Vers une approche statistique pour l'extraction des éléments d'ontologie à partir des textes arabes

MAZARI Ahmed Cherif
Université de Médéa

Résumé
Le travail présenté dans cet article se rapporte à la construction automatique d'une ontologie linguistique arabe. Nous avons proposé une approche qui repose sur l'utilisation des techniques statistiques d'extraction d'informations ou de connaissances appliquées sur des textes écrits en arabe. Parmi celles-ci, nous avons exploités deux méthodes ; la première est du « segment répété » pour repérer les termes pertinents qui peuvent dénoter des éléments d'ontologie (concepts) et la deuxième méthode est de « cooccurrence des termes » pour lier ces nouveaux concepts extraits à l'ontologie par des relations soit de type hiérarchique ou non-hiérarchique. Le traitement est effectué sur un corpus textuel du domaine écrit en arabe constitué et préparé préalablement.

1. Introduction

Les méthodes existantes de construction et d'enrichissement d'ontologies diffèrent principalement selon les éléments qu'elles manipulent (concepts, relations,...) et les techniques d'extraction de ces éléments à partir de textes. Ces techniques sont réalisées soit par des méthodes linguistiques en utilisant des patrons lexico-syntaxiques qui nécessitent des corpus annotés soit par des méthodes statistiques qui n'ont pas besoin de l'annotation du texte. Dans notre approche, nous nous somme orientés vers l'utilisation des méthodes statistiques puisque ces méthodes n'ont pas besoin de ces types de corpus annotés et des analyseurs (lexicale ou syntaxique) de TAL. Nous avons utilisé ces méthodes en s'appuyant sur les deux

critères suivants; la **pertinence** d'un terme par rapport à un domaine qui est défini par le nombre d'occurrence du terme dans le corpus et la **cooccurrence** de deux termes à une forte fréquence.

2. Description de l'approche

Dans notre approche, nous avons initialisé l'ontologie manuellement, par les concepts généraux récupérés à partir de l'ontologie GOLD[1] (*General Ontology for Linguistic Description*) [Far03]. GOLD est une ontologie générale pour la linguistique descriptive, elle est applicable à la plupart des langues humaines.

La table 1 présente les concepts supérieurs (génériques) de GOLD structurées par la relation hiérarchique «is-a».

Table 1 Classes Supérieures de l'ontologie GOLD.

```
Entity
      Abstract
            FeatureValue
            GrammaticalUnit
            LinguisticDataStructure
            LinguisticFeature
            LinguisticSign
            PhonologicalUnit
            SemanticUnit
      Object
            SymbolicString
                  Character

      OrthographicExpression
                  Term
      Process
```

[1] . http://www.linguistics-ontology.org/gold.html

Ensuite, nous avons adopté les méthodes d'extraction d'information à partir les textes à notre objectif. Le processus général se résume en trois principales étapes ; la première est la constitution du corpus du domaine; cette étape est fondamentale car de la qualité du corpus dépendra la qualité des traitements et le corpus doit couvrir entièrement le domaine traité. La deuxième étape est l'extraction des candidats termes (ces termes peuvent être parmi les éléments qui forment l'ontologie : *un concept, une relation, une propriété ou un individu*). Enfin, l'enrichissement de l'ontologie, c'est-à-dire la jonction de ces nouveaux éléments à notre ontologie (on peut faire appel à l'expert du domaine un linguiste par exemple pour la validation).

2.1 Constitution et préparation du corpus

Dans un projet de construction d'ontologies à partir de textes, le corpus, son statut et sa collecte sont d'une importance primordiale à la fois comme source de connaissances pour construire le modèle et comme source de référence tout au long du processus d'élaboration [BoA03]. Donc, les questions qui reviennent dans la conception du corpus comprennent : le type de corpus (spécialisé est un corpus contenant des textes traitant d'un sujet lié à un domaine), l'adéquation pour le projet visé (la qualité des résultats d'un travail sur corpus dépend en grande partie de la qualité du corpus. Ceci implique; que le domaine des textes dans le corpus soit bien défini et délimité; que les textes soient assez représentatifs pour appuyer les conclusions que nous en tirons), la taille (la taille d'un corpus spécialisé est déterminée selon l'intuition des chercheurs, le caractère des textes inclus et leur domaine. Cependant, la taille est souvent limitée par la disponibilité des textes et par des questions de droits d'auteur.), la représentativité (c'est-à-dire, la variété de

textes, d'auteurs, de sources, etc.), l'utilisation de textes complets ou d'échantillons[Mar03].

