Arabic Word Class Tagging Based on the Analysis of Affix Structure

Suleiman H. Mustafa* and Sari M. Awwad**
*Faculty of Information Technology, Yarmouk University, Irbid, JORDAN, smustafa@yu.edu.jo
**Faculty of Information Technology, The Hashemite University, Al-Zarqa, JORDAN, sari@hu.edu.jo

ABSTRACT

This study was based on a major assumption that the lexical structure of Arabic textual words involves semantic content that could be used to determine the class of a given word and its functional features within a given text. Hence, the purpose of the study was to explore the extent at which we can rely on word structure to determine word class without the need for using language glossaries and word lists or using the textual context. The results indicate that the morphological structure of Arabic textual word was helpful in achieving a rate of success approaching 79% of the total number of words in the sample used in the study. In certain cases, the approach adopted in the investigation was not adequate for class tagging due to two major reasons, the first of which was the absence of prefixes and suffixes and the second was the incapability of distinguishing affixes from original letters. It was concluded that the approach adopted in this study should be supplemented by using other techniques adopted in other studies, particularly the textual context.

Keywords *Arabic Language Processing, Word Class Tagging, Part-Of-Speech Tagging, Morphological Analysis*.

1. INTRODUCTION

Arabic differs from other languages, like English, in its morphological and semantic structures. It is a derivative language in which the basis of word formation is a root (usually trilateral or quadrilateral). Lexical and textual words are formed by applying a set of grammatical and morphological rules. The language is characterized by certain features that attract the attention of researchers in the field of natural language processing. But, while these features provide a formal basis for devising sound computational techniques and algorithms, they impose serious limitations in certain aspects of computation.

This study is intended to determine word classes in a non-vocalized Arabic text on the basis of morphological analysis of textual words. Computer-based processing of Arabic involves four major levels of computation: lexical analysis, morphological analysis, syntax analysis, and semantic analysis. Word class tagging is basically related to the first two levels of computation which focus on the structure of words as opposed to the last two levels which emphasize the structure and meaning of sentences in the language.

The major assumption underlying this study is that the lexical structure of Arabic textual words involves semantic content that can be used to determine the class of a given word and its functional features within a given text. Hence, the purpose of the study is to explore the extent at which we can rely on word structure to determine word class without the need for using language glossaries and word lists or using the textual context. Consequently, the study addresses the following three basic research questions:

1. To what extent we can rely on each type of word affixes (be it grammatical or semantic) to determine word class and word linguistic features?
2. When do we need to refer to the standard morphological forms (known as *awzan*) and the associated morphological affixes to determine word class and word linguistic features, and to what extent we can rely on that?
3. When does the word class tagger fails to determine word class based on word affixes, and what kind of problems are encountered?

Given these questions, the study attempts to build a model of morphological analysis, and use it for identifying Arabic word classes within non-vocalized text in accordance with the taxonomy used for classifying Arabic textual words. At the upper level of this taxonomy lies the classical classification of Arabic words into three major classes; namely: verbs, nouns, and particles.

Each class is further subdivided into a number of subclasses. A verb must belong to one three categories: past, present, or imperative. Nouns, on the other hand, are handled, in this study, under the following categories: infinitive nouns, epithetic infinitive, active participle, passive participle, assimilated adjective, superlative nouns, relative nouns, nouns of place, nouns of instrument, generic nouns, and proper nouns.

2. RELATED WORK

Many research studies have addressed the issue of automatic part-of-speech tagging in many languages, especially English. Some of these studies date back to the early days of computer application in the field of computational linguistics.

For instance, Klein and Simons [10] conducted an experiment to classify English words in a large text. When an unknown word was found, its immediate context was tested by applying a set of contextual rules. They reported a success of 90% of words in the analyzed text. Likewise, Baxendale and Clarke [3] adopted a similar approach. Words which were unknown to the algorithm or not found in special-word dictionaries, were classified as verbs or nouns by relying on contextual information such as number agreement.

