JKAU: Eng. Sci., **vol. 13 no. 1,** pp. 71-93 (1421 A.H. / 2001 A.D.)

An Arabic Morphological Analyzer/Synthesizer

M.G. Khayat[*], A. Al-Othman[**] and S. Al-Safran[**]

[*]*Department of Electrical & Computer Engineering,*
KAAU, Jeddah, Saudi Arabia

[**]*KFUPM, Dhahran, Saudi Arabia*

Abstract. Morphology is an essential element in processing natural language. As morphology in Arabic is highly derivational, morphological analysis/synthesis is systematic and can be easily automated.

The objective of this research work is to design and implement a morphological analyzer/synthesizer (MAS) for Arabic. In analysis mode, given a word, MAS determines the following properties of words: 1) type (noun, verb, article), 2) person, number and gender (for verbs and nouns), 3) tense of verb (past, present, imperative), 4) type of article (interrogative, prepositional, etc.), 5) root, and derivation (for verbs and nouns), and 6) type and identity of affixes (prefix, infix, suffix). In synthesis mode, the above properties are given and the corresponding word is constructed.

MAS is based on linguistic principles of Arabic morphology. It is designed as three modules for particles, nouns and verbs respectively. The modules consist of rules that encode the linguistic principles of word construction in Arabic. The mode (analysis or synthesis) of operation is automatically determined by the values associated with the word and its properties. For a word of size n of a particular type (noun, verb or article), the possible derivations (determined according to the linguistic principles) are implemented as ordered (according to their frequencies of occurrence) Prolog predicates. The size of the word and frequency of occurrence of the corresponding derivation are used to minimize the search time.

MAS is currently being used as a component of a natural Arabic understanding system. It can also be used in translation, computer-aided Arabic learning, character recognition and text and speech processing systems.

Introduction

Morphology is an essential element in processing natural language. As morphology in Arabic is highly derivational, morphological analysis/synthesis can be easily systematized. Morphological analysis/synthesis systems can be used in natural language understanding systems, computer-aided-learning of Arabic, sentence generation and spell checking.

The objective of this research work is to design and implement a morphological analyzer/synthesizer (MAS) for Arabic. In analysis mode, given a word, MAS determines the following properties of the word:

1) type (noun, verb, article),
2) person, number and gender (for verbs and nouns),
3) tense of verb (past, present, imperative),
4) type of article (interrogative, prepositional, ... etc.),
5) root, and derivation (for verbs and nouns), and
6) type and identity of affixes (prefix, infix, suffix).

In synthesis mode, the above properties are given and the corresponding word is produced.

Many approaches[1], [2], [3], [4], [5] have been devised to perform morphological analysis of Arabic words. The main disadvantage of these approaches is the use of dictionaries of roots and other types of words. They also do not address the synthesis problem. Furthermore, there is no indication of the implementation of these approaches. With respect to morphological synthesis, a system[6] used two methods of synthesis. The first method used the root and the derivation while the second uses a preliminary word and a set of attributes. The system requires storage for all roots, morphological patterns and standard forms.

In this paper we present a new approach that addresses both the analysis and synthesis problems. Section II of this paper describes the linguistic concepts and principles upon which the design and implementation of the proposed system are based. Section III describes the system design and implementation with some illustrative examples. We then conclude with a summary of the work done and future research areas in the topic.

In our presentation below, we assume the absence of diacritics on Arabic text since most of Arabic text (books, newspaper articles, reports, ... etc.) is non-diacrticized.

Arabic Morphology

In Arabic, like other languages, lexemes can be classified into three types: verbs, nouns, and particles. In general, verbs and nouns are derived from roots according to well-defined rules. Most (over 90%) of the roots are three-letter words while some are four-letter words. The two classes of roots are represented by corresponding patterns as shown in Table 1. The basic set of particles is closed and is divided into separable particles, those which are written as separate words, and non-separable, those which are always one-letter prefixes of words[7]. Table 2 shows the separable particles. Table 3 shows the singleton particles (there are only eight). Note that some of the singleton particles serve more than one purpose.

Table 1. Root patterns and examples.

Examples			Pattern	الوزن
translation	transliteration	Arabic		
go hit decrease	δahaba Daraba naqaSa	ذهب ضرب نقص	fa9ala	فعل
gargle neigh roll	ǥarǥara ħamħama daħraja	غرغر حمحم دحرج	fa9lala	فعلل

Table 2. The basic set of separable particles.

Separable particles ordered in ascending length الحروف المنفصلة	Particle type	نوع الحرف
أن إنَ إيَقد بل	affirmative	توكيد
إن لو ما من أي	conditional	شرط
هل كم	interrogative	استفهام
عن من في رب إذ مذ مع	preposition	جر
ثم أو أم قط	conjunctive	عطف
أي	explicative	تفسير
كلا	negative	نفي
يا وا ها	interjective	نداء
أن كي	infinitive	مصدر
نعم أجل بلى	affirmative	جواب
إذا كيف لما أما متى أين أنى لئن	conditional	شرط
أنى أين متى كيف	interrogative	استفهام

TABLE 2. Contd.

Separable particles ordered in ascending length الحروف المنفصلة	Particle type نوع الحرف
إلى على لدى عند عدا خلا منذ حتى	preposition جر
حتى لكن فقط كذا	conjunctive عطف
كلا	negative نفي
أيا هيا	interjective نداء
لكي إذن	infinitive مصدر
إلا بيد	exceptive استثناء
سوف	futuritive تسويف
أما هلا ألا إما	restrictive تخصيص
لعل كأن لكن	assurative توكيد
لولا لوما كلما أيان	conditional شرط
أما هلا ألا إما إنما أنما	restrictive تخصيص
حاشا	preposition جر
حيثما أينما ريثما كيفما	conditional شرط

TABLE 3. Singleton particles.

