An Ontology Based Formalism for the Arabic Language Using Verbs and their Derivatives

Fatma Zohra Belkredim[1*], Ali El Sebai [2]

[1] Faculty des Sciences et des Sciences de l'Ingenieur
Departement d'Informatique,
Universitie Hassiba Ben Bouali
Chlef, Algeria
fzbelkredim@yahoo.fr

[2] School of Computing, Science and Engineering
Newton Building, University of Salford
Salford M5 4WT, UK
alielsebai@hotmail.com

Abstract. *Developing a strong and complete lexicon is essential for many Natural Language Processing (NLP) related applications such as Information Retrieval, Question Answering Systems, Natural Language Generation and Machine Translation. Such lexicon should be supported by a strong theoretical framework and implemented using tools that will facilitate its implementation and efficient use both in terms of resources and language preciseness. Recognising the derivative and inflexional nature of the Arabic language, the paper presents a formalism based on mathematics and ontologies of part of the morphological knowledge of the Arabic language based on verbs and their derivatives. In this research, we attempt to use the derivations and their patterns to structure the Arabic language and to strongly link the words' morphology to their semantics. We first perform a thorough study of part of the Arabic Language and propose a model that is based on set theory and ontologies. We then show how this model can be used for some applications that include NLP applications.*

1   Introduction and Motivation

1.1   Introduction

Arabic Language can be classified into three types: Classical Arabic (العربية الفصحى), Modern Standard Arabic (العربية الحديثة), and Colloquial Arabic dialects (العربية العامّية) [11]. Classical Arabic is the language of the holy Quran. It could also be viewed as the language of the pre-Islamic poets. This language is fully vowelized and is rarely used in today's everyday writing. Modern Standard Arabic (MSA) is the language of today's Arabic newspapers, magazines, periodicals, letters and modern writers. It is also used as the medium of oral communication in formal speeches and in television and radio broadcasts. MSA could be viewed as classical since there have been no major changes modifying the structure of the classical language. MSA, however, differs from Classical Arabic in two aspects: adopting minor stylistic changes and expanding the lexicon to include new technical terms [3]. Colloquial Arabic dialects, on the other hand, consist of the languages of the different Arab countries. They are used for everyday oral communications by the people of different areas. There are no written transcripts for such dialects. In this paper, we should deal with Arabic text written in Classical or MSA [11].

There is a lot of interest in the study of the Arabic language in the last few years. Most research focused on the morphological analysis of the language and its use in applications such as Information Retrieval and Question Answering Systems. These studies have highlighted Arabic as a highly derivational and inflexional language in which morphology plays a significant role. However, most of these studies saw the derivational and inflexional aspects of the language as a disadvantage rather than an advantage when it comes to the automatic processing of the Arabic language. Most studies did not use the specificities of the Arabic language but used approaches that were successful with roman languages such as English and French. To our knowledge, there was no attempt to formalise the language in such a way that it can be used differently from other languages. In this

---

* Corresponding author

paper we present a formalism based on mathematics and ontologies of part of the morphological knowledge of the Arabic language based on strong (sometimes referred to as sound) verbs and their derivatives. We first perform a thorough study of part of the Arabic language and propose a model that is based on set theory and ontologies. We then show how this model can be used for some applications that include NLP applications.

