageresources.

Elkateb, S., Black, W. and Farwell, D. 2006. Arabic WordNet and the Challenges of Arabic. *Preceedings of The Challenge of Arabic for NLP/MT International Conference at The British Computer Society (BCS)*, London.

Elkateb, S. and Black, W. J. 2001. Towards the Design of English-Arabic Terminological Knowledge Base. *Proceedings of ACL 2000*, Toulouse, France:113-118.

Elliott, J. and Atwell, E. 2000. Is anybody out there?: the detection of intelligent and generic language-like features. *JBIS: Journal of the British Interplanetary Society* 53: pp.7-23.

Elworthy, D. 1995. Tagset design and inflected languages. *In 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL), From Texts to Tags: Issues in Multilingual Language Analysis SIGDAT Workshop*, pp. 1–10. Dublin.

Erjavec, T. 2010. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias (eds.). *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pp. 2544-2547. Valletta, Malta: European Language Resources Association (ELRA).

Escudero, G., Mhrquez, L. and Rigau, G. 2000. A Comparison between Supervised Learning Algorithms for Word Sense Disambiguation. *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning*, pp. 31-36. Lisbon, Portugal: Association for Computational Linguistics, Morristown, NJ, USA.

Eynde, V. E. and Gibbon, D. (eds.) 2000. *Lexicon development for speech and language processing*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Freeman, A. 2001. Brill's POS Tagger and a Morphology Parser for Arabic. *NAACL 2001 Student Rersearch Workshop, Lancaster University*.

Gasser, M. 2010. Expanding the Lexicon for a Resource-Poor Language Using a Morphological Analyzer and a Web Crawler. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pp. 342-347. Valletta, Malta: European Language Resources Association (ELRA).

Glass, K. and Bangay, S. 2005. Evaluating Parts-of-Speech Taggers for Use in a Text-to-Scene Conversion System. *SAICSIT '05: Proceedings of the 2005 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries*, pp. 20--28. White River, South Africa South African Institute for Computer Scientists and Information Technologists.

Gopal, M., Mishra, D. and Singh, D. P. 2010. Evaluating Tagsets for Sanskrit. *Sanskrit Computational Linguistics, Lecture Notes in Computer Science* 6465/2010: 150-161.

Habash, N. 2004. Large Scale Lexeme Based Arabic Morphological Generation. *JEP-TALN 2004, Session Traitement Automatique de l'Arabe*, Fès.

Habash, N., Faraj, R. and Roth, R. 2009. Syntactic Annotation in Columbia Arabic Treebank. *2nd International Conference on Arabic Language Resources & Tools MEDAR 2009*, Cairo, Egypt.

Habash, N. and Rambow, O. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. *Paper presented at the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Ann Arbor, Michigan.

Habash, N. and Roth, R. M. 2009. CATiB: The Columbia Arabic Treebank. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 221–224. Suntec, Singapore: 2009 ACL and AFNLP.

Habash, N. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Hadrich, L. B. and Chaâben, N. 2006. Analyse et désambiguïsation morphologiques des textes arabes non voyellés. *Actes de la 13ème édition de la conférence sur le Traitement Automatique des Langues Naturelles (TALN 2006)*, pp. 493-501. Belgique.

Hajič, J., Smrž, O., Zemánek, P., Šnaidauf, J. and Beška, E. 2004. Prague Arabic Dependency Treebank: Development in Data and Tools. *Proceedings of NEMLAR International Conference on Arabic Language Resources and Tools*, pp. 110–117. Cairo, Egypt.

Halteren, H. v., Zavrel, J. and Daelemans, W. 2001. Improving Accuracy in Word Class Tagging through the Combination of Machine Learning Systems. *Computational Linguistics* 27(2): pp199-229.

Hamada, S. 2009a. المحللات الصرفية للغة العربية "Morphological Analyzers for Arabic". *Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Science and Technology (KACST) and Arabic Language Academy.*, Damascus, Syria.

Hamada, S. 2009b. مقترح لمعايير وضوابط تقييم المحلّلات الصرفية A proposal for evaluating morphological analyzers for Arabic text. *Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Science and Technology (KACST) and Arabic Language Academy.*, Damascus, Syria. 26-28 April 2009.

Hamada, S. 2010. مقترح لمعايير وضوابط تقييم المحلّلات الصرفية Evaluation of the Arabic Morphological Analyzers *Proceedings of The Sixth International Computing science Conference ICCA*, Hammamet, Tunisia.

Hamado, A.-M. B., Belghayth, L. and Sha'baan, N. 2009. "الصرفي للغة العربية لمخبر "ميراكل MORPH, morphological analyzer for Arabic text developed at MIRACL Labs. *Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Science and Technology (KACST) and Arabic Language Academy.*, Damascus, Syria.

Hardie, A. 2003. Developing a tagset for automated part-of-speech tagging in Urdu. *Proceedings of the Corpus Linguistics 2003 conference.*, ed. by D. Archer, Rayson, P, Wilson, A, and McEnery, T. Department of Linguistics, Lancaster University.: UCREL Technical Papers Volume 16.

Hardie, A. 2004. The computational analysis of morphosyntactic categories in Urdu. pp. 477. Lancaster University.

Harmain, H. M. 2004. Arabic Part-of-Speech Tagging. *Paper presented at the The Fifth Annual U.A.E. University Research Conference*, United Arab Emirates.

Harrag, F., Hamdi-Cherif, A. and Al-Salman, A. S. 2010. Comparative Study of Topic Segmentation Algorithms Based on Lexical Cohesion: Experimental Results on Arabic Language. *The Arabian Journal for Science and Engineering* 35.138-202.

Haywood, J. A. and Nahmad, H. M. 1965. *A New Arabic Grammar of the Written Language*. London: Lund Humphries.

Hijjawi, M., Bandar, Z., Crockett, K. and Mclean, D. 2011. An Arabic Stemming Approach using Machine Learning with Arabic Dialogue System. *ICGST International Conference on Artificial Intelligence and Machine Learning (AIML-11)*, Dubai, UAE.

Hu, X. R. and Atwell, E. 2003. A survey of machine learning approaches to analysis of large corpora. In D. Archer, Rayson, P, Wilson, A & McEnery, T (ed.). *Proceedings of SProLaC: Workshop on Shallow Processing of Large Corpora*, pp. 657-661 Lancaster University.

Ingulfsen, T., Burrows, T. and Buchholz, S. 2005. Influence of Syntax on Prosodic Boundary Prediction. *Proceedings, INTERSPEECH 2005*. 1817-1820.

Johansson, S., Atwell, E., Garside, R. and Leech, G. 1986. *The Tagged LOB Corpus*. Bergen, Norway: Norwegian Computing Centre for the Humanities.

Jurafsky, D. and Martin, J. H. 2008. *Speech and Language Processing*. New Jersey: Prentice Hall.

Kammoun, N. C., Belguith, L. H. and Hamadou, A. B. 2010. The MORPH2 new version: A rubust morphological analyzer for Arabic text. *JADT 2010: 10th International Conference on Statistical Analysis of Textual Data*, SAPIENZA, Italy.

Khafaji, R. 2001. Punctuation Marks in original Arabic texts. *Zeitschrift fur Arabische Linguistik* 40(2001): 7-24.

Khalil, H. 1998. *Dirasat fi al-lughah wa al-ma'ajim* " دراسات في اللغة والمعاجم " *Studies of language and lexicons Beiru*t, Lebanon: Dar al-nahdhah al-arabiah.

Khoja, S. 2001. APT: Arabic Part-of-Speech Tagger. *Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001)*, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Khoja, S. 2003. APT: An Automatic Arabic Part-of-Speech Tagger. *Computing Department*, pp. 157. Lancaster, UK: Lancaster University.

Khoja, S., Garside, P. and Knowles, G. 2001. A tagset for the morphosynactic tagging of Arabic. *Corpus Linguistics 2001*, Lancaster University, Lancaster, UK.

Kiraz, G. A. 2001. *Computational Nonlinear Morphology with Emphasis on Sematic Languages*. Cambridge: Cambridge University Press.

Koskenniemi, K. 1983. Two-Level Morphology. University of Helsinki.

Kurimo, M., Virpioja, S. and Turunen, V. T. 2009. Overview and Results of Morpho Challenge 2009. *Proceedings of the workshop of Unsupervised Morpheme Analysis MorphoChallenge at CLEF 2009 (Cross Language Evaluation Forum)*, Corfu, Greece.

Lane, E. W. 1968. An Arabic-English Lexicon. 7: 117-119.

Larkey, L. S. and Connell, M. E. 2001. Arabic Information Retrieval at UMass in TREC-10. *The Tenth Text REtrieval Conference (TREC 2001)* Gaithersburg: NIST, 2001.

Leech, G. and Wilson, A. 1996. EAGLES: Recommendations for the Morphosyntactic Annotation of Corpora.

Leech, G. and Wilson, A. 1999. Standards for Tagsets. In H. v. Halteren (ed.). *Syntactic Wordclass Tagging*, pp. 55-80. KLUWER Academic Publishers.

Liberman, M.Y. and Church, K.W. 1992. Text Analysis and Word Pronunciation in Text-to-Speech Synthesis. *In Advances in Speech Signal Processing*. Furui S. and Sondhi, M.M. (eds.). New York. Marcel Dekker Inc.

