International Journal of Computer Science and Applications,
©Technomathematics Research Foundation
Vol. 12, No. 1, pp. 94 – 107, 2015

AMELIORATION OF THE INTERACTIVE DICTIONARY OF ARABIC LANGUAGE

REQQASS Mohammed, LAKHOUAJA Abdelhak, MAZROUI Azzeddine

Laboratory of computer sciences, University Mohammed First, Faculty of sciences,
Oujda, Morocco
{reqqass.mohammed,abdel.lakh,azze.mazroui}@gmail.com

ATIH Idriss

QatarUniversity,
Doha,Qatar
idrissatih@yahoo.fr

Despite the recognition of the Arabic language by the United Nations and its active development, there are no powerful interactive dictionaries to accompany efficiently this development. In addition, most of the existing dictionaries require knowledge of morphological rules to get the meaning of words.

We studied in this paper the “Interactive Dictionary of Arabic Language”. This is, in our knowledge, the only open source dictionary project that provides morphological information in addition to the meaning of the input Arabic word. By analyzing the dictionary, we noticed that it has a great weakness in the morphological analysis of the input word. Thus, we have improved the morphological analysis step by developing a module based on the lemma extraction step of the system “Al-Khalil 2”. In addition, we expanded the database with the words available in the “Almusotako$if” dictionary.

Keywords: interactive dictionary; database; morphological analysis.

1. Introduction

A dictionary is a collection of words in one or more specific languages, often listed alphabetically (or by radical and stroke for ideographic languages), with usage of information, definitions, etymologies, phonetics, pronunciations, translation, and other informations [Webster's New World College Dictionary, (2002)]. Its construction is based on two principles [Youssfi, (2006)]:

- the comprehensiveness: the dictionary must contain almost all words in the language;
- the ordering: the words in paper dictionaries must follow a certain order to allow users an efficient, fast and universally known consultation method. This principle can be ignored in the case of electronic dictionaries that allow you to use advanced research methods.

According to [Selva, (2004)], several experiments have shown that the standard dictionaries do not improve reading comprehension significantly especially for users not having an advanced knowledge of the language. For this, it has become essential to build electronic dictionaries with the following features:

- advanced search: allow the user to have the meaning of a word according to its various morphological forms;
- interactivity: the user must be able to customize the display and specify the information to display;
- renovation: in a living language, many new words can appear while others are no longer used or change their meaning. Therefore, the dictionary must be easily updated.

This paper is structured around three axes. In the first one, we present the classic dictionaries and recall different experiences taken to build them. In the second axis, we show the operating principles of dictionaries, and we present in the last axis the approach that we have developed to improve the system of “Almuajam” [Redbawi *et al.*, (2011)].

2. Classic dictionaries

The Arabic dictionaries have experienced a noticeable development in terms of collection of lexical entries and their organization. In fact, early dictionaries were built based on "language-mails". However, the new Arabic lexical experiences have included several new words that do not necessarily respect the criteria of the Arabic morphology. Unlike those experiences that target people mastering Arabic language, Lebanese publishers have built dictionaries for students and researchers. They introduced foreign words in their works [Khrish *et al.*, (2013)].

These classical Arabic dictionaries can be classified under four schools according to the organization of lexical entries:

- “AlEayn” (العين): This dictionary organizes the lexical entries according to the voice output letters. Thus, the first letter of the dictionary is the letter “E” (ع) while the last is “m” (م) [Bin Ahmed Al Farahidi, (1984)].
- “AljamHharat” (الجمهرة): This dictionary organizes the lexical entries according to alphabetically Arabic letters [Bin Durai, (1987)].
- “AlSiHaH” (الصحاح): This dictionary organizes the entries under chapters whose names are the letters of the Arabic alphabet. Each chapter consists of one or more sections that have as names the letters of the Arabic alphabet. Each section include the words derived from the roots beginning with the name of the section and ending with the name of the corresponding chapter [Al Fairouz, (1999)]. For example, the word “البعث” (the resurrection) whose root is “b E v” (ب ع ث) is in the section “b” (ب) of chapter “v” (ث).
- “Al>sas” (الأساس): This dictionary organizes entries under chapters. For example, the chapter “b” (ب) contains all the words derived from roots beginning with the letter “b” (ب). The roots are afterward sorted alphabetically [Al Zamkhrashari, (1998)]. For example the word “كثير” (many) which the root is “kvr” (ك ث ر) is in the chapter “k” (ك).

