A MATRIX REPRESENTATION OF THE INFLECTIONAL FORMS OF ARABIC WORDS: A STUDY OF CO-OCCURRENCE PATTERNS

H.E. Mahgoub, M.A. Hashish
IBM Cairo Scientific Centre
56 Gameaat Al Dowal Al Arabeya Street
Mohandessen, Giza, Egypt

A.T. Hassanein
Arabic Department, American University in Cairo

ABSTRACT

A proposed "Matrix" method for the representation of the inflectional paradigms of Arabic words is presented. This representation results in a classification of Arabic words into a tree structure (Fig(1)) whose leaves represent unique conjugational or derivational paradigms, each represented in the proposed "Matrix" form.

A study of about 2,500 stems from a high frequency Arabic wordlist due to Landau <1> has revealed a systematic set of co-occurrence patterns for the enclitic pronouns of Arabic verbs and for the possessive pronouns attached to Arabic nouns. Each co-occurrence pattern represents a subcategorization frame that reflects the underlying semantic relationship.

The key feature that distinguishes these semantic patterns has been observed to be whether the attached suffixes relate to the animate or inanimate. In some cases for verbs, the number of the subject is also a significant feature. These semantic features also extend to non-attached subjects and objects (for verbs) and to possessive noun complements (for nouns). Therefore the semantic classes presented in this paper also assist in syntactic/semantic analysis.

The first application that was developed, based upon the proposed representaion is a stem-based Arabic morphological analyser, from which a spell checker (on a PS/2 microcomputer) emerged as a by-product. Currently, the system is being used to interact with an Arabic syntactic parser and there are plans to use it in a machine assisted translation system.

1. INTRODUCTION

Over the past few years there has been a marked increase in the use of computers in the Arabic speaking countries. Many applications programs in Arabic have been developed, but the field of computational linguistics is relatively new in Arabic and presents a unique challenge, due to the highly inflected nature of the Arabic language.

In the present work, we have attempted to represent the morphological rules governing the inflections of Arabic words in a compact form which can simplify the processing of Arabic words by computers and which is independent of the a particular application. There have been other attempts to show the conjugations of Arabic verbs <2> but the treatment does not delve into sufficient depth and not all enclitics, which are an essential part of Arabic verbs, are considered. Moreover, the treatment in <2> does not extend to nouns.

By studying some 2,500 stems out of a high frequency Arabic wordlist due to Landau <1>, certain systematic co-occurrence patterns governing verb enclitics and noun possessive pronouns have been observed. These patterns are what we call "Matrices" in this paper; each unique "Matrix" reflects a different semantic behaviour.

To summarize Arabic morphology in a nutshell, about 80% of Arabic words can be derived from a sequence of three letters, called a triliteral root.

For example,if we consider the root ك ت ب (K T B), we can form words such as أكتب (?aKTuB - I write) and كتاب (KiTa:B - book), by subjecting the root to various "forms" or "moulds" and by undergoing certain morpho-phonemic (and possibly also morpho-graphemic) changes. For a full discussion of traditional Arabic morphology see <9> and <10>. In this paper, we shall define such an inflected form to be a "STEM".

Thus a stem may contain infixes and certain prefixes which are part of the "mould" but may not contain any suffixes. Suffixes for verbs are subject and object pronouns, while for nouns they are possessive pronouns.

One further definition which is used in the proposed representaion is the "Core"; this is simply the inflected form with all prefixes and suffixes stripped off. The core may or may not be a valid word.

In comparison with other work in the area of traditional Arabic morphology (<3>,<4>), where the concern is with the rules which cause the inflected form to be derived from the ROOT, we have studied the rules governing the derivation of all possible inflected forms from the STEM, as defined above.

2. THE MATRIX REPRESENTATION

Sample "MATRIX PARADIGMS" are shown in Fig(2) for verbs and Fig(3) for nouns. Table(1) gives the keys in English to the columns on the Matrix Paradigms. The inflected form for a given Person/Number/Gender/Mode combination for verbs (obtained from the relevant "row" of the Matrix Paradigm) is constructed by concatenating the prefix, core and both subject and object pronoun column entries. The inflected forms for nouns are similarly constructed for a particular Number/Gender/Case combination.

The various "cells" of the object pronoun columns indicate whether a particular entry is valid (indicated by "١", an Arabic numeral one). Invalid entries are indicated by a "٠", an Arabic zero. It is due to this matrix of ones and zeros that the representation was named the "Matrix Paradigm".

