Towards the Construction of a Comprehensive Arabic Lexical Reference System

Hamza Zidoum, Fatma Al-Rasbi, and Muna Al-Awfi
Department of Computer Science, Sultan Qaboos University, Oman

Abstract: *Arabic is a Semitic language spoken by millions of people in 20 different countries. However, not much work has been done in the field of online dictionaries or lexical resources. WordNet is an example of a lexical resource that has not been yet developed to its full extent for Arabic. WordNet, a lexical database developed by Professor George Miller and his team at Princeton University, has seen life 20 years ago. Ever since then, it has proved to be widely successful and extremely necessary for today's demands. Accordingly, the motivation of developing an Arabic WordNet (AWN) became strong. This project addresses the nominal part of WordNet as the first step towards the construction of a comprehensive AWN. The nominal part means nouns as a part of speech.*

Keywords: *Wordnet, synsets, arabic processing, lexicon.*

Received March 10, 2012; accepted July 28, 2015; published online August 16, 2015

1. Introduction

WordNet is an online lexical reference system, which groups words into sets of synonyms and records the various semantic relations between these synonym sets. It has become an important aspect of NLP and computational linguistics [12]. Many WordNets were constructed for different languages [3, 4, 10, 11, 17, 18].

WordNet is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory [10]. It groups words into sets of synonyms called synsets, which are the basic building blocks of WordNet [10]. A synset is simply a set of words that express the same meaning in at least one context [10, 12]. WordNet also provides short definitions, and records the various semantic relations between these synonym sets. Nouns, verbs and adjectives are organized into synonym sets, each representing one underlying lexical concept [8, 9, 12, 13]. The lexical database is a hierarchy that can be searched upward or downward with equal speed. WordNet is a lexical inheritance system [13]. The success of WordNet is largely due to its accessibility, quality and potential in terms of NLP. WordNet was successfully applied in machine translation, information retrieval, document classification, image retrieval, and conceptual identification [14]. Different WordNets can be aligned, resulting in the possibility of translation between different languages, as so called machine translation. Information retrieval can be achieved by improving the performance of query answering systems using lexical relations. Since, WordNet links between semantically related words, semantic annotation and classification of texts and documents are possible [18]. The visual thesaurus is a dictionary and a thesaurus with an interesting interface. It's an excellent way of learning the English vocabulary and understanding how the English words link together. It has 145,000 words and 115,000 meanings and shows 16 kinds of semantic relationships. The user can as well hear the pronunciation of the word using a British or an American accent. Once the user enters a word, it is kept in the center, and all of the related words surround it. The user can click on a word to bring it to the center, roll over a word to learn more about it, and print the output chart [16]. Another interesting application is "READER". A person reading a text can click on a word, which is linked to a lexical database "WordNet" and reads its meaning in the given context [5]. An English dictionary, thesaurus and word finder program called WordWeb was developed based on the database from Princeton WordNet. It shows synonyms, antonyms, types and parts of a word. It has the advantage of integrating a dictionary and a thesaurus, unlike similar programs, where the dictionary and thesaurus are separate programs [19].

Arabic is the language of millions of speakers around the globe and adopted by the UN as one of the official languages in the world. Surprisingly relatively few efforts have been made to develop an original Arabic WordNet (AWN). As we will see from WordNet's applications, a WordNet is inevitable for any language that aims to be part of today's ever-evolving applications that are becoming increasingly necessary in our daily lives [2]. For instance, the use of AWN as a lexical and semantic resource is becoming increasingly inevitable, where one needs the use of a conceptual representation of the text [6]. Filling this gap by developing an AWN is a challenging and non-trivial task. This project aims towards constructing a Nominal AWN for Modern Standard Arabic, which will be the starting point of developing a complete

AWN. Our goal is to develop an AWN freely distributed to the community. Our objectives are:

1. Producing a Micro AWN, which contain nouns, verbs and adjectives.
2. Collect basic lexical data from available resources.
3. Create a set of computer programs that would accept the user's queries and display output info to the user.

This paper presents Micro AWN. This is a first step towards developing a complete AWN. In this project we concentrate on using a subset of nouns for implementing this system. The other parts of speech e.g., verbs and adjectives, are considered as future work.