2.1.1 Préparation du corpus

Après la constitution du corpus brut, nous devons le préparer pour être traité informatiquement. Cette phase est réalisée par un ensemble d'étapes de prétraitement, pour lever certaines ambiguïté, réduire le nombre d'opérations effectuées et d'adapter le corpus suivant notre objectif.

2.1.1.1. Normalisation

Dans le corpus, nous allons rencontrer des éléments ne portant pas d'information et qui augmentent le délai de traitement. Il s'agit pour l'essentiel des caractères spéciaux, des chiffres, des mots non arabes, des abréviations et de lettres isolées. Ces éléments doivent être supprimés :

- Caractères spéciaux : inclut toute séquence de caractères spéciaux délimitée par des lettres ou des espaces.
- Nombres : regrouper toutes les séquences situées entre deux espaces et contenant des chiffres sous une seule occurrence. Cette méthode a l'avantage aussi de regrouper les dates, les nombres réels et les pourcentages.
- Mots en caractères latins : Les mots en caractères non arabes, essentiellement en caractères latins, sont tout simplement détectés selon leur forme graphique.
- Abréviations et les lettres isolées La liste des mots à une seule lettre dans les textes arabes révèle la présence d'un nombre assez important de ces mots. Ces lettres sont souvent utilisées dans les abréviations. Par exemple ب الفئة, « la catégorie B », numérotation ; أ الفقرة « paragraphe A ». Sigles étrangers,

comme ت.ا.ق abréviation de « TAG ». Ainsi : ت pour تاريخ , «date», م pour ميلادي , ص pour صفحة , «page». Etc. [AbD08].

- Caractère '–'. Les typographes font un usage fréquent du caractère **'-'.** qui permet l'allongement du trait au milieu des mots, pour une meilleure lisibilité, pour limiter les espaces blancs sur une ligne justifiée, voire pour des raisons purement esthétiques. Ce caractère ne faisant pas partie de l'alphabet arabe. Il faut donc recourir à son élimination.
- Dévoyellation : Il s'agit d'enlever les signes de voyellation, qui sont notés sous la forme de signes diacritiques placés au dessus ou au dessous des lettres.
- A cause des variations graphiques qui peuvent exister lors de l'écriture d'un même mot en arabe et par conséquent elles peuvent être des sources de l'ambiguïté. Nous allons opérer certains remplacements comme suit [Dou05] :

Remplacement des lettres إ , آ et أ avec ا .
Remplacement des lettres finales ي , ة par ى, ه .

2.1.1.2 Elimination des mots vides

D'un point de vue linguistique les mots-vides sont par définition des mots "vides" de sens. Donc, un «mot vide» est un mot qui ne doit pas être indexé, qu'il soit mot grammatical ou mot lexical. Ces mots sont très fréquents et ils ne sont pas informatifs [Ver04], ils sont alors regroupés dans une liste répertoriant tous les mots d'outils, pronoms, articles, conjonctions de coordination, prépositions, etc. Exemple :(في،ان هذا الذي، عن، التي، على،).

2.1.1.3. Lemmatisation légère

L'arabe est une langue flexionnelle, fortement dérivable et agglutinante ; les articles, les prépositions et les pronoms collent aux adjectifs, noms, verbes. Pour résoudre l'ambiguïté,

[Bou05] a montré que la lemmatisation est un prétraitement très utile, qui consiste à trouver la racine de chaque mot. Elle effectue une suppression de suffixe et préfixe pour détecter la racine du mot. Ces suffixes et préfixes sont regroupés dans un dictionnaire. Puisque la plupart des mots arabes ont une racine à trois ou quatre lettres, le fait de garder le mot au minimum à trois lettres va permettre de préserver l'intégrité du sens du mot. Nous utilisons la liste de préfixes et de suffixes proposée par [Dar03] Table2, ils ont été déterminés par un calcul de fréquence sur un corpus d'articles arabes.