3. TAXONOMY OF ARABIC WORDS

Fig(1) shows a tree diagram representing the taxonomical classification of Arabic verbs and nouns. There are different "levels" in the tree correspond to different types of variations of the inflected form from one class to another. The first type of variation coincides more or less with the traditional classification and is respresented at levels 2 and 3 for verbs and at level 2 for nouns.

Each Matrix Paradigm also reflects two further types of variation, which can be considered separately from one another. The first is the variation in the core with the different rows; this dimension corresponds, for example, to the traditional study of verb conjugations (see <2>).

The other type of variation is that in the distribution of the Matrix of ones and zeros, which is essentially a variation in the co-occurrence of object pronouns (for transitive verbs) and possessive pronouns (for nouns). This variation is reflected at level 4 of the taxonomy. In the following sections 3.1 and 3.2, we will discuss the study of these co-occurrence patterns in more detail for verbs and nouns separately.

3.1 CO-OCCURRENCE PATTERNS FOR VERBS

On examination of the Landau <1> high frequency wordlist, the following features seemed to distinguish classes of verbs apart:

1- Whether the subject is human or non-human (for both transitive and intransitive verbs).
2- Whether the object is human or non-human (for transitive verbs only).
3- The number of the subject (for intransitive verbs only).

In Arabic, there is a set of object pronouns which refers to a non-human object: (هما,ها,ه) and this will be denoted by -H. This is a subset of the complete set of pronouns +H, which denotes human and non-human. Below, we will discuss the features for transitive and intransitive verbs separately:

(a) Transitive Verbs:

As shown in the table below, there can only be 4 combinations of the features +H and -H. Each of the feature sets in the table has been designated a class code. Only verbs with features corresponding to the feature sets B,C and D have been found in the Landau <1> shortlist examined.

Feature Set code	Subject	Object
B	+H	+H
C	+H	-H
D	+H	-H
?	-H	-H

(b) Intransitive Verbs:

It was found out that the subject number is an additional distinguishing feature for transitive verbs. Moreover, the subject number is significant only in the case of human subjects. For non-human subjects, this feature is not significant.

Based upon the above observations, we will define the distinguishing features for intransitive verbs to be +H(s),+H(dp) and -H, where s denotes singular and dp denotes dual/plural. +H(s) and +H(dp) denote the sets of singular and dual/plural subjects, respectively. By definition +H(s) U +H(dp) -H, where U denotes the union of the two feature sets. The table below shows the possible combinations of these features; only features designated by A,E and F were found for Landau's <1> shortlist.

Feature Set Code	Subject Features
A	+H(s) U +H(dp)
E	-H
F	+H(dp)
?	+H(s)

3.2 CO-OCCURRENCE PATTERNS FOR NOUNS

The same set of object pronouns for verbs denotes the possessive pronouns for nouns, with the exception of a slight difference in form of the first person singular. The -H set is exactly the same. Three distinct classes of Matrix patterns (see level 3 of Fig(1)) have been observed for nouns:

(A) No possessive pronouns can be attached.
(B) All possessive pronouns can be attached.
(C) Only possessive pronouns related to the inanimate (set -H) can be attached.

An additional study was made to determine what Number/Gender (NG) combinations are valid for a particular noun stem. These have been found to be an important feature of Arabic nouns, as not all NG combinations are valid for a stem. Each stem needs to be examined separately and this information is put into the lexicon of stem. The NG combinations are represented at level 3 of the taxonomy, for nouns (see Fig(1)).

Although there is no systematic, theoretical method for determining what all the different NG combinations are for comprehensive coverage of nouns, yet by examining more and more nouns from Landau's <1> wordlist, some form of convergence occurred. For the 2,500 stem shortlist, there were only 17 NG combinations.

This curious feature of Arabic nouns can be mainly attributed to the presence of words of foreign origin and to the pragmatics of the noun in question.

4. APPLICATIONS DEVELOPED

As a first application, an Arabic stem-based morphological analyser has been developed on an IBM PS/2 microcomputer. Morphological features of the word analysed are computed.

As a by-product of the analyser, an Arabic spelling verifier has been developed, by including unification of the morphological and co-occurrence features of the morphemes.

The system is currently being developed for use in the interaction with an Arabic syntactic parser.

ACKNOWLEDGEMENT

The authors sincerely wish to thank Dr. John Sowa for reviewing this paper and for his invaluable comments and suggestions.

REFERENCES

<1> Jacob Landau, "A Word Count of Modern Arabic Prose", American Council of Learned Societies, New York, 1959.

