# An Arabic morphological system

by T. A. El-Sadany
M. A. Hashish

***Nowadays, computers are used in every field in the Arab countries of the middle east. Software systems developed for the European languages are not convenient for the use of Arabic because of the nature of the language and its writing system. Problems arise when trying to use existing software systems, such as spell-checkers and business and office systems, with the Arabic language. These problems are attributable to the fact that the difference between Arabic and the European languages exists not only in character shapes and direction of writing, but also in language structure. In order to successfully use Arabic in software systems, we must, then, analyze the Arabic language word structure—that is, carry out a morphological analysis. Most of the written Arabic texts are nonvowelized, which may lead to ambiguity in meaning or mispronunciation. Moreover, vowelization cannot be avoided in many applications, such as speech synthesis by machines and educational books for children. A two-way Arabic morphological system (analysis/generation) capable of dealing with vowelized, semivowelized, and nonvowelized Arabic words was developed at the IBM Cairo Scientific Center. The system also has the ability to vowelize nonvowelized words. This system consists of three separate modules: computational lexicon, Arabic grammar model module, and analyzer/generator module. The grammar module contains, among others, morphophonemic and morphographemic rules formulated using the conventional generative grammar. Moreover, the developed system covers all of the Arabic language.***

In linguistics, *morphology* is the study of the structure of words.[1,2] In other words, morphology is simply a term for that branch of linguistics concerned with the forms words take in their different uses and constructions.[3–5]

The first true efforts in the research field of Arabic computational linguistics started only a few years ago. Some of the reasons for the relatively late start are the following.

Until recently there has been little interaction between computer scientists and Arabic linguists. At first, most of the systems dealing with the Arabic language were developed by engineers and computer scientists. Thus the systems developed performed small demonstrations that ran only a few examples collected by the system designers, without facing the real problems of the Arabic language itself.

The term *Arabization*, which has come into use, has created confusion among researchers. The term itself is used variously to cover the range from simple character representations in computers, to the translation of messages of already existing products, to font generation, up to complex Arabic computational systems. The computational linguistics portion has been largely ignored.

Another reason for the late start toward computational linguistics is that Arabic, being a member of the semitic languages family, is highly inflected and derived and therefore needs special techniques and

algorithms for solving its morphological problems. Unfortunately, most of the previous work in the Arabic morphological field has applied techniques and algorithms used with Western languages. Of course, this has kept Arabic morphological problems unsolved.

The main objectives of the work described in this paper can be summarized as follows:

- Stress the need for research in the field of Arabic computational linguistics in general and morphology in particular
- Specify and define a set of basic criteria to classify the previous work on Arabic computational morphology and act as a guide for the present and future work in this field
- Model the rules governing the morphological inflection and derivation of the Arabic language (i.e., the Arabic rules of grammar)
- Develop an Arabic computational lexicon that contains all the Arabic language roots with their associated information, to be used by the morphological system and for other linguistic tasks
- Develop a two-way (analysis/generation) morphological system capable of dealing with vowelized, semivowelized, and nonvowelized Arabic texts. The system is to be based on the Arabic grammar model and the computational lexicon. The system must be implemented on a small machine so that it can be used by the largest number of people.

One of the practical applications that the availability of such a morphological system is expected to contribute is that of first-level indexing of Arabic texts, for which software is needed. Another useful expectation is enhancement of the available software search systems for Arabic texts. Inasmuch as available systems use the word itself as the search key—an approach unsuitable for the Arabic language—the presence of a morphological analyzer can modify such search systems by using the Arabic roots as the search keys. The development of an Arabic spell-checking system is clearly a very useful direction.

## Basic criteria for an Arabic morphological system

The following are some proposed criteria to evaluate the previous work in Arabic morphology and to act as a guide for present and future work in this field.

**Classical Arabic language.** Any Arabic morphological system must be capable of dealing with the classical Arabic language, as opposed to the different dialects of the language. The system must be useful for all the Arabic-speaking countries and not merely for a special group.

**Vowelization.** Although most of the written Arabic texts are nonvowelized, the importance of vowelization cannot be ignored in many cases.[6] Vowelization is necessary in resolving ambiguity in the meaning of some words, the correct pronunciation of some words, the teaching of the Arabic language to beginners, and speech synthesis by machines. However, the vowelizing of Arabic text (i.e., the placing of vowels above and below Arabic consonants) is considered a problem even for ordinary native Arabic-speaking people. (This process usually requires a linguist.) Thus, the solution is to develop a system capable of dealing with vowelized, semivowelized, and nonvowelized texts. The system must also be able to vowelize the nonvowelized texts.

**There must be a clear border between the morphological information and the algorithms used for manipulating these data.**

**Coverage.** The system must be able to cover all the words of the Arabic language.

**Separability.** Because of the nature of any interdisciplinary field, there must be a clear border between the morphological information and the algorithms used for manipulating these data. Such separation facilitates the maintenance of the morphological system. To achieve this criterion, the Arabic morphological system must contain the following[7] three separate modules:

- The *Arabic grammar model module* contains all the classical Arabic morphological rules.
- The *Arabic computational lexicon* contains all Arabic roots with their associated features.

- The *analysis/generation module* is an inference engine that uses the Arabic grammar model module and the Arabic computational lexicon for the analysis and generation of Arabic words.