1.2   Research Scope and Limitations

Arabic words are divided into three types: nouns, verbs and particles [2] and a structure of the Arabic language can be summarised as shown in in Fig.1. Derivations can be produced from nouns (verbal nouns more precisely as it will be described in section 3.1) and verbs. In Arabic, verbs are categorised in different ways according to the needs of the grammarians or applications. Verbs can be classified based on the number of characters that form their root. Hence we distinguish verbs with three or four characters referred to as triliteral or tetraliteral verbs respectively[1]. Arabic verbs can also be classified based on the nature of characters forming their root as this will influence their conjugation and the forms of their derivations. Based on the latter approach, we distinguish two major classes: strong and weak verbs. Strong verbs are verbs which root does not contain weak letters (i.e. alef (أ), waw (و), or yaa (ي)). Weak verbs are those which root contains one or more weak letters. Based on the position of the weak letter, the verb can be further classified into the following three subclasses: first weak (called Paradigm), second weak (called Hollow), third weak (called Defective). Some verbs may fall into more than one category. The work reported in this paper concerns only derivations from sound verbs as highlighted in Fig 1.
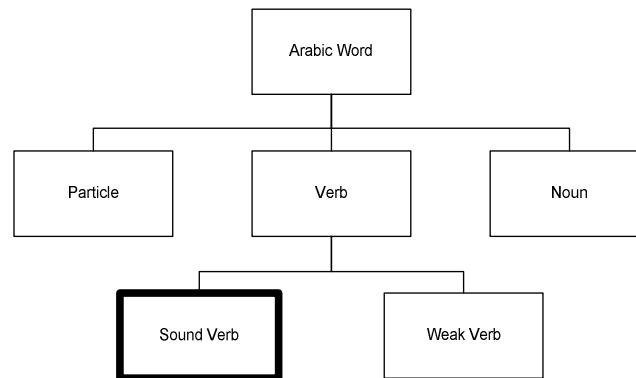

Fig. 1: A structure of the Arabic Language

The remaining of the paper is organized as follows. In section 2, we introduce some characteristics of the Arabic language and its structure. In section 3 we present our model for the Arabic language by identifying the mathematical model behind it and the derivation rules used for the construction of our ontology. In section 4, we propose some possible applications of our model and we conclude in section 5.

2   The Structure of the Arabic Language

As shown in Fig 1, the Arabic language is composed of nouns, verbs and particles. Nouns and verbs are morphemes and derived from a close set of around 10,000 roots [13]. Particles are used to complete the meaning of verbs and nouns. The roots are commonly of three or four letters, referred to as triliteral and tetraliteral roots respectively. Arabic nouns and verbs are derived from roots by applying templates to the roots to generate stems and then introducing prefixes and suffixes [7]. It was reported in ElKateb et al., [2006] that "85% of Arabic words are derived from triliteral roots" [8].

The Arabic verb is any word that indicates the occurrence of an action that is associated with time [4,10,15] . An Arabic verb will have a voice (active or passive), a tense (past, present and imperative), a gender (feminine or masculine) and a number (singular, dual, plural). Verb derivation in the different tenses is achieved using well-behaved morphological rules using Equation 1.

---

[1] There is a very small number of verbs with 5 characters but these are very often ignored in the many studies conducted on the Arabic language.

Verb = Prefix1 + Prefix2 + stem + Suffix1 + suffix2 + suffix3   (1)

The stem is formed by substituting the characters of the root into certain verb forms, called patterns (or patterns). Consider the following example [12]:

Root (ك ت ب) /ktb/
Pattern (تَفَاعَلَ) /tafa:'ala/
Stem (تَكَاتَبَ) /taka:taba/

The pattern is defined by El-Sadany and Hashish [1989] as: "a general mould composed of an ordered sequence of characters". There are 37 patterns for triliteral and tetraliteral verbs [12]. Arabic grammarians modelled the formation of nouns and verbs and their derivatives based on the root concept. The root is a set of the three consonants f 'l (ف ع ل) expressing the idea of the action 'to act' [4]. For example, the three consonants k t b (ك ت ب) expresses the notion of writing and so on. The root is not part of the language. However, to best represent this root Arabic grammarians often use the third masculine person in the past tense of a verb. This is similar in meaning to the infinitive mood in English or French languages. The verb kataba (كَتَبَ) (to write) is derived from the root "ktb" and scaled to fa'ala (فَعَلَ) [4].

All verbs have a pattern which not only provides morphological information, but also provides in many cases semantic and contextual knowledge. Hence, certain patterns can state that the action is performed only once, performed with some intention etc. Examples showing some of this semantic knowledge will be given in section 3. It is therefore, desirable to define a model to represent the Arabic language that not only models the morphology, but also uses this as the primary source for semantic and contextual knowledge. Hence, in this research, we attempt to use derivations and their patterns to structures the Arabic language and then link words' morphology to their semantics. This representation is modelled as an ontology. In the following subsections, we describe the various derivations by providing their patterns and then assign part of the ontology structure.