These early studies laid the foundation for many other research works that came later, most notably those reported by Eklund [5]. The strategy adopted in Eklund's study relied on storing words and their associated features in a special dictionary. Searching in the dictionary for a given word takes place using the full word. If this fails, a stripping technique is applied. The leftmost letter (left stripping) or rightmost letter (right stripping) would be removed and searching is repeated for the remaining. The results indicated that this approach was successful in tagging 94% of the words used in the study.

Automatic Arabic part-of-speech tagging has not received considerable attention in the literature of natural language processing. The earliest attempts in this regard seem to have come in the context of syntax analysis such as the study conducted by Ibrahim [8]. In his study, the author outlined the rules of word tagging that was needed for syntax analysis. He showed that the semantics of grammatical affixes provide a good basis for assigning words to class categories. He also recognized the importance of standard morphological rules in this regard.

A number of studies have recently been reported in the literature. Some of which came in the context of automatic lexicon generation, such as the investigation reported by Abuleil and Evens [2], who used three techniques: finding phrases, affix analysis, and word pattern analysis, and the investigation reported by Farghali and Sennelart [6].

Other studies reported the results of developing Arabic word taggers. The investigators of these studies attempted various strategies, including: lexical database lookup [7][11], lexicon lookup combined with contextual analysis [13], word affix and pattern analysis combined with user feedback and contextual information based on some keywords [1], finite-state machines [12][4], and combined statistical and rule-based techniques [9]. Unlike other studies, Talmon & Winter [12] developed a computational system for morphological tagging of the Holy Quraan only.

Not all authors have reported the level of accuracy achieved in their investigations, but various approaches have demonstrated different levels of success in word tagging. Based on the available figures, the reported levels of accuracy ranged between 90% and 97%. It is beyond the scope of this paper to go into more details about all these studies, but it informative to review a few of them.

Khoja [9] developed a tagger using a combination of statistical and rule-based techniques. Using a corpus of 50,000 words, she derived a lexicon to be used in the initial step of tagging. If a word is not found in the lexicon, it is stemmed. Affixes and patterns were used to help determine the tag of the word. A statistical tagger, based on lexical and contextual probabilities, was used for ambiguous words. Tests showed that this technique achieved an accuracy of 97% using a dictionary of 4,748 roots.

Abuleil, Alsamra & Evens [1] developed a learning system that can identify Arabic nouns and produce their morphological information and their paradigms with respect to gender and number depending on suffix analysis, word pattern analysis, and user feedback. They also used certain keywords in the context to make judgments about the classes of words in the text. The system was able to make 90.2% correct judgments, of which only about 36% as a result of using morphological analysis.

Von Mol & Paulussen [13] adopted a two-step word tagging procedure. For the first step, they built a lexicon of all morphological standard patterns and their variations with respect to morphological and grammatical prefixes and suffixes. A given textual word is first matched with the entries stored in the lexicon in order to retrieve the tags and other lexical features. If this step fails, the tagger uses the available contextual information depending on the location of the word in a given textual context.

3. AFFIX STRUCTURE

As stated earlier, the major assumption underlying this study is that the affix structure of Arabic textual words involves substantial semantic content that can strongly guide our judgment in determining the class of a given word and its functional features within a given text. As Figure 1 shows, Arabic words are composed of three morphological levels: root, lexical word, and textual word.

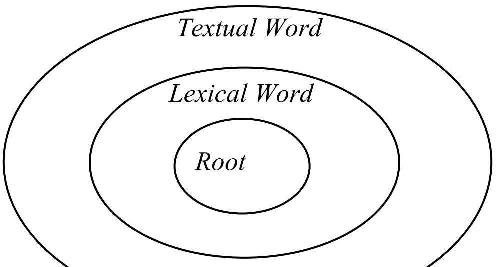

Figure 1: Morphological Levels of Arabic Words

The root represents the basic level from which words of the lexicon are formed according to a set of morphological standard patterns. As Figure 2 shows, a lexical word is composed of a root plus zero or one lexical prefix, zero or one lexical suffix, and zero to two lexical infixes. A textual word is composed a lexical word plus zero to three grammatical or contextual prefixes along with zero to four grammatical or contextual suffixes.

