*Extraction des relations sémantiques à partir du Wiktionnaire Arabe **

Bakhouche Abdelali, Yamina Tlili-Guiassa

Laboratoire LRI/Equipe SRF, Université Badji Mokhtar, Annaba, Algérie

Bakhouche2006@yahoo.fr, guiyam@yahoo.fr

Abstract: *Electronic language resources play a very important role in natural language processing. They are used in several linguistic applications including machine translation, text indexing, automatic summary…etc. The aim of this paper is to create a lexical database for the Arabic language that does not have many of these resources. We exploit web resources like Wikipedia and Wiktionary, which have become an interesting websites for extracting information. In this work, we seek to automatically extract semantic relationships such as synonyms and antonyms from Wiktionary Arabic.*

Résumé : *Les ressources linguistiques électroniques jouent un rôle très importantentraitement automatique du langage naturel.Elles sont utilisées dans plusieurs applications linguistiques notamment la traduction automatique, l'indexation des textes, le résumé automatique…etc.L'objectif de cet article est decréer une base lexicale pour la langue Arabe qui ne dispose pas beaucoup de ces ressources. Nous exploitons les ressources Web comme Wikipédia et le Wiktionnairequi sont devenues des sources intéressantes pour l'extraction d'information.Dans ce travail, nous cherchons à extraireautomatiquement des relations sémantiques notamment les synonymes et les antonymes à partir du Wiktionnaire Arabe.*

Keywords:*Lexical base, Information retrieval, Semantic relations, Arabic language,Wiktionnaire, Wikipédia.*
Mots clés *: Base lexicale, Extraction d'informations, Relations sémantiques, la langue Arabe, Wiktionnaire, Wikipédia*

**Extraction of semantic relation from Arabic Wiktionnaire*

© Revue RIST, vol. 20, n° 2, p.47-56

1 Introduction

Depuis l'essor de l'internet,Le volume d'information ne cesse d'accroître, ce qui augmente le nombre de ressources électronique d'information, notamment les ressources du domaine public ou collaboratives comme Wiktionnaire et Wikipédia qui sont exploitées pour créer les bases de connaissances ainsi que pour les utiliser dans les différentes tâches du traitement automatique du langage naturel.

Wiktionnaire est un dictionnaire multilingue, universel et librement diffusable. Depuis son lancement official par Jimmy Wales et Larry Sanger le 15 janvier 2001. Le nom de «Wiktionnaire» se compose de deux termes «wiki» et «dictionnaire». Un wiki est une application Web permettant l'édition collaborative simplifiée de pages. Ward-Cunninghammis en place le premier système de ce genre en 1995et a inventé le nom de"wiki", le mothawaïen pour "rapide". L'exemple sans doute le plus connu d'une ressource wiki basée sur l'encyclopédie en ligne est Wikipédia. Un dictionnaire est un lexique contenant l'ensemble des mots d'une langue ou d'un domaine d'activité fournissant pour chacun une définition et d'autres informations linguistiques. La version arabe de Wiktionnaireest lancée le 24 mai 2004, elle contient au 11 mars 2012 plus de 48,409articles[1].

Dans cet article, nous exploitons le Wiktionnaire arabe pour extraire les relations sémantiques et nous montrons comment il peut contribuer à la création ou l'enrichissement d'un réseau lexical du domaine public pour l'arabe ; dans la section suivante nous présentons les travaux liés à l'extraction automatique de connaissances lexico-sémantiques et l'utilisation des ressources collaboratives pour l'extraction d'informations. La section 3 présente les différentes caractéristiques des relations sémantiques de la langue arabe, ensuite nous expliquons dans la section 4 l'accès au contenu de Wiktionnaire arabe et comment transformer ce contenu en un format plus convivial, ainsi cette section décrit le processus d'extraction des relations sémantiques, et La section 6 conclu cet article et donne des pointeurs et extensions pour le futur.

