Formalizing Arabic Inflectional and Derivational Verbs Based on Root and Pattern Approach Using NooJ Platform

Ilham Blanchete(✉), Mohammed Mourchid, Samir Mbarki, and Abdelaziz Mouloudi

MIC Research Team, Laboratory MISC, IbnTofail University, Kenitra, Morocco
ilham.blanchete@gmail.com, mourchidm@hotmail.com, mbarkisamir@hotmail.com, mouloudi_aziz@hotmail.com

Abstract. This article presents the inflectional and the derivational model of Arabic verbs based on root and pattern approach, using a linguistic classification that determines and specifies a set of morphological properties.

Our work in NOOJ platform is based on:

- A dictionary that consists of roots, lemmas and patterns.
- Generating all possible verbs inflectional and derivational forms using our NooJ grammars.

Since formalizing Arabic verb's requires the determination of certain unavoidable morphological properties as root and pattern, we started by specifying them; the matching process between them gives twenty thousand verbs entries [1]. Consequently, this work stands on the following principles:

- Categorizing all possible roots and patterns.
- Matching roots with patterns that give Lemma, which is a verb in our case.

Our dictionary considers, in addition to roots and patterns, lemmas as dictionary entries, which allows making an advanced search in a text.

Root and pattern serve to build the meaning of most Arabic word [1]. For instance, matching the root(KTB-كتب) with the pattern فَعَلَ [Fa3aLa], gives the lemma (verb): كَتَبَ(KaTaBa-to write); matching the same root with the pattern فِعَالٌ (Fi3aLun) gives the lemma (noun): كِتَابٌ(KiTaBun – a book); matching the same root with different patterns gives different Arabic words.

The implementation of this work is a dictionary that contains all inflectional and derivational verbs forms.

Adding other Arabic words (nouns, adjectives, infinitives) with their inflectional and derivational forms, in order to complete our dictionary, is considered as our future perspective.

Keywords: ANLP · NooJ platform · Inflection and derivation

S. Mbarki et al. (Eds.): NooJ 2017, CCIS 811, pp. 52–65, 2018.
https://doi.org/10.1007/978-3-319-73420-0_5

1 Introduction

The linguistic analysis must go through a first step of lexical analysis, which consists of testing the membership of each word of the text to the Arabic vocabulary. So, we must begin with the formalization of the Arabic vocabulary [2]. Verbs cover a large part of Arabic vocabulary. Each Arabic verb has a one single root, also called radical, and a single pattern. In addition to these two components, the verb has other morphological properties that we are going to detail.

The verb is classified into several categories according to the number and the nature of its root letters. The verbs that have 3 letters (called trilateral verbs) form 64% of Arabic verbs, like the verb (to write - كَتَبَ- KaTaBa), those that have 4 letters (called quadrilateral verbs) form 33% of Arabic verbs, like the verb (to assure - طَمْأَنَ - TaMAaNa), and those that contain 5 letters, form the rest of the percentage [3], the most frequently ones are those that have 3 radical letters.

In this paper we give an overview about the first and the only dictionary that was implemented in NooJ platform (EL-DiCar), then we present a theoretical study that reflects the linguistic side of our work, and answers the question: why do we have to specify the root and the pattern during the formalization of Arabic vocabulary especially verbs? And, We also give the linguistic classification of Arabic verbs that we have adopted during the formalization process, as a first part, the second part is the practical study which is the implementation of the theoretical study that was enumerated in the first part of the article, using NooJ platform importantly; we are going to detail the dictionary structure, and give an example of our own lexical grammar that we have implemented to generate the inflectional and the derivational forms of all dictionary entries. In addition to that we are going to present the annotations as the result of the text linguistic analysis. Finally, we are going to present the comparison results between the search operation using root and pattern in our dictionary and the lemma search in the previous dictionary.

2 Related Works

As far as the previous works in NooJ platform are concerned, there is only one dictionary called the Arabic EL-DiCar dictionary that is based on lemma approach. EL-DiCar retrieves all inflectional and derivational forms of one entry using lemma search. It contains, in addition to the Arabic words, 10375 fully vowelled entries of verbs; each entry represents third person, singular, masculine and perfect verbs [4]. These entries are not related to each other even if they share the same root. For example: the verb (to write – KaTaBa – كَتَبَ), the verb (to write a letter - iKTataBa – اِكْتَتَبَ), the noun (library – maKTaBa - مَكْتَبَة) and the noun (book – KiTaB- كِتَاب) share the same root that is: [ktb-كتب], but they are not linked to each other in this dictionary; entries have no relation between them whereas the dictionary is based on lemma. This limits the search operation, and makes the application of several operations on the text like retrieving concepts or the auto-correction more complex. Our model that is based on root and pattern approach retrieves all entries with their inflectional and derivational forms that share the same root. It can also we can extract concepts within a text using pattern search even if these concepts do not share the same root.

