# Word Sense Disambiguation Approach for Arabic Text

Nadia Bouhriz
Dept. of Mathematics and computer science
Faculty of Sciences Ben M'sik
Hassan II University
Casablanca, Morocco

Faouzia Benabbou
Dept. of Mathematics and computer science
Faculty of Sciences Ben M'sik
Hassan II University
Casablanca, Morocco

El Habib Ben Lahmar
Dept. of Mathematics and computer science
Faculty of Sciences Ben M'sik
Hassan II University
Casablanca, Morocco

***Abstract*—Word Sense Disambiguation (WSD) consists of identifying the correct sense of an ambiguous word occurring in a given context. Most of Arabic WSD systems are based generally on the information extracted from the local context of the word to be disambiguated. This information is not usually sufficient for a best disambiguation. To overcome this limit, we propose an approach that takes into consideration, in addition to the local context, the global context too extracted from the full text. More particularly, the sense attributed to an ambiguous word is the one of which semantic proximity is more close both to its local and global context. The experiments show that the proposed system achieved an accuracy of 74%.**

***Keywords*—Word Sense Disambiguation; Arabic Text; local context; global context; Arabic WordNet; Semantic Similarity**

## I. Introduction

WSD is a natural language processing (NLP) field. It aims at determining the appropriate sense of an ambiguous word occurring in a given context [1] [2]. It is a task which allows a better understanding, and consequently a better exploitation of the processed linguistic material. It is therefore very essential task for NLP applications, such as Machine Translation (MT), Information Retrieval (IR), Text classification…etc.

The oldest WSD approach proved that two words before and after the ambiguous word are sufficient for its disambiguation [3]. For the Arabic language, the information extracted from this local context is not always sufficient.

To solve this problem, an Arabic WSD system was proposed in this paper that is not only based on the local context, but also on the global context extracted from the full text. The objective is to combine the local contextual information with the global one for a better disambiguation.

More particularly, the proposed system uses the resource Arabic WordNet (AWN) to select word senses. The sense attributed then to an ambiguous word is the one that possesses the closest semantic proximity to the local context, as well as to the global one. This proximity is measured based on the semantic hierarchy offered by WordNet.

The rest of the paper is organized as follows: Section II presents the architecture of WSD systems. Section III exposes some Arabic WSD systems. Section IV displays the description of the proposed system. Section V contains experiments and the obtained results. The last section gives conclusion and some perspectives.

## II. WSD Systems Architecture

In 1949, Weaver [4] discussed the necessity of WSD for MT, and he explained that to realize this process, the ambiguous word must be taken from the context where it occurred. In 1950, Kaplan [3] made experiences to determine in which size the context should be, in order to disambiguate a word. It proved that two words at the right and at the left (size =2) of the ambiguous word are sufficient for its disambiguation; Masterman [5] confirmed this result for the Russian language, while Choueka and Lusignan [6] confirmed it for the French.

Over the years, WSD systems were developed according to different approaches. Actually, these systems have generally an architecture structured around three main steps:

- Sense inventory: consists on selecting the senses of the words.
- Context representation: represents senses and contexts in a formal manner.
- Disambiguation Process: attributes for every ambiguous word its correct sense according to its context.

The sense inventory step is the one that makes the difference from one system to another depending on the adopted approach. Generally, two approaches exist:

The first one, called Knowledge-based approach, is based on the use of external lexical resources. These resources are containing all the words of a language with their senses. These resources can be dictionaries [7], thesaurus [8], or ontologies [9] [10].

Unlike the first approach, the second one doesn't use external lexical resources, but it acquires the necessary information to define words' senses from a corpus; it's called a Corpus-based approach. This information is obtained by the application of statistical language models on this corpus. Three approaches are distinguished in this category, supervised approaches that require annotated corpus [11] [12] [13], unsupervised approaches [14] [10] that require unannotated corpus, and a semi-supervised approaches that require both of the annotated and the unannotated corpus [15].

## III. ARABIC WSD SYSTEMS

### A. Challenges

Arabic presents several challenges for WSD, due essentially to the particularities of this language and also to the lack of resources necessary to the disambiguation process.

Diacritics' missing in Arabic texts is the most challenging characteristic for WSD; because it increases the number of a word's possible senses and consequently makes the disambiguation task more difficult. For example, the word without diacritics (Swt صوت) have 11 senses according to the AWN, while the use of diacritics for the same word (Saw~ata صَوَّتَ), cuts down the number of senses to two.