2 Travaux antérieurs

Les dictionnaires sont des sources intéressantes pour l'extraction automatique des différentes connaissances lexico-sémantiques, dans ce sens plusieurs travaux ont utilisé ces ressources pour extraire des relations sémantiques, dont le but était de créer des grands réseaux sémantiques [01] [02]. D'autres études utilisent le dictionnaire pour la désambigüisation du sens du mot, ainsi que pour l'analyse des textes [13]. La différence entre les ressources en quantité et en nature pose un problème d'extraction d'information, et pour résoudre ce problème des chercheurs ont combiné plusieurs dictionnaires [14].

Quelques études récentes sur l'acquisition d'ontologies et sur l'extraction de relations sémantiques à partir des ressources comme WordNet qui est une base de données lexicales développée au laboratoire des sciences cognitives de l'université de Princeton [03], cette ressource a été utilisée par les auteurs [02] dans le cadre de question/réponse pour construire une ontologie. Il existe d'autres sources, qui ont été utilisées dans la création ou l'enrichissement des bases de données lexicales telles que des corpus ou des ressources collaboratives comme Wikipédia et Wiktionnaire. Ces deux dernières fournissent des dictionnaires électroniques au contenu gratuit, avec des définitions, des exemples et des informations sur la partie du discours (POS), des traductions, des prononciations

[1]http://ar.wiktionary.org/wiki

et l'étymologie, ainsi que des informations sur les relations sémantiques (synonymes, antonymes...) [04]. Qui ont été construits manuellement par des gens non professionnels sur le Web, aujourd'hui le Wiktionnaire contient environs 5 millions d'entrées dans 170 éditions linguistiques [04].

Les Wiktionnaires ont été exploité pour l'extraction des relations sémantiques et les comparer avec celle extraites des autres resources [05]. Par ailleurs, le Wiktionnaire a été utilisé pour l'enrichissement des ressources lexicales existantes ainsi que dans la création automatique des nouvelles ressources lexicales [06] [08].

3 Les relations sémantiques dans la langue arabe

La sémantique occupe une position importante dans le traitement de la langue naturelle. Il n'est pas facile de réaliser des traitements profonds des textes sans informations suffisantes sur la sémantique des termes et les relations sémantiques entre les mots constitutifs des textes, pour cela les chercheurs du traitement automatique de la langue arabe n'ont pas écarté les questions de Synonymie et d'Antonymie et d'autres relations sémantiques, en raison de leurs valeurs ajoutées pour résoudre plusieurs problèmes commel'analyse automatique des textes, la compréhension des textes, la traduction automatique,...

Dans cette section nous allons aborder les deux relations sémantiques (la synonymie et l'antonymie) dans la langue arabe.

3.1 La synonymie

La synonymie est la relation sémantique qu'il existe entre deux items lexicaux qui diffèrent par leur forme mais expriment le même sens ou un sens très proche [09]. Nous devons savoir que les synonymes ne sont pas complètement des mots identiques, sauf s'ils ont le même opposé etla même utilisation en context [10], ce qui n'est pas généralement le cas. Donc on ne peut pas dire que la synonymie est la similitude complète et absolue, mais c'est la similitude de la plupart des traits sémantiques.

- **L'importance de la synonymie**

Les synonymes offrent les avantages suivants:

Offrir à l'utilisateur un lexique très riche et plusieurs termes pour désigner le même sens, donc l'utilisateur a l'occasion de sélectionner ce qui est adéquat avec le contexte.
Stimuler la jouissance et réduire l'ennui du lecteur, car la diversité offre à l'auteur la possibilité de choisir ses termes loin de toute ambiguïté sémantique et donc il peut fixer le sens voulu.

