REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEURE ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE ABOU-BEKR BELKAID – TLEMCEN

THÈSE

Présentée à :

FACULTE DES SCIENCES – DEPARTEMENT D'INFORMATIQUE

Pour l'obtention du diplôme de :

DOCTORAT EN SCIENCES

Spécialité: Informatique

Par :

Mr BENABDALLAH Ali

Sur le thème

Construction semi-automatique des ontologies à partir des documents textuels arabes

---

Soutenue publiquement le 28 juin 2017 à Tlemcen devant le jury composé de :

M[r] CHIKH Azedine	Professeur	Université de Tlemcen	Président
M[r] ABDERRAHIM Mohammed El Amine	MCA	Université de Tlemcen	Directeur de thèse
M[r] BENAMAR Abdelkrim	MCA	Université de Tlemcen	Examinateur
M[r] BELALEM Ghalem	Professeur	Université d'Oran 1	Examinateur
M[r] BOUAMRANE Karim	Professeur	Université d'Oran 1	Examinateur

Laboratoire de Traitement Automatique de la Langue Naturelle
BP 119, 13000 Tlemcen - Algérie

Résumé

La tâche de construction d'une ontologie à partir d'un corpus textuel commence par la phase de conceptualisation, qui consiste à extrait les concepts de l'ontologie. Ces concepts sont reliés par des relations sémantiques. Dans le cadre de cette thèse, nous présentons une contribution pour la construction semi-automatique d'une ontologie à partir d'un corpus textuel arabe, en commençant d'abord par la collecte des documents et le prétraitement du corpus à travers la normalisation, puis la suppression des mots vides et la lemmatisation; Ensuite, pour extraire les termes de notre ontologie, une méthode statistique pour extraire des termes simples et complexes appelée « méthode des segments répétés » est appliquée. Pour sélectionner les segments avec un poids suffisant, nous appliquons deux filtres : un filtre de pondération TF-IDF (Term Frequency-Inverse Document Frequency) et un filtre coupant. Pour relier ces termes par des relations sémantiques, nous appliquons une méthode d'apprentissage automatique des marqueurs linguistiques à partir du texte. Cette méthode nécessite un ensemble de paires de relations, qui sont extraites à partir de deux ressources externes: un dictionnaire arabe de synonymes et d'antonymes et une base de données lexicale Arabe.

A la fin de cette thèse, nous présentons les résultats de notre expérimentation en utilisant notre corpus textuel. L'évaluation de notre approche montre des résultats encourageants en termes de rappel et de précision.

Mots clés :

Traitement Automatique du Langage Naturel (TALN), Prétraitement du texte, construction des ontologies, extraction des termes, méthode des segments répétés, relations sémantiques, marqueurs linguistiques.

Abstract

The task of building an ontology from a textual corpus begins with the conceptualization phase, which consists of extracting ontology concepts. These concepts are connected by semantic relations. In this thesis, we present a contribution to the semi-automatic construction of an ontology from an Arabic text corpus, beginning with the collection and pre-processing of the corpus through normalization, then deletion of stop-words and lemmatization; Then, to extract the terms of our ontology, a statistical method for extracting simple and complex terms called the "repeated segment method" is applied. To select the segments with a sufficient weight, we apply two filters: a TF-IDF (Term Frequency-Inverse Document Frequency) weighting filter and a cutting filter. To link these terms by semantic relations, we apply an automatic method of learning the linguistic markers from the text. This method requires a set of pairs of relationships, which are extracted from two external resources: an Arabic dictionary of synonyms and antonyms and an Arabic lexical database.
At the end of this thesis, we present the results of our experimentation using our textual corpus. The evaluation of our approach shows encouraging results in terms of recall and precision.

Keywords:

Natural Language Processing (NLP), Text Preprocessing, ontology construction, terms extraction, repeated segments method, semantic relations, linguistic markers.

ملخص

مهمة بناء الأنطولوجيات انطلاقا من الملفات النصية تبدأ بمرحلة مهمة وهي مرحلة التصميم، والتي تعتمد على استخراج المفاهيم المكونة لهذه الأنطولوجيات. تربط هذه المفاهيم بواسطة العلاقات المعنوية . من خلال هذه الأطروحة، سنقدم إسهاما في البناء شبه الأوتوماتيكي للأنطولوجيات عن طريق النصوص العربية، بدءا بجمع ومعالجة هذه النصوص العربية ومرورا بتقييمها و إزالة بعض الكلمات (كحروف العطف و حروف الجر و الضمائر المنفصلة) ثم استخراج جذورها.

لاستخراج المصطلحات المكونة للأنطولوجيا ، يتم تطبيق طريقة إحصائية لاستخراج المصطلحات البسيطة والمعقدة والتي تسمى بطريقة "المقاطع المتكررة".

لتحديد المقاطع الأكثر أهمية من بين المقاطع المتحصل عليها ، نستخدم نوعين من المصفيات، الأول يسمى بـ(تردد المصطلحات والتردد العكسي للوثائق) والثاني يسمى بـ(مصفاة القطع).

لربط هذه المصطلحات بواسطة العلاقات المعنوية، نطبق طريقة التعلم التلقائي للعلامات اللغوية عن طريق النصوص العربية. يتطلب هذا الأسلوب مجموعة أزواج من العلاقات، والتي تستخرج اعتمادا على موردين خارجيين وهما : قاموس عربي للمترادفات والأضداد و قاعدة بيانات معجمية عربية .
في نهاية هذا البحث، سنقدم نتائج تجربتنا باستخدام نصوص عربية .
تقييم عملنا أظهر نتائج مشجعة باستخدام مقياسين وهما مقياس التذكير ومقياس الدقة.

الكلمات المفتاحية:

المعالجة الآلية للغة الطبيعية، المعالجة الأولية للنص، بناء الأنطولوجيات، استخراج المصطلحات، العلاقات المعنوية، العلامات اللغوية.

Dédicaces

Je dédie cette thèse :

- A l'esprit de mon père (que Dieu bénisse son âme)
- A ma mère, pour son amour, sa confiance son soutien moral et surtout ses sacrifices...
- A mon oiselle Nour Elhouda ...
- A mes petits-neveux : Safouane, Siraj, Ayoub et Salah-eddine.
- A toute ma famille et mes amis ...

ALi...

Remerciements

Tout d'abord, mes remerciements sont adressés à ***Allah*** *qui m'a donné la puissance et le courage pour achever ce travail*

Je tiens à exprimer ma profonde gratitude à Monsieur **ABDERRAHIM Mohammed El-Amine**, mon directeur de thèse, Maitre de conférences « A » à la Faculté de technologies de l'Université de Tlemcen, de m'avoir encadré avec un grand intérêt et une grande compétence, pour l'intérêt qu'il a voulu porter à mon travail, et surtout pour sa patience avec moi.

Je remercie aussi Mon cher camarade **ABDERRAHIM Mohammed Alaa-Eddine,** Maitre de conférences « B » à l'université de Tiaret, pour son aide, ses conseils, et surtout pour son encouragement qui m'a beaucoup aider pour achever ce travail.

Je remercie **Mr CHIKH Azedine**, Professeur à l'université de Tlemcen de m'avoir fait l'honneur d'être le président de la soutenance de cette thèse et à qui je souhaite exprimer ma profonde reconnaissance.

Un grand merci est adressé aussi à **Mr BENAMAR Abdelkrim,** Maitre de conférences classe « A » à la Faculté de sciences de l'Université de Tlemcen, pour avoir accepté d'examiner mon travail.

Je remercie **Mr BELALEM Ghalem**, Professeur à l'université d'Oran, pour l'intérêt qu'il a porté à mon travail et pour le temps qu'il a consacré pour examiner cette thèse.

J'adresse également mes remerciements à **Mr BOUAMRANE Karim**, Professeur à l'université d'Oran, qui a accepté d'être examinateur et de participer au jury.

Et finalement, je remercie tous ceux qui m'ont aidé de près ou de loin pour achever ce travail.

Table des matières

Chapitre 2 : Etat de l'art sur la construction des ontologies à partir des textes

Chapitre 3 : Notre Contribution

Table des figures

Table des tableaux

ACRONYMES

A Nearly-New Information Extraction System	ANNIE
Apprentissage Naturel Automatique	ANA
Arabic WordNet	AWN
ARCHitecture for ONTological Elaborating	ARCHONTE
DARPA Agent Markup Language	DAML
Differential Ontologies Editor	DoE
General Architecture for Text Engineering	GATE
Intelligence Artificielle	IA
Java Annotation Patterns Engine	JAPE
Java DataBase Connectivity	JDBC
Knowledge Acquisition supported Semi-automated Ontology building	KASO
MAXimum ENTropy	MAXENT
Ontology Design Environment	ODE
Ontology Inference Layer	OIL
Ontology Web Language	OWL
Relational DataBase Management System	RDBMS
Resource Description Framework	RDF
Resource Description Framework schémas	RDFs
Robust Accurate Statistical Parsing	RASP
Semantically Enabled Knowledge Technology	SEKT
Simple Matching Coefficient	SMC
Software for the Extraction of N-ary Textual Associations	SENTA
Suggested Upper Merged Ontology	SUMO
Support Vector Machine	SVM
Term Frequency-Inverse Document Frequency	TF-IDF
Traitement Automatique du Langage Naturel	TALN
Transitioning Applications to Ontologies	TAO
Waikato environment for knowledge analysis	Weka

Introduction générale

1. Contexte et problématique

Depuis son émergence, dans les recherches d'extraction et de modélisation de connaissances, la notion d'ontologie s'est rapidement diffusée dans un grand nombre de domaines de recherche en informatique. Définie comme la représentation formelle et consensuelle au sein d'une communauté d'utilisateurs, des concepts propres à un domaine et des relations qui les relient, la notion d'ontologie apparaît comme un moyen de représenter explicitement et de partager des objets d'un domaine ainsi que leur sémantique. Compte tenu du caractère prometteur de cette notion, de nombreux travaux portent sur l'exploitation des ontologies dans des domaines aussi divers que le TALN (Traitement Automatique de la Langue Naturelle), la recherche d'information, le commerce électronique, le web sémantique, la spécification des composants logiciels, l'intégration de systèmes d'information, etc..

L'efficacité de tous ces travaux dépend de l'existence ou non d'une ontologie de domaine susceptible d'être exploitée. Or, la conception d'une telle ontologie s'avère particulièrement difficile si l'on souhaite qu'elle fasse l'objet de consensus dans une communauté assez large. Un moyen très largement utilisé pour atteindre cet objectif est de partir d'éléments préexistants dans le domaine : corpus de textes, dictionnaire, taxonomies, thésaurus, fragments d'ontologies préexistants, des schémas de bases de données, etc. et de les exploiter comme connaissance a priori pour la construction progressive d'une ontologie du domaine.

Cette tâche correspond à un apport de connaissances et est difficilement automatisable. Dans le cas de la construction d'ontologies à partir de textes, par exemple, il existe néanmoins, et en particulier lorsque des corpus importants sont utilisés, la possibilité de recourir à des outils informatiques pour faciliter l'extraction des *termes* et des *relations*, l'analyse syntaxique et distributionnelle, l'identification des synonymes et homonymes, etc., et cela, jusqu'à la représentation formelle de l'ontologie.

Par ailleurs, s'il existe des outils tels que Protégé[1], OntoEdit[2], etc., utilisés pour éditer formellement une ontologie supposée déjà conçue, et s'il existe également plusieurs outils de

[1] http://protege.stanford.edu/

traitement automatique de la langue (TAL) permettant d'analyser automatiquement les corpus de textes et de les annoter sur les points de vue syntaxique, distributionnel et statistique, il est difficile de trouver une procédure globalement acceptée, ni a fortiori un ensemble d'outils supports permettant de concevoir une ontologie de domaine de façon progressive et explicite à partir d'un ensemble de ressources informationnelles relevant de ce domaine.

Nous nous intéressons dans ce travail de thèse, à la construction semi-automatique d'une ontologie à partir de textes et plus particulièrement une ontologie pouvant représenter un domaine spécifique en langue Arabe.

2. Objectif et contributions

La construction entièrement manuelle d'une ontologie est une tâche rude, complexe et nécessite beaucoup de temps et de ressources. Le recours à des méthodes automatiques ou semi-automatiques est devenu indispensable, toutefois le recours aux experts pour la validation des résultats au cours de ce processus de création permet d'aboutir à une ontologie plus accomplie et plus précise.

Notre objectif dans cette thèse est de proposer et d'implémenter une approche pour la construction semi-automatique d'une ontologie de domaine à partir des textes arabe. L'objectif principal de cette approche est l'extraction des éléments constituant l'ontologie à partir d'un corpus de textes, qui sont principalement les termes et les relations sémantiques reliant ces termes.

Pour atteindre cet objectif, nous avons proposé un processus qui se base principalement sur trois grandes phases : d'abord, nous avons commencé par la collecte et le prétraitement de notre corpus. Dans cette phase, et après la collecte et le filtrage des documents constituant le corpus, le texte passe par trois étapes : la normalisation, la suppression des mots vides et la lemmatisation. Ensuite, dans la deuxième phase, le texte obtenu sera utilisé pour extraire les termes simples et composés par une méthode statistique qui est la méthode des segments répétés. Les segments trouvés dans le texte seront filtrés deux fois : par le filtre TF-IDF (Term Frequency-Inverse Document Frequency) et le filtre coupant. La troisième phase consiste à relier les termes simples et composés trouvé précédemment par trois types de relations sémantiques : Hyperonyme, synonyme et antonyme, en se basant sur les documents textuels

[2] http://www.daml.org/tools/#OntoEdit

de notre corpus, sur un dictionnaire et sur une base de données lexicale. La méthode utilisée dans cette phase se base sur l'apprentissage de marqueurs linguistiques à partir du corpus de texte, des ressources externes et une intervention d'un expert du domaine.

3. Organisation de la thèse

Cette thèse est structurée comme suit :

Le premier chapitre est consacré à tous ce qui concerne les ontologies dont : les constituant, la classification, les langages de représentation, les méthodes de construction et les outils d'édition et de construction des ontologies à partir des textes.

Le deuxième chapitre est un état de l'art, il se décompose en deux parties ; la première concerne les approches et outils d'extraction de termes et aussi de relations ; et la deuxième partie contient des résumés de quelques travaux sur la construction d'ontologies à partir des textes arabes. Il convient de dire que, dans ce chapitre, nous avons mis l'accent sur les outils qui supportent ou qui peuvent être adaptés à la langue arabe.

Le troisième chapitre, est consacré à la présentation de la méthode proposée pour la construction semi-automatique d'ontologies à partir de textes arabes : extraction de termes et de relations. A la fin de chaque phase, nous détaillons les résultats obtenus pour cette phase avec une évaluation et discussion des résultats obtenus.

A la fin, nous présentons une conclusion générale avec quelques perspectives de ce modeste travail de recherche.

Chapitre 1
Construction des ontologies

1. Introduction

La quantité de plus en plus croissante d'information dans tous les domaines a généré un besoin capital d'organisation et de structuration des contenus de documents, disponibles généralement sur le web. Les ontologies en sont un moyen prometteur et qui ne cesse de donner ses preuves. Leurs applications sont multiples : indexation, recherche d'informations, traduction automatique, e-Learning etc. Les principaux buts de la construction des ontologies sont la partageabilité, la portabilité, la réutilisabilité et la capitalisation de la connaissance et de l'expertise d'un domaine.

Parce que l'information n'est pas statique, parce qu'elle se modifie, s'enrichisse, s'altère avec le temps et qu'elle vienne de différentes sources, nous avons besoin d'outils et de modèles qui permettent aux utilisateurs et aux experts du domaine de constituer, consulter et maintenir à jour leurs connaissances du domaine.

2. Qu'est ce qu'une ontologie

Le mot *ontologie* qui vient du grec ***ontos*** =être et ***logos***= études, appartient à la philosophie ancienne grecque, Aristote le définit comme la science de l'Être en tant qu'être [**Welty & Guarino, 2001**]. Il est difficile de définir ce qu'est une ontologie d'une façon définitive. Le mot est en effet employé dans des contextes très différents touchant à la philosophie, la linguistique ou l'intelligence artificielle.

Une définition, au sens strict, est donnée en juin 1993, par Gruber [**Gruber, 1993]**, et qui est la plus citée en informatique plus précisément en intelligence artificielle (IA) : *« An ontology is an explicit specification of conceptualization.»* à savoir : «*Une ontologie est une spécification explicite d'une conceptualisation* ». L'expression *spécification explicite* signifie, que la conceptualisation est représentée dans un langage qu'il soit naturel (arabe, français..) ou formel (logique de description, graphes conceptuels..).

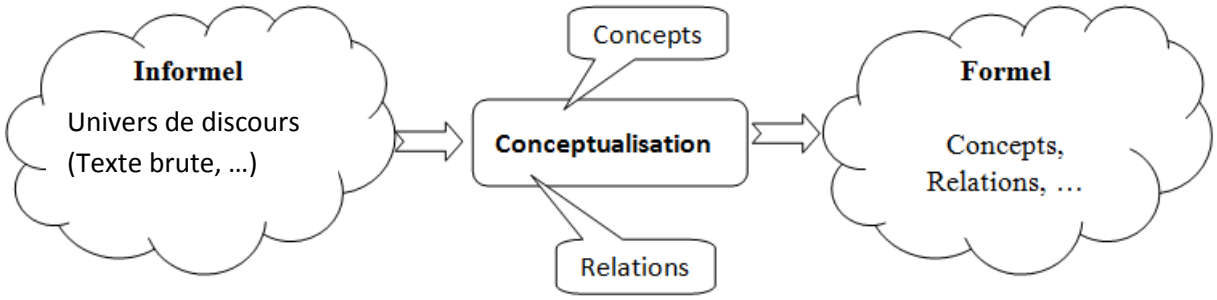

Figure I.1 : La conceptualisation

Une autre définition, peut être plus rigoureuse : « *Une ontologie implique une certaine vue du monde par rapport à un domaine donné. Cette vue est souvent conçue comme un ensemble de concepts (entités, attributs, processus, leurs définitions et leurs interrelations). On appelle cela une conceptualisation* » [**Charlet et al, 1996**]. Une ontologie peut prendre différentes formes mais elle inclura nécessairement un vocabulaire de termes et une spécification de leur signification [**Baneyx & Charlet, 2006**].

3. Constituants d'une ontologie

En informatique, deux types d'ontologies sont fréquemment distingués. Les ontologies de haut niveau fournissent des définitions pour des concepts généraux tels que les notions de *processus*, d'*objet* ou d'*événement* afin de servir de fondement pour des ontologies plus spécifiques dites de domaine **[Ian Niles, 2001, Gangemi et al., 2003]**. Les ontologies de domaine sont liées à un univers du discours particulier. Elles décrivent et représentent la connaissance existant dans le domaine correspondant à cet univers. Les ontologies de domaine sont plus spécifiques que les ontologies noyaux de domaine (appelées core-ontologies) qui englobent les concepts généraux d'une discipline, par exemple le droit **[Teguiak, 2012].**

Dans ce travail de thèse, nous nous intéressons aux ontologies de domaine qui permettent de décrire la sémantique des objets d'un domaine d'étude. Ainsi, par la suite, le mot *ontologie* est utilisé pour désigner une *ontologie de domaine*.

Ces ontologies de domaine représentent la sémantique des concepts d'un domaine en termes de concepts, propriétés, instances et relations. Dans ce qui suit, nous allons aborder ces notions plus en détails :

3.1. Les concepts

Un concept, également appelé *classe* dans certains travaux ou outils, représente l'idée que l'on se fait d'un terme : le contenu. Il est porteur d'une connaissance. Il peut designer un objet concret comme :(**حاسوب** =*ordinateur)* ou abstrait comme :*(* **معلومة** = *information).*

3.2. Les propriétés

La propriété est une caractéristique qui qualifie un concept et qui peut généralement être dotée d'une valeur. Si nous prenons l'exemple précédent (**حاسوب** =*ordinateur*) nous pouvons designer quelques propriétés comme : (**نوع_الحاسوب**=*Type_de_l'ordinateur*), (سعة_التخزين=*capacité_de_stockage*), (سرعة_الحساب=*vitesse_de_calcul*).

3.3. Les facettes

Les facettes sont des restrictions sur les valeurs des propriétés, si nous prenons le concept (**حاسوب_شخصي**= *ordinateur personnel),* la facette de la propriété (**نوع_الحاسوب**=*Type_de_l'ordinateur*) de ce concept sera une liste finie de tous les ordinateurs personnels, en l'occurrence (**حاسوب_مكتبي**=*Ordinateur_de_bureau,* **حاسوب_محمول**=*Ordinateur_portable,* **حاسوب_لوحي** =*Tablette_PC,* etc.)**.** Quant aux facettes de (**سرعة_الحساب**=*vitesse_de_calcul*), ce sera tout simplement un entier suivi d'une unité de mesure comme « **جيقا_هرتز**=GHz » ou «**ميقا_هرتز**=MHz ».

3.4. Les instances

Les instances d'un concept concret sont des éléments singuliers de ce concept, aussi appelées individus dans certains travaux. Les instances ne sont nécessaires que lorsque l'objectif de l'ontologie est de servir à la construction d'une base de connaissances.

3.5. Les relations

Les relations sont un type d'interaction entre deux concepts. La relation la plus utilisée est sans doute celle qui établit la hiérarchie de la structure ontologique, c'est la relation de généralisation-spécialisation (عبارة_عن= *est_un).* B *est_un* A, exprime le fait que le concept B est un sous concept du concept A, dans le sens où B hérite de toutes les propriétés de A et a forcement des propriétés spécifiques.

Exemple : (الحاسوب **عبارة_عن** آلة**),** (الطابعة **عبارة_عن** آلة) et (حاسوب **عبارة_عن** الحاسوب الشخصي)

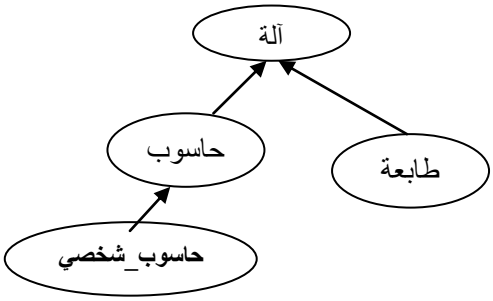

Figure I.2: Exemple de la relation de généralisation-spécialisation

Il existe deux catégories de relations : celles qui sont indépendantes du domaine et celles qui sont étroitement liées au domaine choisi. Les relations indépendantes du domaine sont générales et peuvent être utilisées dans n'importe quel champ de spécialité, les plus connues sont (**عبارة_عن**= *est_un*), (**جزء_من**= *partie_de*), (**مرادف_لـ**= *synonyme_de*), (**ضد**= *opposé_de*) etc. Les relations dépendantes d'un domaine, ont un sens précis dans ce domaine.

Exemple : (**متصل_بـ** = est_connecté_à) dans (**الطابعة متصلة بالكمبيوتر**).

Cette relation ne peut être utilisée que dans certains domaines, et elle a un sens bien précis par rapport à chaque domaine.

4. Classifications des ontologies

Il existe plusieurs méthodes de classification des ontologies qui sont proposées par des groupes de recherche selon l'objectif principal pour lequel l'ontologie a été conçue. La plus courante des classifications est la classification de Gómez-Pérez **[Gomez-Pérez, 2004]**. Elle s'intéresse aux objets que modélisent les ontologies, elle les classe ainsi :

4.1. Ontologies pour la représentation des connaissances

Les ontologies de représentation des connaissances sont utilisées pour formaliser un modèle de représentation des connaissances. On peut par exemple citer l'exemple de l'ontologie de *frame* **[Gruber, 1993]**, qui définit les primitives de représentation des langages à base de frames (classes, instances, slots, facettes, etc.).

4.2. Ontologies de domaine

Elles servent à fournir une modélisation d'un domaine de connaissance donné comme la médecine ou une spécialité en médecine comme l'ophtalmologie. Elles sont réutilisables à l'intérieur d'un domaine donné et modélisent le vocabulaire à l'intérieur de ce domaine

[**Gomez-Perez**, **1999]**. La plupart des ontologies existantes sont des ontologies de domaine [**Psyché et al, 2003]**.

4.3. Ontologies de haut niveau

Une ontologie de haut niveau décrit des concepts très généraux comme l'espace, le temps, la matière, les objets, les événements, les actions, etc. Ces concepts ne dépendent pas d'un problème ou d'un domaine particulier. Un exemple d'une ontologie de haut niveau est : «*Upper Cyc*[3] ».

4.4. Ontologies génériques (méta-ontologie)

Elles ne sont pas spécifiques à un domaine précis, aussi appelée méta-ontologies, elles véhiculent des connaissances génériques qui, bien que moins abstraites que celles modélisées dans l'ontologie de haut niveau, doivent être assez générales pour être réutilisées dans différents domaines. Elles organisent des connaissances factuelles ou des connaissances visant à résoudre des problèmes génériques d'un ou de plusieurs domaines**.** Un exemple d'une telle ontologie : *SUMO*[4] (Suggested Upper Merged Ontology).

4.5. Ontologies de tâches

Elles modélisent le processus d'une tache spécifique ou d'une activité particulière, tel un diagnostic d'une panne par exemple.