Particle الحرف	Particle type نوع الحرف	Examples أمثلة
أ	interrogative استفهام	Is he here? أهو هنا؟
will س	futuritive تسويف	I will go سأذهب
and و by	conjunctive عطف preposition جر	He and I went هو وأنا ذهبنا By God والله
for ل to verily let	preposition جر subjunctive نصب affirmative توكيد jussive أمر	I went for playing ذهبت للعب I went to play ذهبت لألعب Verily you are more feared لأنتم أشد رهبة Let thy heart be at ease ليطب قلبك
like ك	preposition جر	He is like a lion هو كالأسد
with ب	preposition جر	He played with the ball لعب بالكرة
then ف	conjunctive عطف	He went then ran. ذهب فجرى
by ت	preposition جر	By God تالله

Affixes to words in Arabic can be classified into two categories: external and internal. External affixes, typically prefixes and suffixes, are lexemes such as pronouns, conjunction particles, prepositions, or interrogatives. External affixes (excluding the definitive "al" equivalent to "the" in English) represent syntactic entities. Thus, a word can be a phrase or a complete sentence as shown in Table 4. Internal affixes (prefixes, and infixes) are used to produce derivations of nouns and verbs of a root.

TABLE 4. Examples of one-word phrases and sentences.

Translation	Transliteration	Arabic
I hit him	Darabtuhu	ضربته
This is their house	haδa manziluhum	هذا منزلهم
He sat then stood	jalasa faqaama	جلس فقام

Verbs are classified into three classes: past, present, and imperative[7]. Past and present tense verbs can be active or passive. Passive forms are derived from the corresponding active forms by only changing the diacritics. Active past tense single masculine third person forms represent the basic verbal derivations. Table 5 shows all the basic verbal derivations of the two patterns of roots respectively. Other past tense verbal derivations (*e.g.,* dual, plural, feminine, first person, second person) are formed by adding pronouns as (external) suffixes. To produce present tense single derivations, a one-letter prefix (depending on the person) is added to all derivations. In addition, for the present tense dual and plural derivations, pronouns are added as (external) suffixes. Imperative form derivations only apply to the second person (spoken to) and require the addition of pronouns as suffixes and for some derivations the addition of the letter "alef" as a prefix. Table 6 shows the possible derivation patterns of the basic derivation "fa9al".

A noun in Arabic can be a substantive, adjective, numeral adjective, pronoun or proper noun[7]. Pronouns can be demonstrative, relative, personal, interrogative, or indefinite. As the pronouns and the cardinal numbers and (a set of) proper nouns are fixed in number and do not follow any derivation patterns, they can simply be recognized by pattern matching. Substantive and adjective nouns are derivatives. The derivative nouns include the infinitive noun, active voice noun, passive voice noun, noun of assimilation and intensiveness, noun of preeminence, relative adjective, diminutive noun, dual noun, sound plural noun, and broken plural noun[7].

The infinitive nouns as defined in[7] are "abstract substantives, which express the action, passion, or state indicated by the corresponding verb, without any reference to object, subject or time". These include derivations from verb (root), the nouns formed from the derived forms of the verb, nouns that express the doing of an action once, nouns of kind, nouns of place and time, and nouns of instrument. There are 44 infinitive noun derivations from the root verb[7]. Table 7 shows a sample of these derivations. Table 8 shows the infinitive nouns derived from the different forms (Table 5) of the verb.

Table 5. The basic verbal derivation patterns.

Derivation patterns in	الأوزان	Examples		أمثلة
ascending order	حسب عدد الحروف	translation	transliteration	Arabic
a9ala	فعل	to write	kataba	كتب
af9ala	أفعل	to pour out	araaqa	أراق
faa9ala	فاعل	to fight	qaatala	قاتل
fa99ala	فعـــل	to disperse	farraqa	فرق
fa9lala	فعلل	to roll	daħraja	دحرج
inf9ala	انفعل	to be cut off	inqaTa9a	انقطع
ifta9ala	افتعل	to oppose	i9taraDa	اعترض
tafaa9ala	تفاعل	to pretend to cry	tabaaka	تباكى
tafa99ala	تفعّـــل	to speak	takallama	تكلـم
tafa9lala	تفعلل	to roll along	tadħraja	تدحرج
ifa9alla	افعل	to turn black	iswadda	اسود
istaf9ala	استفعل	to ask pardon	istagfara	استغفر
if9aw9ala	افعوعل	to become moist	ixDawDala	اخضوضل
if9anlala	افعنلل	to flow	iθ9anjara	اثعنجر

Active voice nouns are verbal adjectives representing the actor of the verb. There is one derivative for every derivative form of the verb. The passive voice nouns are analogously defined. Table 9 shows the derivations of both types.

Nouns of assimilation and intensiveness "express a quality inherent and permanent in a person or thing with a certain degree of intensity"[7]. Table 10 shows the basic derivation patterns of nouns of assimilation and intensiveness.

Nouns of preeminence have the signification of the comparative and superlative[7] and have only one derivation pattern "af9al". Relative adjectives "denote that a person or thing belongs to or is connected therewith"[7], and are formed by suffixing a word with the letter ya. The diminutive noun has three basic derivational forms. Dual nouns and sound plural nouns are formed by adding a two-letter suffix to the singular form. Table 11 shows the derivation patterns of the noun of preeminence, relative adjective, diminutive noun, and sample dual and sound-plural nouns of a singular derivation "mufaa9il".

Table 6 . The number-gender-person patterns of a verb.