3 Theoretical Study

3.1 Arabic Verb Morphology

Arabic morphology exhibits rigorous and elegant logic. It consists primarily of a system of consonant roots, which interlocks with pattern to form most Arabic words, especially verbs [5]. With the same root we can derive several words by using different patterns. In this context it is necessary to provide a definition of root and pattern:

A root is a relatively invariable discontinuous bound morpheme, represented by two to five phonemes, typically three consonants in a certain order, which interlocks with a pattern to form a stem [5]. For instance, the verb (to write, KaTaBa - كَتَبَ) has the root letters (KTB-كتب), that are interlocked with the pattern (FaAaLa – فَعَلَ). To clarify, the first letter in the root will be substituted with the first letter in the pattern and so on. This process of interlocking aims to form the verb (to write, KaTaBa-كَتَبَ), as well as the quadrilateral roots like the root (TMAN- طمأن) and the pattern (FaALaLa- فَعْلَلَ) that gives the verb (to assure –TaMAaNa - طَمْأَنَ).

The importance of specifying the roots and the patterns in the morphological analysis during building a dictionary, and adopting them as dictionary entries is to make all nouns, verbs and adjectives; in general all Arabic words that share the same root or concepts, retrievable by their root or pattern.

A pattern is a bound and in many cases, discontinuous morpheme that carries meaning [6]. For example, the pattern (FaAaLa- فَعَلَ) refers to an action, matching this pattern with:

- The root (KTB-كتب) gives the verb (to write, KaTaBa –كَتَبَ).
- The root (SANAa-صنع) gives the verb (to make-SaNaAea-صَنَعَ).
- The root (JLS-جلس) gives the verb (to sit-JaLaSa-جَلَسَ).

With a simple search operation using this pattern, we can extract all verbs that refer to an action, even if they do not share the same root.

Matching the pattern (MaFALa- مَفْعَلَةُ) that refers to a place with:

- The root (KTB-كتب) gives the noun (a liberary - maKTaBa - مَكْتَبَة).
- The root (KHBZ-خبز) gives the noun (a bakery shop - miKHBaZa -مَخْبَزَةٌ).

To extract all nouns that refer to a place, we simply apply a pattern search using the previous pattern, the same thing applies to all other patterns.

The following table represents a part of the Arabic lexicon that is formed by the possible matching process between roots and patterns: 1 means that the matching process is possible, while 0 means that the matching process is not possible or there is no Arabic lemma formed from this combination (Table 1).

3.2 Verb Classification

This linguistic classification covers all Arabic verbs categories [6]. Structured as a tree, each leaf represents a verb model that we are going to formalize using NooJ platform; Fig. 1 shows a part of the classification tree that we used.

Table 1. Capture of Arabic lexicon formed from matching several roots and patterns

Root/pattern	فَعَلَ FaAaLa	فَعُلَ FaAoLa	فَعِلَ FaAiLa	فَاعِلُ FaAiL	تَفَاعَلَ taFaALa
حصن HSN	0	1	1	1	0
كتب KTB	1	0	0	1	1
ضرب DRB	0	0	1	1	1

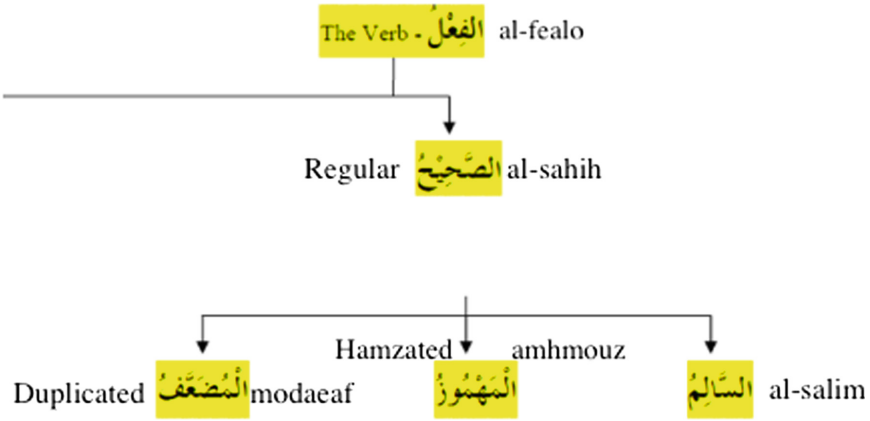

Fig. 1. Arabic regular verb classification

Arabic verbs are divided into two main classes: regular verbs, like the verb (to write), and weak verbs like, the verb (to say) [6]. We are going to detail one of them.

Regular Verbs. Verbs that their roots are free of vowel are also called (الأَفْعَالالْصَحِيحَة- AaFAaL - SaHiHa) like the verb (to write كَتَبَ – KaTaBa).

It is worth mentioning that the vowel letters in Arabic are one of these letters {ي,ا,و} - {YaE, AaLiF, WaW}.

Regular verbs, in their turn, are divided into three sub-classes, as Fig. 1 shows, [sound -سَالم –SaLiM], [Duplicated – مُضَعف – MoDaAaEF] and [hamzated – مَهْمُوز - MaHMoZ].

