Computerizing Arabic Morphology

Sane M Yagi
Sharjah University

1. Introduction

This topic is not novel at all; several similar studies [eg, 1, 3, 4], have been conducted over the past 20 years, but in isolation from one another and without benefiting from each other's achievements. Many theoretical problems in morphology have to be resolved for spell-checking to become an effective feature of Arabic-enabled word processors, yet the very theoretical foundation upon which the morphological parsing of these word processors is based remains enigmatic.

Arabic computational linguistics suffers from two main predicaments; one relates to the traditional grammatical description of the language and the other to the lack of cumulative effort on the part of computational linguists themselves. There is no doubt that Arab grammarians' description of the language is extremely sophisticated, yet it has been characterized by a high degree of ratiocination that complicates an otherwise useful description. Therefore, a systematic effort must be made to render the grammatical description more congenial to computers. Towards this end, several individual attempts have been made [eg, 10, 15]. Arab grammarians need to adopt tools that rely more on empirical techniques and less on introspection. They need to make use of recently developed methods in corpus, functional, and structural linguistics.

As to the other critical problem, the lack of cumulative computational linguistic efforts, achievements –when made– are guarded jealously and concealed from the rest of the research community for potential financial gain. Consequently, there is little genuine knowledge in the public domain and, as a result, any new research will have to start from scratch. This is

true in terms both of the computer code and of the linguistic corpus they use. A researcher who wants to write a concordancer, for example, needs first to write the code for root extraction, and so will the dictionary, automatic translation, and computer-assisted language learning software developers, to cite but some of the most obvious examples. In order toachieve their goals, they will also need to develop their own electronic corpus when other colleagues must have done so before them.

It is a recognized fact that Arab public institutions must take the lead in the efforts to computerize the language. But in the absence of a serious commitment towards that, non-profit oriented academics are the ones who will have to shoulder such a responsibility, and some have begun to do so [eg, 2, 7, 8, 12]. If their individual efforts, however, are to bear fruit, their achievements will have to be disclosed and made available to others to build on. Concealing codes and protecting corpora will only hinder progress in the field. Single-handed efforts hardly ever culminate in monumental achievements, but cumulative group efforts can.

The focus of this paper is to show what needs to be done to overcome the two problems facing Arabic computational linguistics. At first, there will be a statistical description of verb roots to demonstrate how the morphological system can be described in a computer-congenial way. Then an algorithm for root extraction will be presented. This is an exploratory study as part of an on-going project at Sultan Qaboos and Sharjah universities, and it is hoped that the actual code will be put in the public domain as soon as a satisfactory algorithm is established.

2. Morphological Investigation

To study the nature of morphological patterns in the absence of a representative corpus of Modern Spoken Standard Arabic, dictionaries can be used as substitutes. Most modern Arabic dictionaries have good collections of the patterns available to native speakers. The dictionary used for this study is *Al-Mu'jam Al-Arabi Al-Asasi* [9], a contemporary dictionary compiled by the Arab League to account for modern standard terminology and intended to serve second language learners and native speakers alike. Different from most Arabic dictionaries, it is not laden with archaic words. Furthermore, *Al-Mu'jam* contains great many loan words that other dictionaries would brush aside as not belonging to Arabic. It treats their most basic forms (including vowels) as root entries.

A random sample of words was collected from this dictionary. The sample consisted of a list of roots and the number of words derived from each. The root of the first entry in each page was written down, then the derivative entries listed under this root were counted and their sum noted down. The sample consisted of 1229 root entries, a number that corresponds to the number of pages in the dictionary. Our sample constitutes 4.9% of the total number of root entries in the dictionary. When derivative entries are taken into account, the data size swells to 13,633 words in all. The random selection of roots and the data size make our corpus more or less representative of Arabic, to the extent that our dictionary is.