In spite of the improvement of the classical Arabic dictionaries, they are not regularly updated. So, they do not contain all the information that can help the user to understand the word's meaning. In addition, their use requires an advanced knowledge of the Arabic language.

Thus, it became necessary to develop electronic dictionaries which are able to follow the permanent evolution of the Arabic language, and facilitate access to the meaning of a word without necessarily knowing the grammatical rules of the language.

3. Electronic Dictionaries

In addition to the features that a classical dictionary can offer, the electronic dictionaries provide a formal representation of a lexicon, associating with each form its lemma and some grammatical, inflectional and semantic information.

Among the main advantages of an electronic dictionary [Rebdawi *et al.*, (2011)]:

- an easy update of the dictionary entries;
- an easy access to an entry: indeed, the research in a classical dictionary is done by lemma of the word (infinitive verb or singular noun), while in an electronic dictionary we can make a search by a derived form of a word;
- a quick navigation between the dictionary entries (synonym, antonym, ...);
- provide more information than a classical dictionary (morphologic information, examples, ...);
- keep the search's history and specify the amount of information that the user wants.

We quote below the electronic Arabic dictionaries that we have studied:

- "Albahit AlEaraby" (الباحث العربي) [dictionary Albahit AlEaraby]: this is an electronic dictionary based on five classic dictionaries ["lisan AlEarab"(لسان العرب), "maqayis Aluga" (مقاييس اللغة), "AlSiHah fi Aluga" (الصحاح في اللغة), "Alqamus AlmuHit"(القاموس المحيط), "AlEabab Alzaxir"(العباب الزاخر)]. It allows:

  the research in all five dictionaries that it includes;

  the specification of dictionaries that gave the information following the request made by the user.

The major disadvantage of this dictionary is that the search of a word is done in both tables: the table of the entries and that of the meanings of these entries. For example, for the word كتب (write) it returns the meaning of the words الكِتابُ (the book) and الحَبْل (the cord), …

- "AlmaEani" (المعاني) [dictionary AlmaEani]: this is an electronic dictionary based on several classic dictionaries ["Al>E$ab" (الأعشاب), "Alra}id"(الرائد), "Algany"(الغني)…]. It allows:

  the research in all the dictionaries that it includes;

  the specification of dictionaries that gave the information following the request made by the user;

  the search of a word with or without vowel.

- “Almuajam Alwasyt” (المعجم الوسيط) [dictionary Almuajam]: this is an electronic dictionary based on the classic dictionary that bears the same name. It allows the research:
  by the first letter (the first alphabet) of a word;
  of a word with or without vowels ;
  of a word from the dictionary entries and / or in texts giving the meaning of words.

These three dictionaries do not allow searching by the derived form of an entry. For example, they do not return any result for the word مخطوطات (manuscripts), which requires knowledge by the user the of the word's lemma before searching.

- “Almuajam”: this is an electronic dictionary based on “Alwassyt” dictionary. It allows the research:
  by word or root;
  depending on the type of the word (verb, noun, etc.).

This dictionary is distinguished by the possibility to integrate new resources and the richness of offered functions. However, searching for a word in all its morphological forms is not assured (for example: for the word لاعب (player) the dictionary return the meaning but does not give the one of the word لاعبتين (two players)). In the next section, we will give its detailed description.

4. “Almuajam” dictionary

The “Interactive Dictionary of Arabic Language Almuajam” (Almuajam) is an interactive open source web application downloadable from the web site “www.sourceforge.net”. It allows the basic functions that should provide an interactive dictionary. In this section, we describe its data base as well as the different stages of searching for a word.

