وزارة التعليم العالي والبحث العلمي

BADJI MOKHTAR –ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR ANNABA

جامعة باجي مختار- عنابة -

FACULTÉ DES SCIENCES DE L'INGÉNIEUR
DÉPARTEMENT D'INFORMATIQUE
ANNEE UNIVERSITAIRE 2005

كلية علوم المهندس
قسم الإعلام الآلي
السنة الجامعية 2005

MEMOIRE

Présenté en vue de l'obtention du diplôme de **MAGISTER**

THEME

Vers une automatisation de la construction des ontologies

Option

Intelligence Artificielle Distribuée (IAD)

Par :

Mr Allalga A. Wahid

DIRECTEUR DE MEMOIRE

Professeur SELLAMI Mokhtar

DEVANT LE JURY

PRESIDENT :	Mr T. Bensbaa	M. C	Univ. Annaba
RAPPORTEUR	Mr M. Sellami	Pr	Univ. Annaba
EXAMINATEUR :	Mr T. Kimour	M. C	Univ. Annaba
EXAMINATEUR :	Mr T. Khadir	M. C	Univ. Annaba
EXAMINATEUR :	Mme H. Merouani	M. C	Univ. Annaba

À ma mère, mon père, mes sœurs et frères, ainsi que à tous mes amis

REMERCIEMENTS

Après deux années de travail, j'ai beaucoup de personnes à remercier pour leur participation directe ou indirecte à mes travaux.

Je tiens en premier lieu à remercier le professeur MOKHTAR SELLAMI qui a encadré mon travail de recherche durant ces deux années. Sa patience, ses encouragements, son écoute et sa confiance ont été d'un réconfort et d'une aide précieuse. Qu'il trouve dans ces quelques mots l'expression de ma profonde gratitude.

Je tiens à vivement remercier tous les membres du jury :

M[r] TAHAR BENSBAA, maître de conférence à l'université Badji-Mokhtar Annaba, pour m'avoir fait l'honneur de présider ce jury.

M[r] T. Kimour, maître de conférence à l'université Badji-Mokhtar Annaba, de faire partie de mon jury.

M[r] T. Khadir, maître de conférence à l'université Badji-Mokhtar Annaba, pour avoir accepté d'être membre de ce jury.

M[me] H. Merouani, maître de conférence à l'université Badji-Mokhtar Annaba, pour son aimable participation au jury.

Je suis reconnaissant à M[me] M. Babes pour m'avoir aidé à surmonter toutes les difficultés que j'ai rencontrées.

Merci à Redouane et Mourade, nous nous sommes serrés les coudes quand la pente était trop raide. Un merci à Zahra et Akila qui m'ont soutenu comme seuls savent le faire les amis.

Il me faut également remercier tout le reste des collègues de la post-graduation. Impossible de tous les nommer, et j'ai épuisé ma collection de synonymes pour dire merci. Que chacun d'entre eux sache néanmoins que me lever le matin pour aller à la salle PG a toujours été un plaisir, sachant l'ambiance tellement sympathique que j'allais y trouver.

Je tiens aussi à remercier tous ceux, qui ont su ramener si souvent cette question si embêtante « Alors, la soutenance c'est pour quand ? », j'ai enfin la réponse à votre

question, que tous ceux et celles qui ont fait des promesses de célébration les tiennent maintenant !

Merci enfin à tous ceux qui ont fait ce que je suis actuellement : Parents, famille, amis, ma petite amie, professeurs, etc.

ملخص:

رغم الإجماع حول أهمية الأنطولوجيات, كمفهمة مشتركة في مجال معين, وهذا لتطوير برامج الكمبيوتر حيث هناك حاجة لاستخدام المعرفة, إلا إن استعمالهم مازال جد محدود بسبب أن تطويرهم يتم غالبا بطريقة يدوية ذات تكلفة عالية في الوقت و الجهد. انطلاقا من قناعتنا بأن بناء الأنطولوجيات لا يمكن أن يتم بطريقة أوطوماطكية محض, نقترح في هذه الرسالة, طريقة نصف آلية لاستخراج الأنطولوجيات من النصوص و من الأنطولوجيات التي سبق تطويرها. الطريقة التي نقترحها تتميز بكونها تتم بصفة تصاعدية مما يساعد المستخدم أثناء عملية صيانة الأنطولوجية. من ناحية أخرى سنحاول تحديد طريقة تسمح باستخراج أنطولوجيات مزدوجة اللغة. هذا النوع من الأنطولوجيات يعتبر ذا اهمية كبرى بالنسبة لتطبيقات مثل الترجمة الآلية و البحث عن معلومات متعددة اللغة.

الكلمات المفتاحية:

استخراج المعرفة, الأنطولوجية, نصوص, استخراج الألفاظ, استخراج العلاقات, الأنطولوجيات مزدوجة اللغة

Résumé :

Malgré le consensus établi sur l'importance des ontologies, vues comme une conceptualisation partagée d'un domaine, pour le développement et l'interopérabilité entre applications informatiques où il y a une nécessité de manipuler explicitement la connaissance, leur utilisation reste très restreinte, ceci est dû au fait que leur élaboration, qui se fait le plus souvent manuellement, s'avère être une tâche laborieuse, coûteuse et se relevant plus du savoir-faire que de l'ingénierie. Etant convaincu que ce travail de développement ne peut être réalisé d'une manière totalement automatique, nous proposons, dans ce mémoire, les prémisses d'une méthodologie et d'une architecture générale d'un système susceptible d'assister le concepteur tout au long de la construction des différents éléments composants l'ontologie. Notre approche se démarque par le fait que le développement se fait d'une façon incrémentale, ce qui constitue de notre point de vue une aide précieuse pour l'utilisateur dans sa tâche de maintenance de l'ontologie. Dans un autre cadre de travail, nous proposons quelques principes à suivre pour développer des ontologies bilingues, ce type d'ontologies devient une clé de voûte pour la réussite de systèmes de traduction automatique ou de recherche d'informations multilingues.

Mots-clés : acquisition de connaissances, ontologies, corpus textuel, extraction de termes, extraction de relations, ontologies bilingues.

Abstract :

despite the established consensus over the importance of ontologies, considered as a shared conceptualization of a given domain, to the development and interoperability of applications where there is a need to explicitly deal with knowledge, their exploitation is still very limited, this is due to the fact that their elaboration, which is often accomplished manually, seems to be time-consuming, expensive, and demanding a lot of know-how rather than engineering. Being convinced that this work of development cannot be realized automatically, we present, in this thesis, the foundations of a methodology and an overall architecture of system developed to assist the designer throughout the construction of the various elements making up the ontology. Our system has the characteristic that the development process is done in an incremental manner, which will help the user during the maintaining of the ontology. The acquisition of bilingual ontologies will be discussed in the last chapter, this kind of ontologies is very important to systems like machine translation and multilingual information retrieval.

Keywords : Knowledge acquisition, ontologies, textual corpora, term extraction, relation extraction, bilingual ontologies.

TABLE DES MATIERES

TABLE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

« Le commencement de toutes les sciences,
c'est l'étonnement de ce que les choses sont ce qu'elles sont »
ARISTOTE

Introduites dans une optique de représentation de connaissances, les ontologies sont à l'heure actuelle au cœur des travaux menés en Ingénierie des Connaissances. Visant à établir des représentations à travers lesquelles les machines puissent manipuler la sémantique des informations, la construction des ontologies demande à la fois une étude des connaissances humaines et la définition de langages de représentation, ainsi que la réalisation de systèmes pour les manipuler.

Selon Gruber [Gruber, 1993], « une ontologie est une spécification explicite d'une conceptualisation ». Cette définition de la notion d'ontologie laisse déjà entendre le rôle qu'auront les ontologies comme moyen de représentation de connaissances. Cet outil, autrefois réservé aux cercles restreints de l'intelligence artificielle, pénètre de plus en plus dans les diverses sphères d'activité, comme le Web sémantique [Bernard-lee, 2001], la gestion de connaissances, de documents, les systèmes d'informations, etc. De part cette importance, le nombre de ressources ontologiques construites, par exemple [Miller, 1990], [Lindberg et al., 1993], [Sowa, 2004], connaît un développement exponentiel.