**Information distribution.** This criterion is optional, and the designer must decide whether the morphological information should be included in the Arabic computational lexicon or in the Arabic grammar model module.

**If the system is implemented on a small machine, it can be used by a large number of people.**

**Practicality.** If the system is implemented on a small machine, it can be used by a large number of people.[8] This requires efficient system implementation and the use of efficient textual storage techniques for the large amount of data. The system response is also a main factor in evaluating the practical usage of such systems.

### Arabic language terminology

The following are the definitions and terminology that are used in this paper.[9,10]

**Arabic alphabet.** The Arabic alphabet is an ordered set of the following 28 consonant letters: { ا، ب، ت، ث، ج، ح، خ، د، ذ، ر، ز، س، ش، ص، ض، ط، ظ، ع، غ، ف، ق، ك، ل، م، ن، هـ، و، ى }.

**Vowels.** Vowels are special shapes used with Arabic words. They are considered a main component in Arabic writing. There are three short vowels placed above and below Arabic characters,

| | | |
|---|---|---|
| Back close vowel | ُ | Damma (above) |
| Open middle vowel | َ | Fatha (above) |
| Front close vowel | ِ | Kasra (below) |

and the following three long vowels:

ُو، ا، ِى

There is a difference between the letters (و،ى) when used as consonant letters and when located after the short vowels. For the long vowels, a Damma is placed before the (و) and a Fatha before the (ا) and a Kasra before the (ى) to differentiate between them and the consonant letters. Therefore, the Arabic vowels are { ُ، َ، ِ، ُو، ا، ِى }.

**Suku:n.** Suku:n is a singleton set. It is phonetically nothing, yet it is a very important element in the vowelization of Arabic words, namely the { ْ } placed above the Arabic consonants. Suku:n = { ْ }.

**Diacritics.** Diacritics are marks used to distinguish letters or sounds that resemble one another as written. That is, they are special signs used for modifying the pronunciation of the letters. The diacritics used in Arabic language are the following.

*Gemination mark* (ّ) is a sign placed above the Arabic letters that results in repeating the letter at the phonemic level.

*Hamzatulwasl* (ا) has the same shape as that of the long vowel (ا). Arabic words cannot start with a letter having suku:n above it. Hamzatulwasl is used with words of that type, and is therefore called the "ladder of the tongue" (سلم اللسان).

*Madda* (~) is a sign used to prolong the duration of pronunciation of the letter. Note that the points used above and below some letters such as (ت) were considered as diacritics in the past. Nowadays, points are associated with these letters, and they (letters with their associated points) are considered members of the alphabet.

**Nunnation.** Nunnation is a process called "tanwin" in Arabic and is placed with the short vowel above the last letter of the word. Thus nunnation has a phonetic effect of placing (n ن) at the end of the word.

**Marks.** Marks is a set of signs and shapes used with the consonant characters to form the Arabic word. Marks = {vowels,suku:n,diacritics,nunnation}.

**Root (الجذر).** The root is an ordered sequence of valid three or four characters from the alphabet, i.e., the root can be either triliteral or tetraliteral. The root is not a valid Arabic word. For example, (ك ت ب) /ktb/.

**Measure or form (الوزن).** The measure is a general mould composed of an ordered sequence of characters. Some of these characters are constants (instantiated) and some are variables (uninstantiated). The uninstantiated characters are to be substituted (instantiated) with the characters of an Arabic root to generate a word called the "stem." Most of the characters of the measure have fixed marks, as previously discussed {vowels,suku:n,diacritics,nunnation}. There are different measures for the triliteral and tetraliteral roots. Note that the form (measure) is not a valid Arabic word, whereas the stem is a valid word.

**Declinable (معرب) and indeclinable (مبنى).** Declinable means that the word can take three cases (حالات اعراب) according to its position in the sentence. The three cases for the declinable verbs are the following:

Case 1 (مرفوع)
Case 2 (منصوب)
Case 3 (مجزوم)

For nouns, there are the following three cases:

Case 1 (مرفوع)
Case 2 (منصوب)
Case 3 (مجرور)

Indeclinable means that the word has a fixed case, regardless of its position in the sentence.

### Arabic word classification

From the morphological point of view, an Arabic word is classified as shown in Figure 1. Each element in the figure is now discussed in detail.

**Verbs.** Arabic verbs are generated from either triliteral or tetraliteral roots according to the following structure:

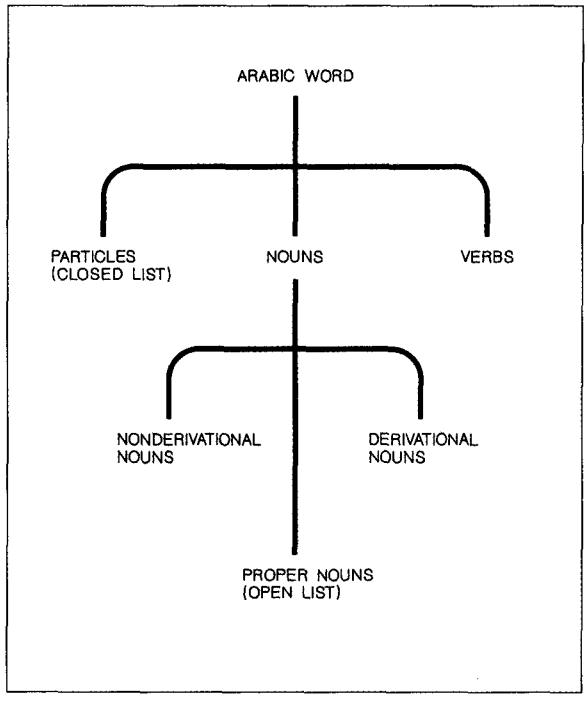
$$\text{Verbs} = \text{Prefix1} + \text{Prefix2} + \text{Stem} + \text{Suffix1} + \text{Suffix2} + \text{Suffix3} \quad (1)$$