On the other hand, the Arabic language is very rich morphologically. This causes an ambiguity during the lexical segmentation, and influences consequently the detection of the words' correct sense during disambiguation process. For example, the word (وجد Wjd) have two possible segmentations; the first one considers that the letter (و W) is a prefix of (جد Jad), while the second considers it as a letter in the word, which gives two totally different words.

### B. State of the Art

The first WSD systems were mostly concerned by Latin languages like English and French since several decades ago. The Arabic language, as for it, didn't get the attention until the last decade.

The first Arabic WSD system was proposed by Mona Diab in 2002 [10]. The author introduced in this work an unsupervised method to annotate Arabic words by their sense using English WordNet and an English-Arabic parallel corpus.

Another contribution was proposed by Elmougy [13] where a Naïve Bayes Classifier was used to disambiguate Arabic words without diacritics.

Merhbene [16] was based on the semantic trees and a measure of collocation to choose the most appropriate sense to an ambiguous word.

Zouaghi [17] have proposed a system of WSD by combining the information retrieval measures with the Lesk algorithm to estimate the most appropriate sense of the ambiguous word.

The most recent work was proposed by Menai [18], in which the author was based on the genetic algorithms. His objective is to exploit the power of these algorithms in the Arabic WSD.

All of the previously mentioned works used only one contextual information to disambiguate. The proposed system, as for it, leans on two contextual informations. The first one is extracted from the local context of the ambiguous word and the second from its global context.

## IV. THE PROPOSED SYSTEM

Before describing the system process, the structure of Arabic WordNet is firstly given.

### A. Arabic WordNet

The Arabic WordNet (AWN) [19] [20] [21] is a lexical resource for modern standard Arabic. It was constructed according to the Princeton WordNet content. It's organized around elements called Synsets, which are a set of synonyms and pointers connecting it with other synsets. So, the AWN is a lexical network in which synsets represent its nodes and the connections between synsets represent its edges.

This resource counts at present 23,481 words organized into 11,269 synsets. A word can belong to one or more synsets.

In this work, the senses of a word are defined by the Synsets to which it belongs in the AWN. Below, some words synsets (i.e. senses) extracted from AWN are presented:

TABLE I. EXAMPLE OF AWN SENSES

| words | Senses (synsets) |
|---|---|
| بحر | Sense 1 = [بحر]<br>Sense 2 = [محيط, بحر] |
| شعر | Sense 1 = [شعر, قصيدة]<br>Sense 2 =[شعر]<br>Sense 3 = [شعر]<br>Sense 4 = [أحس, شعر]<br>Sense 5 = [حس, أحس, شعر]<br>Sense 5 = [أحس, شعر] |
| مال | Sense 1 = [فلوس, ثروة, دراهم, مال]<br>Sense 2 = [نقود, مال]<br>Sense 3 = [مال]<br>Sense 3 = [تمايل, ترنح, مال]<br>Sense 4 = [انحدر, مال]<br>Sense 5 = [نزع إلى, مال]<br>Sense 6 = [أقنع, أمال, مال]<br>Sense 7 = [انحرف, انحنى, مال] |

### B. Description of the proposed system

#### 1) Sense inventory

In this step, a preprocessing phase is applied; it contains a text segmentation process, a stop words removal process, and finally a stemming process to remove words' affixes (prefixes and suffixes).

Afterwards, the obtained words are classified, according to the AWN, into two categories:

- Non ambiguous words: belonging to one Synset, i.e. possessing one sense.
- Ambiguous words: belonging to several Synsets, i.e. possessing several senses.

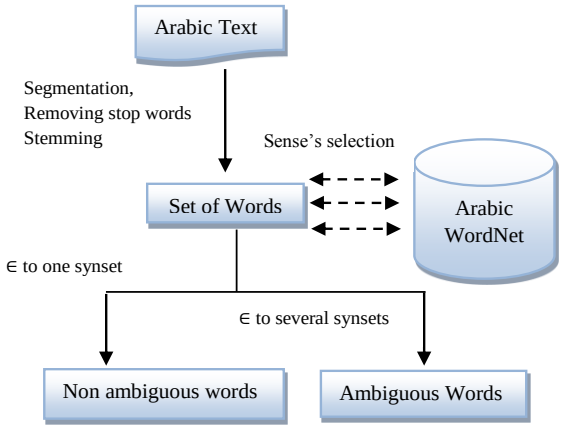


Fig. 1. Sense inventory process

**Sense inventory Algorithm**

**Input**: Arabic Text T
**Output**: List of Ambiguous Words $AW$ and Non Ambiguous Words $NAW$.

1: **Segment** the Text
2: **Remove** stop words
3: **Apply** Stemming process for all obtained words
4: **For each word** $w$ **do:**
5: **If**: $w$ is belonging to one synset in AWN
6: **Then**: Add $w$ to NAW list.
7: **Else if**: $w$ is belonging to a several synsets in AWN
8: **Then**: Add $w$ to AW list.
9: **End**