Person	Gender	Number	Past tense derivation	Patterns	Present tense derivation	Patterns	Imperative derivation	Patterns
Sing.	masc.	1st	fa9altu	فعلت	af9alu	أفعل		
Sing.	fem.	1st	fa9altu	فعلت	af9al	أفعل		
Sing.	masc.	2nd	a9alta	فعلت	taf9alu	تفعل	if9a	افعل
Sing.	fem.	2nd	fa9alti	فعلت	taf9aliin	تفعلين	if9alii	افعلي
Sing.	masc.	3rd	fa9ala	فعل	yaf9alu	يفعل		
Sing	fem.	3rd	fa9alat	فعلت	taf9alu	تفعل		
Dual	masc.	1st	fa9alnaa	فعلنا	naf9alu	نفعل		
Dual	fem.	1st	fa9alnaa	فعلنا	naf9alu	نفعل		
Dual	masc.	2nd	fa9altumaa	فعلتما	taf9alaani	تفعلان	if9alaa	افعلا
Dual	fem.	2nd	fa9altumaa	فعلتما	taf9alaani	تفعلان	if9alaa	افعلا
Dual	masc.	3rd	fa9alaa	فعلا	yaf9alaani	يفعلان		
Dual	fem.	3rd	fa9alataa	فعلتا	taf9alaani	تفعلان		
Plur.	masc.	1st	fa9alna	فعلنا	naf9alu	نفعل		
Plur.	fem.	1st	fa9alna	فعلنا	naf9alu	نفعل		
Plur.	masc.	2nd	fa9altum	فعلتم	taf9aluun	تفعلون	if9aluu	افعلوا
Plur	fem.	2nd	fa9altunna	فعلتن	taf9alna	تفعلن	if9alna	افعلن
Plur.	masc.	3rd	fa9aluu	فعلوا	yaf9aluun	يفعلون		
Plur.	fem.	3rd	fa9alnna	فعلن	yaf9alna	يفعلن		

TABLE 7. A sample of the infinitive nouns.

Derivation pattern	Examples		
	translation	transliteration	Arabic
فعل	escape	harab	هرب
فعلة	mercy	raħmah	رحمة
فعلى	memory	δekraa	ذكرى
فعلان	turbulence	hayajaan	هيجان
فعال	marriage	nikaah	نكاح
فعالة	cleanliness	naĐaafah	نظافة
فعالية	hatred	karaahiyah	كراهية
فعول	acceptance	qabuul	قبول
فعولة	difficulty	Su9ubah	صعوبة
فعولية	privacy	xuSuusiyah	خصوصية
فعيل	departure	raħiil	رحيل
مفعل	entrance	madxal	مدخل

TABLE 8. The infinitive nouns of the verbal derivation patterns.

Verb pattern	Infinitive noun pattern	Examples		
		translation	transliteration	Arabic
فعل	فعل	understanding	fahm	فهم
أفعل	إفعال	honoring	ikraam	إكرام
فاعل	مفاعلة	practice	mumaarasah	ممارسة
فعّل	تفعيل	separation	tafriiq	تفريق
فعلل	فعلال	earthquake	zilzaal	زلزال
انفعل	انفعال	ceasure	inqiTaa9	انقطاع
افتعل	افتعال	objection	i9tiraaD	اعتراض
تفاعل	تفاعل	variation	tafaawut	تفاوت
تفعل	تفعل	bearing	tahhamul	تحمل
تفعلل	تفعلل	rolling	tadahruj	تدحرج
افعل	افعلال	blackening	iswidaad	اسوداد
استفعل	استفعال	inhaling	istinsaaq	استنشاق
افعنلل	افعنلال	gathering	ihrinjaam	احرنجام

TABLE 9. The active and passive voice nouns of the verbal derivations.

Verb pattern	Active voice noun pattern		Examples (active voice)			Passive voice noun pattern		Examples (passive voice)		
			translation	transliteration	Arabic			translation	transliteration	Arabic
فعل	faa9il	فاعل	killer	qaatil	قاتل	mafo9uul	مفعول	killed	maqtuul	مقتول
أفعل	muf9il	مفعل	producer	muntij	منتج	muf9a	مفعل	product	muntaj	منتج
فاعل	mufaa9il	مفاعل	fighter	muqaati	مقاتل	mufaa9a	مفاعل	fought	muqaata	مقاتل
فعــل	mufa99il	مفعل	teacher	mu9allim	معلم	mufa99al	مفعل	taught	mu9allam	معلم
فعلل	mufa9lil	مفعلل	earthshaker	muzalzil	مزلزل	mufa9lal	مفعلل	earthshaken	muzalzal	مزلزل
انفعل	munfa9il	منفعل	loser	munhazim	منهزم	munfa9a	منفعل	led	munqaad	منقاد
افتعل	mufta9il	مفتعل	victor	muntasir	منتصر	mufta9al	مفتعل	prey	muftaras	مفترس
تفاعل	mutafaa9il	متفاعل	responsive	mutajaawib	متجاوب	mutafaa9al	متفاعل	neglected	mutagaafal	متغافل
تفعل	mutafa99il	متفعل	speaker	mutakallim	متكلم	mutafa99al	متفعل	spoken	mutakallam	متكلم
تفعلل	mutafa9lil	متفعلل	rolling	mutadaħrij	متدحرج	mutafa9lal	متفعلل	rolled	mutadaħraj	متدحرج
افعل	muf9il	مفعل	blackener	muswidd	مسود	muf9all	مفعل	blackened	muswadd	مسود
استفعل	mustaf9il	مستفعل	enquirer	mustafsir	مستفسر	mustaf9al	مستفعل	enquired	mustafsar	مستفسر
افعنلل	muf9anlil	مفعنلل	flowing	muth9anjir	مثعنجر	muf9anlal	مفعنلل	flowed	muth9anjar	مثعنجر

TABLE 10. The derivations of the nouns of assimilation and intensiveness.