Once the list of roots was compiled, **it was** entered into the computer for statistical analysis. Each root was classified, for the most part, according to traditional categories, those that are often used to characterize Arabic roots and derivational patterns:

1. Number of root-constituents: Arabic roots are triliteral, non-triliteral, or foreign. Non-triliteral is a label used to cover quadreliterals, quinqueliterals (if there is such a type), adverbs, prepositions, non-finite verbs, and particles. The smallest number of constituents is three; so there are no biliteral or monoliteral roots.
2. Type of root-constituents: saheeh, all constituents are strong consonants or one or two of the constituents are semi-vowels, mu'tall.
3. All-consonantal roots are labeled as mahmouz (containing a glottal stop, a hamza), or muda'af (containing a geminated consonant), or saalim (if consisting of consonants that are neither glottal nor geminated).
4. Glottalized roots are initial glottalized, medial glottalized, final glottalized or twice glottalized, depending on which of the root-constituents is a glottal stop and how many glottal stops there are.
5. Semi-vowel inclusive roots (mu'tall) are initial semi-voweled (mithaal), medial semi-voweled (ajwaf), or final semi-voweled (naqis), according to which of the root-constituents is a semi-vowel.

6. Initial semi-voweled roots are either initial-w (mithaal wawi) or initial-y (mithal ya?i) roots, depending on whether the semi-vowel root-constituent was /w/or /y/.
7. Medial semi-voweled roots are either medial-w roots (ajwaf wawi) or medial-y roots (ajwaf ya?i).
8. Final semi-voweled roots are either final-w (naqis wawi) or final-y (naqis ya?i) roots.

9. Contentious semi-voweled roots (mushtarak) are initial contentious, medial contentious, or final contentious, depending on which root-constituent is the subject of argument.

3. Summary of Statistical Results

In terms of number of root-constituents, the stems that consist of three radicals make up the vast majority (88.73%) of bare stems, whilst non-triliterals and foreign roots make up the rest (6.04% and 5.22% respectively). With regard to the sum of words derived from these roots, however, triliterals account for 98.14%; each root entry produces an average of 12.57 words, whilst non-triliterals and foreign entries produce 2.35 and 1.38 words per entry respectively. Neither non-triliterals nor foreign roots are productive; i.e., only one, two, or three words are found per root.

As for constituent types, about three-quarters of the entries in the corpus are all-consonantal roots and only 25.75% are semi-vowel inclusive roots. The two types, however, are about equal in productivity (the average numbers of words produced by an entry are 12.09 for all-consonantal and 12.41 for semi-vowel inclusive roots).

The all-consonantal root divides into salim, which constitutes more than three-quarters, geminated (12.8%), and glottalized (9.14%). Compared to the overall data, salim alone comprises more than half the root entries (57.94%), whilst geminated and glottalized make 9.52% and 6.79%, respectively, of the total root entries. Salim, geminated, and glottalized are comparably prolific, giving 12.26, 13.03 and 9.32 words per entry respectively.

Glottalized roots have a small share of overall root entries, but they have been classified into four groups: initial glottalized which makes up less than half the glottalized entries (44.16%), medial glottalized about a quarter of them (25.97%), final glottalized less than a quarter (23.38%), and twice glottalized a mere 6.49% of the glottalized entries. Compared to the overall list of entries, their shares are quite small: 3%, 1.96%, 1.59% and 0.44% respectively. The most productive of these roots is the initial glottalized (11.38 words per root), second to it is the final glottalized (10.83), followed by the medial glottalized (5.90) and the least productive is the twice glottalized (3.60 words per root).

Semi-vowel inclusive roots constitute one fourth of the total corpus, but they divide into four categories and the medial semi-voweled comprises almost half of them. Final semi-voweled includes 30.82%, initial semi-voweled 17.47%, and initial-final semi-voweled a mere 2.74% of all semi-voweled roots. Medial-final semi-voweled (e.g., kawa and rawa) behaves as final semi-voweled roots; hence, they are included in the latter category. In relation to the total corpus, initial, medial, final, and initial-final semi-voweled roots comprise 4.50%, 12.61%, 7.94%, and 0.71% respectively. All four categories, however, are comparable in productivity; the average number of derivatives per entry is 13.29 words per entry for the initial semi-voweled, 11.83 for the medial semi-voweled, 13.00 for the final semi-voweled, and 10.37 for the initial-final semi-voweled.