Sound Verbs. Their radical letters are free of hamzaletters(أ-ء-ئ-ؤ), and none of their radicals are identical. In its turn, this type of verbs is divided into three sub classes. The reason of this division is due to the inflections forms. All verbs that endwith the letter [n - ن] are inflected in a similar way to those that end with the letter [t - ت] except the first singular person and the second plural person, to which we have to add *shedda*ّdiacritic.

Duplicated Verbs. They contain all verbs that their second and third root letters are identical. For example, the root letters of the verb (to count -عَدَّ- AaDa) are (ADD-عدد); the diacritic sign ّ refers to the duplication of any letteron which it is placed. This category of verbs is divided into two sub-classes: category * (all verbs that their root letters are free of hamza) like the verb (to count – AaDDa- عَدَّ), its root letters are: [عدد]

free of hamza and the second category is hamzated (verbs that contain hamzaletter in one of their three radicals). These two categories are divided, in their turn, three sub-categories [7].

Hamzated Verbs. The category of the verbs that contain hamzaletter in any of its root letters like the verb (to read - قَرَأَ -KaRaAa), is divided into tree sub-classes, as it is shown in Fig. 1.

3.3 Inflectional and Derivational Forms of Arabic Verbs

Arabic derivations are derived from the combination of a specific pattern and a given root. For example, the matching process between the root (LAAB-لعب) and the pattern (FaAiLon- فَاعِلٌ), that serves to generate the active participle form, gives the noun (player – LaAiBon-لَاعِبٌ). The same thing is applied to the rest of the derivation forms of all Arabic verbs.

Also, the matching process gives the inflections of any Arabic verb as we have explained before, or by applying several morphophonological alternations to this combination.

Arabic verbs inflect for the following grammatical categories:

- *Subject agreement*: in person (1st, 2nd and 3rd), number (singular, dual, and plural) and gender (masculine and feminine)
- *Tense / aspect*: (perfect المَاضِي-imperfect المُضَارِع)
- *Mood*: (indicative المَرفُوع, subjunctive المَنْصُوب , jussive المَجْزُوم, long energetic المُؤَكَد الثَقِيْل, imperative الأَمْر, imperative of long energetic الأَمْرالمُؤَكَد) and
- *Voice* (active المَبْنِي لِلْمَعْلُوم and passive الْمَبْنِي لِلمَجْهُول) [6].

Along with the root and pattern, we have adopted the previous grammatical categories as the text annotation. The result of the text analysis appears as annotations that give all information of the grammatical category, root and pattern of all verbs in the text.

4 Practical Study

We are going to review the previous theoretical study using NooJ platform, in order to formalize a comprehensive model of Arabic verbs, by building a dictionary of verbs that is based on the root and pattern approach. This dictionary will use our lexical grammars to generate all possible verb inflections and derivations.

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, 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 [2].

As we have mentioned before, the following words: ([to study –DaRaSa - دَرَسَ] [a teacher - moDaRiS- مُدَرِس], [a school – maDRaSa - مَدْرَسَة], [a study - DiRaSa- دِرَاسَة], [a lesson – DaRS - دَرْس] and [to teach – DaRraSa - دَرَّسَ] are formed as entries and they share the same root. They share also the same concept. Even though they have different patterns. We can retrieve all of these words with their inflectional and

derivational forms within a text, using the root that they share. We can also apply a search using pattern to extract all words that share the same concept (as we have explained before), while EL-DiCardictionary cannot retrieve concepts, because it is based on lemma and their entries are separated.

Figure 2 shows the result of the search operation in EL-DiCardictionary using lemma [to write – كَتَبَ] as search entry, on a text that contains these words: ([Library - maKTaBa–مَكْتَبَةٌ], [Writer-KaTiB-كَاتِب], [Write a letter-KaTaBa- كَاتَبَ], [To copy - iKTataBa- اِكْتَتَبَ] and [Desk-maKTaB-مَكْتَب]), while Fig. 3 shows the same operation's result on the same text using the root as a search entry. As we can observe the first search operation in Fig. 2 that is based on lemma [to write – كَتَبَ] will retrieve only the active participle because it is formed as a derived form of this entry in EL-DiCar dictionary. Thus, this dictionary provides only inflectional and derivational forms of a given lemma.

Furthermore, unlike

The difference between our provided root and pattern approach and lemma approach, used by EL-DiCar dictionary is very clear by applying a search using lemma [to write – كَتَبَ] on a text that contains words which share the same root: ([Library - maKTaBa–مَكْتَبَةٌ], [Writer-KaTiB-كَاتِب], [Write a letter-KaTaBa- كَاتَبَ], [To copy - iKTataBa- اِكْتَتَبَ] and [Desk-maKTaB-مَكْتَب]) returned only the active participle that is formed as a derived form of the entry (to write – KaTaBa- كَتَبَ) in EL-DiCar dictionary. While the rest of the words are not retrieved as they are not linked with the lemma (to write – KaTaBa- كَتَبَ).