Section 2 gives a definition of Arabic language and its characteristics. Section 3 introduces some system requirements and specifications, general approaches for constructing WordNet, details of other WordNets, reasons of adopting the AWN philosophy and challenges. Section 4 explains the different aspects of system design, where a system architecture, dataflow diagrams, entity-relation diagram, data structures and interface designs are sketched. Section 5 lists the sample data used in the database. It lists our test cases and our observations regarding those cases, and provides the performance tests, system information. Finally, we include a cross checking validation of the requirements and discuss the work still to be done to the system in the future.

2. Arabic WordNet

It is important to state that the most distinctive feature of this work is the insistence on maintaining language specific concepts and the intention of developing an AWN exhibiting its richness rather than be driven by other incentives such as national security.

In the field of lexical semantics terms such as 'word' that we would usually define as "the blocks from which sentences are made" [10], are defined differently. It is therefore necessary to define such terms in order to be able to comprehend the following concepts.

2.1. Word

A word is an association between a lexicalized concept and an utterance (or inscription) that plays a syntactic role [12]. For clarity, "word form" is used to refer to the physical utterance or inscription and "word meaning" to the lexicalized concept. Associations between word forms and word meanings are many: many. Some word forms could have several meanings (Polysemy), and some words meanings could be expressed with several word forms (Synonymy) [12].

2.2. Semantic Relations

Semantic relations are very important in lexical semantics. However, prior to the appearance of WordNet, they were implicit in conventional dictionaries [15]. Now, they are explicit in WordNet, and play as the source of WordNet's richness. Semantic relations associate between sunsets and words. Before they are listed, an important concept must be put forward. It is the distinction between lexical semantic relations as shown in Table 1 and conceptual semantic relations as shown Table 2. The former are between word forms such as, synonymy and antonymy whereas the latter are between synsets such as, hyponymy and meronymy [10].

Table 1. Lexical semantic relations.

Relation	Relation in Arabic	Definition	Example	Type
Synonymy	الترادف	Similarity of meaning; two expressions are synonymous if the substitution of one for the other never changes the truth value of a sentence in which the substitution is made.	Location and Place	Lexical
Antonymy	التضاد	The antonym of a word *x* is sometimes *not-x*, but not always.	Rich and Poor	Lexical

Table 2. Conceptual semantic relations.

Relation	Relation in Arabic	Definition	Example	Type
Hyponymy/ Hypernymy	انضواء/احتواء	ISA relation; a hyponym inherits all the features of the more generic concept and adds at least one feature that distinguishes it from its superordinate and from any other hyponyms of that superordinate.	Maple and Tree	Semantic
Meronymy/ Holonymy	جزء/كل	HASA relation; part-whole relation	Finger and Hand	Semantic

Mona Diab from Stanford University proposed the idea of "Bootstrapping an AWN Leveraging Parallel Corpora and an English WordNet". She studied the feasibility of meaning definition projection of English words onto their Arabic translation [4]. They concluded that it is feasible to automatically bootstrap an AWN taxonomy given less than perfect translations and alignments leveraging off existing English resources. The results were encouraging, as they are similar to those of researchers built EuroWordNet. Supported by the United States Central Intelligence Agency, a group of researchers, some of who were involved in the construction of other WordNets such as, Princeton WordNet and Euro WordNet, decided to undertake the task of developing an AWN [7]; for reasons such as, Arabic being a language spoken in more than 20 countries and the fact that it represents vital interest to US national security [3].

3. Specifications

It is evident from the design of the AWN that is being developed by Black *et al.* [3], is centered on enabling future machine translation between Arabic and other languages that justifies the use of tools such as the SUMO. Some aspects have been adopted in our project because SUMO for instance is a good software engineering practice (increasing the number of users). However, it is necessary to state that the most distinctive feature of this project is the insistence on maintaining language specific concepts and the intention of developing an AWN that exhibits its richness rather than be driven by other incentives.

Input: *The user enters an Arabic word*.
Processing: *The system searches for the word in the Lexical Database* (*AWN*).
Output: *The system displays all the senses of the word* (*synset*). *It will display synonyms, antonyms, hypernyms and hyponyms*.

A WordNet, since it is a lexical database, attempts to approximate the lexicon of a native speaker [10]. The mental lexicon, which is the knowledge that a native speaker has about a language, is highly dense in connectivity, i.e., there are many associations between words. Therefore, constant additions of relations are needed to improve the connectivity of a WordNet. This requires intensive research to discover relations that are not commonly used, since they are the ones with a lower priority of inclusion into the database. Moreover, according to miller "one of the central problems of lexical semantics is to make explicit the relations between lexicalized concepts" [10]. A lexicalized concept is a concept that can be expressed by a word [15].