Initialement, cette construction se faisait manuellement, mais le coût et le temps consacrés à ce développement ne permettaient pas de répondre aux demandes, et ce malgré la présence d'environnements dédiés à l'édition d'ontologies [Sure et al., 2002], [Arpirez et al., 2001]. De plus, on a constaté que ce travail de construction qui relève de l'acquisition de connaissances se heurtait à son tour au problème du goulot d'étranglement [Buchanan & Wilkins, 1993]. Face à cet engouement et pour surmonter les difficultés intrinsèques liés à cette tâche d'acquisition, le besoin de méthodes et outils destinés à la conception (semi) automatique d'ontologies devient pressant.

Le texte est la première source d'acquisition et de plus en plus de travaux font usages de corpus textuel comme source principale d'apprentissage; il s'agit en particulier des travaux de [Aguirre et al., 2000], [Alfonseca & Manandhar, 2002a ; 200b], [Aussenac-Gilles et al.,

2000a ; 2000b], [Bachimont et al., 2002], [Gupta et al., 2002], [Hahn & Markó, 2001], [Hearst, 1998], [Hwang,1999], [Nobécourt, 2000], [Assadi & Bourigault, 1996], [Roux et al., 2000], [Taleb & Sellami, 2003], etc. Ces méthodes sont principalement fondées sur des techniques d'analyse du langage naturel.

Cependant, l'expansion du Web et la quantité énorme des ontologies produites ne peuvent être désormais écartés du processus de construction. Ainsi, plusieurs travaux se sont intéressés à l'exploitation de ces nouvelles ressources pour développer et alimenter les ontologies [Maedche & Staab, 2001], [Velardi et al. 2002], [Faatz & Steinmetz, 2002], [Kietz *et al.*, 2000]. Dans cet esprit, nous proposons, dans ce mémoire, une méthodologie, ainsi que l'architecture générale et les fondements théoriques d'un système que nous développons pour assister l'utilisateur dans la construction des ontologies de domaines, et ce en jumelant l'utilisation de corpus textuels et l'exploitation d'ontologie existantes. Cette tâche est accomplie en profitant de la structure modulaire de l'environnement d'édition d'ontologies Protégé2000 [Noy et al., 2000], qui sera utilisé comme plateforme dans laquelle viendra s'ancrer l'outil OntoLogos [Allalga & Sellami, 2005] que nous développons sous la forme d'un plugin. OntoLogs se veut un outil destiné à la fois au développement et à la maintenance d'ontologies, ceci est fait grâce à la méthodologie sous-jacente à son développement.

Plan du mémoire

Ce mémoire est organisé en deux parties. La première partie, sera consacrée à un état de l'art sur les ontologies et les techniques d'acquisition de terminologie à partir de corpus textuel, cet partie se compose des chapitre 1 et 2

Le chapitre 1, commence par définir la notion d'ontologie. Nous présentons ensuite les éléments composants une ontologie, les dimensions de classification, et les principales méthodologies de construction existantes. Nous verrons aussi les différents environnements d'édition d'ontologie disponibles. Divers points liés à la manipulation des ontologies seront à leur tour abordés.

Le deuxième chapitre présente les différentes techniques d'acquisition de terminologie à partir de textes. Cette tâche comporte une première phase d'acquisition de termes suivie d'une phase d'acquisition des relations qu'ils maintiennent entre eux, nous présentons alors une comparaison entre les outils et méthodes développés pour accomplir ces deux tâches.

La deuxième partie présente notre contribution aux problèmes posés dans ce mémoire, à savoir l'automatisation de la construction des ontologies. Cette partie est faite des chapitres 3 et 4.

Dans le chapitre 3, nous exposons la méthodologie de construction que nous avons développée et l'architecture générale d'un système basé sur cette méthodologie. Nous abordons en détail les différentes étapes du processus de construction.

Le chapitre 4 vient pour poser les fondements d'une acquisition ontologique bilingues, nous y abordons les sources et approches d'acquisition permettant de construire des ontologies bilingues. Dans cette optique, notre contribution va se limiter à définir les principales orientations et tendances dans ce domaine.

Nous concluons ce mémoire en rappelons le problème posé par cette étude et la solution que nous avons apporté. La conclusion est surtout l'occasion de présenter les perspectives d'application et de poursuite de la recherche entreprise.

PARTIE I

ETAT DE L'ART

CHAPITRE 1

L'INGENIERIE ONTOLOGIQUE

« La connaissance s'accroît en la partageant »

Vues comme moyen de représentation de connaissances, les ontologies ont très vite prouvé leur utilité dans diverses applications où elles été destinées à établir un vocabulaire commun et partagé au sein d'une communauté particulière. Leur prolifération rapide a donné naissance à une nouvelle discipline de l'ingénierie de connaissances, connue sous l'appellation d'ingénierie ontologique, qui a pour objectif de fournir des méthodes et outils nécessaires pour faciliter leur construction. Ce chapitre a pour but de présenter sans exhaustivité l'état de l'art en ingénierie ontologique tout en mettant en lumière certaines des principales difficultés rencontrées dans cette nouvelle discipline.

Dans la première partie, nous commençons par définir la notion d'ontologie. Les différents éléments dont elles sont constituées sont ensuite décrits. Enfin, les dimensions de leur classification seront explicitées. La deuxième partie du chapitre est consacrée aux différents aspects de l'ingénierie ontologique, à travers de laquelle nous découvriront les méthodologies développées pour construire, évaluer et maintenir les ontologies. Nous présentons ensuite quelques formalismes de représentation des connaissances qui sont à l'origine des langages permettant de les exprimer. Les principaux langages et outils liés aux ontologies sont alors présentés. Finalement, certains points supplémentaires liés à l'ingénierie ontologique seront discutés.

1. PERSPECTIVE HISTORIQUE SUR L'ONTOLOGIE

Historiquement, l'ingénierie ontologique a émergé de l'ingénierie des connaissances. Cette dernière a longtemps été considérée comme le domaine de prédilection du développement d'expertise en conception de systèmes à base de connaissances. Malgré le fait que l'ingénierie des connaissances ait contribué à accroître cette expertise en l'organisant dans une perspective computationnelle, certains membres de la communauté de l'intelligence artificielle ont éprouvé le besoin de passer à une ingénierie s'appuyant plus solidement sur des fondements théoriques et méthodologiques, afin d'améliorer la conception des systèmes intelligents : l'ingénierie ontologique (IO) permet de spécifier la conceptualisation d'un système, c'est-à-dire, de lui fournir une représentation formelle des connaissances qu'il doit acquérir, sous la forme de connaissances déclaratives exploitables par un agent. Ainsi, l'exploitation par un mécanisme d'inférence, d'une représentation de type déclarative telle que l'ontologie, tout en suivant les règles d'inférence définie dans cette ontologie, est la source de l'intelligence de système.

L'ingénierie de connaissances a ainsi donné naissance à l'ingénierie ontologique, où l'ontologie est l'objet clé sur lequel il faut se pencher. La nécessité d'une ontologie et d'une ingénierie ontologique des systèmes à base de connaissances commence à être comprise et acceptée par la communauté. Fonder l'ingénierie ontologique exige que l'on puisse en définir l'objet et en défendre la spécificité méthodologique. Or, si personne ne conteste que l'objet de l'ingénierie ontologique soit l'ontologie, la définition explicite et la délimitation précise de ce concept soulève un questionnement qui est tout à la fois d'ordre philosophique, épistémologique, cognitif et technique. La prochaine section introduit la notion d'ontologie ainsi que sa genèse dans l'histoire de la philosophie occidentale.

1.1. Origine de l'ontologie

L'Ontologie est un terme philosophique qui signifie être - du grec ancien ôn, onton, participe présent de einai - et discours, étude, science - de logos – [EncyclopædiaUniversalis, 00]. En d'autres termes, l'Ontologie serait la Science ou théorie de l'être. Bien que ce soient les Grecs qui aient inventé cette science, ils ne l'avaient pas appelé Ontologie, le terme étant beaucoup plus récent (XVIIe siècle) que la discipline qu'il désigne [EncyclopædiaUniversalis, 00]. La discipline elle-même a évolué en se rapprochant des sciences cognitives et de l'IA, il y a seulement une vingtaine d'années.

Dans les écrits scientifiques contemporains, le terme ontologie recouvre deux usages dont le premier appartient à la philosophie classique et le second plus récent, aux autres sciences cognitives. De ce fait, la convention veut que la notation Ontologie (avec un O majuscule) soit attribuée au domaine issu de la philosophie et ontologie aux autres.