4.1. *Database structure*

The database of “Almuajam” dictionary contains 65 tables that we can be grouped in three classes:

- morphological information: these tables contain different morphological information that a word can have in Arabic language (root, pattern, type, etc.);
- the semantic entry : each entry is associated with a set of information to identify its semantic field, its degree of use and geographical location of its appearance;
- the lexical entry: each lexical entry possesses meanings and examples.

4.2. *Research method*

The “Almuajam” dictionary allows searching by root or by entry. We describe below the various stages of research followed by the system:

(1) the user enters the search word;
(2) he specified the type of research (by root or by entry);

(3) if the search word is among the dictionary entries, the system:

(i) returns, if the search is by entry, all entries in the dictionary (nouns, verbs, ... etc.) corresponding to the word input by the user;

(ii) and it returns, if the search is by root, all dictionary entries derived from the root seized by the user;

(4) if the search word does not exist then the system perform a morphological analysis to determine the lemma:

(i) if the lemma is among the dictionary entries, the system redone the search using this lemma.

(ii) otherwise, the system will perform spell checking and offers a set of words.

The "Almuajam" dictionary is the most effective and easiest to use among the various existing free Arabic monolingual system. However, it has some shortcomings. The main ones are:

- since in the case of interactive dictionaries, research is done in most cases by inflected form, "Almuajam" dictionary accesses twice to its database. Indeed, the first access consists in searching the entry in its inflected form (e.g. مدرستين: two schools, الكتاب: the book). This inflected form will not be found in the database because the latter contains only lemmas. So, the system performs a morphological analysis to identify the lemma, and resumes the search in the database with this lemma;
- the extractor module of lemma in the "Almuajam" dictionary is a simple segmentation (removal of prefixes and suffixes). This module does not always give the lemma of the word (e.g. it returns no results for the word لاعبتين: two players). In other cases, it returns wrong results (e.g. it returns the word "mabAlig" (مبالغ: amounts) for the word "AlmubAlagat" (المبالغات: the exaggerators) instead of lemma "mubalig" (مبالغ: exaggerator).
- an inflected form can be derived from several lemmas (e.g. the word كتاب may represent the lemma "kitAb" (كتاب : book) and it may represent the inflected form of the lemma "tAba" (تاب : repent) where the letter "k" (الكاف) is considered as a prefix). For this example, the "Almuajam" system only treats the lemma "kitAb" (كتاب : book).
- multiple words are not treated because the database used does not cover all Arabic words.

To remedy these shortcomings, improvements must bear on the one hand on the database that must be enriched (by entries, meanings, examples, etc.) and on the other hand its morphological analysis system must be improved because it often does not identify the lemma of a word.

Our work consists in enrich the resources of the "Almuajam" dictionary by the entries of the "Almusotako$if" dictionary, and also improve the morphological analysis step by using the analyzer "Al-Khalil" [ould Abdallahi Ould Bebah, (2010)].

5. Operating AlMostakchif dictionary

“Almusotako$if” (المستكشف) is presented in [Atih, (2011)] as the first Arabic lexical encyclopedia of the dictionary “AlmuHit” (المحيط) that is classified.

Unlike other classical Arabic dictionaries interested only by the maintenance and the documentation of Arabic language, this dictionary organized the lexical information according to its semantic domain in order to exploit the richness of the Arabic language.

5.1. *Structure “Almusotako$if” in a XML file*

To organize and use the data of “Almusotako$if” dictionary, we generated a XML file from the Word version of this dictionary. We defined entities, fields and their relations as follows:

- the tag ‘scope’: contains the field ‘name’ which allows grouping words that have the same semantic field;
- the tag ‘words’: has the field ‘name’. This field contains the word;
- the tag ‘meaning’ that contains the definition of the word and a set of associated words having a semantic relationship with the main word;
- the tag ‘part’: has the ‘name’ field that contains the name of the part;
- the tag ‘word’: has the field ‘name’ that contains the word and the ‘meaning’ tag that contains the definition of the word.