4.6. Ontologies d'application

Il s'agit du type d'ontologie le plus spécifique [**Psyché et al, 2003]**. Les concepts que l'on trouve dans ce genre d'ontologies modélisent les concepts d'un domaine particulier dans le cadre d'une application donnée, tel que l'aide au diagnostic d'une panne dans une turbine à vapeur utilisée en électricité par exemple [**Klai & Khadir, 2009**].

[3] https://www.cs.auckland.ac.nz/courses/compsci367s1c/resources/cyc.pdf
[4] http://suo.ieee.org developed in the project *IEEE SUO* Working Group.

5. Méthodologie de construction d'une ontologie

5.1. Stratégies de construction d'une ontologie

Il existe trois stratégies de construction d'une ontologie : *ascendante*, *descendante* et *mixte*. Dans l'*ascendante*, on débute par les feuilles c'est-à-dire les concepts terminaux, ce sont les plus spécifiques puis on généralise au fur et à mesure que nous montons dans la hiérarchie conceptuelle. Par exemple, nous allons commencer par (**حاسوب_محمول** vers **حاسوب_شخصي** vers **حاسوب** vers **آلة** vers **جهاز**).

La stratégie *descendante* consiste à commencer par les concepts les plus génériques et développer la structure vers les concepts les plus spécifiques, elle nous permet de nous arrêter au niveau de détail désiré. Dans ce cas nous commençons par (**جهاز** vers **آلة** vers **حاسوب** vers **حاسوب_شخصي** vers **حاسوب_محمول**). Toute fois, nous pouvons combiner les deux stratégies (stratégie mixte) en commençant par des concepts saillants relativement au domaine, puis selon le besoin, étendre vers le spécifique ou le générique. Dans l'exemple précédent si nous considérons que (**حاسوب**) est important ou saillant par rapport à notre domaine, nous commençons par ce concept et on peut aller, soit vers le haut, soit vers le bas. Donc de (**حاسوب** vers **آلة** vers **جهاز**), ou l'inverse de (**حاسوب** vers **حاسوب_شخصي** vers **حاسوب_محمول**). Dans la pratique, aucune stratégie n'est meilleure par rapport aux autres. La stratégie choisie, dépend principalement du point de vue du concepteur par rapport au domaine à modéliser et à l'objectif de l'ontologie.

5.2. Méthodologies de construction

Bien que l'on s'accorde volontiers à dire qu'il n'existe pas de méthodologie de construction d'ontologie, ceci n'est vrai que dans le sens où aucune méthodologie proposée ne fait l'unanimité de la communauté de l'ingénierie ontologique. En effet toutes les méthodologies proposées respectent juste un certain processus de l'objectif pour lequel elles ont été crées. Nous abordons dans la section suivante les méthodologies les plus importantes qui ont réussi à acquérir une certaine crédibilité dans le monde de la conception des ontologies.

5.2.1. Méthodologie de Uschold et Grüninger

Elle propose plusieurs étapes pour construire des ontologies manuellement **[Uschold & Grüninger, 1996]**, ces étapes se résument en:

a) Identifier l'objectif souhaité et spécifier le domaine concerné.

b) Construire l'ontologie et pour cela définir les concepts, les relations clés et produire des définitions textuelles précises et non ambigües de ces concepts.

c) Evaluer le résultat.

d) Documenter le modèle en éditant des recommandations précises pour chaque étape.

5.2.2. La méthode « Methontology »

Proposée par Fernandez et son équipe à l'université de Madrid **[Fernandez et al, 1997]**, cette méthodologie est largement utilisée. Elle se base sur les dix étapes suivantes:

a) Construire le glossaire des termes qui seront inclus dans l'ontologie, préciser leur définition en langage naturel, identifier leurs synonymes et leurs acronymes.

b) Construire des taxonomies de concepts.

c) Construire des diagrammes de relations binaires.

d) Construire le dictionnaire de concepts qui inclut, pour chaque concept, ses attributs d'instance, ses attributs de classe et ses relations.

e) Décrire en détail chaque relation binaire.

f) Décrire en détail chaque attribut d'instance.

g) Décrire en détail chaque attribut de classe.

h) Décrire en détail chaque constante (les constantes donnent des informations sur le domaine de connaissances).

i) Décrire les axiomes formels.

j) Décrire les règles utilisées pour contraindre le contrôle et pour inférer des valeurs aux attributs.

5.2.3. Méthodologie de Guarino et Welty

C'est une méthode qui n'est pas un guide de construction d'ontologies, mais plutôt une étape permettant la vérification et la correction d'une structure ontologique construite un peu anarchiquement, entre autres les règles de subsomptions. Elle est à incorporer dans le cycle de vie d'une ontologie **[Welty & Guarino, 2001]**.

5.2.4. Méthode ARCHONTE

ARCHONTE (ARCHitecture for ONTological Elaborating), est proposée par Bruno Bachimont **[Bachimont, 2000]**, cette méthode comporte trois étapes :

a) Choisir les termes pertinents du domaine et normaliser leur sens en précisant les relations de similarités et de différences que chaque concept entretient avec ses concepts frères et son concept père.

b) Formaliser les connaissances, c'est-à-dire ajouter éventuellement des propriétés et des axiomes à des concepts.

c) Opérationnaliser dans un langage de représentation des connaissances.

6. Langages de représentation des ontologies

Pour pouvoir exploiter une ontologie et la partager par un grand nombre d'utilisateur, il faut l'exprimer dans un langage permettant son utilisation sur différentes applications et plateformes. Ce langage doit répondre aux exigences des utilisateurs potentiels de cette ontologie. Il existe un grand nombre de langages développés à cet effet. Tous sont basés sur la syntaxe XML, bien que XML lui-même ne soit pas un langage de représentation des connaissances ontologiques. Nous allons citer dans cette section les plus utilisés et donc les plus enclins à respecter une certaine standardisation.

6.1. RDF & RDFs

Les initiales RDF correspondent à « Resource Description Framework », ou cadre de description de ressources en français, le « s » de schémas est une extension de RDF. Une ressource est simplement une *chose* : Une personne, un livre, un clavier, un article de publication, un bureau, une idée, toute *chose* qui peut être décrite. RDF est un cadre d'applications utilisant l'architecture du Web pour décrire une ressource. Tel HTML qui permet de relier des documents à d'autres documents sur le Web, RDF permet de relier une ressource à d'autres ressources sur le Web. Comme tous ses prédécesseurs, ce langage se base sur la syntaxe d'XML. Doté d'un schéma de représentation riche, incluant des classes, sous-classes, propriétés, sous-propriétés et des règles d'héritage de propriétés.

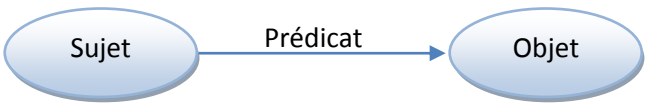

Figure I.3 : Le triplet RDF

Dans un graphe, chaque triplet RDF représente l'existence d'une relation entre les choses symbolisées par les nœuds qui sont joints (voir figure I.3). La structure objet-classe des RDFS

permet de définir des objets du domaine et leurs relations pour rendre compte d'une ontologie [**Baneyx, 2007**].

6.2. OIL

OIL[5] (Ontology Inference Layer) est un langage de représentation d'ontologies qui dérive de RDF. Les principaux fondements du langage OIL sont les langages de frame (tels que OKBC, XOL ou RDF) et les logiques de descriptions. OIL a été défini dans l'objectif de permettre la spécification et l'échange d'ontologies.

6.3. DAML et DAML+OIL

DAML+OIL12 [6] (DARPA [7] *Agent Markup Language*) est un langage permettant la représentation des ontologies. DAML est une combinaison de XML et de RDF permettant de spécifier des objets mais également les relations entre ces objets. La dernière version de DAML se combine avec OIL (DAML+OIL). Ce nouveau langage supporte désormais les types de données primitifs (tels qu'on les trouve dans la norme XML Schéma) et la définition d'un certain nombre d'axiomes comme l'équivalence de classes ou de propriétés [**Baneyx, 2007**].

6.4. OWL

Nous avons vu dans la section 6.1 que RDF et RDFS permettent de définir, sous forme de graphes de triplets, des données ou des métadonnées. Cependant, de nombreuses limitations bornent la capacité d'expression des connaissances établies à l'aide de RDF/RDFS. On peut citer, par exemple, l'impossibilité de raisonner et de mener des raisonnements automatisés sur les modèles de connaissances établis à l'aide de RDF/RDFS. C'est ce manque que se propose de combler OWL.

OWL[8] **(***Ontology Web Language*) a été créé en 2001 par le W3C, il hérite du langage DAML+OIL et doit permettre de représenter des ontologies sur le Web (voir figure I.4) **[Gomez-Pérez, 2004**]. OWL fournit en fait trois sous-langages, d'expressivité croissante, nommés OWL Lite, OWL DL et OWL Full **[Deborah & al., 2004]**.

[5] http://www.ontoknowledge.org/oil/
[6] http://www.daml.org/2001/03/daml+oil-index
[7] DARPA: Defense Advanced Research Projects Agency
[8] http://owl.cs.manchester.ac.uk/

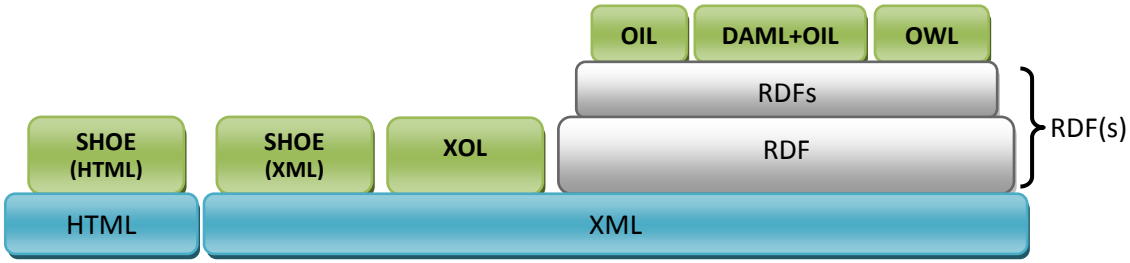

Figure I.4: Les langages d'exploitation des ontologies [Gomez-Pérez, 2004]

OWL est devenu un standard du Consortium W3C qui a publié en 2004 une recommandation définissant le langage OWL fondé sur le standard RDF et en spécifiant une syntaxe XML. Plus expressif que RDFS, il tend à détrôner les autres langages et à s'imposer de plus en plus en maitre absolu **[Deborah & al., 2004]**.

7. Outils de manipulation des ontologies

7.1. Outils d'édition des ontologies

Éditer une ontologie avec un outil adéquat permet son affichage sous forme arborescente, en outre l'intégration de plugins appropriés, permet la visualisation des différents concepts avec toutes les relations qui les relient, cela donne une vue plus globale de la disposition des concepts les uns par rapport aux autres. Certains éditeurs vont plus loin en permettant d'importer ou d'exporter une ontologie d'un format vers un autre, ce qui facilite largement sa portabilité et la génération automatique de fichiers OWL/XML ou RDF **[Patil, 2005]**.

Nous présentons dans cette section les éditeurs les plus importants, certains d'entre eux représentent de véritables plateformes avec de multiples plugins permettant de soumettre des requêtes, de vérifier la consistance et de fusionner des ontologies existantes dans différents formats.

7.1.1 PROTEGE

PROTEGE [9] est un éditeur d'ontologies, distribué en open source par l'institut d'informatique médicale de Stanford **[Kapoor & Sharma, 2010]**. C'est un éditeur hautement extensible, capable de manipuler des formats très divers. Il existe deux moyens pour modéliser une ontologie avec PROTEGE, PROTEGE-Frame et PROTEGE-OWL. Une ontologie en PROTEGE peut être exportée dans différent formats incluant RDF(s), OWL,

[9] http://protege.stanford.edu/

XML schémas. PROTEGE est une plateforme Java, il est flexible et supporte plusieurs langues dont l'Anglais, le Français, l'Arabe, le Chinois, le Russe, etc. Une large communauté de développeurs académiques, de gouvernements et d'entreprises utilise PROTEGE dans divers domaines. L'interface de PROTEGE (voir figure I.5) permet de créer, supprimer, modifier et mettre à jour les concepts, les propriétés, les instances et les relations.

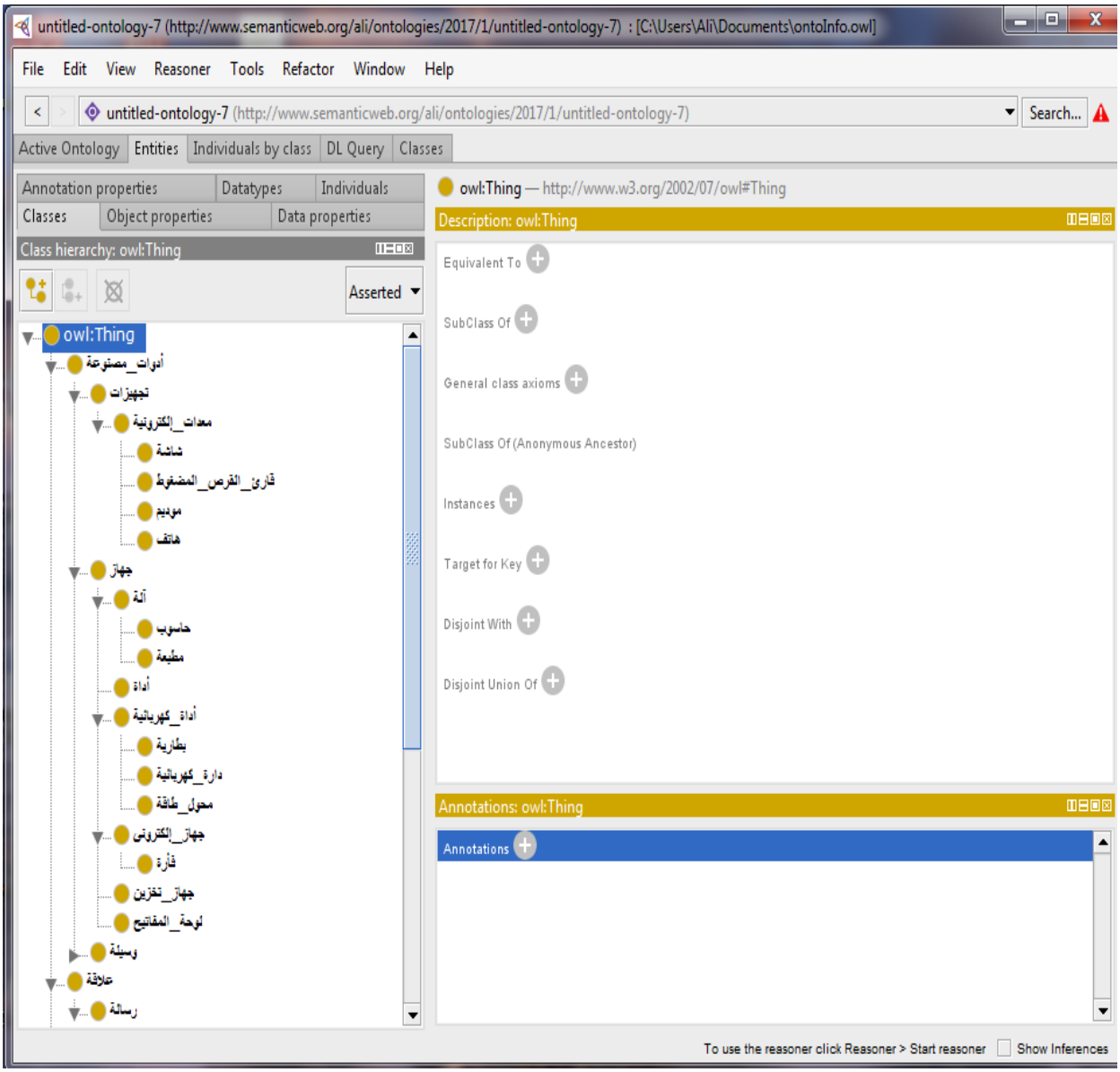

Figure I.5: La hiérarchie Ontologique sous PROTEGE

En plus de la visualisation de la hiérarchie ontologique, PROTEGE permet une visualisation graphique à l'aide de plugins comme OntoGraph (voir figure I.6), il dispose aussi de raisonneurs comme Fact++, Hermitt et Pellet.

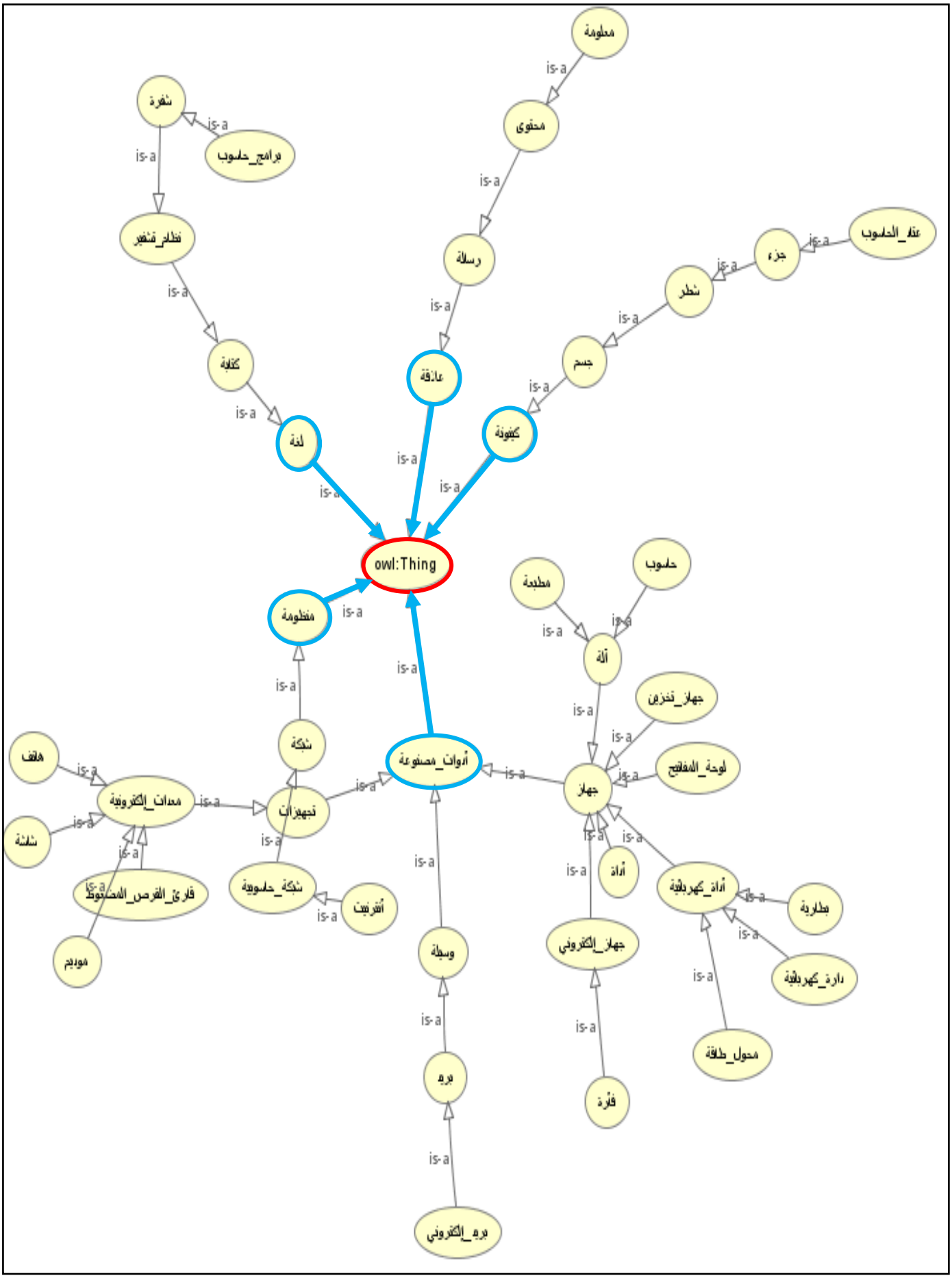

Figure I.6: Visualisation d'ontologie en arabe avec PROTEGE

7.1.2 ODE (ONTOLOGY DESIGN ENVIRONMENT)

C'est un outil de construction des ontologies au niveau connaissances (Méthodologie METHONTOLOGY).

L'objectif de l'ODE est de soutenir l'ontologue tout au long du développement de l'ontologie, de la spécification des exigences, de l'acquisition des connaissances et la conceptualisation, à la mise en œuvre, avec autant d'intégration et d'évaluation que possible.

Pour atteindre cet objectif, ODE cherche à automatiser chaque activité de développement d'ontologie et intégrer automatiquement les résultats de chaque phase à l'entrée de la phase suivante. **[Mariano, 1999]**

7.1.3 JENA

Apache Jena (ou Jena[10] en bref) est une plateforme Java gratuite et open source pour la création des applications Web sémantiques et des applications liées au web sémantique. JENA est composé de différentes API interagissant ensemble pour traiter les données dans des documents RDF, RDFS, OWL et SPARQL. Il fournit un moteur d'inférences permettant des raisonnements sur les ontologies. JENA est maintenant sous Apache Software Licence.

7.1.4 OntoEdit

OntoEdit[11] est un outil mis au point par l'institut AIFB de l'université de Karlsruhe (Allemagne) et qui est maintenant commercialisé par la société Ontoprise GmbH[12]. Il s'inspire de l'approche par frames. OntoEdit est un des seuls éditeurs, avec DOE (voir section 7.1.6), à s'attaquer au problème de la synonymie.

OntoEdit est un éditeur d'ontologie qui a été développé en tenant compte de cinq objectifs principaux **[Sure & al., 2002]** :

1. Facilité d'utilisation.
2. Développement guidé par méthodologie d'ontologies.
3. Développement de l'ontologie à l'aide de l'inférence.
4. Développement d'axiomes ontologiques.
5. Extensibilité grâce à la structure du plug-in.

[10] http://incubator.apache.org/jena
[11] http://www.daml.org/tools/#OntoEdit
[12] http://www.ontoprise.de/

7.1.5 WebOde

WebOde[13] est une plateforme en ligne développée par le groupe Ontological Engineering du département d'intelligence artificielle de la faculté d'informatique de l'université polytechnique de Madrid. C'est un éditeur qui assurait le support de Methontology (voir section 5.2.2). L'éditeur d'ontologie de WebODE permet d'éditer et de naviguer dans les ontologies WebODE et il se base sur des formulaires HTML et des applets Java **[Arpirez & al, 2001]**.

WebOde est construit comme un workbench évolutif, extensible et intégré qui soutient la plupart des activités impliquées dans le processus de développement de l'ontologie (conceptualisation, raisonnement, échange, etc.) et fournit un ensemble complet de services liés à l'ontologie qui permettent l'interopérabilité avec d'autres systèmes d'information. Parmi ces services, les services intégrés de travail pour l'importation et l'exportation de langage ontologique (XML, RDF (S), OIL, DAML + OIL, OWL, Prolog, etc.), le service WAB (WebODE Axiom Builder) Pour la documentation, l'évaluation, l'évolution, l'apprentissage et pour la fusion **[Arpirez & al, 2001]**.

7.1.6 DoE

DoE [14] (Differential Ontologies Editor) a été développé à l'Institut National de l'Audiovisuel par R. Troncy et A. Isaac en 2002 **[Isaac, 2005]**. DOE est un éditeur qui offre des interfaces de création, modification et suppression de concepts et de relations, une représentation graphique de l'arbre ontologique, et des fonctionnalités de recherche et de navigation dans la structure créée. L'ontologie est documentée par des définitions encyclopédiques avec des synonymes et les principes différentiels en plusieurs langues **[Baneyx, 2007]**.

7.2. Outils de construction d'ontologies à partir des textes

Dans le cadre de la construction des ontologies à partir des textes, nous avons trouvé plusieurs outils pour la langue anglaise, française et espagnole (ces outils ne supportent pas la langue Arabe). Parmi ces outils nous avons privilégié trois systèmes opérationnels, et

[13] http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/old-technologies/60-webode/
[14] http://www.eurecom.fr/~troncy/DOE/

disponibles sur la toile, et pouvant exporter au format OWL, ce qui permet d'exploiter l'ontologie obtenue **[Mondary, 2008]** :

7.2.1. Text2Onto

Text2Onto **[Cimiano & Volker, 2005]** est un outil très utilisé dans le domaine de la construction automatique des ontologies à partir de textes (voir figure I.7). C'est une plateforme codée en java qui se décompose en plusieurs modules pour extraire des concepts à partir des textes, des relations entre ces concepts (relation d'équivalence, hiérarchiques, etc.) et même des instances de concepts. Chaque module de cet outil peut utiliser différents algorithmes et combiner leurs résultats : on peut ainsi combiner des patrons d'extraction "à la Hearst" et une ressource externe comme *WordNet* pour construire une hiérarchie. Pour pré-traiter les textes, Text2Onto utilise la boite à outils GATE[15] (General Architecture for Text Engineering) qui est une boîte à outils logicielle écrite en Java pour le traitement automatique du langage naturel dans différentes langues (voir section 3.1.3 du chapitre 2). Les résultats sont dotés d'une mesure de confiance entre 0 et 1 obtenue à l'aide de différentes mesures combinables (TF.IDF, RTF, entropie). Text2Onto se présente comme une boîte à outils, et l'ontologue doit lui-même sélectionne les algorithmes à utiliser. Il peut accepter ou rejeter les résultats obtenus mais pas les modifier ni revenir aux parties des documents dont ils sont issus. Le système KASO (Knowledge Acquisition supported Semi-automated Ontology building) **[Wang et al., 2006]** dont la conception est centrée utilisateur peut être couplé à Text2Onto pour affiner l'ontologie produite à l'aide de méthodes d'acquisition de connaissances.