One of the challenges in this project specific to Arabic is the fact that Arabic texts today tend to be written without diacritics, leaving the task of disambiguation to the reader's natural mental ability, which is a very complicated one when attempted to implement through a computer.

For example, the form (كتاب) could be either intended to mean (كُتَّاب) "kuttab" which is 'a group of writers' or (كِتَاب) "kitab" which is 'a book'. This example clearly demonstrates how missing diacritic marks compound lexical ambiguity.

It is a common complaint that much of the infrastructure for computational linguistics research, or for applications development is lacking. Low involvement of Arab linguists also compounds the challenge, driving some researchers to find alternative means, which sometimes might degrade the quality desired of the outcome.

4. System Architecture

The system has two main components:

1. A User System Component.
2. Lexicographer System Component.

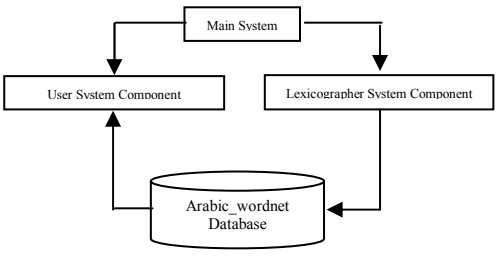

Figure 1. System architecture.

The latter is necessary since there is a lack in electronic Arabic resources and available lexicographers that are necessary for the population of the AWN. The user system component retrieves information from AWN database. It addresses the normal user's need, when interested in finding out the different synsets and relations for a word.

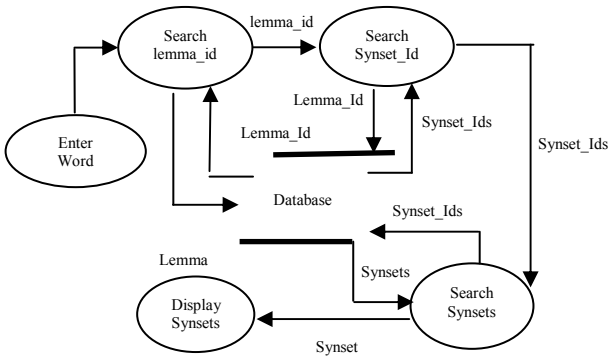

Figure 2. Data flow diagram of displaying overview (level 2).

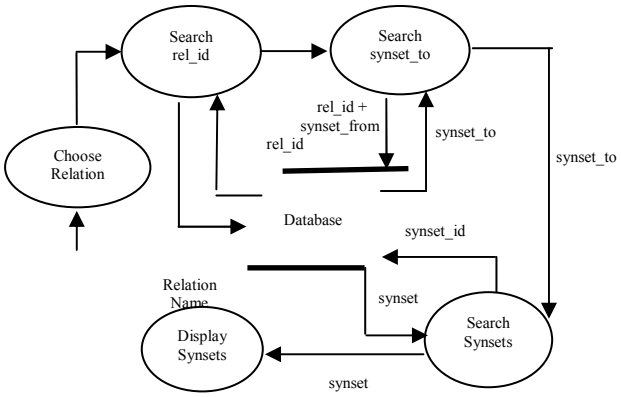

Figure 3. Data flow diagram of displaying synset's relation (level 2).

The lexicographer system component stores information into Arabic_wordnet database. It handles the lexicographer (linguist) requirements, who basically insert new synsets and relations. Data flow starts when the lexicographer chooses an action. The lexicographer could either choose to add a synset or to edit an existing one. Dashed arrows are optional paths which the lexicographer could choose to follow.

There are two main strategies for building WordNets:

1. Expand Approach: Translate English (or Princeton) WordNet synsets to another language and take over the structure. This is an easier and more efficient method. The outcome of this approach is of compatible structure with English WordNet. However, the vocabulary is biased by PWN.
2. Merge Approach: Create an independent WordNet in another language and align it with English WordNet by generation the appropriate translation. This is more complex and requires a lot of work and effort. Language specific patterns can be maintained, but it has different structure from WordNet.

Arabic is a totally different language from English, obviously the expand approach will not be appropriate. Moreover, it is undesirable for the AWN to be biased

by the English WordNet. Therefore, we are going to use the merge approach, since Arabic's specific patterns could be maintained.