The stem is formed by substituting the characters of the root into certain verb forms, called measures. Consider the following example:

Root (ك ت ب) /ktb/
Measure (فَعَّلَ) /ta_a:_a_a/
Stem (كَتَّبَ) /taka:taba/

There are 37 measures for the triliteral and tetraliteral roots.

**Figure 1 Arabic word classification**




Arabic verbs appear in three tenses, past (ماضى), present (مضارع), and imperative (امر).[11] The verb measures are in the past tense. For each verb measure in the past tense, there exists only one corresponding measure in the present tense except for the abstract triliteral measures. To achieve a one-to-one correspondence for the triliteral abstract measures, an additional feature must be used beside the measure. It was observed that the relation between the short vowel on the second character (ع) in the past and present tenses is the key or the missing feature. For each of the three abstract measures in the past tense there are at most three possible measures in the present tense (corresponding to placing Damma or Fatha or Kasra on the ع); thus, there are nine ($3 \times 3 = 9$) clusters representing the different abstract measures in the past and present tenses. However, only six out of the nine clusters[12] are actually used. This clustering is considered a very important feature in the morphology of Arabic verbs and is called the conjugation (الباب).

The derivation of the verbs in the different tenses is achieved using well-behaved morphological rules. Referring to Equation 1, the different prefixes and

suffixes are lists (vectors) of finite length; features are associated with each element in these lists. The properties of the prefixes and suffixes are given as follows:

*Prefix1.* The elements of this list serve as conjunctions. The associated features with each element are used to indicate the tense and the case, as previously discussed, of the verb attached to it. For example, the element (فل) indicates that the verb is in the past tense, and the case may be case 2 (منصوب) or case 3 (مجزوم).

*Prefix2.* The attributes associated with the elements of this list determine the tense of the verb and the features of the subject. The subject features are the person, gender, and number. For example, in (أ), the tense is present and the subject is for the first person, male or female, and singular.

*Suffix1.* The elements of this list are the subject pronouns attached to the verb. The attributes associated determine the tense, the case of the verb, and the subject features. The element (ون), for example, indicates that the verb is in the present tense, the case is case 1, and the subject pronoun is for the male, plural, third person.

*Suffix2.* The elements of this list are the first object pronouns. The elements associated with this list determine the features of the object pronoun (person, gender, and number).

*Suffix3.* The elements of this suffix are the same as those of the Suffix2 list. In this case, however, they represent the second object pronoun.

The last element of the prefixes and suffixes is nil. By substituting the nil for all the prefixes and suffixes in Equation 1, one can conclude that the stem is a valid Arabic word.

*Vowelization of Arabic verbs.* Vowelization is the process of placing the short vowels (ً، ٌ، ٍ) and the no vowel (ْ) above and below the Arabic letters. Arabic verbs may be either declinable or indeclinable. Verbs in the present tense are usually declinable, whereas verbs in the past and imperative tenses are always indeclinable. As mentioned before, the past tense stem is generated from the different measures and these measures are of fixed vowelization. The stem—either active or passive—in the present or imperative tenses can be generated with its vowelization from the past tense according to well-behaved rules. Therefore, knowing the measure, the tense, and the conjugation for the abstract forms, the indeclinable verb stem can be completely vowelized. The vowel on the last character of the declinable verb stem is determined by knowing the position of the word in the sentence.

Vowelization of the different prefixes and suffixes is fixed. However, the vowelization of the last character of the stem is influenced by the suffix attached to it (subject pronoun), according to regular morphological rules. Hence, using the morphological information associated with the verb, we have the following information:

- Measure
- Conjugation number (for the abstract triliteral forms)
- Tense
- Active/passive voice
- Subject pronoun attached to the stem
- Fixed vowelization of the different prefixes and suffixes

Arabic verbs are completely vowelized, except for the declinable ones with Suffix1 = nil, i.e., those for which the subject is not attached to the verb.

*Irregular verbs.* These are the kinds of verbs that need special treatment after being generated using the morphological rules previously given. This means that there is a difference between the irregular verb form that is generated using well-behaved rules and its orthographic realization that appears in written texts. The difference between these two realizations is attributed to three different origins, explained next.

Two similar characters may be written as one character with the diacritic gemination (ّ) above it. This is called (ادغام) in Arabic. Taking as an example the past tense for (ضلل) concatenated with the suffix (وا), the verb when generated using rules is ضَلَلُوا /ḍalalu:/; the verb as it appears in written texts is ضَلّوُا /ḍallu:/.

