# Bootstrapping Tagged Islamic Corpora

**Mahmoud Shokrollahi-Far[1], Behrooz Minaei[2], Issa Barzegar[3], Hadi Hossein-Zadeh[4], Mozhdeh Ghasdi[5], and Salman Hoseini[6]**

[1, 3, 4, 5]University College of Nabi-Akram, Rah-Ahan 1283, P.O. Box 51385-1488, Tabriz, Iran
[1]TilburgUniversity, P.O. Box 90153, NL-5000 LE Tilburg , The Netherlands
[2]Iran University of Science & Technology, P.O. Box 16846-13114, Tehran, Iran
[1&2]Noorsoft: Center for Computer Research on Islamic Sciences, Amin Aven., P.O. Box 37185-3857, Qom, Iran
[6]Hojatiyeh School for Islamic Sciences, Qom, Iran
[1]{msf@ucna.ac.ir, msf@noor.net, m.shokrollahifar@uvt.nl}; [2]b_minaei@iust.ac.ir; [3]Issa_Barzegar@yahoo.com;
[4]hosseinzadeh@fannavaran.com; [5]ghasdi@tashilgostar.com; [6]seyyedsalman.hosseini@gmail.com

## Abstract

Among tagged language resources for Arabic there is a high density for Modern Standard Arabic. Nonetheless, the tagged corpora for Classical Arabic are of very low density. Moreover, such corpora are normally developed applying software that are of serious shortcomings. This paper is elaborating on the tagging approach of the Islamic corpora which are being tagged at Noorsoft, Qom, Iran, exploiting Mobin Expert System of Mahmoud Shokrollahi-Far at University College of Nabi-Akram, Tabriz, Iran. The system relying on the traditional grammar of Arabic where there are just three parts-of-speech, after tokenizing the phrasal words, bootstraps the grammatical tags in the corpora employing the vocalism in the vowelized texts. This gives the opportunity for the system to incorporate a tagset which is morpho-syntactically as diverse as possible. The prepared corpora to be tagged, being in a variety of Islamic genre, consist of 1G of phrasal words, whose tagged output is in xml format.

## Introduction

Recent years of Arabic language resources and tools have enjoyed a good range of tagged corpora on Modern Standard Arabic resulting in various researches on the language; performing similar researches on Classical Arabic, however, has always been suffering the lack of such tagged corpora, the only existing example of which is Tagged Quran by University of Haifa (Judith, 2004).

To create a tagged corpus in Arabic the existing software, say that of Buckwalter (2004), are usually utilized. Nevertheless, they are enduring three main shortcomings:

1. Their output is in need of massive time-consuming and expensive manual editing
2. Their tag-sets are poor in diversity of tags
3. They are tuned for Modern Standard Arabic

An innovative shortcut to 'bootstrap' the corpora, overcoming the shortcomings of the existing approaches, seems to be relying on the vocalism of Arabic. Traditionally Arabic 'alphabet' is believed to be a set of 28 consonants, excluding any vowels. This distinction has from many centuries ago been represented in Arabic written texts where the words are normally strings of just consonants. The first attempt to symbolize non-consonantal phonemes in Arabic texts goes back to the 7th century AC when such characters were manifested as 'diacritics' on the consonants to help Holy Quran be 'disambiguated' grammatically by its readers. Hence these vocalic diacritics in Arabic texts may be regarded as 'traditional text annotations' which still stands.

This paper is reporting the research which aims at re-exploring the significance of this Arabic traditional methodology of text annotation through amplifying the efficiency of opting for these annotations to bootstrap very accurately a diverse range of grammatical tags in Classical Arabic texts, for which Mobin Expert System has been utilized, a system which depends intimately on the traditional grammar of Arabic.

## Existing Tagged Corpora on Arabic

A said before, almost the whole tagged corpora for Arabic is on Modern Standard Arabic, mainly developed at LDC, Pennsylvania, and ELDA, Paris. The only formally existing tagged corpus on Classical Arabic is *Tagged Quran* of Haifa.