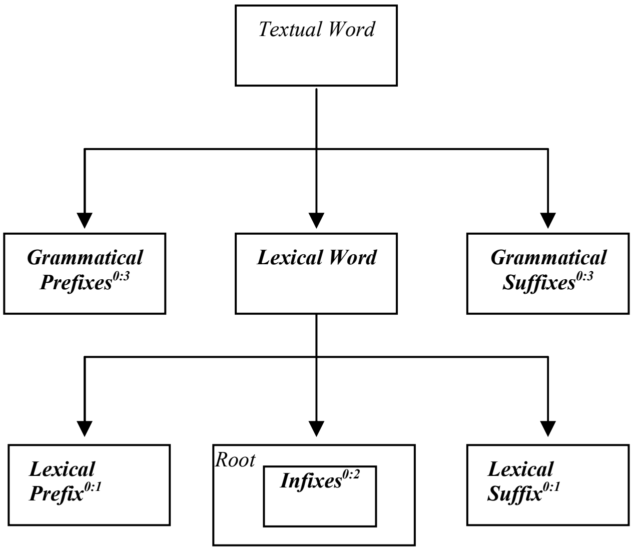

Figure 2: The Morphological Structure of Arabic Textual Words

Given this morphological structure of Arabic textual words, the researchers also recognized the fact that grammatical prefixes and suffixes are combined in different ways where more than one single prefix or suffix is added to a lexical word. As an example, we show in Table 1 and Table 2 how two-part grammatical prefixes and suffixes are formed. The same trend was noticed in respect to the relationships between grammatical prefixes and suffixes. A given grammatical prefix occurs with only a small set of grammatical suffixes.

Table 1: Double Occurrences of Grammatical Prefixes

Prefix	أ	ا	ب	ت	س	ف	ك	ل	ن	و	ي	ال	وا
أ	x		x	x	x	x	x	x	x	x	x		
ب	x											x	
ت													
س	x			x					x		x		
ف	x		x	x	x		x	x	x	x	x	x	
ك	x											x	
ل	x			x					x		x	x	
ن		x											
و	x		x	x	x	x	x	x	x		x	x	
ي													
ال													
وا													

Table 2: Double Occurrences of Grammatical Suffixes[1]

Suffix	ا	ت	ك	ن	ه	ة	و	ي	تم	تما	تن	ها	هما	هن	هم	كم	كما	كن
ا		x	x	x	x				x	x	x		x	x	x	x	x	x
ت				x	x			x				x	x	x	x	x	x	x
ك																		
ن	x				x			x				x	x	x	x			
ه																		
ة																		
و	x		x	x	x							x	x	x	x	x	x	x
ي				x	x	x						x	x	x	x			
ات					x								x	x	x	x	x	x
تم	x						x											
تا				x	x			x					x	x	x	x	x	x
نا			x		x								x	x	x	x	x	x
ون					x							x	x	x	x			
ين					x							x	x	x	x			

Having established these intra- and inter-relationships, the next step was identifying the contribution of these prefixes and suffixes in word tagging. Some of these affixes rarely occur in the formation of textual words and, hence, might not contribute much to the tagging process.

A number of grammatical affixes provide strong indications for certain classes of words, while some others can only be taken as providing weak clues for the word tagger. Table 3 presents an example of the role played by grammatical affixes in identifying verbs and nouns.

Based on this analysis, the authors believed that the information provided by grammatical affixes would be useful but not sufficient to determine the exact part-of-speech label within the two major categories: nouns and verbs. Similar analysis had to be performed on lexical affixes, as determined by the Arabic standard patterns.

[1] Note that some grammatical suffixes are omitted from the table rows, because they are not followed by any other grammatical suffixes

Certain lexical prefixes or suffixes come with certain classes of words. The lexical prefix "*must*" ("مست"), for instance, comes only with nouns. When combined with a given infix, a lexical prefix can help determine the word tag with high certainty.

