Developing a tagset for automated POS tagging in Arabic

SHIHADEH ALQRAINY and ALADDIN AYESH
Centre for Computational Intelligence (CCI) - School of Computing
De Montfort University
Leicester – The Gateway
UNITED KINDOM
{alqrainy , aayesh}@dmu.ac.uk

Abstract: - Arabic language has much more syntactical and morphological information. Diacritics, which are marks placed over and below the letters of Arabic word, play a great role in adding linguistic attributes to Arabic word in part-of-speech tagging system. This paper describes a tagset that were built based on the inflectional morphology system which derived from traditional Arabic grammatical theory. The tagset developed represent an early stage of research related to automatic morphosyntactic annotation in Arabic language. This paper aims to present a general tagset for use in diacritics-based automated tagging system that is underdevelopment by the author.

Key-Words: - Part-of-Speech (POS), Arabic Language, Tagset, Diacritics, Syntactical, Morphological.

1 Introduction

A tag is a code which represents some features or set of features and is attached to the segment in a text. Single or complex information are carried by a tag [8]. In the case of POS Tagging, a POS tagset to categories and mark up the words of the target text is an absolutely necessary preliminary [3]. The development of a tagset to support diacritical based tagging system is at early stage. Little work has been done in developing Arabic tagset. The need for such a tagset comes from the fact that there is no standardized and comprehensive Arabic tagset.

An overview of Arabic language followed by diacritics in Arabic described in this paper. Tagset background and EAGLES guidelines overview presented. Finally we will present our tagset (Analysis and Hierarchy) followed by conclusion and future work.

2 Arabic Language

2.1 Background

The Arabic language is spoken in more than 20 Countries, from Egypt to Morocco and throughout the Arabian Peninsula. It is the native language of over 195 million people. Plus, at least another 35 million speak Arabic as a second language.

Modern Standard Arabic (MSA) is the official language throughout the Arab world, and its written form is relatively consistent across national boundaries. MSA is used in official documents, in educational settings, and for communication between Arabs of different nationalities. However, the spoken forms of Arabic vary widely, and each Arab country has its own dialect. Dialects are spoken in most informal settings, such as at home, with friends, or while shopping.

The Arabic language belongs to the Semitic family of languages, written from right to left. Arabic has been a literary language since the 6th century A.D., and is the liturgical language of Islam in its classical form.

The Arabic writing system is quite different from the English system. The Arabic alphabet consists of 28 letters that change shape depending on their position within a word and the letters by which they are surrounded. Some Arabic letters must be connected to other letters; others may stand alone. Arabic vowels are indicated by marks (Diacritics) above and below the consonants. In many cases, these diacritics play the role of vowels in English and thus influence pronunciation. Additionally, there are no special forms, such as the use of capital letters in English, to indicate proper nouns or the beginning of a sentence [10].

2.2 Diacritics in Arabic

Diacritics are marks placed over and below the letters of Arabic word. This feature plays a great role in adding linguistic attributes to Arabic words which help us to assign the most likely tag of the word in POS tagging system and in indicating pronunciation and grammatical function of the words. It is particularly of interest for the purpose of this paper. Table 1 shows Arabic vowel diacritics.

The pronunciation of diacritized languages words cannot be fully determined by spelling their characters only; special marks are put above or below the characters (Diacritics) to determine the correct pronunciation and indicate the grammar function of the word within the sentence. For example, the word "**كتبت** " without mark (Diacritic) may be pronounced to mean "*He wrote*" , "*It was written*", "*books*". The reader may refer to the context the word appears in to decide which of the words is actually intended. In such languages, two different words may have identical spelling whereas their pronunciations and meanings are totally different [2].

In Arabic, short vowels are not apart of the Arabic alphabet. They are used in both Noun and Verb in Arabic Language. They indicate the case of the noun and the mood of the verb.