### Corpora on Modern Standard Arabic

The corpora tagged at LDC and ELDA are usually on the texts of Arabic newspapers published internationally.

**Corpora of LDC**

The oldest corpus reviewed here is *Arabic Newswire Part 1* published by LDC. This project has been done by David Graff and Kevin Walker in 2001. As the website of LDC explains, "the Arabic Newswire Corpus is composed of articles from the Agence France Presse (AFP) Arabic Newswire. The source material was tagged using TIPSTER-style SGML and was transcoded to Unicode (UTF-8). The corpus includes articles from May 13, 1994 to December 20, 2000. The data is in 2,337 compressed (zipped) Arabic text data files. There are 209 Mb of compressed data (869 Mb uncompressed) with approximately 383,872 documents containing 76 million tokens over approximately 666,094 unique words." *Prague Arabic Dependency Treebank 1.0* is another similar corpus released by Jan Hajic, Otakar Smrz, Petr Zemanek, Petr Pajas, Jan Snaidauf, Emanuel Beska, Jakub Kracmar and Kamila Hassanova at LDC: "Prague Arabic Dependency Treebank (PADT) not only consists of multi-level linguistic annotations over the language of Modern Standard Arabic, but even provides a variety of unique software implementations designed for general use in Natural Language Processing (NLP). The PADT project might be summarized as an open-ended activity of the Center for Computational Linguistics, the Institute of Formal and Applied

Linguistics, and the Institute of Comparative Linguistics, Charles University in Prague, resting in multi-level annotation of Arabic language resources in the light of the theory of Functional Generative Description. The project is a younger sibling to Prague Dependency Treebank for Czech, and is maintained upon co-operation with the Linguistic Data Consortium, University of Pennsylvania, who release non-annotated corpora of Arabic newswire and develop an independent Penn Arabic Treebank. The corpus of PADT 1.0 consists of morphologically and analytically annotated newswire texts of Modern Standard Arabic, which originate from the Arabic Gigaword and the plain data of Penn Arabic Treebank, Part 1 and Penn Arabic Treebank, Part 2. The PADT 1.0 distribution comprises over 113,500 tokens of data annotated analytically and provided with the disambiguated morphological information. In addition, the release includes complete annotations of MorphoTrees resulting in more than 148,000 tokens, 49,000 of which have received the analytical processing." *Arabic Gigaword* is another corpus which was first released in 2003 and its third edition has been released in 2007. The project has been done by David Graff at LDC: "Arabic Gigaword Third Edition is a comprehensive archive of newswire text data acquired from Arabic news sources by the LDC at the University of Pennsylvania. Arabic Gigaword Third Edition includes all of the content of Arabic Gigaword Second Edition as well as new data collected after the publication of that edition. Also, an archive from a new newswire source -- Assabah -- has been included in the third edition. The six distinct sources of Arabic newswire represented in the third edition are:

- Agence France Presse (afp_arb)
- Assabah (asb_arb)
- Al Hayat (hyt_arb)
- An Nahar (nhr_arb)
- Ummah Press (umh_arb)
- Xinhua News Agency (xin_arb)

This release contains 547 files, totalling approximately 1.8GB in compressed form (6,673 MB uncompressed) and 1,994,735 K-words. Certain data and formatting issues observed in previous releases of Arabic Gigaword have been normalized in the third edition." The last corpus is *Arabic Treebank: Part 3 v 1.0* which has been released by Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Hubert Jin at LDC in 2004: "this publication is the third part of a corpus of 1,000,000 words of Arabic Treebank, designed to support language research and development of language technology for Modern Standard Arabic. Part one was released in 2003 as Arabic Treebank: Part 1 v 2.0, having the source data extracted from Agence France Press stories. Part two was released in 2004 as Arabic Treebank: Part 2 v 2.0, having the source data extracted from Al-Hayat distributed by Ummah. The current Arabic Treebank: Part 3 v 1.0 corpus consists of stories from An Nahar News Agency. This corpus includes 600 stories from the An Nahar News Text. There are a total of 340,281 words (counting non-Arabic tokens such as numbers and punctuation) in the 600 files - one story per file. New features of annotation include complete vocalization (including case endings), lemma IDs, and more specific POS tags for verbs and particles. The corpus contains 293,035 Arabic-only word tokens (prior to the separation of clitics), of which 290,842 (99.25%) were provided with an acceptable morphological analysis and POS tag by the morphological parser, and 2,193 (0.75%) were items that the morphological parser failed to analyze correctly."