We can schematize this structure with the following graph:

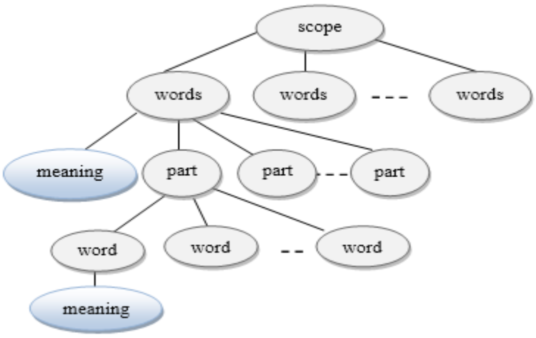

Fig. 1. XML schema of “Almusotako$if”.

Below is an excerpt of the XML file that we have built:

```
<scope name='السوائل'>
<words name='الماء'>
    <meaning>
    سائل شفاف بلا طعم أو رائحة، مكون كيميائيا من (ذرتين من الهيدروجين وذرة واحدة من
    الأكسجين)، وهو أهم السوائل المشروبة، ترتبط به حياة الإنسان وجميع الكائنات الحية على
    الأرض.
    </ meaning>
    <part name='أنواعه ومتعلقاته '>
        <word name='الجُلاَّبُ'>
            <meaning>ماءُ الوَرْدِ</ meaning>
        </word>…
    </part>…
</words>…
</scope>…
```

This step allowed us to clear the Word version of “Almusotako$if” and show the possible links between different dictionary entries.

5.2. *Operating “Almusotako$if” dictionary*

To extract data from XML file and to keep the database’s structure of “Almuajam” dictionary, we proceeded as follows:

(1) extract all the words and their definitions from the generated XML file;

(2) analyze words using the “Al-Khalil 2” analyzer in order to get the different morphological information for each word;

(3) keep the words that do not exist in the database of “Almuajam” dictionary and for which “Al-Khalil 2” returns only one result (to avoid ambiguity);

(4) extend the “Almuajam” dictionary database by the new morphological information and the new words. Each entry is added as follows:

(i) Check if the morphological information associated to the word exist or not in the database of "Almuajam" dictionary. In the case where these information do not exist, we add them in the database;

(ii) add and link the word to the key of different associated morphological information;

(iii) add a semantic entry associated to the word ;

(iv) add the meaning of the word.

This approach allowed us to extend the database of the “Almuajam” dictionary by 3588 new words obtained from the “Almusotako$if” dictionary.

6. Improvement of the system of morphological analysis

As the interactive dictionaries have to look for the word in all its morphological forms, they require a morphological analysis of the word.

We noticed from our study of “Almuajam” dictionary that its morphological analyzer cannot deal with all the morphological forms derived from the word. This led us to seek to improve the morphological analysis stage of a word. Our approach is described in the following paragraphs.

6.1. *Construction of a set of tests*

To identify the different classes not treated by the “Almuajam” dictionary, we built a corpus of words derived from different classes of roots. We show in the following tables the different roots randomly selected.

The table below contains some healthy roots (الجذور الصحيحة).

Table 1. Healthy roots

Saalim (سالم)	**modaaf** (مضعف)	**Mahmooz al fae** (مهموز الفاء)	**Mahmooz al ain** (مهموز العين)	**Mahmooz al lam** (مهموز اللام)
بتر - btr	صدد - Sdd	ءبد – 'bd	بءج – b'j	نبء - nb'
بثر - bvr	ضخخ - Dxx	ءبق – 'bq	تءم - t'm	ملء - ml'
بذر - b*r	طبب - Tbb	ءتن – 'tn	ثءر - v'r	لجء - lj'
جرس - jrs	عمم- Emm	ءرج – ‘rj	مءق - m'q	قنء - qn'
حجز - Hjz	فخخ - fxx	ءزم – 'zm	لءم - l'm	كلء - kl'
خدع - xdE	قضض - qdd	ءزق – 'zq	كءب - k'b	قرء - qr'
درس - drs	لتت - ltt	ءكم – 'km	فءر - f'r	صدء - Sd'
رسم - rsm	هدد - hdd	ءكل – 'kl	ضءن - D'n	طرء - Tr'
زرع - zrE	بشش – b$$	ءمر – 'mr	شءم - $'m	ظمء - Zm'
سعل - sEl	حسس - Hss	ءنب – 'nb	سءل - s'l	فتء - ft'

The table below contains some weak roots (الجذور المعتلة).