Short Vowels (Diacritics)

Name	*Fatha*		*Damma*		*Kasra*	
Symbol	ـَ	/a /	ـُ	/u /	ـِ	/i /
Explanation	Written above the consonant.		Written above the consonant.		Written below the consonant.	
Example	بَ		بُ		بِ	
Pronunciation	ba		Bu		bi	

Nunation " Tanween" (Diacritics)

Name	*Tanween Fath*		*Tanween Damm*		*Tanween Kasr*	
Symbol	ً	/an/	ٌ	/un/	ٍ	/in/
Explanation	Written above the consonant.		Written above the consonant.		Written below the consonant.	
Example	بًا		بٌ		بٍ	
Pronunciation	ban		bun		bin	

Shadda & Sukun (Diacritics)

Name	*Shadda*	*Sukun*
Symbol	ّ	ْ
Explanation	Written above the consonant.	Written above the consonant.
Example	بّ	بْ
Pronunciation	bb	b

Table 1: Arabic vowel diacritics

3 Arabic Tagset and EAGLES guidelines

EAGLES [9] guidelines outline a set of features for Tagsets, these guidelines were designed to help standardize tagsets for what were then the official languages of the European Union.

EAGLES tags are defined as sets of morphosyntactic attribute-value pairs (e.g. Gender is an attribute that can have the values Masculine, Feminine or Neuter)[3]. The tagset discussed here is not being developed in accordance with the EAGLES guidelines for morphosyntactic annotation of corpora. Arabic is very different from the languages for which EAGLES was designed, and belongs to the Semitic family rather than the Indo-European one.

Following a normalized tagset and the EAGLES recommendations would not capture some of Arabic relevant information, such as the jussive mood of the verb and the dual number that are integral to Arabic. Another important aspect of Arabic is inheritance, where all subclasses of words inherit properties from the classes from which they are derived. For example, all subclasses of the noun inherit the "Tanween" nunation when in the indefinite which is one of the main properties of the noun [7].

3.1 Previous work on POS tagsets

There are numbers of popular tagsets for English, such as : 87-tag tagset used Brown Corpus , 45-tag Penn Treebank tagset and 61-tag C5 tagset, TOSCA tagset, ICE tagset, LUND tagset [5][3]. For Arabic also very small number of tagset had been built, El-Kareh S, Al-Ansary [1] described the tagset ,they classifying the words into three main classes, Verbs are sub classified into 3 subclasses; Nouns into 46 subclasses and Particles into 23 subclasses. Shereen Khoja [7] described more detail tagset. Her tagset contains 177 tags, 57 Verbs, 103 Nouns, 9 Paricles, 7 residual and 1 punctuation.

3.2 Proposed Arabic Tagset: Analysis

It is necessary to have a model of the language to create the linguistic categories of a tagset. An ideal approach would be to derive this model from the grammatical description of the language.

Since the grammar of Arabic has been standardized for centuries, it is logical to derive our morphosyntactic Arabic tagset from this grammatical tradition that has been used for around fourteen centuries by all students of Arabic.

Arabic grammarians and linguists have always used the Arabic system of inflectional morphology called "**الإعراب**" when teaching Arabic grammar to students. For example, given the sentence " **لعبَ الولدُ** " "the boy played", students would have to say that the first word is the indeclinable, indicative, perfect verb, while the second word is the nominative subject[7][3].
The proposed Arabic tagset in this paper is based on the inflectional morphology system. Arabic grammarians traditionally analyses all Arabic words into three main parts-of-speech. These parts-of-speech are further sub-categorised into more detailed parts-of-speech which collectively cover the whole of the Arabic language [4]. These are:
• **Noun:** A noun in Arabic is a name or a describing-word for a person, a thing or an idea. This includes not only the English equivalent of a noun, but also adjectives, proper nouns and pronouns.
• **Verb:** Verb: Verbs are the same in Arabic as they are in English in that they denote actions.
• **Particle:** Particles include prepositions, conjunctions, Exceptions, Vocative, Annulment, Subjunctive, and Jussive.

