STANDARD ARABIC VERBS INFLECTIONS USING NOOJ PLATFORM

Mohammed Mourchid [1] Ilham Blanchete [2] and Abdelaziz Mouloudi[3]

[1] MIC search team, Laboratory MISC, Ibn Tofail University Kenitra- Morocco
[2]Department of Computer Science FSK, Ibn Tofail University, Kenitra, Morocco
[3]MIC search team, Laboratory MISC, Ibn Tofail University Kenitra- Morocco

ABSTRACT

This article describes the morphological analysis of a standard Arabic natural language processing, as a part of an electronic dictionary-constricting phase. A fully 3-lettered inflected verbs model are formalized based on a linguistic classification, using NOOJ platform, the classification gives certain representative verbs that will considered as lemmas, this verbs form our dictionary entries, they are also conjugated according to our inflection paradigm relying on certain specific morphological properties. This dictionary will be considered as an Arabic resource, which will help NLP applications and NOOJ platform to analyse sophisticated Arabic corpora.

KEYWORDS

Morphological analysis, NOOJ, ANLP & Arabic verb inflections

1. INTRODUCTION

The Arabic natural language applications need a fully and automatic Arabic dictionary to analyse the sophisticated corpora, as a first phase of building this dictionary we started by formalizing the trilateral verbs based on a linguistic verbs classification [6]. The linguistic analysis must go through a first step of lexical and morphological analysis, which consists in testing membership of each word of the text to the Arabic vocabulary [1] we started from a basic kind of verbs, which are called trilateral, verbs that contain three letters. Using a specific linguistic classification of these verbs we guarantee that we are going to cover all Arabic trilateral verbs [7], this verbs will also attached to their inflectional paradigms to cover all conjugated forms, in this paper we give examples of our implemented dictionary and grammars in NooJ platform as figures.

2. DEFINITIONS

2.1. Nooj Platform

NooJ is a linguistic developmental environment, which can analyze texts of several million words in real time. It includes tools to construct, test and maintain large coverage of lexical resources,

DOI: 10.5121/ijnlc.2017.6103

as well as morphological and syntactic grammars. Dictionaries and grammars are applied to texts in order to locate morphological, lexicological and syntactic patterns, remove ambiguities, and tag simple and compound words [5]. NooJ platform works on cascade model; the result of each analysis step is the input of the next one. For more information please consult the official NooJ website. We adopted this platform because it allows us to:

- Implement all linguistic analysis phases: morphological, syntactical and semantic analysis.
- Create our own corpora and apply search option using special queries.
- To implement our grammars and dictionary using its linguistic engine.
- To analyse our text by giving morphological, syntactical and semantic properties of each word/sentence.

2.2.Nooj Architecture

NooJ platform is Programmed using C#/.net Framework. NooJ follows a component-based software approach, which is a step beyond the object oriented programming paradigm . The system consists of three modules, corpus handling, lexicon and grammar development that are integrated into a single intuitive graphical user interface (command line operation is also available). NooJ processes texts and corpora (i.e. sets of text files) at the Orthographical, Lexical, Morphological, Syntactic and Semantic levels. All linguistic information (at any level) is represented by annotations that are stored in the Text Annotation Structure (TAS)[5]. We use this platform to formalize the Arabic 3-lettered Verbs model as a first step of Arabic dictionary constructing phase; starting by building our dictionary that contains the previous verb category and linking it with the our productive grammars that give all inflectional forms for each dictionary entry, we will detail this in next sections.

2.3. Natural language

Is a human spoken and/or written languages like Arabic, French, and English.

2.4. Natural Language Processing

Is a subfield of Artificial Intelligence and linguistic, devoted to make computers understand the Statements or words written in human languages. A natural language also known as a spoken or written language by people for general-purpose communication [2].