To sum up, here is a table where the frequency of each stem type is indicated in relation to the whole set of data. Major categories that include subcategories (e.g., all-consonantal, semi-voweled, and glottalized) have not been included here, since listing the elements within a category compensates for this.

Stem Type	**%**
Salim	57.9
Geminated	9.52
Initial Glottalized	3.00
Medial Glottalized	1.96
Final Glottalized	1.59
W-Initial Semi-Voweled	4.00
Y-Initial Semi-Voweled	0.62
W-Medial Semi-Voweled	7.91

Y-Medial Semi-Voweled	4.35
W/Y-Medial Semi-Voweled	0.27
W-Final Semi-Voweled	2.58
Y-Final Semi-Voweled	3.20
W/Y-Final Semi-Voweled	1.95

Table 1: Frequency of Morphological Patterns

4. Discussion and Implications

The results above are informative in assessing the degrees of prevalence and productivity of each root type and subtype. They reiterate that Arabic roots are predominantly triliteral and, more importantly, that they constitute 88.73% of Arabic bare stems. Furthermore, they indicate that triliterals are the most productive roots in the Arabic language, with their derivatives constituting 98% of the derivatives in the corpus. It is recommended, therefore, that school curricula emphasize triliterals and pay the highest attention to them since they are responsible for the vast majority of words in the language.

Looking at root constituents from the standpoint of their type, it has been established that roots with semi-vowel constituents are a substantial minority of Arabic roots; hence, they should receive a proper morphophonemic description that accounts for the transformation of the vowel [aa] into a semi-vowel, [w] or [y]. Special attention needs to be given to medial and final semi-voweled roots since they constitute the most frequent categories of this root type.

Studying the Arabic morphological system with a view to the frequency of the various morphological patterns allows the linguist to decide on the relative importance of the derivational rules they formulate. They can rank-order rules in terms of their productivity, so that language teachers, for example, can concentrate their efforts on those with the widest range of application.

Arabic derivational patterns, as a case in point, should be treated differentially in accordance with their importance. Of Arabic entries in the dictionary, 89% are triliteral, approximately 4% are quadriliteral, and about 5% are loan words; if quadriliterals are emphasized as much as triliterals in the linguistic treatment of derivational patterns, the morphological system will be complicated unnecessarily. The rules dealing

with triliterals are clearly of greater importance and, hence, they deserve more attention on the part of the linguist than those relevant to quadrilaterlas. The traditional approach to morphology confounds predominant derivational patterns with a labyrinth of rules contrived for the sake of but a small minority of word types. Whilst linguistics must account for all types of stems, the description must not be lopsided.

Linguistic description must be based on widely used language patterns, and the majority of conclusions drawn from linguistic analysis must likewise be pragmatic and applicable to the language as a whole. If some abnormal patterns are used occasionally, by a minority of speakers or in special circumstances, the linguist need not obscure a clear system with such patterns since either they are infrequent or limited in applicability.

This will prevent the invincible tendency of linguists to indulge in abstract argumentation and sophistry of little pragmatic value. Whether Arabic basic morphemes are best characterized by masdars, nouns, or past tense verbs, for example, is of little significance for the language user or the computer programmer. Similarly, modern speakers of Arabic are untroubled by many cases of metathesis, eg, [?ayisa] for [ya?isa]; either such words are not in their repertoire, or they are simply unaware of the metathesis in them.