Table 2. Liste des préfixes et des suffixes.

Préfixes							
لا	فبـ	للـ	كمـ	بمـ	وتـ	بتـ	والـ
با	وا	ليـ	فمـ	لمـ	ستـ	يتـ	فالـ
	فا	ويـ	الـ	ومـ	نتـ	متـ	بالـ
Suffixes							
ا	ـه	ين	نا	هم	ته	وه	ـات
ون	ـي	يه	تك	هن	تم	ـان	وا
					ها	كم	تي

2.2 Extraction automatique des candidats termes

Après la préparation du corpus nous passons à l'étape d'extraction d'éléments d'ontologie. Le traitement se fait en deux passages ; dans le premier, nous allons extraire tous les termes (un ou plusieurs mots) qui servent à dénoter des concepts du domaine en utilisant la méthode des « segments répétés » en s'appuyant sur les prépositions suivantes :

- Un terme significatif sera utilisé à plusieurs reprises dans un texte spécialisé.
- Les termes peuvent êtres complexes, c'est-à-dire qu'ils sont composés de plusieurs mots utilisés isolément (ex. جملة اسمية).

- Les termes complexes se construisent au moyen d'un nombre fini de séquences de mots.

Dans le deuxième passage. Nous chercherons les couples de termes qui cooccurrent plus souvent dans le corpus. Le résultat de ce traitement nous fournit une liste de couples de termes qui sera utilisée pour la mise à jour de l'ontologie.

Donc, l'objectif du premier passage est de détecter les termes qui dénotent les concepts liés au domaine, en revanche le deuxième passage est de repérer parmi ces termes les couples qui ont des liens avec les éléments de l'ontologie.

2.2.1 Application de la méthode des segments répétés

La méthode des segments répétés est une technique statistique d'extraction d'information à partir de textes non étiquetés, il s'agit des segments de texte qui se répètent plusieurs fois à l'intérieur d'un corpus, La répétition de ces segments indique que ces segments peuvent servir à dénoter des concepts liés au domaine du corpus. Un segment de texte est constitué d'un ou plusieurs mots (en choisissant 4 sur le principe qu'un terme dénotant un concept contient au maximum 4 mots) et les délimiteurs sont les signes de ponctuation ou les espaces.

A ce stade un grand nombre de segments sont extraits dont certains sont incorrects. L'ensemble de ces segments est ensuite filtré pour éliminer les segments indésirables et ne conserver que ceux qui seront retenus comme candidats-termes. Dans notre approche, nous avons recours à deux filtres : un filtre par pondération [Her06] et un filtre coupant[2]. Le

[2]. Utilisé dans MANTEX (Extracteur de terminologie à partir de textes non étiquetés) [RoF02].

filtre par pondération permet de sélectionner les termes possédant un poids suffisant par rapport à cette pondération. Le poids est calculé par la fréquence totale d'un terme. Si cette *fréquence* est supérieure à un *seuil global* donc, le terme fait parti du domaine.

Le filtre coupant permet de supprimer les segments comportant certains mots comme des verbes, des entités nommées, des nombres en lettres ou autres. Les mots du filtre coupant peuvent être présents au début, à la fin et à l'intérieur du segment. La liste des mots du filtre peut être aisément adaptée et complétée par l'utilisateur en fonction des spécificités du corpus traité.

2.2.2 Application de la méthode de cooccurrence

La technique est fondée sur l'extraction des cooccurrents en couples de termes qui se rencontrent l'un de l'autre de manière plus fréquente que par hasard et que ces deux termes faisaient partie de la liste trouvée dans la phase précédente. On commence par repérer les cooccurrents d'un terme donné dans une fenêtre de taille fixe (exemple 10 termes) et dans une même phrase, en examinant les cooccurrents par rapport au terme cible.