Maamouri, M. and Bies, A. 2004. Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004).*

Maamouri, M., Bies, A., Buckwalter, T. and Mekki, W. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. *NEMLAR Conference on Arabic Language Resources and Tools,*, Cairo, Egypt.

MacKinlay, A. 2005. The Effects of Part-of-Speech Tagsets on Tagger Performance. *The Department of Computer Science and Software Engineering*, pp. 44. Melbourne, Australia: University of Melbourne.

Marques, N. C. and Lopes, G. P. 2001. Tagging with Small Training Corpora. *Advances in Intelligent Data Analysis*, pp. 63-72. Springer Berlin / Heidelberg.

Marsi, E., Bosch, A. v. d. and Soudi, A. 2005. Memory-based morphological analysis generation and part-of-speech tagging of Arabic. *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pp. 1-8. Ann Arbor: Association for Computational Linguistics.

Mazroui, A. e., Meziane, A.-w., Lakhouaja, A.-H., Bebaha, M., Boudlal, A.-R. and Belhabeeb, R. 2009. محلل صرفي للكلمات العربية خارج النص وداخله Morphological analyzer for Arabic text in-context and out of context. *Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Science and Technology (KACST) and Arabic Language Academy.*, Damascus, Syria.

McCarthy, J. and Prince, A. 1990a. Foot and word in prosodic morphology: The Arabic broken plurals. *Natural Language & Linguistic Theory* 8: 209–282.

McCarthy, J. and Prince, A. 1990b. Prosodic morphology and templatic morphology. In M. Eid and J. McCarthy (eds.). *Perspectives on Arabic Linguistics: Papers from the Second Symposium*, pp. 1–54. Amsterdam: Benjamins, Amsterdam.

Melamed, D. and Resnik, P. 2000. Tagger Evaluation Given Hierarchical Tag Sets. *Computers and the Humanities* 34: 79-84.

Monachini, M. and Calzolari, N. 1996. Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. *A common proposal and applications to European languages.* Istituto di Linguistica Computazionale -CNR.

Mousser, J. 2010. A Large Coverage Verb Taxonomy For Arabic. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pp. 2675 - 2681. Valletta, Malta: European Language Resources Association (ELRA).

Nicolas, L., Sagot, B., Farré, J. and Clergerie, É. d. L. 2008. Computer aided correction and extension of a syntactic wide-coverage lexicon. *Proceedings of COLING 2008 22nd International Conference on Computational Linguistics*, Manchester, UK.

Ooi, V. B. Y. 1998. *Computer corpus lexicography* Edinburgh: Edinburgh University Press.

Paikens, P. 2007. Lexicon-Based Morphological Analysis of Latvian Language. *Proceedings of the 3rd Baltic Conference on Human Language Technologies*, pp. 235–240. Kaunas.

Pauw, G. D. and Schryver, G.-M. D. 2008. Improving the Computational Morphological Analysis of a Swahili Corpus for Lexicographic Purposes. *Lexikos 18 (AFRILEX-reeks/series 18: 2008)*: 303-318.

Petasis, G., Karkaletsis, V., Dimitra Farmakiotou, Samaritakis, G., Androutsopoulos, I. and Spyropoulos, C. D. 2001. A Greek Morphological Lexicon and its Exploitation by Greek Controlled Language Checker. In Y. Manolopoulos and S. Evripidou (eds.). *Proceedings of the 8th Panhellenic Conference in Informatics*, pp. 80–89. Nicosia, Cyprus.

Porter, M. F. 1980. An algorithm for suffix stripping. *Program* 14(3): 130−137.

Roark, B. and Sproat, R. W. 2007. *Computational Approaches to Morphology and Syntax*. Oxford University Press.

Rodríguez, H., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M. and Martí, M. A. 2008. Arabic WordNet: Semi-automatic Extensions using Bayesian Inference. *the 6th Conference on Language Resources and Evaluation LREC2008*, Marrakech (Morocco).

Russell, G. J., Pulman, S. G., Ritchie, G. D. and Black, A. W. 1986. A dictionary and morphological analyser for English. *Proceedings of the 11th coference on Computational linguistics*, Bonn, Germany: Association for Computational Linguistics.

Ryding, K. C. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge University Press.

Sabir, M. and Abdul-Mun'im, A.-M. i. 2009. برنامج (مداد) للتحليل الصرفي للكلمات العربية MIDAD morphological analyzer for Arabic text. *Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Science and Technology (KACST) and Arabic Language Academy.*, Damascus, Syria.

Sagot, B. 2005. Automatic acquisition of a Slovak Lexicon from a Raw Corpus. *Lecture Notes in Artificial Intelligence (© Springer-Verlag)* 3658 156-163.

Sagot, B. 2010. The Lefff, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias (eds.). *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pp. 2744-2751. Valletta, Malta: European Language Resources Association (ELRA).

Sagot, B. , Clement, L., Clergerie, E. V. d. L. and Boullier, P. 2006. The Lefff 2 syntactic lexicon for French: architecture, acquisition, use. *Proceeding of the fifth international conference on Language Resources and Evaluation, LREC 2006*, Genoa - Italy: European Language Resources Association (ELRA).

Sánchez León, F., and Nieto Serrano, AF. 1997. Retargeting a tagger. Corpus Annotation, ed. by Garside, Leech & McEnery, 163-64. London: Longman.

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*, Manchester, UK.

Sawalha, M. and Atwell, E. 2009a. Linguistically Informed and Corpus Informed Morphological Analysis of Arabic. *Proceedings of the 5th International Corpus Linguistics Conference CL2009*, Liverpool, UK.

Sawalha, M. and Atwell, E. 2009b. توظيف قواعد النحو والصرف في بناء محلل صرفي للغة العربية (Adapting Language Grammar Rules for Building Morphological Analyzer for Arabic Language). *Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Science and Technology (KACST) and Arabic Language Academy.*, Damascus, Syria.

Sawalha, M. and Atwell, E. 2010a. Constructing and Using Broad-Coverage Lexical Resource for Enhancing Morphological Analysis of Arabic. *Language Resource and Evaluation Conference LREC 2010*, Valleta, Malta: European Language Resources Association (ELRA).

Sawalha, M. and Atwell, E. 2010b. Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text. *Language Resource and Evaluation Conference LREC 2010* Valleta, Malta: European Language Resources Association (ELRA).

Sawalha, M. and Atwell, E. 2011a. Accelerating the Processing of Large Corpora: Using Grid Computing Technologies for Lemmatizing 176 Million Words Arabic Internet Corpus. *Advanced Research Computing Open Event*, University of Leeds, Leeds, UK.

Sawalha, M. and Atwell, E. 2011b. Corpus Linguistics Resources and Tools for Arabic Lexicography. *Workshop on Arabic Corpus Linguistics*, Lancaster University, Lancaster, UK.

Sawalha, M. and Atwell, E. 2011c. التحليل الصَّرفي لنصوص اللغة العربية الحديثة والكلاسيكية "Morphological Analysis of Classical and Modern Standard Arabic Text". *7th International Computing Conference in Arabic (ICCA11)*, Imam Mohammed Ibn Saud University, Riyadh, KSA.

Sawalha, M. and Atwell, E. Under review. A Theory Standard Tag Set Expounding Traditional Morphological features for Arabic Language Part-of-Speech Tagging. *Word structure journal, Edinburgh University Press.*

Schmid, H. and Laws, F. 2008. Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. *COLING'08*, Manchester,UK.

Sharoff, S. 2006. Creating General-Purpose Corpus Using Automated Search Engine Queries. In M. Baroni and S. Bernardini (eds.). *WaCky! Working papers on the Web as Corpus*, pp. 63-98. Bologna: GEDIT.

Sharoff, S., Kopotev, M., Erjavecy, T., Feldmanz, A. and Divjak, D. 2008. Designing and Evaluating a Russian Tagset. *LREC 2008: In Proceedings of the sixth international conference on Language Resources and Evaluation.*

Smrz, O. 2007. Functional Arabic Morphology: Formal System and Implementation. *Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics*, pp. 104. Prague: Charles University in Prague.

Smrž, O. 2009. ElixirFM Functional Arabic Morphology: Case Studies. *Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Science and Technology (KACST) and Arabic Language Academy.*, Damascus, Syria.26-28 April 2009.

Smrž, O., Bielický, V., Kouřilová, I., Kráčmar, J., Hajič, J. and Zemánek, P. 2008. Prague Arabic Dependency Treebank: A Word on the Million Words. *Proceedings of the Workshop on Arabic and Local Languages (LREC 2008)*, pp. 16–23. Marrakech, Morocco.

Sonbul, R., Ghnaim, N. and Dusouqi, M. S. 2009. نظام تحليل صرفي موّجه بالتطبيقات An Application Oriented Arabic Morphological Analyzer. *Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Science and Technology (KACST) and Arabic Language Academy.*, Damascus, Syria.26-28 April 2009.

Soudi, A., Bosch, A. v. d. and Neumann, G. (eds.) 2007. *Arabic Computational Morphology. Knowledge-based and Empirical Methods*. Dordrecht, The Netherlands: Springer.

Soudi, A., Cavalli-Sforza, V. and Jamari, A. 2001. A Computational Lexeme-Based Treatment of Arabic Morphology. *ACL/EACL 2001 Workshop on Arabic NLP.*, Toulouse, France, Friday 6 July 2001.

Tadi, M. and Fulgosi, S. 2003. Building the Croatian morphological lexicon. *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, Budapest, Hungary: Association for Computational Linguistics.