Table 2. Weak roots

Assimilated (معتل الفاء)	**Hollow** (معتل العين)	**Defective** (معتل اللأم)	**Separate mixed**	**Assimilated** (معتل الفاء)
وهب - whb	بوج -bwj	بدو - bdw	وشي – w$y	نوي - nwy
يقظ - yqz	توب - twb	بزو - bzw	وطء - wT'	مءي - m'y
يهم - yhm	جوب - jwb	تلو - tlw	وفي - why	لءي - l'y
يسر - ysr	خيب -xyb	جثو -jvw	وكء - wk'	فيء - fy'
يبس - ybs	دين – dyn	حدو - Hdw	وهي - why	غوي - gwy
ومض - wmD	ريب - ryb	حوي - Hwy	يدي - ydy	عوي - Ezy
ولد -wld	ريق - ryq	خشي –x$y	ءبو - 'bw	طوي - Twy
ولج - wlj	زيت - zyt	دري - dry	ءدو - 'dw	شيء - $y'
وكب -wkb	سيب - syb	رحي - rHy	ءخو - 'xw	سوء - sw'
وقد -wqd	شيد - $yd	سبي - sby	ءثو - 'vw	زوي - zwy

Each root is associated with a set of their derived words. These words are stored in a file as follows (we have adopted the signs used in the analyzer “Alkhalil 2”):

We remember their meanings in the following form:

Table 3. Codes of used morphological information

indication	Signification	indication	Signification
فا	Active participle اسم فاعل	1	Singular masculine nominative مفرد مذكر مرفوع
مف	Passive participle اسم مفعول	2	Singular feminine مفرد مؤنث مرفوع nominative
مفا	Intensive Active Participle مبالغة اسم الفاعل	3	Dual masculine nominative مثنى مذكرمرفوع
آ	Instrumental noun اسم آلة	4	Dual feminine nominative مثنى مؤنث مرفوع
زمك	Noun of place and timeاسم زمان أو مكان	5	Plural masculine nominative جمع مذكر مرفوع
نك	Indefiniteness نكرة	6	Plural feminine nominative جمع مؤنث مرفوع
إض	Annexed مضاف	7	Singular masculine accusative مفرد مذكر منصوب
فض	Elative noun اسم تفضيل	8	Singular feminine accusative مفرد مؤنث منصوب
وش	Adjective صفة مشبهة	9	Dual masculine accusative مثنى مذكر منصوب
صأ	Gerund مصدر أصلي	10	Dual feminine accusative مثنى مؤنث منصوب
صم	Gerund with initiative mime مصدر ميمي	11	Plural masculine accusative جمع مذكر منصوب
صر	Gerund of instance مصدر مرة	12	Plural feminine accusative جمع مؤنث منصوب
صه	Gerund of state مصدر هيئة	13	Singular masculine genitive مفرد مذكر مجرور
جا	Primitive noun اسم جامد	14	Singular feminine genitive مفرد مؤنث مجرور
صن	Gerund of profession مصدر صناعي	15	Dual masculine genitive مثنى مذكر مجرور
نس	relative نسبة	16	Dual feminine genitive مثنى مؤنث مجرور
		17	Plural masculine genitive جمع مذكر مجرور
		18	Plural feminine genitive جمع مؤنث مجرور

6.2. *Untreated categories*

In this step we determine the categories of words that deals the “Almuajam”, but fails to treat their derivative forms. It consists in:

- extract all vocalized words and their morphological information;
- change the system code of “Almuajam” dictionary to identify entries that the system cannot deal. This step allowed us to treat a large number of words in an automatic way;