La méthode mesure ainsi l'attirance au sein de couples (les termes dans un certain ordre) et non au sein de paires. La paire {جملة, اسم} correspond à deux couples <اسم , جملة > (جملة est le premier terme et اسم apparaît plus à gauche dans le texte) et < جملة, اسم > (c'est cette fois جملة qui apparaît à gauche dans le texte).

Enfin, nous allons sélectionner les cooccurrents dont la fréquence dépasse de manière statistiquement significative la

fréquence due au hasard. Un seuil numérique de 80%[3] est défini a priori pour estimer qu'une relation entre deux termes est significative.

2.3 Enrichissement de l'ontologie

Le traitement précédent fournie une liste de couples des termes (un terme peut un seul mot ou composé par plusieurs mots) pertinents relatifs au domaine. Cette liste est traitée et exploitée pour enrichir l'ontologie. Le principe de l'approche, est de comparer le couple de candidats termes extraits <t1,t2> avec les labels de concepts de l'ontologie, nous allons trouver les cas possibles suivants; t1 (t2) appartient à des labels de l'ontologie et t2 (t1) n'appartient pas, t1 et t2 appartiennent au même temps à des labels de l'ontologie et ni t1 et ni t2 n'appartiennent à des labels de l'ontologie.

2.3.1 Relations par marqueur linguistique

Afin de déceler des relations entre termes, nous allons étudier le contexte autour de ces termes dans une fenêtre de petite taille (exemple : quatre mots) [Koo03]. A partir de ce contexte la méthode va chercher des éléments lexico-syntaxiques permettant de repérer une relation entre ces derniers. Ces éléments sont appelés Marqueurs linguistiques[4].

Exemple « T1 est-un T2 », « T1 partie-de T2 » ,...

Mais comme une même relation peut s'exprimer par différents marqueurs, donc ces derniers sont organisés en catégories ou listes distinctes selon le type de relation à extraire, qui seront incrémentées au fur et à mesure.

[3] Seuil numérique utilisé dans l'extracteur Xtract est 80% [Sma93].

[4] Utilisé dans CAMELEON (un logiciel de recherche de relations lexicales à partir de marqueurs linguistiques [Ség01].)

Ainsi nous aurons, dans chaque liste, une sorte de paradigme d'unités linguistiques dont les catégories sont parfois hétérogènes (noms, verbes, mots outils ou grammaticaux, etc.) mais qui remplissent toujours les mêmes fonctions pour le type de relation.

- Relation hyponymie généralisation « est-un » : la liste = { **هو، هي ، هم، ...**}
- Relation meronymie partie-de : la liste={ **تتألف-من ، تنقسم-الى ، تتكون-من،** }

En conséquence à la particularité de la morphologie arabe au niveau de la vocalisation et d'agglutination, la liste des marqueurs doit avoir toutes les formes agglutinées ainsi que les autres variantes morphologiques susceptibles d'être rencontrées dans les textes.

Nous pouvons toujours rajouter des nouvelles relations ainsi de mettre à jour les listes des relations préexistantes.
Le procédé de mise à jour sera comme suit:

- Dans le cas où un seul terme du couple est retrouvé parmi les labels de concepts de l'ontologie, le second terme du couple sera proposé pour être un nouveau concept dans l'ontologie, et il sera lié au premier concept par une relation définie par le marqueur linguistique.
- Dans le cas où les deux termes sont parmi des labels de concepts de l'ontologie et il n'y avait pas une relation entre ces deux concepts ; une nouvelle relation sera proposée à partir du marqueur linguistique.

- Dans le cas où ni le premier terme et ni le second n'appartient à des labels de l'ontologie. Le procédé ne fait rien et laisse ces cas pour les prochaines exécutions.