2.5. Arabic Natural Language

Arabic is a Semitic language spoken by more than 330 million people as a native language, in an area extending from the Arabian/Persian Gulf in the East to the Atlantic Ocean in the West. Arabic is a highly structured and derivational language where morphology plays a very important role [2]. Morphology is central in working on Arabic NLP because of its important interactions with both orthography and syntax. Arabic’s rich morphology is perhaps the most studied and written about aspect of Arabic. As a result, there is a wealth of terminology, some of it inconsistent that may intimidate and confuse new researchers [3].

2.6. Arabic Natural Language Processing

Over the last few years, Arabic natural language processing (ANLP) has gained increasing importance, and several state-of-the-art systems have been developed for a wide range of applications, including machine translation, information retrieval and extraction, speech synthesis and recognition, localization and multilingual information retrieval systems, text to speech, and tutoring systems. These applications had to deal with several complex problems pertinent to the nature and structure of the Arabic language. Most ANLP systems developed in the Western world focus on tools to enable non-Arabic speakers make sense of Arabic texts. Since understanding Arabic language becomes a point of interest for non Arabic speakers, funding became available for companies and research centers to develop tools such as named entity recognition, machine translation, especially spoken machine translation, document categorization, etc [4].

3. NLP STEPS

There are 3 phases involved in natural language processing: Morphological Analysis, Syntactic Analysis and Semantic Analysis. The first step will be detailed in section 6. , we will define briefly the other steps.

3.1. Syntactic Analysis

This involves analysation of the words in a sentence to depict the grammatical structure of the sentence. The words are transformed into structure that shows how the words are related to each other Eg. “The girl the go to the school”. This would definitely be rejected by the English syntactic analyzer [2].

3.2. Semantic Analysis

This abstracts the dictionary meaning or the exact meaning from context. The structures, which are created by the syntactic analyser, are assigned meaning. There is a mapping between the syntactic structures and the objects in task domain. Eg. “Colorless blue idea”. The analyser would reject this as colorless blue do not make any sense together [2]. NooJ helps us to implement this steps, using their linguistic engine to build our dictionary and grammars rules - as it will explained in next sections - that gives the inflectional forms for each dictionary,

4. 3-LETTEREDVERBS

Most Arabic words are derived from three-letter Verbs or 3- lettered verbs, we started by formalizing this kind of verbs standing on their linguistic classification as Figure 1 shows; this figure will be detailed in section 5.