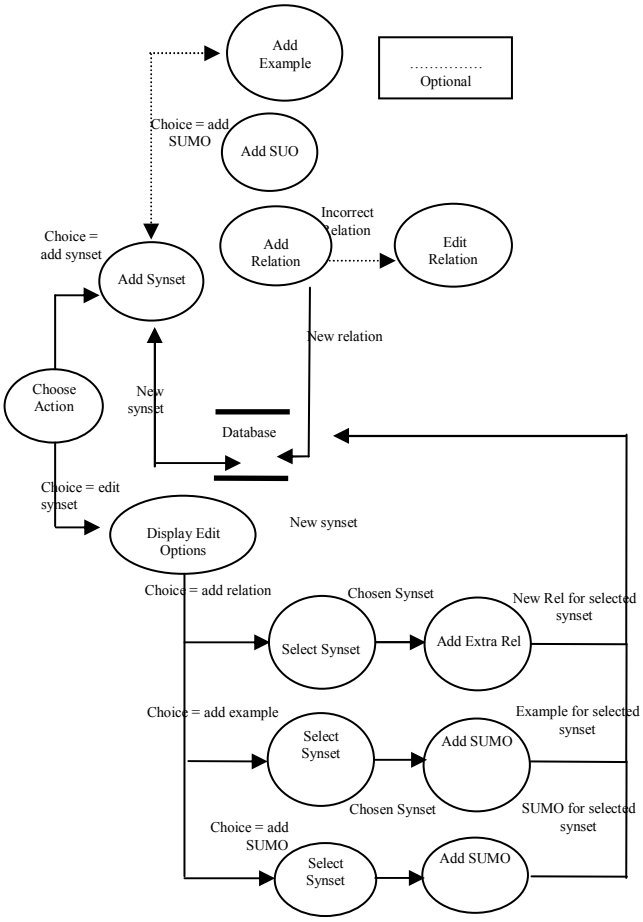

Figure 4. Lexicographer system component.

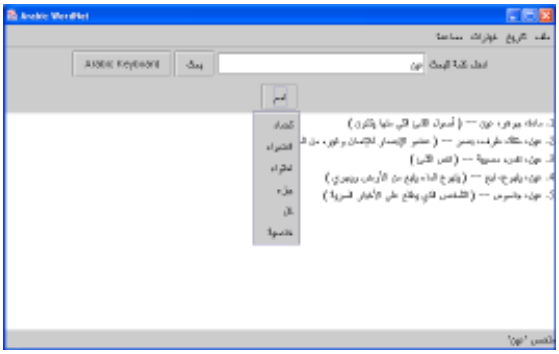

Figure 5. AWN browser-click noun button.

5. System Implementation and Validation

In this project we used MySQL as a database management system for many reasons. MySQL is a widely used software by many large companies for keeping thousands even millions of records. Since there are thousands of words in a language, the need for software to handle such big number of records has emerged. MySQL is also web accessible. Querying MySQL is straight forward and easy. MySQL Query Browser is a free software, which renders MySQL database with an interface similar to Microsoft Access DBMS. It also checks for users' query correctness.

The programming language used is Java Programming Language accessed by SunOne Java. Java is platform independent. It allows creating attractive graphical interfaces. It manipulates Unicode characters as well. As we are going to manipulate Arabic words, ASCII characters will not suffice our purpose. Unicode characters are the solution for writing and reading Arabic text. Being an object oriented language; Java can provide a better structure and interface of the system being implemented. Java is also popular for its rich library which facilitates string manipulation.

5.1. Testing

Testing data was carefully chosen to cover all test cases. The test cases are shown in Table 3.

Table 3. Test cases results.

Test Case	Covered in System	Proved
Input is Correct	Yes	Yes
Input is Wrong	Yes	Yes
Input is Correct and Relation Exist	Yes	Yes
Input is Correct and Relation does not Exist	Yes	Yes

The functionality of the system is designed to handle all cases and all errors. Testing data was carefully chosen to cover all test cases. The test cases are shown in Table 4. Apparently, the functionality of the system is well designed to handle all cases and all errors.

Table 4. Test cases results.