Talmon, R. and Wintner, S. 2003. Morphological Tagging of the Qur'an. *In Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, an EACL'03 Workshop*, Budapest, Hungary.

Teahan, B. 1998. Modeling English Text. *Department of Computer Science*, New Zealand: University of Waikato.

Teufel, S., Schmid, H., Heid, U. and Schiller, A. 1996. Study of the relation between tagsets and taggers. Stuttgart, Germany Institut für maschinelle Sprachverarbeitung, Universität Stuttgart

Thabet, N. 2004. Stemming the Qur’an. *COLING 2004, Workshop on computational approaches to Arabic script-based languages.August 28,2004*, pp. 85-88.

Tlili-Guiassa, Y. 2006. Hybrid Method for Tagging Arabic Text. *Journal of Computer Science* 2(3): 245-248.

Taylor, P. and Black, A.W. 1998. Assigning Phrase-Breaks from Part-of-Speech Sequences. *In Computer Speech and Language*. 12.2: 99-117.

Voutilainen, A. 2003. Part-of-Speech Tagging. In R. Mitkov (ed.). *The Oxford Handbook of Computational Linguistics*, pp. 219-232. Oxford University Press.

Wald Abah, M. A. 2008. تاريخ النجو العربي في المشرق والمغرب *History of Arabic Grammar in the East and the West*. Beirut, Lebanon: Dar Al-Kutub Al-Alamyyah.

Wright, W. 1996. *A Grammar of the Arabic Language, Translated from the German of Caspari, and Editted with Numerous Additions and Corrections*. Beirut: Librairie du Liban.

Ya‘qūb, I. B. 1996. *Mu‘jam al-awzān al-sarfiyah* معجم الأوزان الصرفية. Beirut, Lebanon: ‘ālam al-Kutub

Yonghui, G., Baomin, W., Changyuan, L. and Bingxi, W. 2006. Correlation Voting Fusion Strategy for Part of Speech Tagging. *8th International Confenerance on Signal Processing Proceedings, ICSP2006.*

Yousfi, A. 2010. The morphological analysis of Arabic verbs by using the surface patterns. *IJCSI International Journal of Computer Science Issues* 7(3(11)): 33-36.

Yusof, R. J. R., Zainuddin, R. and Baba, M. S. 2010. Qur'anic Words Stemming. *The Arabian Journal for Science and Engineering* 35(2C): 37-49.

Zaenen, A., Carletta, J., Garretson, G., Bresnan, J., Koontz-Garboden, A., Nikitina, T., O'Connor, M. C. and Wasow, T. 2004. Animacy encoding in English: Why and how. *In Proceedings of the ACL-04 Workshop on Discourse Annotation*.

Zaied, M. 2009. تقرير في المحللات الصرفية العربية "Report on Arabic Morphological Analyzers". *Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Science and Technology (KACST) and Arabic Language Academy.*, Damascus, Syria.

Zarrouki, T. and Kebdani, M. 2009. مشروع أية –سبل القاموس العربي للتدقيق الإملائي مفتوح المصدر، تجربة وآفاق Aya-Spell Project, An Open-source Arabic Spell Checker Dictionary, experience and Future Work. *Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Science and Technology (KACST) and Arabic Language Academy.*, Damascus - Syria.

Zeman, D. 2008. Reusable Tagsets Conversion Using Tagset Drivers. *Proceedings of the Sixth conference on International Language Resources and Evaluation (LREC'08)*, pp. 213-218. Marrakech, Morocco: European Language Resources Association (ELRA).

Zerrouki, T. and Balla, A. 2009. Implementation of infixes and circumfixes in the spellcheckers. *2nd International Conference on Arabic Language Resources and Tools*, Cairo - Egypt.

Zibri, C. B. O., Torjmen, A. and Ahmad, M. B. 2006. An Efficient Multi-agent system Combining POS-Taggers for Arabic Texts. *CICLing 2006,* LNCS 3878(pp.121-131).

Zolfagharifard, E. 2009. Anti-terror technology tool uses human logic. *The Engineer*.

Appendix A
The SALMA Tag Set for Arabic text

The SALMA Morphological Features Tag Set (SALMA, Sawalha Atwell Leeds Morphological Analysis tag set for Arabic) captures long-established traditional morphological features of Arabic, in a compact yet transparent notation. First, we introduce Part-of-Speech tagging and tag set standards for English and other European languages, and then survey Arabic Part-of-Speech taggers and corpora, and long-established Arabic traditions in analysis of grammar and morphology. A range of existing Arabic Part-of-Speech tag sets are illustrated and compared; and we review generic design criteria for corpus tag sets. For a morphologically-rich language like Arabic, the Part-of-Speech tag set should be defined in terms of morphological features characterizing word structure. We describe the SALMA Tag Set in detail, explaining and illustrating each feature and possible values. In our analysis, a tag consists of 22 characters; each position represents a feature and the letter at that location represents a value or attribute of the morphological feature; the dash "-" represents a feature not relevant to a given word. The first character shows the main Parts of Speech, from: noun, verb, particle, punctuation, and Other (residual); these last two are an extension to the traditional three classes to handle modern texts. The characters 2, 3, and 4 are used to represent subcategories; traditional Arabic grammar recognizes 34 subclasses of noun (letter 2), 3 subclasses of verb (letter 3), 21 subclasses of particle (letter 4). Others (residuals) and punctuations are represented in letters 5 and 6 respectively. The next letters represent traditional morphological features: gender (7), number (8), person (9), inflectional morphology (10) case or mood (11), case and mood marks (12), definiteness (13), voice (14), emphasized and non-emphasized (15), transitivity (16), rational (17), declension and conjugation (18). Finally there are four characters representing morphological information which is useful in Arabic text analysis, although not all linguists would count these as traditional features: unaugmented and augmented (19), number of root letters (20), verb root (21), types of nouns according to their final letters (22). The SALMA Tag Set is not tied to a specific tagging algorithm or theory, and other tag sets could be mapped onto this standard, to simplify and promote comparisons between and reuse of Arabic taggers and tagged corpora.

The SALMA tag structure consists of 22 characters. Figure 1 shows a sample of tagged sentence from the Qur'an and it shows the morphological categories and the attributes of a selected word in more details.

Word		Morphemes			Tag
wa waaṣṣaynā And We have enjoined	ووَصَّيْنا	وَ	*wa*	*And*	`p--c------------------`
		وَصَّيْ	*waṣṣay*	*Have enjoined*	`v-p---mpfs-s-amohvtt&-`
		نَا	*nā*	*We*	`r---r-xpfs-s----hn----`
al-'insāna (on) man	الْإِنسَانَ	الْ	*al-*	*The*	`r--d------------------`
		إِنسَانَ	*'insāna*	*man*	`nq----ms-pafd---htbt-s`
bi- wālidayhi His parents	بِوَالِدَيْهِ	بِ	*bi*	*To*	`p--p------------------`
		وَالِد	*wālida*	*Parents*	`nu----md-vgki---htot-s`
		يْ	*y*	*Both*	`r---r-xdts-s----------`
		هِ	*hi*	*His*	`r---r-msts-k----------`
ḥusn^{an} Kindness	حُسْنًا	حُسْنَ	*ḥusn*	*kindness*	`ng----ms-vafi---ndst-s`
		اً	*an*		`r---k------f----------`

Figure A.1 Sample of Tagged document of vowelized Qur'an Text using SALMA Tag Set

Main category	Main Part-of-Speech	Part-of-Speech: Noun	Part-of-Speech: Verb	Part-of-Speech: Particle	Part-of-Speech: Other	Punctuation marks	Gender	Number	Person	Inflectional morphology	Case and Mood	Case and Mood marks	Definiteness	Voice	Emphasized and non-emphasized	Transitivity	Rational	Declension and Conjugation	Unaugmented and Augmented	Number of root letters	Verb root	Noun Finals
Position	1	2	3	4	5	6	7	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
إِنسَانَ	n	q	-	-	-	-	m	s	-	p	a	f	d	-	-	-	h	t	b	t	-	s
Attributes	**Noun**	**Generic noun**	-	-	-	-	**Masculine**	**Singular**	-	**Non-declinalbe**	**Accusative**	***fatha***^{h}	**Defined**	-	-	-	**Rational**	**Primitive / Concrete noun**	**Augmented by two letters**	**Triliteral**	-	**Sound**