Pris dans son sens le plus large, le terme ontologie est plus ou moins synonyme de : théorie ou conception du réel. Dans cette acception, très large, la recherche ontologique n'est nullement quelque chose dont la philosophie aurait le monopole, comme nous le verrons par la suite dans l'historique de l'ontologie en intelligence artificielle.

Dans la section suivante, nous nous intéressons à son sens philosophique, le plus étroit et le plus théorique, où l'Ontologie est définie comme la théorie de l'être en tant qu'être.

1.2. Fondement métaphysique : La science de l'être

Dans la philosophie classique, l'Ontologie correspond à ce qu'Aristote appelait la Philosophie première, protè philosopha, c'est-à-dire la science de l'être en tant qu'être, par opposition aux philosophies secondes qui s'intéressaient, elles, à l'étude des manifestations de l'être (les étants) [Graf, 96]. D'après, le constat fondamental d'Aristote, influencé par Parménide, l'étant se dit de multiples façons.

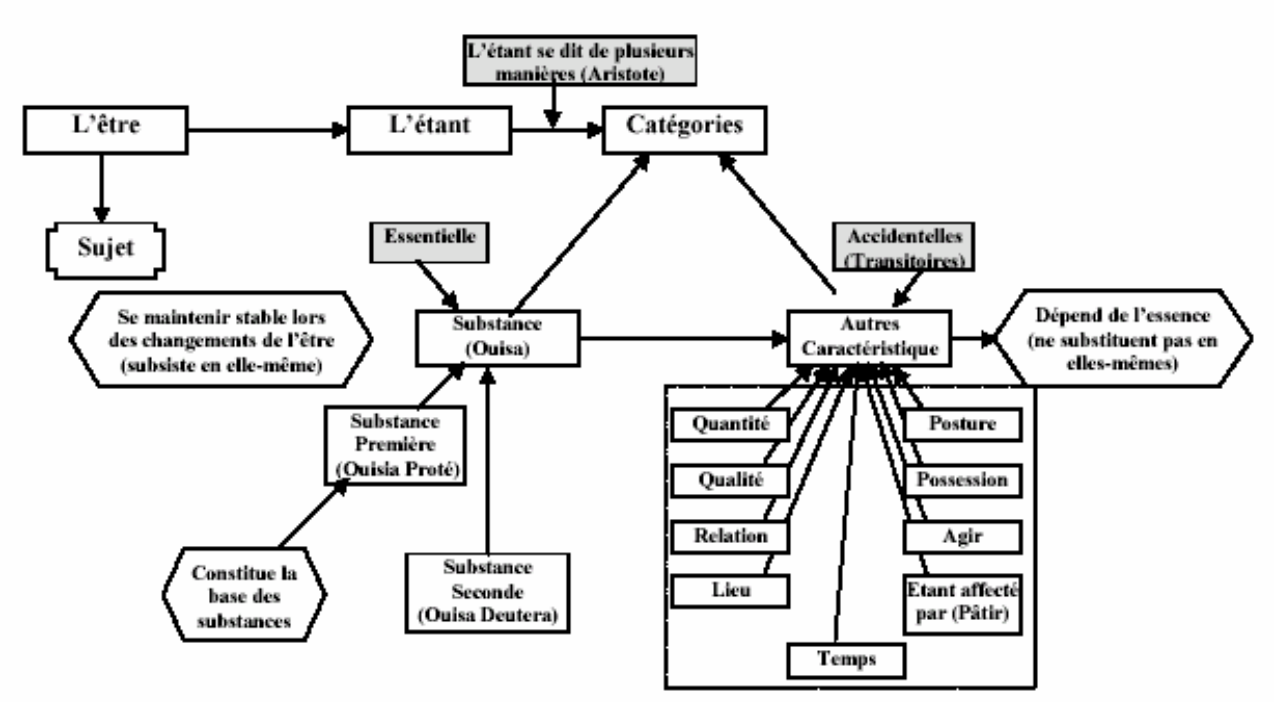


Figure 1 : Ontologie de l'être selon Aristote

Les catégories constituent les différentes descriptions associées aux manifestations de l'être dans le monde, traduites par des propositions. Aristote a ramené l'ensemble des formes possibles de manifestations de l'être à dix catégories : 1) Substance ; 2) Quantité ; 3) Qualité (quels attributs); 4) Relation (plus x que, etc.) ; 5) Lieu (où) ; 6) Temps (quand) ; 7) Posture (positionné comment) ; 8) Possession (avec quoi) ; 9) Action (en faisant quoi) ; 10) Pâtir (subissant/étant affecté par). Parmi les dix catégories, la substance a une importance prépondérante car : 1) elle constitue l'essence sans quoi une entité ne peut pas subsister, et qui par le fait même, individualise et différencie une entité par rapport à toutes les autres ; 2) elle assure une structure qui demeure stable à travers les changements continuels du monde. Ainsi, il est possible de reconnaître un être (un certain individu par exemple), comme étant en essence le même, en dépit des changements qu'il subit dans toutes les autres propriétés (les catégories dites accidentelles ou transitoires) au fil du temps, modifiant par exemple son apparence/qualité (nourrisson, enfant, adulte, vieillard), sa posture, ses actions, ses possessions, etc. La Figure 1 (modélisée à l'aide du langage MOT1) résume cette ontologie de l'être selon Aristote.

1.3. Fondements épistémologiques

Le terme Ontologie aurait été introduit sous sa forme latine au XVIIe siècle par Goclenius (Rudolph Göckel, dans Lexicon philosophicum, 1613-1615) pour désigner la science de l'être en général. Il correspond par conséquent à cette recherche sur l'être en tant qu'être (on hè on, en grec ancien) qu'Aristote avait assignée, parmi d'autres objets propres, à la philosophie première, appelée par la suite, métaphysique [EncyclopædiaUniversalis, 00].

Par la suite, Johann Clauberg attribue la même signification au terme dans ses œuvres Metaphysica (1646) et Ontosophie sive ontologia (1656), où il l'emploie pour faire référence à une sorte de métaphysique générale qui aurait pour objet les caractéristiques essentielles communes à tous les êtres, à savoir : substance, existence, essence, etc. [Auroux, 84], [Graf, 96]. Selon Clauberg dans son œuvre Metaphysica : "Le nom seul est nouveau ; quant à cette science, elle existait déjà chez les scolastiques avec la même définition : on appelait Transcendentia ces déterminations communes à tous les êtres".

[1] MOT est un langage pour la modélisation informelle qui permet de décrire différents types de modèles (hiérarchie de classes, modèles procéduraux, théories, arbres de décision, structures de contrôle, processus et méthodes). Le langage permet d'intégrer dans un même modèle (sur plusieurs niveaux) des connaissances factuelles, procédurales et de contrôle (sous la forme de principes et de liens de régulation) [Paquette, 96].

La diffusion du terme est due à l'Ontologia de [Wolff, 1729], qui, dans le concept scolaire de métaphysique, rangeait l'Ontologie en tant que métaphysique générale, puisqu'elle traitait de l'être en général, et la distinguait des trois sciences métaphysiques spéciales (metaphysica specialis) que sont la psychologie rationnelle (l'être de l'âme intellective), la cosmologie rationnelle (l'être du monde) et la théologie rationnelle (l'être de Dieu), chacune traitant d'une région déterminée de l'être.

La différence entre la conception wolffienne de l'être et la conception classique dépend selon [Couturat, 1903], de ses prémisses leibniziennes, qui veulent que le possible précède le réel, si bien que l'être est défini comme, ce qui veut exister, soit qu'il existe effectivement, soit qu'il n'existe pas, l'existence apparaissant comme le complément de la possibilité.

Les principes suprêmes de l'Ontologie sont le principe de non-contradiction et le principe leibnizien de raison suffisante. Les déterminations internes de l'être sont ses attributs essentiels. Pour le reste, l'Ontologie étudie une série de couples conceptuels, comme quantité et qualité, nécessité et contingence, simplicité et composition, finitude et infinitude, identité et diversité, cause et effet, etc.

Kant a conçu son analytique transcendantale - première partie de la logique transcendantale, dans la Critique de la raison pure [Kant, 1781] - d'une telle manière qu'elle put prendre la place de la vieille Ontologie. Hegel (1770-1831) procéda de manière analogue avec la logique qu'il identifie à la métaphysique, lorsqu'il affirme dans l'un des textes introductifs à la Science de la logique : "la logique objective prend [...] la place de la métaphysique d'autrefois" [Hegel, 1812].