Table 3: The Role of Grammatical Prefixes and Suffixes in Identifying Verbs and Nouns

Class	Prefixes		Suffixes	
	High Certainty	Low Certainty	High Certainty	Low Certainty
Verb	أسأ، أست، أسن، أسي، أفن، أفي، ألت، ألي، أوت ، سأ، ست، سي، فسأ، فست، فسي، وسأ، وست، وسي	ا، أت، أف، أو، ف، ل، ول،و، ن، ون، ي، وي، ولي، ولن، ت، أي،أوي، س، لن، وسن، أن، فلن، فن، فسن، ألن، أوأ، وأ، أفأ، أأ، ولأ، لأ، ولت، فلأ، فلت، فلي، في، فت، أفت، فأ، لت، لي	تما، تمو، تن، ناك، ناكم، ناكما، ناكن، ناه، ناها، ناهم، ناهما، ناهن، نه، نها، نهم، نهما، نهن، ني، وا، ونه، ونها، ونهم، ونهما، ونهن، ينه، ينها، ينهم، ينهما، ينهن، يه، يها، يهم، يهما، يهن	ت،تك، تكم، تكما، تكن، ته، نا، ه، ها، هم، هما، هن، ون، ي، ين، ان، ته ،تهم، تهما، تهن، اه، وك، وكم، وكما، وكن، وه، وها، وهم، وهما، وهن، اك، اكم، اكما، اكن، اها، اهم، اهما، اهن، ن
Noun	أل، أف، أو، ف، ل، و، ول، ك، وك، أك، فك، وال، ، ب، وب، فك، فب، أول ، أفب، أب، أوب،	أبأ، أبال، أكال،ألل، بأ، بال، فال، فكال،فل، فلل، كأ، كال، لل، وبال، وكال، ولل	اتك، اتكم، اتكما، اتكن، اته، اتها، اتهم، اتهما، اتهن، تاك، تاكم، تاكما، تاكن، تان، تاهم، تاهما، تاهم، تاهن، تاي	ا، ان، ون، ين، تا، تاه، تاها، ك، كم، كما، كمو، كن، كي، ات، ة، ية

However, as Table 4 indicates, the analysis of lexical affixes showed that none of the lexical prefixes or suffixes can be a determining factor in the tagging process. Therefore, we had to further examine the relationships between grammatical prefixes and lexical prefixes and how they can contribute to the word tagging process. Likewise, we examined the corresponding relationship between the lexical suffixes and grammatical suffixes. Furthermore, we explored the kind of information that can be provided by the presence of combinations of prefixes and suffixes.

Table 4: The Role of Lexical Prefixes in Identifying Word Classes

Class	Class	High Certainty	High Certainty
Verb	فعل	-	تم، است، أ، ان، ت
Infinitive noun	مصدر	-	است، أ، ان
Epithetic infinitive	مصدر صناعي	-	است،ان
Passive participle	اسم مفعول	-	مست، مت، متم، من، م
Active participle	اسم فاعل	-	مست، مت، متم، من، م
Assimilated adjective	صفة مشبهة	-	أ، ت
Superlative noun	اسم تفضيل	-	أ
Relative noun	اسم منسوب	-	-
Noun of place	زمان ومكان	-	م
Noun of instrument	اسم آلة	-	م
Generic noun	الاسم العام	-	م

4. THE TAGGING PROCEDURE

The research approach adopted in this study is based on four basic steps of analysis that makeup the framework of the processing heuristic (see Figure 3).

Given a textual word W for which the class tag is to be determined, the analysis process starts by checking if W is one of the particles listed in the list of articles, pronouns, and similar words that have no morphological basis in Arabic. If W is not a particle, its grammatical prefixes and suffixes as defined by grammarians (such as connected conjunctions, prepositions, and the definite article) are identified and used in the analysis.

If this type of affix analysis fails to provide a satisfactory result, W is exposed to further affix analysis based on the lexical structure entailed by the standard Arabic morphological forms, a structure which involves prefixes, infixes, and suffixes. In certain cases, the length of W is needed to determine its standard form.

The approach described above is supported by a number of data tables that are essential for identifying the class of a given word and its semantic features. The tables included a list of prefixes, a list of suffixes, a list of standard forms, a list of relations between prefixes and suffixes, and a list of particles. All lists are represented in such a way that can lead to efficient access (namely, binary search).