3.2 L'antonymie

L'antonymie est la relation sémantique qui existe entre deux items lexicaux dont les sens s'opposent [09]. Elle est capable d'enrichir les dictionnaires très fortement, l'opposé du mot éclaire le sens de son opposé malgré qu'il n'est généralement pas complet.Il y a des raisons derrière l'antonymie des mots, citons:

Ressemblance *Exemple:* (بشرة) pour la peau de l'être humain et pour la plante

Opposition*Exemple*: (عسس) pour l'avènement de la nuit (إقبال الليل) et pour retour de la nuit (إدبار الليل)

***Différents sens de même champ sémantique* :** *Exemple* (السرحان) pour le lion et pour le loup

Classe de parole : *Exemple* (أجم) pour le verbe (اقترب) *s'approcher* et pour le nom (كبش بلا قرون) *bélier qui n'a pas des chélicères*

***La brièveté et la métaphore* :** *Exemple* « مكتب » pour le meuble « ما يكتب عليه », pour la salle du bureau « حجرة المكتب » et pour l'équipe de travail « مكتب الإدارة »

Les aspects de l'antonymie

Les aspects de cette notion sont:

1. ***L'antonymie complémentaire*** *:* la différence entre les deux mots est totale, il n'y a pas d'aspects ou des degrés de rapprochement entre les deux :

(أنثى ″ féminin″-ذكر ″masculin″), (ميت ″*mort*″-حي ″*vivant*″), (متزوج ″*marié*″-أعزب ″*célibataire*″), (امرأة ″femme″-رجل ″ homme″)

2. ***L'antonymie scalaire*** *:* entre les deux mots existent des aspects de rapprochement sémantique, exemple : (صعب ″difficile″-سهل ″facile″) il y a entre les deux des nuances de facilitation de difficulté, d'autres exemples

(قريب ″proche″-بعيد ″ loin″), (حار ″ chaud″-بارد ″ froid″), (قوي ″fort″-ضعيف ″ faible″)

3. ***L'antonymie des conversifs*** *:* entre deux mots accompagnes exemple :

(أب″père″-ابن ″fils″**), (**زوج″époux″-زوجة ″épouse″**), (**مهزوم ″vaincu″-فائز″gagnant″**)**

4. ***L'antonymie duale*** *:* il s'agit des antonymes spatiaux comme : (تحت ″ sous″-فوق ″sur″), (شرق ″est″-غرب″ouest″), (شمال ″nord″-جنوب ″sud″), les antonymes temporels comme : (نهار ″jour″ - ليل ″nuit″) et culturels comme (سؤال ″question″ -جواب ″réponse″)

– **Le rôle de l'antonymie dans l'enrichissement du lexique**

L'antonymie est un phénomène linguistique intéressant dans l'éclaircissement du sens, tel que l'opposé du mot éclaire son sens, malgré que l'antonymie ne soit généralement pas complète que dans des rares cas, mais il l'est dans certains traits.

Exemple : (ميت) ″mort″ son opposé est (حي) ″vivant″, donc l'opposé est désigné dans un seul trait qui est le trait de vie « الحياة », mais pour le mot « ميت » il y a plusieurs traits comme « الحياة » ″ la vie″, « الجنس » ″ le sexe″, « النوع » ″ le type″….

L'antonymie est l'un des types de la synonymie négative (les synonymes jouent un rôle de l'aspect positif pour le champ sémantique, alors que les antonymes jouent le rôle de l'aspect négatif). Et de ce fait l'antonymie montre les aspects sémantiques, d'un coté et enrichir le lexique, d'un autre coté, de surcroît les liaisons sémantiques entre les champs [10].

4 Architecture de notre approche

Dans cette section, nous décrivons la méthode d'extraction des synonymes et des antonymes à partir de Wiktionnaire arabe. Notre approche est composée de 3 parties, la première est une phase de préparation (prétraitement et extraire les définitions), la deuxième phase est l'analyse des vocabulaires de ces définitions ainsi l'extraction des relations sémantiques. La dernière phase consiste à créer une base lexicale.La figure 1 montre l'architecture de notre approche.