C'est dans le cadre du développement de la phénoménologie que le terme d'Ontologie a recommencé à investir le discours philosophique : d'abord Husserl, dont le projet de phénoménologie pure le conduit à parler d'ontologies régionales ou sciences idéales de genres d'être qui empiriquement sont l'objet de plusieurs sciences (par exemple, l'ontologie régionale de la nature physique, etc.), puis Heidegger et Hartmann. L'école existentialiste, avec Sartre, développera ensuite sa propre vision de l'Ontologie.

Dans la philosophie analytique, l'Ontologie a été étroitement liée à la logique et à la philosophie du langage. Selon Quine, les engagements ontologiques du discours (plus exactement d'une théorie scientifique) ne sont pas tant déterminés pas ses assertions d'existence que par le type de variables sur lesquelles le langage admet la quantification :

ainsi, une position nominaliste – pour qui il n'existe que des individus – admettra seulement la quantification sur des variables individuelles (et non pas, par exemple, sur des variables prédicatives).

L'Ontologie est donc déterminée par la sémantique de son langage, et coïncide de fait avec les aspects généraux de cette sémantique. Un courant significatif de la philosophie analytique poursuit la construction d'une ontologie formelle, c'est-à-dire d'une théorie formelle des modes d'être. La construction d'une telle théorie coïncide avec la définition d'une sémantique pour un langage logique, dans laquelle peuvent trouver place les types d'entités que la théorie admet (par exemple, des individus, ou bien des individus et des classes, ou bien des propriétés, etc.), et où sont définies les relations entre les différents types d'entités.

Une telle ontologie formelle implique de soumettre à une re-formulation dans le langage logique toutes les théories traditionnelles de l'être substantiel (idéalités mathématiques, réalités phénoménales des sciences naturelles, etc.). Cela constitue une réduction des ontologies des théories de la substance à l'ontologie fondamentale proposée.

La section suivante présente l'appropriation des ontologies par les chercheurs contemporains en IA.

2. VISION CONTEMPORAINE EN INTELLIGENCE ARTIFICIELLE

Dans les milieux de l'intelligence artificielle, il semblerait que l'ontologie ait été abordée pour la première fois par John McCarthy qui reconnut le recoupement entre le travail fait en Ontologie philosophique et l'activité de construire des théories logiques de systèmes d'intelligence artificielle. McCarthy affirmait déjà en 1980 que les concepteurs de systèmes intelligents fondés sur la logique devaient d'abord énumérer tout ce qui existe, en construisant une ontologie de notre monde.

Cette vision de McCarthy, inspirée par les lectures de Quine, fut reprise par Patrick Hayes, en 1985 dans son travail sur la Physique Naïve. La signification du terme a évolué, et pendant que les champs de l'ingénierie des connaissances, de la modélisation conceptuelle, et de la modélisation du domaine commençaient à converger, la signification du terme a fait de même.

Au début des années 1990, l'usage du terme était déjà bien répandu dans chacun des sous-domaines de l'intelligence artificielle. [Neeches, 91] et ses collègues, ont présenté leur vision

en ces termes : "An ontology defines the basic terms and relations to define extensions to the vocabulary".

En 1993, Gruber propose sa définition : "An ontology is an explicit specification of a conceptualization" [Gruber, 93], qui est jusqu'à présent la définition la plus citée dans les écrits en intelligence artificielle. Depuis la définition de Gruber, beaucoup de définitions de l'ontologie ont été proposées dans la littérature.

En 1997, [Borst, 97] modifia légèrement la définition de Gruber en énonçant que : "Une ontologie est définie comme étant une spécification formelle d'une conceptualisation partagée".

Ces deux définitions ont été expliquées par [Studer, 98] et ses collègues comme suit : Conceptualisation réfère à un modèle abstrait d'un phénomène dans le monde, en ayant identifié les concepts appropriés à ce phénomène. Explicite signifie que le type de concepts utilisés et les contraintes liées à leur usage sont définis explicitement. Formel réfère au fait que l'ontologie doit être traduite en langage interprétable par une machine. Partagé réfère au fait qu'une ontologie capture la connaissance consensuelle, c'est-à-dire non réservée à quelques individus, mais partagée par un groupe ou une communauté.

En 1995, [Guarino & Giaretta, 95] ont choisi sept définitions dont ils ont fourni des interprétations syntaxiques et sémantiques. D'après [Gómez, 99], des auteurs ont également fourni une définition fondée sur la méthodologie qu'ils ont utilisée pour construire leur ontologie. Pour [Swartout, 97] par exemple; "An ontology is a hierarchically structured set of terms for describing a domain that can be used as a skeletal foundation for a knowledge base". Dans le même ordre d'idées, [Bernaras, 96] proposent la définition suivante : "an ontology provides the means for describing explicitly the conceptualization behind the knowledge base".

Les approches associées à ces définitions sont explicitées par la suite dans notre section sur les méthodologies de l'ingénierie ontologique.

3. BRIQUES DE BASE DES ONTOLOGIES

Comme nous l'avons dit plus haut, les ontologies fournissent le vocabulaire propre à un domaine et fixent - avec un degré de formalisation variable - le sens des concepts et des relations entre ceux-ci. Ces concepts (et ces relations) sont organisés via la relation de

subsomption pour former une taxinomie. L'usage montre que cette taxinomie - elle n'est jamais absente, et quelques fois la définition des ontologies se résume même à sa seule construction - constitue la porte d'accès à l'ontologie. Nous détaillons ci-après l'ensemble de ses composantes :

- **Les concepts,** aussi appelés termes ou classes de l'ontologie, correspondent aux abstractions *pertinentes* d'un segment de la réalité (le domaine du problème), retenues en fonction des objectifs qu'on se donne et de l'application envisagée pour l'ontologie. Selon [Gómez, 99] ces concepts peuvent être classifiés selon plusieurs dimensions : 1) *niveau d'abstraction* (concret ou abstrait) ; 2) *atomicité* (élémentaire ou composée) ; 3) *niveau de* réalité (réel ou fictif).
- **Les propriétés** (ou attributs). Les propriétés sont les caractéristiques attachées aux concepts ; elles sont valuées.
- **Les liens organisant les concepts.** La relation de subsomption *is-a,* qui définit un lien de généralisation, est la plus utilisée pour structurer les ontologies. Mais ce n'est pas la seule et, surtout, pas la plus utile dans certain cas [Charlet, 2003]. Dans le domaine de l'audiovisuel, par exemple, il est indispensable de décrire les séquences composant une émission ou les émissions composant une grille des programmes. On a alors besoin de relations de partie-tout (ou *méronymie)* [Winston, 1987].
- **Les relations** traduisent les associations (pertinentes) existant entre les concepts présents dans le segment analysé de la réalité. Ces relations incluent les associations suivantes: 1) Sous-classe-de (généralisation – spécialisation) ; 2) Partie-de (agrégation ou composition) ; 3) Associée-à ; 4) Instance de, etc. Ces relations nous permettent d'apercevoir la structuration et l'interrelation des concepts, les uns par rapport aux autres**.** Les **fonctions** constituent des cas particuliers de relations, dans laquelle un élément de la relation, le nième (extrant) est défini en fonction des n-1 éléments précédents (intrants).
- **Les axiomes**. Les axiomes sont utilisés pour modéliser des assertions toujours vraies. Ils peuvent intervenir dans la définition des concepts et des relations, ou alors sous forme de règles.
- **Les individus** (ou instances). constituent la définition extensionnelle de l'ontologie ; ces objets véhiculent les connaissances (statiques, factuelles) à propos du domaine du problème.

Cette énonciation nous permet de clarifier encore un peu plus l'objet informatique « ontologie », mais elle appelle quelques remarques. La formalisation ontologique repose sur le fait que le monde ressemble à la façon dont il est décrit dans la théorie des modèles : on suppose implicitement qu'il existe des individus pour les énumérer explicitement. La difficulté est donc centrée sur l'élaboration du modèle. Des choix de modélisation sont ainsi effectués durant les différentes étapes de l'élaboration de l'ontologie. Ils doivent être assumés par le concepteur d'ontologies et idéalement dictés par la méthodologie de construction. Un exemple d'un choix de conception est de décider si une connaissance doit être modélisée dans une propriété ou à l'aide d'une relation pointant sur un autre concept. Un critère souvent retenu consiste à dire qu'il s'agit d'une propriété dès lors que les valeurs possibles sont d'un type dit primitif (entier, chaîne de caractères), alors que les valeurs possibles d'une relation sont d'un type dit complexe (c'est-à-dire un autre concept de l'ontologie). Mais cette frontière peut évidemment être remise en question. Enfin, la présence ou non des individus dans l'ontologie est sujette à débat. En effet, la définition des concepts du domaine peut faire appel à des objets ou des assertions individuelles, ce que [Euzenat, 2002] nomme de la *connaissance de contexte.* Cette connaissance est souvent partagée et nécessaire pour la compréhension du domaine, donc incluse dans l'ontologie. Mais là encore, la frontière entre ce type de connaissances et les assertions émises à l'aide de l'ontologie est une question de choix, dictés par l'application à l'origine de l'ontologie construite.