**Corpora of ELDA**

As ElDA's website reports, *Le Monde Diplomatique* "contains 102,960 vowelised, lemmatised and tagged words (58 texts from Le Monde Diplomatique Arabic). 3 files are associated to each text: raw text in Arabic, vowelized text in Arabic, one XML file containing the morphological annotation of the text. Each text word associates a certain number of information, such as word size, rank of the word in the text, paragraph number where the word was found, etc. Each word associates a node in the XML file. Each node contains the following positional features of the word in the text:

- Paragraph number in the text, i.e. paragraph where the word can be found,
- Sentence number in the paragraph,
- Sentence number in the text,
- Rank of the word in the text,
- Rank of the first character of the word in the text,
- Word size."

*NEMLAR Written Corpus* has been produced within the NEMLAR project by ELRA: "the NEMLAR Written Corpus consists of about 500,000 words of Arabic text from 13 different categories, aiming to achieve a well-balanced corpus that offers a representation of the variety in syntactic, semantic and pragmatic features of modern Arabic language. The different categories are:

- Political news: 48,000 words
- Political debate: 30,000 words
- Islamic text (Preaching and others): 29,000 words
- Phrases of common words: 8,500 words
- Text from broadcast news: 5,500 words
- Business: 20,000 words
- Arabic literature: 30,000 words
- General news: 100,000 words
- Interviews: 56,000 words
- Scientific press: 50,000 words
- Sports press: 50,000 words
- Dictionary entries explanation: 52,000 words
- Legal domain text: 21,000 words

The corpus is provided in 4 different versions:

- Raw text
- Fully vowelized text
- Text with Arabic lexical analysis
- Text with Arabic POS-tags

Diacritics, lexical analysis and POS-tags were generated by RDI's tool Fassieh©. The accuracy of the automatic analysis is around 95%. To reach about the 99% accuracy rate as defined for this corpus, the linguists used the visual revision mode of Fassieh© where the linguist has to either approve the 1st most likely analysis (most of the time) or select another one manually (in the 4% minority of the cases)."

## *Tagged Quran* of Haifa

The work which is more comparable with those reported in this paper is Morphological Tagging of the Qur'an

(Judith, 2004), which has been done in Haifa University by Judith Dror, Dudu Shaharabani, Rafi Talmon and Shuly Wintner: “we present a computational system for morphological tagging of the Qur’an, for research and teaching purposes.The system facilitates a variety of queries on the Qur’anic text that make reference not only to the words but also to their linguistic attributes. The core of the system is a set of finite-state based rules which describe the morpho-phonological and morpho-syntactic processes of the Qur'anic language. Using a finite-state toolbox we apply the rules to the Qur'anic text and obtain full morphological tagging of its words. The results of the analysis are stored in an efficient database and are accessed through a graphical user interface which facilitates the presentation of complex queries. The system is currently being used for teaching and research purposes. The results of the analysis are stored in a database in a form that encodes, for each analyzed word, its morphological features and their values. For example, an analysis such as:

swr+fu&lat+Noun+Triptotic+Fem+Sg+Nom

is converted to a record structure of the form:

| | |
|---|---|
| root | swr |
| pattern | fu&lat |
| part of speech | Noun |
| case marking | triptotic |
| gender | Fem |
| number | Sg |
| case | Nom” |

## The Tagset

Since the aimed corpora of the present research are being tagged at word level, it is premier to define ‘word’ in Arabic, which has been characterized as a part inside Arabic phrasal word in this approach.