Derivation pattern		Examples أمثلة		
		translation	transliteration	Arabic
fa9aal	فعال	baker	xabbaaz	خباز
mifa9aal	مفعال	talkative	miqwaal	مقوال
fa9uul	فَـعول	shy	xajuul	خجول
fa9iil	فعيل	sick	mariiD	مريض
fa9il	فَعِل	rough	xashin	خشن
faa9uul	فاعول	rocket	Saaruux	صاروخ
fi99iil	فعّيل	alcoholic	sikkiir	سكير
mif9iil	مفعيل	poor	miskiin	مسكين
fu9alah	فُعَلة	breaking in pieces	ħutamah	حطمة
fu99aal	فُعّال	very large	kubbaar	كبار
af9al	أفعل	red	ahmar	أحمر
fa9laan	فعلان	thirsty	aTsaan	عطشان
fa9aal	فَعَال	cowardly	jabaan	جبان
fu9aal	فُعَال	brave	sujaa9	شجاع
fay9al	فيعل	dead	mayyit	ميت
fa9l	فَعل	easy	sahl	سهل
fi9l	فعل	child	tifl	طفل
fu9l	فُعل	steel	sulb	صلب

TABLE 11. The derivations of the nouns of preeminence, relative adjective, diminutive, dual, and sound plural nouns.

Type of noun	Derivation patterns		Examples أمثلة		
			translation	transliteration	Arabic
preeminence	af9al	أفعل	better	aħsan	أحسن
Relative adjective	fa9aliy	فعلي	mountainous	jabaliy	جبلي
demunitive	fu9ayl fu9ay9il fu9ay9iil	فعيل فعيعل فعيعيل	hill booklet sparrow	jubay kutayyib 9usayfiir	جبيل كتيب عصيفير
dual	mufaa9ilaan	مفاعلان	two fighters	muqaatilaan	مقاتلان
sound plural	mufaa9iluun	مفاعلون	fighters	muqaatiluun	مقاتلون

The broken plural noun has 39 derivations from the three-letter root and three derivations from the four-letter root[7]. Table 12 shows a sample of these derivations.

TABLE 12. Sample derivations of the broken plural noun

Broken plural noun derivation patterns	Examples	أمثلة	
	translation	transliteration	Arabic
fu9al فعل	knees	rukab	ركب
fu9ul فعل	books	kutub	كتب
fi9al فعل	tents	xiyam	خيم
fi9aal فعال	men	rijaal	رجال
fu9uul فعول	souls	nufuus	نفوس
afa9aal أفعال	feet	aqdaam	أقدام
fawaa9il فواعل	stamps	Tawaabi9	طوابع
fa9aail فعأْئل	pronouns	Damaair	ضمائر
fi9laan فعلان	neighbors	jiiraan	جيران
fu9laan فعلان	horsemen	fursaan	فرسان
fu9alaa فعلاء	poets	su9araa	شعراء
af9ilaa أفعلاء	friends	aSdiqaa	أصدقاء
fa9iil فعيل	slaves	9abiid	عبيد
fa9aalil فعالل	tables	jadaawil	جداول

The verbal and nominal derivation patterns discussed above are basic and can be further affixed by (external) prefixes and suffixes. Table 13 shows the basic set of prefixes, which are the singleton particles (shown earlier in Table 3 with examples) in addition to the definitive "al" equivalent to "the" in English. Table 14 shows the basic set of suffixes, the type of word (particle, noun, or verb) they affix to and examples.

When some derivations are applied to roots that contain vowels (typically one or two vowels), new patterns result as a consequence of deleting or changing the vowels. In addition, when combinations of certain letters occur in a derivation of a root, some letters are substituted according to phonological rules to ease the pronunciation of the word. These actions are manifested by well-defined rules[7], [8]. Table 15 illustrates some examples of both phenomena. In this paper, we refer to the non-vowel roots as normal.

TABLE 13. The basic prefixes.

Prefix	Types of words prefixed
أ	noun, verb, particle
ب	noun
ت	noun
س	verb
ف	noun, verb, particle
ك	noun
ل	noun, verb, particle
و	noun, verb, particle
ال	noun

TABLE 14. The basic suffixes.

Sufffix	Types of words prefixed	Examples
ا	noun, verb	صاحبا ، صدفا
ت	verb	صدقت
ة	noun	ذاهبة
ك	noun, verb, particle	كتابك ، ضربك ، عنك
ن	verb	صدقن
ه	noun, verb, particle	كتابه ، أخرجه ، فيه
و	noun, verb	مهندسو ، سألتمونيها
ي	noun, verb , particle	كتابي ، اكتبي ، عني
ات	noun	سيدات
ان	noun, verb	مدرسان ، يكتبان
تم	verb	ذهبتم
كم	noun, verb , particle	منكم,كتابكم ، ضربكم
كن	noun, verb , particle	كتابكن ، دخلن ، عنكن
نا	noun, verb, particle	كتابنا ، ضربنا ، فينا
ني	verb	أعطاني
ها	noun, verb , particle	كتابها ، دخلها ، منها
هم	noun, verb , particle	بيتهم ، نصرهم ، فيهم
هن	noun, verb , particle	بيوتهن ، بايعهن ، عنهن
وا	verb	صدقوا

Table 14. Contd.