<2> Peter F. Abboud, Ernest N. McCarus (Eds.), "Elementary Modern Standard Arabic", Parts 1 & 2 (2nd. Edition), Cambridge University Press, 1986.

<3> T. El-Sadany & M. Hashish, "Arabic Morphological System", IBM Natural Language Processing Conference, Thornwood, New York, October 1988.

<4> T. El-Sadany & M. Hashish, "Arabic Morphological System", IBM Systems Journal 25th. Anniversary for Scientific Centres issue, Vol. 28, No. 4, 1989.

<5> فؤاد نعمة، ملخّص قواعد اللغة العربية، الطبعة الخامسة، المكتب العلمى" للتأليف والترجمة.

<6> عبد العليم ابراهيم، الاملاء والترقيم فى الكتابة العربية، مكتبة غريب، القاهرة.

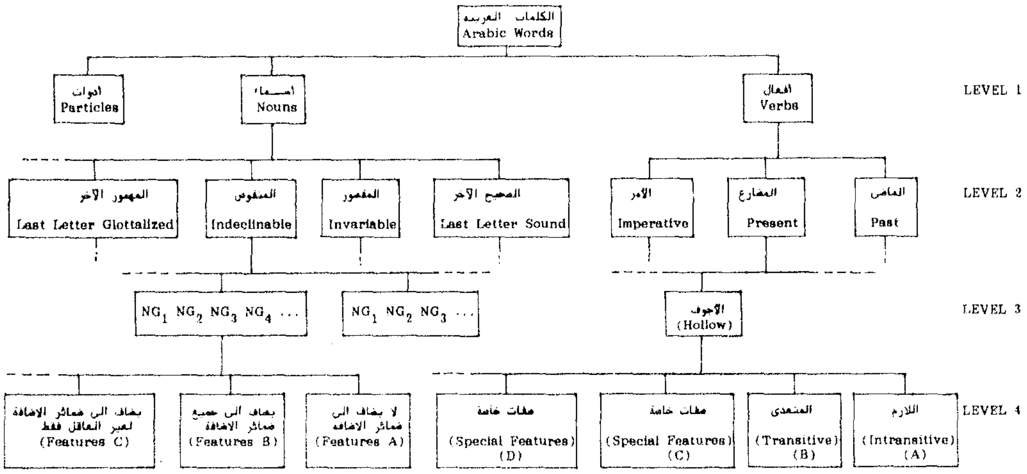
<7> عبد العليم ابراهيم، دليل الاملاء، القاهرة، ١٩٧٠

<8> د. أحمد طاهر حسنين، نظريّة الاكتمال اللغوى" عند العرب، هجر، القاهرة، ١٩٨٧

<9> الحملاوى، شذا العرف فى فن الصرف، طبعة ١١

<10> ابن عصفور، الممتع فى التصريف، تحقيق د. فخر الدين قباوة، الطبعة الثالثة، بيروت، ١٩٦٠

FIG(1): Tree Structure Classification of the Arabic Language



FIG(2): Sample Paradigm for Present Tense Verb : الفعل المضارع الأجوف الألفي المتعدي

<---- 6									---->	<-5->	<--- 4 -->	<--3 ->	<-- 2 ->	<------ 1 ----->
هن	هم	هما	ه	كن	كم	كما	ك	نا	ني	ضمير فاعل	الصيغة او جزء منها	سوابق	شخص، عدد،نوع	الحالة الاعرابية
١	١	١	١	١	١	١	١	٠	٠	x	خاف	ا	انا	I & S رفع و نصب
١	١	١	١	١	١	١	١	٠	٠	x	خف			J جزم
١	١	١	١	١	١	١	١	٠	٠	x	خاف	ن	نحن	I & S رفع و نصب
١	١	١	١	١	١	١	١	٠	٠	x	خف			J جزم
١	١	١	١	٠	٠	٠	٠	١	١	x	خاف	ت	انت	I & S رفع و نصب
١	١	١	١	٠	٠	٠	٠	١	١	x	خف			J جزم
١	١	١	١	٠	٠	٠	٠	١	١	ين	خاف	ت	انت	I رفع
١	١	١	١	٠	٠	٠	٠	١	١	ي	خاف			S & J نصب و جزم
١	١	١	١	٠	٠	٠	٠	١	١	ان	خاف	ت	انتما	I رفع
١	١	١	١	٠	٠	٠	٠	١	١	ا	خاف			S & J نصب و جزم
١	١	١	١	٠	٠	٠	٠	١	١	ون	خاف	ت	انتم	I رفع
١	١	١	١	٠	٠	٠	٠	١	١	وا/و	خاف			S & J نصب و جزم
١	١	١	١	٠	٠	٠	٠	١	١	ن	خف	ت	انتن	ISJ رفع-نصب-جزم
١	١	١	١	١	١	١	١	١	١	x	خاف	ي	هو	I & S رفع و نصب
١	١	١	١	١	١	١	١	١	١	x	خف			J جزم
١	١	١	١	١	١	١	١	١	١	x	خاف	ت	هي	I & S رفع و نصب
١	١	١	١	١	١	١	١	١	١	x	خف			J جزم
١	١	١	١	١	١	١	١	١	١	ان	خاف	ي	هما	I رفع
١	١	١	١	١	١	١	١	١	١	ا	خاف			S & J نصب و جزم
١	١	١	١	١	١	١	١	١	١	ان	خاف	ت	هما	I رفع
١	١	١	١	١	١	١	١	١	١	ا	خاف			S & J نصب و جزم
١	١	١	١	١	١	١	١	١	١	ون	خاف	ي	هم	I رفع
١	١	١	١	١	١	١	١	١	١	وا/و	خاف			S & J نصب و جزم
١	١	١	١	١	١	١	١	١	١	ن	خف	ي	هن	ISJ رفع-نصب-جزم