The second origin is related to the hamza (أ). The hamza is changed to other different realizations, due to the influence of the vowels before and after the hamza. The different realizations for the hamza are (إ، ئ، ؤ، ء، آ). These different realizations are called the allographs for the hamza. For example, in the present tense for (أمن) for the form (أفعل), the verb when generated using rules is يُأمِن /yu?min/; the verb as it appears in written texts is يُؤمِن /yu?min/.

The third origin is the most elaborate case. The Arabic characters ( و ، ا، ى ) are called the weak characters ( حروف العلة ). These characters can be deleted or replaced by one another. The reason for these modifications may be explained in light of phonological Arabic constraints. This case is considered to be a very important topic in the Arabic linguistics theory ( الاعلال والابدال ). The replacement of (و) by (ا) can be shown for the past tense of (قول) in the abstract form in the following example: the verb when generated using rules is قَوَلَ /qawala/; the verb as it appears in written texts is قَالَ /qa:la/.

The example for deletion of the imperative tense of ( قول ) in the abstract form is the following: the verb when generated using rules is اِقْوُلْ /?iqwul/; the verb as it appears in written texts is قُلْ /qul/.

These problems are solved by finding out all the rules governing these morphophonemic and morphographemic changes.[13]

**Derivational nouns.** Derivational nouns are those derived from Arabic verbs. Thus they were originally derived from Arabic roots. The derivational nouns have the following properties:

- Derivational nouns are semantically related to the root.
- The measures for every category of the derivational nouns are fixed.

For each derivational noun, there are certain measures at which the letters of the root are substituted to generate the stem. The total number of measures for the derivational nouns is of the order of 400. A derivational noun is formed by concatenating the stem with valid prefixes and suffixes according to the following structure:

$$\text{D.Noun} = \text{Prefix1+Prefix2+Prefix3+Stem} + \text{Suffix1+Suffix2+Suffix3} \quad (2)$$

The properties of the different prefixes and suffixes are now given.

*Prefix1* elements serve only like conjunctions.

*Prefix2* elements and their associated elements determine the case of the noun. For example, the presence of the element (ب) indicates that the noun is of case 3 (مجرور).

*Prefix3* elements are used to indicate whether the derivational noun is declared with « ال » (معرف بال).

*Suffix1* elements determine whether the noun is feminine with « ة » (مؤنث بالتاء).

*Suffix2* elements and their associated attributes indicate the case of the noun and whether it is dual ( مثنى ), masculine safe plural ( جمع مذكر سالم ), or feminine safe plural ( جمع مؤنث سالم ). For example, ( و ن ) indicates case 1 and the noun is for masculine safe plural.

*Suffix3* elements are the pronouns attached to the derivational nouns. The elements and their features are the same as those of the object pronouns.

*Irregular derivational nouns.* As in the case of verbs, there are some occasions in which the derivational nouns generated from the well-behaved morphological rules differ from their orthographical realizations that appear in written texts. The difference between the two realizations is attributed to the same reasons stated in the case of irregular verbs.

A different approach from that used with irregular verbs was carried out for solving this problem. Artificial measures were created for the surface realizations of the Arabic nouns. Hence, the total number of measures used with the derivational nouns has largely increased. The increase has been from about 400 standard measures to 1200 measures.

**Nonderivational nouns.** Nonderivational nouns are also nouns derived from Arabic roots. Nonderivational nouns have the property that they are semantically related to the root. The difference between the derivational and the nonderivational nouns is that the second property of the derivational nouns (fixed measures) is relaxed in case of the nonderivational ones.

Nonderivational nouns are formed by concatenating the stem with the different affixes (prefixes and suffixes) according to Equation 2, which is used for the derivational nouns. The different affixes concatenated with the nonderivational nouns are the same ones used with the derivational nouns.

**Particles.** The particles in Arabic language form a closed list, the number of which is relatively small and their features are known.

**Proper nouns.** The proper nouns are the nouns that are not derived from valid Arabic roots and are not particles. They form a large, open list. Based upon frequency analysis and statistical language models,

the most frequently-used Arabic proper nouns can be selected.

### Arabic computational lexicon

An Arabic computational lexicon has been developed on the IBM PS/2 Model 60. The Arabic lexicon is a special-purpose computational lexicon tailored to provide lexical information for the Arabic morphological system. The lexicon is also used to check the validity of the possible analyses produced by the morphological analyzer.

The head of the lexical entry in the Arabic lexicon includes the *root* features and attributes that are associated with each lexical entry. The following three features are of these types:

- Morphological
- Syntactic
- Semantic

Inasmuch as the purpose of this lexicon is to provide information for the morphological system, the syntactic and semantic features associated with each entry are only the ones needed for the morphological processing.

An example from the Arabic computational lexicon for the extracts of the entry " خدم " follows:

```
(ROOT( خدم ))
    (POS(VERB))
    (MORPH(INFLECTION(REG)))
                (CONJUNCTION فعّل يفعل ))
                (MEASURE استفعل )))
                (TENSE (PAST)))
                       (PRES)))
                       (IMPV)))
    (SYNTACTIC(TRANS1(OBJ1 NOUN)))
    (SEMANTIC(CONTEXT(SUBJ ANIMATE)))
                       (OBJ1 NONANIMATE)))
```

The first line is the root. In this case, the root is a triliteral one called " خدم ". Next, the *pos* gives the part of speech (verb in this case), which is followed by a set of morphological features and attributes. The first one of the morphological features states that the *inflection* of this verb is regular. Next, there are two morphological attributes that have specific values. The *conjunction* is ( فعّل يفعل ), and the measure at which the root is substituted to produce a stem is ( استفعل ). The final morphological feature is the *verb tense*, i.e., the verb can appear in the past, present, or imperative tenses.