Figure A.2 SALMA tag structure

Table A.1 SALMA Tag Set categories

Position	Morphological Features Categories		
1	Main Part-of-Speech	أَقسام الكلام الرئيسيَّة	*'aqsām al-kalām ar-ra'īsiyya*^{t}
2	Part-of-Speech: Noun	أقسام الكلام الفرعيَّة (الاسم)	*'aqsām al-kalām al-far'iyya*^{t} *(al-'ism)*
3	Part-of-Speech: Verb	أقسام الكلام الفرعيَّة (الفعل)	*'aqsām al-kalām al-far'iyya*^{t} *(al-fi'l)*
4	Part-of-Speech: Particle	أقسام الكلام الفرعيَّة (الحرف)	*'aqsām al-kalām al-far'iyya*^{t} *(al-ḥarf)*
5	Part-of-Speech: Other	أقسام الكلام الفرعيَّة (أخرى)	*'aqsām al-kalām al-far'iyya*^{t} *('uḫrā)*
6	Punctuation marks	أقسام الكلام الفرعيَّة (علامات الترقيم)	*'aqsām al-kalām al-far'iyya*^{t} *('alāmāt at-tarqīm)*
7	Gender	المُذَكَّر والمُؤَنَّث	*al-mudakkar wa al-mu'annat*
8	Number	العدد	*al-'adad*
9	Person	الاسناد	*al-'isnād*
10	Inflectional Morphology	الصَّرف	*aṣ-ṣarf*
11	Case or Mood	الحالة الإعرابية للاسم أو الفعل	*al-ḥāla*^{tu} *al-'i'rābiyya*^{tu} *lil-'ism 'aw al-fi'l*
12	Case and Mood Marks	علامة الإعراب أو البناء	*'alāmāt al-'i'rāb wa al-binā'*
13	Definiteness	المَعْرِفَة والنَّكِرَة	*al-ma'rifa*^{ti} *wa an-nakira*^{ti}
14	Voice	المَبْني لِلمَعْلُوم و المَبْني لِلمَجْهُول	*al-mabnī lil-ma'lūm wa al-mabnī lil-mağhūl*
15	Emphasized and Non-emphasized	المُؤكَّد وغيرُ المُؤكَّد	*al-mu'akkad wa ğayir al-mu'akkad*
16	Transitivity	اللازم والمتعدي	*al-lāzim wa al-muta'addi*
17	Rational	العاقل وغير العاقل	*al-'āqil wa ğayir al-'āqil*
18	Declension and Conjugation	التَّصريف	*at-taṣrīf*
19	Unaugmented and Augmented	المجرَّد والمزيد	*al-muğarrad wa al-mazīd*
20	Number of Root Letters	عَدَد أحْرُف الجَذْر	*'adad 'aḥruf al-ğadr*
21	Verb Root	بُنية الفعل	*bunya*^{tu} *al-fi'l*
22	Noun Finals	أقسام الأسم تبعاً للفظ آخره	*'aqsām al-'ismi tib'*^{an} *li-lafẓi 'āḫirhi*

A.1 Position 1; Main part-of-speech

Table A.2 Main part-of-speech category attributes and tags at position 1

Position	**Feature Name**				**Tag**
1	**Main Part-of-Speech** أَقسام الكلام الرئيسيَّة *'aqsām al-kalām ar-r'īsiyya*^{t}				
	Noun	اسم	*'ism*	كِتَاب *kitāb* 'book'	**n**
	Verb	فعل	*fi'l*	كَتَب *katab* 'wrote'	**v**
	Particle	حرف	*ḥarf*	عَلَى *'alā* 'on'	**p**
	Other (Residual)	أخرى	*'uḫrā*	كاتِبَةٌ *kātiba*^{tun} 'writer / Fem'	**r**
	Punctuation	علامة ترقيم	*'alāmat tarqīm*	قالَ: أنا ذَاهِبٌ *qāla : 'anā dāhib*^{un} 'he said: I am leaving'	**u**

A.2 Position 2; Part-of-Speech Subcategories of Noun

Table A.3 Part-of-Speech subcategories of Noun attributes and their tags at position 2

Position	**Feature Name**				**Tag**
2	**Part-of-Speech: Noun** (الاسم) أَقسام الكلام الفرعيَّة *’aqsām al-kalām al-far’iyya*^{t} *(al-‘ism)*				
	Gerund / Verbal noun	المصدر	*al-maṣdar*	ضَرْب *ḍarb* ‘hitting’	**g**
	Gerund/ verbal noun with initial *mīm*	المصدر الميمي	*al-maṣdar al-mīmī*	مَوعِد *maw‘id* ‘date’	**m**
	Gerund of instance	مصدر المرَّة	*maṣdar al-marra*^{h}	نَظْرَة *naẓra*^{h} ‘one look’	**o**
	Gerund of state	مصدر الهيئة/ مصدر النوع	*maṣdar al-hay’a*^{h} */ maṣdar al-naw’*	جِلْسَة *ǧilsa*^{h} ‘sitting position’	**s**
	Gerund of emphasis	مصدر التوكيد	*maṣdar al-tawkīd*	حطّمتُ الخزانةَ تحطيماً *hattamtu al-hizāna*^{ta} *taḥṭīm*^{an} ‘I completely destroyed the wardrobe’	**e**
	Gerund of profession	المصدر الصناعي	*al-maṣdar al-ṣināī*	فُروسيَّة *furūsiyya*^{h} ‘Horsemanship’	**i**
	Pronoun	الضمير	*al-ḍamīr*	هو *huwa* ‘He’	**p**
	Demonstrative pronoun	اسم الإشارة	*’ism al-’šāra*^{h}	هذا *hāḏā* ‘This’	**d**
	Specific relative pronoun	اسم الموصول الخاص	*’ism al-mawṣūl al-ẖāṣ*	الذي *al-laḏī* ‘Who’	**r**
	Non-specific relative pronoun	اسم الموصول المشترك	*’ism al-mawṣūl al-muštarak*	مَنْ *man* ‘Who’	**c**
	Interrogative pronoun	اسم الاستفهام	*’ism al-’istfhām*	مَنْ *man* ‘Who?’	**b**
	Conditional noun	اسم الشرط	*’ism al-šarṭ*	أينما *aynamā* ‘where ever’	**h**
	Allusive noun	الكناية	*al-kināya*^{h}	كذا *kaḏā* ‘as well as’	**a**
	Adverb	الظَّرف	*aẓ-ẓarf*	يوم *yawm* ‘day’	**v**
	Active participle	اسم الفاعل	*’ism al-fā‘il*	ضارِب *ḍārib* ‘hitter’	**u**
	Intensive Active participle	مبالغة اسم الفاعل	*mubālaġa*^{t} *’ism al-fā‘il*	جَرَّاح *ǧarraḥ* ‘Surgeon’	**w**
	Passive participle	اسم المفعول	*’ism al-mf‘ūl*	مَضْرُوب *maḍrūb* ‘Struck’	**k**
	Adjective	الصِّفة المشبَّهة	*aṣ-ṣifa*^{h} *al-mušabbaha*^{h}	طويل *ṭawīl* ‘tall’	**j**
	Noun of place	اسم المكان	*’ism al-mkān*	مَكْتَب *maktab* ‘office’	**l**
	Noun of time	اسم زمان	*’ism zamān*	مَطْلِع *maṭla‘* start time	**t**
	Instrumental noun	اسم الآلة	*’ism al-’āla*^{h}	مِنْشار *minšār* ‘Saw’	**z**
	Proper noun	اسم العلم	*’ism al-‘alam*	فاطِمَة *fāṭima*^{h} ‘Fatima’	**n**
	Generic noun	اسم الجنس	*’ism al-ǧins*	حِصان *hiṣān* ‘Horse’	**q**
	Numeral	اسم العدد	*’ism al-‘adad*	ثلاثة *ṯalāṯa*^{h} ‘Three’	**+**
	Verb-like noun	اسم الفعل	*’ism al-fi‘l*	هيهات *hayhāt* Wishing	**&**
	Five nouns	الأسماء الخمسة	*al-’asmā’ al-ẖamsa*^{h}	أَبٌ *‘ab*^{un} ‘Father’	**f**

Position	Feature Name				Tag
2	**Part-of-Speech: Noun** (الاسم) أَقسام الكلام الفرعيَّة *’aqsām al-kalām al-far’iyya^{t} (al-‘ism)*				
	Relative noun	اسم منسوب	*’ism mansūb*	عِلْمِيّ *‘ilmiyy^{un}* Scientific	*
	Diminutive	اسم تصغير	*’ism taṣġīr*	شُجَيْرَة *šuğayra^{h}* ‘Bush’	y
	Form of exaggeration	صيغة مبالغة	*ṣīġat al-mubālaġah*	جَبَّار *ğabbār* ‘Tremendous’	x
	Collective noun	اسم جمع	*’ism ğam‘*	قوم *qawm* ‘Folk’	$
	Plural generic noun	اسم جنس جمعي	*’ism ğins ğam‘ī*	تفاح tuffāḥ ‘Apple’	#
	Elative noun	اسم تفضيل	*’ism tafḍīl*	أفضل *’afḍal* ‘Better’	@
	Blend noun	اسم منحوت	*’ism manḥūt*	بسملة basmalah ‘bismallah’	%
	Ideophonic interjection	اسم صوت	*’ism ṣawt*	آه *’āh* ‘Ah’	!