*2) Context representation*

This step consists of representing words' senses as vectors. For this purpose, the set of all non ambiguous word senses($S_1$, $S_2$, $S_3$,…, $S_n$) is firstly considered, afterwards, the vector space, spanned by the standard basis $B = \{e_i\}_{i=1..n}$, where $e_1 = (1,0,0,\dots,0)$, $e_2 = (0,2,0,\dots,0)\dots, e_n = (0,0,0,\dots,1)$ are respectively the unit vector of the sense $S_i$, is built.

Using this space, words' senses will be represented by the vector $V = \sum_{i=1}^{n} a_i\, e_i$ where $a_i$ is the $i^{th}$ coordinate representing the semantic distance between the word sense and the sense $S_i$ in AWN. To calculate this distance, the Wu and Palmer (wu-p) measure is used [2].

The global context will be afterwards defined by the sense vectors set of non ambiguous words present in the full text: $Contx_{Global} = \{V_1, V_2, \dots, V_n\}$, while the local context will be defined by the sense vectors set of non ambiguous words present only locally: $Contx_{Local} = \{V_i$ where $V_i$ is the vector sense of $i^{th}$ non ambiguous word present locally $\}$.

Finally, an ambiguous word aw that has m senses will be represented by the set of its sense vectors:

$aw = \{W_1, W_2, \dots, W_m\}$.

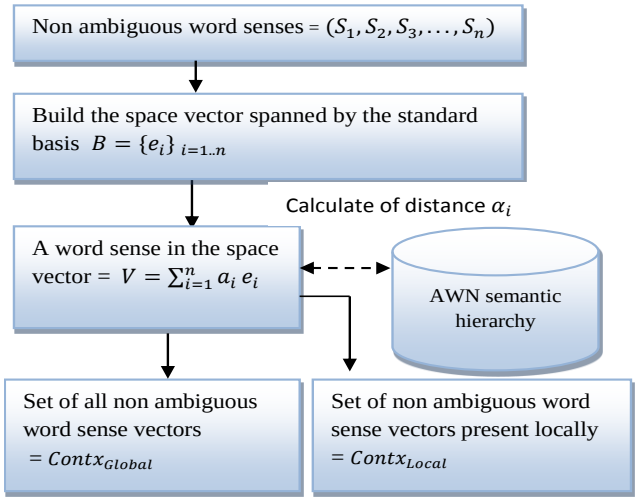


Fig. 2. Context representation

**Context representation Algorithm**

**Input:** list of $AW$ and $NAW$
**Output:** $Contx_{Global}$, $Contx_{local}$,

1: **For all** words in $NAW$ **do:**
2: **Extract** associated senses ($S_1$, $S_2$, $S_3$,…, $S_n$)
3: **End**
4: **For each** word **do:**
5: **For each** senses **do:**
6: **For each** $S_i$ **do:**
7: **Calculate** the wu-p semantic distance $a_i$ between word sense and the sense $S_i$.
8: **End**
9: **Calculate** word sense vector $V = \sum_{i=1}^{n} a_i\, e_i$
10: **End**
11: **End**
12: **Construct** $Contx_{Global}$
13: **Construct** $Contx_{Local}$

*3) Disambiguation process:*

This last step consists of attributing for each ambiguous word its appropriate sense. This is done by choosing the sense with the closest semantic proximity to its local and global context.

Sense semantic proximity with a context is defined by the percentage of vectors in this context that are similar to the vector of this sense.