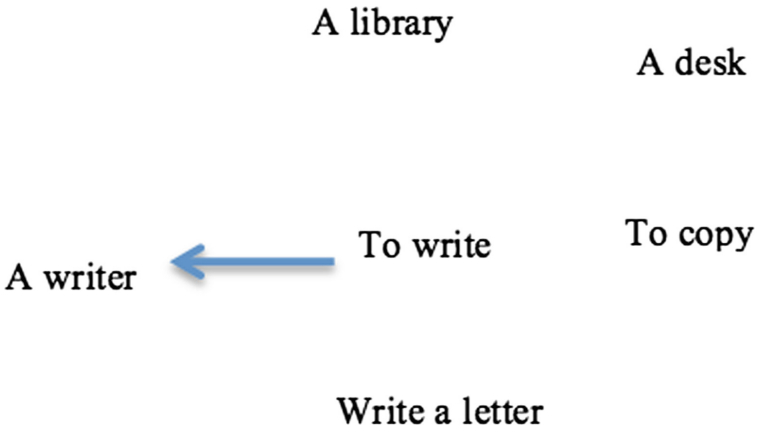

Fig. 2. The result of lemma search in EL-DiCardictionary.

Figure 3 shows the result of our dictionary that is based on root and pattern approach. All words that share the same root will be retrieved using root search.

Now, we move to explain the internal structure of EL-DiCar dictionary. To clarify In lemma based dictionary, the entry that represents a verb takes the form دَرَّسَ, V + Tr + FLX = V_darrasa + DRV = D_darrasa:FLXDRV:(دَرَسَ –DaRaSa- to teach) is an entry; V: Verb, Tr: Transitive, V_darrasa: the inflectional paradigm that gives 122

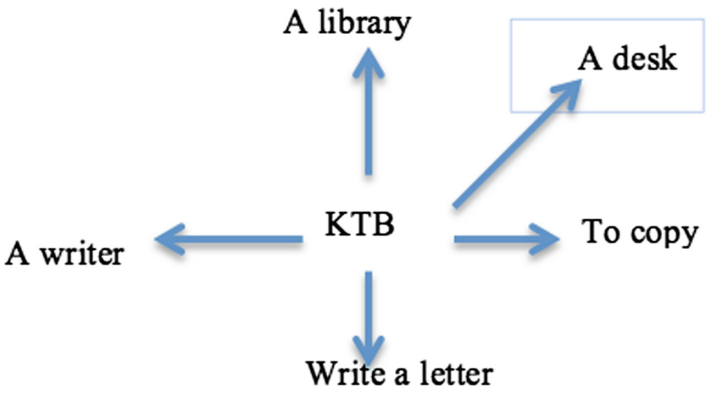

Fig. 3. The result of root search in our dictionary.

inflectional transformations, for example: [I thought – DaRaSTo - دَرَّسْتُ] and [you thought – DaRaSTa - دَرَسْتَ], and DRV = D_darrasa: derivational paradigm that gives the active participles (مُدَّرِسٌMoDaRiSon teacher) and the passive participles (MaDRouS – مَدْرُوسٌ – has been taught) of the entry, and other plural forms [7]. As we have mentioned before, both the active and the passive participles are derived from the same entry. Conducting a search using this entry will retrieve both of them if they exist within the text.

The following words: دِرَاسَةٌ, N + FLX = N_drassa2 (study - DiRaSa-دِرَاسَة) and مَدْرَسَةٌ, N + FLX = N_mdrassa, (a school – maDRrSa - مَدْرَسَةٌ) share the same root. Making a search using (DiRaSa-دِرَاسَةٌ - a study) or any of their inflectional or derivational formswill neither give (a school – maDRrSa - مَدْرَسَةٌ) nor their derivations and inflections if they exist within the same text.

Unlike EL-DiCardictionary, in our dictionary, we have added more morphological features: root and pattern. For example our dictionary gives the following forms of entries:

- دَرَسَ,درس,V+Tr+درس+فَعَلَ+FLX=flx1+DRV=drv1:flx2.
- مَدْرَسَة,درس,N+درس+مَفْعَلَة+FLX=N_mdrassa+DRV=drv2:flx3.
- دِرَاسَة,درس,N+درس+فِعَالَة+FLX=N_drassa+DRV=drv3:flx4.درس :root of the entries, [(فَعَلَ-FaAaLa), (مَفْعَلَة-maFAaLa),(فِعَالَة-FiAaLa)]: patterns of the entries, FLX and DRV the inflectional and the derivational paradigms, that serve to generate the inflections and the derivations of the entries.

The difference between the internal structure of EL-DiCardictionary and our dictionary lies in the adopted approach as we have previously explained.

Furthermore, unlike our dictionary EL-DiCar dictionary cannot make a root and pattern search. This is because EL-DiCar dictionary lacks the morphological features (root and pattern) that we are going to use to make an advanced search.