A.3 Position 3; Part-of-Speech Subcategories of Verb

Table A.4 Part-of-Speech subcategory of verb attributes and their tags at position 3

Position	Feature Name				Tag
3	**Part-of-Speech: Verb** (الفعل) أَقسام الكلام الفرعيَّة *’aqsām al-kalām al-far’iyya^{t} (al-fi’l)*				
	Perfect verb	فعل ماضٍ	*fi‘l māḍ^{in}*	كَتَبَ *kataba* ‘He wrote’	p
	Imperfect verb	فعل مضارع	*fi‘l muḍāri‘*	يَكْتُبُ *yaktubu* ‘He is writing’	c
	Imperative verb	فعل الأمر	*fi‘l al-’amr*	اكْتُبْ *’uktub* ‘write’	i

A.4 Position 4; Part-of-Speech Subcategories of Particle

Table A.5 Part-of-speech subcategories of Particles attributes and their tags at position 4

Position	Feature Name				Tag
4	**Part-of-Speech: Particle** (الحروف) أَقسام الكلام الفرعيَّة *’aqsām al-kalām al-far‘iyya^{t} (al-ḥarf)*				
	Jussive-governing particle	حرف جزم	*ḥarf ğazim*	لَمْ *lam* ‘No’	j
	Subjunctive-governing particle	حرف نصب	*ḥarf naṣib*	كَيْ *kay* ‘So that’	o
	Partially subjunctive-governing particle	حرف النصب الفرعي	*ḥarf naṣib far‘ī*	حتى *ḥattā* ‘till’	u
	Preposition	حرف جر	*ḥarf ğarr*	إلى *’ilā* ‘To’	p
	Annulling particle	حرف ناسخ	*ḥarf nāsiẖ*	ما *mā* ‘No’	a
	Conjunction	حرف عطف	*ḥarf ‘aṭif*	و *wa* ‘And’	c
	Vocative particle	حرف نداء	*ḥarf nidā’*	يا *yā* ‘Oh’	v
	Exceptive particle	حرف استثناء	*ḥarf ’stiṯnā’*	إلاَّ *’illā* ‘Except’	x
	Interrogative particle	حرف استفهام	*ḥarf ’stifhām*	هل *hal* ‘Is?’	i
	Particle of futurity	حرف استقبال	*ḥarf ’stiqbāl*	سوف *sawfa* ‘will’	f
	Causative particle	حرف تعليل	*ḥarf ta‘līl*	كي *kay* ‘To’	s
	Negative particle	حرف نفي	*ḥarf nafī*	لَمْ *lam* ‘No’	n
	Jurative particle	حرف قسم	*ḥarf qasam*	بِ *bi* ‘sware’	q
	Yes/No response particle	حرف الجواب	*ḥarf ğawāb*	نعم *na‘am* ‘Yes’	w

Position	Feature Name				Tag
4	**Part-of-Speech: Particle** (الحروف) أَقسام الكلام الفرعيَّة *’aqsām al-kalām al-far‘iyya^{t} (al-ḥarf)*				
	Jussive-governing conditional particle	حرف شرط جازم	*ḥarf šart ğāzim*	إنْ *’in* ‘If’	**k**
	Particle of incitement	حرف تحضيض	*ḥarf taḥḍīḍ*	هلاَّ *hallā* ‘would’	**m**
	Gerund-equivalent particle	حرف مصدري	*ḥarf maṣdarī*	أنْ *’an* ‘To’	**g**
	Particle of attention	حرف تنبيه	*ḥarf tanbī^{h}*	ألا *’alā* ‘careful’	**t**
	Emphatic particle	حرف توكيد	*ḥarf tawkīd*	إنَّ *’inna* ‘emphasis’	**z**
	Explanatory particle	حرف تفسير	*ḥarf tafsīr*	أي *’ay* ‘i.e’	**d**
	Particle of comparison	حرف تشبيه	*ḥarf tašbī^{h}*	كأنَّ *ka’anna* ‘similar’	**l**
	Non-governing particles	حرف غير عامل	*ḥarf ğayr ‘āmil*	قَدْ *qad* ‘already or perhaps’	**b**

A.5 Position 5; Part-of-Speech Subcategories of Other (Residuals)

Table A.6 Part-of-speech subcategories of Other (Residuals) attributes and their tags at position 5

Position	Feature Name				Tag
5	**Part-of-Speech: Other** (أُخْرَى) أقسام الكلام الفرعيَّة *’aqsām al-kalām al-far‘iyya^{t} (’uḫrā)*				
	Prefix	زيادة في أول الكلمة	*ziyāda^{h} fī ’awal al-kalima^{h}*	استكتبني *’istaktabanī* ‘he employed me as a writer’	**p**
	Suffix	زيادة في آخر الكلمة	*ziyāda^{h} fī ’āḫir al-kalimah*	أَصدقاء *’aṣdiqā’* ‘Friends’	**s**
	Suffixed pronoun	ضمير متصل	*ḍamīr mutaṣil*	كِتابُهُ *kitabahu* ‘his book’	**r**
	tā' marbūṭa^{h}	تاء مربوطة	*tā’ marbūṭa^{h}*	كاتِبَة *kātiba^{tun}* ‘she-writer’	**t**
	Relative *yā'*	ياء النسبة	*yā’ an-nisba^{h}*	عَرَبِيّ *‘arabiyy* ‘Arabian’	**y**
	tanwīn	تنوين	*tanwīn*	كِتابٌ *kitāb^{un}* ‘a book’	**k**
	tā' of femininization	تاء التأنيث	*tā’ al-ta’nīṯ*	كَتَبَتْ *katabaṯ* ‘she wrote’	**f**
	nūn of protection	نون الوقاية	*nūn al-wiqāya^{h}*	سَألَنِي *sa’alanī* ‘he asked me’	**n**
	Emphatic *nūn*	نون التوكيد	*nūn al-tawkīd*	يَضْرِبَنَّ *yaḍribanna* ‘They are hitting’	**z**
	Imperfect prefix	حرف مضارعة	*ḥarf muḍāra‘a^{h}*	يَسْأَلُ *yas’alu* ‘He is asking’	**a**
	Definite article	أداة تعريف	*’adā^{t} ta‘rīf*	الكتاب *al-kitāb* ‘The book’	**d**
	Masculine sound plural letters	حروف جمع المذكر السالم	*ḥurūf ğam‘ al-muḏakkar as-sālim*	الكاتبون *al-kātibūn* ‘The writers (MAS)’	**m**
	Feminine sound plural letters	حروف جمع المؤنث السالم	*ḥurūf ğam‘ al-mu’nnaṯ as-sālim*	الكاتبات *al-kātibāt* ‘The writers (FEM)’	**l**
	Dual letters	حروف المثنى	*ḥurūf al-muṯannā*	الكاتبان *al-kātibān* ‘The two writers’	**u**
	Imperative prefix	حروف الأمر	*ḥurūf al-’amr*	اكتب *’uktub* ‘Write’	**I**

Position	Feature Name				Tag
5	**Part-of-Speech: Other** (أُخْرَى) أقسام الكلام الفرعيَّة *'aqsām al-kalām al-far'iyya*^{t} *('uẖrā)*				
	Number (digits)	رَقَم	*raqam*	(+325461) (-897,653) (0.986)	**g**
	Currency	عُمْلَة	*'umla*^{t}	(1,500د.أ) (2,927ر.س) ($250)	**c**
	Date	تَارِيخ	*tārīẖ*	(27/09/2011) (2011 أيلول 27)	**e**
	Non-Arabic word	كَلِمَة غَيْر عَرَبِيَّة	*kalima*^{t} *ġayr 'arabiyya*^{h}	windows, photoshop, games, download	**w**
	Borrowed (foreign) word	كَلِمَة مُعَرَّبَة	*kalima*^{t} *mu'arraba*^{h}	كُوزْمُوبُولِيْتَان *kuzmūbūlītān* 'cosmopolitan'	**x**

A.6 Position 6; Part-of-Speech Subcategories of Punctuation Marks

Table A.7 Part-of-speech subcategories of Punctuation Marks attributes and their tags at position 6

Position	Feature Name				Tag
6	**Punctuation Marks** (علامات الترقيم) أقسام الكلام الفرعية *'aqsām al-kalām al-far'iyya*^{t} *('alāmāt at-tarqīm)*				
	Full stop	نقطة	*nuqṭa*^{h}	(.)	**s**
	Comma	فاصلة	*fāṣila*^{h}	(،)	**c**
	Colon	نقطتان	*nuqṭatān*	(:)	**n**
	Semi colon	فاصلة منقوطة	*fāṣila*^{h} *manqūṭa*^{h}	(؛)	**l**
	Parentheses	قوسان	*qawsān*	(( ))	**p**
	Square brackets	قوسان حاصرتان	*qawsān ḥāṣiratān*	([ ])	**b**
	Quotation mark	علامة اقتباس	*'alāma*^{tu} *'iqtibās*	(" ")	**t**
	Dash	شرطة معترضة	*šarṭa*^{h} *mu'tariḍa*^{h}	(–)	**d**
	Question mark	علامة استفهام	*'alāma*^{tu} *'istifhām*	(؟)	**q**
	Exclamation mark	علامة تعجب	*'alāma*^{tu} *ta'ağğub*	(!)	**e**
	Ellipsis mark	علامة حذف	*'alāma*^{tu} *ḥaḏf*	(…)	**i**
	Continuation mark	علامة التَّابعية	*'alāma*^{tu} *at-tabi'yya*^{h}	(=)	**f**
	Other punctuations	عَلامَات أُخْرَى	*'alāmāt 'uẖrā*	/	**o**

A.7 Position 7; Morphological Feature of Gender

Table A.8 Morphological feature of Gender attributes and their tags at position 7

Position	Feature Name				Tag
7	**Morphological Gender** المُذَكَّر والمُؤَنَّث *al-muḏakkar wa al-mu'annaṯ*				
	Masculine	مذكر	*muḏakkar*	رجل *rağul* 'man'	**m**
	Feminine	مؤنث	*mu'annaṯ*	امرأة*'imra'a*^{h} Woman	**f**
	Common gender	مذكر أو مؤنث	*muḏakkar 'aw mu'annaṯ*	ملح *milḥ* 'Salt' روح *rūḥ* 'Soul'	**x**

A.8 Position 8; Morphological Feature of Number

Table A.9: Morphological feature of Number attributes and their tags at position 8