A syntactic feature then follows, which states that the verb is transitive for one object and the value specified with this object is *noun.*

Finally, a semantic feature is stated that is of contextual nature rather than inherent and states that this verb takes an animate subject and inanimate object.

The Arabic computational lexicon consists of 5700 distinct lexical entries (roots). An efficient direct accessing of the lexicon information is achieved by combining a minimal hashing function and an indexing technique. The technique is based on the fact that most of the Arabic roots are triliteral.

**The morphological analyzer accepts unvowelized, partially vowelized, or completely vowelized input Arabic words.**

### Arabic morphological analyzer

The Arabic morphological analyzer is one part of the two-way Arabic morphological system. The analyzer is implemented on an IBM PS/2 Model 60 using logic programming language.[14,15] The analyzer gives a complete morphological analysis and vowelization for Arabic verbs, derivational nouns, nonderivational nouns, and particles. A simplified block diagram of the analyzer is shown in Figure 2. The modules shown in the figure are discussed below.

**Unvowelizing and base-shaping module.** The morphological analyzer accepts unvowelized, partially vowelized, or completely vowelized input Arabic words. The objective of this model is to remove the vowels from the input stream of characters and convert the characters to their base-shaped representation (i.e., position-independent codes).

**Analysis module.** The objective of the analysis module is to extract the root and the associated morphological features from the input unvowelized word. It

can be easily observed from Equations 1 and 2 that the Arabic word is formed from a stem concatenated with affixes (suffixes and prefixes). The presence of the affixes in a word indicates specific morphological features associated with this word.

The first step in the analysis is to remove the prefixes and the suffixes attached to the word. Augmented transition networks (ATN)—with slight modification to suit the Arabic language—are used to represent the ordering and linking of the affixes to each other and to the stem. The removal of the affixes and the extraction of the stem and the morphological features is achieved by traversing the augmented transition networks shown in Figure 3.

In the transition networks shown in Figure 3, certain features can be observed. Some of the labels above the arcs of the networks are terminal symbols (VPr1, VPr2, ...) that stand for verb prefixes and some are nonterminal symbols (Verb,Noun). The terminal symbols (VPr1,VPr2, ...) are used to indicate the elements of the prefix and suffix (affix) lists that are to be attached to the different stems. Feature registers are associated with different terminal symbols (affixes).

Conditions and actions are associated with the labels on the arcs of the networks. The condition on the arc must provide that the beginning of the word is to be matched with the affix list. For example, for the noun prefix Npr1, the beginning of the word is checked with the elements of the Prefix1 list of nouns until a match is found.

No contradictions may exist between the features associated with the word and those associated with the matched affix.

In case of condition fulfillment, one action to be performed is to remove the matched affix from the word to produce the new word. Another action is that features associated with the word are logically added with those associated with the affix. The results are to be assigned as the new features of the new word. Action arcs are shown in Figure 3A and 3B. No condition is to be satisfied in this case; only an action must be taken. The action is to jump to the end of the word to start the suffix-matching process.

The networks in Figure 3 are traversed by the input unvowelized word. The output from traversing these networks is a set of stems with their associated morphological features.