3   A Model for the Arabic Language

Let L be the universe of discourse (UoD) of the Arabic language. This UoD is composed of all the words of the Arabic Language. Let R be the relation "Has the same root as" and let x, y, and z be elements of L. R is an equivalent relation on L because:

R is reflexive: *for any x of L*   xRx
R is symmetric: *for any x and y of L*   xRy => yRx
R is transitive: *for any x, y and z of L*   xRy and yRz => xRz

R divides the set L into subsets where each subset is an equivalent class. Each class $cl_i(w)$ contains all the words derived from the word "w" which in our current research is chosen to be the verb.

Verbs have a specific role in the Arabic language and particularly in grammar. Indeed, most Arabic sentences begin with a verb [14] and the subject cited or implicit is linked to the verb. Although Grammarians agree that triliteral verbs are derived from verbal nouns (مَصْدَر), which literally means in Arabic the source, they also agree that verbal nouns cannot be structured and classified in a logical manner.[4,9,10]. Hence, to develop an efficient computing system for the Arabic language it is preferable to use verbs. Verbs are very often used as stems for deriving nouns and other verbs. This is the approach we will adopt in our research. This view is also supported by Koufa Grammarians that favour deriving from verbs.[1]

The verb is also used as the representative of the equivalent class that identifies it. Verbal nouns could not be used as they are numerous and have different patterns example (صَرَخَ صُراخ صَريخ) (to shout)). In Arabic, we use the verb in the third masculine person past tense to represent the basic form of a verb and this is also used to identify an equivalent class in our research. Hence for the basic form f ' l represented by fa'ala, the cl(fa 'ala) uses derivation and inflexion rules to built all the nouns and verbs derived from (fa 'ala).

Furthermore, in order to relate the word's morphology to its semantic and contextual knowledge, we model each equivalent class as ontology. In the following subsections, we describe the main derivations obtained from the verb and the associated part of the ontology associated with them.

3.1 Derived Nouns

Nouns attached to verbs are called in Arabic "Derived Named Nouns" (الأَسْماءُ المُشَبَّهَةِ بِالفِعْلِ). Ghalayani [10] has identified 9 categories; however three of these are adjectives. We favour in our research the use adjectives as a single category and then identify subcategories for the adjective category. We have also grouped Time and Space nouns into one single category as most of them are using the same derivation rules. The following describes our defined six categories:

- **The verbal noun (المَصْدَر)**: Verbal nouns are nouns that are formed directly as an inflexion of a verb or a verb stem [16]. However, in Arabic the word "مَصْدَر" means source and most grammarians state that triliteral verbs are derived from verbal nouns. However, in practice, it is much easier to organise the ontology based on the verb and to link the verbal nouns to the verb as verbal nouns do not have a logical organisation and they have too many patterns [4,10]. Verbal nouns refer to various meanings such as the act of doing something (الضَّحِك) or the result of doing something (الخُروج). For triliteral verbs, there are 44 types [4]. Examples of these forms include: فَعْل فِعْل فَعَل فِعِل فِعَل فَعَال فَعَال فَعِيل فَعَلان فُعول فَعَالَة فُعولَة etc.فُعْلان فِعْلان فُعْلَى فَعَلوت

- **The active participle (اِسْمُ الفَاعِلِ)**: Active participle represents the subject
performing the action described by the verb. Each verb in Arabic has one active participle and this include the verb's derivations. Table 1 summarises the patterns of the active participle for some verb patterns with examples. Hence, it is not only possible to identify that a particular word is an active participle, but we can also identify the root, hence the meaning of the verb from the ontology.