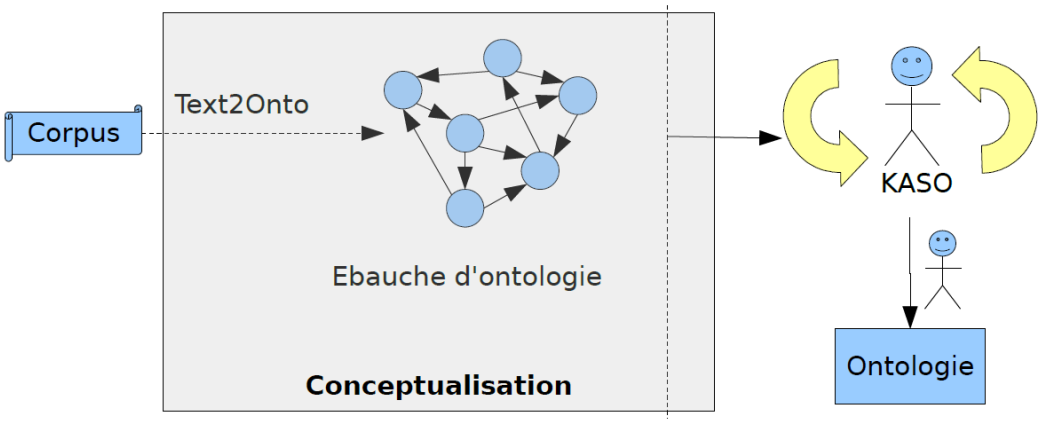

Figure I.7 : Text2Onto + KASO

[15] http://gate.ac.uk

7.2.2. OntoGen

OntoGen **[Fortuna et al., 2006]**, est un outil codé en .net, qui implémente une approche pour la construction semi-automatique d'ontologies à partir de collections de documents (voir figure I.8). OntoGen est un outil interactif qui propose des concepts à l'expert du domaine sous la forme de classes de documents, il propose une dénotation pour ces concepts et leurs associe automatiquement des instances (c-à-d des documents). Il permet une visualisation de l'ontologie en cours de construction. OntoGen utilise des algorithmes de fouille de textes non supervisés (tels que : k-means ou LSI) ou supervisés (tels que SVM (Support Vector Machine) ou active learning) mais il travail toujours selon une approche descendante. A chaque étape le classifieur travaille sur la sous-collection associée au concept en cours de construction.

OntoGen suggère à l'expert plusieurs propositions, et c'est à ce dernier de choisir la proposition correcte parmi celles qui lui sont suggérées. OntoGen suit une approche semi-automatique de la conceptualisation : les outils utilisés pour la classification des documents sont utilisés pour préparer le travail de la conceptualisation, l'expert du domaine est guidé dans une démarche descendante. C'est lui qui construit les concepts et choisit quelles zones de l'ontologie à modifier. Il est à souligner qu'OntoGen se focalise sur la construction d'une ontologie dont les instances des concepts sont les documents.

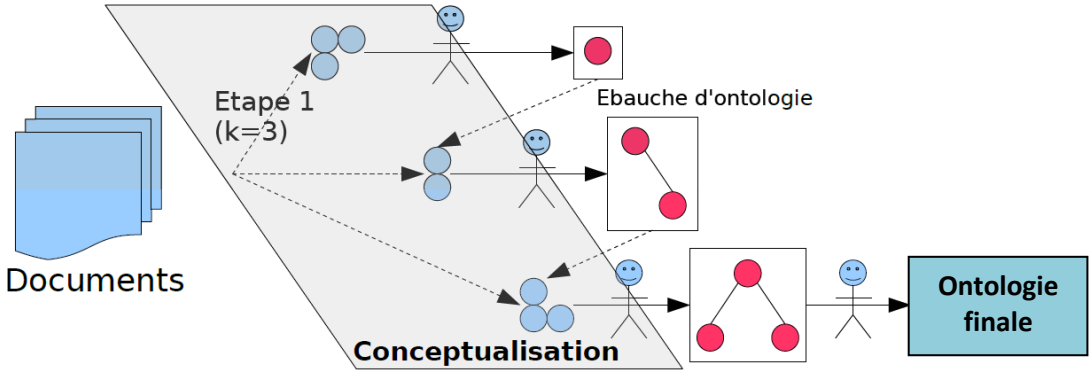

Figure I.8 : OntoGen

7.2.3. Terminae

Terminae **[Aussenac-Gilles et al., 2008]** est un outil logiciel (voir figure I.9) qui peut guider l'ontologue dans la conception de l'ontologie. Pour extraire des éléments de

l'ontologie, Terminae utilise les résultats des outils du TAL(Traitement Automatique des Langues) tels que :

- ✓ Un extracteur de termes
- ✓ Un concordancier
- ✓ Un détecteur de synonymie
- ✓ Un analyseur syntaxique

Les occurrences d'un terme dans le corpus sont affichées sur une fiche terminologique. Il est possible aussi d'afficher plusieurs définitions en langage naturel des concepts terminologiques et les termes synonymes. Le concept dénoté par le terme est alors construit par l'ontologue. La construction des concepts, en utilisant ces fiches, reste entièrement manuelle mais elle est assistée : l'ontologue peut visualiser différentes vues sur le corpus de texte et Terminae offre une traçabilité entre les concepts de l'ontologie en cours de création et le corpus. L'approche implémentée dans Terminae n'impose pas de stratégie de construction contrairement à l'approche implémentée dans OntoGen qui impose une stratégie de construction descendante.

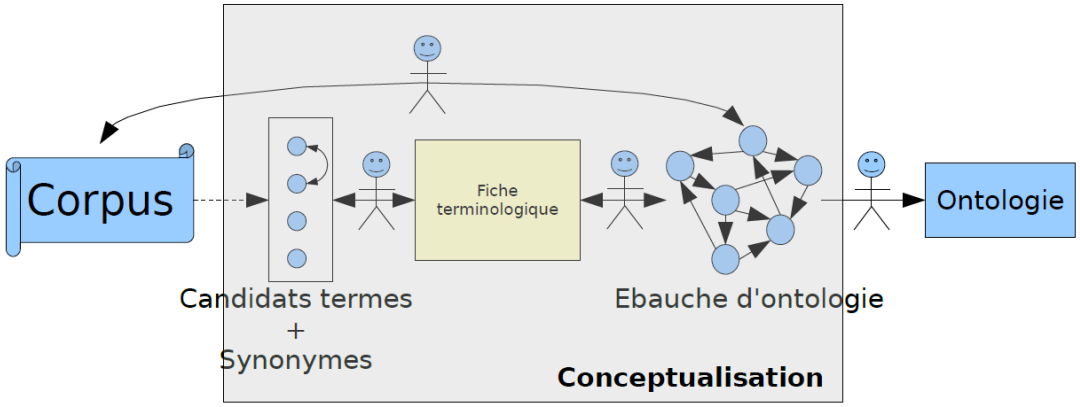

Figure I.9 :Terminae

8. Conclusion

Dans ce chapitre nous avons essayé de faire un tour d'horizon sur les ontologies et sur les méthodes et techniques permettant de les créer. D'abord, nous avons commencé par définir le terme « ontologie » et ses constituants, ensuite nous avons fait un survol sur la classification des ontologies. Ensuite, nous avons décrit les différentes méthodologies de construction des ontologies et les langages de représentation des ontologies les plus connues et les plus utilisés. Et finalement, nous avons examiné les outils d'édition et de construction des ontologies à partir des textes.

Dans le chapitre suivant, nous allons présenter un état de l'art sur les travaux de la construction des ontologies à partir des textes.

Chapitre 2
Etat de l'art
sur la construction des ontologies à partir des textes

1-Introduction

Parmi les sous-domaines de l'ingénierie des ontologies on distingue la construction d'ontologies à partir de documents textuels. Ces ontologies peuvent êtres utilisées dans plusieurs domaines tels que la recherche d'informations, l'annotation sémantique de ressources, l'indexation automatique des documents, les résumés automatiques des textes,…etc**.**

Dans ce chapitre nous allons présenter un état de l'art sur la construction des ontologies à partir des textes dans toutes les langues, en commençant par un état de l'art sur le domaine de l'extraction des termes et des relations, et nous allons finir par quelques approches de la construction des ontologies à partir des textes *arabes*.

2-Les étapes de construction d'une ontologie à partir de textes

La construction d'une ontologie à partir d'un corpus de textes, nécessite une suite logicielle outillant une méthodologie globale de construction. La figure II.1 donne une idée sur les différentes étapes nécessaires pour la construction automatique d'ontologie du domaine. Nous explicitons dans cette section brièvement chaque étape.

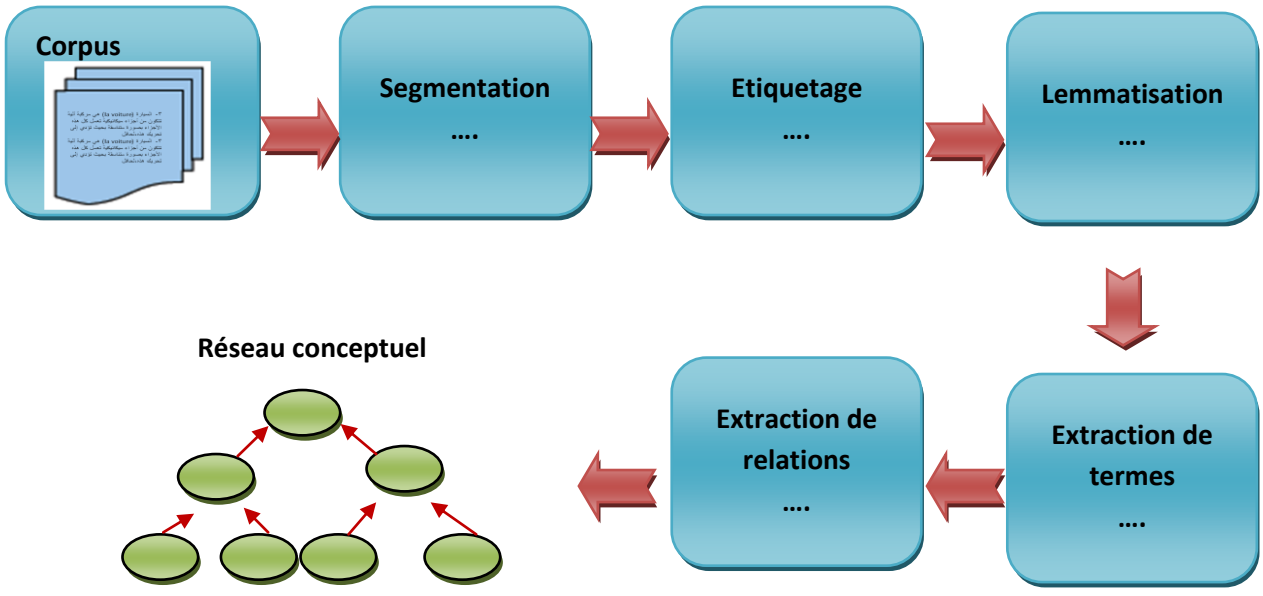

Figure II.1: Les étapes de la construction d'ontologies à partir de textes

2.1- Corpus

Dans un projet de construction d'une ontologie à partir de textes, la tâche de construction du corpus est à la fois primordiale et délicate. Puisque, d'une part, le corpus est la source d'information essentielle pour tout le processus de construction de l'ontologie et que, d'autre part, il restera, une fois le processus achevé, l'élément de documentation de la ressource construite, il doit être composé avec un maximum de précautions méthodologiques. Dans ce domaine, il n'est hélas pas encore possible de définir a priori des instructions méthodologiques très précises pour encadrer la tâche de sélection des sources textuelles qui viendront constituer le corpus. Au-delà des problèmes techniques ou juridiques de disponibilité des textes, cette collecte doit se faire avec l'aide des spécialistes et en fonction de l'application cible visée. Il convient en effet de s'assurer auprès des spécialistes que les textes choisis ont un statut suffisamment consensuel pour éviter toute remise en cause ultérieure de la part d'utilisateurs ou de leur part. Par ailleurs, il convient de prévoir d'emblée une boucle de rétroaction au cours de laquelle une première version du corpus sera modifiée et enrichie en fonction d'une première phase d'analyse des résultats fournis par les outils de TAL sur cette version initiale. Le critère de la taille est évidemment important, même s'il est impossible de donner un chiffre idéal. Le choix est ici encore un compromis. Le corpus doit être suffisamment « gros » pour justifier que des outils de traitement de la langue soient nécessaires pour le dépouiller de façon efficace. Mais il doit être suffisamment petit et/ou redondant pour pouvoir être appréhendé de façon globale par l'analyste, même à l'aide d'outils de TAL. Une fourchette entre 50 000 et 200 000 mots semble raisonnable. **[Bourigault et al., 2003]**

Les projets prenant le Web comme source de textes font rapidement exploser ces chiffres, posant par la même des problèmes spécifiques, comme celui de la définition d'un « échantillon » pertinent pour l'étude. Enfin, dans la majorité des cas, le corpus sera hétérogène dans le sens où il aura été constitué en rassemblant des textes d'origine variée. Il est alors absolument nécessaire de procéder à un balisage du corpus qui permettra aux outils d'analyse, et ainsi qu'à l'analyste, de repérer les différents sous-corpus pour procéder éventuellement à des analyses contrastives. **[Bourigault et al., 2003]**

Apres la construction du corpus et avant de passer à l'étape suivante, ce corpus doit être prétraité. Le prétraitement de corpus est une étape préliminaire pour identifier les données lexicales à partir des textes [**Harrathi, 2009**]. Les prétraitements des données textuelles

consistent à normaliser les diverses manières d'écrire un même mot, à corriger les fautes d'orthographe évidentes ou les incohérences typographiques et à expliciter certaines informations lexicales exprimées implicitement dans les textes.

2.2- Segmentation

La segmentation est une étape quasiment obligatoire avant l'extraction d'information. Elle permet de découper le texte en unités linguistiques suffisamment élémentaires pour qu'elles soient traitées **[Dubois et al., 1994]**. C'est une étape qui permet de découper un texte d'abord en section puis en phrase et enfin en mot.

Exemple : .../ به / المتصلة / الحواسيب / شبكات / يهمج / عالمي / نظام / هي / الإنترنت

2.3- Etiquetage

L'étiquetage permet l'identification de la catégorie grammaticale (nom, verbe, adjectif, particule…) de chaque mot. Un texte étiqueté ressemblera grossièrement à ceci :

.../ به / المتصلة / الحواسيب / شبكات / يهمج / عالمي / نظام / هي / الإنترنت

/ préposition / nom / nom / nom / verbe / nom / nom / préposition / nom

2.4- Lemmatisation

Un lemmatiseur est un programme de traitement automatique du langage qui permet de passer d'un mot portant des marques de flexion (pluriel, forme conjuguée d'un verbe...) à sa forme de référence (lemme ou forme canonique).

Exemple : الحواسيب → " حسب "

2.5- Extraction de termes

L'extracteur terminologique analyse le contenu des documents d'un domaine et recherche la terminologie disponible dans un corpus choisi.

2.6- Extraction de relations sémantiques

Il s'agit de repérer des relations sémantiques entre des termes préalablement extraits (dans l'étape d'extraction de termes), telle la synonymie, l'hyponymie, la méronymie ou d'autres types de relations sémantiques.

Exemple : Une phrase telle que: الحاسوب هو آلة مثل باقي الآلات…

Dans cet exemple l'extracteur de relation génère une relation d'hyperonymie (est-un) entre les deux termes « الحاسوب » et « آلة ».

Remarque : Le concordancier est utilisé aussi dans le processus de construction d'une ontologie à partir des textes (malgré que nous n'avons pas le mentionner dans la figure II.1), c'est un programme qui, pour un mot donné, recherche dans un texte toutes ses concordances, c'est-à-dire les phrases ou les groupes de mots dans lesquels il apparaît.

3-Les approches et les outils d'extraction de termes

L'extraction de terminologie consiste à identifier des termes potentiels dans un document de texte spécifique ou un ensemble de document de textes (corpus) ainsi que les informations pertinentes liées à l'emploi de ces termes ou aux concepts auxquels ils renvoient (définition, contexte, etc.).

L'extraction de termes représente une étape importante dans la construction d'une ressource ontologique à partir de corpus. Les termes sont des mots ou des expressions ayant un sens précis dans un contexte donné et représentent les supports linguistiques des concepts.

Dans la littérature, les différents travaux d'extraction des termes à partir des corpus textuels utilisent deux approches : l'analyse *statistique* ou numérique et l'analyse *linguistique* ou structurelle **[Claveau, 2003]**. L'analyse statistique se base sur l'étude des contextes d'utilisation et les distributions des termes dans les documents. L'analyse linguistique exploite des connaissances linguistiques, telles que les structures morphologiques ou syntaxiques des termes. D'autres travaux couplent ces deux approches et constituent une approche dite «approche *hybride* ou mixte».

3.1- Les approches linguistiques

Ces méthodes sont qualifiées de linguistique puisqu'elles font appel à des techniques d'analyse se basant sur les connaissances de la *langue* et de sa structure. La majorité de ces méthodes exploitent des connaissances syntaxiques, lexicales ou morphologiques.

Les méthodes linguistiques considèrent que la construction des unités terminologiques obéit à des règles de syntaxe plus ou moins stables, ce sont principalement des syntagmes formés de noms et d'adjectifs. Se basant sur ces connaissances, ces systèmes procèdent à

l'extraction de candidats termes à l'aide de schémas syntaxiques [**Malaisé, 2005**]. On peut aussi utiliser des grammaires et un lexique acquis en cours d'analyse ou par le biais d'une collaboration avec des spécialistes pour générer l'ensemble des termes potentiels d'un domaine [**Drouin, 2002**].

Ces outils nécessitent donc un prétraitement du corpus par un analyseur syntaxique. La qualité des résultats dépend étroitement de la qualité de ces analyseurs. Ils ont l'inconvénient de dépendre directement de la langue des textes traités et nécessitent des ressources linguistiques (dictionnaires, liste de stop-word etc.). De plus ils ne sont efficaces que sur de *petits corpus*.

Les premiers travaux dans ce sens, sont ceux de David et Plante. Le premier outil spécifiquement dédié à la construction de bases terminologiques est *Termino* [**David & Plante, 1990**]. Il a été élaboré dans le cadre d'une collaboration entre l'Office de la langue française du Québec et une équipe du Centre d'ATO de l'Université du Québec à Montréal. La première version de ce logiciel, qui se nommait donc *Termino*, a depuis été remplacée par un nouveau système nommé *Nomino* [**Perron 1996**].

La majorité des outils, pour ne pas dire la totalité, pour le traitement de l'Arabe n'ont pas été construits à l'origine pour accomplir cette tache **[Zaidi & Laskri, 2009]**. Une grande partie de ces outils, a été développée pour le traitement de l'Anglais, du Français ou autre langue. Après des années de raffinement et d'amélioration l'idée ou le besoin s'est ressenti de l'adapter au traitement de la langue Arabe.

La littérature regorge d'ouvrages présentant ou recensant les outils dédiés à l'extraction de termes, nous citons à titre d'exemples : **[Assadi H.& Bourigault D., 1996], [Beguin & al, 1997], [Condamines & Rebeyrolle, 1997], [Daille, 1994], [Daoust, 1992], [Garcia, 1998], [Hearst, 1992], [Le Priol & al, 1998], [Morin, 1999a], [Smadja, 1993], [Biebow & Szulman, 2000], [Gandon, 2002]** et **[Voutilainen, 1993].** Les outils que nous avons testé supportent ou peuvent être modifiés pour supporter la langue arabe, ils seront décrits dans la section suivante:

3.1.1- UNITEX

Unitex[16] est un ensemble de logiciels permettant de traiter des textes en langue naturelle, en utilisant des ressources linguistiques. Ces ressources se présentent sous la forme de dictionnaires électroniques, de grammaires et de tables de lexique-grammaire. UNITEX manipule des textes Unicode.

Le prétraitement du texte consiste à lui appliquer les opérations suivantes :

- ✓ normalisation des séparateurs.
- ✓ découpage en unités lexicales.
- ✓ normalisation de formes non ambiguës.
- ✓ découpage en phrases et application des dictionnaires [**Paumier, 2009**].

La figure II.2 représente l'interface graphique de « UNITEX 3.0 » que nous avons obtenu après l'application sur un texte arabe.

[16] http://www-igm.univ-mlv.fr/~unitex

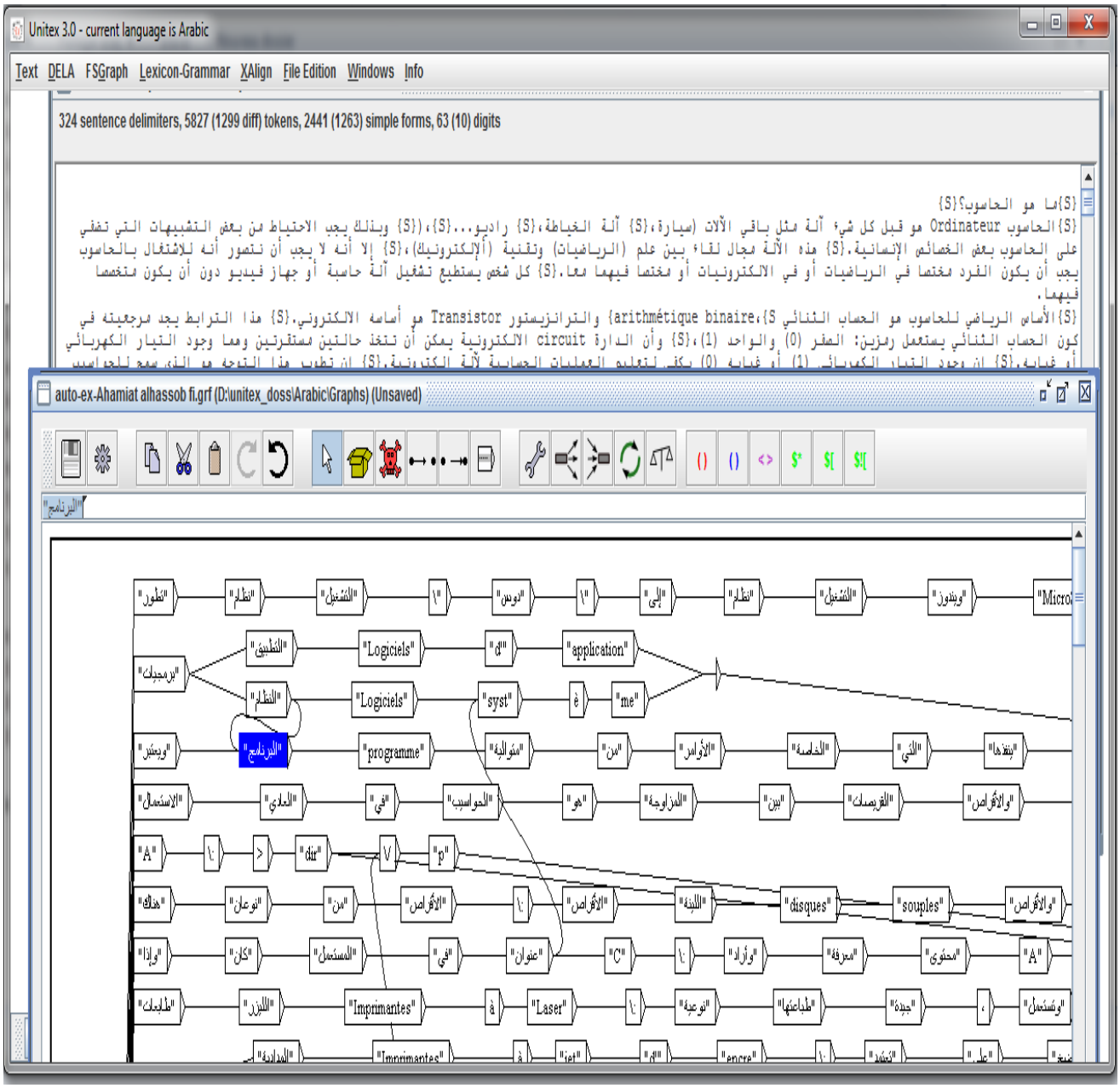

Figure II.2: Extraction d'information arabe avec UNITEX

En plus de l'extraction d'information, UNITEX dispose d'un concordancier permettant la localisation d'une entité dans le texte d'origine, en affichant ses contextes droit et gauche. L'outil peut être facilement adapté à l'extraction de termes arabes.