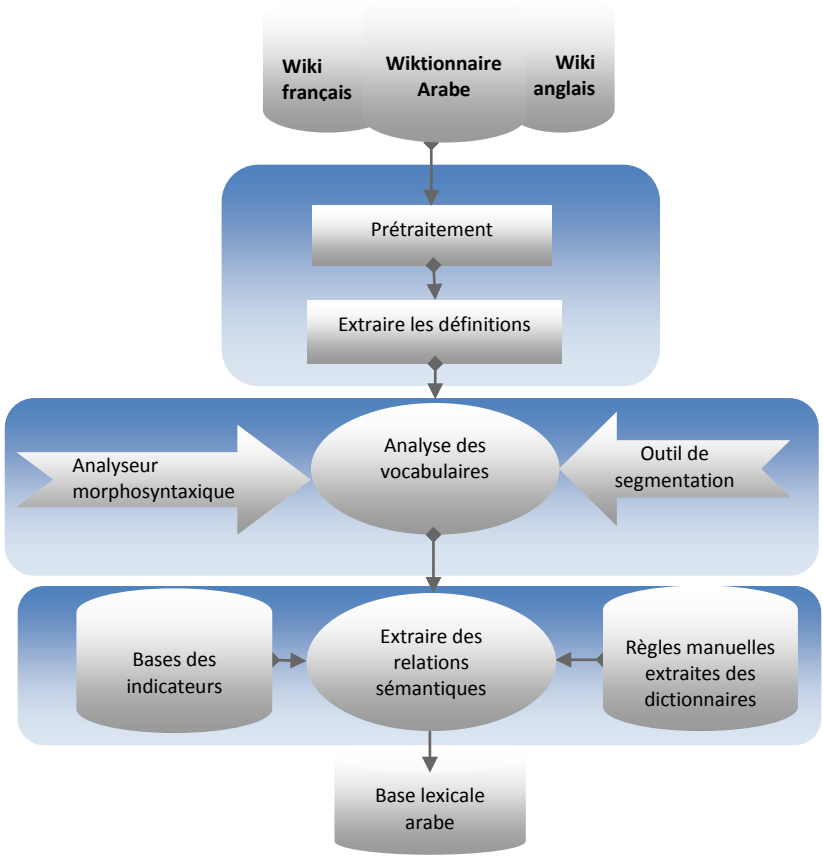

Figure 1: Architecture de l'approche

4.1 Prétraitement et Extraction des définitions

L'extraction des informations lexico-sémantiques disséminés dans les bases de connaissances collaboratives nécessite des outils d'accès automatiques à son fichier XML (Fig.2). Ces outils sont disponibles pour les langues comme l'anglais, l'allemand [07] et le portugais [05], dans cet objectif nous avons développé un outil pour analyser la structure du fichier XML pour la langue arabe. Qui peut supprimer les lettres latines, les chiffres, les caractères spéciaux,…

```
<page>
<title>حَاسُوْب</title>
.....
<textxml:space="preserve">wikipedia
=عَرَبِيَّة=
==المعاني==
# [[حَوَاسِيب]] الحَاسُوْب اسم مذكر يُجمع جمع تكسير على
آلة[[ الكترونية]]تقوم بعمليات[[ حِسَابِيَّة ]]سريعة بواسطتها يمكن للآلة القيام بمهمات مختلفة
على[[ البَيَانَات والمَعْلُومَات]].اِسْتِخْدَامٌ حَدِيْثٌ #
=== المرادفات ===
# '''1''': [[حَاسِبَة]]
# '''2''': [[آليحَاسِب]]
.....
</page>
```

Figure 2: Les informations contenues dans le fichier XML pour l'entrée حَاسُوْب dans le Wiktionnaire arabe

Cet outil peut exporter toutes les définitions et les transformer en format convivial où chaque ligne contient l'entrée, sa partie de discours et sa définition. Comme le montre l'exemple suivant (Fig.3):

الحَاسُوْب	اسم	مذكر يُجمع جمع تكسير على حَوَاسِيب	
الحَاسُوْب	اسم	آلة الكترونيةتقوم بعمليات حِسَابِيَّة سريعة بواسطتها يمكن للآلة القيام بمهمات مختلفة على البَيَانَات والمَعْلُومَاتاِسْتِخْدَامٌ حَدِيْثٌ	
الحَاسُوْب	صفة	حَاسِبَة	
الحَاسُوْب	اسم	حَاسِبآلي	