### Arabic Phrasal Word

When processing texts of most languages, say English, a word is considered to comprise one or a sequence of few concatenated morphemes mapped, in a given co-text, with only one part-of-speech. In these texts words are usually delimited by spaces, which results in ease of word recognition. Moreover, in these languages consecutive words are grouped together to construct a syntactic constituent known as a ‘phrase’ whose grouping decision normally leads to the ‘parsing problem’. In Arabic texts, however, space delimiter usually discriminates a sequence of consecutive concatenated morphemes each of which has an independent POS which includes, in Arabic, ‘noun’, ‘verb’, and ‘particle’. This sequence can be referred to as a phrase where the problem of parsing is frequently reduced to the problem of ‘segmentation’. Therefore, not only the base of such phrasal string, but even its prefix and suffix, as bound morphemes always attached to the base, are words on their own. Whereas the base word may be any of the three parts of speech, the prefix or suffix word, if any, can never be a verb.

On this view, the general structure of an Arabic phrasal word which is tokenized in the tagged corpora is as follows,

*(prefix-word(s))base-word(suffix-word(s))*

Whose components has been elaborated in Table 1.

### Attributes and Values in the Tagset

The Arabic phrasal word categorized in Table 1 consists of one or more actual words each of which possesses different grammatical features which are listed as attributes in Table 2 where each attribute is mapped with some of the potential values it may take for each token in the Arabic phrasal word. Table 3, Table 4 and Table 5 are the complete list of the attributes and values in Table 2, which are acronymed in the outputted tagged corpora of this research.

## Applying Mobin Expert System

The tagset is applied to tune the tagger of Mobin Expert System whose knowledge-base comprises Arabic morpho-syntactic rules on Arabic traditional grammar which is highly dependent on morphological templatic patterns of vowels and consonants of three parts-of-speech of verb, noun and particle at phrasal level.

Mobin, recently developed by Mahmoud Shokrollahi-Far and his fellow researchers at University College of Nabi-Akram, Tabriz, Iran, outputs an XML file where the vocalism of the vowelized inputted Arabic text leads it to tokenize the phrasal words and generate the tags predicted in the tagset. The system has been evaluated on Nobi (Shokrollahi-Far, Hoseini & Isazadeh, 2007), and its very high F-score was preliminarily reported in (Shokrollahi-Far *et al*, 2007). The system is going to be elaborated on very soon in a journal paper.

## The Tagged Islamic Corpora

The target input for Mobin is the Classical Arabic corpora being developed on Islamic texts at Noorsoft, Qom, Iran. These corpora include Quran, its commentaries, Islamic Hadith, Islamic theology, Islamic philosophy, Arabic dictionaries, Islamic Mythology, Arabic literature, Bibliographies, Islamic history, and geography of Islamic lands, whose present size is exceeding 1G phrasal words. This produces a huge size of tagged corpora. Quran with nearly 100K phrasal words, for instance, whose input plain text file size is 746KB gets the output tagged xml file size of 31.5MB. A sample output of Mobin on just two phrasal words of Quran is as follows:

1. a verb

```
- <Phr Ent="فَسَيَكْفِيكَهُمُ">
- <Word Ent="فَ" Afix="Pref" POS="Particle">
  <PAtrib Decl="NDc" Spc="NVCom" Opr="Nut"
Sem="Adt" />
  </Word>
- <Word Ent="سَ" Afix="Pref" POS="Particle">
  <PAtrib Decl="NDc" Spc="VSp" Opr="Nut" Sem="Fut"
/>
  </Word>
- <Word Ent="يَكْفِي" Afix="Base" POS="Verb">
  <VAtrib Decl="Dc" Alt="Nag" Aspect="ImPrf"
Mod="Moz" Time="Futr" Genr="Genrt" Prsn="NAdrse"
Num="Sing" Gend="Masc" Case="Jusv" Voic="Act" />
- <Morph>
  <Pref Ent="يَ" Sem="InflP" />
  <Stem Ent="كْفِي" Struct="Simp3" Temp="yaFAaLu"
Root="ك+ف+ي|و" />
  </Morph>
  </Word>
- <Word Ent="كَ" Afix="Suf" POS="Noun">
```

```
  <NAtrib Decl="NDc" DeclT="Pron" Alt="Nul"
Prsn="Adrse" Num="Sing" Gend="Masc"
Genr="NGenrt" Def="Deft" Case="Acus" Derv="Nul"
Func="" />
  </Word>
- <Word Ent="هُمْ" Afix="Suf" POS="Noun">
 <NAtrib Decl="NDc" DeclT="Pron" Alt="Nul"
Prsn="NAdrse" Num="Plr" Gend="Masc"
Genr="NGenrt" Def="Deft" Case="Acus" Derv="Nul"
Func="" />
  </Word> </Phr>
```