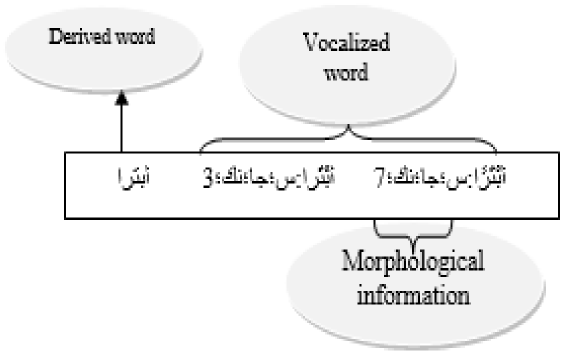

Fig. 2. Extract of file.

- calculate the total number of words in each category and that of untreated words in this category.

To achieve these tasks, we tested the original version of the dictionary on a representative sample in order to identify classes of words with a high percentage of untreated words. The results are given in the table below.

Tables 4. Untreated categories

Categories with high percentage of untreated words	**number of words**	**untreated words**	**Percentage of untreated words**
نسبة – مضاف –جمع مذكر مرفوع relative - annexed - plural masculine nominative	**281**	**279**	**% 99**
اسم فاعل – نكرة – مثنى مؤنث مرفوع active participle – indefiniteness – Dual feminine nominative	**421**	**413**	**% 98**
اسم فاعل – نكرة – مثنى مؤنث مجرور active participle – indefiniteness – Dual feminine nominative genitive	**420**	**414**	**% 98**
اسم فاعل – نكرة – مثنى مؤنث منصوب active participle – indefiniteness - Dual feminine nominative accusative	**420**	**414**	**% 98**
اسم آلة – نكرة – مثنى مؤنث منصوب Instrumental noun – Indefiniteness – Dual feminine accusative	**136**	**133**	**% 97**

We note that the words of some categories are not treated. Therefore, it was necessary to improve the step of morphological analysis.

6.3. *Improvement of the morphological analysis step*

We remind that if the "Almuajam" system doesn't find the searched word among its database of entries, it analyses the word in order to get its lemma and restarts the research with the lemma. If it cannot find the lemma of the word in its database, it offers alternative solutions.

The extraction method of the lemma in this system is based on the possible segmentations of the word in prefix, suffix and base. However, this method does not always give the lemma of the word and not ensure that the returned segment corresponds to the true lemma.

Our task is to develop an extraction module of the lemma based on the program "Al-Khalil 2" [Boudchiche, (2014)] which was developed in LaRI laboratory (the laboratory of computer sciences LaRI Oujda). The method we have developed through the following steps:

(1) the user enters the search word;
(2) the system looks for the word in the database and returns the result when it exists;
(3) the system uses the lemma extractor module, and when it manages to extract one or more lemmas, the system searches them in the database;
(4) if it finds them in its database, it returns the results;
(5) if the system does not find any solution after the search in the database of the word and its potential lemmas, it offers suggestions.

This approach allowed us to reduce the number of accesses to the database as well as the response time of a search performed. It also allows more possible results (example: if you look up the word كتاب our system reconsider the meaning of the word تاب in addition to كتاب while the original system is limited to the first word كتاب).

6.4. *Evaluation*

In order to evaluate the contribution of our new search system, we tested the search results of a word in the two systems. To do this, it is necessary to have a corpus allowing to highlight the relationship of each inflected form with its lemma. After several searches, we have unfortunately not found a standard Arabic corpus that allows us to do this assessment. Thus, we decided to build two test corpus composed of words with their lemma. Since our objective is the evaluation of the research system, we are limited to the lemmas belonging to the “Almuajam” database.