The information given under each affix entry includes: affix letters, affix type, length, possibility of conflict with original letters, word class, gender, number, state, and relationships with other affixes.

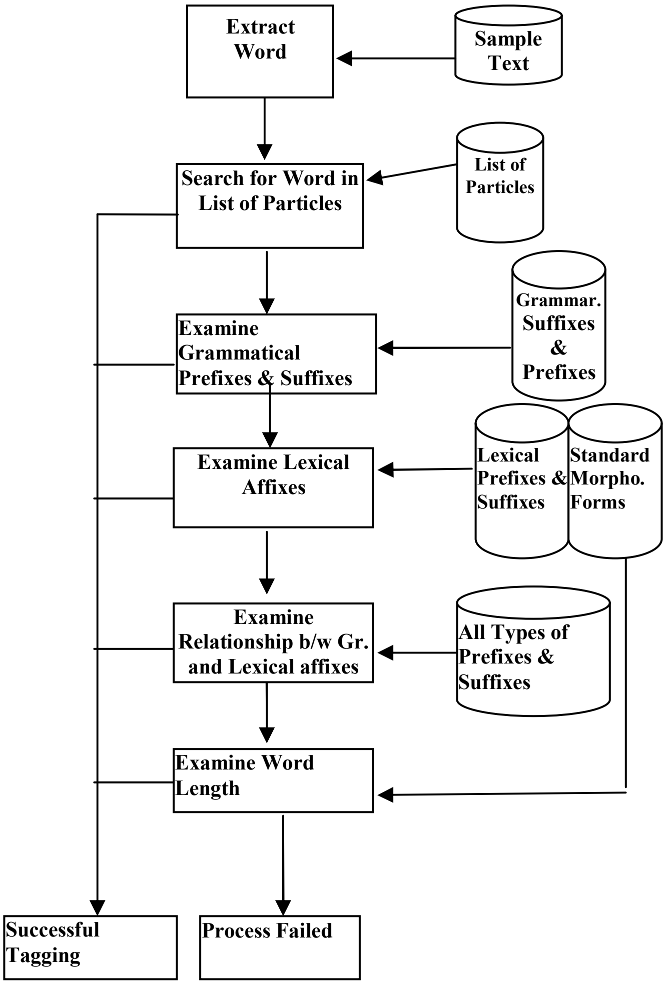
Extract Word
Sample Text
Search for Word in List of Particles
List of Particles
Examine Grammatical Prefixes & Suffixes
Grammar. Suffixes & Prefixes
Examine Lexical Affixes
Lexical Prefixes & Suffixes
Standard Morpho. Forms
Examine Relationship b/w Gr. and Lexical affixes
All Types of Prefixes & Suffixes
Examine Word Length
Successful Tagging
Process Failed

Figure 3: The word tagging procedure used in the study

5. RESULTS AND DISCUSSION

To test the assumptions underlying this study, the class tagging approach outlined above was translated into a program written in C++. Two samples of Arabic text were selected for analysis and testing, the first comprised 2000 words, while the other sample comprised 1,500 words. To verify the results of the program, the two samples were analyzed by hand. Table 5 and Table 6 show the results of this manual analysis.

Table 5: Classes of words in the sample as determined by manual analysis

Word Class	Number	%
Past Verb	٤٥٥	١٣.0
Present Verb	٢٥٠	٧.١٤
Imperative Verb	٠	٠.0
Infinitive	٤٦٤	١٣.٢٦
Generic Noun	٤٣٢	١٢.٣٤
Proper Noun	٢١٧	٦.٢
Relative Noun	٢٩٧	٨.٤٨
Assimilated Adj.	١٣٦	٣.٨٨
Active Participle	١٧٨	٥.٠٨
Passive Participle	٨٨	٢.٥١
Superlative Noun	١٨	٠.٥١
Noun of Place	٣٨	١.٠٨
Epithetic Infinitive	٤	٠.١١
Noun of Instrument	0	0.0
Particles	٩٢٣	٢٦.٣٧
TOTAL	**٣٥٠٠**	١٠٠