Making a search using the root [DRS-درس] in our dictionary, will retrieve all entries with their inflectional and derivational forms that share the same root like: ([to study – DaRaSa - دَرَسَ] – [a teacher - moDaRiS- مُدَرِس] – [a school – maDRaSa - مَدْرَسَة] - [a study - DiRaSa- دِرَاسَة] - [a lesson – DaRS - دَرْس] – [to teach – DaRraSa - دَرَّسَ]).

Figure 4 contains a text that we are going to analyze in NooJ platform, using our dictionary that is based on root and pattern approach, and linked to our lexical grammars that generate the possible inflectional and derivational verb forms.

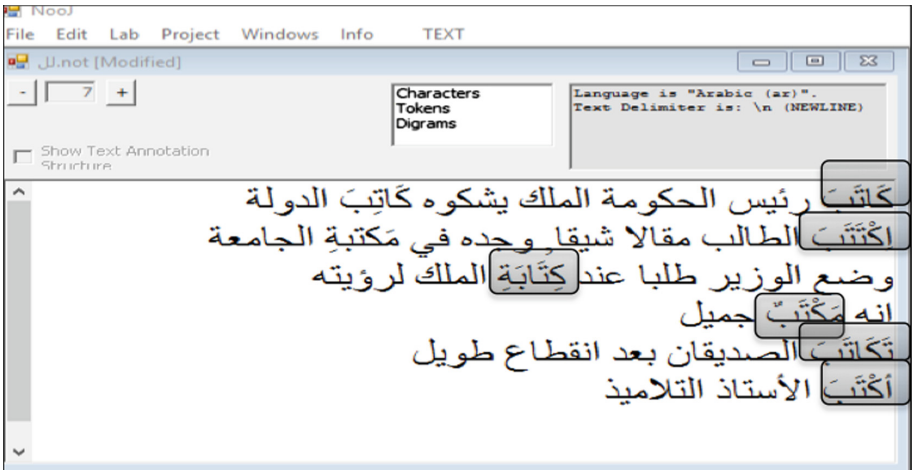

Fig. 4. Text to be analyzed in NooJ platform using EL-DiCar dictionary

The text: The prime minister *wrote* to the king to complain about his State *Secretary*, A student *copies* an interesting article that he found at the University *Library*; the minister submits a demand to the king *secretary* asking for an appointment to meet the king; it is a beautiful desk; two friends *wrote* to each other after a long time; The teacher *dictates* to his students. Figure 5 shows the "locate pattern" or the search wizard in NooJ platform [8]; the lemma (to write – KaTaBa- كَتَبَ) should return all inflectional and derivational forms of this entry. Note that NooJ Platform allows us to make a search with full, semi or even without diacritics [8].

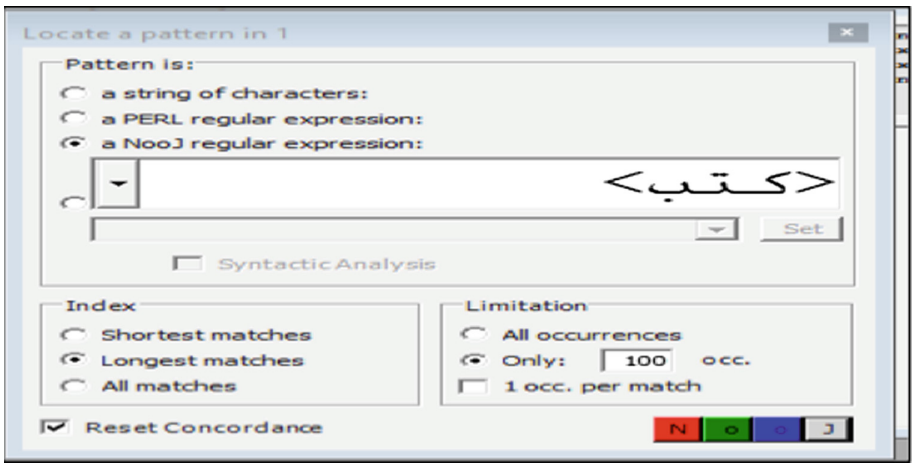

Fig. 5. Lemma search in NooJ platform using DiCar dictionary-EL

In the result of Fig. 6, only the derivational form (active participant) of the previous entry is retrieved, although the text contains seven words that share the same root [KTB-كتب].