3.1.2- NOOJ

NOOJ[17] est un système de traitement de corpus, reprenant et améliorant les fonctionnalités d'INTEX [**Koeva & al, 2007**], conçu pour l'enseignement des langues et de la linguistique. Le système intègre des outils TAL (analyse morphologique de mots simples et complexes,

[17] http://www.nooj4nlp.net

élaboration de transducteurs d'états finis, grammaires locales) qui permettent un prétraitement du corpus par l'enseignant, et des procédures de recherche et d'entraînement pour l'étudiant. La nouvelle mouture du logiciel INTEX (appelée "NOOJ") a été réécrite à partir de zéro, en particulier pour répondre aux besoins des utilisations pédagogiques.

Le système NOOJ présente des fonctionnalités de TAL qui paraissent prometteuses pour l'enseignement des langues et de la linguistique. Parmi ces fonctionnalités nous citons :

- ✓ La description de la morphologie et de la syntaxe des langues.
- ✓ Analyse et traitement de corpus.
- ✓ Applications d'extraction de l'information à partir des corpus.
- ✓ Des mini-applications pédagogiques utilisées pour l'enseignement des langues.
- ✓ Construction, édition et gestion sophistiquées de concordances

Le principal avantage de NOOJ est sa simplicité d'utilisation : il permet à la fois à l'enseignant non spécialiste de TAL de constituer des ressources linguistiques (à l'aide d'interfaces simples) et de les paramétrer afin de constituer des projets pédagogiques destinés aux apprenants [**Silberztein & Tutin, 2004**]. Après construction d'un dictionnaire et d'une grammaire nous pouvons extraire de l'information à partir de textes arabes. La figure II.3 ci-dessous, présente l'interface d'utilisation du logiciel NOOJ pour un texte Arabe.

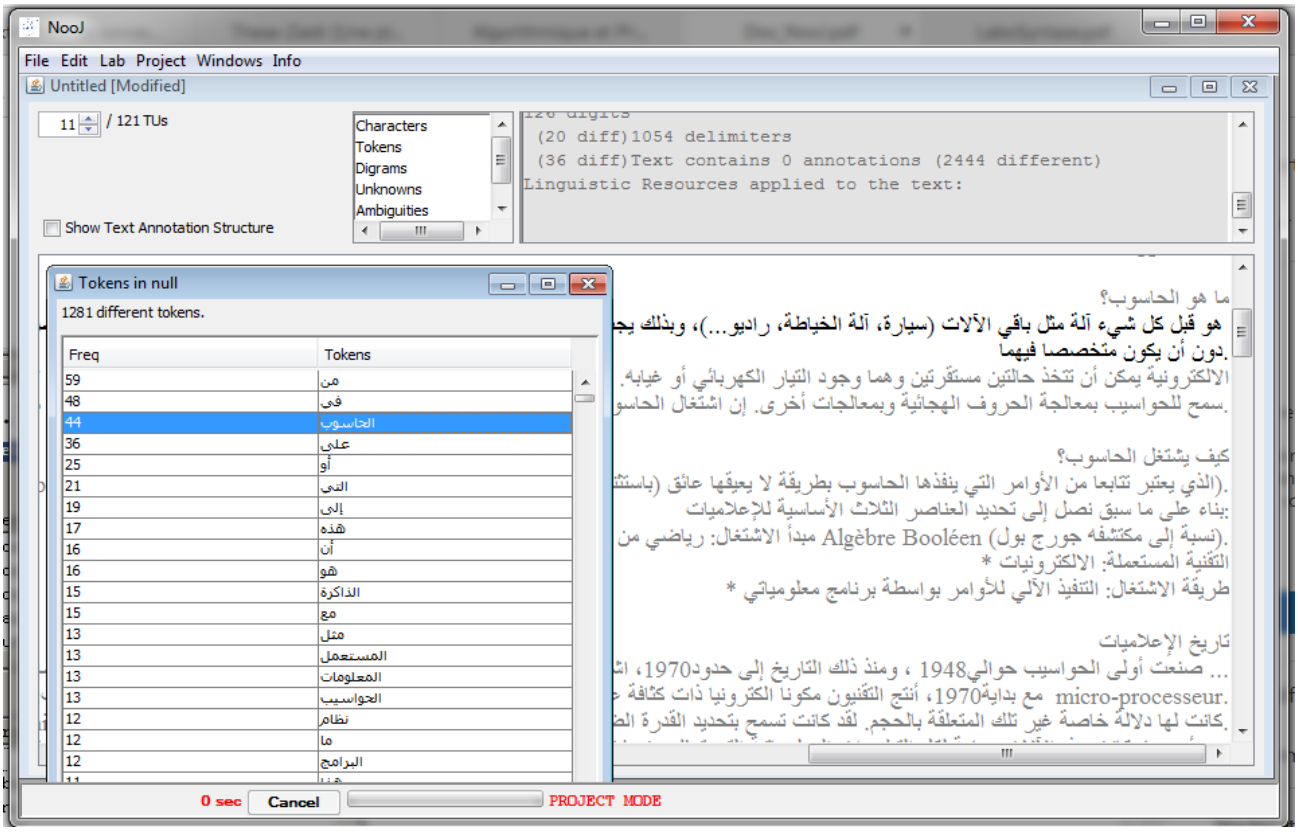

Figure II.3: Interface d'utilisation du logiciel NOOJ pour un texte Arabe.

Il est aussi possible d'utiliser un concordancier sous NOOJ, ceci nous permettra avec l'aide d'une liste de termes simples, de chercher des termes composés, les sauvegarder dans une base de données par exemple et soumettre des requêtes sur des suites de mots respectant un certain marqueur.

3.1.3- GATE

GATE[18] (General Architecture for Text Engineering) est une boîte à outils logicielle écrite en Java à l'université de Sheffield, à partir de 1995. Il est très largement utilisé à travers le monde, par de nombreuses communautés (scientifiques, entreprises, enseignants, étudiants) pour le traitement du langage naturel dans différentes langues. La communauté de développeurs et de chercheurs autour de GATE est impliquée dans plusieurs projets de recherche européens comme TAO (Transitioning Applications to Ontologies) et SEKT (Semantically Enabled Knowledge Technology).

GATE offre une architecture, une interface de programmation d'applications(API) et un environnement de programmation graphique. Il comporte :

[18] http://gate.ac.uk

- ✓ Un système d'extraction d'information, ANNIE (A Nearly-New Information Extraction System, pour système quasi nouveau pour l'extraction d'information),
- ✓ Un analyseur lexical, un gazetteer (lexique géographique),
- ✓ Un segmenteur de phrases (avec désambiguïsation),
- ✓ Un étiqueteur,
- ✓ Un module d'extraction d'entités nommées,
- ✓ un module de détection de coréférences.
- ✓ Des plugins pour faire l'apprentissage automatique (Weka[19], RASP[20], MAXENT[21], SVM[22] light),
- ✓ Des plugins pour la construction d'ontologies (WordNet[23]),
- ✓ Des plugins pour l'interrogation de moteurs de recherche comme Google et Yahoo

GATE accepte en entrée divers formats de texte comme le texte brut, HTML, XML, Microsoft Word (Doc), PDF, ainsi que divers formats de bases de données comme Java Serial, PostgreSQL, Lucene, Oracle, grâce à RDBMS[24] et JDBC[25].

GATE utilise également le langage JAPE (Java Annotation Patterns Engine) pour bâtir des règles d'annotation de documents. On trouve aussi un debugger et des outils de comparaison de corpus et d'annotations [**Cunningham & al, 2002**].

La *figure II.4* montre l'interface d'utilisation de GATE pour l'extraction des entités nommées en arabe.

[19] *Weka* (acronyme pour Waikato environment for knowledge analysis, en français : « environnement Waikato pour l'analyse de connaissances » *www.cs.waikato.ac.nz/ml/weka/*
[20] RASP (Robust Accurate Statistical Parsing) Est un système d'analyse robuste pour l'anglais.
[21] MAXENT (MAXimum ENTropy) est une méthode de classification automatique.
[22] SVM (Support Vector Machine) est un algorithme d'apprentissage automatique.
[23] *WordNet* est une base de données lexicale développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton.
[24] RDBMS (Relational DataBase Management System) est un système de gestion de base de données
[25] JDBC (Java DataBase Connectivity) est une interface de programmation pour les programmes utilisant la plateforme Java.

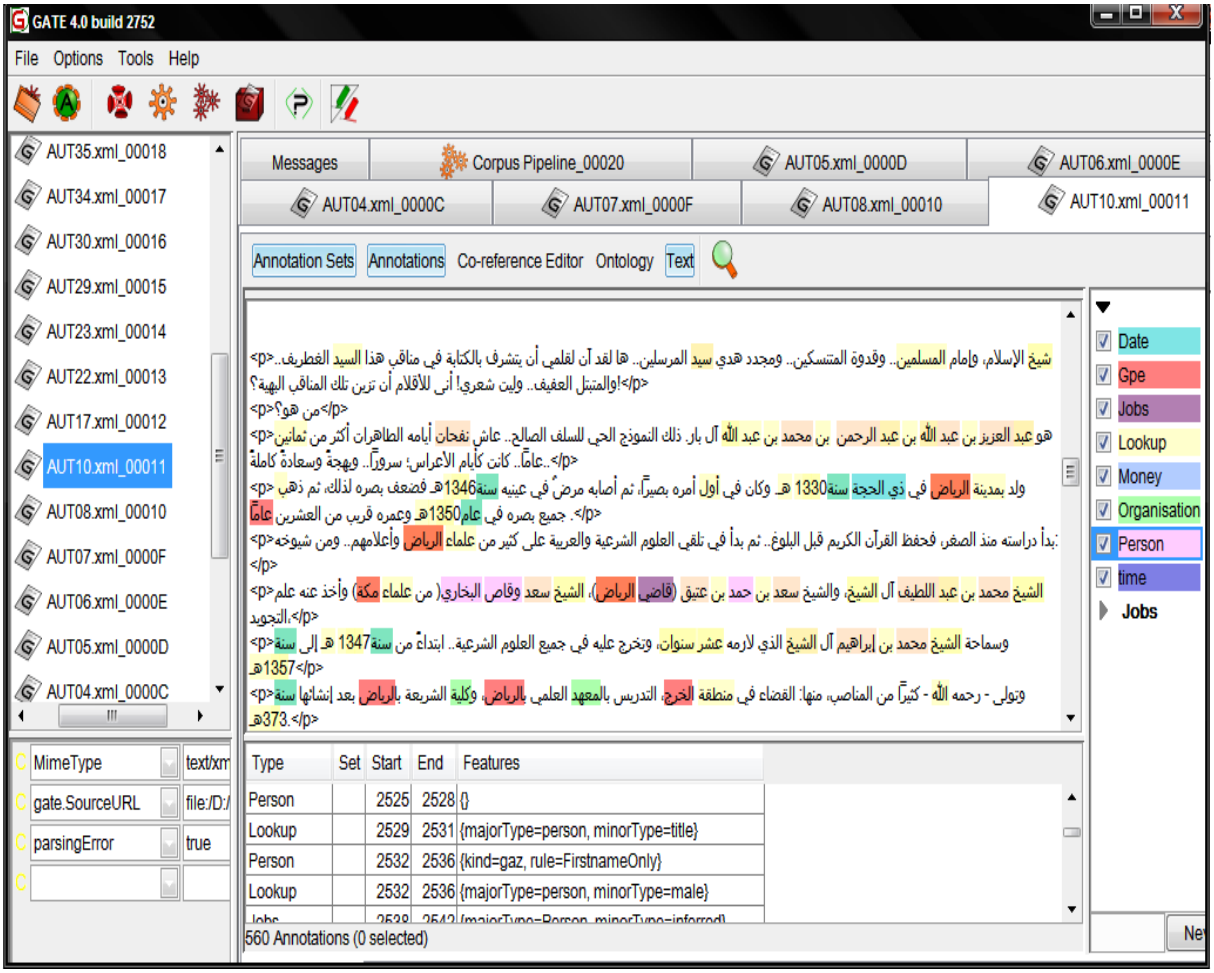

Figure 10: Extraction d'entités nommées arabes avec Gate [Amari, 2009]

3.1.4 Discussion : les approches linguistiques

Les résultats obtenus par les méthodes linguistiques sont jugés pertinents. Cependant l'utilisation de ces approches nécessite une maîtrise complète de la langue du corpus étudié. L'extraction des termes simples et des termes composés nécessite une connaissance parfaite des règles syntaxiques de dérivation dans la langue du corpus. Les méthodes linguistiques sont basées sur des propriétés linguistiques de la langue naturelle. Ces propriétés sont intrinsèques à la langue du corpus d'étude (dans notre cas la langue Arabe). Elles ne sont pas, de ce fait, généralisables à d'autres langues.

Il est à souligner que les propriétés et les règles utilisées dans ces méthodes sont issues d'un traitement manuel du corpus d'étude. Ces éléments sont difficiles à dégager à partir des corpus volumineux. En effet, pour dégager une règle il est indispensable de feuilleter la quasi-totalité du corpus d'étude. Cette tâche n'est pas aisée dans le cas ou le corpus est de grande taille.

En conclusion les approches linguistiques trouvent leurs performances dans des corpus bien spécifiques sur lesquels une étude linguistique détaillée a été réalisé. Ces approches ne peuvent pas être généralisées sur des corpus de *langue* différente, de *taille* différente et de *spécialité* différente **[Harrathi, 2009]**.

3.2. Les approches statistiques

Les méthodes statistiques ou numériques sont basées sur des techniques quantitatives. Ces méthodes sont souvent utilisées pour les traitements des corpus volumineux. Avec l'évolution incessante des nouvelles technologies, les documents numériques sont devenus facilement disponibles facilitant ainsi la constitution de ces corpus volumineux. De ce fait ces méthodes continuent à connaitre un grand succès. Elles présentent l'avantage de ne pas nécessiter de connaissances linguistiques a priori et s'appliquent sur des corpus pour lesquels aucune ressource externe (dictionnaire, stop liste, ontologie…) n'a été élaborée. **[Harrathi, 2009]**

L'objectif des méthodes statistiques est de représenter un texte, qui est habituellement vu comme un enchaînement de formes simples, comme une succession de formes simples et de segments répétés (SR). Ces derniers sont définis comme des « *suites de formes graphiques non séparées par un caractère délimiteur de séquence, qui apparaissent plus d'une fois dans ce corpus de textes* » [**Drouin, 2002**].

Il est à souligner que ces méthodes exploitent des mesures de similarité. Dans la littérature il existe plusieurs méthodes statistiques appliquées à l'extraction de termes, la plupart sont fondées sur l'information mutuelle ou le coefficient de Dice[26] [**Velardi, 2001**]. Le principe est que l'association récurrente de deux mots ne peut être due qu'au fruit du hasard. Par conséquent, elle est forcément significative **[L'Homme, 2001]**.

3.2.1. Les mesures de similarité pour l'extraction des termes

3.2.1.1. Le tf-idf

Tf-idf (*Term Frequency-Inverse Document Frequency*) est une méthode de pondération souvent utilisée en recherche d'information. C'est une mesure statistique qui permet d'évaluer l'importance d'un terme contenu dans un document, relativement à un corpus. Le poids augmente avec le nombre d'occurrences du mot dans le document et varie aussi en fonction de

[26] **Coefficient de dice** : est un indicateur statistique qui mesure la similarité de deux échantillons.

la fréquence du mot dans le corpus. Il existe plusieurs variantes de la formule originale. Elles sont aussi utilisées dans d'autres domaines comme l'extraction des termes.

a) La fréquentielle (tf)

La représentation dite fréquentielle, notée *tf*, est une extension de la représentation binaire qui ne considère que la présence ou l'absence du mot dans le document. Tf est plutôt basée sur le nombre d'occurrences d'un terme *i* dans un document *j* .

b) Le idf

La fréquence inverse de document est une mesure de l'importance du terme dans l'ensemble du corpus. Dans le schéma *tf-idf*, elle vise à donner un poids plus important aux termes les moins fréquents, considérés comme plus discriminants. Cette pondération issue du domaine de la recherche d'informations tire son inspiration de la loi de *Zipf*[27] [**Zipf, 1949**], introduisant le fait que les termes les plus informatifs d'un corpus ne sont pas ceux apparaissant le plus dans ce corpus. Ces mots sont la plupart du temps des mots outils. Par ailleurs, les mots les moins fréquents du corpus ne sont également pas les plus porteurs d'informations. Ces derniers peuvent en effet être des fautes d'orthographe ou encore des termes trop spécifiques à quelques documents du corpus étudié. Le *tf-idf* peut se décrire formellement comme suit : pour un terme *i* dans un document *j* parmi les *N* documents du corpus [**Bechet, 2009]** :

$$w_{ij} = tf_{ij} \times idf_i \qquad \textbf{Avec} \quad idf_i = log\,\frac{N}{n_i}$$

Où :
- n_i : est le nombre de documents dans lesquels apparaît le terme *i*.
- N : le nombre total de documents.

3.2.1.2. La fréquence d'un couple

La fréquence d'une séquence s est le nombre d'apparition de s. Cette séquence peut être un lexème[28], un lemme[29], un mot, un terme, etc. Cette mesure est utilisée dans tous les modèles

[27] **La loi de Zipf** : est une observation empirique concernant la fréquence des mots dans un texte.
[28] *Le lexème* : étant une entrée lexicale, issue de l'analyse lexicale qui décompose le texte en unités lexicales, selon des grammaires. Ces unités sont généralement des chaînes alphabétiques.
[29] *Un lemme* : permet de définir une forme canonique pour les entrées lexicales (les lexèmes). Cette forme est représentée par l'infinitif pour les verbes et par le masculin singulier pour les substantifs. Grâce à cette étape de

statistiques, ce qui explique le soin apporté pendant les calculs de cette mesure. Ces modèles utilisent souvent quatre fréquences **[Daille, 1994]**:

1) La fréquence d'un couple de séquences (S_i , S_j) dans un document et/ou dans un corpus.
2) La fréquence des couples de séquences (S_i , S_j), où la séquence S_i apparait comme premier élément d'un couple.
3) La fréquence des couples de séquences (S_i , S_j), où la séquence S_j donné apparait comme deuxième élément d'un couple
4) La fréquence totale des couples (pour chaque couple (S_i ,S_j)) dans un document et/ou dans un corpus.

3.2.1.3. Critères d'associations

B. Daille **[Daille, 1994]** considère que les lemmes qui forment un couple sont considérés comme des variables qualitatives pour lesquelles elle teste le degré d'association ou de liaison. Ainsi, les données définies à partir des fréquences citées précédemment, sont représentées sous forme d'un tableau croisé, dit tableau de contingence. Dans ce tableau on associe à chaque couple de lemmes, les valeurs ***a, b, c*** et ***d*** qui décrivent les fréquences du couple.

	Lj	**Lj' avec j'≠j**
Li	a	b
Li' avec i'≠i	c	d

Tableau II.1 : Tableau de contingence du couple de lemmes

- **a** est la fréquence du couple (*Li, Lj*) Li est le premier élément et *Lj* le second.
- **b** est la fréquence des couples où *Li* est le premier élément d'un couple et *Lj* n'est pas le deuxième.
- **c** est la fréquence des couples où *Lj* est le deuxième élément du couple et *Li* n'est pas le premier.

---

lemmatisation, il est possible d'établir la correspondance entre les formes conjuguées des verbes et entre des dérivés morphologiquement distincts.

- **d** est la fréquence de couples où ni *Li* ni *Lj* n'apparaissent.

- La somme a+b+c+d, notée N est le nombre total d'occurrences de tous les couples trouvés.

La majorité des mesures statistiques exploitent les données du tableau de contingence afin de déterminer le degré de liaison de deux lemmes donnés. En résumé, il s'agit de tester l'indépendance des lexèmes pris deux à deux.

Les mesures statistiques qui seront présentées par la suite, sont les plus utilisées dans le domaine de l'extraction de terminologie. Cependant, dans la littérature on trouve de nombreuses autres mesures qui ont déjà été évaluées dans des travaux ultérieurs **[Daille, 1994]**. Dans les mesures ci-dessous, a, b, c et d représentent les fréquences déjà données dans le Tableau II.1.

- **Coefficient de Proximité simple** (**SMC** : **S**imple **M**atching **C**oefficient)

SMC est un coefficient qui se varie de 0 à 1, il est calculé par l'équation suivante :

$$\textbf{SMC}=\frac{a+b}{d}$$

- **Coefficient de φ2 (PHI)**

Cette mesure est utilisée dans les travaux de W.Gale **[Gale et al, 1991]** pour l'alignement de mots dans les phrases.

$$\textbf{PHI}=\frac{(ad-bc)^2}{(a+b)(a+c)(b+c)(b+d)}$$

- **Score d'association ou l'information mutuelle (IM)**

Il s'agit d'un score d'association d'un couple de lexèmes (*Li*, *Lj*), noté IM. Cette mesure a été décrite par P. Brown **[Brown et al, 1988] [Brown et al, 1990]** et par K. Church **[Church et al, 1990]** dans le cadre d'extraction des termes à partir des corpus bilingues et monolingues. L'information mutuelle permet de comparer la probabilité d'observer ces deux lexèmes *Li* et *Lj* ensemble avec la probabilité de les observer séparément. IM se définit comme suit :

$$\textbf{IM}=log_2\left(\frac{a}{(a+b)(a+c)}\right)$$

Si l'information mutuelle IM est fortement positive, cela signifie que *Li* et *Lj* apparaissent très souvent ensemble. Si IM est proche de 0, alors *Li* et *Lj* n'ont aucun rapport et enfin, si IM est fortement négative, alors *Li* et *Lj* ont des distributions complémentaires.

- **Coefficient de vraisemblance : Loglike**

Cette mesure introduite par T. Dunning **[DUNNING, 1993]**, représente le rapport de vraisemblance appliqué à une loi binomiale. Ce score s'exprime de la manière suivante :

$$\text{LogLike} = a \log(a) + b \log(b) + c \log(c) + d \log(d) - (a+b) \log(a+b) - (a+c) \log(a+c) - (b+d) \log(b+d) - (c+d) \log(c+d) + N \log(N)$$

3.2.2. Les travaux de L. Lebart et A. Salem

La méthode présentée dans les travaux menés par L. Lebart et A. Salem **[Lebart et al, 1988]** consiste à repérer des séquences de mots qui se répètent plus d'une fois côte à côte dans un texte. Les auteurs étudient les segments répétés dans un corpus afin d'extraire un ensemble de termes dits « termes complexes ou termes composés ». Le texte est alors considéré comme étant un enchainement de mots et de segments répétés.

Un segment répété est une séquence de deux ou plusieurs mots voisins et qui apparaissent plus d'une fois dans le texte. En pratique il s'agit de compter le nombre d'occurrences d'un couple (l1, l2), afin de vérifier si ce nombre est supérieur à une valeur de seuil fixée expérimentalement. Si c'est le cas, la séquence formée par (l1, l2) est considérée comme étant un terme composé et il sera repris dans le processus. Ce processus s'arrête si aucune nouvelle séquence n'a été repérée. Le nombre d'occurrences d'un couple (l1, l2) correspondant à la valeur dans le tableau de contingence. Afin de regrouper des séquences qui diffèrent d'un point de vue graphique (par exemple : phénomène fréquent, phénomènes fréquents), les auteurs utilisent des corpus textuels lemmatisés.

3.2.3. Les travaux de Church

Dans **[Church et al, 1990]**, les auteurs proposent une méthode d'extraction des termes composés basée sur une mesure statistique appelée l'information mutuelle. Les auteurs considèrent que les mots qui apparaissent souvent ensemble d'une manière statistiquement significative ont une grande chance de former des termes complexes. Ainsi, ils évaluent la probabilité d'apparition des mots ensemble en la comparant à la probabilité d'apparition de ces mots séparément.

3.2.4. Les travaux de R. Oueslati

Dans ces travaux de thèse, R. Oueslati **[Oueslati, 1999]** reprend le principe des segments répétés présentés précédemment. L'objectif de l'auteur est la réalisation d'un système d'aide à la construction de la terminologie d'un domaine spécialisé, tel que la médecine. La méthode proposée fait appel aux travaux sur les segments répétés durant l'étape d'extraction des termes. Les termes extraits sont validés par un linguiste ou terminologue. Ensuite, il cherche à construire des classes de termes sémantiquement proches on utilisant la distribution contextuelle.

3.2.5. Les travaux de Kurshid

(Kurshid, 1996) propose un indice qu'il nomme coefficient d'étrangeté (co-efficient of weirdness) et qui consiste à évaluer le rapport entre la fréquence relative d'une forme dans un corpus non spécialisé et la fréquence relative de la même forme, au sein d'un corpus technique. Les formes qui apparaissent dans le corpus technique, mais qui ne sont pas représentées dans le corpus non technique, se voient attribuer une valeur infinie. Ils sont donc considérés comme «étranges». Les formes dont le coefficient d'étrangeté est particulièrement élevé sont spécifiques au corpus technique.

3.2.6. Les travaux de Heitz (le système EXIT)

L'approche de Heitz comprend les étapes suivantes :

- Une opération de lemmatisation
- Une étape de recherche des termes déjà formés avec un tiret '-' par le rédacteur du texte d'origine, qui permettra d'affecter à leurs unités des poids plus importants que les unités qui apparaissent seules dans le texte d'origine.
- Après l'extraction des unités, différentes mesures de pertinence peuvent être testées par l'expert afin de déterminer la mieux adaptée. Les résultats sont une liste de termes candidats classés par pertinence décroissante.