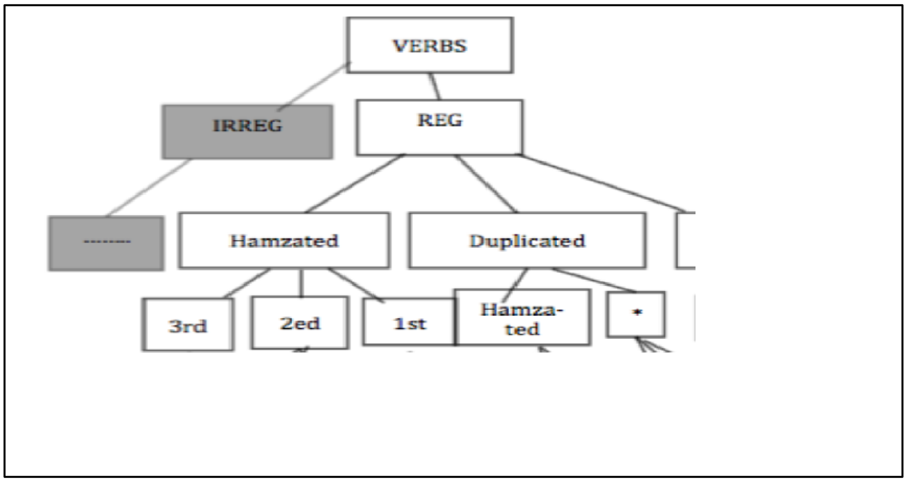


Figure 1. Arabic 3-lettered verbs classification, regular verbs part

As each Arabic verb has its morphological prosperities like root and pattern [8], we attached each verb/dictionary entry with this properties, and with their conjugation form, for instance the verb (to write - كَتَبَ - KaTaBa) takes ك ت ب as roots letters (in Arabic the root letters are separated فَعَلَ faEala as pattern and يَفْعُلُ yafEalo as conjugation form . The conjugation form of the previous verb: كَتَبَ KaTaBa is يَكْتُبُ yaKToBo according to the matching process between the pattern and the conjugational form (فَعَلَ faEala and يَفْعُلُ yafEalo), There are 3 types of the 3-lettered verbs patterns in Arabic they are distinguished according to the second letter diacritic:

(fatha ◌َ kasra ◌ِ , dama ◌ُ).

فَعَلَ– faEala is a pattern of the verb: (to write كَتَبَ) .

فَعُلَ faEula is a pattern of the verb : (to grow – كَبُرَ).

عِلَفَ –faEila is a pattern of :(to play- لَعِبَ).

May takes three conjugation forms as table 1 shows.

Table 1. All possible patterns with their conjugation forms in 3-letter Arabic verbs

pattern	Conjugation form	Conjugation form	Conjugation form
faEala عَلَفَ	yafEalo يَفْعُلُ	yafEilo يَفْعِلُ	yafEolo يَفْعُلُ
faEula فَعُلَ	yafEalo يَفْعُلُ	yafEilo يَفْعِلُ	yafEolo يَفْعُلُ
faEila فَعِلَ	yafEalo يَفْعُلُ	yafEilo يَفْعِلُ	yafEolo يَفْعُلُ

For instance: the verb (to write - kataba - كتب) is the result of matching its root with its pattern, by switching root letters with patterns one; without any changing on the pattern diacritics : كتب → (كتـــب X فَعَلَ), as their conjugation form is(yafEolo يَفْعُلُ) it takes this model : (يَكْتُبُ X كتب) to be conjugated.

5. Arabic 3-Letterd Verbs Classification:

The following classification covers all 3-lettered verbs in standard Arabic language, each representative verbs may takes the three previous patterns (faEala فَعَلَ / faEula فَعُلَ/ faEila فَعِلَ) , and each pattern may take 3 conjugation forms (yafEolo يَفْعُلُ/ yafEalo يَفْعَلُ / yafEilo يَفْعِلُ) ,for example: The path in the following classification as shown in figure 1 : [verb] [regular verbs] [*]: represents the verbs that their last letter neither a (n)ن nor a (t ت) character , and it takes all this representative verbs:

(فَتَحَ -FaTaHa-to open) as : (faEala فَعَلَ - yafEalo يَفْعَلُ).
(كَتَبَ- KaTaBa-to write) as : (faEala فَعَلَ - yafEolo يَفْعُلُ).
(جَلَسَ- JaLaSa- to set) as : (faEala فَعَلَ - yafEilo يَفْعِلُ).
(كَبُرَ-KaBoRa- to grow) as : (faEula فَعُلَ- yafEolo يَفْعُلُ).
(عَلِمَ-AaLiMa- to know) as : (faEila فَعِلَ - yafEalo يَفْعَلُ).

Figure 1 gives a part of the Arabic 3-lettered verbs classification; here are the definitions of some used abbreviation.

REG verbs: regular verbs (afeal sahiha –الأَفْعَال الصَحِيْحَة) contains three verbs kind [Hamzated verbs – duplicated verbs – salim verbs].
Hamzated verbs (mahouza – المَهْمُوز) verbs that contain theأ hamza character

Duplicated verbs (modaEafa – مضاعفة) verbs that contain a duplicated character.