It is disturbing to find modern Arabic morphology textbooks containing examples of archaic words, of abandoned derivational patterns, and of rules that modern users refuse to observe. For example, the imperative forms of [waqaa] and [wa'aa], [qi] and ['i] respectively, are not in use. Alternative forms are used in order to avoid saying monosyllabic utterances or writing one letter words. Archaic derivational patterns like the mazeeds of triliterals and quadriliterals, iFMawwaLa and iFManLaLa respectively, are abandoned by the language user. It is virtually impossible to hear modern speakers saying words like [ijlawwaza] and [i'lawwaTa] on the one hand, or [ihranjama] on the other; nevertheless, morphology textbooks continue to cite these very same examples to illustrate the abandoned derivational patterns [5]. Perhaps here lies the answer to Subhi Al-Salih's question [6]: Why do **we** see our creative writers killing by neglect patterns in use while reviving the quaint, the obsolete, the abandoned? It may be because they were taught such patterns at school without being told that they were archaic. If the aim is to preserve the language, it is doubtful that ignoring present usage and insisting on

obsolete rules will achieve that. Languages evolve in order to reflect their speakers' changing ways.

When the linguistic description is based on common usage, it quickly becomes apparent which patterns are favored by speakers. The short vowels that follow the medial radical in Arabic stems are unpredictable; consequently, speakers pay little attention to them when they derive present tense forms. Any derivational rules for the present tense that will make reference to the medial stem-radical's short vowel will not be heeded. They will remain part of the linguistic heritage but not relevant to usage or to the computerization of the language.

A statistical analysis gives empirical evidence that morphological studies need to take into account. It is not adequate in the present information age to use intuition as evidence of prevalence. Neither will intuition do for judgments about the well-formedness of expressions, especially when massive corpora of texts and discourse that were actually produced by speakers of the language in natural contexts of communication are available to the linguist. This enormous body of linguistic material needs to be harnessed by the linguist, and statistical evidence must be derived to show the direction of current usage, the most dominant linguistic patterns, and the latest stages of linguistic evolution. Without such corporal and statistical evidence, linguistic description will only be groping in the dark and linguistic theorization will be divorced from reality. If machines are ever to become efficient in processing Arabic natural language, its linguistic description will need to be made more congenial to computers.

Peculiarities, rule-deviance, infrequent patterns, and unique usage must all be identified and classified separately from dominant patterns, as was done in the statistical table above. They can still be studied and generalizations can be made about them, but these generalizations will only be applicable to members in the same class. Furthermore, they will carry less weight in the overall description of the language than generalizations that apply to dominant patterns.

When all these observations are taken into account, the description of the morphological system will be rendered machine-friendly. Computational linguists need to know the degree of probability of occurrence for all linguistic patterns, the degree of productivity of linguistic rules, and the extent of their applicability to language as spoken or written by modern

users. With a statistical account of all patterns and a corpus-based description of the language, computational linguists can write algorithms that will account for the general, and design filters that will specifically apply to the peculiar, rule-deviant, or infrequent. But if computational linguists are forced to operate with a speculative linguistic description, they too will be groping in the dark, and they will write unnecessarily cumbersome algorithms that will fail to account for prevalent patterns.

5. Algorithm

On the basis of the statistical conclusions above, an algorithm was designed. Here is a brief description:

- Get the *word* for which the root is to be extracted.
- Filter this *word* through the graphological purification filter. This clears the word of terminators and any possible punctuation marks

that may be attached. This step is necessary because the shape of Arabic letters changes in accordance with whether they are word-initial, medial, or final, whether they are the connecting or non-connecting type, what vowels are adjacent to them, and whether diacritics are marked or not. The graphological filter standardizes word forms and transforms the Arabic non-linear writing system into a linear one.

- Match the *word* with a list of rootless words. If a match is found, then return it as a root. Rootless words include all function words, aplastic nouns (*jaamid*), and all possible combinations of these.
- Otherwise, search for the *word* in our dictionary database. This is root-based like traditional Arabic dictionaries, but it differs in at least two ways: (i) the database includes all foreign words that are in use, and (ii) entries include no allographs. If a match is found in the dictionary, then return *word* itself as the root.
- If the *word* is not in a form identical to its root, then it must have been derived from the root by the addition of prefixes and/or suffixes according to a specific derivational pattern; therefore, to extract the root, the *word* is sent through an affix-filter, where attachments are identified. The filter contains both single affixes (eg, the imperfect verb marker of the feminine third person singular, [ta]) and sequences of affixes (eg, the sequence of future and imperfect verb markers for the feminine third person singular, [sata]). Long prefixes and suffixes consisting of three or more

letters get marked before short affixes. If removing the identified affixes would leave behind only three consonants, then most likely the remaining letters are the word's root constituents. Otherwise, the affixes are simply marked and the *word* is outputted from the filter intact.