Table 1. Patterns of some active participles with examples

Verb pattern	A.P. pattern	Example
فَعَلَ / فَعِلَ/ فَعُـَل	فاعِل	كَتَبَ كَاتِب
فَـَعَّـَل	مُفَعِّل	كَسَرَ كَسَّرَ مُكَسِّر
أفْعَل	مُفْعِل	جَلَسَ أجْلَسَ مُجْلِس
فَاعَل	مُفاعِل	تَبَعَ تابَعَ مُتابِع
تفَاعَل	مُتَفاعِل	سَبَق تَسابَقَ مُتَسابِق
تَفَعَّـَل	مُتَفَعِّل	كَسَرَ تَكَسَّرَ مُتَكَسِّر
إنْفعَل	مِنْـُفعِل	خَدَعَ إنْخَدَعَ مُنْخَدِع

- **The passive participle (اِسْم المَفْعُول)**: Represents the object upon which the action is performed. The passive participle takes the form مَفْعُول. For example مَكْتُوب (what is written, destiny) is derived from كَتَبَ (to write). We note that not all verbs will have a passive participle [16].

- **Substantive Adjectives:** Adjectives are divided into three subcategories:
  - Adjectival Noun, (الصِـِفَة المُشَبِه): These are nouns that in a particular context are regarded as adjectives and describe the state of a person or a thing. From triliteral verbs, adjectives with the following patterns are derived: فَعْل فَعَل فَعال فَاعِل فَعِل فَعَل فَعْلان فَعِيل فَعِل and فُعْل. For example سَمِيع (someone who can hear) is derived from سمع (to hear).
  - Overstated form Adjective (صِيَغُ المُبَالَغَةِ): These are used to overstated the qualities or descriptive of a noun. They have the following three patterns: فَعِيل فَعَّال and فَعُول. For example صَدِّيق (trustworthy) is derived from صدق (to trust) [4,10].
  - Elative adjective (اِسْم التَفْضيـِل): Also known as superlative adjectives indicate that something has some feature to a greater degree than anything it is being compared to in a given context. They are also used for comparison. These are only derived from triliteral verbs and take the pattern أفْعَل. For example أكْبَر (the greatest) is derived from the verb كَبُرَ (to grow).

- **Time and Space Nouns (اِسْم الزَمانِ, اِسْم المَكَان):** These nouns refer to places or periods in time. The following patterns are used for time nouns, space nouns or both (مَفْعَل,مَفْعِـَل) for example مَسْكَن (the place where someone lives) is derived from the verb سكن (to inhabit) and مَشْرَق (when the sun rises) is derived from the verb شرق (to rise). There are three patterns that are used to generate only space nouns these are: مَفْعَـُلَة and مِفْعَلَة for example مَكْتَبَةٌ (library) derived from the verb كَتَبَ (to write)

- **Instrument noun:** This names category denotes instruments used to perform the action described by the verb[2]. The patterns of instrument nouns are: مِفْعَلَة , مِفْعَل and مِفْعَال . For example, مِفْتاح (the key) is derived from فَتَحَ (to open).