Sufffix	Types of words prefixed	Examples
ون	noun, verb	مكذبون ، يكتبون
ين	noun	مدرسين
تما	verb	ذهبتما
كما	noun, verb	كتابكما ، أخرجكما
هما	noun, verb	منزلهما ، أخرجهما

Table 15. Vowel verbs and substitutions.

Derivation pattern		Root		Actual derivation		
				translation	transliteration	Arabic
if9al	افعل	qawala	قول	say	qul	قل
fa9ala	فعل	qawala	قول	he said	qaala	قال
efta9ala	افتعل	Daraba	ضرب	he agitated	iDTaraba	اضطرب
efta9ala	افتعل	axaδa	أخذ	took for himself	ettaxaδa	اتخذ

The Morphological Analyzer/Synthesizer (MAS)

As words in Arabic are classified into nouns, verbs and particles, MAS consists of three word-modules for nouns, verbs and particles respectively, and a control module. If the type of the word is already determined (*e.g.* by a syntax analyzer/synthesizer), the corresponding module can be directly called. If the type is unknown (applicable in analysis mode), the control module is invoked. The control module applies heuristic criteria to restrict the search space and time as follows. First, the word is checked against the basic set of particles shown in Table 2, the basic set of pronouns and a set of proper nouns defined by the user. Second, the particles module is called since their number is limited. Third, the nouns and verbs modules are called in that order according to their frequencies of occurrence, 57% and 11% respectively as given in[9]. If at this stage, the word can not be recognized the system returns failure.

It is noteworthy that some of the affixes cannot be determined (in synthesis mode) by morphological rules as the affixes depend on their syntactic function in the context in which they occur. In such cases, it is assumed that an end-case or syntax synthesizer[10],[11] provides the affixes. In fact, this strategy is adopted in the natural Arabic understanding system (NAUS) which uses MAS as a morphological component.

Each word-module is divided into a set of rules based on the number of letters in the word and the set of possible affixes. For each module, the patterns have been grouped in terms of word size. This approach minimizes the number of rules as words can be analyzed/synthesized in terms of shorter words and affixes. However, the compatibility of possible concurrent affixes must be checked.

The particles module processes separable particles. The inseparable particles are recognized/synthesized as prefixes in all three modules. The length of particle words spans from two to seven letters. Table 16 shows the possible constructions for each length with examples.

The length of verbal words spans from one to twelve. Table 17 shows a representative sample of possible constructions of verbal words with examples. The Table shows possible constructions of verbal words of size one, two, three, four, ten, eleven, and twelve.

For verbal words of size n, $4 \leq n \leq 6$, the word can be an n-letter verbal derivation, an (n-1)-letter verb prefixed with a one-letter preposition or interrogative, an (n-1)-letter verb suffixed with a one-letter pronoun, an (n-2)-letter verb with a two-letter prefix, a (n-2)- letter verb with a two-letter suffix, or an (n-3)-letter verb with a three-letter suffix. For verbal words of size n, $7 \leq n \leq 12$, the word can be an (n-1)-letter verb prefixed with a one-letter preposition or interrogative, an (n-1)-letter verb suffixed with a one-letter pronoun, an (n-2)-letter verb with a two-letter prefix, a (n-2)-letter verb with a two-letter suffix, or an (n-3)-letter verb with a three-letter suffix.

The length of nominal words, excluding proper nouns, spans from two to fourteen. Table 18 shows a representative sample of constructions of nouns with examples. The Table shows possible constructions of words of size two, three, four, five, ten and fourteen.

A nominal word of length $5 \leq n \leq 9$ can be a noun derivative of length n, an (n-1)- letter word with a one-letter prefix, an (n-1)-letter word with a one-letter suffix, an (n-2)-letter word with a two-letter suffix, or an (n-3)-letter suffixed with a three-letter pronoun. A nominal word of length $10 \leq n \leq 14$ can be an (n-1)-letter word with a one-letter prefix, an (n-1)-letter word with a one-letter suffix, an (n-2)-letter word with a two-letter suffix, or an (n-3)-letter suffixed with a three-letter pronoun.

Having determined a root of a word, the analyzer checks its validity according to the phonological properties of the letters of the Arabic alphabet. The letters are grouped according to their location of occurrence in the human speech system. Those letters of the same group, for example, the letters (ħ, 9and h), can never be adjacent in a word.

TABLE 16. Particle word constructions.

Word size	Constructions	Examples		أمثلة
		translation	transliteration	Arabic
2	two-letter particle	from	min	من
	one-letter particle with a one-letter suffix	for him	lahu	له
3	three-letter particle	when	mata	متي
	two-letter particle-word with a one-letter prefix	and from	wamin	ومن
	two-letter particle with a one-letter suffix	from him	minhu	منه
	one-letter particle with a two-letter suffix	for her	lahaa	لها
4	four-letter particle	whenever	ayyaan	أيان
	three-letter particle-word with a one-letter prefix	and for her	walahaa	ولها
	three-letter particle-word with a one-letter suffix	and from him	waminhu	ومنه
	two-letter particle with a two-letter suffix	from her	minhaa	منها
	one-letter particle with a three-letter suffix	for both of them	lahumaa	لهما
5	five-letter particle	wherever	aynamaa	أينما
	four-letter particle-word with a one-letter prefix	and from her	wa-min-haa	ومنها
	four-letter particle-word with a one-letter suffix	and whenever	wa-ayyaan	وأيان
	three-letter particle-word with a two-letter suffix	and from her	wa-min-haa	ومنها
	two-letter particle with a three-letter suffix	from both of them	min-humaa	منهما
6	five-letter particle-word with a one-letter prefix	and from both of them	wa-min-humaa	ومنهما
7	six-letter particle-word with a one-letter prefix	Is ... from both of them ...?	a-wa-min-humaa	أومنهم

TABLE 17. Sample verbal word constructions.