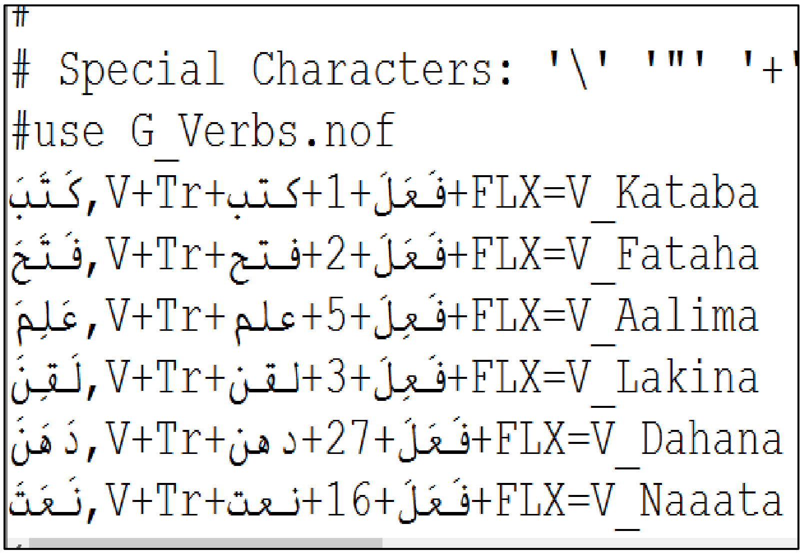
Salim verbs: (salim سالم) verbs that are neither hamzated nor duplicated.
* : verbs that their last letter neither a (n)ن nor a (t ت) character , n : last character is ن , t : last character is ت.

1st,2nd,3rd: the first and second and third character in the verb.

IRRG : verbs that contains one of the Arabic long characters : alif الف , yae ياء , waw واو or the hamza character. In this paper I am going to present only the regular verbs.
Each representative verb considered as dictionary entry that will be assigned to a unique inflectional paradigm, only and only if the verb accept to be conjugated in standard Arabic, then all verbs that are conjugate in the same manner, will take the same inflectional paradigm, tow verbs are conjugated with the same manner or with the same inflectional paradigm if they have the same conjugation form.

6. Morphological Analysis

The lexicon of a language is its vocabulary that includes its words and expressions, while morphological Analysis involves dividing a text into paragraphs, words and the sentences and its main role is to represent the Atomic Language Unit (ALUs) which is the smallest elements that make up the sentence, we are going to define these ALUs/3-lettered verbs as dictionary entries that represent the language vocabulary these entries are associated with their morphological properties which enrich it with linguistic information like: s means singular, p means plural , as basic properties while we add our specific verb morphological properties like:(Root/Pattern/Category Numb),that will be used in advanced analysis phases ,Figure 2 shows our constructed dictionary that calls at first our inflectional grammars G_Verbs ,as it is shown in figure 3, the dictionary contains the language vocabulary with their special morphological properties: V: verb, Tr : transitive verb , 1: verb category which determines verbs conjugation form(يَفْعُلُ - yafEolo), pattern (فَعَلَ : faEala).



```
#
# Special Characters: '\' '"' '+'
#use G_Verbs.nof
كَتَبَ,V+Tr+كتب+1+فَعَلَ+FLX=V_Kataba
فَتَحَ,V+Tr+فتح+2+فَعَلَ+FLX=V_Fataha
عَلِمَ,V+Tr+علم+5+فَعِلَ+FLX=V_Aalima
لَقِنَ,V+Tr+لقن+3+فَعِلَ+FLX=V_Lakina
دَهَنَ,V+Tr+دهن+27+فَعَلَ+FLX=V_Dahana
نَعَتَ,V+Tr+نعت+16+فَعَلَ+FLX=V_Naaata
```

Figure 2. Our constructed dictionary

The FLX paradigm represents all inflectional forms in active and passive voice for each dictionary entry, this FLXs are represented using our defined rules graphs as Figure 3 shows, we preferred to describe this inflectional paradigms using NooJ’s graphical rules interface that is equivalent to the textual rule editor, here is our inflectional paradigms that are assigned to dictionary, each dictionary entry has an inflectional paradigm, that generate all its inflectional and derivational forms.

