

Towards Arabic Electronic Dictionary

Oumayma Al Dakkak, Ahmad Zein
Electromechanical department
HIAST
Damascus-SYRIA
odakkak@hiast.edu.sy

Abstract— Textual processing involves morphological, syntactic and semantic analysis. Many unilingual dictionaries exist for many languages. However, Arabic has no such an electronic one. Besides, to look for a word in an Arabic-Arabic dictionary is rather a tedious job; one has to know the root of the word if it does exist, know the origin of "Alef" and the origin of "Hamza",...

A pioneer work has been done in HIAST, where the data from "Al Wasseet Dictionary" has been computerized in a form of a Database. Other fields have been added by Arabic linguists [1] to enrich the dictionary and help in morphological, syntactic and semantic analysis. Data from other dictionaries has also been added.

In a second step, a Graphical User Interface is built, enabling the search of a word without the need of human prior processing. The work can make the core for An Arabic Electronic Dictionary, with the help of other linguists from other Arab countries to complete the dictionary and update it regularly.

Keywords- Electronic Arabic Arabic Dictionary; Semantics; Morphological Analysis; Pruning

Introduction

In November 2006, an Expert meeting was held, in the Arabic Academy in Damascus, which gathered Experts from the Arabic countries and from the world, interested by launching "The Arabic Electronic Dictionary" project. The aim is to build a unilingual Arabic dictionary with the following characteristics:

- Oriented to people who speak and learn Arabic.
- Open source, working on all platforms.

The lexical terms are taken from Al Wasseet and other dictionaries, and from contemporary Arabic writings. It must incorporate treasures of synonyms and opposites and Idioms, in addition to lexical and semantic information about the terms.

In fact, such a dictionary is not available, and when we search for electronic Arabic dictionaries we find rather bilingual dictionaries, having Arabic as one of its languages [2], [3] and [4].

A lexical database has already been built in HIAST. It has all the data of Al Wasseet dictionary, and from other dictionaries. The data is structured and completed by Arabic linguists. This work can make the core of the fetched

dictionary. In the present paper, we will present the structure of the data base, followed by a strategy to search the meaning of a root or of a word according to its root. Then we present a pruning algorithm of extended search, to search for similar words which are not explicitly present in the database. A conclusion will end the paper showing the future applications that can be undertaken with the help of the present project.

I. THE LEXICAL DATABASE

In spite of the presence of 29 lexical databases in the region; we don't know a free lexical database for researchers and applications. The goal is to build a database gathering the information of "Al wasseet dictionary", in a structured form, using Microsoft Access software, and then to use it with the GUI to form a kernel of an Electronic Arabic Dictionary and then to use it to build more sophisticated Arabic natural processing applications such as advanced text-to-speech, talking dictionary, continuous speech recognition and dialog systems.

The database is formed of sixteen tables, covering the morphological categories in Arabic and related information: verbs, nouns, infinitives, plural of nouns, particles, special combinations (idioms), examples of use ...etc.

In fact, the 16 tables are divided in verb tables, noun tables, idiom tables, particles tables and additional two tables.

Verb tables include: verbs entries, non standard infinitive forms for trilateral verbs only, verbs examples table, analogous (assimilate) adjectives, analogous (assimilate) adjectives plural, exaggeration forms of active participle, exaggeration forms of active participle plural, non standard cases of adjectives with 'faiil' pattern, meaning passive participle or meaning active participle.

The nouns tables include: nouns entries, non standard plural, noun examples. In addition to that, we have the idioms table, the instrumental nouns, the sound nouns, and verbal nouns. Figure (1) shows the relationship between different tables. Each table consists of a different number of appropriate fields. Table (1) shows the sixteen tables with their corresponding fields. The total number of the studied items exceeds 200 thousands.

Concerning the fields in the tables, the spreading tag equals 1 for common items and equals 2 for less popular ones (according to Arabic specialists). The verb pattern "Wazn" is a digit between 1 and 6 indicating the morphological category of trilateral verb, according to the mid-vowel in the past tense and in the present tense, while for non trilateral verbs, we give the

explicit pattern. Only verbs in the past tense are present in the database, other tenses can be easily deduced from any morphological synthesizer.

TABLE I. THE TABLES OF THE LEXICAL DATA BASE.

Tables	Nb. Of fields	fields
verbs	10	a key, spreading tag, root, the verb, pattern, transitivity, the subjects, the objects, the prepositions that follow, the semantic content.
Non standard Infinitives	5	a verb key, spreading tag, the infinitive, pattern, gender.
Verbs examples	2	The verb key, the example
Analogous adjectives	6	The verb key, spreading tag, adjective sub-key, the adjective, its pattern, its gender.
Analogous adjectives plural.	5	The verb key, spreading tag, adjective sub-key, the adjective plural, its pattern
Exaggeration forms of active participle.	6	The verb key, spreading tag, adjective sub-key, the exaggeration form, its pattern, its gender.
Exaggeration forms of active participle plural.	5	The verb key, spreading tag, adjective sub-key, the exaggeration form plural, its pattern
Adjectives with 'faail' pattern, passive.	3	the verb key, the spreading tag, and the adjective itself.
Adjectives with 'faail' pattern, active.	3	the verb key, the spreading tag, and the adjective itself.
Nouns	9	a noun key, spreading tag, root, the noun, its pattern, its gender, its type, its meaning, its origin
Noun plurals	4	The noun key, spreading tag, the plural word, its pattern.
Nouns examples	2	The noun key, the example
idioms	5	A key, root of the word in idiom, the idiom, its meaning, spreading tag
Particles	5	A key, the particle, its function, its meaning, its spreading tag.
Verbal nouns	4	A key, the noun, its meaning, its spreading tag.
sound nouns	4	A key, the noun, its meaning, its spreading tag.