2. a noun

```
- <Phr Ent="فَبِالْبَاطِلِ">
- <Word Ent="فَ" Afix="Pref" POS="Particle">
  <PAtrib Decl="NDc" Spc="NVCom" Opr="Nut"
Sem="Adt" />
- <Word Ent="بِ" Afix="Pref" POS="Particle">
  <PAtrib Decl="NDc" Spc="NSp" Opr="Genty"
Sem="GentP-Rel" />
  </Word>
- <Word Ent="الْ" Afix="Pref" POS="Noun">
  <NAtrib Decl="NDc" DeclT="Nut" Alt="Nul"
Prsn="Nul" Num="Sing" Gend="Nut" Genr="NGenrt"
Def="Deft" Case="Gent" Derv="Nul" Func="" />
  </Word>
- <Word Ent="بَاطِل" Afix="Base" POS="Noun">
  <NAtrib Decl="Dc" DeclT="Nut" Alt="Slim"
Prsn="Nul" Num="Sing" Gend="Masc" Genr="Genrt"
Def="Deft" Case="Gent" Derv="Dervd" DervT="AgntN"
Func="" />
- <Morph>
  <Stem Ent="بَاطِل" Struct="Comp3" Temp="Faa:AiL"
Root="ب+ط+ل" SEP="١" />
  <Suf Ent="ِ" Sem="Vowel" />
  </Morph>
  </Word> </Phr>
```

| branch | Sub branch | Part-Of-Speech | Morphemes | Morphemes Sub branch |
|---|---|---|---|---|
| ✓ Phrase | ⊕ Prefix Word | ⊕ Particle$_1$ | - | - |
| | | ⊕ Particle$_2$ | - | - |
| | | ⊕ Noun | - | - |
| | ✓ Base word | ○ Verb | ⊕ Prefix | - |
| | | | ⊕ Stem | - |
| | | | ⊕ Suffix | ⊕ S$_1$ |
| | | | | ⊕ S$_2$ |
| | | | | ⊕ S$_3$ |
| | | ○ Particle | - | - |
| | | ○ Noun | ⊕ Stem | - |
| | | | ⊕ Suffix | - |
| | ⊕ Suffix Word | ⊕ Particle | - | - |
| | | ⊕ Noun$_1$ | - | - |
| | | ⊕ Noun$_2$ | - | - |

Table 1: The structure of Arabic phrasal Word

(The bullets in the Table:

✓Means that it will certainly be matched in the phrasal word

○ Means that just one of them may be matched

⊕ Means that it’s match is possible)

| Part-Of-Speech | Attributes | Sample Values |
|---|---|---|
| Verb | Declination | Non-Declined ׀ Declined |
| | Alternation | Slim ׀ Mah ׀ Muz ׀ Ajvaf ׀ Nag ׀ Laf Ebd , … |
| | Aspect | Perfect ׀ Imperfect |
| | Mood | Imperative ׀ Nah ׀ Moz ׀ Maz ׀ Mok ׀ |
| | Time | Past ׀ Future ׀ Present |
| | Generation | Generative ׀ Non-Generative |
| | Person | Non-Addressee ׀ Addressee ׀ Speaker |
| | Number | Singular ׀ Dual ׀ Plural |
| | Gender | Masculine ׀ Feminine ׀ Neutral |
| | Case | Indicative ׀ Subjunctive ׀ Jussive |
| | Voice | Active ׀ Passive |