Figure 3: Définitions obtenus à partir de l'entrée dans la figure 1

4.2 Analyse des vocabulaires

Pour extraire les relations sémantiques à partir des définitions. Nous avons utilisé l'outil de segmentation AraSeg [11], pour découper les textes des définitions en unités lexicale: paragraphe, phrase, syntagme, mot graphique, mot-forme, morphème,…etc. ensuite l'analyseur morphosyntaxique [12] utilise les différentes ressources de connaissance de la langue arabe pour extraire les lemmes et les classes grammaticales des mots constituant les définitions, le résultat de ces outils est présenté dans le tableau suivant.

Mot	Lemme	Catégorie grammaticale
آلة	آلة	إسم ، مبتدأ
الكترونية	إلكترون	إسم، خبر
تقوم	قام	فعل
عمليات	عمل	إسم، مجرور
……..	……..	……

Tableau 1 : Exemple de résultat de l'analyseur morphosyntaxique de définition du mot الحاسوب

4.3 Extraction des relations sémantiques

L'extraction des relations sémantiques est basée sur plusieurs informations morphosyntaxiques, marqueurs et des règles inspirées du dictionnaire. La phase actuelle est très importante car elle affecte les résultats du système. Elle a pour but de déterminer la relation sémantique entre les entrées de wiktionnaire et les mots constituant leurs définitions.Ce système fait l'extraction des relations à base des indicateurs et des règles conçues manuellement.Lesindicateurs et ces règles ont été inspirés à partir des dictionnaires (لسان العرب, القاموس المحيط)[2]. Les indicateurs sont présentés sous forme d'une base contenant par exemple les pronoms personnels comme (هو، هي،...) et autres indicateurs comme(مرادفه، معناه، ضده،.....). Cependant les règles représentent les structures des phrases.

Nous avons obtenus 8321 relations sémantiques induites à partir de 25037 Définitions. Les relations sont représentées sous forme de triplets (A, R, B) dont Aest un mot dans la définition, B est l'entrée de Wiktionnaire et R est le nom de la relation.

L'entrée de wiktionnaire «Représenté par le caractère B »	**Relation Représenté par le caractère « R»**	**Mot dans la définition « Représenté par le caractère A»**
الحاسوب	Synonyme	آلة حاسبة
العَيْن	Synonyme	مُعَايَنَة
العَيْن	Antonyme	نَهْر
……..	……..	……

Tableau 2: Exemple de représentation des relations sémantiques

4.4 Création de la base lexicale

La dernière étape consiste à mettre en relation les mots et les connaissances sémantiques trouvées dans l'étape précédente afin de construire la structure générale des données. La base lexicale est représentée par le modèle entité-association telle que les entités représente les mots et les associations représentent les relations sémantiques entre eux. On travaille toujours sur la base et sa représentation et pour des raisons d'analyse on représente la base sous cette forme (figure 4).