FIG(3): Sample Paradigm for a Noun
الإســـم الصحيح الآخر غير المهموز

<---- 6										---->	<- 5 ->	<--- 4 --->	<-- 8 ->	<--- 1 --->	<- 7 ->
هن	هم	هما	ها	ه	كن	كم	كما	ك	نا	ي	لواحق	الصيغة او جزء منها	معرفة/ نكرة	الحالة الاعرابية	العدد/ النوع
١	١	١	١	١	١	١	١	١	١	١	x	كاتب	نكرة	رفع	مفرد (A) مذكر
١	١	١	١	١	١	١	١	١	١	١	x/ا	كاتب		نصب و جر	
٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	x	كاتب	معرفة	رفع	
٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	x	كاتب		نصب و جر	
١	١	١	١	١	١	١	١	١	١	١	ان/ ا	كاتب	نكرة	رفع	مثنى (B) مذكر
١	١	١	١	١	١	١	١	١	١	٠	ين/ ي	كاتب		نصب و جر	
٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	ان	كاتب	معرفة	رفع	
٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	ين	كاتب		نصب و جر	
١	١	١	١	١	١	١	١	١	١	١-	ون/ و	كاتب	نكرة	رفع	جمع (C) مذكر سالم
١	١	١	١	١	١	١	١	١	١	٠	ين/ ي	كاتب		نصب و جر	
٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	ون	كاتب	معرفة	رفع	
٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	ين	كاتب		نصب و جر	
١	١	١	١	١	١	١	١	١	١	١	ة/ت	كاتب	نكرة	رفع	مفرد (D) مؤنث
١	١	١	١	١	١	١	١	١	١	١	ة/ت	كاتب		نصب و جر	
٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	ة	كاتب	معرفة	رفع	
٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	ة	كاتب		نصب و جر	
١	١	١	١	١	١	١	١	١	١	١	تان/تا	كاتب	نكرة	رفع	مثنى (E) مؤنث
١	١	١	١	١	١	١	١	١	١	٠	تين/تي	كاتب		نصب و جر	
٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	تان	كاتب	معرفة	رفع	
٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	تين	كاتب		نصب و جر	
١	١	١	١	١	١	١	١	١	١	١	ات	كاتب	نكرة	رفع	جمع (F) مؤنث سالم
١	١	١	١	١	١	١	١	١	١	١	ات	كاتب		نصب و جر	
٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	ات	كاتب	معرفة	رفع	
٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	ات	كاتب		نصب و جر	

TABLE (1)

KEY TO COLUMNS ON MATRIX PARADIGMS

KEY	DESCRIPTION
1	MODE for Verbs: Indicative (I): رفع Subjunctive (S): نصب Jussive (J): جزم CASE for Nouns: Nominative (N): رفع Accusative (A): نصب Genitive (G): جر
2	Person-Number-Gender (Verbs only)
3	Prefix
4	Core
5	Subject Pronoun (for verbs) Case Ending (for nouns)
6	Object Pronoun (for verbs) Possessive Pronoun (for nouns)
7:	Number and Gender (Nouns only):
A	Masculine Singular
B	Masculine Dual
C	Masculine Plural (Sound)
D	Feminine Singular
E	Feminine Dual
F	Feminine Plural (Sound)
8	Definite/Indefinite (Nouns only): Definite: معرفة Indefinite: نكرة