Input	عين+علاقة جزء
Expected Synsets Output	١. مادة، جوهر، عين -- (أصول الشئ التي منها يتكون) ٢. عين، مقلة، طرف، بصر -- (عضو الإبصار للإنسان وغيره من الحيوان) ٣. عين، نفس، مصيبة -- (نفس الشئ) ٤. عين، ينبوع، نبع -- (ينبوع الماء ينبع من الأرض ويجري) ٥. عين، جاسوس -- (الشخص الذي يطلع على الأخبار السرية)
Real Synsets Output	١. مادة، جوهر، عين -- (أصول الشئ التي منها يتكون) ٢. عين، مقلة، طرف، بصر -- (عضو الإبصار للإنسان وغيره من الحيوان) ٣. عين، نفس، مصيبة -- (نفس الشئ) ٤. عين، ينبوع، نبع -- (ينبوع الماء ينبع من الأرض ويجري) ٥. عين، جاسوس -- (الشخص الذي يطلع على الأخبار السرية)
Expected Related Synsets Ouput	١. مادة، جوهر، عين -- (أصول الشئ التي منها يتكون) ٢. عين، مقلة، طرف، بصر -- (عضو الإبصار للإنسان وغيره من الحيوان) ٣. عين، نفس، مصيبة -- (نفس الشئ) ٤. عين، ينبوع، نبع -- (ينبوع الماء ينبع من الأرض ويجري) ٥. عين، جاسوس -- (الشخص الذي يطلع على الأخبار السرية)
Real Related Synsets Output	عين، مقلة، طرف، بصر -- (عضو الإبصار للإنسان وغيره من الحيوان) => جسم، جسد، بدن -- (كل شخص يدرك من الإنسان والحيوان والنبات)

Input	عيت
Expected Synsets Output	الكلمة غير موجودة
Real Synsets Output	الكلمة غير موجودة
Expected Related Synsets Ouput	-
Real Related Synsets Output	-

5.2. Performance

In this section, we will discuss the approximation of data retrieval in our system. Another test was made to approximate processing times on Intel(R) Pentium(R) 4 CPU 3.20GHz 3.19GHz, 0.99GB of RAM. The results using the timer in the source code were summarized in Table 5.

Table 5. Performance test using java timer.

Number of Test Case	Retrieval Time of Synsets	Retrieval Time of Relations
Case 1	15 milliseconds	78 milliseconds
Case 2	0 milliseconds	-
Case 3	16 milliseconds	31 milliseconds
Case 4	16 milliseconds	15 milliseconds

The results using the timer in MySQL Query Browser are given in Table 6.

Table 6. Performance test using MySQL timer.

Number of Test Case	Retrieval Time of Synsets	Retrieval Time of Relations
Case 1	0.0043 seconds	0.0050 seconds
Case 2	0.0096 seconds	-
Case 3	0.0125 seconds	0.0454 seconds
Case 4	0.0125 seconds	0.0107 seconds

5.3. Future Work

Our current system tackles nominal singular input. In the future, we are planning to implement verbs and adjectives as well as plural input. There will be separate tables in the database for verbs and adjectives. Also, an algorithm will be designed to generate the singular form of a given plural word since only the singular forms will be stored in the database. Moreover, to make the system comprehensive it is planned that the system be equipped with an algorithm that is similar to a morphological analyzer but doesn't generate the root. Instead, it generates the sound form of the word if given an inflected form or a derived form.

Another issue that is planned to be tackled in the future is the problem of diacritics, a problem unique to Arabic. Lemmas with the same orthographical representation when stripped of diacritic marks will have to be disambiguated if the user attempts to search for one of them.

To enrich our database as much as possible, it is desired in the future to cover Classical Arabic in addition to the Modern Standard Arabic which is currently being covered.

For additional functionality, and specific to Semitic languages, it would be convenient to have a search by roots or to display the words derived from the same root.

Also, an important plan for our system is that we are planning to upgrade our system to be a web application. To facilitate this upgrade, we have taken all the necessary precautions. We have used open source tools like MySQL and Java. To transform the application we developed to a servlet or an applet is not a big challenge. The operating system that will be used is Linux with Apache as a server. It is notable that the domains: Arabicwordnet.com, arabicwordnet.org and arabicwordnet.net have been registered.

As well as the system being a textual-based system, we are looking forward to implement a graphical-based AWN. The synsets and relations can be represented as hierarchies, trees or even radial diagram.