2.3.2 Relation hiérarchique

Si les marqueurs linguistiques sont absents dans le contexte des termes, l'approche reposera sur une relation parent-enfant où le terme parent est plus général que le terme enfant. Cette relation entre termes est extraite d'après la cooccurrence asymétrique de termes. La relation est caractérisée par les deux règles suivantes :

$$P(x/y) \geq 0.8. \quad (1)$$

$$P(y/x) < P(x/y). \quad (2)$$

où p(x/y) est la probabilité d'apparition du terme x puis du terme y, inversement pour p(y/x). La première règle assure que les deux termes apparaissent suffisamment dans le corpus (en l'occurrence dans 80% des cas)[5]. D'après la deuxième règle, x subsume y. Le terme apparaissant le plus souvent est choisi comme parent. Les relations extraites à partir des deux règles citées sont ensuite nettoyées en supprimant les relations redondantes par rapport à la propriété transitive de la relation. Si les relations a subsume b, a subsume c et b subsume c sont extraites, la relation a subsume c peut être supprimée parce qu'elle est déductible des deux autres [Her06].

Le procédé de mise à jour sera comme suit :

- Premier terme du couple est retrouvé parmi les labels de concepts de l'ontologie et le second terme du couple n'appartient pas, alors ce dernier sera proposé pour être un

[5] Le seuil de 80% est expérimenté dans les travaux de [HeM06].

nouveau concept fils lié au premier concept par la relation de subsumption «est-un».

- Deuxième terme est parmi des labels de concepts de l'ontologie et le premier terme du couple n'appartient pas, alors ce dernier sera proposé pour être un nouveau concept père du deuxième concept par la relation de subsumption (« est-un »).
- Deux termes sont parmi des labels de concepts de l'ontologie et il n'y avait pas une relation entre ces deux concepts ; une nouvelle relation de subsumption (« est-un ») sera proposée.

- Tous les deux n'appartiennent pas à des labels de l'ontologie. Le procédé ne fait rien et laisse ces cas pour les prochaines exécutions.

3. Expérimentation et résultat

Nous avons programmé l'approche à l'aide du langage de programmation Python dû à sa puissance et grâce à sa bibliothèque NLTK[6] (Natural Language Toolkit).

3.1 Constitution du corpus

Nous avons constitué un échantillon de corpus à partir des documents écrits en arabe recherchés dans les ressources suivantes: des livres sur la linguistique arabe, des articles des revues N°7 et N°08 (AL-LISANIYYAT Linguistique) publiées par le CRSTDLA[7] et via le Web en introduisant des mots clés spécifiques lié au domaine dans le moteur de recherche par les requêtes les suivantes :

[6]. http://nltk.sourceforge.net/index.php/Main_Page

[7]. CRSTDLA Centre de Recherche Scientifique et Technique pour le Développement de la Langue Arabe (Alger)

نظرية اللسانية ،علم الدلالـة، الألفاظ في النحو العربي، اللسانيات الحديثة قواعد في اللغة العربية، الأبنية ودورها في اللغة العربية ، الأوزان ، خصائص اللغة العربية، النحو العربي .

Les documents trouvés sont téléchargés, sélectionnés et préparés manuellement (en supprimant les tableaux les schémas et les graphes).

La table 3 nous montre les caractéristiques du corpus.

Table 3. Caractéristique techniques du corpus

Nombre total des documents	57
Nombre total des mots	468 554
Volume total en K octet	2 742 Ko

3.2. Phase du prétraitement

3.2.1. Segmentation et normalisation

Segmentation les textes à des séquences de mots en détectant les délimiteurs de mot tels que l'espace ou la ponctuation. Nous avons utilisé la liste des symboles de ponctuation spécifique à l'arabe: ["،",".","؟",",",":","؛" ...]

Dans la normalisation nous avons éliminé tous les éléments qui ne portent pas d'information et qui augmentent le délai de traitement. Il s'agit des caractères spéciaux, des chiffres, des caractères latins, des abréviations et des lettres isolées et la dévoyellation. Nous avons recensé tous les caractères spéciaux qui sont utilisés dans les textes arabes ("–","/",",",",","«","+","%"…).