- Next, infixes need to be identified, so the *word* is sent to a vowel extraction filter. Vowels are identified as potential infixes; they are potential because Arabic represents vowels and semi-vowels in identical forms; only the latter may be root constituents. Because the vowels [u] and [i] and the semi-vowels [w] and [y] are represented by the same symbols, and since semi-vowels are root constituents in around 25% of Arabic bare stems, the simplistic procedure of removing all vowels will also remove all semi-vowels, thus resulting in root extraction failure in one quarter of Arabic stems. To avoid this, the *word* goes through a morphological pattern matching process that will be commented on shortly.
- Once extra vowels have been marked and semi-vowel constituents have been identified, a check is performed to know if one of the constituents in the attempted root just derived is the pure vowel [aa]. If so, it is coverted to [w]. This decision emanates from the principle that Arabic roots can never contain vowel constituents. The decision to convert it to the semi-vowel [w] is purely arbitrary; nevertheless, renowned grammarians from the Middle Ages support a similar arbitrary decision. Ibn Jinni (1993) says, "If ambiguous [aa] occurs in a medial position in a root, then upon the recommendation of Seebawayhi... it must be viewed as originating from a [w]". I am of the opinion, however, that it is more pragmatic and easier to explain if all [aa] root constituents were to be treated as allographs of the grapheme [w]; hence, they have been encoded as such in the dictionary database of the present root extractor. Our statistical analysis has shown that most semi-voweled roots contain [w] constituents anyway; consider the following table:

Semi-Voweled	% (total corpus)	% (Semi-Voweled
W-inclusive	14.49	58.24
W/Y-inclusive	2.22	8.92
Y-inclusive	8.17	32.84

Table 2: Frequency of Vowel-Inclusive Patterns

- What remains unmarked of the *word* processed so far is most likely root constituents. To verify that, the *word* is matched against an exhaustive list of morphological patterns where derivational prefixes, suffixes, and infixes are marked and root constituents are represented as variable letters. The processing that flagged all inflectional and some derivational affixes in the *word* together with the labeled morphological pattern will help in identifying the affixes, consonants, and semi-vowels that function as root constituents. If the word matches a morphological pattern but has some extra attachments at the front or end and these are marked as potential affixes, then they are definitely inflectional affixes. If the *word* contains letters that match the elements marked as affixes in the morphological pattern, then they must be derivational affixes. If the unmarked letters in the word match in position the variable letters in the morphological pattern, then they are definitely root constituents.

- To counteract the possibility of extracting the wrong root, the extraction program prompts the user to either confirm the extracted root or enter what they think is right. The correct root will then be listed into the dictionary database as the main entry and the original word will be entered as a subentry. This is a useful facility because it gives the algorithm the ability to learn new words and to update its own database.

When this algorithm is implemented, the roots of all the words in one novel [11] were extracted producing an output as exemplified below (next page):