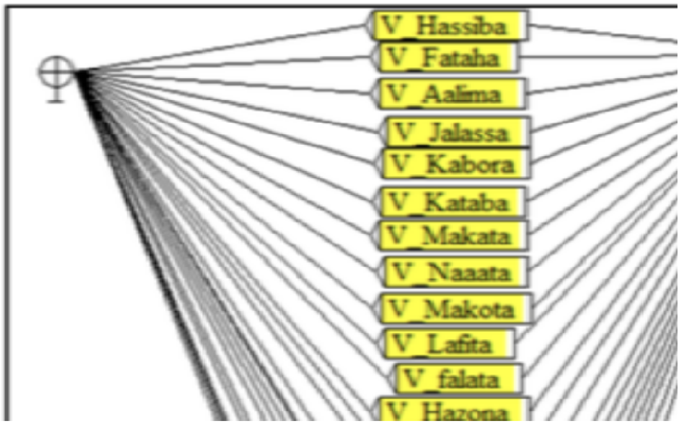


Figure 3: Verbs inflectional paradigm

For instance the lexical entry (to write – كتب kataba) has an inflectional paradigms: FLX=V_Kataba that matches any form in the set of {they write كتبوا KaTaBo, he writes كتب KaTaBa, we write نكتب NaKToBo, you wrote كَتَبْتَ

KaTaBTa,..} NooJ recognizes all this forms even if they are semi or non- vowelized. Each verb is conjugated in 12 deferent tenses as it is shown in Figure 4. This figure shows all possible verb inflections in both of active and passive voice. ACC_Kataba presents the past tens, we used the pre-defined NOOJ operators to define the inflections of this entry here is some special operators <Z>,<T> AND <M> that are defined only for Arabic language for more information please read NOOJ manual.

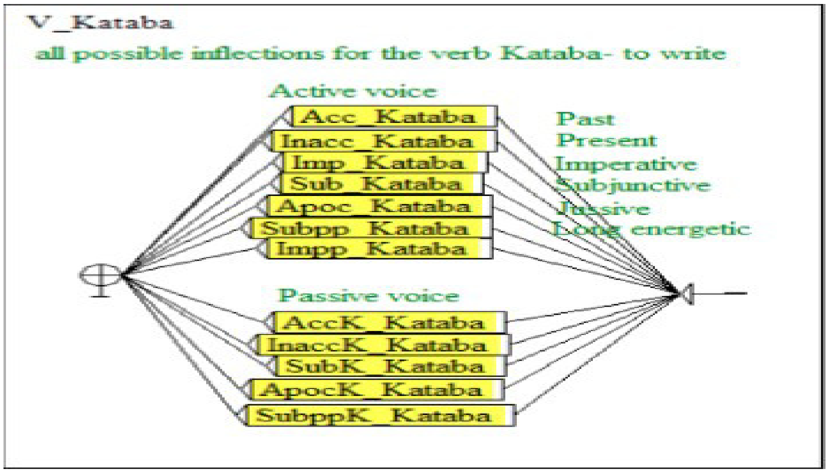


Figure 4: all possible verb inflections

Figure 5 shows how we easily define the past tens for the verb (to write) using NOOJ predefined operators <B> : will erase the last character, <Z> will add ت character while the last character of the given verb is not a ت, else it will add تُّ , chadda appearance refers to a duplication for instance : the conjugated form of the verb (to write) for the first person is I Wrote كَتَبْتُ so the operator <Z> adds ت character while the last character of the given verb is not a ت else it will add تُّ for all verbs that finished with ت like ثبت will converted to ثَبَتُّ while its last character is ت . As its mentioned above this graph recognizes all appearances of the past conjugation set{ I wrote , you wrote ,...}, and annotate them with the morphological defined properties which appear under each node for instance { A+I+1+s } : the first singular person in active voice . A: active voice, I: past / 1: first person and s : singular .