Table 6: Classifying words in the sample with respect to the presence of affixes

Word Category	Num	%
Words Starting with Grammatical Prefixes	١٦٥٩	٤٧.٤٠
Words Ending with Grammatical Suffixes	١١٦٥	٣٣.٢٨
Words with no Grammatical Affixes	٦٧٦	١٩.٣0
Words Starting with Lexical Prefixes	٧٠٩	٢٠.٦0
Words Ending with Lexical Suffixes	٥٥	١.٥٧
Words with no Prefixes and Suffixes	٣٩٧	١١.٤٠

The results of the program showed that the word class tagging approach used in this study succeeded in categorizing the majority of words into the three classical categories: verbs, nouns, and particles. As Table 7 shows, it was able to recognize 88.46% of the nouns contained in the text and 84.11% of the verbs. Since particles were recognized on the basis of a lookup table, the automatic tagging procedure reported complete success in handling this class of words.

But, when further analysis was carried out to determine the exact word class within these categories, the program was able to achieve a rate of success approaching 79% of the total number of words in the sample. Table 8 shows the results of the automatic tagging procedure developed in this study in comparison with results determined by the manual analysis. It was able, for instance, to achieve a rate of success in identifying infinitives of about 78%, of which about 57% with high level of accuracy and about 21% with low level of accuracy.

Table 7: Results of the tagging procedure in recognizing the major word classes

Class	**Manual Analysis**	**Automatic Analysis**	
	Number	**Number**	**Accuracy Rate**
Nouns	١٨٧٢	١٦٥٦	٨٨.٤٦%
Verbs	٧٠٥	٥٩٣	٨٤.١١%
Verbs/Nouns	-	٩٨	٣.٨0%
Particles	٩٢٣	٩٢٣	١٠٠%

Table 8: Performance of the tagging procedure in determining noun categories with high certainty or low certainty

Nouns	Manual Analysis	Automatic High Certainty		Automatic Low Certainty		Total	
	Num	Num	Accuracy %	Num	Accuracy %	Num	%
Infinitive	٤٦٤	٢٦٦	٥٧.٣٢	٩٧	٢٠.٩٠	٣٦٣	٧٨.٢٣
Generic Noun	٤٣٢	٩٤	٢١.٧٥	١٩٥	٤٥.١٣	٢٨٩	٦٦.٨٩
Relative Noun	٢٩٧	٥٩	١٩.٦٨	١٨٢	٦١.٢٧	٢٤١	٨١.١٤
Assimilated Adj.	١٣٦	٥١	٣٧.٥	٤١	٣٠١٤	٩٢	٦٧.٦٤
Active Participle	١٧٨	٢	١.١٢	١١٧	٦٥.٧٣	١١٩	٦٦.٨٥
Passive Participle	٨٨	٤٤	٥٠	٣٣	٣٧.٥	٧٧	٨٧.٥0
Superlative Noun	١٨	٠	٠	١٤	٧٧.٧٧	١٤	٧٧.٧٧
Noun of Place	٣٨	٢٢	٥٧.٨٩	٠	٠	٢٢	٥٧.٨٩
Proper Noun	٢١٧	٣٦	١٦.٥٨	٠	٠	٣٦	١٦.٥٨
Epithetic Infinitive	٤	٠	٠	٤	١٠٠	٤	١٠٠.00
Total	١٨٧٢	٥٧٤	٣٠.٦٦	٦٨٣	٣٦.٤٨	١٢٥٧	٦٧.١٥

As we examine the errors made by the tagging procedure, the results indicate that the morphological structure of Arabic textual word is not adequate for class tagging in certain cases. This can be attributed to a number of reasons. The first is the ambiguity resulting from having certain prefixes or suffixes that are identical to original letters in some textual words.

Errors in this category constituted 7.75% of the total number of erroneous results. The second category of errors (which constituted about 11.4% of the number words in the sample) resulted from the absence of grammatical and morphological affixes to be used in determining word classes. The third reason is that some prefixes and suffixes are applicable to both nouns and verbs, which reduces the possibility of making the right judgment.