**Figure 2 Simplified block diagram for the Arabic *morphological analyzer***

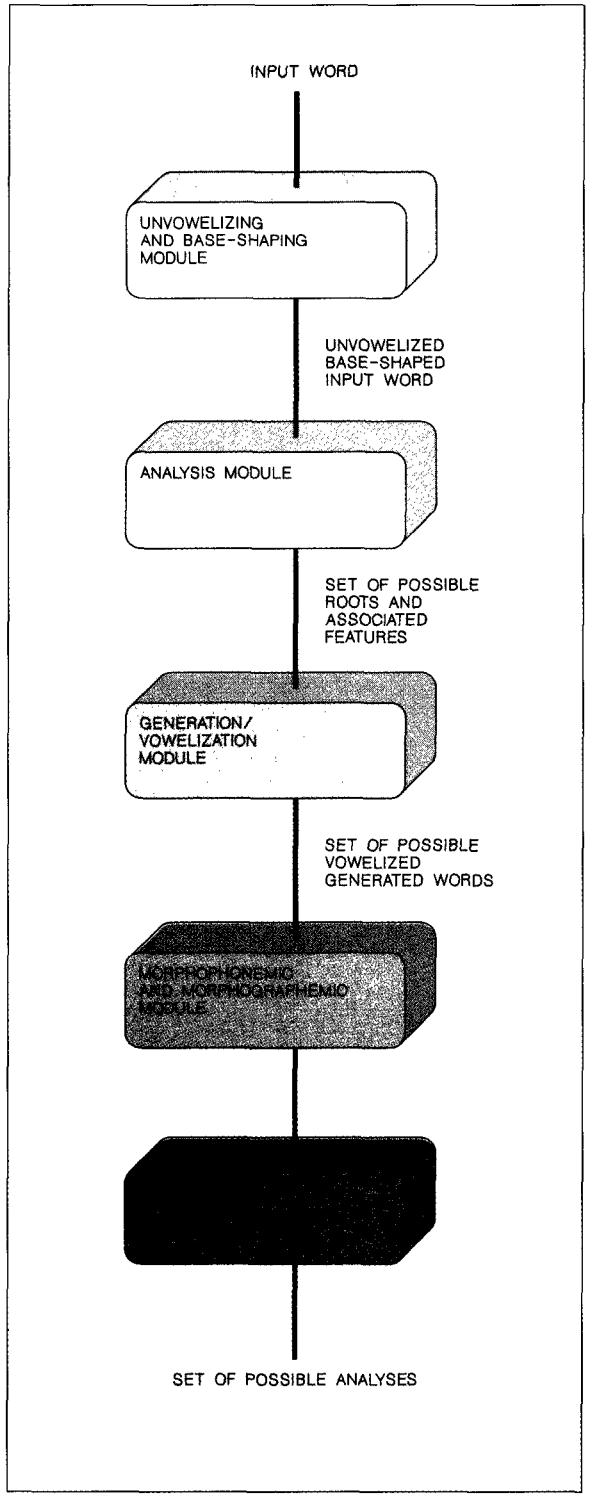
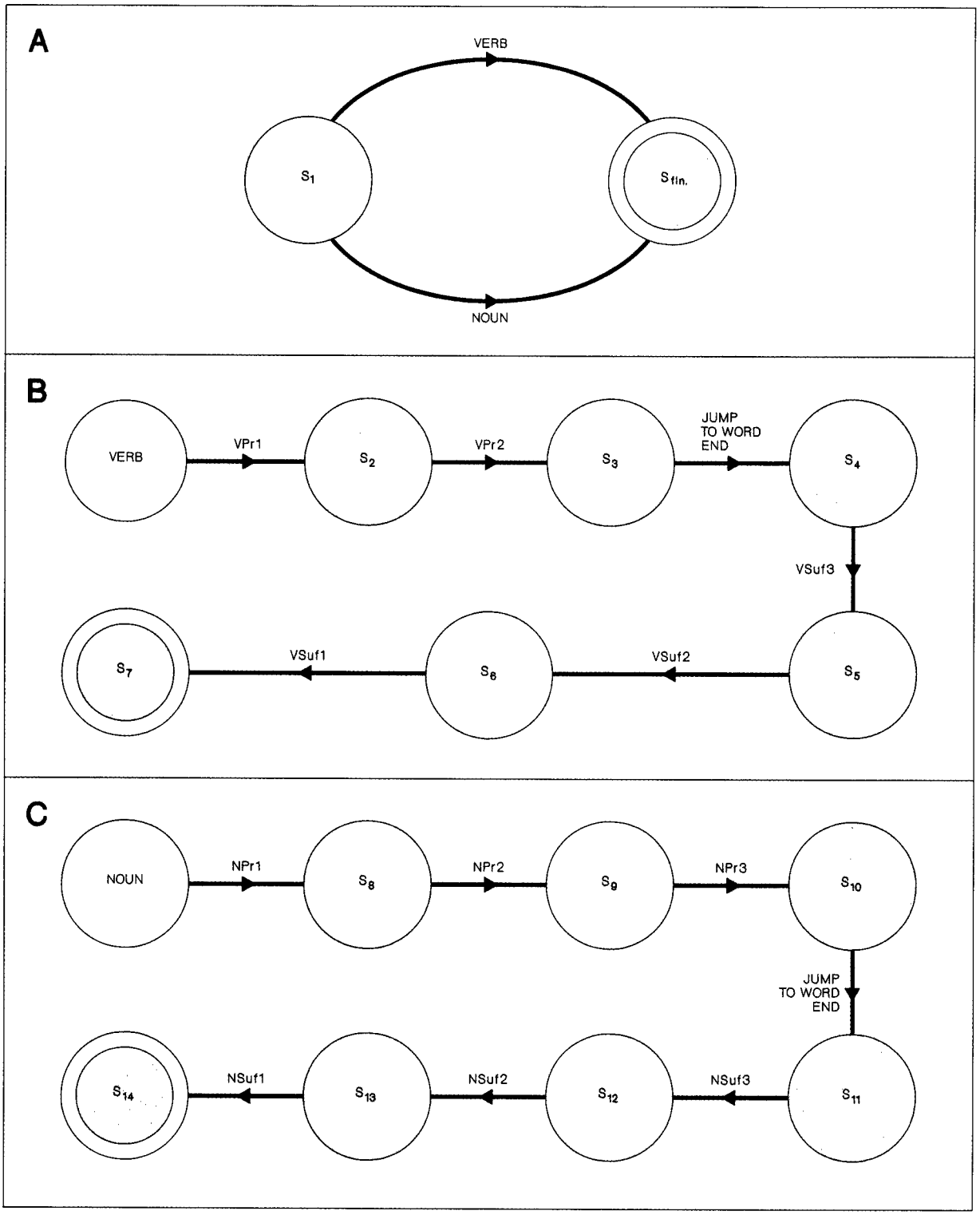


**Figure 3 ATN for Arabic language**



The second step in the analysis is to extract the roots from the output stems. A unification routine is implemented to extract the root from the stem by matching the stem with the different measures of verbs, derivational nouns, and nonderivational nouns.