Word size	Constructions	Examples		أمثلة
		translation	transliteration	Arabic
1	singular masculine imperative of two-vowelled root	protect	qi	ق
2	singular masculine imperative of one-vowelled root	take	xuδ	خذ
	one-letter verb with a one-letter suffix	protect him	qihi	قه
3	Past tense three-letter normal verb	he drank	shariba	شرب
	Present tense of one-vowelled root	we promise	na9id	نعد
	Past tense of one-vowelled root	I came back	9ud-tu	عدت
	two-letter verbal word with a one-letter suffix	take him	xuδ-hu	خذه
	one-letter verb with a two-letter suffix	protect them	qi-him	قهم
4	derivable verb	he fought	qaatil	قاتل
	three-letter verbal word with a one-letter prefix	and he drank	wa-shariba	وشرب
	three-letter verbal word with a one-letter suffix	he advised him	nasah-hu	نصحه
	one-letter verb with a three-letter suffix	protect both of them	qi-hima	قهما
10	nine-letter verbal word with a one-letter prefix	do we give it to you	a-nu9tikumuuhaa	أنعطيكموها
	eight-letter verbal word with a two-letter suffix	will you use her	a-satastakhdima-haa	أستستخدمها
	seven-letter verbal word with a three-letter suffix	and he used both of them	wa-staxdama-huma	واستخدمهما
11	nine-letter verbal word with a two-letter suffix	you gave it to me	a9Taytumuunii-ha	أعطيتمونيها
12	eleven-letter verbal word with a one-letter prefix	did you gave it to me	a-a9Taytuumuniiha	أأعطيتمونيها

TABLE 18. Sample nominal word constructions.

Word size	Constructions	Examples		أمثلة
		translation	transliteration	Arabic
2	non-derivable noun derivable vowelled noun	they blood	hum dam	هم دم
3	non-derivable noun	we	naħnu	نحن
	derivable noun	escape	harab	هرب
	two-letter nominal word with a one-letter prefix	and they	wa-hum	وهم
	two-letter nominal word with a one-letter suffix	his hand	yadu-hu	يده
4	non-derivable noun	you	antum	أنتم
	derivable noun	killer	qaatil	قاتل
	three-letter nominal word with a one-letter prefix	and we	wa-naħnu	ونحن
	two-letter nominal word with a two-letter suffix	her hand	yadu-haa	يدها
5	non-derivable noun	you	antumaa	أنتما
	derivable noun	fighter	muqaati	مقاتل
	four-letter nominal word with a one-letter suffix	his killer	qaatilu-hu	قاتله
	three-letter nominal word with a two-letter suffix	her escape	harabu-haa	هربها
	two-letter nominal word with a three-letter suffix	their blood	damu-humaa	دمهما
10	nine-letter nominal word with a one-letter prefix	and by the teachers	wa-bilmudarrisiin	وبالمدرسين
	nine-letter nominal word with a one-letter suffix	and with his infirmation	wabima9auumati-hi	وبمعلوماته
	eight-letter nominal word with a two-letter suffix	and with her keys	wabimafaatiiħi-haa	وبمفاتيحها
	seven-letter nominal word with a three-letter suffix	their information	ma9aluumaatu-humaa	معلوماتهما
14	Thirteen-letter nominal word with a one-letter prefix	and with both colonies?	a-wabilmusta9maratayin	أوبالمستعمرتين

In implementing the rules of each of the three modules, the words are grouped according to their lengths and properties, and the properties of their prefixes. Whenever any of the rules implies the concatenation of affixes, the affixes are checked for compatibility. When a property of a word assumes any of a set of possible values, the property is left undefined in order to match any possibility later through unification in Prolog. The rules are ordered in conformation to the frequencies of occurrence of the different derivations as given in[9]. In addition, due to the absence of diacritization, as assumed earlier, a single derivation may by satisfied by a number of rules as a word can be interpreted in a number of ways in the absence of diacritics, particularly for verbs. In such cases, the desired choice is assumed to be made by the user (when prompted by the program), or any of the syntax, end-case, or semantic analyzers of the natural Arabic processing system by backtracking and forcing the morphological component to present the next possible construction of the word or to reprocess the word.

Figure 1 shows sample rules of MAS. The predicate *npre_test9* is used to recognize a possible construction of a nine-letter noun. The noun has a three-letter prefix represented by the variables I, H, and G in order. Note that Arabic is read from right to left. The remaining six letters are recognized by the predicate *nsuf_test6* as a six-letter noun. The predicate *conca* is used to match in analysis mode (or construct in synthesis mode) the variables G, H, and I with (from) any of the possible prefixes represented by the variable M. The predicate *ifthen* checks if the rule is being used in synthesis mode, in which case the derivation DEE and the prefix PRE of the remaining six-letter noun are determined in order to synthesize the noun using the predicate *nsuf_test6*. Next the compatibility of the prefix and suffix is guaranteed by assuring that the suffix is not incompatible with the prefix. The predicate *concat* is only useful in analysis mode and has no effect in synthesis mode.

The predicate *nsuf_test8* recognizes a possible construction of an eight-letter noun. The noun has a three-letter suffix represented by the variables A, B, and C in order. The predicates *member* and *conca* check the suffix as being one of two possibilities that imply that the word is a feminine dual noun. The remaining five letters are recognized by the predicate *npre_test5* as a five-letter noun.