	oldwrd	purified	pattern	root	prefix	infix	suffix
1	معربة	م ع ر ب ة	م؟؟؟ة	ع ر ب	م		ة
2	واصطلاحات	و ا ص ط ل ا ح ا	؟؟؟ا؟	ص ط ل	وا	ا	ات
3	محلية	م ح ل ي ة	م؟؟؟	ح ل ل	م		يّة
4	قليلة	ق ل ي ل ة	؟؟ي؟	ق ل ل		ي	ة
5	يمكن	ي م ك ن	؟؟؟	م ك ن	ي		
6	العودة	ال ع و د ة	؟؟؟	ع د د	ال	و	ة
7	إليها	ء ل ي ه ا		ء ل ي ه ا			
8	في	ف ي		ف ي			
9	المعجم	ال م ع ج م	م؟؟؟	ع ج م	الم		
10	المخصص	ال م خ ص ص	م؟؟؟	خ ص	الم		
11	لها	ل ه ا		ل ه ا			
12	في	ف ي		ف ي			
13	آخر	ء ا خ ر	؟؟؟	ء خ ر			
14	هذه	ه ذ ه		ه ذ ه			
15	الرواية	ال ر و ا ي ة	؟ا؟؟	ر و و	ال	ا	يّة
16	هذر	ه ذ ر	؟؟؟	ه ذ ر			
17	بدون	ب د و ن	؟؟و؟	ب د ن		و	
18	سفر	س ف ر	؟؟؟	س ف ر			
19	لا	ل ا		ل ا			
20	يهمني	ي ه م ن ي	؟؟؟	ه م م	ي		ني
21	كيف	ك ي ف		ك ي ف			
22	محنة	م ح ن ة	؟؟؟	م ح ن			ة
23	وبدعة	و ب د ع ة	؟؟؟	ب د ع	و		ة
24	محنتي	م ح ن ت ي	؟؟؟	م ح ن			تي
25	من	م ن		م ن			

Figure 1: Snap Shot of Output

For debugging purposes, the original word, purified word, morphological pattern, attempted root, and the affixes that were removed from the word have all been listed in the same row.

The output was analyzed linguistically then statistically, and it was found that the algorithm had a success rate of 84%. This is obviously a very

unsatisfactory result, yet within the broader context of this research it is extremely encouraging. To explain why this is so, let us study some of the cases of failure.

In the output snap shot above, rows have been numbered for easy reference to some of the cases of failure; i.e., the highlighted numbers (2), (6), (15), and (17). The algorithm has failed to account for morphophonemic transformations, as in (2), some cases of semi-voweled roots, as in (6) and (15), and for some rootless words, as in (17).

The algorithm was able to remove from waiSTilaaHaat in (2) the prefix conjunction [wa], the infix [aa], and the suffix [aat] but was still unable to extract S L H as the right root. It assumed that the word was quadriliteral (i.e., comprised of S T L H root constituents). The reason is that the morphophonemic transformation that the velarized letter [T] had undergone disguised the word. Prior to the morphophonemic change, the word would have had an alveolar [t] instead, so it would have matched with its true morphological pattern, iFtiMaaL. The algorithm would have been able to remove all affixes: the prefix [i] and the infixes [ti] and [aa] in one step had it not been for the morphophonemic transformation.

The algorithm did not handle properly such semi-voweled words as al'wdah in (6) and alriwaayah in (15). In the first, it successfully identified and removed the prefix [al] and the suffix [ah]. It failed, however, to identify the remaining letters (i.e., ' w d) as constituents of the root and instead considered the [w] as a pure vowel and removed it. When it ended up with two consonants and Arabic roots can never be biliteral[1], it had to apply a rule that doubles the final root constituent, giving ' d d as the root. In the second semi-voweled word, it correctly identified the prefix [al] and infix [aa] and removed them, but it confused the root constituent [y] with that in the relative adjective suffix [yah] and wrongly removed it. This resulted in a root with two constituents, which is not possible in Arabic, so it had to apply the rule that doubles the final constituent.

The failure exemplified in biduun, in (17), is the easiest to remedy. Here the non-derivational word duun was prefixed by the preposition bi, so the algorithm could not recognize it as rootless and attempted a root for it. It

is evident that the list of rootless words must not only include the basic form but also combinations of the word and all possible affixes.

It is clear from this sample of output that the present algorithm had difficulty with what language speakers have difficulty with, morphophonemic transformation and semi-voweled root patterns. Rules have to become complicated and the algorithm extremely complex before the remaining 16% of Arabic words are accounted for.