| | | |
|---|---|---|
| Particle | Declination | Non-Declined |
| | Specific | Noun Specific l Verb Specific l Noun & Verb Common |
| | Operation | Accusativity l Genitivity l Jussitivally l Neutral |
| | Semantic | Genitive Particle l Semi-Verb l Exclusion l Future l Negativel… |
| Noun | Declination | Non-Declined l Declined |
| | Declination Type | Pronoun l Proper Noun l Neutral |
| | Alternation | Slim l Mah l Muz l Ajvaf l Nag l Laf Ebd , |
| | Person | Non-Addressee l Addressee l Speaker l Nul |
| | Number | Singular l Dual l Plural |
| | Gender | Masculine l Feminine l Neutral |
| | Generation | Generative l Non-Generative |
| | Definiteness | Definite l In Definite |
| | Case | Nominative l Accusative l Genitive |
| | Derivation | Non-Derivative l Derivative l Nul |
| | Derivation Type | Agent Noun l Patient Noun l Instrument Noun l Time Nounl … |
| | Function | Subject l Agent l Patient l Appended l Addressed l Excluded l Add |

Table 2: The grammatical attributes and values that each token may take in Arabic phrasal word

| Row | Attributes | Acronyms | Row | Attributes | Acronyms |
|---|---|---|---|---|---|
| 1 | Affixation | Afix | 13 | Gender | Gend |
| 2 | Alternation | Alt | 14 | Generation | Genr |
| 3 | Aspect | Aspect | 15 | Mood | Mod |
| 4 | Case | Case | 16 | Number | Num |
| 5 | Declination | Decl | 17 | Operation | Opr |
| 6 | Declination | Decl | 18 | Person | Prsn |
| 7 | Declination Type | DeclT | 19 | Semantic | Sem |
| 8 | Definiteness | Def | 20 | Specific | Spc |
| 9 | Derivation | Derv | 21 | Stem Extra Particle | SEP |
| 10 | Derivation Type | DervT | 22 | Structure | struct |
| 11 | Entry | Ent | 23 | Time | Time |
| 12 | Function | Func | 24 | Voice | Voic |

Table 3: the acronyms of the attributes

| Row | Attributes | Acronyms | Row | Attributes | Acronyms |
|---|---|---|---|---|---|
| 1 | Particle Attributes | PAtrib | 6 | Morpheme | Morph |
| 2 | Phrase | Phr | 7 | Prefix | Pref |
| 3 | Word | Word | 8 | Suffix | Suf |
| 4 | Noun Attribute | NAtrib | 9 | Stem | Stem |
| 5 | Verb Attributes | VAtrib | | | |

Table 4: the acronyms of the sub-attributes

| Row | values | Acronyms | Row | values | Acronyms |
|---|---|---|---|---|---|
| 1 | Accept | Acc | 51 | Jussitivally | Jusy |
| 2 | Active | Act | 52 | Masculine | Masc |
| 3 | Accusative | Acus | 53 | Non-Addressed | NAdrs |
| 4 | Accusative Particle | AcusP | 54 | Non-Declined | NDc |
| 5 | Accusativity Particle | Acusy | 55 | Non-Derivative | NDervd |
| 6 | Added | Add | 56 | Negative | Neg |
| 7 | Address | Adrs | 57 | Non-Generative | NGenr |