3.2.1 Noun

A noun in Arabic indicates a meaning by itself without being connected with the notion of time and refers to a person, place, thing, event, substance or quality.
Nouns are also divided into the following types: (Common, Demonstrative, Relative, Personal, Adverb, Diminutive, Instrument, Conjunctive, Interrogative, Proper, and Adjective).
The linguistic attributes of nouns that have been used in this tagset are:

• **Case:** Arabic nouns have three cases: nominative, accusative and genitive. For example, the words " **الدرسُ ، الدرسَ ، الدرسِ** " which mean "*the lesson*", indicate the above three cases respectively.
Without the case marker associated with the last letter of the above words (e.g short vowels), it's difficult to determine the case of that word.

• **State**: Arabic nouns are marked for definiteness and indefiniteness. Definiteness is marked by the article "**ال**", which means" *the*". For example, the words "**الكتاب**" and "**كتاب**" which mean" *the book*"," *a book*" indicate definiteness/indefiniteness respectively.

• **Number:** Arabic has three numbers: singular, dual, and plural. For example, the words "**ولد**","**ولدان**" and "**أولاد**" which mean " *a boy* ", " *two boys* " and "*boys* " indicate singular, dual, and plural respectively.

• **Gender:** Arabic nouns have three genders: masculine, feminine and neuter. Most common noun ends with "Tanween". Most feminine singular nouns end with a round Ta (marbuta). For example, the words "**ملك**","**طائرة**" and "**جماعة**" , which mean " *a king* ", " *a plane* " and " *group of people* " indicate masculine, feminine and neuter respectively.

• **Person:** Arabic nouns have three persons: the speaker (First person), the individual spoken to (Second person), and individual spoken of (third person). For example, the personal noun and "**أنا**" which mean " *I* ", " *You* " and " *He* " indicate First, Second, and third person respectively.

3.2.2 Verb

Arabic verbs are deficient in tenses. Moreover, these tenses do not have accurate time significances as in Indo-European languages [6].

The verb in the Arabic language implies a state or action and a notion of time combined with them and has several aspects: Perfect, Imperfect and Imperative.
The Perfect verb indicates a state or a fact in the past. For example, the word "**كتبَ**" which means "*He wrote*".
The Imperfect verb expresses an action still unfinished at the time to which reference is being made. For example, the word "**يأكلُ**" which means "*He is writing*".
The Imperative verb indicates an action demanded to be carried out in the future. For example, the word "**أكتب**" which means "*you write*".

The linguistic attributes of Verbs that have been used in this tagset are:

• **Mood:** Arabic Verbs have three moods: Indicative, Subjunctive and Jussive. In Verbs, the words "**كتبَ**", "**كتبتُ**" and "**كتبتِ**" which mean " *He wrote* " , " *I wrote* " and " *You wrote* " indicate Indicative, Subjunctive, Jussive mood respectively.

• **Number:** Arabic has three numbers: singular, dual, and plural. For example, the words "**قرأ**", "**يقرأن**" and "**قرأوا**" which mean " *He read* ", " *(two people) read* " and " *they read* " indicate singular, dual, and plural number respectively.

• **Gender:** Arabic verbs have two genders: masculine, feminine. For example, the words "كتبَ" and "كتبتْ" which mean ” He wrote ”and ” She wrote ”.

• **Person:** Arabic verbs have three persons: the speaker (First person), the individual spoken to (Second person), and individual spoken of (third person).
For example, the words which mean the words "كتبَ", "كتبتُ" and "كتبتِ" which mean ” *He wrote* ” , ” *I wrote* ” and ” *You wrote* ” indicate First, Second, and third person respectively.

3.2.3 Particle

In Arabic, particles are classified as one of the three main categories as part of speech, some of the particles used with Verbs and effective the mood of verb when precedes the Verb word. For example, the particles "لم" (Jussive), "كي" (Subjunctive), some of them used with Nouns. For example, the particles "في" (Preposition), "الا" (Exception), and some used with both the noun and the verb. For example, the particle "و" (Conjunction).

3.3 Proposed Arabic Tagset: Hierarchy

We have based our Arabic tagset on inflectional morphology system. The traditional description of Arabic grammarians consider as a base to create the linguistic categories of Arabic tagset. Arabic grammarians describe Arabic as being derived from three main categories: noun, verb and particle. Figure 1 shows the tagset hierarchy.