The predicate *vpre_test7* is used to recognize a possible construction of a seven-letter verb. The verb has a one-letter prefix represented by the variable G. The remaining six letters are recognized by the predicate *vpre_test6* as a six-letter verb. The predicate *conca* is used to match in analysis mode (or construct in synthesis mode) the variables G, H, and I with (from) any of the possible prefixes represented by the variable M. The rule identifies the tense of the verb as present. This conclusion is forced by the fact that the first letter (prefix) applies

```
% In the rules below the list [A, B, C, ...] represents the letters of the word being processed.
% RO = root, DE = derivation, TY = type of verb (past, present, imperative)
% SDP = number (singular, dual, plural) , MF = gender, PSN = person
% PR = prefix, IN = infix, SU = suffix

   npre_test9([A,B,C,D,E,F,G,H,I],RO,DE,SDP,MF,PR,IN,SU) :-
      member(M,[$وال$,$فلل$,$بال$,$كال$,$فال$]),
      conca([G,H,I],M), ifthen( (var(A)),(conca(DEE,M,DE),conca(PRE,M,PR)) ),
      nsuf_test6([A,B,C,D,E,F],RO,DEE,SDP,MF,PRE,IN,SU),
      not(member(SU,[$هن$,$نا$,$ها$,$هم$,$كم$,$كن$])),
      concat(PRE,M,PR), concat(DEE,M,DE).

   nsuf_test8([A,B,C,D,E,F,G,H],RO,DE, SDP,MF, PR,IN,SU) :-
      member(SU,[$تان$,$تين$]), conca([A,B,C],SU),
      npre_test5([D,E, F,G,H,],RO,DE,SDP,MF,PRE,IN,$$),
      SDP = $مثنى$, MF =$مؤنث$.

   vpre_test7([A,B,C,D,E,F,G],RO,DE,TY,SDP,MF,PSN,PR,IN,SU) :-
      member(G,[$س$,$ل$]), not(member(F,[$و$,$ل$,$س$,$ف$])),
      vpre_test6([A,B,C,D,E,F],RO,DE,TY, MF,PSN,PRE,IN,SU),
      conca([PRE,G],PR), TY = $مضارع$.

   vsuf_test6([A,B,C,D,E,F,],RO,DE,TY,SDP,MF,PSN,PR,IN,SU) :-
      member(F,[$أ$,$ا$]), member(SU,[$ها$,$هن$,$هم$]),
      conca([A,B],SU), conca([C,D,E],RO),DE = $فعل$,TY = $أمر$,
      SDP = $مفرد$, MF = $مذكر$, PSN = $مخاطب$, PR= F, IN = $$.

   art_test4([A,B,C,D],[Oword,TC,Root,Type,X,SU]) :-
       member(D,[$و$,$ف$]), ifthen( (var(A)),(conca(PR,D,X)) ),
      find_art3([A,B,C],TC,Root,Type,PR,SU),
      concat([A,B,C,D],Oword), conca(PR,D,X).
```

FIG. 1. Sample rules of MAS.

only to present tense verbs, and by assuring that the second letter, represented by the variable F is not incompatible with the prefix G.

The predicate *vsuf_test6* is used to recognize a possible construction of a six-letter verb. The verb has a one-letter prefix represented by the variable F. The verb also has a two-letter suffix recognized by the predicate *member* as the variable SU. The predicate *conca* is used to match in analysis mode (or construct in synthesis mode) the variables A, B, and C with (from) any of the possible suffixes represented by the variable SU. The rule identifies the type of the verb as imperative, number as singular, gender as masculine and person as second.

The predicate *art_test4* is used to recognize a possible construction of four-letter particles. The particle has a one-letter prefix represented by the variable D. The remaining three letters are recognized by the predicate *find_art3* as a three-letter particle. The predicates *ifthen, conca* and *concat* are used as mentioned earlier.

The Appendix shows sample output of the program. It is notable that some of the output fields are left undefined in order to match any of a number of possibilities as mentioned earlier. The program was written in Prolog. The number of rules is 80, 150, 200 for particles, verbs, and nouns respectively.

Conclusion

In this paper we have presented a morphological analyzer/synthesizer (MAS) of Arabic words. MAS is based on linguistic principles of Arabic morphology, statistical frequencies of occurrence of words and their derivations, and artificial intelligence techniques.

MAS may produce more than one result for a word since no diacritization is assumed. One can obtain the desired result by rejecting solutions as the analyzer will continue the analysis through backtracking until a solution is accepted. MAS currently validates the produced roots of words according to the phonological properties of letters as mentioned earlier. As a result, a root that is not in use may be produced. However, this approach accommodates the possibility of new roots as the language expands. In addition, since the number of roots in Arabic is between 3000 and 4000[8], a dictionary of roots can be used for validation. Another approach for root validation can be based on the theory of associating semantics with letters[12], and using these semantic properties to validate the roots.

MAS is currently being used as a component of a natural Arabic understanding system NAUS. The syntax module directly calls the modules. MAS can further be used to teach Arabic morphology and in translation, speech, text processing, and character recognition systems. It can also be used in translation, computer-aided Arabic learning, character recognition and text and speech processing systems.

References

[1] **Thalouth, B.** and **Al-Dannan, A.** Hypothesized Algorithms for Decomposition of Modern Arabic Words. *The 1985 Annual Report, IBM Kuwait Scientific Center, Safat, Kuwait.*

[2] **Hilal, Y.** Morphological Analysis of Arabic Speech. *Proceedings of the International Workshop on Computer-Aided Translation, Riyadh, 1985.*

[3] **Hegazi, N.** and **El-Sharkawi, A.** Natural Arabic Language Processing. *Proceedings of the 9th NCC, Riyadh, 1986,* 1-17.