Nous reviendrons sur ces choix de modélisation dans la section 5.3 qui présente quelques méthodologies et outils de construction d'ontologies. Mais avant cela, nous allons présenter dans la prochaine section une classification des ontologies.

4. DIMENSIONS DE CLASSIFICATION

Les ontologies peuvent être classifiées selon plusieurs dimensions. Parmi celles-ci, nous en examinerons quatre : 1) Objet de conceptualisation ; 2) Niveau de détail; 3) Niveau de complétude; 4) Niveau de formalisme de représentation. Ces dimensions de classification sont illustrées à la Figure 2.

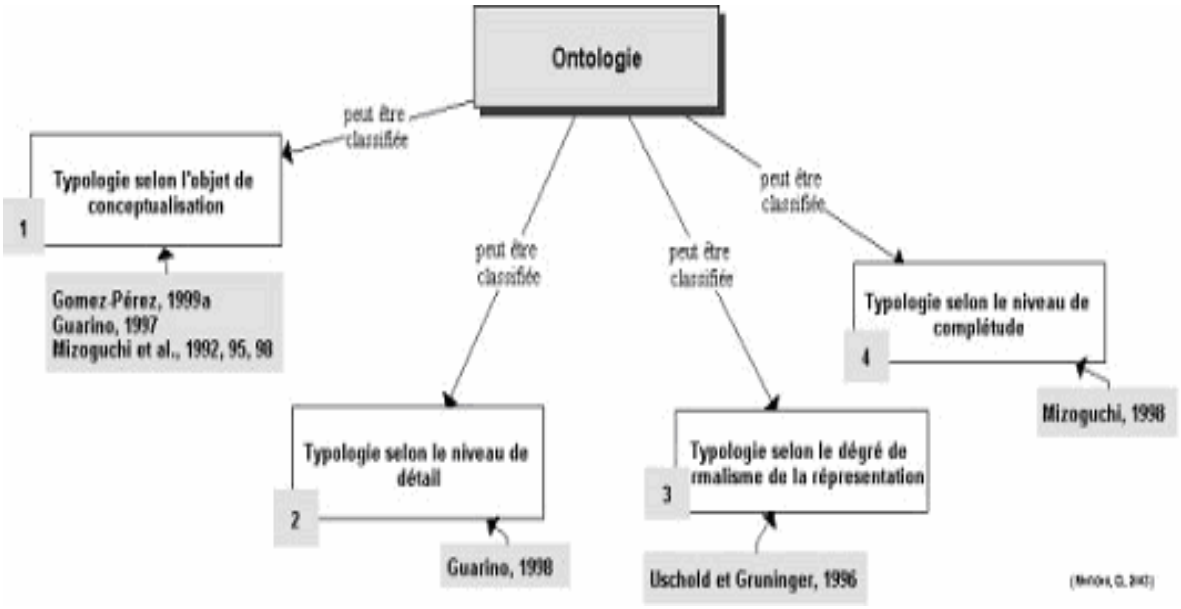


Figure 2 : Typologies d'ontologies selon quatre dimensions de classification

4.1. Typologie selon l'objet de conceptualisation

Les ontologies classifiées selon leur objet de conceptualisation par [Gómez, 99], [Guarino, 97b], [Mizoguchi, 96; 98], [Van, 97], [Vanwelkenhuysen, 94], [Vanwelkenhuysen, 95], [Wieliinga, 93], le sont de la façon suivante : 1) Représentation des connaissances; 2) Supérieure/ Haut niveau; 3) Générique ; 4) Domaine ; 5) Tâche; 6) Application.

- **Ontologie de représentation des connaissances [Gómez, 99a]**, **[Van, 97]**. Ce type d'ontologies regroupe les concepts (primitives de représentation) impliqués dans la formalisation des connaissances. Un exemple est l'*ontologie de Frame* qui intègre les primitives de représentation des langages à base de *frames* : classes, instances, facettes, propriétés/*slots*, relations, restrictions, valeurs permises, etc.
- **Ontologie supérieure ou de Haut niveau [Guarino, 97a]**, **[Sowa, 95a ; 95b]**. Cette ontologie est une ontologie générale. Son sujet est l'étude des catégories des choses qui existent dans le monde, soit les concepts de haute abstraction tels que: les entités, les événements, les états, les processus, les actions, le temps, l'espace, les relations, les propriétés. L'ontologie de haut de niveau est fondée sur : la théorie de l'identité, la méréologie *(theory of whole and parts role)* et la théorie de la dépendance. Guarino et Sowa ont poursuivi chacun indépendamment des recherches sur la théorie de l'ontologie. Tous deux intègrent les fondements philosophiques comme étant des principes à suivre pour concevoir l'ontologie de haut niveau ou supérieure. Sowa introduit deux concepts importants, *Continuant* et *Occurrent*, et obtient douze

catégories supérieures en combinant sept propriétés primitives. L'ontologie supérieure de Guarino consiste en deux mondes : une ontologie des *Particuliers* (choses qui existent dans le monde) et une ontologie des *Universels* comprenant les concepts nécessaires à décrire les *Particuliers*. La conformité aux principes de l'ontologie supérieure a son importance, lorsque le but est de standardiser la conception des ontologies.

- **Ontologie Générique [Gómez, 99 ; 99a], [Van, 97]**. Cette ontologie aussi appelée, *méta-ontologies* ou *core ontologies*, véhicule des connaissances génériques moins abstraites que celles véhiculées par l'ontologie de haut niveau, mais assez générales néanmoins pour être réutilisées à travers différents domaines. Elle peut adresser des connaissances factuelles (*Generic domain ontology*) ou encore des connaissances visant à résoudre des problèmes génériques (connaissances procédurales) appartenant à ou réutilisables à travers différents domaines (*Generic task ontology*). Deux exemples de ce type d'ontologies sont : 1) l'ontologie méréologique [Borst, 97] contenant des relations, *Partie-de* et 2) l'ontologie topologique contenant des relations, *Associé-à*.
- **Ontologie du Domaine [Mizoguchi, 00]**. Cette ontologie régit un ensemble de vocabulaires et de concepts qui décrit un domaine d'application ou monde cible. Elle permet de créer des modèles d'objets du monde cible. L'ontologie du domaine est une méta-description d'une représentation des connaissances, c'est-à-dire une sorte de méta-modèle de connaissance dont les concepts et propriétés sont de type déclaratif. La plupart des ontologies existantes sont des ontologies du domaine. Selon Mizoguchi, l'ontologie du domaine caractérise la connaissance du domaine où la tâche est réalisée.
- **Ontologie de Tâches [Mizoguchi, 00]**. Ce type d'ontologies est utilisé pour conceptualiser des tâches spécifiques dans les systèmes, telles que les tâches de diagnostic, de planification, de conception, de configuration, de tutorat, soit tout ce qui concerne la résolution de problèmes. Elle régit un ensemble de vocabulaires et de concepts qui décrit une structure de résolution des problèmes inhérente aux tâches et indépendante du domaine. Selon Mizoguchi, l'ontologie de tâche caractérise l'architecture computationnelle d'un système à base de connaissances qui réalise une tâche.

- **Ontologie d'Application**. Cette ontologie est la plus spécifique. Les concepts dans l'ontologie d'application correspondent souvent aux rôles joués par les entités du domaine tout en exécutant une certaine activité [Maedche, 02].

4.2. Niveau de détail de l'ontologie

Par rapport au **niveau de détail** utilisé lors de la conceptualisation de l'ontologie en fonction de l'objectif opérationnel envisagé pour l'ontologie, deux catégories au moins peuvent être identifiées :

- **Granularité fine** : correspondant à des ontologies très détaillées, possédant ainsi un vocabulaire plus riche capable d'assurer une description détaillée des concepts pertinents d'un domaine ou d'une tâche. Ce niveau de granularité peut s'avérer utile lorsqu'il s'agit d'établir un consensus entre les agents qui l'utiliseront;
- **Granularité large** : correspondant à un vocabulaire moins détaillé comme par exemple dans les scénarios d'utilisation spécifiques où les utilisateurs sont déjà préalablement d'accord à propos d'une conceptualisation sous-jacente [Fürst, 02], [Guarino, 97b]. Les ontologies de haut niveau possèdent une granularité large, compte tenu que les concepts qu'elles traduisent sont normalement raffinés subséquemment dans d'autres ontologies de domaine ou d'application.