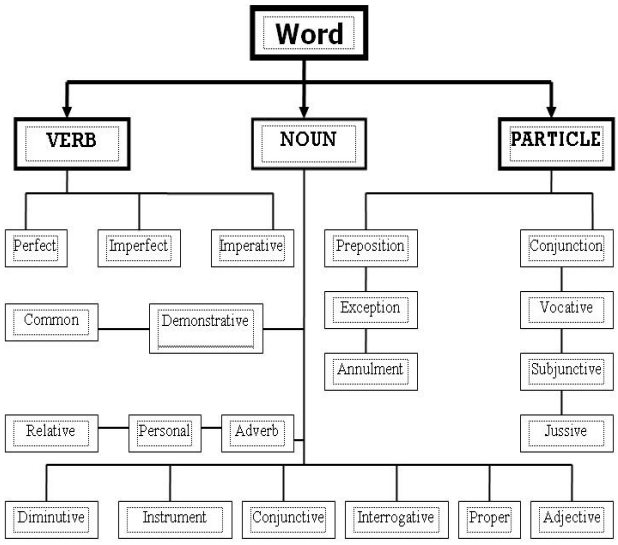

Fig. 1: Tagset Hierarchy.

The tagset has the following main formula:
[T , S , G , N , P , M , C , F] , Where:
T (Type) = {Verb, Noun, Particle}
S = Sub-Class {*Common, Demonstrative, Relative, Personal, Adverb,Diminutive,Instrument, Conjunctive, Interrogative, Proper and Adjective*}
G (gender)= {Masculine, Feminine, Neuter}
N (Number) = {Singular, Plural, Dual}
P (Person) = {First, Second, Third}
M (Mood) = {Indicative, Subjunctive, Jussive}
C (Case) = {Nominative, Accusative, Genitive}
F (State) = {Definite, Indefinite}

Figure 2 shows the Abbreviations which was used to define the words in our tagset.
A sample of our tagset shown in Table 2.

Word	Abb	Word	Abb
Verb	Ve	Annulment	An
Noun	Nu	Subjunctive	Sb
Particle	Pr	Masculine	Ma
Perfect	Pe	Feminine	Fe
Imperfect	Pi	Neuter	Ne
Imperative	Pm	Singular	Sn
Common	Cn	Plural	Pl
Adjective	Aj	Dual	Du
Demonstrative	De	First	Fs
Relative	Re	Second	Sc
Personal	Ps	Third	Th
Diminutive	Dm	Indicative	Dc
Instrument	Is	Subjunctive	Sj
Proper	Pn	Jussive	Js
Adverb	Ad	Nominative	Nm
Interrogative	In	Accusative	Ac
Conjunction	Cj	Genitive	Ge
Preposition	Pp	Definite	Df
Vocative	Vo	Indefinite	Id
Conjunction	Co		
Exception	Ex		

Fig. 2: Tagset Abbreviations

Let us try to explain the symbols of the tagset formula for a moment.
The symbols **[T , S , G , N , P , M]** consider as linguistic attributes for class Verb, while the symbols **[T , S , G , N , P , C , F]** consider as linguistic attributes for class Noun. For example , the word " كتبَ " which means ” *He wrote* ” has the following tag [**VePeMaSnThSj**], which means [*Perfect Verb , Masculine Gender , Singular Number , Third Person , Subjunctive Mood*].

4 Conclusion and Future Work

In this paper, we described a morphosyntactic tagset that is derived from the ancient Arabic grammar, which is based on Arabic system of inflectional morphology. The tagset represent an early stage for use in a word-class based automated tagging system that is underdevelopment by the author. The tagset does not follow the traditional Indo-European tagset that is based on Latin but is instead based on the Semitic tradition of analyzing language.

These tags contain a large amount of information and add more linguistic attributes to the word. Also, we are currently expanding our tagset to cover most categories word in Arabic.