| Row | values | Acronyms | Row | values | Acronyms |
|---|---|---|---|---|---|
| 8 | Addressed | Adrsd | 58 | Nominative | Nomn |
| 9 | Addressee | Adrse | 59 | Noun-Specific | NSp |
| 10 | Addition | Adt | 60 | Null | Nul |
| 11 | Agent | Agnt | 61 | Neutral | Nut |
| 12 | Agent Noun | AgntN | 62 | Neutral Particle | NutP |
| 13 | Alternation | Alt | 63 | Noun/Verb Common Particle | NVCom |
| 14 | Answer | Ans | 64 | Past | Past |
| 15 | Appended | Apnd | 65 | Patient | Pat |
| 16 | Approximation | Aprox | 66 | Patient Noun | PatN |
| 17 | Attribute | Atrib | 67 | Place Noun | PlcN |
| 18 | Base | Base | 68 | Plural | Plr |
| 19 | Case | Case | 69 | Part-of-Speech | POS |
| 20 | Chapter | Chap | 70 | Premonition | Prem |
| 21 | Comparison Noun | CmpN | 71 | Perfect | Prf |
| 22 | Consonant | $C_n$ | 72 | Person | Prsn |
| 23 | Compound Trilateral | Comp3 | 73 | Present | Prst |
| 24 | Compound Quadrilateral | Comp4 | 74 | Passive | Psi |
| 25 | Compound Pentaliteral | Comp5 | 75 | Question | Ques |
| 26 | Condition | Cond | 76 | Reject | Rej |
| 27 | Declined | Dc | 77 | Relative | Rel |
| 28 | Definite | Deft | 78 | Reason | Res |
| 29 | Dual | Dual | 79 | Restraining | Restr |
| 30 | Excitation | Excit | 80 | Root | Root |
| 31 | Exaggeration Noun | ExgN | 81 | Semantic | Sem |
| 32 | Excluded | Exld | 82 | Semi-Verb | SemiV |
| 33 | Exclusion | Exls | 83 | Silence | Sile |
| 34 | Feminine | Fem | 84 | Similar Adjective | SimAdj |
| 35 | Future | Fut | 85 | Simple Trilateral | Simp3 |
| 36 | Gender | Gend | 86 | Simple Quadrilateral | Simp4 |
| 37 | Generative | Genrt | 87 | Simple Pentaliteral | Simp5 |
| 38 | Genitive | Gent | 88 | Singular | Sing |
| 39 | Genitive Particle | GentP | 89 | Specific | Spc |
| 40 | Genitivity Particle | Genty | 90 | Speaker | Spkr |
| 41 | Imperative | Imper | 91 | Structure | Struct |
| 42 | Imperfect | ImPrf | 92 | Subject | Subj |
| 43 | Indefinite | InDeft | 93 | Subjunctive | Subjn |
| 44 | Indicative | Indv | 94 | Template | Temp |
| 45 | Infinitive | Inf | 95 | Time | Time |
| 46 | Inflection Particle | InflP | 96 | Time Noun | TimN |
| 47 | Instrument Noun | InstN | 97 | Verse | Vers |
| 48 | Interpretation | Int | 98 | Voice | Voic |
| 49 | Jussive Particle | JusP | 99 | Verb-Specific | VSp |
| 50 | Jussive | Juss | 100 | | |

Table 5: the acronyms of the values of the attributes

## Acknowledgements

This research has been supported by a grant from High Council of Information in Iran.

## Bibliographical References

Buckwalter, T. (2004). Buckwalter Arabic Morphological Analyzer Version 2. Linguistic Data Consortium.

Judith, D., Shaharabani, D., Talmon, R. and Wintner, S. (2004). Morphological Analysis of the Qur'an. Literary and Linguistic Computing. 19(4).

Shokrollahi-Far, M. & Saraei, J. (2007). An Enhanced Knowledge-Based Tagger for Arabic Verbs. At Computational Linguistics in the Netherlands (CLIN 18). Nijmegen, the Netherlands.

Shokrollahi-Far, M (2007). Developing a Regular Grammar for Arabic Morphology. At NATO Advanced Studies Institute on Advances in Language Engineering for Low- and Middle-Density Languages. Batumi, Georgia.

Shokrollahi-Far, M, Isazadeh, A., Barzegar, I. & Soltani, R (2007). Knowledge-Base on Holy Qur’an for Tagging Arabic Verbs. Proceedings of DCCA2. Jordan.

Shokrollahi-Far, M, Hoseini, S.S., Isazadeh, A. (2007). Nobi: A Manually Tagged Corpus of Qur’an. CLIN 17. Leuven, Belgium.