- Sélection de 417 059 mots et élimination 51 495 (11%)

3.2.2 Elimination des mots vides (1)

Nous avons constitué cette liste des mots vides à partir du corpus sur deux principes; leurs fréquences et leurs contenus informationnels. Nous avons trié les mots les plus utilisés dans le corpus suivants leurs fréquences et puis nous avons sélectionné manuellement parmi eux les mots qui ne portent pas d'informations liées au domaine. Au total nous avons trié 455 mots vides.

3.2.2.1 Résultat

La liste n'est pas exhaustive, donc nous devons toujours la mettre à jour par des nouveaux mots ou des nouvelles formes morphologiques d'un même mot, puisque le résultat du traitement (segments répétés) est fortement dépend de cette étape. 116 137 mots sont éliminés soit 27,9%.

3.2.3. Lemmatisation légère (Light-Stemming)

En supprimant des préfixes et des suffixes suivant une lise prédéfinie sauvegardée sur des fichiers.

3.2.3.1 Résultat

Nous avons trouvé des cas où un même mot apparaît encore sur plusieurs formes morphologiques ce qui va diminuer les performances du traitement.

3.2.3.2 Suggestion

Pour remédier ce problème, nous pouvons utiliser un analyseur morphologique pour la lemmatisation complète qui va améliorer nettement la qualité du traitement.

3.2.4 Elimination des mots vides (2)

Nous avons besoin d'éliminer les mots vides nouveau, puisque dans le résultat de la lemmatisation légère nous avons trouvé encore des mots non significatifs après la suppression des préfixes et des suffixes :

Exemple (des cas sont présents: بعده-بعد ، اخرى-الاخرى)

3.2.4.1 Résultat. 39 207 mots sont éliminés (13%).

3.3 Phase du traitement

3.3.1Extraction des segments répétés

Nous devons fixer les paramètres suivants :

- Taille maximale du segment = 4 mots. désigne la taille maximale d'un terme complexe, en arabe est constitué au maximum de 4 mots.
- Seuil de pondération: Le poids d'un terme est calculé par sa fréquence totale. Le seuil de pondération d'un mot simple est égal à 100 et le seuil de pondération d'un terme composé est égal à 20. Le nombre de 100 et 20 sont choisi au hasard et relativement à la taille du corpus.

3.3.1.1. Résultat.

Le programme extrait au total 281 200 segments différents mais il sélectionne uniquement une liste de 445 segments conformément aux seuils définis précédemment. En analysant cette liste, nous avons distingué les remarques suivantes:

1. Apparition des mots qui n'ont pas de liens avec notre domaine (noms de personnes, noms d'objets...). Nous

pouvons ajouter ces mots à la liste des mots vides et refaire le traitement.

2. Deux formes morphologiques d'un même mot sont repérées comme deux segments différents. Exemple (لغه ، لغوى، لغات ، للغه) (حرف ، حروف) (عنصر ، عناصر). Nous pouvons regrouper les différentes formes morphologiques sous une même forme et les remplacer automatiquement dans le corpus et refaire le traitement.

La table 4 présente un échantillon des segments sélectionnés.

Table 4. Segments sélectionnés par le programme

Segment	Fréquence	Segment	Fréquence	Segment	Fréquence
لغه	5071	فاعل	592	لفظ معنى	84
دلال	2449	ظاهر	579	مفعول مطلق	83
فعل	2042	ضمير	575	جمل اسم	78
اسم	1938	مفعول	540	فعل ماض	76
…	…	…	…	…	…

3.3.2 Extraction des cooccurrents

Nous avons fixé les paramètres suivants :

- Fenêtre de cooccurrence =10 mots.
- Seuil de cooccurrence = 80 %.
- Seuil de co-fréquence =100 est le nombre de fois minimum de cooccurrence.

Le programme fournit le résultat dans un fichier balisé où chaque ligne contient les cooccurrents leurs fréquences et leur coofréquence. Comme l'exemple suivant :

< t1="نصب" t2="فتح" Ft1=672 Ft2=129 CF=211/>

Suggestion. Ce fichier doit être validé par un expert di domaine (un linguiste).