Tag	Description
VePeMaSnThSj	Verb, Perfect, Masculine, Singular, Third Person, Subjunctive
VePeMaSnFsDc	Verb, Perfect, Masculine, Singular, First Person, Indicative
VePeMaSnSeSj	Verb, Perfect, Masculine, Singular, First Person, Subjunctive
VePeFeSnSeJs	Verb, Perfect, Feminine, Singular, Second Person, Jussive
VePeFeSnThJs	Verb, Perfect, Feminine, Singular, Third Person, Jussive
VePeNeDuSeSj	Verb, Perfect, Neuter, Dual, Second Person, Subjunctive
VePeMaDuThSj	Verb, Perfect, Masculine, Dual, Third Person, Subjunctive
VePeFeDuThSj	Verb, Perfect, Feminine, Dual, Third Person, Subjunctive
VePeMaPlFsSj	Verb, Perfect, Masculine, Plural, First Person, Subjunctive
VePeMaPlSeJs	Verb, Perfect, Masculine, Plural, Second Person, Jussive
VePeFePlSeJs	Verb, Perfect, Feminine, Plural, Second Person, Subjunctive
VePeFePlThJs	Verb, Perfect, Feminine, Plural, Third Person, Subjunctive
VePeMaPlThDc	Verb, Perfect, Masculine, Plural, Third Person, Indicative
VePiMaSnThDc	Verb, Imperfect, Masculine, Singular, Third Person, Indicative
VePiMaSnFsDc	Verb, Imperfect, Masculine, Singular, First Person, Indicative
VePiFeSnThDc	Verb, Imperfect, Feminine, Singular, Third Person, Indicative
VePiNePLFsDc	Verb, Imperfect, Neuter, Plural, First Person, Indicative
VePiMaDuThJs	Verb, Imperfect, Masculine, Dual, Third Person, Jussive
VePiFeDuSeJs	Verb, Imperfect, Masculine, Dual, Third Person, Jussive
VePiMaPlThSj	Verb, Imperfect, Masculine, Plural, Third Person, Subjunctive
VePiFePlThSj	Verb, Imperfect, Feminine, Plural, Third Person, Subjunctive
VePmMaSnSeJs	Verb, Imperative, Masculine, Singular, Second Person, Jussive
VePmNeDuSeSj	Verb, Imperative, Neuter, Dual, Second Person, Subjunctive
VePmFePlSeSj	Verb, Imperative, Feminine, Plural, Second Person, Subjunctive
VePmMaPlSeSj	Verb, Imperative, Feminine, Plural, Second Person, Subjunctive
NuAjMsSnNmId	Adjective Noun, Masculine, Singular, Nominative, Indefinite
NuAjMsSnAcId	Adjective Noun, Masculine, Singular, Accusative, Indefinite
NuAjMsSnGeId	Adjective Noun, Masculine, Singular, Genitive, Indefinite
NuAjMsSnNmDf	Adjective Noun, Masculine, Singular, Nominative, Definite
NuAjMsSnAcDf	Adjective Noun, Masculine, Singular, Accusative, Definite
NuAjMsSnGeDf	Adjective Noun, Masculine, Singular, Genitive, Definite
NuAjMsDuGeId	Adjective Noun, Masculine, Dual, Genitive, Indefinite
NuAjMsDuGeDf	Adjective Noun, Masculine, Dual, Genitive, Definite
NuAjFeSnNmId	Adjective Noun, Feminine, Singular, Nominative, Indefinite
NuAjFeSnAcId	Adjective Noun, Feminine, Singular, Accusative, Indefinite
NuAjFeSnGeId	Adjective Noun, Feminine, Singular, Genitive, Indefinite
NuAjFeSnNmDf	Adjective Noun, Feminine, Singular, Nominative, Definite
NuAjFeSnAcDf	Adjective Noun, Feminine, Singular, Accusative, Definite
NuAjFeSnGeDf	Adjective Noun, Feminine, Singular, Genitive, Definite
NuAjFeDuGeId	Adjective Noun, Feminine, Dual, Genitive, Indefinite
NuAjFeDuGeDf	Adjective Noun, Masculine, Dual, Genitive, Definite
NuAjMaPlAcId	Adjective Noun, Masculine, Plural, Accusative, Indefinite
NuAjMaPlGeId	Adjective Noun, Masculine, Plural, Genitive, Indefinite
NuAjMaPlNmId	Adjective Noun, Masculine, Plural, Nominative, Indefinite