[4] **Geith, M.** and **El-Sadany, T.** An Arabic Morphological Analyzer on a Personal Computer. *Proceedings of the First KSU Symposium on Computer Arabization, Riyadh, 1987,* 55-65.

[5] **Al-Fadaghi, S.** and **Al-Anzi, F.** A New Algorithm to Generate Arabic Root-Pattern Forms. *Proceedings of the 11th NCC, Dhahran, 1989,* 391-400.

[6] **Hilal, Y.** Arabic Morphological Generation. *Proceedings of the 9th National Computer Conference, Riyadh, 1986.*

[7] **Wright, W.** *A Grammar of the Arabic Language, Volume 1.* Cambridge, 1896.

[8] **Al-Othman, A**. An Arabic Morphological Analyzer. MS Thesis, KFUPM, Dhahran, 1990.

[9] A**l-Khuli, M. A.** *a-taraakiib al-ssai9ah fi allugat al9arabiyat - dirasat Iħsa'iyah* (التراكيب اللغوية الشائعة في اللغة العربية – دراسة إحصائية). Dar Al-Uloom, 1982.

[10] **Al-Safran, S.** An Arabic Sentence Generator. *MS Thesis, KFUPM, Dhahran,* 1992.

[11] **Al-Sawadi, A.** and **Khayat, M. G.** An Arabic End-Case Analyzer of Arabic Sentences. *KSU Journal (Computer Division),*V. **8,** No. 1, 1996, 21-52.

[12] **Ibn Jinni, A.** *Al-Khasa'is* (الخصائص). Daar Al-hady, Beirut, Lebanon.

Appendix

The following particle lists have the following form: [word, root, type, prfix , infix, suffix]

[هم , , , حرف جر, إلى , إليهم]
[, , , حرف استفهام, هل , هل]
[نا, , , حرف شرط, إن , إننا]
[, , , حرف نفي, لا , لا]
[ه , , ل , حرف توكيد, أن , لأنه]

The following noun lists have the following form: [word, root, derivation, type, gender, number, person, definite/indefinite, prefix, infix, suffix].

[, ا , , نكرة , غائب , مفرد , مذكر , اسم , فعال, جود , جواد]

[ه , و , , نكرة , غائب , جمع , | , اسم , فعوله, درس , دروسه]

[ة , ,بال , معرفة , غائب, مفرد , | , اسم , الفعلة, لعب , باللعبة]

[ات ,ا ,ال , معرفة , غائب , جمع , مؤنث , اسم , الفعالات, سمو , السماوات]

[, , , معرفة , غائب , مفرد , مذكر , اسم علم , الله, الله , الله]

[,ي ,ل , نكرة , غائب , مفرد , | , اسم , لفعيل, كرم , لكريم]

The following verb lists have the following form: [word, root, derivation, type, gender, number, person, prefix, infix, suffix]

[, ,ي , غائب , مفرد , مذكر , مضارع , يفعل, لعب , يلعب]

[كموها , ,ن , متكلم , جمع , , مضارع , نفعلكموها, لزم , نلزمكموها]

[تك , ,أ , متكلم , مفرد , | , ماضي , أفعلتك, عطي , أعطيتك]

[تني , , , غائب, مفرد , | , ماضي , فعلتني, ظنن , ظننتني]

[ن , ,لأ ,متكلم , مفرد , | , مضارع , لأفعلن, فعل , لأفعلن]

محلل ومركب صرفي عربي

محمد غزالي خياط[*] ، عبد العزيز العثمان[**] و صفران الصفران[**]

[*] *قسم الهندسة الكهربائية وهندسة الحاسبات ، جامعة الملك عبد العزيز*
جدة - المملكة العربية السعودية

[**] *جامعة الملك فهد للبترول والمعادن ، الظهران - المملكة العربية السعودية*

المستخلص . يمثل الصرف عنصرًا أساسيًا في معالجة اللغة العربية آليا . وحيث أن للصرف في اللغة العربية قواعد واضحة فإنه يمكن برمجة التحليل والتركيب الصرفي بسهولة . والهدف من هذا البحث هو تصميم وتطوير محلل ومركب صرفي عربي . وفي حالة التحليل يقوم المحلل بتحديد الخصائص التالية للكلمة : النوع (اسم ، فعل ، حرف) ، والضمير والعدد والجنس (للأسماء والأفعال) ، وحالة الفعل (ماضي ، مضارع ، أمر) ، نوع الحرف (استفهام ، جر ، ... إلخ) ، والجذر ، والوزن (للأسماء والأفعال) ، والزوائد (قبلية ، وسطية ، بعدية) . وفي حالة التركيب يقوم البرنامج بتركيب الكلمة من الخصائص المذكورة أعلاه .

لقد تم تطوير البرنامج بناء على قواعد الصرف العربي . وتم تصميم البرنامج كثلاث وحدات للحروف والأسماء والأفعال . وتتكون كل وحدة من قواعد برمجية تمثل قواعد الصرف العربي . ويحدد البرنامج الحالة (تحليل أو تركيب) تلقائيا من المعطيات . وقد تم تمثيل الأوزان المختلفة لكلمة تتكون من عدد س من الحروف كقواعد لغة برولوج مرتبة وفقا لتردد استخدام الوزن . ويستخدم عدد الحروف التي تتكون منها الكلمة وتردد الوزن لتقليل وقت البحث عن التركيب أو التحليل الصحيح في البرنامج .

هذا ويتم استخدام البرنامج المطور حاليا كوحدة في نظام لفهم اللغة العربية . كما يمكن استخدام البرنامج في الترجمة الآلية ، والنظم الآلية لتعليم اللغة العربية ، ونظم التعرف على الكلام المكتوب ، ونظم معالجة الكلام المنطوق .