---

[2] Note that not all verbs have a derived instrument noun

The part of the ontology that models the derived nouns is shown in Fig 2 which shows the main six categories of the derived nouns and the three subclasses of adjectives.
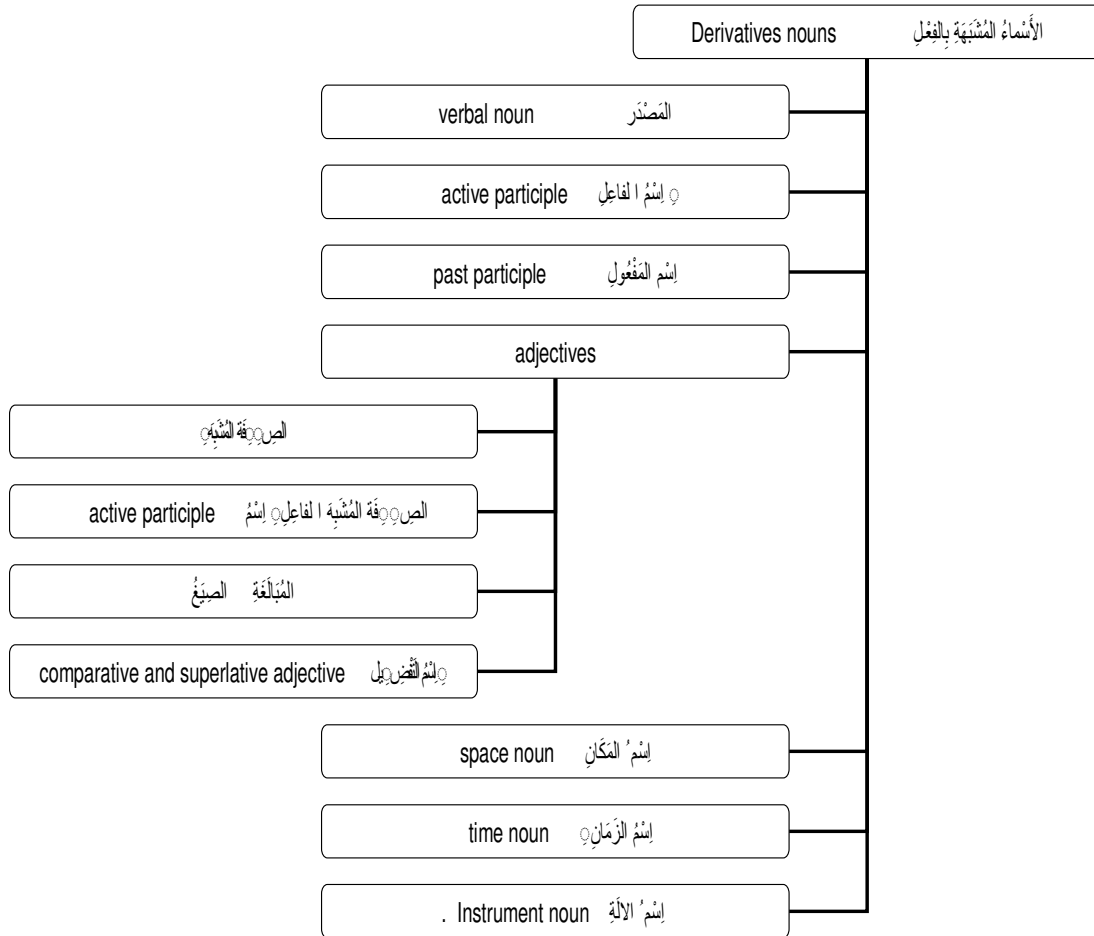

Fig. 2. Part of the Ontology representing Derived Nouns

3.2   Derived Verbs

The basic patterns of triliteral and tetraliteral verbs are as follow:
- Triliteral verbs have the following three basic patterns: (فَعَلَ /فَعِلَ /فَعُلَ).
- Tetraliteral verbs have only one basic pattern which is represented as: فَعْلَلَ.

From these basic forms, many derivatives are produced and these are based on the number of consonants in the verb. The derivation is composed of the basic consonants forming the root to which we add one or more consonants. The derivatives forms of the triliteral and tetraliteral verbs are described in the following subsections.

3.2.1 Primary Derivations from Triliteral Verbs

There are twelve primary derivations from triliteral verbs and these are classified into three categories. The first category contains those derivations that have 4 characters (The three basic characters from the root add one) and they have the following three patterns (فَعَّلَ / أَفْعَلَ / فَاعَلَ). The second category contains those derivations with 5 characters (the three basic characters add two) and they have the following five patterns (/ تَفَعَّلَ / تَفَاعَلَ إِنْفَعَلَ / إِفْعَلَّ / إِفْتَعَلَ) and the third category contains the 6 characters derivations and they have the following four patterns (إِسْتَفْعَلَ / إِفْعَالَّ / إِفْعَوَّلَ / إِفْعَوْعَلَ).

3.2.2 Derivations from Tetraliteral Verbs

There are three derivations from tetraliteral verbs and similar to triliteral verbs, the classes are defined based on the number of characters composing the derivations. The first category contains the one five character derivation and has the form (تَفَعْ⊝لَّ). The second category contains two derivations with 6 characters and they have the patterns (إِفْعَلَّ / إِفْعَنْلَلَ).