Position	Feature Name				Tag
8	**Number** العدد *al-‘adad*				
	Singular	مفرد	*mufrad*	قلم *qalam* ‘A pen’ فلّاح *fallāḥ* ‘Farmer’ منارة *manāra*^{h} ‘A minaret’	**s**
	Dual	مثنى	*muṯannā*	(قلم: قلمان، قلمين) (qalam: qalamān, qalamayn) ‘(A pen: two pens)’ (منارة: منارتان، منارتين) (*manāra*^{h}*: manāratān, manāratayn)*(A minaret: two minarets)	**d**
	Sound plural	جمع سالم	*ğam‘ sālim*	(فلّاح: فلّاحون، فلّاحين) *(fallāḥ: fallāḥūn, fallāḥīn)* (A farmer: Farmers)’ (منارة: منارات) (*manāra*^{h}*: manārāt)* (A minaret: minarets)	**p**
	Broken plural	جمع تكسير	*ğam‘ taksīr*	(قلم: أقلام) (*qalam: ’aqlām)* ‘(A pen: pens)’	**b**
	Plural of paucity	جمع قلة	*ğam‘ qilla*^{h}	(حرف: أحرف) *(ḥarf: ’aḥruf)* (A letter: letters)	**m**
	Plural of multitude	جمع كثرة	*ğam‘ kaṯra*^{h}	(حرف: حروف) *(ḥarf: ḥurūf)* (A letter: letters)	**j**
	Ultimate plural	منتهى الجموع	*munthā al-ğumū‘*	(مسجد: مساجد) *(masğid: masāğid)* (A mosque: mosques)	**u**
	Plural of plural	جمع الجمع	*ğam‘ al-ğam‘*	(بيت: بُيوت، بُيوتات) *(bayt: buyūt, buyūtāt)* ‘(A home: homes)	**l**
	Undefined	غير مُعرَّف	*ğayr mu‘arraf*	كَتَبَ الطَّالِبُ الدَّرْسَ *katab aṭ-ṭālibu ad-darasa* ‘the student wrote the lesson’; كَتَبَ الطَّالِبَانِ الدَّرْسَ *katab aṭ-ṭāliban ad-darsa* ‘the two students wrote the lesson’; كَتَبَ الطَّلابُ الدَّرْسَ *kataba aṭ-ṭulābu ad-darsa* ‘the students wrote the lesson’	**x**

A.9 Position 9; Morphological Feature of Person

Table A.10 Morphological feature of Person category attributes and their tags at position 9

Position	Feature Name				Tag
9	**Person** الاسناد *al-’isnād*				
	First Person	المُتكَلِّم	*al-mutakallim*	كَتَبتُ *katabtu*‘I wrote’	**f**
	Second Person	المُخاطَب	*al-muḫāṭab*	كَتَبْتُمَا *katabtumā* ‘You wrote’	**s**
	Third Person	الغائِب	*al-ğā’ib*	كَتَبْنَ *katabna*‘They Wrote’	**t**

A.10 Position 10; Morphological Feature of Inflectional Morphology

Table A.11 The morphological feature category of Inflectional Morphology attributes and their tags at position 10

Position	Feature Name				Tag
10	**Inflectional Morphology** الصَّرف *aṣ-ṣarf*				
	Declined (noun) Conjugated (verb)	مُعرب	*mu‘rab*	يَغيبُ *yaġību* ‘Miss’	**d**
	Triptote / fully declined	مُعرب – منصرف	*mu‘rab - munṣarif*	غائبٌ *ġā’ib* ‘Absent’	**v**
	Non-declinable	مُعرب – ممنوع من الصرف	*mu‘rab - mamnū’ mina aṣ-ṣarf*	عُثمانُ *‘uṯmānu* ‘Othman’	**p**
	Invariable (v, n)	مبني	*mabnī*	هؤُلاءِ *hā’ulā’i* ‘Those’ فَعَلَ *fa‘ala* ‘Did’ لَيْتَ *layta* ‘Wish’	**s**

A.11 Position 11; Morphological Feature Category of Case or Mood

Table A.12 The morphological feature of Case or Mood category attributes and their tags at position 11

Position	Feature Name						Tag
11	**Case or Mood** الحالة الإعرابية للاسم أو الفعل *al-ḥāla*^{tu} *al-’i‘rābiyya*^{tu} *lil-’ism ’aw al-fi‘l*						
	Nominative	Indicative	مرفوع	*marfū‘*	يَكْتُبُ *yaktubu* ‘He is writing’	الكتابُ *al-kitābu* ‘The Book’	**n**
	Accusative	Subjunctive	منصوب	*manṣūb*	لن يَكْتُبَ *lan yaktuba* ‘He will not write’	الكتابَ *al-kitāba* ‘The Book’	**a**
	Genitive	--------	مجرور	*mağrūr*	**-------**	الكتابِ *al-kitābi* ‘The Book’	**g**
	-------	Imperative or jussive	مجزوم	*mağzūm*	لَمْ يَكْتُبْ *lam yaktub* He did not write’	**-----**	**j**

A.12 Position 12; The Morphological Feature of Case and Mood Marks

Table A.13 The morphological feature category of Case and Mood Marks attributes and tags at position 12

Position	Feature Name				Tag
12	**Case and Mood Marks** علامة الإعراب أو البناء *‘alāmāt al-’i‘rāb wa al-binā’*				
	ḍamma^h	الضمة / الضم	*al-ḍamma*^{h} / *al-ḍamm*	قدِمَ الوزيرُ *qadima al-wazīru* ‘The minister arrived’ يَصومُ أحمد *yaṣūmu aḥmad* ‘Ahmad fasts’	**d**
	fatḥa^h	الفتحة / الفتح	*al-fatḥa*^{h} / *al-fatḥ*	أكرمَ صالحٌ الوزيرَ *’akrama ṣāliḥun al-wazīra* ‘Salih honored the minister’ لنْ نَصبِرَ على الذُّلِّ *lan naṣbira ‘alā aḏ-ḏulli* ‘We are not standing the humiliation’	**f**
	kasra^h	الكسرة	*al-kasra*^{h} / *al-kasr*	خلق الله السماواتِ والأرضَ *ẖalaqa allahu as-samāwāti wa al-’arḍa* ‘God created the skys and the earth’	**k**
	sukūn (Silence)	السكون	*as-sukūn*	لَمْ أُسافرْ إلى المدينةِ *lam ’usāfir ’ilā al-madīnati* ‘I did not travel to the city’	**s**
	wāw	الواو	*al-wāw*	إذا جاءَكَ المنافقونَ *’iḏā ğā’aka al-munāfiqūn* ‘If the Hypocrites come to thee’	**w**
	alif	الألف	*al-’alif*	التقى الفريقانِ *’iltaqā al-farīqān* ‘The two teams have met’	**a**
	yā’	الياء	*al-yā’*	ذهَبتُ إلى أخيكَ *ḏahbtu ’ilā ’aẖīka* ‘I went to your brother’	**y**
	Inflectional *nūn*	ثبوت النون	*ṯubūt an-nūn*	المرشحان يتقدمانِ الإنتخابات *al-muraššḥāni yataqddamāni al-’intiẖābāt* ‘ Both candidates are ahead of elections’	**n**
	Deletion of *nūn*	حذف النون	*ḥaḏf an-nūn*	المسلمون لنْ يَصبروا على الذُّلِّ *al-muslimūn lan yaṣbirū ‘’alā aḏ-ḏulli* ‘Muslims will not stand to the humiliation’	**o**
	Deletion of vowel letter	حذف حرف العلّة	*ḥaḏf ḥarf al-‘illa*^{h}	لَمْ يخشَ صالحٌ إلا الله *lam yahša ṣāliḥ ’illā allaha* ‘Salih does not afraid except of God’	**v**

A.13 Position 13; The Morphological Feature of Definiteness

Table A.14 The morphological feature of Definiteness category attributes and their tags at position 13

Position	Feature Name				Tag
13	**Definiteness** المَعرِفة والنَّكِرَة *al-ma‘rifa*^{h} *wa an-nakira*^{h}				
	Definiteness	مَعرِفة	*ma‘rifa*^{h}	الكتاب *al-kitāb* ‘The book’	**d**
	Indefiniteness	نَكِرَة	*nakira*^{h}	كتاب *kitāb* ‘A book’	**i**

A.14 Position 14; The Morphological Feature of Voice

Table A.15 The morphological feature of Voice category attributes and their tags at position 14

Position	Feature Name				Tag
14	**Voice** المَبْني لِلمَعْلُوم و المَبْني لِلمَجْهُول *al-mabnī lil-ma‘lūm wa al-mabnī lil-mağhūl*				
	Active voice	مَبْنِي للمَعْلُوم	*mabnī lil-ma‘lūm*	كَتَبَ *kataba* ‘He wrote’	**a**
	Passive voice	مَبْنِي للمَجْهُول	*mabnī lil-mağhūl*	كُتِبَ *kutiba* ‘it was written’	**p**

A.15 Position 15; The Morphological Feature of Emphasized and Non-emphasized

Table A.16 The morphological feature of Emphasized and Non-emphasized category attributes and their tags at position 15