6.4.1. Construction of the test set:

- Quranic corpus: it is based on “The Quranic Arabic Corpus” [Dukes and Habash, (2010)]. This corpus represents the morphological and syntactic information for all the words of the Quran. It is written with Buckwalter transliteration. We give below as an example the representation of the verse « بسم الله الرحمن الرحيم » [Alfatiha, verse 1] :

```
(1:1:1:1)   bi        P     PREFIX|bi+
(1:1:1:2)   somi      N     STEM|POS:N|LEM:{som|ROOT:smw|M|GEN
(1:1:2:1)   {ll~ahi   PN    STEM|POS:PN|LEM:{ll~ah|ROOT:Alh|GEN
(1:1:3:1)   {l        DET   PREFIX|Al+
(1:1:3:2)   r~aHoma`ni      ADJ
      STEM|POS:ADJ|LEM:r~aHoma`n|ROOT:rHm|MS|GEN
(1:1:4:1)   {l        DET   PREFIX|Al+
(1:1:4:2)   r~aHiymi        ADJ
STEM|POS:ADJ|LEM:r~aHiym|ROOT:rHm|MS|GEN
```

To generate “Quranic corpus” we proceeded as follows:

Combine the parts of each word and associate it with its lemma, without yet keep other morpho-syntactic information. Thus, the example above becomes:

bisomi	{som
{ll~ahi	{ll~ah
{lr~aHoma`ni	r~aHoma`n
{lr~aHiymi	r~aHiym

Convert the writing of this corpus in Arabic characters; So we get for the block above:

بِسْمِ	ٱسْم
اللَّهِ	ٱللَّه
الرَّحْمَانِ	رَحْمَان
الرَّحِيمِ	رَحِيم

Keep only the inflected forms whose lemmas exist in the database of “Almuajam” dictionary. Thus, for the block above it keeps only:

الرَّحِيمِ	رَحِيم

The corpus that we have built in this manner contains 38198 couples (inflected form and its corresponding lemma).

- Sarf corpus: it based on the data base of Sarf - Arabic Morphology System [Sarf]. We recall that “Sarf” is a system of derivation and conjugation of Arabic words. It allows generating the Arabic words from a root. To construct a test set using this system we proceeded as follows:

Generate all inflected forms from the roots that exist in the database “Almuajam” dictionary. We than grouped the different inflected forms according to their lemmas. Below is an excerpt from the generated file relating to lemma ولج:

L.وَلَجَ
وَلَجَ وَلَجْتُنَّ وَلَجْتُمْ وَلَجْتُمَا وَلَجْتِ وَلَجْتَ وَلَجْنَا وَلَجْتُ
تَلِجْنَ تَلِجُونَ تَلِجَانِ تَلِجِينَ تَلِجُ نَلِجُ أَلِجُ
أَلِجَ يَلِجْنَ يَلِجُونَ تَلِجَانِ يَلِجَانِ تَلِجُ يَلِجُ

Keep only the sets which their lemmas belong to “Almuajam” database.

Given the large size of the generated corpus, we are limited to 40000 word couples (inflected form and associated lemma) representing the various categories of Arabic words.

6.4.2. Evaluation Method of the two systems:

To evaluate the two systems, we considered that a system succeeds the process of searching for a word if the lemma corresponding to this word is one of the returned results. We give in the Table 5 the success rate of each system:

Tables 5. Statistics of result

	success rate in Quranic corpus	**success rate in Sarf corpus**
Almuajam dictionary	59%	55%
Our system	89%	100%

We note that the performance of our system compared to those of “Almuajam” dictionary have improved by 29% and 40% respectively on the Quranic corpus and Sarf corpus. This performance is due to the quality of the morphological analyzer that we have integrated. However this analyzer must be improved to better analyze the 11% of the words for which the corresponding lemma is not one of the outputs of the morphological analysis. The perfect rate of 100% that we obtained on the corpus Sarf can be explained by the fact that the database “SARF” system is a part of that of the morphological analyzer “Al-khalil 2”.

In addition, we sought to compare the coverage rate of the two systems (given a word, verify if the systems give the set of all possible solutions) . So, we conducted another experience in order to evaluate the average number of solutions returned by each system. For this purpose, we conducted this experience on the words recognized by the two systems in the previous evaluation. We selected among these words a test set containing 1322 non-vocalized words. We found that our system returns 5394 results for this set which represent 4.08 results per word, while “Almuajam” dictionary returns 4298 results which means 3.25 results per word.