3.2.3 Secondary Derivations from Triliteral Verbs

The secondary derivations of the triliteral verbs are obtained as follow. We add one; two or three characters to the triliteral verb to give it one of the four patterns of the tetraliteral verb respectively (فَعْلَلَ, تَفَعْ⊝لَلَ, إِفْعَنْلَل, إِفْعَلَّ) and different variations of the form having the same pattern are than constructed following very strict rules [10]. The first secondary class contains the verbs constructed using the (فَعْلَلَ) pattern and the following seven forms are obtained[3] :

- (فَعْلَلَ   two last consonants are the same and fixed to ل ل),
- (فَعْوَلَ   the third consonant is fixed to و),
- (فَوْعَلَ   the second consonant is fixed to و)
- (فَيْعَلَ   the second consonant is fixed to ي)
- (فَنْعَلَ   the second consonant is fixed to ن)
- (فَعْيَلَ   the third consonant is fixed to ي)
- (فَعَ⊝لَىَ   the last consonant is fixed to ى)

The second secondary class contains 5 characters' verbs constructed using (تَفَعْلَلَ) pattern. There are six patterns in this class and these are constructed as follows[4]:

- (تَفَعْلَلَ   two last consonants are the same and fixed to ل ل),
- (تَفَعْوَلَ   the fourth consonant is fixed to و),
- (تَفَوْعَلَ   the third consonant is fixed to و),
- (تَفَعَيْلَ   the fourth consonant is fixed to ي),
- (تَفَيْعَلَ   the thirst consonant is fixed to ي),
- (تَفَعَلَىَ   the last consonant is fixed to ى).

The third secondary class contains 6 characters verbs constructed using the (إِفْعَلَّ,فْعَنْلَل) forms. There are 3 patterns in this class and these are constructed as follows:

- (إِفْعَنْلَل   two last consonants are the same ل ل),
- (إِفْعَنْ⊝لَىَ   the last consonants is ى),
- (إِفْتَعَلَىَ   the third consonant is fixed to ت and the last consonant is fixed to ى).

The part of the ontology that models triliteral verb derivations is given in Fig 3. Each verb will have a list of primary verb derivations and a list of secondary verb derivations. The primary derivations contain those verbs which patterns are described in section 3.2.1 and the secondary derivations contain the verbs described in section 3.2.3. Each verb will then have a list of derived nouns as shown in Fig 2 and a list of conjugated forms. The overall structure of the ontology is given in Fig 4 which includes the derivations for tetraliteral verbs.

3.3 Conjugation

As mentioned earlier, the conjugation in Arabic has three tenses, the past tense or perfective (ا لماضي), the present and future tense or the imperfective[5] (المضارع), the imperative tense (الامر) and two voices passive and active. The passive form is usually referred to as: (مبني للمجهول). The past third masculine person is used to represent the infinitive form of the verb. The Arabic language has separated and attached personal pronouns. We will not refer to the conjugation in details in this current research.