4.3. Typologie selon le niveau de complétude

Le niveau de complétude a été abordé par [Mizoguchi, 98] et [Bachimont, 00]. À titre d'exemple, nous décrivons la typologie de Bachimont. Ce dernier propose la classification sur trois niveaux suivante :

- **Niveau 1 (Sémantique)**: Tous les concepts (caractérisés par un terme/libellé) doivent respecter les quatre principes différentiels : 1) Communauté avec l'ancêtre; 2) Différence (spécification) par rapport à l'ancêtre; 3) Communauté avec les concepts frères (situés au même niveau); 4) Différence par rapport aux concepts frères (sinon il n'aurait pas lieu de le définir). Ces principes correspondent à l'**engagement sémantique** qui assure que chaque concept aura un sens univoque et non contextuel associé. Deux concepts sémantiques sont identiques si l'interprétation du terme/libellé à travers les quatre principes différentiels aboutit à un sens équivalent.

- **Niveau 2 (Référentiel)** : Outre les caractéristiques énoncées au niveau précédent, les concepts référentiels (ou formels) se caractérisent par un terme/libellé dont la sémantique est définie par une extension d'objets. L'**engagement ontologique** spécifie les objets du domaine qui peuvent être associés au concept, conformément à sa signification formelle. Deux concepts formels seront identiques s'ils possèdent la même extension (ex : les concepts *d'étoile du matin* et *d'étoile du soir* associés à Vénus).
- **Niveau 3 (Opérationnel)** : Outre les caractéristiques énoncées au niveau précédent, les concepts du niveau opérationnel ou computationnel sont caractérisés par les opérations qu'il est possible de leur appliquer pour générer des inférences (**engagement computationnel**). Deux concepts opérationnels sont identiques s'ils possèdent le même potentiel d'inférence.

4.4. Typologie selon le niveau du formalisme

Par rapport au **niveau du formalisme de représentation** du langage utilisé pour rendre l'ontologie opérationnelle, [Uscholod, 96b] proposent une classification comprenant quatre **catégories** :

- **Informelles :** ontologies opérationnelles dans un langage naturel (sémantique ouverte);
- **Semi-informelles :** utilisation d'un langage naturel structuré et limité;
- **Semi-formelles :** langage artificiel défini formellement;
- **Formelles** : utilisation d'un langage artificiel contenant une sémantique formelle, ainsi que des théorèmes et des preuves des propriétés telles la robustesse et l'exhaustivité.

Selon Studer, *"il y a différents types d'ontologie et chaque type remplit un rôle différent dans le processus de construction du modèle du domaine"*. [Studer, 98]. La section suivante aborde les fondements de ce processus.

5. FONDEMENTS DE L'INGENIERIE ONTOLOGIQUE

Le processus de construction d'ontologies, appelé ingénierie ontologique, peut être décrit selon les principes qui le gouvernent, et les méthodologies et les outils qui le soutiennent.

5.1. Principes

Gruber pose cinq critères que doit suivre une ontologie destinée à être partagée ou réutilisée par des agents (cogniticiens, systèmes, processus, etc.) [Gruber, 93] :

- **Clarté et Objectivité** : L'ontologie doit fournir la signification des termes définis en fournissant des définitions objectives ainsi qu'une documentation en langage naturel.
- **Cohérence** : Une ontologie cohérente doit permettre des inférences conformes à ces définitions.
- **Extensibilité monotonique maximale** : De nouveaux termes généraux et spécialisés devraient être inclus dans l'ontologie d'une façon qui n'exige pas la révision des définitions existantes.
- **Minimalité des postulats d'encodage** : ce permet d'assurer une bonne portabilité
- **Engagements ontologiques minimaux** : Ce principe invite à faire aussi peu de réclamations que possible au sujet du monde représenté. L'ontologie devrait spécifier le moins possible la signification de ses termes, donnant aux parties qui s'engagent dans cette ontologie la liberté de spécialiser et d'instancier l'ontologie comme elles le désirent.

Il existe des conflits entre ces critères : par exemple, l'objectivité et l'expressivité sont antagonistes, c'est au concepteur de l'ontologie de faire des choix de construction en fonction de l'utilisation qui en sera faite. Ces choix sont également contraints par la conceptualisation sous-jacente. Outre les critères introduits par Gruber, d'autre principes ont fait leurs preuves dans le développement des ontologies et peuvent être résumés comme suit :

- **Principe de distinction ontologique [Borgo, 96] :** les classes dans une ontologie devraient être disjointes. Le critère utilisé pour isoler le noyau de propriétés considérées comme invariables pour une instance d'une classe est appelé le critère d'*Identité*.
- **Modularité [Bernaras, 96]** : Ce principe vise à minimiser les couplages entre les modules.
- **Diversification des hiérarchies** : Ce principe est adopté pour augmenter la puissance fournie par les mécanismes d'héritage multiple[2]. Si suffisamment de connaissances sont représentées dans l'ontologie et que suffisamment de différentes classifications

[2] Certains chercheurs comme Mizoguchi *et al.,* sont opposés à l'idée d'héritage multiple en ingénierie ontologique.

de critères sont utilisées, il est plus facile d'ajouter de nouveaux concepts (puisqu'ils peuvent être facilement spécifiés à partir des concepts et des classifications de critères préexistants) et de les faire hériter de propriétés de différents points de vue.

- **Distance sémantique minimale** : Il s'agit de la distance minimale entre les concepts enfants de mêmes parents. Les concepts similaires sont groupés et représentés comme des sous-classes d'une classe, et devraient être définis en utilisant les mêmes primitives, considérant que les concepts qui sont moins similaires sont représentés plus loin dans la hiérarchie.
- **Normaliser les noms** : Ce principe indique qu'il est préférable de normaliser les noms aussi autant que possible. Cet ensemble de critères et de processus est généralement accepté pour guider le processus d'ingénierie ontologique. Ces trois derniers critères ont été proposés dans [Arpirez, 98].

5.2. Cycle de vie d'une ontologie

Selon [Fernandez, 97], lorsqu'une ontologie devient importante le processus de création d'une ontologie doit être considéré comme un projet à part entière, en conséquence des méthodes de managements doivent être utilisées.

Par ailleurs, les ontologies étant destinées à être utilisées comme des composants logiciels dans des systèmes répondant à des objectifs opérationnels différents, leur développement doit s'appuyer sur les mêmes principes que ceux appliqués en génie logiciel. En particulier, les ontologies doivent être considérées comme des objets techniques évolutifs et possédant un cycle de vie qui nécessite d'être spécifié. Les activités liées aux ontologies sont d'une part des activités de gestion de projet (planification, contrôle, assurance qualité), et d'autre part des activités de développement (spécification, conceptualisation, formalisation) ; s'y ajoutent des activités transversales de support telles que l'évaluation, la documentation, la gestion de la configuration [Blazquez, 98]. Un cycle de vie inspiré du génie logiciel est proposé dans [Dieng, 01]. Il comprend une étape initiale d'évaluation des besoins, une étape de construction, une étape de diffusion, et une étape d'utilisation. Après chaque utilisation significative, l'ontologie et les besoins sont réévalués et l'ontologie peut être étendue et, si nécessaire, en partie reconstruite.

La phase de construction peut être décomposée en 3 étapes : conceptualisation, ontologisation, opérationnalisation. L'étape d'ontologisation peut être complétée d'une étape d'intégration au cours de laquelle une ou plusieurs ontologies vont être importées dans

l'ontologie à construire. Fernandez insiste sur le fait que les activités de documentation et d'évaluation sont nécessaires à chaque étape du processus de construction, l'évaluation précoce permettant de limiter la propagation d'erreurs. Le processus de construction peut être intégré au cycle de vie d'une ontologie comme indiqué dans la Figure 3 [Gandon, 02]. La section suivante va être plus spécifiquement consacrée aux méthodologies mises en oeuvre lors de la phase de construction afin, en particulier, de guider les choix délicats de conceptualisation et de représentation.