Position	Feature Name				Tag
15	**Emphasized and Non-emphasized** المُؤكَّد وغيرُ المُؤكَّد *al-mu’akkad wa ğayir al-mu’akkad*				
	Emphatic verb	فعل مُؤكَّد	*fi‘l mu’akkad*	لأكتُبَنَّ *la’aktubanna* ‘I will write’	**n**
	Non-emphatic verb	فعل غَيْر مُؤكَّد	*fi‘l ğayr mu’akkad*	أكْتُبُ *’aktubu* ‘I am writing’	**m**

A.16 Position 16; The Morphological Feature of Transitivity

Table A.17 The morphological feature of Transitivity category attributes and their tags at position 17

Position	Feature Name				Tag
16	**Transitivity** اللازم والمتعدي *al-lāzim wa al-muta‘addi*				
	Intransitive	لازِم	*lāzim*	نامَ الولَدُ *nāma al-waladu* ‘The boy slept’	**i**
	Singly transitive	مُتعدٍّ إلى مَفعُول واحِد	*muta‘add^{in} ’ilā maf‘ūl^{in} wāḥid*	فَتَحَ الرجلُ الباب *fataḥa ar-rağulu al-bāba* ‘The man opened the door’	**o**
	Doubly transitive	مُتَعدٍّ إلى مَفْعُولَين	*muta’add^{in} ’ilā maf‘ūlayn*	أعطاه دينارًا *’a‘ṭāhu dīnār^{an}* ‘He gave him a dinar’	**b**
	Triply transitive	مُتَعدٍّ إلى ثَلاثَة مَفاعِيل	*muta‘add^{in} ’ilā ṯalāṯati mafā‘īl*	أنبأتُهُ الخبرَ صحيحًا *’anb’tuhu al-ẖabara ṣaḥīḥ^{an}* ‘I announced him the correct news’	**t**

A.17 Position 17; The Morphological Feature of Rational

Table A.18 Morphological feature category of Rational attributes and their tags at position 17

Position	Feature Name				Tag
17	**Rational** العاقل وغير العاقل *al-‘āqil wa ğayir al-‘āqil*				
	Rational	عاقِل	*‘āqil*	قَرَأَ *qara’a* ‘read’	**h**
	Irrational	غَيْر عَاقِل	*ğayr ‘āqil*	نَبَحَ *nabaḥa* ‘bark’	**n**

A.18 Position 18; The Morphological Feature of Declension and Conjugation

Table A.19 The morphological feature of Declension and Conjugation category attributes and their tags at position 18

Position	**Feature Name**				**Tag**
18	**Declention and Conjugation** التَّصريف *at-taṣrīf*				
	Non-Inflected (n, v)	غير مُتصَرِّف	*ǧayr mutaṣarrif*	هُوَ *huwa* ‘him’	**n**
	Primitive / Concrete noun	مُتَصَرِّف – جامِد– اسم ذات	*mutaṣarrif – ǧāmid – ’ism ḏāt*	شَجرةٌ *šaǧarah* ‘A tree’	**t**
	Primitive / Abstract noun	مُتَصَرِّف – جامِد– اسم معنى	*mutaṣarrif – ǧāmid – ’ism ma‘nā*	ذَكاءٌ *ḏakā’un* ‘Intelligence’	**a**
	Inflected / Derived noun	مُتَصَرِّف – اسم مُشْتَقّ	*mutaṣarrif –’ism muštaqq*	كتابٌ *kitāb*^{un} ‘a book’ مكتبةٌ *maktaba*^{tun} ‘a library’	**d**
	Non-conjugated / restricted to the perfect	فعل جامِد– ملازم للماضي	*fi‘l ǧāmid-mulāzim lil-maḍī*	نَعِمَ *na ‘ima* ‘be happy’	**p**
	Non-conjugated / restricted to the imperfect	فعل جامِد– ملازم للمضارع	*fi‘l ǧāmid-mulāzim lil-muḍāri‘*	يَهِيطُ *yahīṭu* ‘scream’	**c**
	Non-conjugated / restricted to the imperative	فعل جامِد– ملازم للأمر	*fi‘l ǧāmid-mulāzim lil-’amr*	هَبْ *hab* ‘suppose’	**i**
	Conjugated / fully conjugated verb	مُتَصَرِّف – فعل تام التَّصريف	*mutaṣarrif – fi‘l tām at-taṣarīf*	يكتبُ *yaktubu* ‘he is writing’	**v**
	Conjugated / partially conjugated verb	مُتَصَرِّف – فعل ناقص التَّصريف	*mutaṣarrif –fi‘l nāqiṣ at-taṣarīf*	كادَ *kāda* ‘close; near or almost’	**m**

A.19 Position 19; The Morphological Feature of Unaugmented and Augmented

Table A.20 The morphological feature of Unaugmented and Augmented category attributes and their tags at position 19

Position	**Feature Name**				**Tag**
19	**Unaugmented and Augmented** المُجَرَّد والمَزِيد *al-muğarrad wa al-mazīd*				
	Unaugmented	مُجَرَّد	*al-muğarrad*	كَتَبَ *kataba* ‘wrote’	**s**
	Augmented by one letter	مَزيْد بِحرف	*mazīd bi-ḥarf*	كَاتَبَ *kātaba* ‘wrote’	**a**
	Augmented by two letters	مَزيْد بِحَرفَيْن	*mazīd bi-ḥarfayn*	اكْتَتَبَ *’iktataba* ‘Subscribed’	**b**
	Augmented by three letters	مَزيْد بِثلاثَةِ أحرف	*mazīd bi-ṯalāṯa*^{t} *’aḥruf*	اسْتَكْتَبَ *’istaktaba* ‘registered’	**t**
	Augmented by four letters	مَزيْد بأربعة أحرف	*mazīd bi-’arba‘a*^{ti} *’aḥruf*	استقبال *’istiqbāl* ‘Reception’	**q**

A.20 Position 20; The Morphological Feature of Number of Root Letters

Table A.21 The morphological feature of Number of Root Letters category attributes and their tags at position 20

Position	Feature Name				Tag
20	**Number of Root Letters** عَدَد أحْرُف الجَذْر *adad ’aḥruf al-ǧaḏr*				
	Triliteral	ثُلاثِي	*ṯulāṯī*	ك ت ب *k t b* ‘wrote’	**t**
	Quadriliteral	رُباعِي	*rubāʻī*	د ح ر ج *d ḥ r ǧ* ‘rolled’	**q**
	Quinqueliteral	خُماسِي	*ẖumāsī*	ز ب ر ج د *z b r ǧ d* ‘chrysolite’	**f**

A.21 Position 21; The Morphological Feature of Verb Root

Table A.22 The morphological feature of Verb Root category attributes and their tags at position 21

Position	Feature Name			Tag
21	**Verb Root** بُنية الفعل *bunya^{tu} al-fiʻl*			
	Intact verb	صحيح	*saḥīḥ*	**a**
	Doubled verb	مضعف	*muḍaʻʻaf*	**b**
	Initially-hamzated verb	مهموز الفاء	*mahmūz al-fā’*	**c**
	Initially-hamzated and doubled verb	مهموز الفاء مضعَّف	*mahmūz al-fā’ muḍaʻʻaf*	**d**
	Initially and finally hamzated verb	مهموز الفاء ومهموز اللام	*mahmūz al-fā’ wa mahmūz al-lām*	**e**
	Medially-hamzated verb	مهموز العين	*mahmūz al-ʻayn*	**f**
	Finally-hamzated verb	مهموز اللام	*mahmūz al-lām*	**g**
	wāw-initial verb	مثال واوي	*miṯāl wāwī*	**h**
	wāw-initial and doubled verb	مثال واوي مضعف	*miṯāl wāwī muḍaʻʻaf*	**i**
	wāw- initial and medially-hamzated verb	مثال واوي مهموز العين	*miṯāl wāwī mahmūz al-ʻayn*	**j**
	wāw-initial and finally-hamzated verb	مثال واوي مهموز اللام	*miṯāl wāwī mahmūz al-lām*	**k**
	yā'-initial verb	مثال يائي	*miṯāl yā’ī*	**l**
	yā'-initial and doubled verb	مثال يائي مضعف	*miṯāl yā’ī muḍaʻʻaf*	**m**
	yā'- initial and medially-hamzated verb	مثال يائي مهموز العين	*miṯāl yā’ī mahmūz al-ʻayn*	**n**
	Hollow with *wāw*	أجوف واوي	*’ağwaf wāwī*	**o**
	Hollow with *wāw* and initially-hamzated verb	أجوف واوي مهموز الفاء	*’ağwaf wāwī mahmūz al-fā’*	**p**
	Hollow with *wāw* and finally-hamzated verb	أجوف واوي مهموز اللام	*’ağwaf wāwī mahmūz al-lām*	**q**
	Hollow with *yā'*	أجوف يائي	*’ağwaf yā’ī*	**r**
	Hollow with *yā'* and initially-hamzated verb	أجوف يائي مهموز الفاء	*’ağwaf yā’ī mahmūz al-fā’*	**s**
	Hollow with *yā'* and finally-hamzated verb	أجوف يائي مهموز اللام	*’ağwaf yā’ī mahmūz al-lām*	**t**
	Defective with *wāw* verb	ناقص واوي	*nāqiṣ wāwī*	**u**
	Defective with *wāw* and	ناقص واوي مهموز الفاء	*nāqiṣ wāwī mahmūz al-fā’*	**v**