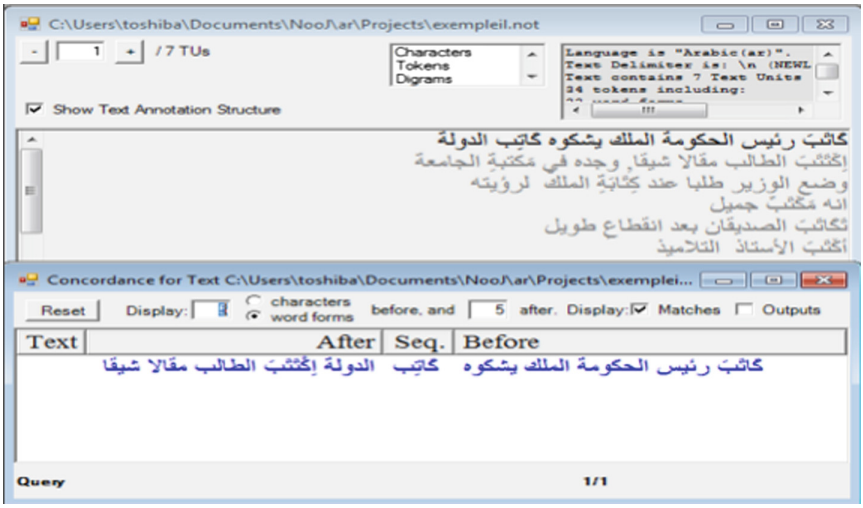

Fig. 6. Lemma search result in NooJ platform using dictionary DiCar-EL

Figure 7 shows the search operation using the root (KTB-كتب) on the previous text, using our dictionary:

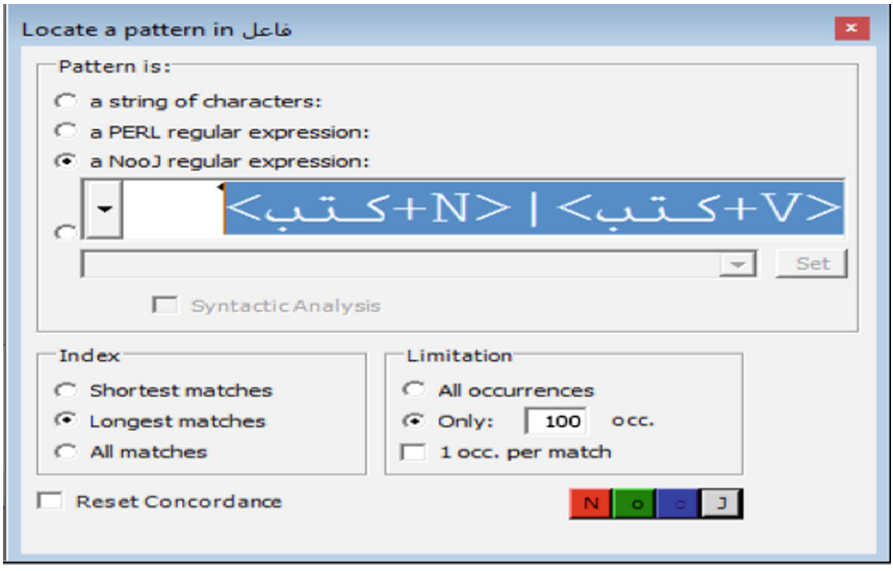

Fig. 7. Root search in NooJ platform

The query above retrieves all verbs and nouns that have the root KTB. The result is as follows:

All entries with their inflectional and derivational forms that share the root KTB have been retrieved, as it is shown in Fig. 8.

Fig. 8. Root search result in NooJ platform using our dictionary.

While patterns care about meaning, we can also apply a pattern search in a text, to retrieve all words that share the same concept. For instance, the pattern [FiAaLa – فِعَالَة] that refers to a craft will retrieve all crafts that occur in the text that Fig. 9 shows:

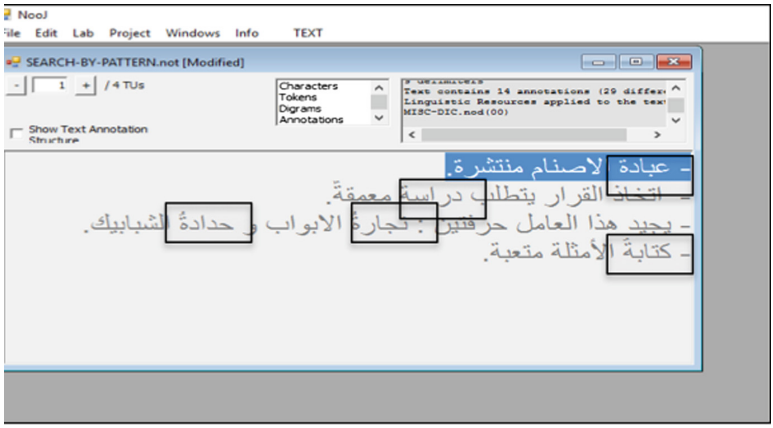

Fig. 9. Text to be analyzed in NooJ.

- The *worship* of idols is widespread.
- Decision-making requires an in-depth *study*.
- The worker mastered two craft: wood *carpentry* and *wild fences*.
- *Writing* examples is exhausting.

As we can see, the text contains five words that share the same pattern [FiAaLa – فِعَالَة]. Making a search using this pattern will retrieve all crafts that occur in the text. Figure 10 shows the search operation using the previous pattern.