6. CONCLUSION

If a human is given an Arabic text and asked to find out the part of speech for each word in the text, s/he will use a number of sources of information to do so. The most important of these sources is the cognitive lexicon. Other sources evolve around the information provided by the lexical structure of words, the syntactic structure and the textual context. Previous research studies have attempted to use all these sources for word tagging. In this study, an attempt was made to see how far we can rely on the affix structure of Arabic words in automatic part-of-speech tagging.

Maintaining a huge lexicon of lexical and semantic information involves a high cost. On the other hand, the information embedded in the text itself could be treated as a major source of information in word tagging. Starting from this assumption, the investigation was carried out using a procedure devised by the authors based on the well established taxonomy of Arabic words.

Given the size and type of sample used and the program implementation of the algorithm, we may conclude that the morphological analysis of word affix structure can be used successfully for determining a high percentage of word classes.

However, in certain cases, the approach adopted in this investigation was not adequate for class tagging due to two major reasons, the first of which was the absence of prefixes and suffixes and the second was the incapability of distinguishing affixes from original letters. The results suggest that we should supplement this approach by using other techniques adopted in other studies, particularly the textual context.

REFERENCES

1. Abuleil, S., Alsamara, K., and Evens M., (2002). Acquisition System for Arabic Noun Morphology. (available at:

http://www.cs.um.edu.mt/~mros/WSL/papers/ abuleil:alsamara: evens.pdf)

2. Abuleil, S., and Evens, M. (1998). Discovering Lexical Information by Tagging Arabic Newspaper Text. (available at: http://citeseer.nj.nec.com/ correct/583651)

3. Baxendale, P.B. and Clarke, D.C. (1966). Documentation for an Economical Program for the Limited Parsing of English: Lexicon, Grammar, and Flowcharts. San Jose (CA): IBM San Jose Research Laboratories.

4. Diab, Mona, Hacioglu, Kadri and Jurafsky, Daniel (2004). *Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks.* Proceedings of HLT-NAACL. [Available as pdf at: http://www1.cs.columbia.edu/~mdiab].

5. Eklund, R. (1993). A Probabilistic Word Class Tagging Module Based on Surface Pattern Matching, pp. 83–95. (available at: http://www.ida.liu.se/~g-robek/ nodalida93tag.pdf)

6. Farghali, A., and Senellart, J. (2003). Intuitive Coding of the Arabic Lexicon, MT summit IX Workshop on Machine Translation for Semitic Languages, U.S.A.. (available at: http://www.systransoft.com/Technology/ 2003_MTIX_AR. pdf)

7. Habash, N.(2004). Large Scale Lexeme Based Arabic Morphological Generation, University of Maryland Institute for Advanced Computer Studies, U.S.A. (available at: http://www.nizarhabash.com/ publications/taln-04-1.pdf)

8. Ibrahim, F. (1986). A Syntactically-Based Preprocessor for A Limited Experimental Arabic Document Retrieval System, Ph.D. Thesis. Loughborough University of Technology.

9. Khoja, S. (2001). APT: Arabic Part-of-Speech Tagger, Proc. Of the Student Workshop at NAACL. (available at: http://archimedes.fas. harvard.edu/mdh/arabic/NAACL. pdf)

10. Klein, S. & Simons, R.F. (1963). A Computational Approach to Grammatical Coding of English Words. ACM, pp. 334-347.

11. Tahir,Y., Chenfour, N., and Harti, M.(2003). Realization of A Morphological Analyzer for Arabic Language Text. Workshop on Information Technology, Rabat, Morroco.

12. Talmon, R., and Wintner, S.(2001) Morphological Tagging of the Qur'an. (available at: http:// cs.haifa.ac.il/~shuly/publications/talmon-wintner-eacl03.pdf)

13. Van Mol, M., and Paulussen, H. (2004). The Semi-Automatic Tagging of Arabic Corpora. (available at: http://www.ilt.kuleuven.ac.be/ilt/ arabic/_pdf/ talnfez2004.pdf)