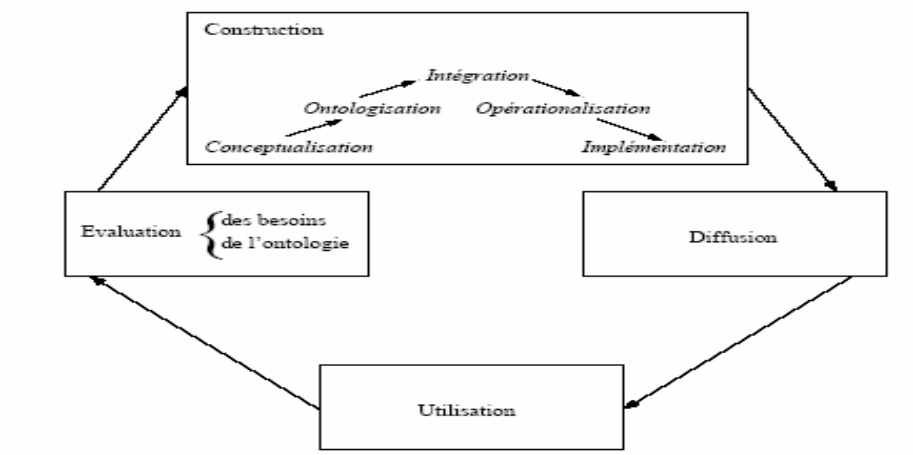


Figure 3 : Cycle de vie d'une ontologie.

5.3. Méthodologies de construction existantes

La conceptualisation d'une ontologie commence par l'organisation des concepts et des relations d'un domaine en taxinomies. Mais malgré ce consensus, les ontologies produites sont très diverses comme en témoigne une des premières études comparatives consacrées aux méthodologies de construction [Noy & Hafner, 1997]. Le problème est qu'il n'est pas aisé pour l'ontologiste de déterminer ce qui doit apparaître dans les taxinomies, ni d'ailleurs de s'assurer que le résultat produit est bien conforme à la conceptualisation souhaitée du domaine. Nous présentons ci-après quelques éléments méthodologiques pouvant apporter des réponses à ces questions. Nous devons préciser que parmi les - nombreux - travaux en la matière, nous nous sommes surtout intéressés à ceux qui cherchaient à prendre en charge la construction *ex nihilo* des ontologies.

Uschold et Gruninger : La méthodologie présentée dans [Uschold & Gruninger, 1996] se fonde sur l'expérience de la construction de plusieurs ontologies pour la modélisation d'entreprises. Elle propose les étapes suivantes :

1. identifier clairement le domaine concerné, le but et la portée de l'ontologie;
2. construire l'ontologie et coder les connaissances en les intégrant à d'autres ontologies;
3. évaluer le résultat ;
4. documenter l'ontologie en établissant des directives pour chacune des étapes.

La deuxième étape est donc la phase de construction proprement dite. Elle aboutit à la spécification d'une ontologie en langage formel. Le modèle n'est pas construit directement et [Uschold & Gruninger, 1996] suggèrent de passer par une étape intermédiaire qui consiste à identifier un ensemble de questions de compétences. Celles-ci constituent l'élément clé qui permet de catégoriser les connaissances que doit inclure une ontologie. Un mécanisme de composition de ces questions de compétences et de leurs réponses permet ensuite de répondre à des questions plus complexes et facilite l'intégration avec d'autres ontologies.

Methontology : La méthodologie Methontology, proposée par l'équipe du LAI de l'Université Polytechnique de Madrid, insère l'activité de construction d'ontologies dans un processus complet de gestion de projet (planification, assurance qualité), de développement (spécification, conceptualisation, formalisation, implémentation, maintenance) et de support (intégration, évaluation, documentation) [Fernández-López *et al.,* 1997], [Blázquez *et al.,* 1998]. Elle permet de spécifier l'ontologie au niveau des connaissances dans la mesure où, contrairement à la méthodologie précédente, il est suggéré de passer par des *représentations intermédiaires* lors de la conceptualisation de l'ontologie. La formalisation et l'implémentation ne sont alors, dans l'idéal, que des phases de traduction quasi-automatique du modèle précédent.

Ce cadre méthodologique est soutenu par l'outil **WEBODE**, que nous présenterons dans la section 5.6, et a permis de mettre au point plusieurs ontologies. Il rejoint également pour partie la méthodologie décrite par F. Gandon et développée au sein de l'équipe ACACIA de l'INRIA [Gandon, 2002].

Terminae : L'acquisition des connaissances d'un domaine est un préalable à sa conceptualisation. Cette tâche étant difficile en général, [Aussenac-Gilles *et al.,* 2003] proposent de l'assister par des outils de traitement du langage naturel opérant sur des textes.

Pour ce faire, une « démarche de corpus » et des outils terminologiques sont utilisés afin de commencer à modéliser le domaine. Ces outils, reposent pour la plupart sur la recherche de formes syntaxiques particulières manifestant les notions recherchées comme des syntagmes nominaux pour des candidats termes, des relations syntaxiques marqueurs de relations sémantiques, ou des proximités d'usage - par exemple, contextes partagés - pour des regroupements de notions [Charlet, 2003]. Le système **Terminae** développé par [Aussenac-Gilles *et al,* 2003] implémente cette méthodologie en intégrant des outils de repérage de candidats termes (**SYNTEX**), de regroupement de contextes (**UPERY**) et de repérage de relations (**YAKWA**).

Guarino et Welty : Les travaux menés par [Guarino & Welty, 2000a], [Guarino & Welty, 2000b] et concernent le « nettoyage » de taxinomies, souvent construites de manière anarchique en utilisant abusivement de la relation de subsomption. Il s'agit plus d'une étape de correction à intégrer dans le processus de construction d'ontologies qu'un réel guide méthodologique complet. L'idée est de donner une valeur à des méta-propriétés - ou *propriétés formelles* - pour chacune des propriétés présentes dans l'ontologie. Les auteurs définissent ainsi les notions de *rigidité, identité, unité* et *dépendance* et donnent les combinaisons valables qui permettront à l'ontologiste de vérifier les règles de subsomption dans la taxinomie. Le module **ODEClean** de l'environnement WebODE propose une implémentation de cette méthodologie, mais celle-ci reste tout de même réservée à des spécialistes compte tenu de sa complexité.

Au final, nous constatons que les méthodologies de construction proposées s'intéressent à l'ensemble du processus de conception d'ontologie en se concentrant sur des questions comme celles du cycle de vie des ontologies construites ou de l'ordonnancement des étapes auquel l'ontologiste devra se soumettre pour valider son travail. Répondre à ces questions est bien évidemment important si l'on veut voir un jour émerger un véritable « génie ontologique », mais nous pensons que la phase de conceptualisation proprement dite, celle où les concepts de l'ontologie sont dégagés, définis par un certains nombres de propriétés et organisés entre eux, gagnerait à être guidée de manière plus précise qu'elle ne l'est dans ces réflexions. Par exemple, la méthode Methontology, alors qu'elle propose un grand nombre de *représentations intermédiaires* [Blázquez *et al.,* 1998] afin de mieux conduire la construction des ontologies au niveau des connaissances, n'insiste pas sur la manière de structurer ses *arbres de classification de concepts,* l'une de ces représentations.

5.3.1. Problème non résolu: comment obtenir les taxinomies?

Parmi les méthodologies présentées, peu proposent des aides concrètes pour guider les utilisateurs à organiser les concepts entre eux : tout repose sur leur bonne intuition du domaine. Ce constat est somme toute normal puisqu' aucune des méthodologies exposées - à l'exception du système Terminae - ne prend réellement en charge l'explicitation des concepts sous sa forme la plus naturelle : le langage. En effet, si à peu près toutes préconisent l'utilisation de celui-ci pour tenter de préciser le sens des concepts manipulés, que ce soit dans des commentaires présents dans l'ontologie elle-même ou dans des documents produits lors du processus de modélisation, elles sont peu à énoncer des consignes précises pour la rédaction de ces compléments [Isaac, 2001]. En fait, malgré les efforts de commentaires, les termes utilisés comme primitives de connaissances peuvent toujours être sujets à des interprétations multiples, ce qui nuit fortement à la compréhension de l'ontologie et à son utilisation.