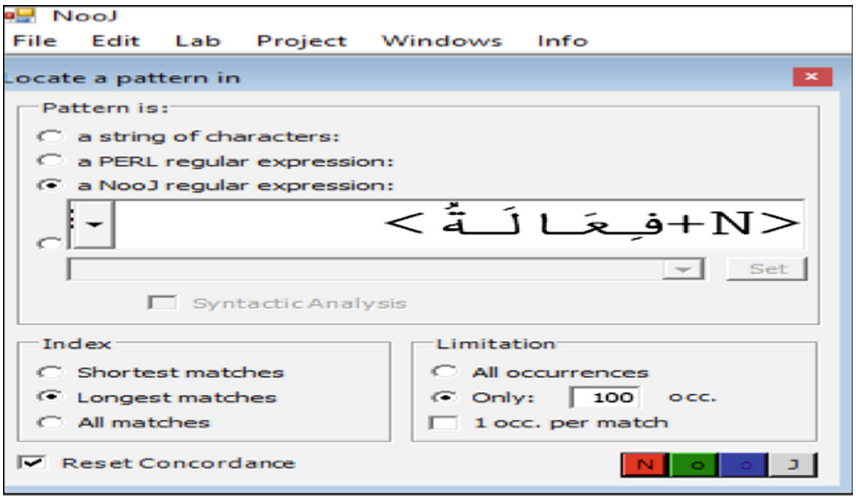

Fig. 10. Pattern search.

The result is as follows (Fig. 11):

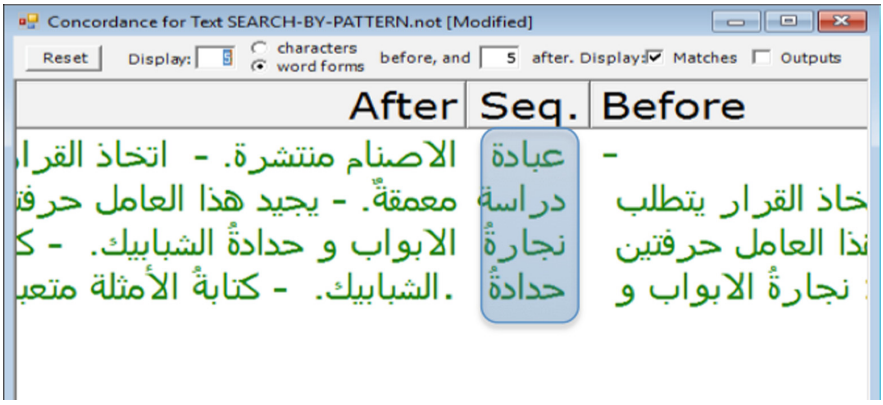

Fig. 11. Pattern search result.

We can retrieve also all nouns that refer to an actor using the pattern [FaAiLفَاعِل] or all nouns that refer to an action or a place using their patterns.

Text annotations take the form shown in Fig. 12. The text contains 213 inflectional forms of the verb (to store – KhaZaNa - خَزَنَ); its root is [KHZN-خزن], voice = passive, tense = time = I, while I means perfect, subject agreement = 1st person.

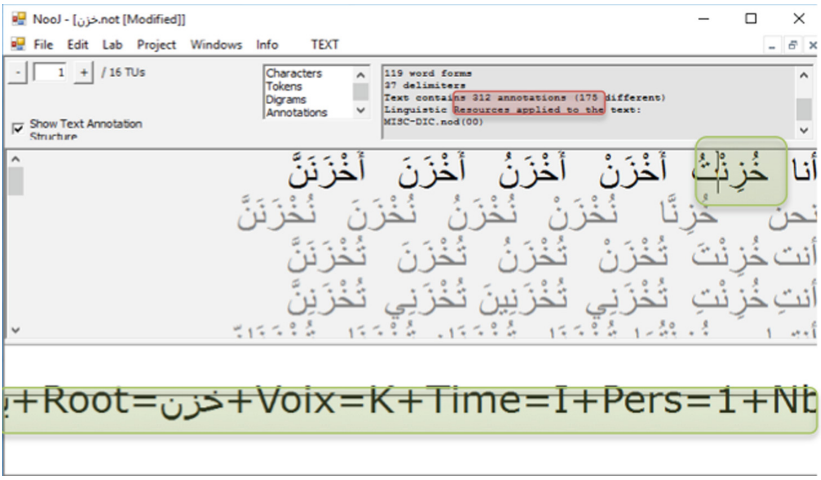

Fig. 12. Text annotation form.

Lexical grammar takes the same hierarchy of the verb classification, in NooJ platform; Fig. 13 gives an example of a lexical grammar of the regular category. Yellow color means that the case includes is a sub-graph.