**The last step in the analysis module is to check the possible set of roots and their associated features.**

The last step in the analysis module is to check the possible set of roots and their associated features with the Arabic computational lexicon. The checking results in a smaller set in which the matching with the lexicon succeeded.

The output from the analysis module would be a set of valid Arabic roots with their valid associated morphological features.

**Generation/vowelization module.** This module accepts the set of possible roots with their morphological features. The objective of this module is the generation of a vowelized word using the root and the associated morphological features. Using only the morphological information associated with the word, semiautomatic vowelization can be achieved.

A rule-based system is used for implementing the rules of the generation of vowelized words. The following is a sample production rule written in Prolog to generate a vowelized verb stem.

```
    /* Root */ /* Vowelized Stem */
vowelization([X,Y,Z],[X,'َ','ا',Y,'َ',Z,'ُ']):-
    features(tense,1),          /* past tense */
    features(measure,2),        /* measure فاعل */
    features(active,1),         /* active */
    features(suffix1," وا "),!.  /* subject pron. وا*/
```

The vowelized stem is then concatenated with the prefixes and suffixes. Since the vowelization of the affixes is fixed, the vowelized word can be generated. This process is repeated for the rest of the input set producing a new set of generated vowelized words.

**Morphophonemic and morphographemic module.** The objective of this module is to transfer the generated vowelized word into its surface realization, i.e., written form, in Arabic texts. The generated internal representation of Arabic words is usually different from their surface representation, due to morphophonemic and morphographemic changes. These changes are attributed to the irregular behavior of some Arabic words.

There are two approaches for solving such problems: (1) a rule-based approach; and (2) creating additional artificial measures for each word (not just the standard Arabic measures). The first approach was implemented with the Arabic verbs. The problem faced was that the rules governing morphological changes are not well defined and are not complete in the Arabic literature. Linguists used heuristic approaches based on human intuition and experience. Thus, a given internal representation is transferred to its corresponding surface realization (known to linguists) by selecting the appropriate set of rules in the appropriate order. The selection of the set of rules turns out (by experiments with several linguists) to be nonunique. Therefore, finding for the sake of machine analysis a definite unique sequence of rules to transfer a given internal representation to its surface realization without knowing the surface realizations beforehand is considered a great problem. Research was carried out to develop the required ordered set of rules.

The morphophonemic and morphographemic rules are also represented in a rule-based form. The rules are classified into six groups, wherein each group contains an ordered set of rules. In transferring the generated vowelized word to its surface realization, at most one rule is applied from each group.

The second approach is applied for the derivational and nonderivational nouns. Artificial measures are created for the different surface realizations of Arabic nouns.

By using this approach, the total number of measures (standard plus artificial) is largely increased. Thus the system speed compared with that for verbs is greatly reduced.

**Figure 4 Sample results from the morphological analyzer**

| | |
|---|---|
| الكلمة المراد تحليلها : قنا | التحليل الثانى : |
| التحليل الأول : | الكلمة المشكلة : قِنَا |
| الكلمة المشكلة : قَنَا | الجذر : و ق ى |
| الجذر : ق ن و | فعل امر |
| فعل ماضى مبنى للمعلوم | من باب فعَل يفعِل |
| من باب فعَل يفعُل | لفيف مفروق |
| ناقص وادى | ثلاثى مجرد |
| ثلاثى مجرد | متعدى لمفعولين |
| متعدى لمفعول واحد | ضمير الفاعل للمذكر المفرد المخاطب |
| ضمير الفاعل للمذكر المفرد الغائب | ضمير المفعول (نا) للمذكر والمؤنث المثنى والجمع المتكلم |

Another disadvantage concerning this approach is that the system cannot be used for an explanation of the morphophonemic and morphographemic changes occurring to the word. The morphophonemic and morphographemic rules present an important part of the education of Arabic morphology.

**Decision module.** The last step in the analyzer is to use a correlation procedure to match the resulting set of words with the given input word. The match results in a smaller set that is considered to give the possible analyses for this input word. Each one of the different analyses contains the root from which the word is derived, the morphological analysis associated with this solution, and the vowelization of the input word.

Sample results from the morphological analyzer are shown in Figure 4. Note that the different examples for the same word are the possible analyses for it.

### Arabic morphological generator

The Arabic morphological generator is the second part of the two-way Arabic morphological system. The objective of the generation module is to generate an Arabic vowelized word using the root and the morphological features. It can be easily observed that two main modules of the simplified block diagram of the morphological analyzer shown in Figure 2 (generation/vowelization module and morphophonemic and morphographemic module) are also the main modules used in the generation system.

A rule-based expert system for generating Arabic verbs and derivational nouns from roots using the associated morphological information was developed on an IBM PS/2 Model 60 using logic programming. The main units of this expert system are now discussed.

**Knowledge base.** The knowledge base of this system consists of two parts. One part is the Arabic com-

**Figure 5 Sample result from the morphological generator**

المطلوب :

ماهو جمع المؤنث السالم لاسم الفاعل من الجذر درس على وزن فعّل؟

الاجابة :

مُدَرِّسات

putational lexicon and the other is the generation module (generation/vowelization module and the morphophonemic and morphographemic module). As discussed before in the analysis module, the knowledge is represented in a rule-based form.