---

[3] Referred to by some Arabic literature as "مُلْحَقات دَحْرَجَ" because they have the pattern of the verb " دَحْرَج".
[4] Referred to by some Arabic literature as "مُلْحَقات تَدَحْرَجَ" because they have the pattern of the verb " تَدَحْرَجَ".
[5] The imperfective tense can also be conjugated using different moods.
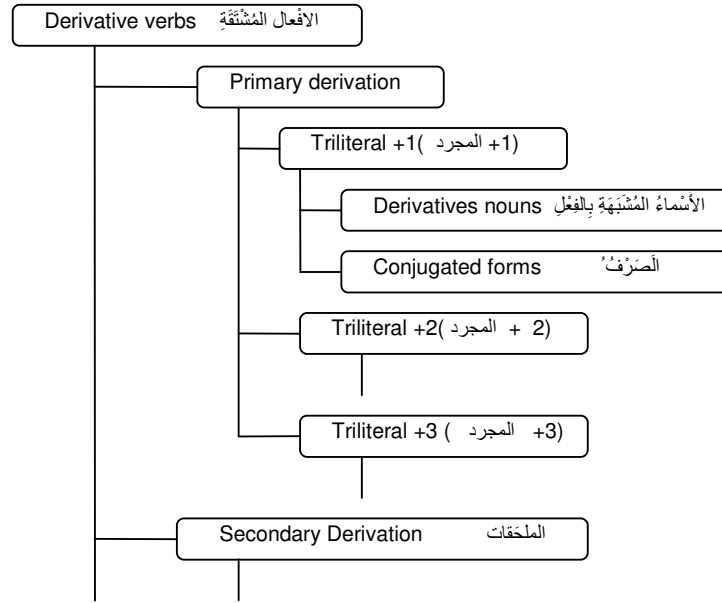
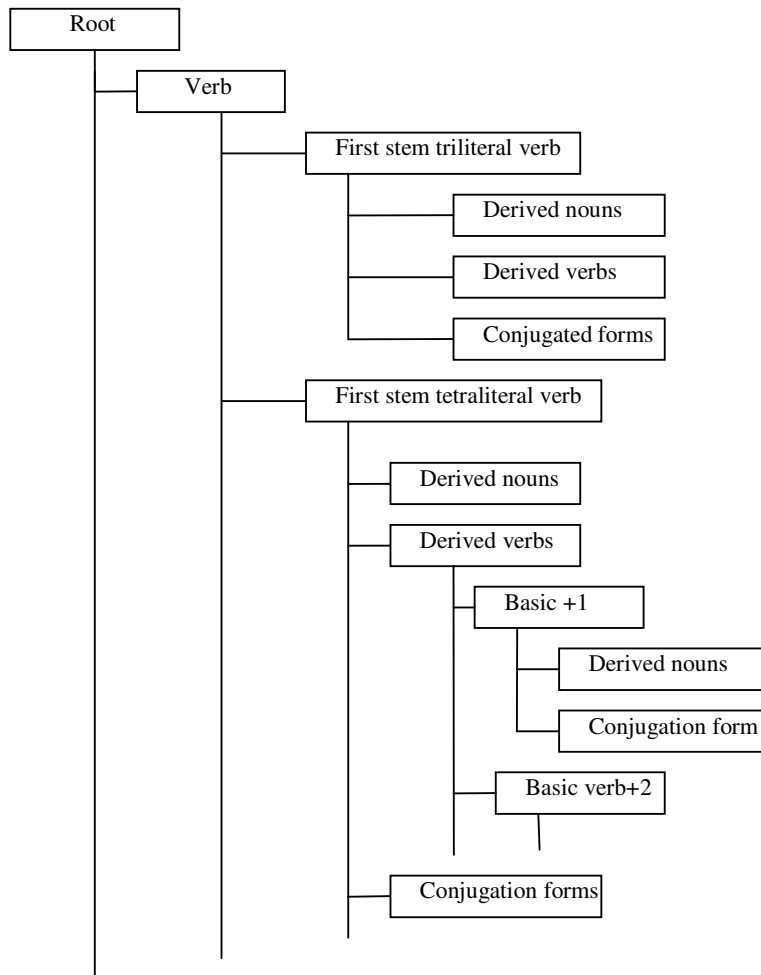**Fig. 3**. Part of the Ontology Representing Derived Verbs

Fig. 4. Part of the Arabic Derivational Ontology

4   Applications of the Proposed Model

The ontology developed in this research is unique to Arabic as it is based on the rules that govern the language. Very often, approaches to Arabic NLP were based on methods developed for roman languages [5,8] or use transliteration [6] and this has created unnecessary problems and complexity. Our model formalise the Arabic language using its derivation and inflection rules. We strongly believe that such a model can be used successfully in many Arabic applications whether related to NLP or not. We summarise here some of the applications where our model can be used.

- **Automatic Ontology Construction:** With our model, we can use the list of existing Arabic verbs to generate all its derivations and populate the ontology in an easy and straightforward manner. As not all verbs have all the derivations, it is necessary to use an electronic dictionary to eliminate nouns that are not found in the dictionary.