Position	Feature Name			Tag
21	**Verb Root** بُنية الفعل *bunya^{tu} al-fi‘l*			
	initially-hamzated verb			
	Defective with *wāw* and medially-hamzated verb	نافص واوي مهموز العين	*nāqiṣ wāwī mahmūz al-‘ayn*	**w**
	Defective with *yā'* verb	ناقص يائي	*nāqiṣ yā’ī*	**x**
	Defective with *yā'* and initially-hamzated verb	ناقص يائي مهموز الفاء	*nāqiṣ yā’ī mahmūz al-fā’*	**y**
	Defective with *yā'* and medially-hamzated verb	ناقص يائي مهموز العين	*nāqiṣ yā’ī mahmūz al-‘ayn*	**z**
	Adjacent doubly-weak verb	لفيف مقرون	*lafīf maqrūn*	*
	Adjacent doubly-weak and initially-hamzated verb	لفيف مقرون مهموز الفاء	*lafīf maqrūn mahmūz al-fā’*	**$**
	Separated doubly-weak verb	لفيف مفروق	*lafīf mafrūq*	**&**
	Separated doubly-weak and medially-hamzated verb	لفيف مفروق مهموز العين	*lafīf mafrūq mahmūz al-‘ayn*	@

A.22 Position 22; The Morphological Feature of Noun Finals

Table A.23 The morphological feature of Noun Finals category attributes and their tags at position 22

Position	Feature Name				Tag
22	**Noun Finals** أقسام الأسم تبعاً للفظ آخره *’aqsām al-’ismi tib$^{‘an}$ li-lafẓi ‘āẖirhi*				
	Sound noun	الاسم صحيح الآخر	*al-’ism ṣahīh al-’āir*	جبل *ǧabal* ‘mountain’ نهر *nahr* ‘river’ درهم *dirham* ‘Dirham (currency)’	**s**
	Semi-sound noun	الاسم شبه الصحيح	*al-’ism šibh aṣ-ṣaḥīḥ*	دَلْو *dalw* ‘bucket’بهو *bahw* ‘hall’	**i**
	Noun with shortened ending	الاسم المقصور	*al-’ism al-maqṣūr*	بُشْرَى *bušrā* ‘glad tidings’	**t**
	Noun with extended ending	الاسم الممدود	*al-’ism al-mamdūd*	سَمَاء *samā’* ‘sky’	**e**
	Noun with curtailed ending	الاسم المنقوص	*al-’ism al-manqūṣ*	القَاضِي *al-qāḍī* ‘the judge’	**c**
	Noun with deleted ending	الاسم محذوف الآخر	*al-’ism maḥḏūf al-’āẖir*	يَدْ *yad* ‘hand’, سَنَة *sana^{h}* ‘year’, and لُغَة *luġa^{h}* language’.	**d**

Appendix B
Summary of Arabic Part-of-Speech Tagging Systems

Tagger	**Corpus used**	**Algorithm (Methodology)**	**Tagset & tagset size**	**Evaluation method**	**Evaluation Metrics**
1- APT: Arabic Part-of-Speech tagger by KHOJA	• 59,040 words of the Saudi `` Al-Jazirah" newspaper, dated 03/03/1999. • 3,104 words of the Egyptian `` Al-Ahram" newspaper, date 25/01/2000. • 5,811 words of the Qatari `` Al-Bayan" newspaper, date 25/01/2000. • 17,204 words of Al-Mishkat, an Egyptian published paper in social science, April 1999.	Statistical and rule-based techniques. Statistical tagger uses the Viterbi algorithm.	The tagset developed by Khoja contains 177 tags: 103Nouns 57 Verbs 9 Particles 7 Residual 1 Punctuation	Stemmer evaluated using a dictionary of 4,748 trilateral and quadrilateral roots.	The test of the stemmer shows an accuracy of 97%.
	<u>Lexicon:</u> 50,000 words extracted from Jazirah newspaper were tagged, and used to derive the lexicon, which contains 9,986 words.				Statistical tagger achieved an accuracy of around 90%

Tagger	**Corpus used**	**Algorithm (Methodology)**	**Tagset & tagset size**	**Evaluation method**	**Evaluation Metrics**
2- POS Tagging of Dialectal Arabic by Duh and Kirchhoff.	1- The CallHome Egyptian Colloquial Arabic (ECA) corpus 2- The LDC Levantine Arabic (LCA) corpus, 3- The LDC MSA Treebank corpus,	LCD-distributed Buckwalter stemmer. Internal stem lexicon combined with rules for affixation. The baseline tagger was a statistical trigram tagger in the form of a hidden Markov model (HMM).	They mapped both sets of tags, the LDC ECA annotation and and the Buckwalter stemmer to a unified, simpler tagset consisting only of the major POS categories. **17 categories.**	ECA Evaluation set Systems: CombileData CombineLex Interpolate – λ Interpolate – λ (t_i) JointTrain(1:4) JointTrain(2:1) JointTrain(2:1) + affix w/ECA+LCA w/ECA+MSA	Accuracy was 58.47% 66.61% improved using affix features and to 68.48% by joint training.
3- Memory-based morphological analysis and part-of-speech tagging of Arabic by Bosch, Marsi, and Soudi	Arabic Treebank version 3.0 **Lexicon** They created a lexicon that maps every word to all analyses.	Memory-based learning (k-nearest neighbor classification) morphologically analyzes and PoS tags unvoweled written Arabic and analyzes it using Tim Buckwalter's Arabic Morphological analyser which is rule-based. They employed the MBT memory-based tagger-generator and tagger. http://ilk.uvt.nl/	They used the same tagset in the Penn Arabic TreeBank.	They evaluated on the complete correctness of all reconstructed analysis in terms of recall, precision and F-score.	The accuracy of the tagger on the held-out corpus was 91.9%. On the 14155 known words it was 93.1%. on the 947 unknown words it was 73.6%

Tagger	**Corpus used**	**Algorithm (Methodology)**	**Tagset & tagset size**	**Evaluation method**	**Evaluation Metrics**
4- Brill's POS tagger and a Morphology parser for Arabic by Freeman	Large corpus of Modern Standard Arabic text. All input Arabic text was assumed to be Windows CP-1256 text using the transliteration scheme devised by Tim Buckwalter and Ken Beesely at Xerox.	Brill's "transformation-based" or "rule-based" tagger.	119 tagset	The system was not evaluated	The system was not evaluated
5- Automatic Tagging of Arabic Text by Diab, Hacioglu and Jurafsky.	The data was transliterated in the `Arabic TreeBank` into Latin based ASCII characters using the Buckwalter transliteration scheme.	Support Vector Machine (SVM) based approach	24 collapsed tags available in the `Arabic TreeBank` distribution. This collapsed tag set is a manually reduced form of the 135 morpho-syntactic tags created by `AraMorph`.	A standard SVM with a polynomial kernel, of degree 2 and C=1.7 Standard metrics of Accuracy (Acc), Precision (Prec), Recall (Rec), and the F-measure, $F_{\beta=1}$, on the test set are utilized	95.49%
6- Part-of-Speech Tagging by Habash and Rambow	The data they used comes from the Penn Arabic Treebank. They used the first two releases of the ATB, ATB1 and ATB2, which are drawn from different news sources. They used the *ALMORGEANA morphological analyzer* which uses the databases (i.e.,lexicon) from the Buckwalter Arabic Morphological Analyzer.	SVM-based Yamcha (which uses Viterbi decoding) rather than an exponential model.	They used a reduced POS tagset (15 tags) along with the other orthogonal linguistic features.	They mapped their best solutions to the English tagset and they assumed gold standard tokenization. Then evaluated against the gold standard POS tagging which is mapped similarly.	On their own reduced POS tagset, evaluating on TE1, they obtained an accuracy score of 98.1% on all tokens.

Tagger	**Corpus used**	**Algorithm (Methodology)**	**Tagset & tagset size**	**Evaluation method**	**Evaluation Metrics**
7- Arabic Part-of-Speech Tagging by Harmain.	(42000 HTML document = 316 MB) mostly from Al-Hayat Arabic newspaper **Dictionary**: they used Buckwalter's dictionary available from the Linguistic Data Consortium (LDC).	Rule-Based	Tagset is unknown.	He did not show any evaluation for his system.	No evaluation done.
8- Hybrid Method for Tagging Arabic Text by Tlili-Guiassa	Texts extracted from educational books in first stage and some Qur'anic text that was tagged using a small tag set.	Hybrid method of based-rules and a machine learning method	The tag set used is the tag set derived from APT	All experiments are performed on texts extracted from educational books in first stage and some Qur'anic text that was tagged using a small tag set and retagged with more detailed tag set.	85%
9- A Hidden Markov Model –Based POS Tagger for Arabic by Al-Shamsi and Guessoum	A training corpus of Arabic news articles has first been stemmed using the stemming component and then tagged manually with their proposed tag set. They examined LDC's Arabic TreeBank corpus (LDC, 2005) that consists of 734 news articles. They have developed a 9.15 MB corpus of native Arabic articles, which were manually tagged using the developed tag set.	They used Buckwalter's stemmer to stem the training data. They constructed trigram language models and used the trigram probabilities in building the HMM model	**55 tagset** They selected the tags that were rich enough to allow a good training and a good performance of the HMM-based POS tagger. At the same time, they tried carefully to make the tag set small enough to make the training of the POS tagger computationally feasible.	They used the F-measure to evaluate POS tagger performance. They computed the *F-measure* as : [2 x Precision x Recall] / [Precision + Recall] where Precision = Ncorrect / Nresponse Recall = Ncorrect / Nkey	97%.