Conclusion

Dans cet article nous avons présenté notre approche pour la construction automatique de l'ontologie à partir d'un corpus du domaine « linguistique arabe ». L'approche a réutilisé les techniques d'extraction d'information pour extraire les nouveaux termes qui pourront dénoter des éléments de l'ontologie (concept, relation). Pour analyser les textes du corpus, nous avons employé deux méthodes statistiques ; la méthode des *segments répétés* pour repérer les candidat-termes et la méthode de *cooccurrence* pour faire la mise à jour de l'ontologie. Nous avons constitué ainsi un corpus du domaine linguistique par la récupération des textes à partir des articles à travers les revues, mémoires et livres du domaine et aussi par la récolte des documents via le Web. Ce corpus est a été prétraité pour lever certaines ambiguïté, réduire le nombre d'opérations effectuées et d'adapter le corpus suivant notre objectif.

Nombreuses perspectives s'offrent suite de notre travail. Parmi elles; nous avons proposé une ontologie basée sur un schéma hiérarchique des concepts qui représentent les notions fondamentales de la linguistique arabe, cette ontologie pourra être utile pour développer des outils de TAL qui analysent des textes écrits en arabe. Une deuxième perspective serait d'exploiter nos techniques et méthodes statistiques d'extraction de connaissances pour d'autres travaux qui manipulent des textes arabes (exemple: Extraction terminologique, création des dictionnaires électroniques ou thésaurus, etc.).

Références

[AbD08] RAMZI Abbès, Joseph DICHY « Extraction automatique de fréquences lexicales en arabe » JADT 2008 :« 9eme Journées internationales d'Analyse statistique des Données Textuelles » Université Lumière Lyon 2, ICAR-CNRS.

[BoA03] Didier BOURIGAULT et Nathalie AUSSENAC-GILLES. «Construction d'ontologies à partir de textes ». Actes de la conférence sur le traitement automatique des langues (TALN), France, Juin 2003.

[Dar03] DARWISH K « *Probabilistic methods for searching OCR-Degraded Arabic Text* » Thèse de Doctorat Université de Maryland 2003.

[Dou05]. F. S. DOUZIDIA, G. LAPALME « *Un système de résumé de texte en arabe* » université de Montréal exposé en deuxième conférence International de "l'Ingénierie de la Langue et Ingénierie de l'Arabe " Alger 2005.

[Far03] : FARRAR, William D. Lewis, and D. TERENCE « *An Ontology for Linguistic Annotation* » Department of Linguistics, University of Arizona 2003.

[HeM06] N. HERNANDEZ, J. Mothe « **TtoO**: une méthodologie de construction d'ontologie de domaine à partir d'un thésaurus et d'un corpus de référence » IRIT, Toulouse, 2006.

[Her06] Nathalie HERNANDEZ « Ontologies de domaine pour la modélisation du contexte en recherche d'information » Thèse de Doctorat à l'Université Paul Sabatier France 2006.

[Koo03] S. KOO, S.Y. LIM, S.J. Lee, « Building an Ontology based on Hub Words for Informational Retrieval », In Proceedings of the IEEE/WIC International Conference on Web Intelligence, 2003.

[Mar03] Elizabeth MARSHMAN «Construction et gestion des corpus : Résumé et essai d'uniformisation du processus pour la terminologie » Janvier 2003, "Observatoire de linguistique Sens-Texte" (OLST) de l'Université de Montréal.

[RoF02] F. ROUSSELOT et P. FRATH, « Terminologie et Intelligence Artificielle » (12èmes rencontres linguistiques), Presses Universitaires de Caen, 2002.

[Sma93] Frank. SMADJA, « Retrieving collocations from text: Xtract, Computational Linguistics », université de Columbia 1993.

[Ség01] Patrick SEGUELA « Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques » thèse présentée à l'UNIVERSITÉ TOULOUSE III. 2001.

[Ver04] Jacques VERGNE « Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource » JADT 2004 :« 7eme Journées internationales d'Analyse statistique des Données Textuelles » GREYC – Université de Caen.