[2]http://www.baheth.info/

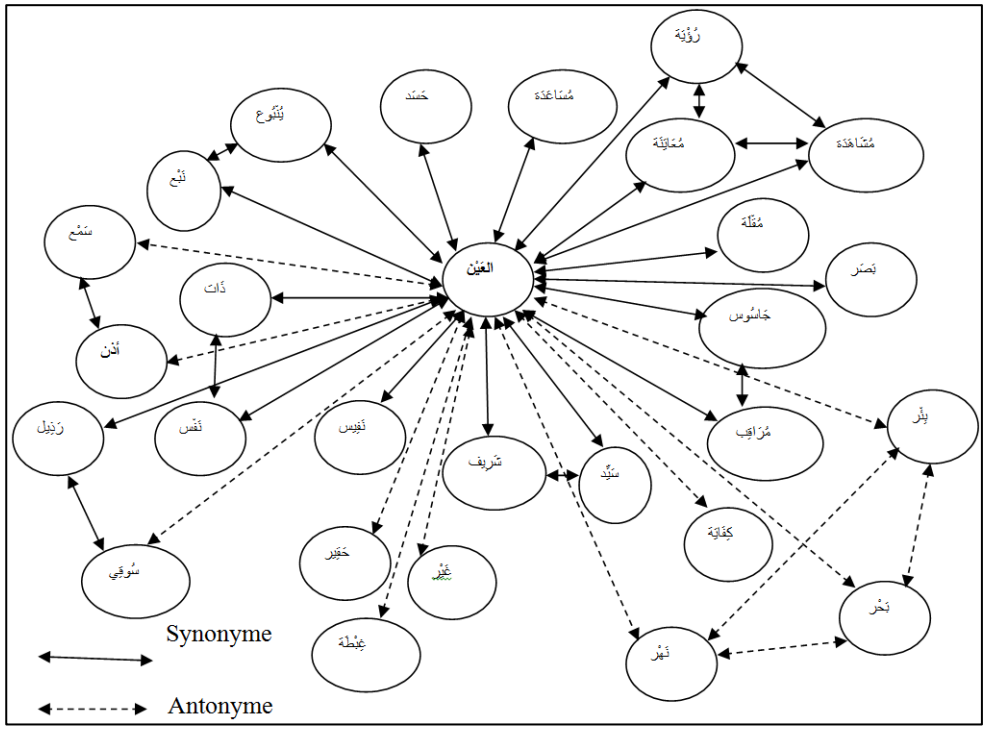

Figure 4 : Exemple d'une base lexicale

5 Résultats

Une fois le processus d'extraction est appliqué à des définitions extraites du Wiktionnaire arabe, nous avons obtenu 8321 triples relationnels, répartis dans le tableau 3. Ce tableau présente, le nom de la relation sémantique, le nombre des relations extraites de chaque type. Le pourcentage par rapport le nombre global. D'après le résultat on remarque que la synonymie représente le pourcentage le plus élevé.

Relations	**Nombre de relations**	**Pourcentage**
Synonymie	6784	81%
Antonymie	1537	19%

Tableau 3 : Résultat du system

6 Conclusion et perspectives

Le Wiktionnaire arabe est une ressource riche en connaissances lexico-sémantiques, dans cetarticle nous avons exploité cette ressource pour créer une base lexicale. On a proposé des étapes d'extraction des relations sémantiques aprèsavoir convertile contenu de Wiktionnaire à un format plus convivial.Les résultats sont en cours d'analyse.

Ce travail est une partie d'un projet plus vaste dont le but est de créer une vaste ressource lexicale pour la langue arabe.On va utiliser d'autres sources comme Wikipédia qui contient beaucoup de connaissances qui sont mal organisées, pour cette raison nous proposons de sélectionner les entrées les plus pertinentes pour chaque mot dans le réseau lexical, puis extraire les relations sémantiques à partir des résumés de ces entrées. L'utilisation des ressources de collaboration pour extraction des informations morphosyntaxiques est un avantage et un défi car son contenue est en augmentation permanente.

7 Références

[01] Atserias,J., andall.(1997). Multiple Methods for the Automatic Construction of Multilingual WordNets.ArXiv: cmp-lg/9709003v2 16 Sep 1997

[02]Lahsen A., Karim, B. and Paolo R. (2008).Construction de l'ontologie Amine Arabic WordNet dans lecadre des systèmes Q/R. In Proceedings de la Journée Scienfiquesur les Technologies de l'Information et de la Communication, JOSTIC 08, Rabat, Maroc, Novembre, 2008.

[03] Bond,F.etPaik,K.(2012).A Survey of WordNets and their Licenses.In Proceedings of the 6th Global WordNet Conference (GWC 2012).Matsue, Pages 64–71

[04] Meyer, C. etGurevych,I. (2012).Wiktionary: a new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography.In Electronic Lexicography, Oxford University Press, 2012.