Cette méthode est supportée par un logiciel qui s'appel EXIT[30] (EXtraction Itérative de Terminologies). C'est un logiciel (voir figure II.5) destiné à être utilisé, après un étiquetage grammatical des mots, malgré qu'il soit fondé sur une méthode statistique. Tout texte étiqueté

[30] http://thomasheitz.free.fr/text.mining.researcher/#softwares

est analysable par ce programme à condition de modifier les expressions régulières dans le panneau « *Expressions* » de son interface **[Heitz, 2008]**.

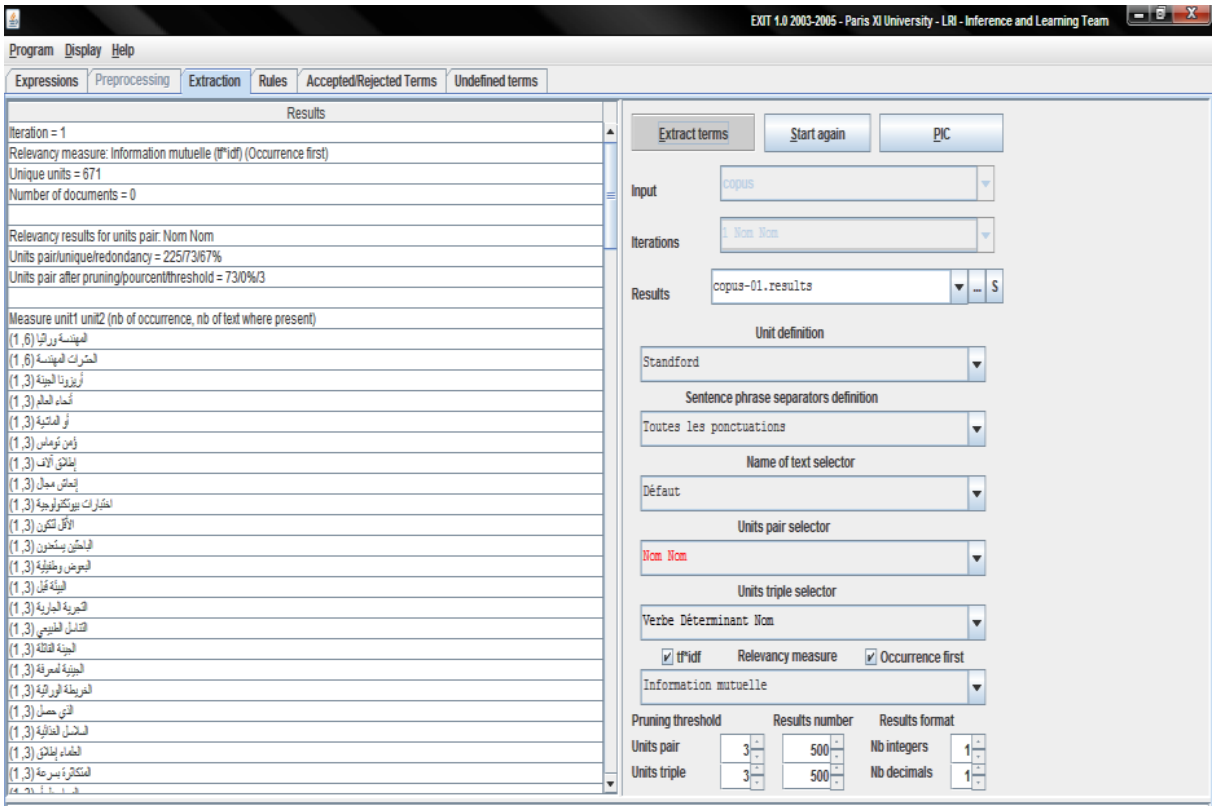

Figure II.5: Extraction de collocations avec Exit [Lalaouna, 2009]

Pour pouvoir utiliser l'outil EXIT sur des textes arabes, il faut disposer d'un étiqueteur fonctionnant sur du texte arabe translittéré. Nous pouvons utiliser l'ancienne version de l'analyseur morphologique de Abdelali[31] **[Abdelali, 2007].**

3.2.7. Les travaux de Enguerhard (le système ANA)

Enguerhard **[Enguerhard, 1993]** propose une méthode d'apprentissage automatique permettant de reconnaitre des termes sous plusieurs variations morphologiques. Cette méthode est supportée par le système ANA[32] (Apprentissage Naturel Automatique) qui a été développée à l'aide du langage *Lisp* et d'un système de gestion de base de données objets, sur une station *Sun*.

ANA est un système de détermination automatique de terminologie pour la construction du thésaurus d'un domaine. Les taux de rappel et de précision ne sont pas égaux à 1, Cependant,

[31] http://crl.nmsu.edu/Resources/lang_res/arabic.html
[32] http://pagesperso.lina.univ-nantes.fr/info/perso/permanents/enguehard/recherche/ana/iln.htm

d'après son auteur **[Enguerhard, 1993]**, les performances de ce système sont satisfaisantes. Il présente l'avantage d'être simples et non spécifiques à une langue particulière.

3.2.8. Les travaux de Dias (Le système SENTA)

La méthode implémentée dans le système SENTA (Software for the Extraction of N-ary Textual Associations) permet de trouver des associations lexicales N-aires de mots, qui ne sont pas forcément contigus. SENTA fonctionne selon un calcul de probabilités conditionnelles et un indice de « cohésion » lexicale entre différents mots d'une fenêtre de mots mobile : cet indice révèle potentiellement un terme du domaine lorsqu'il atteint un pic **[Dias, 2002]**.

3.2.9. Discussion : les approches statistiques

Les méthodes statistiques présentent l'avantage d'être rapides et simples à mettre en œuvre. En effet, ces méthodes s'appuient sur des formules statistiques et sur de simples calculs des fréquences. Ces méthodes ne nécessitent ni de connaissances spécifiques des langues des corpus, ni des domaines couverts par ces corpus. Les approches statistiques peuvent être qualifiées d'autonomes du fait qu'elles n'utilisent pas des ressources linguistiques externes au corpus (dictionnaire, stop liste…). Ces ressources sont généralement constituées manuellement et nécessitent beaucoup de temps et d'effort.

Cependant il est à noter que malgré leurs autonomies, les résultats obtenus par les approches statistiques sont fortement reliés aux corpus étudiés et ne peuvent pas être généralisés en dehors de ce contexte. Ces approches sont performantes sur des corpus de taille suffisamment grande. Elles ne sont pas applicables sur des corpus de petites tailles **[Harrathi, 2009].**

3.3. Les approches hybrides

Dans les modèles hybrides ou mixtes, les approches *statistiques* et les approches *linguistiques* sont associées ou couplées. L'ordre dans lequel cette association est effectuée varie d'un système à un autre. En effet, dans certains systèmes les résultats obtenus par une analyse linguistique sont validés et filtrés par une analyse statistique, tandis que dans d'autres systèmes les résultats de l'analyse statistique sont validés par une analyse linguistique.

Dans le contexte de la langue arabe, nous n'avons pas trouvé beaucoup de travaux qui ont testé ces approches. Par ailleurs il existe plusieurs travaux pour les autres langues, nous

pouvons citer ceux de : **[Smadja, 1993], [Daille, 1994], [Frantzi & Ananiadou, 1997]** et **[Frantzi & al, 1999]**. Dans la section suivante, nous allons décrire les seuls travaux sur la langue arabe de **[Boulaknadel, 2008]** que nous avons trouvés dans la littérature.

3.3.1 Système proposé par Boulaknadel

Le travail présenté dans **[Boulaknadel, 2008]**, s'inspire d'ACABIT **[Daille, 1994]**, dédié plutôt aux langues *françaises* et *anglaises*. Celui de Boulaknadel traite la langue arabe et il tourne autour du thème d'extraction de termes. Son objectif est d'améliorer l'indexation et la recherche d'information en arabe. L'auteur propose une plateforme intégrant divers composants, dont l'identification de termes complexes sur corpus, qui produit des résultats de bonne qualité en terme de précision, et ce en s'appuyant sur une approche mixte qui combine modèle statistique et données linguistiques. Pour la découverte des termes complexes et leurs variantes, Boulaknadel utilise une analyse partielle qui permet une formalisation des spécifications linguistiques. Elle utilise l'analyse morphologique pour permettre l'identification de certaines variantes des termes complexes relevant de la morphologie.

L'identification des patrons typiques s'effectue par la recherche de certains types de syntagmes nominaux en tenant pour acquis que ceux-ci se composent de séquences de parties de discours. Il s'agit donc de définir les patrons admissibles sous forme de règles et de localiser les séquences correspondantes. Apres l'analyse linguistique et l'extraction des candidats termes potentiels, la liste de ces termes est soumise à diverses mesures statistiques. Ces mesures permettent de calculer le potentiel terminologique de la séquence rencontrée. Un calcul de fréquence est utilisé pour valider la liste obtenue linguistiquement.

3.3.2. Discussion : les approches hybrides

Les approches hybrides fournissent des résultats de qualité. Elles présentent un compromis entre les méthodes statistiques et les méthodes linguistiques. L'idée d'associer ces deux dernières méthodes est pertinente. En effet, cette association profite de la finesse des analyses linguistiques et de la robustesse des analyses numériques. La puissance des méthodes hybrides provient de l'adoption de modèles traitant de l'information comme étant un ensemble de variables qualitatives **[Daille, 1994]**, offrant ainsi la possibilité de traitement des corpus de taille volumineux. En plus, les méthodes linguistiques permettent un filtrage des résultats obtenus afin de diminuer le bruit.

3.4. Evaluation des systèmes d'extraction des termes

L'évaluation des systèmes d'extraction des termes se focalise sur la qualité de la terminologie obtenue par ce système. Elle ne prend pas en compte de nombreux autres facteurs tels que la vitesse de traitement, la portabilité et la robustesse **[Paroubek & Rajman, 2000], [Daille, 2002].** Ces méthodes d'évaluation se basent toutes sur un corpus, une liste de référence et des mesures statistiques.

3.4.1. Le corpus de référence

Le corpus de référence pour l'évaluation doit couvrir un domaine unique. Les documents du corpus doivent être monolingues et suffisamment variés afin d'être représentatifs du domaine de spécialité du corpus.

3.4.2. La liste de référence

Il s'agit, d'une liste contenant des termes dits, *de référence* avec lesquels les résultats obtenus par les systèmes d'extraction des termes sont comparés **[Daille, 2002]**. Cette liste peut être construite à partir d'un dictionnaire spécialisé de même domaine que le corpus. Elle peut être aussi obtenue par l'extraction manuelle des termes du corpus d'étude, celle-ci est effectuée par des experts du domaine.

Cependant, un jugement humain d'un expert peut remplacer la liste de référence, dans le cas où il s'agit d'évaluer un seul outil. En effet, si plusieurs systèmes sont mis en compétition il est impossible de juger si l'expert n'a pas été influencé par les résultats des évaluations précédentes.

3.4.3. Les mesures statistiques

Traditionnellement, les mesures utilisées pour juger la justesse de l'extraction des termes sont la *précision* et le *rappel*.

$$Précision = \frac{(Nombre\ de\ termes\ extraits\ et\ qui\ sont\ présents\ dans\ la\ liste\ de\ référence)}{(Le\ nombre\ de\ termes\ extrait)}$$

$$Rappel = \frac{(Nombre\ de\ termes\ extraits\ et\ qui\ sont\ présents\ dans\ la\ liste\ de\ référence)}{(Le\ nombre\ de\ termes\ de\ la\ liste\ de\ référence)}$$

La précision permet d'évaluer le nombre correct de termes extraits et le rappel permet d'évaluer la proportion des termes corrects qui n'ont pas été extraits **[Daille, 2002]**.

4- Les approches d'extraction de relations

La majorité des travaux liés à l'extraction des relations sémantiques à partir des corpus textuels, ont été effectuées dans des cadres de construction et d'enrichissement des ontologies ou des thésaurus. Ils s'intéressent à l'extraction de deux types de relations : les relations *hiérarchiques* et les relations *non- hiérarchiques* **[Punuru, 2008]**.

Comme pour les approches d'acquisition des termes, Il existe également deux approches principales pour l'acquisition de relations entre termes. Les méthodes dites « *externes* » ou *statistiques*, se basent sur la comparaison des contextes d'occurrence, tandis que les méthodes dites « *internes* » ou *linguistiques*, reposent sur la structure morphologique des mots ou la structure lexicale des expressions [**Bernhard, 2006]**.

Ou encore comme les définit Séguéla [**Séguéla, 2001**], celles qui procèdent par étude de distributions de contextes autour des termes et celles dont le principe opérationnel d'extraction repose sur des formules linguistiques caractéristiques de relations lexicales.

4.1. Extraction des relations hiérarchiques

Les techniques existantes d'extraction et de repérage des relations hiérarchiques se basent sur des patrons syntaxiques ou lexico-syntaxiques. Dans un premier temps, un ensemble de patrons lexico-syntaxiques est défini (un pour chaque relation). Dans un deuxième temps, ces patrons seront projetés sur le corpus de texte afin de repérer les instances des relations. La construction des patrons lexico-syntaxiques est alors une étape préliminaire afin de découvrir les relations dans un corpus. Précisément, il s'agit d'une acquisition des marqueurs de relations à partir du corpus étudié **[Harrathi, 2009].**

4.1.1. Les travaux de M. Hearst

M. Hearst **[Hearst, 1992]** dans ses travaux sur l'extraction des liens d'hyperonymie à partir de textes, propose la méthode itérative suivante :

1. Sélectionner le type de relation R,
2. Etablir une liste de termes pour lesquels on a identifié cette relation,

3. Trouver dans le corpus des phrases où les termes reliés sont co-occurrents,
4. Trouver les régularités dans ces phrases et faire l'hypothèse que ces phrases sont la base de formules ou patrons qui indiquent la relation étudiée,
5. Si un nouveau patron a été repéré et validé, utiliser ce patron pour trouver d'autres couples en relation et revenir en (2).

Par exemple :

N° du Parton syntaxique	Patron Syntaxique	Relation d'hyperonymie $\forall\ NP_i \qquad 1 \leq i \leq n$
1	NP_0 such as {NP_0 , … , (and | or)} NP_n	hyperonymie (NP_i, NP_0)
2	Such NP_0 as {NP_i*, (and | or)} NP_n	hyperonymie (NP_i, NP_0)
3	NP_1 as {, NP_i} * {,} (or | and)} other NP_{n+1}	hyperonymie (NP_i, NP_{n+1})
4	NP_0 {,} (including | especially) {NP_i,}*(or|and) NP_n	hyperonymie (NP_i, NP_0)

Tableau II.2 : Les patrons utilisés par Hearst pour l'extraction de l'hyperonymie

Le *Tableau II.2* présente les patrons utilisés dans **[Hearst, 1992]** pour l'extraction de la relation d'hyperonomie. Dans ces patrons NP désigne un groupe nominal.

Par exemple, la phrase: «*The bow lute, such as the Bambara ndang, is plucked and has an individual curved neck for each string*», satisfait le patron 1 du *Tableau II.2*. Dans cette phrase, NP_0 correspond a «*bow lute*» et NP_n correspond a «*Bambara ndang*». La relation ainsi extraite est :

Hyperonymie (« *Bambara ndang* », « *bow lute* »)

La méthode, présentée par M. Hearst fournit des résultats jugés pertinents pour la relation d'hyperonymie. Cependant, l'auteur signale les difficultés pour la généralisation de ce type de méthode à d'autres relations comme la relation de méronymie et souligne qu'elle obtient de bons résultats pour l'identification de relations spécifiques.

La méthode présentée par M. Hearst a été reprise dans de nombreux travaux d'extraction des relations à partir du corpus : **[Rousselot & al, 1996]**, **[Morin, 1999]**, **[Seguela & Aussenac-Gilles, 1999]**, **[Condamines & Rebeyrolles, 2000]**. Ces travaux partent du même

principe : la découverte de schémas lexico-syntaxique dans un corpus. Ils effectuent une recherche itérative dans le corpus textuel des marqueurs d'une relation donnée et des couples de termes qui entrent dans cette relation.

4.1.2. Les travaux de E. Morin et C. Jaquemin

Dans le même but d'extraire des relations d'hyperonymie, le système présenté par E. Morin et C. Jaquemin **[Morin & Jacquemin, 2004]** est une association de (voir figure II.6):

1. PROMOTHEE : outil de structuration de termes simples en réseaux sémantiques **[MORIN, 1999a]**

2. ACABIT : outil d'extraction de termes composés **[Daille, 1996]**

3. FASTR : outil de détection des variations morphosyntaxiques des termes dans le corpus **[Jacquemin, 1996]**

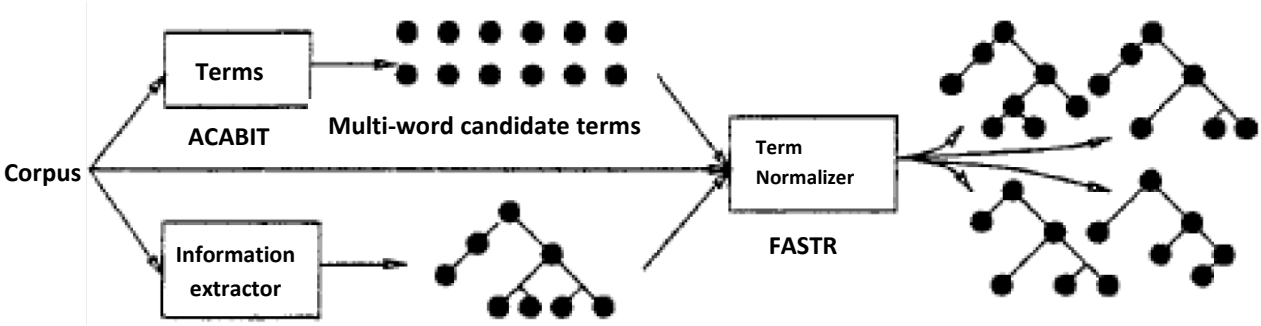

Figure II.6: vue d'ensemble du système proposé par E. Morin et C. Jaquemin [Morin & Jacquemin, 2004]

Pour trouver les relations entre les termes dans différentes phrases, le système tente d'identifier les variations des termes pour lesquels les relations sont déjà déterminées. Par exemple, si la relation hiérarchique entre « *fruits* » et « *pomme* » est connue, alors la relation entre les termes composés «*jus de fruits*» et «*jus de pomme*» est également marquée comme une relation hiérarchique. Les relations sémantiques entre les termes composés t1t2 et t1't2', se référant à des relations sémantiques entre les termes simples qui les constituent, sont marquées si l'une des trois contraintes suivantes est satisfaite :

1. une relation sémantique est connue entre t1 et t2 et/ou t1'et t2',

2. il existe un schéma de relation dans lequel t1 et t2 sont des têtes et t1'et t2'sont des arguments,

3. il existe une relation sémantique connue entre t1t2 et t1't2'.

4.1.3. Les travaux de R. Snow

Dans **[Snow & al, 2004]**, R. Snow propose une méthode d'apprentissage supervisée qui utilise les dépendances des chemins afin de chercher des patrons syntaxiques pour l'extraction des relations d'hyperonymie. Ces dépendances des chemins sont générées par des parseurs d'arbres de dépendance. Un parseur de dépendance produit un arbre des dépendances qui représente les relations syntaxiques entre les termes d'une liste de la forme: (terme1 : catégorie1 : Relation : catégorie2 : terme2) **[Lin & Pantel, 2001].**

Dans cette liste :

- les termes sont les formes singulières (les lemmes) des termes trouvés dans les phrases, par exemple « auteurs » devient « auteur », et ils correspondent à un nœud dans l'arbre de dépendance.

- les catégories sont les catégories grammaticales des termes considérés, par exemple nom et préposition.

- les relations sont les relations syntaxiques réalisées entre les termes, par exemple, la relation « objet » et la relation « modifier », et correspondent à des liens spécifiques dans l'arbre.

Dans l'arbre de dépendance, l'ensemble des plus courts chemins de longueur inférieure à cinq définit l'ensemble des patrons syntaxiques des relations sémantiques. La *Figure II.7* montre l'arbre de dépendance pour le fragment de la phrase « ...such authors as Herrick and Shakespeare» générés par le parseur MINIPAR[33] **[LIN, 1998]**.

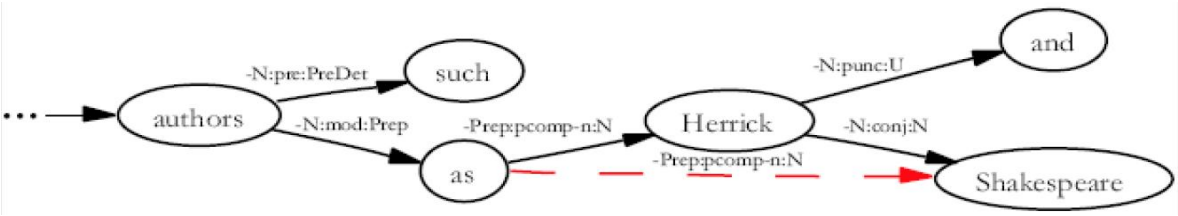

Figure II.7: **Exemple d'arbre de dépendance généré par MINIPAR** [SNOW & al, 2004]

[33] http://www.cs.ualberta.ca/~lindek/minipar.htm

Ils existent d'autres travaux sur l'extraction de relations hiérarchiques à partir de corpus. Nous trouvons les travaux présentés dans **[Ryu & Choi, 2004]** et **[Kashyap et al, 2004]** qui sont spécifiques aux corpus spécialisés couvrant le domaine de la *médecine*. Par ailleurs, les travaux de **[Fotzo & Gallinari, 2004]** utilisent des règles de subsumption dans une collection de documents afin de trouver les relations hiérarchiques. Pour repérer la relation d'hyponymie entre deux termes t1 et t2, les auteurs utilisent la fréquence relative. Cette fréquence relative consiste à comparer le nombre des documents contenant t1 et t2 au nombre des documents contenant t2 seul.

4.2. Extraction des relations non-hiérarchiques

En général, l'identification des relations non-hiérarchiques consistent à trouver dans un premier temps les paires ou les couples de termes qui forment les arguments d'une relation. Et dans un deuxième temps l'identification de l'étiquette pour la relation sémantique qui relie les termes arguments de la relation. Par exemple, dans le couple (« société », « produit »), l'étiquette de la relation peut être : « vendre », « fabrication », ou « consommer ».

Les travaux menés sur l'extraction des relations non-hiérarchiques à partir de corpus textuels, se sont limités à un certain nombre de relations. Dans la suite nous présentons deux relations : la relation de causalité et la relation partie-de.

4.2.1. La relation de causalité

Le système COATIS élaboré par D. Garcia **[Garcia, 1998]** a pour but le repérage des relations de causalité dans le corpus textuel. Ce système utilise des schémas de relations comprenant vingt-cinq relations de causalité, par exemple «créer», «empêcher», «faciliter» ou «pousser-à» dont l'élaboration se base sur le modèle proposé pour l'anglais par L. Talmy **[Talmy, 1988]**. La technique utilisée consiste à déclarer puis repérer un ensemble d'indicateurs linguistiques de la causalité, appelés « marqueurs de la relation ». Ces marqueurs sont en général des verbes, tels que « provoquer » ou « causer ». Et aussi des verbes tels que « gêner », « modifier » ou « contribuer », dont la valeur sémantique causale est confirmée par la coprésence dans le texte d'indices linguistiques complémentaires aux indicateurs. Les termes arguments, cause et effet, sont identifiés de la même façon, mais en utilisant d'autres indicateurs linguistiques.

Cette même démarche a été reprise par E. Cartier **[Cartier, 1997]** pour l'identification des définitions et par B. Goujon **[Goujon, 1999]** pour la veille technologique en anglais.

Dans **[GIRJU et al, 2002]**, R. Girju présente une technique semi-automatique d'extraction des patrons syntaxiques de la relation cause-effet. Cette technique relie un corpus volumineux à WordNet. La méthode proposée consiste à sélectionner à partir de WordNet un ensemble de couples de noms pour lesquels la relation cause-effet est identifiée. Par la suite, l'ensemble des couples est projeté dans le corpus afin de repérer les phrases dans lesquelles un couple est présent. Les phrases repérées sont de la forme < NP1 verbe | verbe expression NP2 >, où NP1 et NP2 sont des groupes nominaux. Un filtrage des couples de noms est effectué. Il ne conserve que les couples dont le second argument appartient à l'une des classes de WordNet «action de l'homme», «phénomène», «état», «fonction psychologique», et «événement». Les noms qui correspondent à NP1 doivent être une sous-classe de la classe «agent causal».

4.2.2. La relation partie-de

De nombreux travaux **[Berland et al, 1999] [Girju et al, 2003] [Turney, 2006]** ont été intéressés par l'extraction de ce type de relation. Ils se basent tous sur les patrons syntaxiques. Ces travaux diffèrent par la manière avec laquelle s'effectue l'extraction des patrons.