- **Arabic Morphology:** The model proposed can be used as the perfect Arabic morphology analyser. For each work of a specific equivalence class, the root and all related words can be found using the ontology structure. This is a computationally efficient and linguistically complete approach. A strong morphology system will help the development of many applications such as information retrieval.

- **Arabic Language Understanding:** The ontology developed was based of the various patterns associated with verb derivations. As shown in the previous sections, most of these patterns also carry semantic and contextual knowledge that can be used to improve existing Arabic language understanding systems.

- **Arabic Language Development:** Such Ontology can be used to teach the Arabic language. Once the meaning of a verb and its derivations rules known, it is easy to develop learner's vocabulary and understanding of the Arab language.

5   Conclusion and Future Developments

The model developed in this research is based on an Ontology using the derivation rules of the Arabic language. Although, in the Arabic language, triliteral verbs are derived from verbal nouns (مَصْدَر), their complexity, different variations and lack of logical structures makes them extremely difficult to use as the root for deriving verbs. As this study shows, we did find it much easier to derive from verbs as the list of Arabic verbs is known and is finite (countable).

The model we developed in this research is at an early stage and can be improved in many ways. We have only used strong verbs hence further developments should include weak verbs, nouns and particles. The implementation and use of this model with some real application and a comparison with more traditional approaches would be the way to evaluate this model. The model we developed in this research has been evaluated by the linguists and computer sciences team directed by Dr Elkalak [17].

References.

1. Abdelwahab Hassen Hamed D. **كينونة   الفعل**. (verb existence) 2003/2004.
    http://www.minshawi.com/other/abdalwahab13.htm

2. Abdul-Al-Aal, Abdul-Monem, An-Nahw Ashamil, Maktabat Annahda Al-Masriya, Cairo, Egypt, 1987.

3. Aljohar B., A Portable Natural Language Interface from Arabic to SQL, PhD thesis, University of Sheffield, UK, 1999.

4. Blachere R. and. Gaudefroy-Demombynes M., Grammaire de l'Arabe classique, (3rd edition) Maisonneuve & Larose Paris, 1975

5. Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C., Introducing the Arabic WordNet project, Proceedings of the 3rd Global Wordnet Conference, Jeju Island, Korea, South Jeju, January 22-26, 2006.

6. Buckwalter, T., Arabic Morphological Analysis, http://www.qamus.org/morphology.htm

7. Darwish K., Building a Shallow Arabic Morphological Analyzer in One Day, Proceedings of the Computational Approaches to Semitic Languages, A workshop affiliated with ACL-2002, University of Pennsylvania, 2002.

8. Elkateb, S., Black, W., Rodriguez, H, Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C., Building a WordNet for Arabic, Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006)., 2006.

9. Elbaydhani, S نزهة الطرف شرح بناء الأفعال في علم الصر وتيسيره (verb explanation and structure in sarf science and to render it easy) -**http://www.almeshkat.net/books/archive/books/altef .zip** . 2000

10. Ghayalani M., جامع الدروس العربية دار ابن الهيثم القاهرة مصر, 2005.(Encyclopedia of Arabic courses) Cairo Egypt 2005.

11. Harmain, H.M. and Aljohar B, A Framework for an Arabic Information Extraction System, Proceedings of the Second Arab Conference on IT, 13-15-11/2001 , Jordan.

12. El-Sadany T.A. and Hashish M.A., An Arabic morphological system, IBM Systems Journal, 28(4):600-612, 1989.

13. 13. Ibn Manzour, Lisan Al-Arab. (Arabic Tongue) www.muhaddith.org.

14. Pashova T., The VS/SV alternation in modern written Arabic from a textual perspective, Zeitschrift für arabische Linguistik (Z. arab. Linguist.), no42, pp. 7-40, ISSN 0170-026X , 2003.

15. Shaalan K., Arabic GramCheck: A Grammar Checker for Arabic, Software Practice and Experience, 35(7):643-665, June 2005.

16. .Wikipedia www.wikipedia.org

17. http://sourceforge.net/projects/sarf (Arabic system morphology)