**Expert shell.** The expert shell constitutes the user interface and the inference engine that deals with the knowledge base. The user interface system is designed to communicate in a friendly way with the user, using Arabic natural language. The input sentence given by the user to the system is analyzed using keyword identification. The system has been programmed to recognize specific keywords. In this sense, the grammatical structure is not important because the program does not analyze relationships between words. The stems (words after removing affixes) of the input sentence are the ones to be matched with the specific keywords, i.e., a morphological keyword matching.

The inference engine uses the information gained from recognizing the keywords of the input sentence and starts to match this information with the lexicon in the knowledge base. The result of matching is always one out of the following three cases:

*Case 1.* A contradiction occurs between the user's requirements and the lexicon. Therefore, a message is sent to the user that the given requirements cannot be achieved.

*Case 2.* The morphological information given by the user to generate the word is found to be incomplete. Control is then returned to the user interface system to ask the user about the missing information. The cycle then continues until all the information required to generate a word is completed.

*Case 3.* The morphological information given by the user is complete and matches the lexicon. The set of the possible vowelized words having this morphological information is generated using the generation/vowelization and the morphophonemic and morphographemic modules in the knowledge base.

A sample result from the developed expert system is shown in Figure 5. This expert system can be used for teaching Arabic morphology to native Arabic-speaking students.

## Concluding remarks

A two-way Arabic morphological system (analysis/generation) capable of dealing with vowelized, semivowelized, and nonvowelized Arabic texts has been developed on an IBM PS/2 Model 60 at the IBM Cairo Scientific Center. The system has been written using Prolog.

The output of the system is either a possible different morphological analysis for a given word and the corresponding vowelization, or a set of vowelized words generated according to specified morphological information given by the user in an Arabic natural form. The system was designed to meet prespecified criteria. Morphophonemic and morphographemic rules are classified into six groups for the Arabic verbs. Work is still in progress to achieve similar results for the Arabic nouns.

## Cited references


1. S. R. Anderson, "Inflectional morphology," *Language Typology and Syntactic Description*, T. Shopen, Editor, Cambridge University Press, New York (1985), pp. 150–201.
2. S. R. Anderson, "Morphological theory," *Linguistics: The Cambridge Survey*, F. Newmeyer, Editor, Cambridge University Press, New York (1986).
3. S. R. Anderson, "Inflection," *Theoretical Morphology*, M. Hammond and M. Noonan, Editors, Academic Press, Inc., New York (1987).
4. S. R. Anderson, "Morphological change," *Linguistics: The Cambridge Survey*, F. Newmeyer, Editor, Cambridge University Press, New York (1986).
5. P. H. Matthews, *Morphology, An Introduction to the Theory of Word-Structure*, third edition, Cambridge University Press, New York (1982).
6. T. A. El-Sadany and M. A. Hashish, "Semi-automatic vowelization of Arabic verbs," *Proceedings of the 10th NCC Conference*, King Abdul-Aziz University, Jeddah, Saudi Arabia (1988).
7. T. A. El-Sadany and M. A. Hashish, *Arabic Morphological System*, IBM Conference on Natural Language Processing, Thornwood, New York (1988).
8. M. Gheith and T. El-Sadany, "Arabic morphological analyzer on a personal computer," *Proceedings of the First King Saud University Symposium on Computer Arabization* (April, 1987).
9. محى الدين عبد الحميد ، شرح ابن عقيل، دار الكتاب البنانى ، القاهرة ، مصر .
10. شذا العرف فى فن الصرف ، أحمد الحملاوى ، المكتبة الثقافية ، بيروت ، لبنان .
11. د. فؤاد نعمة ، ملخص قواعد اللغة العربية ، ١٩٨٦.
12. أنطوان الدحداح ، معجم قواعد اللغة العربية ، مكتبة لبنان ، بيروت ، ١٩٨٥ .
13. د. كامل الخويسكى، الزائد فى الصيغ فى اللغة العربية، ١٩٨٥ .
14. F. C. Pereira and S. M. Shieber, *Prolog and Natural Language Analysis*, CSLI, Stanford University, Stanford, CA (1987).
15. L. Sterling and E. Shapiro, *The Art of Prolog*, MIT Press, Cambridge, MA (1986).



**Tarek A. El-Sadany** *IBM Egypt, Cairo Scientific Center, 56 Gamiette El-Dowal Al Arabia, Mohandiseen, Giza, Egypt.* Mr. El-Sadany received his B.Sc. in electrical engineering (computers and control section) from Ain Shams University in June 1985 and his M.Sc. in electrical engineering (natural-language processing) from Al-Azhar University in February 1989. He joined the IBM Cairo Scientific Center as an assistant scientist in 1986. Mr. El-Sadany's research interests include natural-language processing and machine translation.

**Mohamed A. Hashish** *IBM Egypt, Cairo Scientific Center, 56 Gamiette El-Dowal Al Arabia, Mohandiseen, Giza, Egypt.* Professor Hashish received his B.Sc. in communications and electronics engineering from Cairo University in June 1964, his M.Sc. in electrical engineering (circuit theory) from Alexandria University in 1968, and his Dr.-Ing. (cybernatics) from the German Democratic Republic in 1972. He joined the IBM Cairo Scientific Center in April 1983 and is currently the manager of the Center. Dr. Hashish's research interest is the application of system theory tools to signal processing, pattern recognition, and natural-language processing.



Reprint Order No. G321-5377.