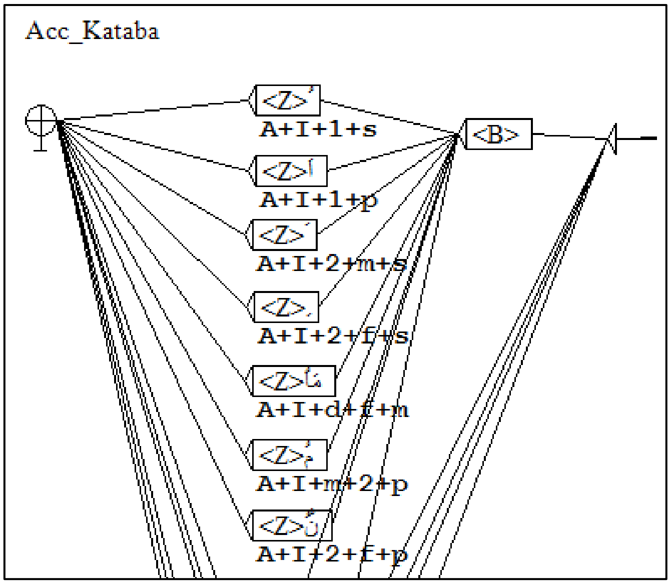
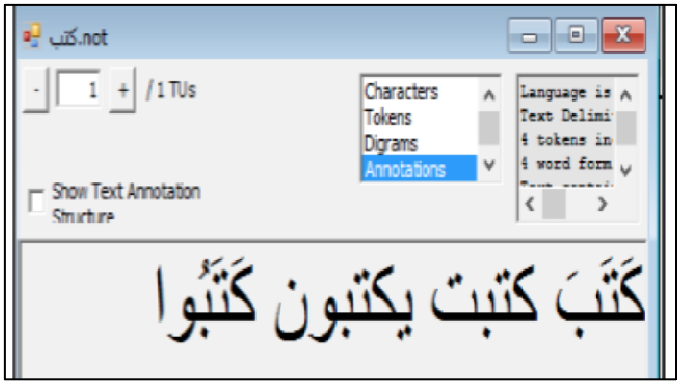


Figure 5: inflectional graph in active voice

7. ANNOTATIONS

Once our dictionary is compiled, we can use it as resource to analyze any Arabic text, that contains only 3-letterad verbs or their inflectional forms, when parsing a text or a corpus, NooJ builds a Text Annotation Structure (TAS) in which each linguistic unit is represented by a corresponding annotation. An annotation stores the position in the text of the text unit to be Represented, its length, and linguistic information, (Silberztein 2007). NooJ adds annotations to the TAS automatically at various stages of the analyses phase, morphological and syntactic parsers provide tools to add, remove and export annotations to TAS in morphological level, and the morphological parser typically applies dictionaries to the text to produce annotations. When the parser recognizes any lemma or inflectional/derivational form in the text it produces the corresponding morphological annotations according to the morphological properties that we have assigned before, the figure 6 shows our text that will be morphologically analyzed on NOOJ platform, it contains different conjugated forms of the verb (to write) vowelized , semi or non vowelized .



Language is
Text Delimi
4 tokens in
4 word form

كَتَبَ كتبت يكتبون كَتَبُوا

Figure 6: text to be analyzed

The vowelized ALU like كَتَبَ KaTaBa he writes will take an expected annotation as it refers to [the third person, masculine singular], but in case of the non vowelized ALU, Nooj Table Annotation Structure gives all possible annotations ,this annotations triggered when the parser matches any inflectional form of a dictionary lemma that we are defined in the inflectional paradigm with the text ALU's, for instance the ALU كتبـــت takes all possible annotations, as it is non- vowelized and the annotations of this inflectional forms {كَتَبَ,كتبت,يَكْتُبُونَ,كَتَبُوا .....}Will be triggered, as they are defined in the inflectional paradigm V_Kataba . The first annotation form is dedicated for the fully diarized verb كتب , the annotation shows that this entry is a verb V , and it is a transitive verb Tr, gives their root كبــت , and their pattern is فعل , NOOJ TAS gives all this properties as Representative verbs that are considered as dictionary entries, all verbs that have the same conjugation form will be assigned to the same inflectional paradigm. Once the dictionary is compiled we can easily use it as a linguistic recourse to analyze the sophisticated corpora.