Similarity measurement between two vectors $V=(v_1, v_2, \dots, v_n)$ and $W = (w_1, w_2, \dots, w_n)$ can be calculated by three distances which are; dot product, cosines, and Jaccard defined respectively as follows:

$$\text{dotProduct}(V.W) = \sum_{i=1}^{n} v_i.w_i$$

$$\cos(V,W) = \frac{\sum_{i=1}^{n} v_i.w_i}{\sum_{i=1}^{n} {v_i}^2 . \sum_{i=1}^{n} {w_i}^2}$$

$$\text{Jaccard}(V,W) = \frac{\sum_{i=1}^{n} v_i.w_i}{\sum_{i=1}^{n} {v_i}^2 . \sum_{i=1}^{n} {w_i}^2 - \sum_{i=1}^{n} v_i.w_i}$$

According to the previous definitions, the local and global semantic proximity are measured for each ambiguous word sense; as a result, a pair of percentages representing respectively each of the semantic proximity is obtained. The sense with the better average of its two percentages will be assigned finally to the ambiguous word.

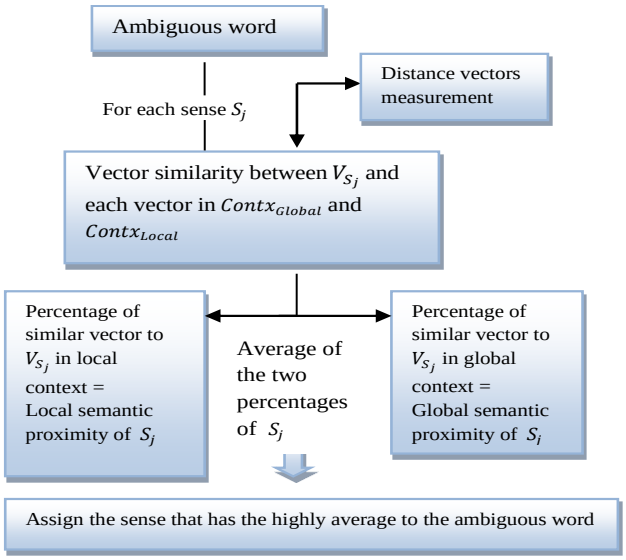


Fig. 3. Disambiguation Process

**Disambiguation process Algorithm**

**Input:** ambiguous word
**Output**: sense of the ambiguous word

1: **For each** sense $S_j$ of the ambiguous word **do:**
2: **Calculate** local semantic proximity of $S_j$
3: **Calculate** global semantic proximity of $S_j$
4: **Calculate** average semantic Proximity of $S_j$
5: **End**
6: $BestSense = S_1$
7: **For each** sense $S_i$ of the ambiguous word **do:**
8: **If: (** $average(S_i) > average(S_1)$**)**
9: **Then:** $BestSense = S_i$
10: **End**
11: **Assign** the sense $BestSense$ to the ambiguous word.

## V. Experimentations and Results

To evaluate the proposed system, a test corpus is constructed by collecting texts from various fields (news, sport, medicine, religions, etc.); afterwards, each word is annotated manually by its correct sense according to the AWN.

The Java language was used to implement the system, and to access the XML AWN database the 'Java API for Arabic WordNet'[1] was used. Finally, the application of the stemming process is based on SAFAR platform[2].

For measuring the system's efficiency, the precision measurement was used; it consists of the number of words correctly disambiguated divided by the number of all ambiguous words. Experiment results have shown a precision of 74%.

Another experiment have shown that the use of a stemming process during the sense inventory phase increases the system's efficiency. More particularly, results (Table II) show firstly that the use of this process increases efficiency by 30%, moreover, they have shown that the use of AlKhalil Analyzer [23] is better than Buckwalter [24] by 4%:

TABLE II. Importance of the Stemming Process

| Morphological analyzer | Without Stemming | Buckwalter | AlKhalil |
|---|---|---|---|
| System precision | 40% | 70% | 74% |

The table below (Table III) show some disambiguated words from this piece of text:

وفي سياق متصل قال المتحدث الرسمي في الرئاسة العامة للأرصاد وحماية البيئة حسين القحطاني انه يوجد بالرئاسة مركز للبلاغات ورقم مجاني (988) لاستقبال بلاغات الكوارث الطبيعية والبحرية، كما أن هناك خطة وطنية للاستجابة ومكافحة التلوث تضم في عضويتها الجهات ذات العلاقة بالتلوث البحري.