[05] Pérez,L., Oliveira,H. andGome,P.(2011).Extracting Lexical-Semantic Knowledge fromthe Portuguese Wiktionary.InEPIA'2011, pp. 704–717

[06]Zesc,T., Müller,C. andGurevych,I. (2008).Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary.In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08),May 2008

[07] Zesch, T., Müller, C. andGurevych, I. (2008).Extracting lexical semantic knowledge fromWikipedia and Wiktionary.In: Proceedings of 6th International Language Resourcesand Evaluation (LREC'08). Marrakech, Morocco

[08] Weal,T., Brew,C. andFosler-Lussier,E. (2009).Using the Wiktionary Graph Structure for Synonym Detection. In Proceeding People's Web '09 Proceedings of the 2009 Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources.

[09] Nizar, H. (2010) Introduction to Arabic Natural language processing. Synthesis Lectures on Human Language Technologies, Print 1947-4040 Electronic 1947-4059

[10] Salwa El-sayed, H. and Omar M. (2006). Automated semantic processing of the Arabic language to build a lexical database of semantics relationships between words, 2006 PHP-Nuke, N° 3

[11] Zoubeir Mouelhi (2008). AraSeg : un segmenteur semi-automatique des textes arabes In JADT 2008 : 9es Journées internationales d'Analyse statistique des Données Textuelles.

[12] El-Shishtawy, T.and El-Ghannam, F. (2012).An Accurate Arabic Root-Based Lemmatizer for InformationRetrieval Purposes.In IJCSI (International Journal of Computer Science Issues), pages 58-66, Vol. 9, Issue 1, No 3, January 2012

[13] Apidianaki, M. (2008). Acquisition automatique de sens pour la désambigüisation et la sélection lexicale en traduction, thèse doctorat, université paris Diderot(Paris 7), 2008

[14]Ide,N. andVéronis,J.(1995). Knowledge Extraction from Machine-Readable Dictionaries: An Evaluation. InProceedings of the Third International EAMT Workshop on Machine Translation and the Lexicon.

[15]García,R., Scholl,P. andRensing,C. (2011).Supporting Resource-based Learning on theWeb using automatically extracted Large-scaleTaxonomies from multiple Wikipedia versions.In Proceedings of the 10th

International Conference on Web-based Learning, ICWL2011, LNCS 7048, pp. 309-314, Springer, December 2011.

[16]Ducatel,B.etall.(2011). CORON : Plate-forme d'Extraction de Connaissances dans les Bases de Données. ArXiv:1111.5687v1 [cs.DB],24 Nov 2011

[17]Mohammed, A.andall. (2010). An automatically built Named Entity lexicon for Arabic.In: LREC 2010 - 7th conference on International Language Resources and Evaluation, 17-23 May 2010, Valletta (Malta).

[18]Mohammed, A.andall., (2010). Automatic Extraction of Arabic Multiword Expressions.In LREC 2010, 7th Conference on Language Resources and Evaluation, 17-23 May 2010, Valletta (Malta).

[19]Hayssam,T.(2009).Arabic Named Entity Extraction: Local Grammar-Based Approach.InProceedings of the International Multiconference onComputer Science and Information Technology, pp. 139 –143

[20]Navarro,E., Sajous, F.andGaume,B. (2009). Wiktionary and NLP: Improving synonymy networks. In proceedings of the 2009 Workshop on the People's Web Meets NLP, ACL-IJCNLP 2009, pages 19–27, Suntec, Singapore, 7 August 2009.

[21]Nichols, E., Bond, F. andFlickinger, D. (2005).Robust ontology acquisition from machine readable dictionaries.In Proceedings of 19th International Joint Conference on Artificial Intelligence (IJCAI), 2005.

[22]Schwab,D. (2004). Base lexicale sémantique basée sur les vecteurs conceptuels LIRMM - Laboratoire d'informatique, de Robotique et de Micro électronique de Montpellier Montpellier - France.