8. CONCLUSION AND PRESPECTIVES

Using the given 3-lettered linguistic classification, we constructed fully inflected verbal Arabic recourses of the previous verbs category, using NOOJ platform.□Lemma –based verbs are used as dictionary entries; an inflectional paradigm may be assigned to a dictionary entry that gives all possible conjugated forms of the entry. The dictionary contains representative verbs for each leaf of the given linguistic classification tree, each leaf has at most 9 representative verbs that are considered as dictionary entries, all verbs that have the same conjugation form will be assigned to the same inflectional paradigm. Once the dictionary is compiled we can easily use it as a linguistic recourse to analyze the sophisticated corpora. The perspective opened over this work is to extend our dictionary to the inflections of the rest verbs categories, nouns, adjectives also to add some morphological grammars in order to generate broken plural, ALU's with affixes and other sophisticated morphological phenomena like ibdal and ielal and idgham .

REFERENCES

[1] D. Revuz ,(1991)"Dictionnaires et lexiques",Thesis, Paris 7 University. : methodes et algorithmes (Doctoral dissertation).

[2] A. Chopra, A. Prashar and C. Sain, "Natural Language Processing", the International journal of technology enhancement and emerging engineering research, vol. 1, issue 4, ISSN 2

[3] N. Habash, (2010) Introduction to Arabic Natural Language Processing, Morgan & Claypool Publishers series.

[4] K. Shaalan, A. Farghaly, (2009) " Arabic Natural Language Processing: Challenges and Solutions", ACM Transactions on Asian Language Information Processing (TALIP) 8.4,14.

[5] S. Max,"Nooj Manual" , www.nooj-association.org.

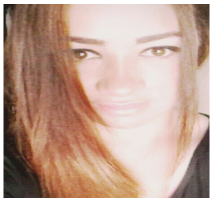
[6] M.El-Ghalayani,(2004) "Jamie aldorous alaarabia",al moassassa al haditha lilkitab, Tripoli,Lebanon,.

[7] M Mohamed, "Generation Morphological and Applications", Specialty thesis of 3rd round, Mohammed V University in Rabat-Morocco, 1999.

[8] A.Yousfi, (2010) "The morphological analysis of Arabic verbs by using the surface patterns", IJCSI International Journal of Computer Science Issues,7(3(11)): p. 33-36.

Authors

M. Mourchid Doctorate Degree In Computer Science In 1999; Associate Professor At The Computer Science Department At The Faculty Of Sciences, Ibn Tofail University In Kenitra Morocco; On Going Research Interests: Natural Language Processing, Web Semantic, And Information Systems.



I. Blanchete, Phd Student At Ibn Tofail University Department Of Computer Science, Laboratory Of Misc Kenitra , Morocco ;Graduated From Damascus University Faculty Of It Engineering 2010.

A. Mouloudi was born in 1959 in Morocco. He received the B.S degreein applied mathematics from Mohamed V University in Morocco at 1982;Master in Computer Science from the University Mohammed V, at 1984; Ph.D in Computer Science from the same university, at 1988. He obtains Habilitation to direct academic research in Computer Science, from Ibn Tofail University in Morocco, at 2008. Currently, A. MOULOUDI is the Director of the laboratory MISC (Information Modelling and Communication System), at the sciences faculty, Ibn Tofail University, in Morocco.