Dans **[Berland et al, 1999]**, M. Berland présente une technique d'extraction de la relation partie-de à partir d'un large corpus textuel anglais. L'auteur utilise deux indicateurs linguistiques : «basement» et «building», pour extraire les phrases dans lesquelles ces indicateurs sont présents. A partir de ces phrases, l'auteur extrait les patrons des relations. Après une validation manuelle, deux patrons ont été retenus. Les patrons sont ensuite projetés dans le corpus pour extraire d'autres paires reliées par la même relation. Les paires extraites sont triées en utilisant une métrique statistique se basant sur la probabilité conditionnelle.

Les travaux par R. Girju **[Girju et al, 2003]** peuvent être présentés, comme une extension des travaux de M. Berland **[Berland et al, 1999]**. R. Girju fait une analyse syntaxique du corpus. Cette analyse permet l'extraction de trois patrons de la relation partie-de. Ces patrons sont représentés dans le *Tableau II.3*.

N° du patron syntaxique	*patron syntaxique*
1	*NP1* of *NP2*
2	*NP1*'s *NP2*
3	*NP1* Verb *NP2*

Tableau II.3: Les patrons extraits par R.Girju

Pour identifier les paires valides susceptibles d'être des arguments de la relation « partie-de», l'auteur extrait les phrases du corpus satisfaisant l'un des patrons retenus. Ensuite, il utilise une technique d'apprentissage supervisée basée sur l'algorithme de l'arbre de décision « *C4.5* » **[Quinlan, 1993]** pour l'apprentissage des contraintes sémantiques. En cas d'ambigüité, l'auteur remplace les termes ambigus par des classes plus spécifiques de WordNet.

4.3. Outils d'extraction de relations

Parmi les travaux cités dans la section précédente, nous avons trouvé quelques outils pour l'extraction de relations à partir de textes, mais la majorité ne supporte pas la langue Arabe. Nous donnons tout de même des références pour ces outils d'extraction de relations dans d'autres langues principalement l'Anglais ou le Français comme les travaux de (Sundblad, 2002), la méthode CAMELEON (Séguéla, 2001), PROMETHEE (Morin, 1999), COATIS (Garcia, 1998), STARTEX (Rousselot, 1996), SEEK (Jouis, 1993).

Nous présentons par la suite un système que *l'on peut adapter* à la l'Arabe, qui est OntoBuilder.

4.3.1. OntoBuilder

MHiri décrit dans **[Mhiri & al, 2006]** une démarche pour la construction d'une ontologie dédiée à la modélisation des systèmes d'information (SI). Il procède à la détermination automatique des relations sémantiques entre les concepts de l'ontologie à partir d'une représentation conceptuelle (RC). Il présente un prototype nommé OntoBuilder pour la construction d'ontologies dédiées exclusivement aux SI. (voir figure II.8)

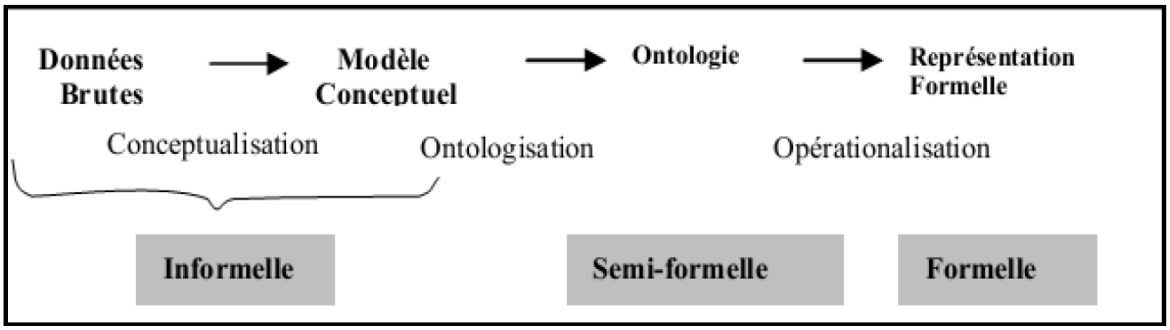

Figure II.8: Processus de construction d'ontologies (Mhiri & al, 2006)

Les données brutes dans OntoBuilder, sont sous la forme de représentations conceptuelles (RC), exprimées avec des diagrammes de classes UML. Les relations conceptuelles prises en compte sont: la généralisation-spécialisation, l'agrégation et la composition, les relations non nommées et celles ayant un nom. Les concepts et leurs relations, obtenus sont transformés en un diagramme appelé diagramme de concepts. Toute fois l'auteur ne donne *aucune évaluation* concernant le processus d'extraction de relations sémantiques.

4.4. Discussion : les approches d'extraction de relations

Malgré le grand nombre de travaux qui se sont intéressés à l'extraction des relations sémantiques entre les termes et entre les concepts, cette tâche reste toujours une tâche difficile à réaliser. Les différentes techniques proposées dans ces travaux, sont basées sur les patrons syntaxiques des relations. Ces patrons doivent être définis manuellement, et ensuite projetés dans le corpus spécialisé afin d'extraire d'autres patrons à partir des phrases satisfaisant les patrons de départ. La contrainte majeure de ces approches est qu'elles nécessitent un effort manuel non négligeable pour chaque domaine. Elles ne sont pas donc adaptables à d'autre domaine **[Jacquemin, 1996]**.

5- Les travaux sur la construction d'ontologies à partir des textes arabes

Concernant les travaux sur la construction des ontologies à partir des textes arabes, on trouve dans la littérature quelques approches proposées et pas de plateformes réelles, parmi ces travaux on cite les suivant:

- Les travaux de Mazari et al

[Mazari et al, 2012] ont proposé une approche de construction automatique des ontologies à partir des textes arabes en passant par les deux phases suivantes :

Pour identifier les termes pertinents qui dénotent les concepts associés au domaine, ils ont adopté d'abord une technique statistique d'extraction des termes qui s'appelle *« repeated segments »* (une fenêtre de 4 mots au maximum a été utilisée pour désignée un terme) , en suite les redondances seront éliminés en supprimant les segments inclus dans les autres avec le même nombre d'occurrence, et pour filtrer les segments obtenus ils ont appliqué deux filtres : le filtre de poids et le « cuting filter ». Ce dernier filtre supprime les segments contenant quelques mots tels que les verbes, les entités nommées, les chiffres en lettres …

Les nouveaux concepts extraits seront reliés à travers des relations hiérarchiques ou non hiérarchiques en utilisant deux méthodes : la méthode des marqueurs syntaxiques et la méthode des co-occurrences :

1-Pour extraire les relations sémantiques entre les termes, **[Mazari et al, 2012**] ont étudié le contexte du terme dans une fenêtre de quatre mots, et à partir de ce contexte la méthode va chercher des éléments lexico-syntaxiques pour identifier une relation entre termes. Ces éléments sont appelés les marqueurs linguistiques.

Pour chaque relation on peut trouver plusieurs marqueurs qui sont organisés par catégories dans des listes séparées qui dépendent du type de la relation à extraire. Ces listes seront incrémentées progressivement. Par exemple :

- Hyponymy or Generalization relation « is-a » : list = {**هو، هي ، هم، ...**}
- Meronymy relation «part-of » : list= { **تتألف-من ، تنقسم-الى ، تتكون-من،** }

A cause de la morphologie spécifique de la langue arabe telle que la vocalisation et l'agglutination, les listes des patrons syntaxiques doivent regrouper toutes les formes morphologiques susceptibles d'être rencontrées dans les textes arabes.

2- Si les marqueurs linguistiques sont absents dans le contexte des mots, l'approche sera basée sur les relations parent-fils : où le terme parent est plus général que le terme fils.

Cette relation entre termes est extraite à partir de la cooccurrence asymétrique du terme.

La relation est caractérisée par les deux règles suivantes :

- $P(x/y) \geq 0.8$
- $P(y/x) < P(x/y)$; *P(x/y) Est la probabilité d'apparition du terme 'x' puis le terme 'y', et inversement pour P(y/x)* **[Hernandez et Mothe, 2006]**

La première règle veut dire que les deux termes appariaient souvent ensemble (80% de cas). Selon la deuxième règle, x subsume y lorsque la probabilité de l'occurrence de x avant y

est supérieur à l'inverse. En utilisant la transitivité des relations, on pourra supprimer quelque relation : c.-à-d. si les relations "a" subsume "b», "a" subsume "c" et "b" subsume "c" sont extraites, la relation "a" subsume "c" peut être supprimée car elle peut être déduite à partir des autres **[Hernandez ,2006]**.

- Les travaux de Benaissa

La contribution de **[Benaissa, 2012]** s'est focalisée sur une ontologie lexicale, en prenant comme modèle *l'ontologie WordNet* et comme source d'entrée, *« les verbes arabes »* d'un dictionnaire monolingue contemporain sous forme d'une base de données lexicale. Le verbe, pivot de la phrase, est leur objectif pour la création des concepts en s'appropriant *les synsets* comme modèle de représentation du sens.

L'algorithme de *Markov pour le clustering* du graphe, généré par des différents définissants obtenue par une fermeture transitive, a été utilisé par les auteurs pour détecter les verbes semblables et d'identifier ainsi pour une entrée verbale donnée l'ensemble de ces synonymes.

- Les travaux de Zaidi et al

[**Zaidi–Ayad, 2013]** ont proposé une approche statistique pour l'extraction de termes simples à partir de corpus arabes (le saint Coran). Cette dernière est basée sur une méthode statistique d'extraction de termes simple suivi d'un filtrage par tf-idf , puis une approche hybride pour l'extraction des termes composés. Et pour les relations sémantiques, les auteurs proposent une méthode hybride d'extraction de relations à partir du texte Coranique. Ils ont utilisé d'abord une approche linguistique basée sur les règles JAPE[34] pour extraire des relations entre termes simples, puis ils ont validé les résultats obtenus par une approche statistique basée sur l'information mutuelle.

[**Zaidi–Ayad, 2013]** ont proposé par la suite une formalisation des concepts à l'aide de la logique de description pour permettre d'un coté la vérification des inconsistances et l'opérationnalisation de l'ontologie produite et son intégration dans d'autres applications. L'ontologie créée peut alors être utilisée dans l'amélioration de la recherche d'information,

[34] JAPE (Java Annotation Pattern Engine) est une composante de la plate-forme open source GATE. C'est un transducteur d'état fini qui fonctionne sur des annotations basées sur des expressions régulières.

l'indexation ou la traduction automatique ou dans toutes autres applications relevant du Web sémantique.

La contribution originale de [**Zaidi–Ayad, 2013]**, réside dans le fait que pour la première fois, l'accent est mis sur les outils d'extraction de termes et de relations à partir de textes arabes et plus spécifiquement le texte *coranique*, et aussi dans la formalisation de concepts arabes à l'aide de la *logique de description*.

6- Conclusion

Dans ce chapitre, nous avons présenté un état de l'art sur la construction des ontologies à partir des textes dans toutes les langues, mais nous avons toujours insisté sur les travaux qui supportent ou qui peuvent être adaptés à la langue arabe. Nous avons commencé par présenter un état de l'art sur le domaine de l'extraction des termes et des relations dans toutes les langues, ensuite nous avons présenté quelques approches de la construction des ontologies à partir des textes arabes.

Pour les approches d'extraction de termes, nous avons présenté trois familles d'approches : linguistiques, statistiques et hybrides, et nous avons vu les méthodes d'évaluation de ces approches. Et pour les relations, nous avons classé les approches d'extraction des relations en deux grandes familles : hiérarchiques et non hiérarchiques.

Nous avons remarqué qu'il y a beaucoup de travaux et d'outils sur l'extraction de termes et de relations pour les autres langues, nous avons souligné plus particulièrement le fait que les travaux les plus innovateurs datent de plus d'une décennie, en effet la majorité des travaux plus récents sont des reprises d'idées avec quelques améliorations relativement à l'approche adoptée ou une hybridation d'approches déjà existantes.

Nous avons vu malheureusement que la majorité de ces outils ne supportent pas la langue arabe, mais nous avons ressenti une petite recrudescence de travaux menés par des chercheurs dans le domaine du traitement automatique de la langue arabe et dans celui de la construction des ontologies à partir des textes arabes plus spécialement.

Le chapitre suivant présente notre contribution dans le domaine de la construction des ontologies à partir d'un corpus de textes arabes.

Chapitre 3
Contribution

1. Introduction

Parmi les sous-domaines de l'ingénierie des ontologies, on distingue la construction des ontologies à partir des documents textuels. Ces ontologies peuvent être utilisés dans plusieurs domaines de traitement automatique du langage naturel tels que la traduction automatique, la recherche d'information, l'annotation sémantique des ressources, l'indexation sémantique, les résumés automatiques de textes, les mémoires de traduction, etc.

L'utilisation des *textes* dans le processus de construction de l'ontologie est justifiée par deux arguments: premièrement, les textes sont souvent porteurs de connaissances stabilisées et partagées par les communautés de pratique. De plus, même si elles ne les remplacent pas complètement, les textes sont plus facilement disponibles que les experts qui n'ont pas souvent du temps pour participer au processus de construction **[Mondary & al, 2008]**. Mais il convient de noter que l'intervention des experts est nécessaire, que se soit dans le processus d'extraction des éléments de l'ontologie ou aussi dans l'évaluation et la validation de l'ontologie finale.

Le reste de ce chapitre est structuré comme suit: la première partie porte sur notre contribution et les étapes à suivre pour construire l'ontologie, et sur les méthodes utilisées pour préparer notre corpus; Dans la deuxième partie, nous discuterons les résultats de l'application de notre approche sur un corpus de textes arabes et nous finissons chaque partie par une évaluation des résultats obtenus.

2. Objectif

L'objectif de notre approche est de fournir une plateforme d'extraction de termes et de relations sémantiques pour la construction semi-automatique d'une ontologie à partir des textes arabes. Tout d'abord, nous commençons par l'étape de collection et de prétraitement du corpus. C'est une étape préliminaire avant tous processus d'extraction d'informations à partir de textes. Ensuite, nous appliquons une méthode statistique d'extraction des termes simples et composés à partir du corpus préparé. Et pour relier ces termes, des relations sémantiques sont extraites à partir des textes en se basant sur un ensemble de paires de relations, qui sont

utilisées comme un ensemble d'exemples pour l'apprentissage automatique des marqueurs linguistiques.

3. Approche proposée

Pour construire notre ontologie à partir d'un corpus de texte arabe, nous avons adopté un processus d'extraction de termes et de relations à partir de documents textuels basés sur trois phases; La première est la collecte et la préparation du corpus. Cette phase est très importante car la qualité de l'ontologie obtenue dépendra principalement de la qualité du corpus traité, de la méthode de prétraitement et de la complétude de la couverture du domaine. La deuxième phase est l'extraction de termes simples et complexes. Pour ce faire, nous avons choisi d'utiliser une méthode statistique, la méthode des "segments répétés" en recueillant des mots et des phrases fréquemment répétées et en filtrant ceux qui représentent vraiment des concepts de notre domaine, en utilisant le filtre TF-IDF (Term Frequency-Inverse Document Frequency) et le filtre coupant. Dans la dernière phase, nous avons utilisé une méthode d'apprentissage de marqueurs linguistiques pour relier les concepts extraits dans la deuxième étape avec des relations sémantiques.

Cette dernière phase se base sur le texte brute et un ensemble de listes préliminaires de paires (pour chaque type de relation) proposés par un expert du domaine, afin d'extraire les paires de relations existants dans le corpus de texte. Ces paires de relation sont utilisés comme un noyau pour apprendre d'autres paires à partir du corpus de texte en utilisant deux ressources externes : un dictionnaire de synonymes et d'antonymes et une base de données lexicale arabe. **[Benabdallah & al, 2017]**

La figure III.1 représente notre processus de construction de l'ontologie à partir du corpus de textes et ces deux ressources externes.

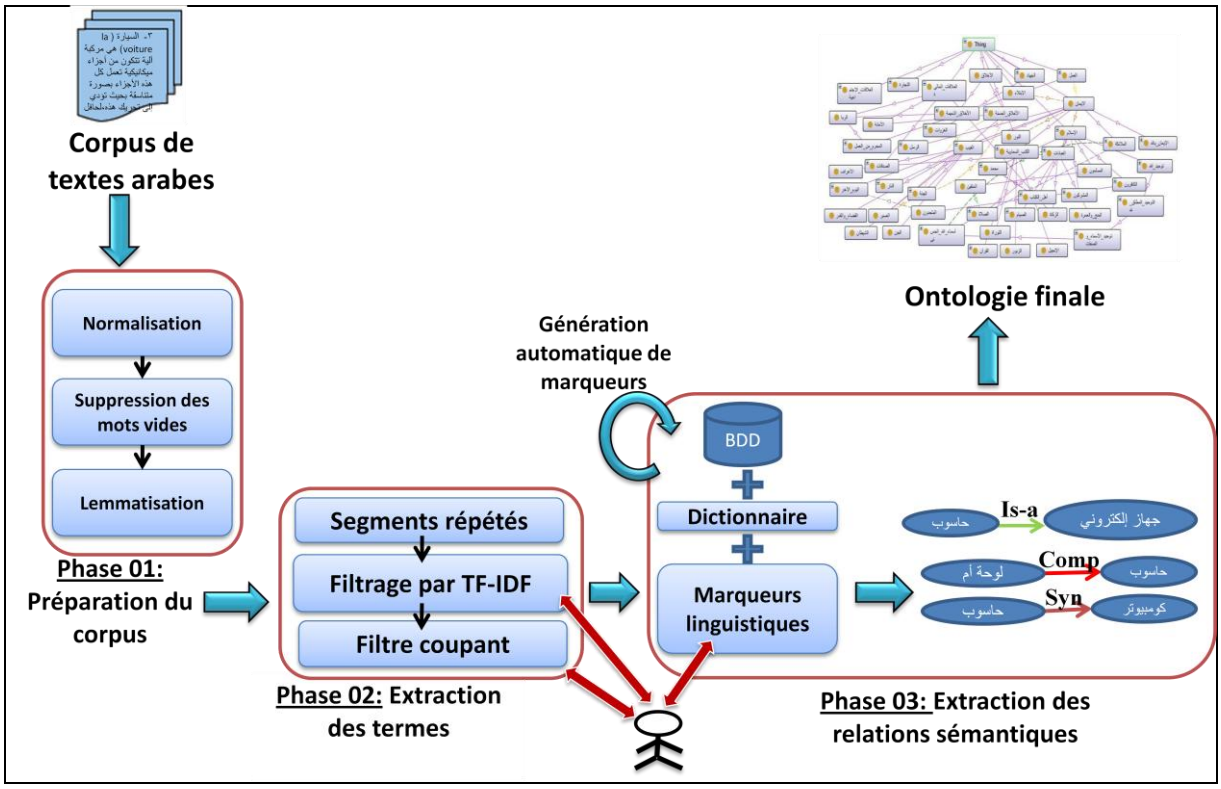

Figure III.1: **Architecture de notre système proposé** [Benabdallah, 2016]

3.1. Corpus

Dans le processus de construction d'une ontologie à partir de textes, la phase de la collecte et de la préparation du corpus est à la fois cruciale et délicate (voir section 3.1 du chapitre II), car le corpus est la source d'information essentielle pour le processus de construction d'une ontologie **[Bourigault & Aussenac-Gilles, 2003]**.

Les questions qui se posent dans la conception de tout corpus comprennent:

- le type de corpus (un corpus «spécialisé» est un corpus contenant des textes sur un sujet lié à un domaine de connaissance, par exemple, dans notre cas, « domaine de l'informatique en arabe ») .
- L'aptitude au projet prévu.
- La possibilité de réutiliser ce corpus.
- La taille (nombre de mots).
- La représentativité (c'est-à-dire la variété des textes, des auteurs, des sources, etc.).
- Et l'utilisation de textes ou d'échantillons complets , ... etc. [**Marshman, 2003**].

Afin de tester notre approche, nous avons collecté notre corpus de texte à partir d'un ensemble de documents textuels en Arabe, contenant des définitions de termes techniques en informatique en arabe. Le tableau III.1 contient quelques exemples de documents de notre corpus.

Titre du document en Arabe	Titre en français	Taille en Kilo octets	Type du document
المعجم الموسوعي في الكمبيوتر والإلكترونيك	Dictionnaire encyclopédique en informatique et en électronique	10 877	PDF
شبكة الحاسوب	Réseau informatique	176	HTML
موسوعة مصطلحات الانترنت والكمبيوتر	Encyclopédie des termes de l'internet et de l'informatique	98	HTML
مصطلحات المعلوماتية	Les termes informatiques	175	DOCX
أهمية الحاسوب في البحث العلمي - علم النفس المعرفي	L'importance de l'informatique dans la recherche scientifique - Psychologie cognitive	100	DOCX
مفاهيم الأنترنت	Les concepts de l'Internet	347	PDF
…	…	…	…

Tableau III.1 : Exemples de documents de notre corpus

Ces documents (en format "HTML" ou "PDF") sont téléchargés, sélectionnés, préparés manuellement (en supprimant des tableaux, graphiques, images ...) et convertis au format "TXT".

Après un filtrage manuel par un expert, nous avons choisi les documents les plus pertinents et les plus représentatifs de notre domaine. Le résultat de cette opération est un corpus de 37 documents avec 304 665 mots et une taille de 7720 Ko.

3.2 Ressources externes

Notre processus d'extraction de relations sémantiques se base sur un algorithme d'apprentissage de marqueurs linguistiques qui nécessite une base d'exemple de paires de relations. Cette base d'exemples peut être construite manuellement par un expert, ou à travers des ressources existantes. Pour notre approche, nous avons préféré de profiter de ces ressources, ce qui nous a poussé à rechercher des ressources de textes externes au corpus. Les

deux ressources de textes que nous avons choisis sont un dictionnaire de synonymes et d'antonymes et une base de données lexicale :

3.2.1 Le dictionnaire

Pour avoir une liste de paires de relations, pour les deux relations : synonyme et antonyme, nous avons opté pour le dictionnaire arabe : "قاموس الطالب في المرادفات والأضداد"[35] (en français : « *Dictionnaire de l'étudiant pour les synonymes et les antonymes* ») **[Khaled & Saad, 2012]**. C'est un dictionnaire qui contient environ 450 paires pour les deux relations. *Le tableau III.2* donne quelques exemples de ces paires de relations en arabe et leurs traductions en français.

Nom de la relation	Synonyme (المرادفات)		Antonyme (الأضداد)	
Les paires de la relation	ابتكار (innovation)	اختراع (invention)	اقفال (fermeture)	افتتاح (ouverture)
	ارتباط (corrélation)	اتصال (connexion)	الإسراع (accélération)	التأخير (retardement)
	عداد (Compteur)	حاسب (ordinateur)	الممتلئ (rempli)	الفارغ (vacant)
	إغلاق (verrouillage)	اقفال (fermeture)	التجميع (assemblage)	التقسيم (division)
	حاسوب (ordinateur)	كمبيوتر (ordinateur)	جزئي (partiel)	كلي (total)
	...	...	...	...

Tableau III.2 : Quelques exemples de paires de relations du dictionnaire.

3.2.2 La Base de données lexicale

Pour la relation de généralisation ou spécilisation (hyperonyme ou hyponyme), nous avons choisi d'utiliser quelques paires existantes dans la base de données lexicale WordNet Arabe[36] (350 paires de relation).

Wordnet Arabe (en anglais **AWN** : **A**rabic **W**ord**N**et) est une base de données lexicale. Sa conception basé sur Princeton WordNet (La version anglaise) est construite suivant des méthodes développées pour EuroWordNet est reliée avec l'ontologie SUMO (Suggested Upper Merged Ontology). Arabic WordNet a été développé par DOI / REFLEX (2005-2007) **[Black & al, 2006].**

[35] http://saaid.net/book/16/8190.pdf visité le 10 mai 2016
36 http://globalwordnet.org/arabic-wordnet/awn-browser/ visité le 10 mai 2016

Arabic WordNet comporte beaucoup de termes et donc beaucoup de relations d'hyperonymie entre ces termes, et pour notre approche, nous nous intéressons seulement aux quelques paires de relations qui concernent notre domaine. Et pour filtrer ces paires de relation, un expert du domaine a choisi manuellement les paires adéquates. Le tableau III.3 comporte quelques exemples de ces paires de relation.

Nom de la relation	**Hyperonyme**	
	Concept n°1	**Concept n°2**
Les paires de la relation	آلة (Machine)	حاسوب (ordinateur)
	جهاز (Appareil)	آلة (Machine)
	شيء مصنوع (une chose fabriquée)	جهاز (Appareil)
	جسم (corps)	شيء مصنوع (une chose fabriquée)
	شفرة (code)	برنامج حاسوب (Logiciel)
	...	...

Tableau III.3: Quelques exemples de paires de relations de la BDD lexicale ArabicWordnet.

La figure III.2 montre l'interface graphique de « AWN », avec un exemple d'une relation de généralisation entre les deux termes : ordinateur (حاسوب) et machine (آلة).