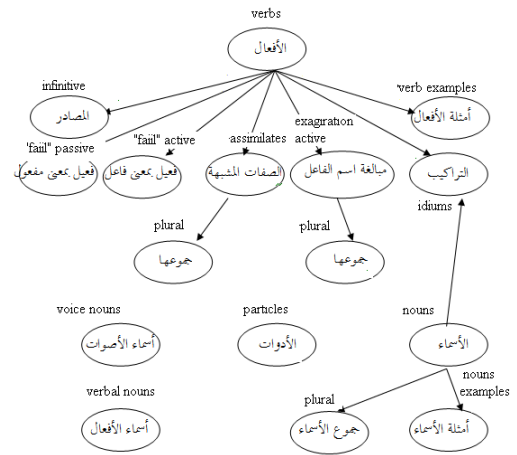


Figure 1. The tables of the database

II. STRATEGY OF SEARCH

We built a Graphical User Interface (GUI), through which we doped three search strategies: According to the root, according to roots and similar words, and according to the word together with the root.

A. According to the root حسب الجذر

The search algorithm is as follows:

- Remove the short vowels, and put "hamza" on the line.
- Search in the Verbs table, all the verbs whose roots are the entered word, sort them according to the verbs, display them with their patterns and give the possibility to show the related semantics.
- Do the same with the Noun table, the Particle table, the Sound Nouns table, and the Verbal Nouns table.
- The search results are displayed in lines, a double click on each line gives the semantic contents of the item. Results of each table are displayed apart; the choice of the table is done through buttons labeled by the table name. Buttons labeled with table names having no results of the actual search are disabled.

Figure (2) shows the results of Verbs table search of the root "Daraba -ضرب", meaning "hit". The columns, from right to left, show the verbs, the roots, the patterns, and the number of items, which verify the same constraints. We also have available results from the Nouns table, while we have no results from the other tables (their buttons are disabled).

A double click, on the first line, for instance, gives the detailed semantics of the 49 entries, one at a time. We use the arrows buttons to move forth and back. Concerning the items in the Verbs table, results from other tables: Analogous, Exaggeration, "faail" are also available when they do exist. (see figure (3)).



Figure (5) Results of Verb table search according to second algorithm for the augmented verb "istaqbala" which means "receives"

C. According to the word and the root حسب الكلمة والجذر

The search algorithm is very much like the first one, where we looked for the word as if it was a "root", then we look for it as if it was a verb, a noun,..etc, discarding the short vowels in these words and during the search process (this is the main difference from the 2nd algorithm).

This algorithm enables us to find the semantic of words with relatively complicated patterns. See example in Figure (6)

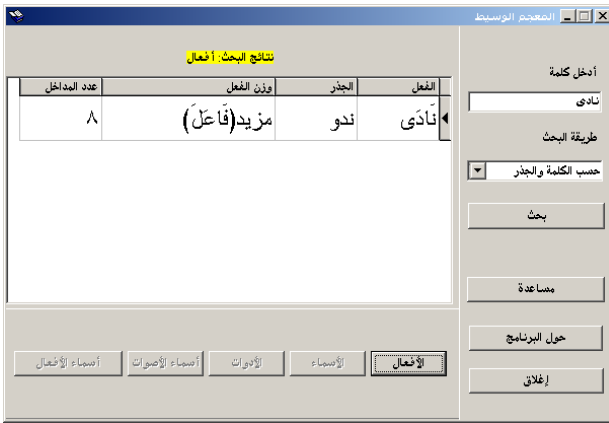


Figure (6) Results of Verb table search according to 3rd algorithm of the augmented verb "naada" which means "calls".

This algorithm eliminates the words which are far in meaning from the entered ones.

III. EXTENDED SEARCH AND PRUNING

To improve the performances of the above algorithms, and let them able to give the meaning of words with prefixes, suffixes, infixes and of transformed words, we applied two pre-processings: search in other sub-tables and an application of a pruning strategy, before searching the meaning of the entered word.

First of all, we look for the entered word, not only in the principle tables, but also in the sub-tables: plural of nouns, infinitives, fails, assimilates with their plurals, and exaggerates with their plurals.