TABLE III. Example of Words Disambiguated

| Ambiguous word | Senses | Local Semantic Proximity | Global Semantic Proximity | Average Semantic Proximity | Sense selected |
|---|---|---|---|---|---|
| رئاسة | رئاسة,إدارة | 0% | 1.88% | 0.94% | قيادة,دور قائد,رئاسة,دور رئيس |
| | قيادة,دور قائد,رئاسة,دور رئيس | 16.6% | 11.3% | 13.95% | |
| | فترة,رئاسية,رئاسة,إدارة | 0% | 0% | 0% | |
| | قيادة,زعامة,رئاسة | 0% | 0.9% | 0.45% | |
| حماية | حماية,جهاز,وقاية,جهاز حماية,سلامة | 16.6% | 18.8% | 17.7% | حماية,جهاز,وقاية,جهاز حماية,سلامة |
| | حماية,تركيب دفاعي,دفاع | 16.6% | 18.8% | 17.7% | |
| | حماية,رعاية,حراسة,عناية | 0% | 1.88% | 0.94% | |
| | حماية,رعاية,وقاية | 16.6% | 16.9% | 16.75% | |
| | حماية,حفظ | 0% | 0% | 0% | |
| | حماية,اهتمام,رعاية,وقاية | 16.6% | 11.3% | 13.95% | |
| مركز | مركز,موقع,مكانة | 0% | 0% | 0% | مركز |
| | مركز,بؤرة | 0% | 11.32% | 5.66% | |
| | مركز,موقع | 0% | 18.8% | 9.4% | |
| | مركز | 16.6% | 20.75% | 18.67% | |
| | مركز,مكان | 0% | 0% | 0% | |
| | مركز,بؤرة,محرق | 0% | 11.32% | 5.66% | |
| | مركز,منتصف,وسط | 0% | 11.32% | 5.66% | |
| | موضع,مركز,موقع,علاقة مكانية | 0% | 3.77% | 1.88% | |
| | موضع,مركز,مكان,موقع | 0% | 11.32% | 5.66% | |
| | مادة,تركيز,المنطقة,مركز | 0% | 5.66% | 2.83% | |
| رقم | رقم,رقم,تعيين الهوية | 0% | 5.66% | 2.83% | رقم,عدد |
| | رقم,عدد | 0% | 11.32% | 5.66% | |
| | رقم | 0% | 0% | 0% | |
| | علم,وضع علامات الترقيم,رقم,نقط | 0% | 0% | 0% | |
| خطة | استراتيجية,خطة,مخطط | 0% | 0% | 0% | استراتيجية,خطة,مخطط |
| | نظام,برنامج,خطة | 0% | 0% | 0% | |
| | خطة | 0% | 0% | 0% | |
| | مؤامرة,خطة | 0% | 0% | 0% | |
| استجابة | إجابة,رد,استجابة,جواب | 0% | 18.8% | 9.4% | إجابة,رد,استجابة,جواب |
| | رد فعل,استجابة | 0% | 0% | 0% | |
| | تلبية,إجابة,رد,استجابة,جواب | 0% | 11.32% | 5.66% | |

The last experiment results show that the proposed approach is better by 0.34% than the classical method (based on local context). This is due to some challenges described as follows:

- The non-recognition of named entities (persons' names, locations, organizations…etc.). These last should not be separated during segmentation process. Experiments show that words like: عبد الله, أبو ظبي have not been recognized as a named entity.

[1]https://sourceforge.net/projects/javasourcecodeapiarabicwordnet/
[2] http://sibawayh.emi.ac.ma/safar/publications.php

- Another similar challenge that decreases the system efficiency is the incapability of multiword expression recognition such as قاعدة بيانت, الأمم المتحدة…etc.
- The absence of a component that allows disambiguating senses with the same average semantic proximity.
- The absence of a part-of-speech tagging that allows categorizing words in verbs and names allowing consequently studying names and verbs in a separate way.
- The last challenge is relying on the lexical resource used. The AWN doesn't cover all Arabic words, which has consequently an impact on the system efficiency. For example the word منطق doesn't belong to the AWN structure.

## VI. Conclusion

In this paper, a WSD system for Arabic texts was presented. The proposed system, unlike other systems, takes into consideration two types of context during disambiguation process. The first one is the local context defined by the words in the neighborhood of the ambiguous word, and the second is the global context defined by the full text.

Experiments have shown an accuracy of 74% for the proposed system.

The incorporation of a named entities and a multiword expression component in the process will be necessary done in the future for a better results, as well as a raise of all the challenges previously mentioned.

As for the future, this method will be integrated in a semantic indexing process to help enhancing Arabic information retrieval system.

## References

[1] E. Agirre, P. Edmonds, ''Word sense disambiguation: algorithms and applications''. Springer, 2006.