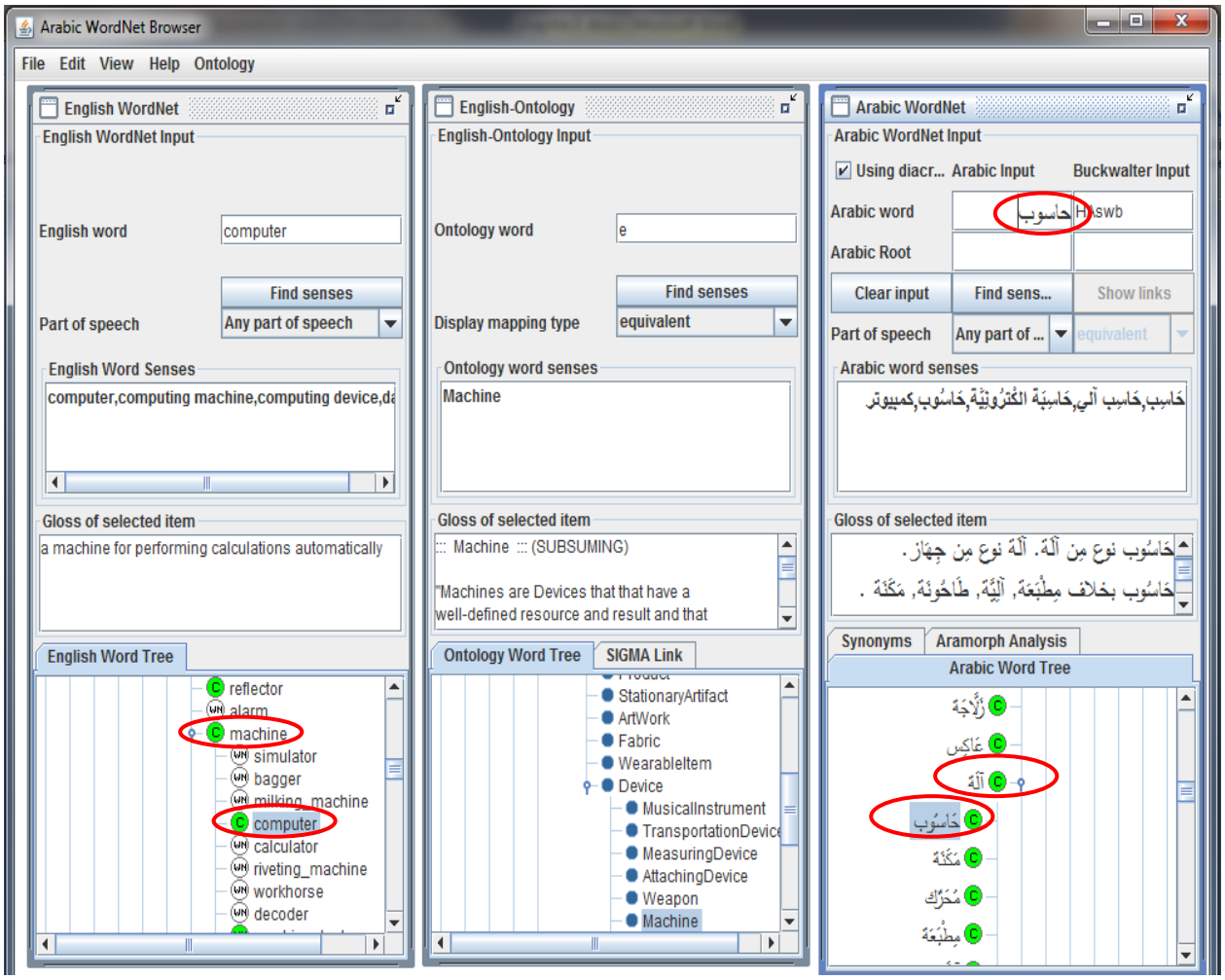

Figure III.2: Exemple de deux paires de la relation hyperonyme sous AWN

4. Prétraitement du corpus

Après la collecte des documents textuels du corpus, ce dernier doit être préparé pour les traitements ultérieurs. Cette phase est réalisée en se basant sur un ensemble d'étapes de prétraitement, afin de supprimer une certaine ambiguïté, réduire la quantité des traitement ultérieurs et adapter le corpus à l'objectif final "extraction des candidats-termes ".

Dans cette phase de prétraitement, le texte du corpus passe par trois étapes:

a. **Normalisation (ou standardisation) :** elle consiste à convertir le document en un format standard facilement manipulé. Avant la lemmatisation, le document est normalisé comme suit:

✓ Supprimez les caractères et les chiffres spéciaux, par exemple: ٢,٣,٤,٥,(,؟,+, »...

- ✓ Supprimer les mots et les caractères latins: les caractères latins sont détectés par leurs graphiques.
- ✓ Supprimez les lettres simples: les mots d'une lettre en arabe et les abréviations. Par exemple: la numérotation (paragraphe "B" ب "الفقرة"), abréviations de date (م: ميلادي ه-: هجري,), voyelles (و "حروف العلة"), etc.

b. Suppression des mots vides: cette étape consiste à éliminer tous les mots vides(en anglais : stop-word), comparer chaque mot reconnu avec les éléments de la liste des mots vides. Il s'agit d'une liste de tous les mots d'outil, de liaison et d'articulations (pronoms (الضمائر المنفصلة), prépositions (حروف العطف), conjonctions (حروف الجر), etc.) par exemple: التي, أن, على في, من, إلى, عن , هذا, هذه, بين, بعد, مع, الذي انه, منذ, ما, لم ...

Généralement, les mots vides qui sont très courants (presque la moitié des occurrences des mots du texte) ne sont pas indexés car ils ne sont pas informatifs **[Vergne, 2004]**.

c. Lemmatisation: c'est une tâche délicate car l'arabe est une langue inflexible; Le manque de signes diacritiques (représentations écrites des voyelles) dans la plupart des textes écrits en arabe crée une ambiguïté et exige donc des règles morphologiques complexes. Plus, les mots en lettres majuscules ne sont pas utilisés en arabe, ce qui rend difficile l'identification des noms propres, des acronymes et des abréviations.

Pour résoudre l'ambiguïté, Aljlayl et Frieder **[Aljlayl & Frieder, 2002]** ont montré que la lemmatisation légère (une approche basée sur la suppression des suffixes et des préfixes) surpasse la lemmatisation basée sur la racine dans le domaine de la recherche d'informations.

5. Extraction de termes

Après le prétraitement du corpus, nous passons à l'étape d'extraction des termes de l'ontologie. Tout d'abord, nous détectons tous les termes différents du corpus par la méthode des segments répétés; Ensuite, nous appliquons deux filtres pour supprimer les termes qui ne sont pas considérés comme des concepts du domaine.

La méthode des "segments répétés" est basée sur la détection de tous les segments possibles du texte, composés de morceaux, apparaissant plusieurs fois dans le même texte. Il

s'agit d'une technique statistique pour extraire des informations à partir de textes. La répétition des segments dans un texte indique que ceux-ci peuvent être utilisés pour décrire les concepts spécifiques au domaine du corpus. Les segments de texte peuvent être séparés par des espaces ou par des signes de ponctuation, et peuvent être simples (un seul mot) ou complexes.

Les termes complexes sont identifiés dans une fenêtre de quatre mots dans la même phrase (le numéro quatre est choisi selon le principe: un segment représentant un concept contient généralement quatre mots en maximum **[Mazari, 2013]**).

5.1. Calcul du poids des termes par la formule tf-idf

La méthode des «segments répétés» repose sur la proposition suivante: *«Un terme pertinent est utilisé plusieurs fois dans un corpus de texte spécialisé»*. Pour cette raison, nous utilisons *"un filtre de pondération"* pour sélectionner des termes avec un poids suffisant. Le poids est mesuré par la méthode de pondération TF-IDF (Term Frequency - Inverse Document Frequency). C'est une mesure statistique numérique destinée à refléter l'importance d'un mot ou d'un segment pour un document dans une collection ou dans un corpus.

La fréquence du terme (TF) est simplement le nombre d'occurrences de ce terme dans le corpus divisé par le nombre total de termes dans le corpus.

La fréquence inverse de document (IDF) est une mesure qui donne plus de poids aux termes moins fréquents considérés comme plus discriminants. Il calcule le nombre de documents contenant un terme cible, puis divise ce nombre par le nombre total de documents dans le corpus. Enfin, le poids TF-IDF est le produit de deux statistiques, de la fréquence des termes et de la fréquence inverse des documents. **[Sparck, 1972]**

Le *tf-idf* peut se décrire formellement comme suit : pour un terme *i* dans un document *j* parmi les *N* documents du corpus.

$$w_{ij} = tf_{ij} \times idf_i \qquad \textbf{Avec} \quad idf_i = \frac{n_i}{N}$$

Où :
- n_i : est le nombre de documents dans lesquels apparaît le terme *i*.
- N : le nombre total de documents.

Si le TF-IDF est supérieur à un seuil: f_{min} (un seuil indiquant la pertinence du terme), le terme sera conservé pour la prochaine étape du processus de construction, sinon il sera ignoré. (f_{min} est choisi empiriquement par un expert et dépend de la taille et la nature du corpus).

5.2. Application du filtre coupant

Dans cette phase, les segments qui seront supprimés sont les segments contenant certain mots tel que les mots qui n'ont pas de racine dans la langue arabe (entités nommés), les numéros en lettres et autres. Les mots du filtre coupant peuvent être présent au début, à la fin ou à l'intérieur d'un segment. La liste des mots du filtre coupant ou *"cut filter"* peut être consultée et modifiée par l'utilisateur suivant la spécificité du corpus de texte traité. Après cette phase les termes restants ne contenant aucun mot de la liste du *"cut filter".*

5.3. Résultats

Pour extraire les termes de notre ontologie, nous définissons les paramètres suivants:

- Taille maximale du segment = 4 mots. Cela indique la taille maximale d'un terme complexe. Exemple: بروتوكول نقل البريد البسيط (protocole de transfert de courrier simple)
- Seuil de pondération f_{min}: Le poids d'un terme est calculé par TF-IDF. Le seuil de pondération d'un mot simple est 4E-05. Le seuil de pondération d'un terme complexe est 5E-06. Ces seuils sont sélectionnés par un expert par rapport à la taille du corps.

Notre système extrait 528692 segments différents (184738 termes simples et 343954 termes complexes), mais il ne sélectionne qu'une liste de 741 segments en fonction des seuils définis ci-dessus (552 termes simples et 189 termes complexes). Le tableau III.4 présente quelques exemples de segments sélectionnés:

Segments	Nombre d'occurrences dans le corpus	TF	Nombre de documents contenant le segment	IDF	TF-IDF
(Ordinateur) كمبيوتر	3045	0,009983128	27	0,72972973	7,28E-03
(Information) معلومات	2233	0,004223631	35	0,945945946	4,00E-03
(Internet) إنترنت	5278	0,009983128	14	0,378378378	3,78E-03
(Site) موقع	1624	0,003071732	15	0,405405405	1,25E-03
…	…	…	…	…	…
(réseau mondial) الشبكة العالمية	203	0,000383966	10	0,27027027	1,04E-04
courrier) البريد الإلكتروني (électronique	101	0,000191038	11	0,297297297	5,68E-05
…	…	…	…	…	…

Tableau III.4 : **Exemples de segments extraits à partir de notre corpus** [Benabdallah & al, 2017]

5. 4. Evaluation

Dans cette section, nous présentons les résultats des évaluations effectuées sur notre corpus textuel pour l'extraction des termes simples et composés. Le tableau III.5 résume les résultats de l'application de notre approche sur l'extraction des termes sur un sous-ensemble de mots de notre corpus de texte.

Les mesures utilisées généralement, pour juger la pertinence de l'extraction des termes sont la précision, le rappel et le F1-mesure. La précision permet d'évaluer le nombre correct de termes extraits alors que le rappel permet d'évaluer la proportion des termes corrects qui n'ont pas été extraits. Le F1-mesure correspond à la moyenne harmonique entre la précision et le rappel. Ces mesures sont calculées par les formules suivantes :

$$\text{La précision} = \frac{\text{le nombre d'entités correctes extraites par notre système}}{\text{le nombre total d'entités extraites par notre système}}$$

$$\text{Le rappel} = \frac{\text{le nombre d'entités correctes extraites par notre système}}{\text{le nombre total d'entités correctes dans le corpus}}$$

$$\text{La F1-mesure} = \frac{2 * \text{rappel} * \text{Précision}}{\text{rappel} + \text{précision}}$$

Pour obtenir le rappel et la précision dans le cadre de cette évaluation, nous devons savoir combien de termes et de relations correctes existe réellement dans le corpus de teste. Pour découvrir cela, un expert du domaine a lu notre corpus de teste et il a étiqueté manuellement

toutes les entités correctes existantes. Sur la base des résultats obtenus par ce traitement manuel et des résultats obtenus par notre système, la précision et le rappel ont été calculés comme indiqué dans le tableau III.54:

Elements de l'ontologie	Précision	Rappel	F1-mésure
Termes simples	0,94	0,88	0,90
Termes composés	0,84	0,76	0,79

Tableau III.5: Évaluations des résultats de la reconnaissance des termes

Les résultats présentés dans le tableau III.5 concernant les deux types de termes simples et complexes montrent une précision égale à 89% en moyenne, et un rappel égal à 82% en moyenne, ce qui est un bon niveau pour ce type de tâche. On notera en particulier le haut niveau de la précision qui caractérise un niveau de fiabilité très important.

6. Extraction de relations

Les travaux existants dans le domaine de l'extraction des relations sémantiques à partir du texte peuvent être divisés en deux familles: nous distinguons généralement les travaux qui se basent sur l'aspect *fréquentiel* ou statistique du corpus à partir duquel ils extraient des paires de relations, et ceux qui exploitent des indices structurels pour détecter les éléments qui peuvent être reliés entre eux, c'est-à-dire, suivant une approche *symbolique* **[Claveau & Sébillot, 2004]**.

Les méthodes symboliques d'extraction des relations sont basées sur la détection du contexte des occurrences des mots pour décider de leur acquisition ou non; Le classificateur symbolique est souvent un ensemble de règles basées sur des indices lexicaux, morphologiques, catégoriques, syntaxiques ou autres. Ces techniques elles-mêmes peuvent être divisées en deux familles principales **[Claveau & Sébillot, 2004]**:

- Les approches *linguistiques* dans lesquelles les indices structurels donnés a priori (par analyse linguistique, par exemple) sont les facteurs fondamentaux du processus décisionnel.
- Et les approches basées sur une notion d'*apprentissage* de *marqueurs* ou de *relations*.

Dans le cadre de ce travail, nous avons opté pour une approche reposant sur une méthode d'apprentissage des marqueurs linguistiques à partir du texte.

Dans les méthodes d'extraction des relations par marqueurs linguistiques, le principe est de définir initialement, un ensemble de listes de marqueurs lexicaux et syntaxiques (une liste pour chaque relation). Ensuite, ces marqueurs seront projetés sur le corpus original pour identifier les instances de relations.

Le tableau III.6 donne quelques exemples de marqueurs linguistiques de la langue arabe:

Type de relation	***Marqueurs linguistiques de la langue arabe***
Hyperonyme et hyponyme (généralisation ou spécification)	هو نوع من، صنف من، هو، هي، هم، وغيره من أنواع، .....
Méronyme (partie de)	هو جزء من ، تتكون من ، تنقسم إلى ، تتألف من ، .....
Antonyme (opposé)	ضد، عكس، بخلاف.....
Synonyme	مرادف لـ ، يعني، مرادفه هو، له نفس معنى ،.....
….	…

Tableau III.6 : Liste de certains marqueurs linguistiques arabe.

Ces listes nous permettent d'extraire un certain nombre d'occurrences de relations à partir du corpus, mais ces occurrences de relations restent insuffisantes par rapport à la taille du corpus et par rapport au nombre de concepts extraits dans l'étape précédente (plusieurs concepts restent isolés, c'est-à-dire non reliés avec l'ontologie).

La construction de *marqueurs linguistiques* est alors une étape préliminaire pour identifier les relations dans le corpus. En raison de la morphologie spécifique de l'arabe comme la vocalisation et l'agglutination, les listes de marqueurs d'extraction des relations doivent regrouper toutes les formes morphologiques qui peuvent être rencontrées dans les textes arabes.

La solution consiste alors, à apprendre ces marqueurs *automatiquement* à partir du texte.

6.1 Apprentissage des marqueurs

Dans la plupart des travaux sur l'acquisition de relations sémantiques par apprentissage des marqueurs linguistiques, l'objectif est d'identifier des marqueurs ou des indices sur une relation sémantique à partir d'un ensemble d'exemples dans un corpus et de les réutiliser pour extraire de nouvelles unités de relation. C'est une approche initiée par Hearst **[Hearst, 1992]** pour acquérir des liens *hyperonymes* pour la langue *anglaise* (voir section 5.1.1 du chapitre 2, pour plus de détail).

Dans le cadre de ce travail, nous avons repris l'algorithme de Hearst, et nous l'avons appliqué sur la langue arabe avec trois types de relations : hyperonymes, synonymes, et antonymes **[Benabdallah & al, 2017]**. L'algorithme d'identification des paires de relations par apprentissage de marqueurs linguistiques à partir du corpus de texte est comme suit:

Algorithme : Extraction de relations sémantiques.

Entrées :

- ✓ Un corpus de textes.
- ✓ Un dictionnaire.
- ✓ Une base de données lexicale.

Sorties :

- ✓ Ensemble de paires de relation.

Debut

1. Pour chaque relation cible $\mathcal{R}$ dans l'ensemble {hyperonyme, synonyme, antonyme, méronyme, holonyme, ...} faire:

2. Choisissez une relation cible $\mathcal{R}$ (dans notre cas, nous avons testé uniquement les trois relations: hyperonyme, synonyme et antonyme);

3. Assembler un ensemble de paires de relation de $\mathcal{R}$ (ces paires peuvent être extraites à partir d'un thésaurus, d'une base de connaissances, d'un dictionnaire ou construites manuellement par un expert du domaine);

4. Trouver toutes les phrases du corpus original (non prétraité) contenant ces paires et enregistrer leurs contextes;

5. Trouver les points communs entre ces contextes et supposer qu'ils forment un $\mathcal{R}$-design;

6. Utilisez les résultats trouvés dans **5** pour obtenir de nouvelles paires et revenir à **4**.

Fin.

Contrairement aux approches *linguistiques* qui utilisent des connaissances a priori pour extraire des relations, les approches d'*apprentissage* sont basées sur une analyse d'exemples

pour apprendre des marqueurs d'extraction pour détecter de nouvelles paires de relations. La base d'exemples contenant ces paires de relations peut être construit manuellement ou à partir d'une ressource externe comme: un thésaurus, une base de données lexicale, un dictionnaire ou autre.

Parmi les exemples de travaux qui utilisent des ressources externes pour extraire des relations sémantiques on trouve : Girju et Moldovan **[GIRJU & al, 2002]** qui ont proposé une technique d'extraction semi-automatique de pattrons syntaxiques en utilisant les relations existantes dans WordNet anglais (voir section 5.2.1 du chapitre 2, pour plus de détaille).

Dans notre approche, nous utilisons deux types de ressources externes pour obtenir des paires de relations: un dictionnaire et une base de données lexicale.

Pour les relations de synonymie et d'antonymie, nous utilisons un dictionnaire contenant des paires de relations les plus connues en langue arabe (voir section 3.4.1).

Pour la relation de généralisation-spécialisation ou Hyperonymie, nous utilisons des paires existantes dans WordNet Arabe (voir section 3.4.2).

6. 2. Résultats

En appliquant l'algorithme d'apprentissage de marqueurs précédant sur notre corpus de texte arabe, notre application a détecté un ensemble de marqueurs et de paires pour chaque relation sémantique. Le tableau III.7 présente quelques exemples de ces résultats:

Relations	**Les *paires* de relations et *marqueurs* extraits à partir du corpus**		
	Terme1	**Marqueurs détectés** (Séquence de mots)	**Terme2**
Terme1 Hyperonyme du **Terme2**	الملفات (fichiers)	[من],[غيرها],[و] (et les autres)	الصور (images)
	لغات البرمجة languages de programmation	[من],[بين],[الأصناف],[الأكثر],[انتشارا], [ل] (parmi les types les plus populaires des)	جافا (java)
	أنظمة التشفير (systèmes de codages)	[شكل], [من],[أشكال] (est une forme de)	يونيكود (Unicode)
	الحواسيب (ordinateurs)	[له],[تأثير],[كبير] [على],[باقي] (a un impact significatif sur le reste des)	الخادم (serveur)
	...	...	...
Synonym of	البيانات (données)	[في],[بعض],[الأحيان],[هي],[مرادفة] [ل] (parfois est un synonyme de)	المعطيات (données)
	شيفرات (codes)	[يمكن],[أن],[تسمى],[ب] (peuvent être appelés)	رموز (symboles)
	الكمبيوتر (ordinateur)	[هي],[كلمة],[تعني] (est un mot qui signifie)	الحاسوب (ordinateur)
	...	...	...
Antonym of	الدائمة (permanente)	[تعني] , [عكس] (signifie le contraire de)	العشوائية (aléatoire)
	المستفيد (client)	[بخلاف] , [كلمة] (Est opposé au mot)	الخادم (serveur)
	المجاهيل (inconnues)	[على] , [خلاف] (sont en désaccord avec)	المعطيات (données)
	...	...	...

Tableau III.7: **Quelques exemples d'instances de relations sémantiques extraites à partir de notre corpus** [Benabdallah & al, 2017]

6. 3. Evaluation

Nous présentons dans cette section, les résultats des évaluations effectuées sur notre corpus pour l'extraction des relations sémantiques. Le tableau III.8 résume les résultats de l'application de notre approche sur l'extraction des relations : hyperonyme, synonyme et antonyme, sur un sous-ensemble de mots de notre corpus de texte.

Les mesures utilisées ici sont les mêmes mesures de la section qui concernent l'extraction des termes, à savoir : la précision, le rappel et le F1-mesure. Pour calculer ces valeurs un expert du domaine a lu notre corpus de teste et il a identifié manuellement toutes les relations correctes existantes. Sur la base de ce traitement manuel et les résultats obtenus par notre système, nous avons calculé les valeurs du tableau III.8.

Eléments de l'ontologie		Précision	Rappel	F1-mesure
Relations	« synonyme »	0,90	0,58	0,71
	« antonyme »	0,91	0,67	0,77
	« hypéronyme »	0,90	0,65	0,75

Tableau III.8: Évaluations des résultats de la reconnaissance des relations sémantiques

Comme dans le cas de la reconnaissance des termes, l'évaluation des relations extraites par notre système se caractérise par une précision élevée et un rappel moyen. La différence entre la précision et le rappel est également plus accentuée dans ce cas que pour la reconnaissance des termes. Nous pouvons donc dire que les relations produites par notre système ont généralement une bonne fiabilité, mais que les marqueurs linguistiques qui sont automatiquement appris ne couvrent pas toutes les formes possibles de la langue arabe (car les formes des marqueurs linguistiques de la langue arabe sont très variés), et il est donc nécessaire que notre approche soit appliquée à d'autres corpus de textes différent de notre corpus, afin d'apprendre plus de marqueurs et donc extraire plus d'instances de relations. On peut remarquer aussi dans le tableau III.8, la différence entre le rappel de la relation de synonymie (*0.58*) et les autres relations, car le nombre de relations de synonymie correctes extraite par notre système n'est pas assez grand par rapport aux autres relations. Nous pouvons interpréter ce rappel par le manque de marqueurs linguistiques appris automatiquement pour la relation de synonymie, car notre corpus de textes n'est pas riche en termes de marqueurs de cette relation.

7. Conclusion

L'approche que nous venons de présenter dans ce chapitre consiste à extraire des éléments constituant une ontologie de domaine de la langue arabes. Ces éléments sont extraits à partir d'un corpus de texte d'un domaine spécifique.

Nous avons commencé d'abord, par la présentation de notre objectif ainsi que l'architecture de l'approche que nous avons proposé pour résoudre notre problématique. En suite, nous avons montré les trois ressources de textes utilisées dans le processus de construction de l'ontologie, qui sont : le corpus de texte, le dictionnaire et la base de données lexicale.

Par la suite, nous avons expliqué en détaille les étapes de notre processus de construction de l'ontologie qui sont : le prétraitement du corpus, l'extraction des termes et l'extraction des relations sémantiques. Et finalement, nous avons présenté les résultats de l'application de notre système sur le corpus de textes ainsi que la méthode d'évaluation de notre approche avec une discussion de cette évaluation pour chaque étape.

Finalement, nous pouvons dire que les résultats obtenus par notre approche d'extraction de termes et de relations pour la construction d'une ontologie peuvent être jugées comme satisfaisants. Et il convient de dire aussi que notre approche peut être appliquée à d'autre corpus de texte d'un domaine différent de notre domaine et même aussi à une langue différente de la langue arabe.

Conclusion générale

Conclusion générale

Actuellement, la surcharge d'information dans tous les domaines a généré un besoin capital de structuration des contenus des documents, disponibles généralement sur le web ou sur autres sources d'information. Les ontologies en sont un moyen prometteur et qui ne cesse de donner ses preuves. Leur construction manuelle s'est avérée trop onéreuse et les ontologies obtenues sont très difficiles à réutiliser. La construction semi-automatique commence à donner des résultats encourageants, vu la facilité relative à les mettre au point et à être plus partageables et plus réutilisables. Les ontologies en langue *Arabe* sont peu existantes, pourtant l'Arabe est une langue parlée par plus de 300 millions de personnes dans plus de 22 pays.

Dans le cadre de cette thèse, nous nous sommes intéressés aux méthodes d'extraction des constituants les plus importants d'une ontologie à partir des textes arabes.