Concerning the pruning process, it has the following steps:

- Eliminate short vowels,
- Apply the extended search on tables and sub-tables,
- Find out if the beginning of the word consists of 3, 2, or 1 prefix [5], [6], [7] (see if the first letter is a possible prefix and if the second letter or the third letter are also possible prefixes). If so, remove the prefixes, apply the extended search on the remaining word and then add the resulting words to a buffer list to look for their meanings.
- Find out if the end of the word consists of 3, 2, or 1 suffixes (see if the last letter is a possible suffix and if the before last letter or the letter in the second position from the end are also possible suffixes). If so, remove the prefixes, apply the extended search on the remaining word and add the resulting words to the same buffer list to look for their meanings.
- Find out if the middle of the word has infixes, remove the infixes one-by-one and add the resulting words to the same buffer list to look for their meanings.
- Look for all the words in the buffer list, using one of the algorithms mentioned in part II.

See Figure (7) for the search of a word "alqatla", which means "the killed people" with the prefixes "a-l" and the transformed letter "final a".



Figure (7) The search of a word "alqatla" with the prefixes "a-l" and the transformed letter "final a".

The program gives a sub-window, telling in red, that the searched word does not exist, and suggesting some similar words.

The first choice gives the word without its prefixes. Here, the user must interact with the program to find the most suitable words.

The user can choose the nearest word in the smallest window, choose the search algorithm, and finally choose among the three buttons, from right to left, one of the following choices:

"closing",

"having more results" or

"going on with the chosen options".

If we choose the first choice which is the nearest word, we find the window shown in Figure(7), which tells the root of the word to find out the meaning. The program finds out the root of the word which is "qatala" which means "kill", and then we find the meaning of it.



Figure (8) going on with the first line of the smallest window of Figure (7).

The left window, tells in "red" writing that the wanted word does not exist, gives another word which is very similar to the searched word according to the search according to the root and the word.

Once again, we can choose among the three search strategies mentioned above (search according to the root, search according to the word and the root, and search according to the root and similar words).

The last three bottom buttons, from right to left, permit successively to:

close the sub-window,

seek more search results, or

continue the search in the same category.

We just recall the first window which we find when we launch the search program, shown in figure (9):

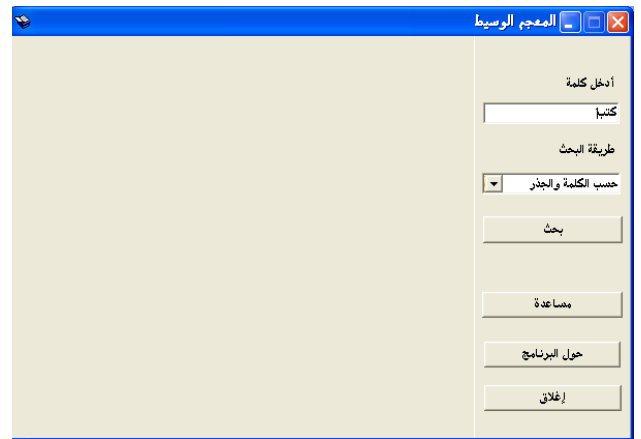


Figure (9) The application window at launch

In this window, the up sub-window permits to enter a word to search its meaning, the second one permits to select the search strategy among the three methods: (search according to the root, search according to the word and the root, and search according to the root and similar words).

The remaining four buttons from up to down are successively:

to launch the search,

to see a help documentation file,

to give information about the program and

to close the application.

IV. CONCLUSION

The application that we describe in this paper, can be used at schools, to help students to look for words in an electronic Arabic-Arabic dictionary. This dictionary can be updated and improved regularly. In addition, it can be linked to answering systems, in future syntactic and semantic Arabic analyzers. It can also be used to evaluate other applications such as syntactic analyzers. The information linking verbs with their subjects and objects can help in future translation systems and others.

The following step is to link the program with a morphological synthesizer, to integrate derived words with their roots, meanings and contexts.

ACKNOWLEDGMENT

Special thanks to Mr. Marwan BAWAB, expert in Arabic, who launched the project in HIAST, to Mme Safa ATTAR who built the lexical database, and who is doing permanent revisions and improvements.

REFERENCES

- [1] S. Attar, M. Bawab, and O. Al Dakkak, "Arabic Lexical Database", paper nb. 338, Workshop on Arabic Natural Language Processing, ICTIS 2007, April 2007, fes, Morocco
- [2] <http://www.tcc-quatoriom/almawrid-BAS-1875.htm>
- [3] <http://language-resource.co.uk/dictionaries.html>.
- [4] <http://babylon.com>
- [5] M. Bawab "prefixes, suffixes and infixes in Arabic" internal report, HIAST.
- [6] M. Z. Alfarkh "Al Wadeh in Grammer and Syntax" In Arabic. " الواضح في القواعد والإعراب" تأليف محمد زرقان الفرخ" نشر دار هبة وهدى".
- [7] A. Aljazem and M. Amin "The Clear Syntax in the Arabic Language" in Arabic. " النحو الواضح في قواعد اللغة العربية" تأليف علي الجازم ومصطفى أمين، نشر دار المعارف في مصر و دار المعارف في لبنان".