We therefore conclude that changes that we brought significantly improve the coverage rate.

Conclusion

Given the growing use of computer systems, the development of an interactive dictionary for the Arabic language has become a major need. The exploitation of the computer power in terms of storage space and data processing makes possible an easy, fast and accurate use of interactive dictionaries.

In this paper, we introduced the “Almuajam” dictionary. We have added entries 3588 new words from the “Almusotakchif” encyclopedia. We also improved the morphological analysis stage by exploiting “Al-Khalil 2” analyzer in order to identify the various morphological forms of the word.

These changes allowed a significant improvement in performance of the dictionary both at the level of accuracy and the cover.

References

Al Fairouz, A. (1999): “muxotar AlSiHaH” dictionary (مختار الصحاح). 5st edition. Beirut, Leban.

Atih, I. (2011): “Almusotako$if” dictionary (المستكشف). Publication of the Islamic Educational, Scientific and Cultural Organization (ISESCO).

Al Zamkhashari, J. A. M. (1998): “Al>sas Albalagap” dictionary (أساس البلاغة). 1st edition, Beirut, Leban.

Bin Duraid, A. (1987): “AljamHharat Allugap” dictionnaire (جمهرة اللغة). 1st edition. Beirut, Leban.

Bin Ahmed Al Farahidi, A. (1984): “AlEiyn” book (كتاب العين), Bagdad, Irag.

Boudchiche, M.; Mazroui, M.; Ould Abdallahi Ould Bebah, M.; Lakhouaja, A. (2014): “L’analyseur Morphosyntaxique Alkhali Morpho Sys 2” JDILA’14, 8 Fevrier 2014, Rabat, Maroc.

Dictionnaire « Al Bahet Al Arabi» : http://www.baheth.info, (last visited 25/10/2014).

Dictionnaire « Al-Maani » : http://www.almaany.com (last visited 25/10/2014).

Dictionnaire « Al-Mujam Al Wasset »: http://kamoos.reefnet.gov.sy, (last visited 25/10/2014).

Dukes, K.; Habash, N. (2010) : Morphological Annotation of Quranic Arabic. The Language Resources and Evaluation Conference (LREC 2010), Malta, 2010.

Khrisha, E H.; Al-Maqableh, M. A.; As'sud Al-Mizeed, K. M. (2013): Efforts of the Lebanese Lexicographers in Authoring and Developing the Arabic Lexicon. International Journal of Business and Social Science. Vol. 4 No. 12, Special Issue – September 2013.

Ould Abdallahi Ould Bebah, M.; Boudlal, A.; Lakhouaja, A.; Mazroui, A.; Meziane A.; Shoul, M. (2010): Alkhalil Morpho Sys: A Morphosyntactic analysis system for Arabic texts. ACIT'2010.

Rebdawi, G.; Ghneim, N.; Desouki, M.; Sonbol, R. (2011): An Interactive Arabic Dictionary. in Proceedings of the 2011 IIT Conference, UAE.

Sarf, morphological system : http://sourceforge.net/projects/sarf, (last visited 25/10/2014).

Selva, T.; Chanier, T. (1998): Apport de l’informatique pour l’acces lexical dans les dictionnaires pour apprenants : projet Alexia. Fontenelle, T. Introduction: Dictionaries, Thesauri and Lexical-Semantic Relations. International Journal of Lexicography. vol. 13, no 4, page 229-231.

Webster's New World College Dictionary, Fourth Edition, 2002.

Yousfi, A. (2006): Traitement automatique de la langue Texte et parole. Page 10-11. Edition et impression Bouregreg, Rabat 2006, ISBN : 9954-423-98-2.

Zaafarani, R. (2004): un dictionnaire électronique pour apprenant de l'arabe (langue seconde) basé sur corpus. JEP-TALN 2004. Traitement Automatique de l’Arabe, Fès, 20 avril 2004.