Dans la première partie de cette thèse, nous avons tout d'abord présenté une vue globale des approches existantes pour la construction d'ontologies à partir des textes dans toutes les langues, puis nous avons détaillé les approches qui, comme celle que nous avons présentée dans cette thèse, utilisent des textes en langue Arabe. Nous avons évoqué que notre approche se distingue des autres à la fois par l'utilisation des ressources externes au corpus de textes, mais aussi par la manière dont nous reconnaissons les termes et les relations sémantiques. Ainsi, la contribution de ce travail se résume selon trois principales phases :

La première est la collecte et la préparation des documents textuels en langue arabe constituant notre corpus. La qualité de l'ontologie obtenue dépendra principalement de la qualité de ce corpus, de la méthode de préparation et de la complétude de la couverture du domaine. Les sous étapes de préparation sont : la standardisation, la suppression des mots vides et la lemmatisation. Ensuite, La deuxième phase consiste à extraire les termes simples et complexes. Et pour ce faire, nous avons choisi d'utiliser une méthode statistique, qui s'appel méthode des "segments répétés", en recueillant des mots et des phrases fréquemment répétées et en filtrant ceux qui représentent vraiment des concepts de notre domaine, en utilisant deux filtres : le filtre TF-IDF (Term Frequency-Inverse Document Frequency) et le filtre coupant. La dernière phase consiste à utiliser une méthode d'apprentissage de marqueurs linguistiques par l'application d'un algorithme, pour relier les concepts extraits dans la deuxième étape avec des relations sémantiques de type : généralisation-spécialisation, synonymie et antonymie.

Cette dernière phase se base sur le corpus de texte brute et un ensemble de listes préliminaires de paires (pour chaque type de relation) proposés par un expert du domaine, afin d'extraire les paires de relations existantes dans le corpus de texte. Ces paires de relations sont utilisées comme un noyau pour apprendre d'autres paires à partir du corpus de texte en utilisant deux ressources externes : un dictionnaire de synonymes et d'antonymes et une base de données lexicale arabe.

Et pour évaluer notre travail, nous avons présenté, d'abord, les résultats de l'application de notre système sur le corpus de textes pour chaque phase, ensuite, nous avons décrit la méthode d'évaluation de notre approche avec une discussion de cette évaluation pour chaque étape. Finalement, nous pouvons dire que les résultats des évaluations en termes de précision et de rappel de notre approche d'extraction de termes et de relations sémantiques pour la construction d'une ontologie peuvent être jugées comme satisfaisants.

Par ailleurs, comme perspectives ou améliorations à ce qui a été déjà accompli et comme travail futur, nous prévoyons d'apporter quelques modification dans les étapes de construction de l'ontologie, par exemple, dans la phase de filtrage des termes par le filtre coupant, au lieu de travailler avec une liste préliminaire des entités nommées qui sont entrées manuellement, nous pouvons automatiser cette opération par l'extraction automatique de ces entités à partir du texte. Aussi dans la phase de l'extraction des relations sémantiques, nous pouvons apprendre ces relations à partir du texte par l'une des techniques d'apprentissages artificiels comme les réseaux de neurones, les arbres de décision, la technique SVM, etc. Ces dernières techniques peuvent être utilisées pour extraire les relations sémantiques ou même les marqueurs linguistiques à partir d'une base d'exemples extraits à partir du corpus de texte ou à partir des ressources externes tels qu'un thésaurus ou un dictionnaire.

Et comme travail futur aussi, notre approche peut être appliquée à d'autre corpus de texte d'un domaine différent de notre domaine, mais toujours dans la langue arabe, comme par exemple le domaine de la linguistique et la littérature arabe, mais ça nécessite la présence d'un expert spécialiste et des ressources externes dans la spécialité du corpus traité. Notre approche peut être appliquée aussi à une langue différente de la langue arabe pour voir quels résultats vont être obtenus et obtenir une autre évaluation de cette approche.

Nous espérons finalement par ce travail réalisé, apporter une contribution significative au projet de construction d'une ontologie pour la langue arabe.

Références bibliographiques

Références

Abdelali A., **Cowie J.R., Farwell D., Ogden W.C.,** (2004). *UCLIR: a Multilingual Information Retrieval Tool*. Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial 8(22): 103-110.

Aljlayl M, Frieder O (2002), « On Arabic search: Improving the retrieval effectiveness via a light stemming approach». In: Proceedings of the eleventh international conference on information and knowledge management, ACM Press, New York, NY, USA, pp 340-347, ACM DL Digital Library, http://dl.acm.org/citation.cfm?id=584848. *Visité 1 mai 2016*

Amari S., (2009). « *Extraction d'information à partir des textes arabes à l'aide de l'outil Gate »,* mémoire de master, Université d'Annaba.

Assadi H., Bourigault D., (1996). « Ac*quisition de connaissances à partir de textes: Outils informatiques et éléments méthodologiques »*. Actes du dixième congrès Reconnaissance de Formes et Intelligence Artificielle (RFIA' 96), pp 505-514. Rennes.

Bachimont B. (2000). *Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances*. In R. Teulier, J. Charlet & P. Tchounikine, Coordinateurs, Ingénierie des connaissances, chapitre 19. Paris

Baneyx A., & Charlet, J., (2006). *Évaluation, évolution et maintenance d'une ontologie en médecine: état des lieux et expérimentation*, Revue I3 ; SI 2006 special issue on Ontological ressources.

Baneyx A., (2007). *Construire une ontologie de la pneumologie, aspects théorique, modèles et expérimentations*. Thèse de doctorat, Université Pierre et Marie Curie.

Béchet N., (2009). « *Extraction et regroupement de descripteurs morpho-syntaxiques pour des processus de Fouille de Textes »*, Thèse de doctorat, Université de MontpellierII.

Beguin A., Jouis C., Widad M., (1997). « *Evaluation d'outils d'aide à la construction de terminologie et de relations sémantiques entre termes à partir de corpus »*. Premières Journées Scientifiques et Techniques (JST) du réseau Francophone de l'ingénierie de langue de l'AUPELF-UREF, pp 419-425. Avignon.

Benabdallah A, (2016) *« Application de la méthode des patrons lexico-syntaxiques pour la Construction semi-automatique des ontologies à partir des textes arabes »*, 7ème Colloque International en Traductologie et TAL (TRADETAL 2016), Oran, Algérie.

Benabdallah A, Abderrahim M.A, Abderrahim M.E (2017) *« Extraction of terms and semantic relationships from Arabic texts for Automatic Construction of an Ontology »*. In Int J Speech Technol (2017). doi:10.1007/s10772-017-9405-5.

Benaissa B, (2012). *Construction semi-automatique d'ontologies à partir de textes arabes* Mémoire de magister en « informatique », Université de Tlemcen – Algérie

Berland M., Charniak E. (1999), *"Finding parts in very large corpora"*. In Annual meeting of Association of Computational Linguisitcs.

Bernhard D. (2006), « *Apprentissage de connaissances morphologiques pour l'acquisition automatique de ressources lexicales »*, Thèse de doctorat, Université Joseph Fourier – Grenoble.

Biebow B., Szulman S., (2000). *Une approche terminologique pour la construction d'ontologie de domaine à paritr de textes : TERMINAE*. Actes du douzième congrès Reconnaissance de Formes et Intelligence Artificielle (RFIA' 2000), pp 81-90. Paris.

Black W.J., Elkateb S, Fellbaum C, Alkhalifa M, Pease A, Rodríguez H, Vossen P, (2006) *« Introducing the Arabic WordNet project Proceedings of the 3rd Global Wordnet Conference »*, Jeju Island, Korea, January, 2006.

Boulaknadel S. (2008). « *Traitement Automatique des Langues et Recherche d'Information en langue arabe dans un domaine de spécialité : Apport des connaissances morphologiques et syntaxiques pour l'indexation »* , Thèse de doctorat, Université de Nantes.

Bourigault D., Aussenac-Gilles N. (2003) « Construction d'ontologies à partir de textes ». TALN 2003, Batz-sur-Mer, 11-14 juin 2003. http://www.atala.org/taln_archives/TALN/TALN-2003/taln-2003-tutoriel-002.pdf. visité le15 Decembre 2015.

Brown P., Cocke J., Della Pietra S., Della Pietra V., Jelinek F., Mercer R., Roossin P. (1988). *"A statistical approach to language translation"*. In: Proceedings of the 12th conference on Computational linguistics . Budapest, Hungry .

Brown P., Cocke J., Della Pietra S., Della Pietra V., Jelinek F., Mercer R., Roossin P. (1990) *"A statistical calcul approach to machine translation"*. Computational linguistics .volume 16, n°2 .

Cartier E. (1997), *« La définition dans les textes scientifiques et techniques : présentation d'un outil d'extraction automatique de relations définitoires »*. Actes des deuxièmes rencontres Terminologie et Intelligence Artificielle (TIA'97), pp 127-140. Toulouse.

Charlet J., Bachimont B., Bouaud J., Zweigenbaum P., (1996). *Ontologie et réutilisabilité :expérience et discussion*. In N. Aussenac-Gilles, P. Laublet & C. Reynaud,

Coordinateurs, Acquisition et ingénierie des connaissances : tendances actuelles, chapitre 4, p. 69–87. Cepaduès-éditions.

Church .K. W., Hanks P. (1990). *"Word association norms, mutual information and lexicography"*. Computational Linguistic, vol 1, Mars 1990, pp: 22-29.

Cimiano P. & Volker J. (2005). «Text2onto - a framework for ontology learning and data-driven change discovery». *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)*, volume 3513 of Lecture Notes in Computer Science, p.227–238, Alicante, Spain : Springer.

Claveau V. (2003). « Acquisition automatique de lexiques sémantiques pour la recherche d'information », Thèse de doctorat, Université de Rennes 1.

Claveau V, Sébillot P (2004) *« Apprentissage semi-supervisé de patrons d'extraction de couples nom-verbe »*, TAL (Traitement Automatique des Langues). Volume 45 – n° 1/2004, pp 153-182. http://people.irisa.fr/Vincent.Claveau/publications.html#2004. Visité le 15 mai 2016

Condamines A., Rebeyrolle J., (1997). *Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode*. Actes des journées d'Ingénierie des Connaissances et Apprentissage Automatique (JICAA' 97), pp 191-206. Roscoff.

Condamines A, Rebeyrolles J. (2000), *« Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode »*. In CHARLET J, ZACKLAD M., KASSEL G. & BOURIGAULT D. éds. Ingénierie des connaissances .

Cunningham H., Maynard D., Bontcheva K., and Tablan V., (2002), « GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications ». In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).

Daille B., (1994), *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. Thèse de doctorat, Université de Paris 7.

Daille B. (1996), *"Study and implementation of combined techniques for automatic extraction of terminology"*. In J. KLAVANS & P. RESNICK, Eds., The Balancing Act :Combining Symbolic and Statistical Approaches to Language, p. 49–66. MIT Press .

Daille B. (2002), « *Découvertes linguistiques en corpus* », Mémoire d'Habilitation à Diriger des Recherches en Informatique, Université de Nantes.

Daoust F., (1992). *SATO (Système d'Analyse de Textes par Ordinateur) version 3.6 : Manuel de Référence*. Centre ATO Université du Québec à Montréal.

David S., Plante P. (1990), « *De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes* ». Intelligence Artificielle et Sciences Cognitives au Québec, 3(3), 140–154.

Dias G., (2002), « *Extraction automatique d'associations lexicales à partir de corpora* ». Thèse de Doctorat, Université d'Orléans.

Drouin P., (2002). « Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés », Thèse de doctorat, Université de Montréal.

Dubois J., Guespin L., Giacomo M., Marcellesi C., **MÉVEL J.**, (1994), « Dictionnaire de linguistique et des sciences du langage ». Collection Trésors du Français, Larousse. Paris. 1994.

Dunning T. (1993). *"Accurate Methods for the Statistics of Surprise and Coincidence"*, Computational Linguistics, vol. 19, n°1, pp: 71-74, Mars 1993.

Enguehard C., (1993). *ANA, Apprentissage Naturel Automatique d'un réseau sémantique*. Thèse de doctorat. Université de Compiègne.

Fernandez M., Gomez-Pérez A., Juristo N., (1997). *Methontology: from ontological art towards ontological engineering*. In Spring Symposium Series on Ontological Engineering, National Conference of the American Association on Artificial Intelligence (AAAI).

Fortuna B., Grobelnik M. & Mladenic D. (2006). Semi-automatic data driven ontology construction system. In Proceedings of the 9th International multiconference Information Society IS-2006, Ljubljana, Slovenia.

Fotzo H. N., Gallinari P. (2004), *"Information access via topic hierarchies and thematic annotations from document collections"*. In International Conference on Enterprise Information Systems, pages 69-76.

Frantzi K. T., Ananiadou S. (1997). *« Automatic Term Recognition Using Contextual Cues », Proceedings of the 3rd DELOS Workshop, Zurich, 8 p.*

Frantzi K. T., Ananiadou S., Tsujii J. (1999). *« Classifying Technical Terms », Proceedings Third ICCC/IFIP Conference on Electronic Publishing, Ronneby, p. 144-155.*

Gale W., Church K. (1991) *"A program for aligning sentences in bilingual corpora"* , Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, 1991, Berkley, California, p. 177-184 .

Gandon F. (2002). *Ontology Engineering : a Survey and a Return on Experience*. Rapport interne 4396, INRIA. 181 p., ISSN 0249-6399

Gangemi, A., Guarino, N., Masolo, C., and Oltramari, A. (2003). Sweetening wordnet with dolce. AI Magazine, 24(3):13–24.

Garcia D., (1998). *Analyse automatique des textes pour l'organisation causale des actions, Réalisation du système informatique COATIS*. Thèse de doctorat, Université de Paris -Sorbonne.

Girju R., Moldovan D. (2002), « *Text mining for causal relations* ». In 15sup th international Florida Artificial Intelligence Research Society Conference, pp: 360-364.

Girju R., Badulescu A., Moldovan D. (2003), *"Learning semantic constraints for the automatic discovery of part-whole relations"*. In Human Language Technologies and

Gomez-Perez A, (1999), "Ontological Engineering: a State of the Art", Expert Update. British Computer Society. Vol. 2. nº 3. pp. 33 – 43.

Gomez-Pérez A. (2004). *Ontology Evaluation,* In S. Staab & R. Studer, Coordinateurs, Handbook on Ontologies, chapitre, p. 251–275. Handbooks in Information Systems. Springer.

Goujon B. (1999), « *Extraction d'informations techniques pour la veille par exploration de notions indépendantes d'un domaine* ». Terminologies nouvelles n° 19. pp 33-42.

Gruber T. (1993). A *translation approch to portable ontology specifications*. Knowledge acquisition, 5(2), 199–220.

Harrathi F., (2009). *Extraction de concepts et de relations entre concepts à partir des documents multilingues : approche statistique et ontologie*, thèse de doctorat, Institut national des sciences appliquées de Lyon.

Hearst M., (1992). *Automatic Acquisition of Hyponyms from Large Text Corpora,* In Proceedings of the 13th international Conference On Computational Linguistics (COLING), pp 539-545. Nantes.

Heitz T., (2008). *Une méthode pour le prétraitement des textes : dépendances entre traitements et leur intelligibilité,* Thèse de doctorat, Université Paris-Sud 11.

Hernandez N., Mothe J., (2006) « TtoO: une méthodologie de construction d'ontologie de domaine à partir d'un thésaurus et d'un corpus de référence » IRIT, Toulouse.

Hernandez N., (2006) « Ontologies de domaine pour la modélisation du contexte en recherche d'information » *Thèse de Doctorat à l'Université Paul Sabatier* France.

Ian Niles, A. P. (2001). Towards a standard upper ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS'01), pages 2–9.

Isaac, A., (2005). Conception et utilisation d'ontologies pour l'indexation de documents audiovisuels, Thèse de doctorat, Université Paris IV – Sorbonne.

Jacquemin C. (1996), *« A Symbolic and Surgical Acquisition of Terms Through Variation»*. In Wermter S., Riloff E., Scheler G. (eds.), Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing . Springer, Heidelberg, pp. 425–438.

Kashyap V., Ramakrishnan C., Thomas C., Bassu D., Rind-Esch T. C., Sheth A. (2004), *"Taxaminer: An experimental on framework for automated taxonomy bootstrapping"*. Technical report, University of Georgia.

Khaled W, Saad D (2012) Student's Dictionary of Synonyms and Opposites. Beirut, Lebanon: alrouqy −Verlag.

Klai S., Khadir M-T., (2009). *Data based Ontology Construction coupled to Expert System for Steam Turbine Aided Diagnostic*, Published in ewic journal: Références et bibliographie Electronic Workshops in Computing Series (eWiC: http://ewic.bcs.org, ISSN 1477-9358), The British Computer Society (BCS).

Koeva S., Maurel D., Silberztein M., (2007). *Formaliser les langues avec l'ordinateur : de Intex à Nooj* Presses Univ. Franche-Comté, 2007 - 438 pages.

L'Homme M.-C. (2001). *Nouvelles technologies et recherche terminologique. Techniques d'extraction des données terminologiques et leur impact sur le travail du terminographe.* In L'impact des nouvelles technologies sur la gestion terminologique, Toronto.

Lebart L., Salem A. (1988). « Analyse statistique des données textuelles : questions ouvertes et lexicométrie ». Paris: Dunod .

Le Priol F., Chevallet J -P., Brunadet M-F., Desclès J-P., (1998). *Intégration d'un système statistique (IOTA) et d'un système sémantique (SEEK) dans une chaîne de traitement permettant l'extraction de terminologies*. Actes Ingénierie des Connaissances (IC' 98), pp 33-40. Pont-à-Mousson.

Lin D. (1998), *"Dependency-based Evaluation of MINIPAR. Workshop on the Evaluation of Parsing Systems"*, Granada, Spain.

Lin D., Pantel P. (2001), *« Discovery of Inference Rules for Question Answering »*. Natural Language Engineering, 7(4), pp. 343–360.

Malaisé V., (2005). « Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles à partir de corpus textuels », Thèse de doctorat, Université Paris 7 – Denis Diderot France.

Mariano Fernandez Lopez, et al. (1999) "*Building a Chemical Ontology Using Methontology and the Ontology Design Environment"*. Polytechnic University of Madrid. IEEE intelligent systems.

Marshman E, (2003) Construction et gestion des corpus : Résumé et essai d'uniformisation du processus pour la terminologie, In: Observatoire de linguistique Sens-Texte (OLST). University of Montréal. http://olst.ling.umontreal.ca/pdf/terminotique/corpusentermino.pdf. Accessed 15 January 2016**.**

Mazari A.C., Aliane H., and Alimazighi Z., (2012) « Automatic construction of ontology from Arabic texts » ICWIT'2012 *« International Conference on Web and Information Technologies , Sidi Bel Abbes, Algeria »* ICWIT, volume 867 of CEUR Workshop Proceedings, page 193-202.

Mazari A.C, (2013) *« Vers une approche statistique pour l'extraction des éléments de l'ontologie à partir des textes arabes »*. In: RML (Revue Maghrébine des langues), ISSN: 2253-0673, 8th edition, Oran Algeria, pp 39-56.

Mhiri M, Gargouri F, Benslimane D*,* (2006). *Détermination automatique des relations sémantiques entre les concepts d'une ontologie,* In Proceedings of INFORSID'2006. pp.627~642

Mondary T., Després S., Nazarenko A., Szulman S. (2008) « Construction d'ontologies à partir de textes : la phase de conceptualisation » IC2008 : *« 19èmes Journées Francophones d'Ingénierie des Connaissances (IC 2008), Nancy : France »* LIPN - UMR 7030 Université Paris 13 – CNRS.

Morin E., Jacquemin C. (2004), *"Automatic Acquisition and Expansion of Hypernym Links".* Computers and the Humanities (CHUM), Kluwer, 38(4), p: 363–396 .

Morin E. (1999), *« Des patrons lexico-syntaxiques pour aider au dépouillement terminologiques »,* Traitement Automatique des Langues, volume 40, Numéro 1, pages 143-166 .

Morin E., (1999a). *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques.* Thèse de doctorat, Université de Nantes.

Oueslati R. (1999). *« Aide à l'acquisition de connaissances à partir de corpus ». Rapport interne, Université Louis Pasteur Strasbourg. Thèse de Doctorat en Informatique.*

Paroubek P., Rajman M. (2000). *« Etiquettage morphosyntaxique , danss Ingenierie des Langues , Collection Information Commande Communication »* , aux Editions Hermes Science ISBN 2-7462-0113-5, october 2000 pp 131-148.

Paumier S., (2009), « Unitex2.0, user manual », UniversitéParis-EstMarne-la-Vallée.

Patil L., Dutta D., Sriram R. (2005).*Ontology formalization of product semantics for product lifecycle management.* Proc. ASME/IDETC CIE Conf., Long Beach, CA

Perron, J. (1996), « ADEPTE-NOMINO : un outil de veille terminologique », dans Terminologies nouvelles, no 15, juin et décembre, Bruxelles, RINT, p. 32-47.

Pery-Woodley M.P. (1995), « *Quels corpus pour quels traitements automatiques ? »* Traitement Automatique de la Langue (TAL), volume 36, n° 1et 2 .p : 213-232.

Psyché V, Mendes O, Bourdeau J, (2003), "Apport de l'ingénierie ontologique aux environnements de formations à distance", revue sticef.org, Volume 10,2003.

Punuru J. R. (2008), « *Knowledge-Based Methods for Automatic Extraction of Domain-Specific Ontologies »*. Phd thesis, Louisiana State University, degree of Doctor of Philosophy.

Quinlan R. J. (1993), *«C4.5: Programs for Machine Learning»*. Morgan Kaufmann.

Rousselot F., Frath P. et Oueslati R. (1996), « *Extracting concepts and relations from corpora, Proceedings workshop on Corpus-Oriented Semantic Analysis »*, Proceddings of the 12th European Conference on Artificial Intelligence (ECAI'96).

Ryu P., Choi K. S. (2004), "*Measuring the specificity of terms for automatic hierarchy construction*". In European Conference on Artificial Intelligence Workshop on Ontology Learning and Population.

Seguela P., Aussenac-Gilles N. (1999), « *Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine »*, Actes de la conférence Ingénierie des Connaissances (IC'99), pp 79-88, Paris.

Séguéla P., (2001). « *Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques »*, Thèse de doctorat, Université de ToulouseIII.

Silberztein M. et Tutin A., (2004). « NooJ : un outil TAL de corpus pour l'enseignement des langues et de la linguistique Une application à l'étude des impersonnels », Université de Franche-Comté.

Smadja F. (1993). *Retrieving Collocations from Text: Xtrac*, Computational Linguistics 19(1), pp. 143-177.

Snow R, Jurafsky D., Andrew Y. (2004), « *Learning syntactic patterns for automatic hypernym discovery »*. In Advances in Neural information Processing Systems.

Sparck Jones K (1972) A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation, vol. 28, no 1, pp 11–21.

Talmy L. (1988), "*Force Dynamics in Language and Cognition"*. In Cognitive Science 12, pp 49-100.

Teguiak H. V. (2012) *Construction d'ontologies à partir de textes : une approche basée sur les transformations de modèles,* Thèse de DOCTORAT de L'ECOLE NATIONALE SUPERIEURE DE MECANIQUE ET D'AEROTECHNIQUE Université de Poitiers- France.

Turney P. D. (2006), "*Expressing implicit semantic relations without supervision*". In 21sup st international conference on computational linguistics, pages 313-320.

Uschold M. & Grüninger M. (1996). *Ontologies : Principles, methods and applications*. Knowledge Engineering Review, 11(2).

Velardi P., Missikof M., Fabriani P. (2001). *Using text processing techniques to automatically enrich a domain ontology*. In Proceeding of ACM-FOIS.

Vergne J, (2004) *«Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource»*, 7eme Journées internationales d'analyse statistique des données textuelles, GREYC – University of Caen.

Voutilainen A. (1993). *Nptool, a detector of English noun phrases*, In Proceedings of the Workshop on Very Large Corpora, June, Columbus, Ohio State University, p.48-57.

Wang Y., Volker J. & Haase P. (2006). Towards semi-automatic ontology building supported by large-scale knowledge acquisition. In AAAI Fall Symposium On Semantic Web for Collaborative Knowledge Acquisition, volume FS-06-06, p. 70–77, Arlington, VA, USA : AAAI AAAI Press.

Welty C., & Guarino, N., (2001). *Supporting Ontological Analysis of Taxonomic Relationships* Data and Knowledge Engineering (39), pages 51-74, 2001.

Zaidi, S., Laskri (2009). « Review of textual terminology tools for ontologies building », In proceedings, MIC'09 Management International conference, 25- 28 november, 2009 Sousse Tunisia.

Zaidi, S., Laskri, M-T, Abdelali, A. (2010). *Arabic collocations extraction based on JAPE rules,* In Proceedings, Acit Arab Conference of Information and Technology, Benghazi, Libya.

Zaidi–Ayad S. (2013) *Une plateforme pour la construction d'ontologie en arabe : Extraction des termes et des relations à partir de textes (Application sur le Saint Coran),* Thèse de DOCTORAT en Informatique, Université de Annaba – Algérie

Zipf. G. K., (1949). *Human Behavior and the Principle of Least Effort,* New York, Harper, réédition 1966.