Since this attempt at root extraction was intended to be exploratory, another approach has been contemplated. This time, root extraction will come via a reversed process of word generation. A database of all Arabic roots will be established, and each root will be matched with the morphological pattern that it works with. Then precise grammatical rules will chart how a root uses a morphological pattern to generate the words that derive from it. This way, cases of morphophonemic transformation, metathesis, affixation, gemination, spelling, etc. will be accounted for at the root entry level. The root extraction algorithm will then work with definitiveness and the accuracy will improve greatly.

6. Conclusion

It has been demonstrated in this paper how an algorithm informed by empirical information about linguistic phenomena can be, on the one hand, simple and, on the other, reasonably successful. A liberated approach to grammatical description coupled with a commitment to sharing experiences amongst computational linguists can make the computerization of Arabic far more feasible, and the learning of Arabic easier and simpler.

Note

[1] Most traditional grammarians concur on that. The author is aware of only Georgie Zaidan, who was in the early 1900's, of the opinion that biliteral roots ought to be acceptable in Arabic since they are frequent in other Semitic languages

References

[1] M. Al-Bawwaab Y. MirAlam and M. Al-Tayyaan (1987). "A Computational System for Arabic Word Derivation", *Al-Lisaaniyat Al-Arabiya wal I'laamiya*. Tunis: University of Tunis, Center for Economic and Social Studies and Research.

[2] M. Al-Hannaash (1989). "*Al-Mu'jam Al-Iliktroni Al-Arabi*", First Kuwait Computer Conference Proceedings, Kuwait Computer Society, pp.173-183.

[3] A. Alneami, and J.J. McGregor (1996). "The Arabic Computational Lexicon", ICMC Proceedings, Cambridge: Cambridge University Press, pp. 3/31-22, 1996

[4] H.M. Al-Omari, T.M.T Sembok, T.M.T. and M. Yusoff, (1995). "Almas: An Arabic Language Morphological Analyser System", *Malaysian Journal of Computer Science*, 8(2), pp. 30-50

[5] . A. Al-Rajhi, (1984). *Al-Tatbeeq Al-Sarfi*, Beirut: Dar Al-Nahda Al-Arabiya.

[6] S. Al-Salih, (1983) . *Diraasaat fi Fiqh Al-Lugha*, 3rd Ed. Beirut: Daar Al-Ilm Lilmalaayeen.

[7] A. S. Al-Salman,(1983). *An Arabic Programming Environment (Compilers, User Interface)* , Ph.D thesis., Oklahoma State University.

[8] K. S. Alsamara, (1996). *An Arabic Lexicon to Support Information Retrieval, Parsing, and Text Generation*, Ph.D., thesis, Illinois Institute of Technology.

[9] Arab Organization for Education (1989). *Culture, and Sciences. Al-Mu'jam Al-Arabi Al-Asasi*. Tunis: Arab League.

[10] A. Bakkoush, (1992). *Al-Tasreef Al-Arabi min Khilaal Ilm Al-Aswaat Al-Hadeeth*, 3rd Ed. Tunis: Arab Press.

[11] M. R. Hamzaoui (1998). *Safar wa Hathar: Haaribun min Khitaab Assidq*, Paris: L'Harmattan.

[12] N. Hegazi and A. Sharkawi, (1985). "An Approach to a Computerized Lexical Analyzer for Natural Arabic Text", Workshop on Computer Processing and Transmission of the Arabic Language, Kuwait.

[13] A. H. Moussa, (1996). "Database for Major Arabic Dictionaries", ICEMCO Proceedings. Cambridge: Cambridge University Press, pp. 3/10/1-7.

[14] Ibn Jinni (1993). *Sirru Sina'atil I'raab*, vol.2. (Manuscript editor: Hasan Hindawi), Damascus: Darul Qalam.

[15] M. Mrayati, (1985). "Statistical Studies in Arabic Linguistics", Conference Proceedings of the Arab School of Science and Technology, Zabadani, Syria.

[16] G. Sarhan, I. Dawa, and M. M. Aboul-Ela, (1997). *A New Approach to the Design of Arabic Lexicon*, Modern Arabic Seminar, Cairo: Higher Council for Culture and Sciences.