[2] R. Navigli, ''Word sense disambiguation: a survey''. ACM Comput Surv 41(2):1–69, 2009.

[3] A. Kaplan, ''An experimental study of ambiguity and context''. Mechanical Translation 2(2), 39–46, 1955.

[4] W. Weaver, ''Translation''. in Machine Translation of Languages, MIT Press, Cambridge, MA, 1949.

[5] M. Masterman, ''Semantic message detection for machine translation, using an interlingua". International Conference on Machine Translation of Languages and Applied Language Analysis, Her Majesty's Stationery Office, London, 437-475, 1961.

[6] Y. Choueka, S. Lusignan, ''Disambiguation by short contexts''. Computers and the Humanities, 19, 147-158, 1985.

[7] M.E. Lesk, ''Automatic sense disambiguation using machine readable dictionaries : How to tell a pine cone from a nice cream cone''. In Proceedings of the SIGDOC Conference, Toronto. 1986.

[8] D. Yarowsky. ''Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora''. Proceedings of the 14th International Conference on Computational Linguistics (COLING-92), 454 –460, 1992.

[9] P. Resnik. ''Disambiguating noun groupings with respect to WordNet senses'', in S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann & D. Yarowsky, eds, 'Natural Language Processing Using Very Large Corpora', Kluwer Academic Publishers, Boston, M.A, pp. 77–98, 1999.

[10] M. Diab, P. Resnik ''An unsupervised method for word sense tagging using parallel corpora''. in Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL-02), Philadelphia, PA, pp. 255–262, 2002.

[11] E.F. Kelly, P.J. Stone. ''Computer recognition of english word senses''. North-Holland Publishing. North-Holland, Amsterdam. 1975.

[12] T. Pedersen. ''Learning Probabilistic Models of Word Sense Disambiguation''. PhD thesis, Southern Methodist University, Dallas 1998.

[13] S. Elmougy, T. Hamza, H.M. Noaman. ''Naive Bayes classifier for Arabic word sense disambiguation''. In: Proceedings of INFOS 2008, Cairo, pp 27–29, 2008.

[14] H. Schutze. ''Automatic word sense discrimination. Computational Linguistics''. Special Issue on Word Sense Disambiguation, 24 (1), 97–123, 1998.

[15] D. Yarowsky. ''Unsupervised word sense disambiguation rivaling supervised methods''. In 33th Annual Meeting of the Association for Computational Linguistics, pp 189-196, 1995.

[16] L. Merhbene, A. Zouaghi, M. Zrigui. ''Approche basée sur les arbres sémantiques pour la désambiguïsation lexicale de la langue arabe en utilisant une procédure de vote''. proceeding de 21ème conférence sur le Traitement Automatique des Langues Naturelles, Marseille 2014.

[17] A. Zouaghi, L. Merhbene, M. Zrigui. ''Combination of information retrieval methods with LESK algorithm for Arabic word sense disambiguation''. Artif Intell Rev 38:257–269, 2012.

[18] M.E. Menai. ''Word sense disambiguation using evolutionary algorithms – Application to Arabic language''. Computers in Human Behavior 41 : 92–103, 2014.

[19] W. Black, S. El-Kateb. ''A Prototype English-Arabic Dictionary Based on WordNet''. http://www.fi.muni.cz/gwc2004/proc/95.pdf . 2004.

[20] C. Fellbaum, W. Black, S. Elkateb, A. Marti, A. Pease, H. Rodriguez, P. Vossen. ''Constructing Arabic WordNet in Parallel with an Ontology''. http://www.globalwordnet.org/AWN/meetings/meet20050901/Fellbaum .ppt , 2005.

[21] S. Elkateb, W. Black, P. Vossen, D. Farwell, A. Pease, C. Fellbaum. Arabic WordNet and the Challenges of Arabic. http://www.mt-archive.info/BCS-2006-Elkateb.pdf, 2006.

[22] Z. Wu, M. Palmer. ''Verb semantics and lexical selection''. Proceedings of 32nd annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, pp 27-30, 1994.

[23] A. Boudlal, A. Lakhouaja, A. Mazroui, A. Meziane, O.B.M Ould Abdallahi, and M. Shoul. ''Alkhalil Morpho SYS1: A Morphosyntactic Analysis System for Arabic Texts''. In the proceedings of the 11th International Arab Conference on Information Technology. Benghazi, Libya 2010.

[24] T. Buckwalter. ''Buckwalter {Arabic} Morphological Analyzer Version 1.0'' , 2002.