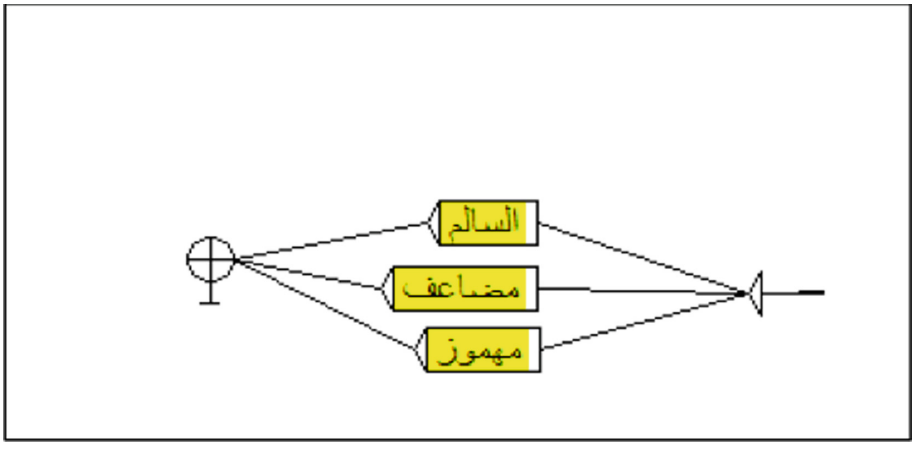

Fig. 13. A sub graph of the regular category

As we have mentioned before, this category contains tree sub categories: sound-duplicated-hamzated, (see Sect. 3.2).

An example of a sub-graph for the duplicated category, as Fig. 14 shows: (FaaAaLa-فَاعَلَ); (taFaAaLA-تَفَاعَلَ) and (inFaAaLa- انْفَعَلَ): several sub-graphs that generate all possible inflectional and derivational forms of a given duplicated root;

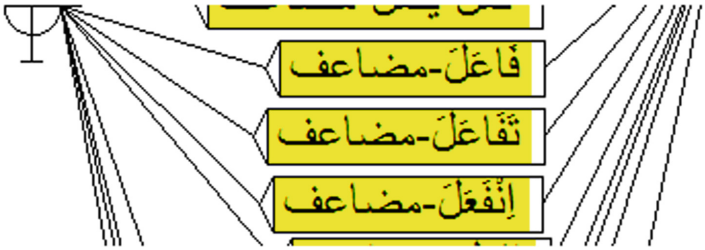

Fig. 14. All possible patterns for each duplicated root

Finally, the grammar takes the form that Fig. 15 shows. Each of these operators serves to reform the root in the inflectional or the derivational form. (For more operators see NooJ Manuel).

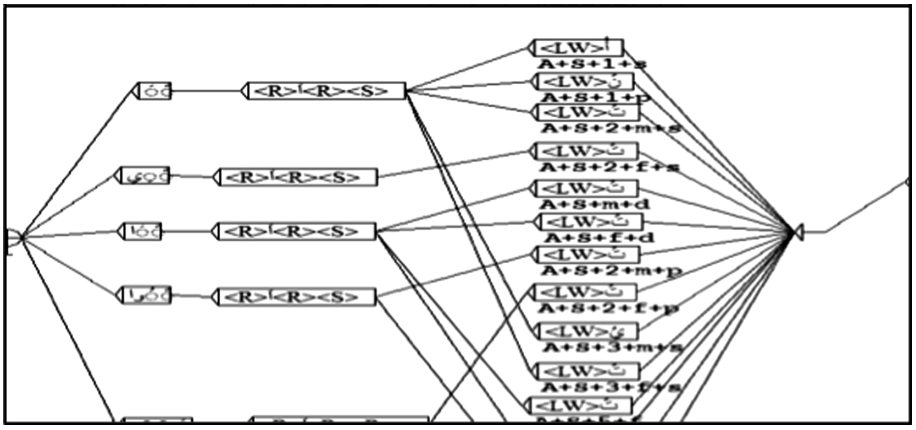

Fig. 15. The possible inflections in the jussive case of a duplicated root

5 Conclusion and Perspectives

Our dictionary that is based on root and pattern approach contains 14500 entries, generated from 295 verbs models. These models cover all Arabic verb categories, each entry contains all possible inflectional and derivational forms, that are generated using our grammar. The adopted approach allows us to extract all words within a text that

share the same root. We can also extract concepts using a pattern search. The result of the linguistic analysis of the text appears as annotations that give all morphological details about the analyzed text words. In future work we will cover the other Arabic words like nouns and adjectives. Also adding Arabic morphophonological alternation as new grammars.

References

1. Mourchid, M.: Génération morphologique et applications. Thèse de Spécialité de 3ème Cycle, Université Mohammed V, Juillet (1999)
2. Slim, M.: Standard Arabic formalization and linguistic platform for its analysis. In: The Challenge of Arabic for NLP/MT, LASELDI, Franche-Comté University, France (2006)
3. Nabil, A.: Arabic language and computer. Ta'areeb (1988). (in Arabic)
4. Nooj association. http://www.nooj-association.org/index.php?option=com_k2&view=item&id=2:arabic-resource&Itemid=611
5. Karin, C.: A Reference Grammar of Modern Standard Arabic. 2nd edn. Cambridge University Press, New York (2005)
6. Antwan, D.: A dictionary of universal Arabic grammar. Lebanon liberary, dar el nacher wa almaajem (1999). (in Arabic)
7. Slim, M.: Analyse morpho-syntaxique automatique et reconnaissance des entites nommees en arabe standard. Specialty thesis of 3rd round, Doctoral thesis, LASELDI, Franche-Comté University, France (2008)
8. Silberztein, M.: Nooj Manual. www.nooj-association.org