6. Conclusions

We have realized after doing research on the possibility of collecting validated data, that it is extremely difficult to populate the database because of the lack of machine readable dictionaries and available lexicographers. We decided therefore to include a lexicographer interface in addition to the original intended user interface. We have also built the system in a structure that enhances scalability. After upgrading the system to a web application (as it has been discussed in the previous chapter) we will aim to contact lexicographers in universities around the world to contribute in the construction of the AWN. In the analysis phase we collected some user, lexicographer and system requirements and analyzed them. A general idea of the system architecture has been developed. The design phase included the system architecture, data flow diagrams, entity-relation diagram, data structures and interface design. The implementation phase discussed the different tools used in implementing the system. We define different functions in this phase as well as stating their pseudocode. The testing phase included database data as well as testing data. Then, a discussion of these tests results is concluded. We also tested the performance of the system and stated the statistics. We hereby anticipate the realization of a comprehensive AWN once it has been published on the web (current project) and the lexicographers contacted.

References

[1] Abu-Absi S., "The Arabic Language, Glossary of Linguistic Terms," available at: http://www.sil.org/linguistics/GlossaryOfLinguisticTerms, last visited 2015.

[2] Black W. and El-Kateb S., "A Prototype English-Arabic Dictionary Based on WordNet," *in Proceedings of the 2nd Global WordNet Conference*, Czech Republic, pp. 67-74, 2004.

[3] Black W., Elkateb S., Rodriguez H., Alkhalifa M., Vossen P., Pease A., and Fellbaum C., "Introducing the Arabic WordNet Project," available at: http://vossen.info/docs/2006/arabic.pdf, last visited 2006.

[4] Diab M., "Feasibility of Bootstrapping an Arabic WordNet Leveraging Parallel Corpora and an English WordNet," available at: http://www.cs.columbia.edu/~mdiab/papers/bootstrappingArabicWN.pdf, last visited 2004.

[5] Educational Uses of WordNet. READER: A Lexical Aid, available at: http://wordnet.princeton.edu/reader.

[6] Elberrichi Z. and Abidi K., "Arabic Text Categorization: A comparative Study of Different Representation Modes," *the International Arab Journal of Information Technology*, vol. 9, no. 5, pp. 465-470, 2012.

[7] Elkateb S., Black, W., Rodriguez H, Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C., "Building a WordNet for Arabic," available at: http://nlp.lsi.upc.edu/papers/fellbaum06.pdf, last visited 2015.

[8] Fellbaum C., "English Verbs as a Semantic Net," *International Journal of Lexicography*, vol. 3, no. 4, pp. 278-301, 1990.
[9] Fellbaum C., Gross D., and Miller K., "Adjectives in WordNet," *International Journal of Lexicography*, vol. 3, no. 4, pp. 265-277, 1990.
[10] Fellbaum C., *WordNet: An Electronic Lexical Database*, MIT Press, 1998.
[11] Mihaltz M. and Proszeky G., "Results and Evaluation of Hungarian Nominal WordNet v1.0," *in Proceedings of the 2nd Global WordNet Conference*, Brno, pp. 175-180, 2004.
[12] Miller G., Beckwith R., Fellbaum C., Gross D., and Miller K., "Introduction to WordNet: An On-line Lexical Database," *International Journal of Lexicography*, vol. 3, no. 4, pp. 235-244, 1990.
[13] Miller G., "Nouns in WordNet: A Lexical Inheritance System," *International Journal of Lexicography*, vol. 3, no. 4, pp. 245-264, 1990.
[14] Morato J., Marzal M., Llorens J., and Moreiro J., "WordNet Applications," *in Proceedings of Global WordNet Conference*, Brno, pp. 270-278, 2004.
[15] Niles I. and Pease A., "Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology," *in Proceedings of International Conference on Information and Knowledge Engineering*, pp 412-416, 2003.
[16] Visual Thesaurus., available at: http://www.darwinmag.com/read/buzz/column.html?ArticleID=576.
[17] Vossen P., "EuroWordNet General Documentation," available at: http://wordnet.princeton.edu/man/wnstats.7WN
[18] Wintner S. and Yona S., "Resources for Processing Hebrew," available at: http://www.cs.cmu.edu/~alavie/Sem-MT-wshp/Wintner+Yona_presentation.pdf, last visited 2003.
[19] WordWeb., "International English Thesaurus and Dictionary for Windows," available at: http://wordweb.info, last visited 2015.

Hamza Zidoum received his Ms, and PhD degrees from the University of France-Comté, France. He is a faculty member in the Department of Computer Science in Sultan Qaboos University since 2002. He is the head of artificial intelligence research group.

Fatma Al-Rasbi graduated from the Department of Computer Science Department, Sultan Qaboos University.

Muna Al-Awfi graduated from the Department of Computer Science Department, Sultan Qaboos University.