NuAjMaPlNmDf	Adjective Noun, Masculine, Plural, Nominative, Definite
NuAjMaPlAcDf	Adjective Noun, Masculine, Plural, Accusative, Definite
NuAjMaPlGeDf	Adjective Noun, Masculine, Plural, Genitive, Definite
NuAjFePlNmId	Adjective Noun, Feminine, Plural, Nominative, Indefinite
NuAjFePlAcId	Adjective Noun, Feminine, Plural, Accusative, Indefinite
NuAjFePlGeId	Adjective Noun, Feminine, Plural, Genitive, Indefinite
NuAjFePlNmDf	Adjective Noun, Feminine, Plural, Nominative, Definite
NuAjFePlAcDf	Adjective Noun, Feminine, Plural, Accusative, Definite
NuAjFePlGeDf	Adjective Noun, Feminine, Plural, Genitive, Definite
NuIsMaSnNmId	Instrument Noun, Masculine, Singular, Nominative, Indefinite
NuIsMaDuGeId	Instrument Noun, Masculine, Dual, Genitive, Indefinite
NuIsMaPlNmId	Instrument Noun, Masculine, Plural, Nominative, Indefinite
NuIsMsSnNmDf	Instrument Noun, Masculine, Singular, Nominative, Definite
NuIsMsSnAcDf	Instrument Noun, Masculine, Singular, Accusative, Definite
NuIsMsSnGeDf	Instrument Noun, Masculine, Singular, Genitive, Definite
NuIsMaDuGeId	Instrument Noun, Masculine, Dual, Genitive, Indefinite
NuIsMaPlNmDf	Instrument Noun, Masculine, Plural, Nominative, Definite
NuIsMaPlAcDf	Instrument Noun, Masculine, Plural, Accusative, Definite
NuIsMaPlNmDf	Instrument Noun, Masculine, Plural, Genitive, Definite
PrPp	Preposition Particle
PrVo	Vocative Particle
PrCo	Conjunction Particle
PrEx	Exception Particle
PrAn	Annulment Particle
PrSb	Subjunctive Particle

Table 2: Sample of Arabic Tagset

Reference:

[1] El-Kareh and Al-Ansary, An Arabic interactive multi-feature pos tagger. *In Proceedings of the, ACIDCA conference*, Monastir, Tunisia, 2000, pp 204- 210.

[2] M. A. Elaraby 2000, A large scale computational processor of the Arabic morphology and application. *(Master's thesis), Cairo University, Egypt.*

[3] Andrew Hardie. Developing a tagset for automated Part-of-speech tagging in Urdu. *Proceedings of the Corpus Linguistics 2003 conference,* Lancaster University, UK, 2003.

[4] J. A. Haywood and H. M. Nahmad. *A new Arabic Grammar: of the written language*, LUND HUMPHRIES , USA, 2005.

[5] Daniel Jurafsky & James H.Martin. Speech and language processing: *An introduction to natural language processing, computational linguistics, and speech recognition.* Prentice-hall, USA., 2000.

[6] S. KHOJA, Apt: Arabic part-of-speech tagger. *Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001),* Carnegie Mellon University, Pittsburgh, Pennsylvania, no. 2, 2001 .

[7] Graside, Khojah and Knowels, A tagset for the morphosyntactic tagging of Arabic. *Paper presented at Corpus Linguistics 2001, Lancaster University, Lancaster, UK, March 2001*, and to appear in a book entitled "A Rainbow of Corpora: Corpus Linguistics and the Languages of the World", edited by Andrew Wilson, Paul Rayson, and Tony McEnery; Lincom-Europa, Munich., 2001.

[8] B Megyesi. Brill's rule-based part of speech tagger for Hungarian. D-level thesis *(Master's thesis) in Computational Linguistics, Stockholm University, Sweden*. 1998.

[9] Leech G, Wilson A 1996 *Recommendations for the Morphosyntactic Annotation of Corpora EAGLES Report.* http://www.ilc.pi.cnr.it/EAGLES96/annotate/

[10] *Transparent Language* http://www.transparent.com/