Le problème réside dans le fait qu'aucune méthodologie ne force l'ontologiste à expliciter clairement le sens qu'il attribue aux concepts : les commentaires restent très informels. B. Bachimont propose, dans une méthodologie introduite dans le cadre du projet Menelas [Menelas, 1994], de contraindre l'utilisateur à un engagement sémantique [Bachimont, 1996], [Bachimont, 2000a] en introduisant une normalisation sémantique des termes manipulés dans l'ontologie. Cette méthodologie propose trois étapes Figure 5 dont nous allons à présent rappeler les principes. Cette méthodologie a été mise en œuvre dans l'éditeur d'ontologies DOE (pour Differential Ontology Editor) que nous présenterons dans la section 5.6.

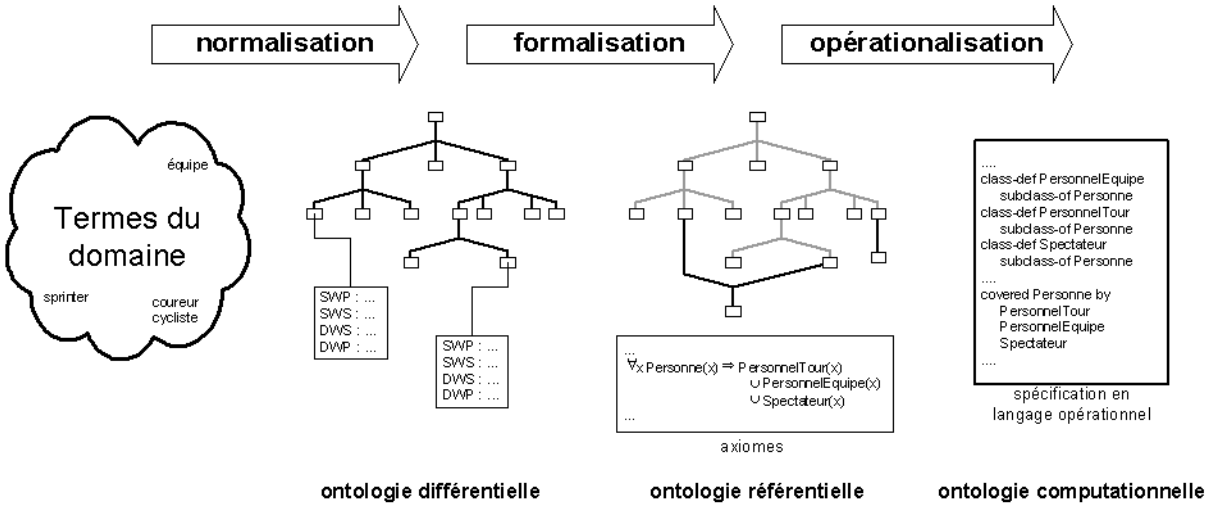


Figure 4 : Méthodologie de construction d'ontologies de B. Bachimont

Étape 1 : la normalisation sémantique

L'étape de normalisation sémantique a pour but d'expliciter le sens des libellés linguistiques pour qu'ils fonctionnent comme des primitives. Ces libellés doivent avoir leurs significations normées pour pouvoir être utilisés. L'usage de la langue naturelle est l'accès privilégié aux connaissances d'un domaine. Il paraît donc logique de commencer par utiliser des procédés d'extraction terminologique afin de trouver des « candidats-termes » (ces candidats seront par exemple les résultats d'un outil comme **LEXTER** [Bourigault, 1994] ou son évolution **SYNTEX**).

Après avoir isolé un certain nombre de libellés, il faut veiller à leurs donner une signification précise et donc utiliser une théorie sémantique appropriée. B. Bachimont propose dans [Bachimont, 2000a] d'utiliser la sémantique différentielle présentée par F. Rastier [Rastier *et al*., 1994]. Dans ce paradigme, le sens est intralinguistique : il se construit par des relations d'opposition entre les unités du système linguistique. En fait, il s'agit d'attribuer un sens aux termes grâce à des traits sémantiques ou *sèmes*. Ces sèmes que l'on va associer à une unité sont regroupables en deux catégories :

- **les sèmes génériques** qui permettent de regrouper les unités entre elles ;
- **les sèmes spécifiques** qui permettent à une unité de se distinguer de celles avec qui on l'a regroupée.

Il faut cependant signaler que ce sens différentiel est fortement dépendant du contexte, qui encadre l'*interprétation*, le mécanisme d'attribution des sèmes aux termes. Or, pour obtenir une primitive réellement exploitable, il faut contrôler cette interprétation : il s'agit bien d'une *normalisation sémantique*.

Comment concrètement parvenir à cette normalisation du sens des unités manipulées? Le réseau d'identités et de différences évoqué précédemment reste à structurer, si on veut réellement le rendre opérationnel, pour aboutir à la construction de ce que l'on va appeler une *ontologie différentielle*. Tout d'abord, on observe qu'une unité (dorénavant, on parlera de *notion*) se définit en premier lieu par une relation de subsomption avec une autre unité : celle dont elle hérite tous ses sèmes génériques. En pratique, le créateur d'ontologies doit donc pouvoir exprimer les identités et les différences de chaque notion dans son voisinage proche :

la notion-parente et les notions-sœurs. Il obtiendra finalement une taxinomie[3] de notions où la signification d'une unité s'obtiendra en collectant les identités et les différences permettant de caractériser les notions rencontrées sur le chemin qui mène de la notion racine (la plus générique) à cette unité.

B. Bachimont a proposé les quatre principes, mentionnés dans la section 4.3, permettant d'expliciter ces informations :

- ***Communauté avec le père*** *(similarity with parent* ou **SWP**) : on explicite pourquoi le fils hérite des propriétés de la notion qui le subsume ;
- ***Différence avec le père*** *(différence with parent* ou **DWP)** : on explicite la différence qui permet de distinguer le fils du père ;
- ***Différence avec la fratrie*** *(différence with siblings* ou **DWS**) : on indique ici la propriété qui permet de distinguer la notion considérée de ses notions-sœurs ;
- ***Communauté avec la fratrie*** *(similarity with siblings* ou **SWS**) : on donne enfin la propriété - admettant plusieurs valeurs exclusives - qui a permis de caractériser les notions d'une même fratrie, justifiant ainsi le principe précédent.

Par le processus de normalisation sémantique exposé ci-dessus, on passe d'un candidat-terme à une notion dont la signification, ancrée dans le domaine et l'application envisagée, est invariable et peut donc fonctionner comme une primitive exprimant une connaissance.

Étape 2 : la formalisation des connaissances

L'arbre ontologique construit dans la première étape a permis de fixer un certain nombre de notions en normalisant leurs sens. Cependant, le travail n'est toujours pas terminé: on ne peut pas vraiment utiliser cet ensemble de notions dans un véritable SBC. Il faut en effet introduire des concepts obéissant à une sémantique formelle et extensionnelle pour que ceux-ci servent en tant que primitives dans un langage formel de représentation des connaissances.

Les notions dont on dispose dans l'ontologie différentielle sont de nature linguistique, toujours soumises à des processus d'interprétation (même si à présent cette interprétation est adaptée à la pratique ciblée lors de la construction de l'ontologie). Le passage à une sémantique extensionnelle va permettre de lier ces notions à un ensemble de référents dans le

[3] Il s'agit d'un arbre : l'héritage multiple est interdit dans la mesure où l'on associe aux notions d'une fratrie des sèmes spécifiques qui sont en opposition.

monde, donnant lieu à ce que l'on va appeler dans la suite une *ontologie référentielle,* composée de *concepts* qui vont agir comme des primitives formelles.

Ces primitives ne sont plus définies par les principes différentiels. Elles n'ont donc pas intrinsèquement de sens interprétatif; elles n'en acquièrent que parce qu'elles sont reliées - par leur libellé - aux notions différentielles. Cependant, ce sont des primitives et elles peuvent, grâce aux mécanismes de composition de sens d'une sémantique extensionnelle, servir à définir de nouveaux concepts formels. Chaque concept étant lié par référence à un ensemble d'objets du domaine (son *extension),* on peut avoir recours à des opérations de composition de sens qui utilisent les opérations qui existent pour les ensembles (réunion, intersection ou complémentaire).

La comparaison des extensions permet de définir une relation d'héritage extensionnelle entre les concepts : un concept sera subsumé par un autre si et seulement si son extension est incluse dans celle de son parent. On peut alors se poser la question de la structure de la hiérarchie de subsomption obtenue. S'il paraît naturel d'affirmer que les relations d'héritage définies dans l'ontologie différentielle tiendront toujours, on peut se demander dans quelle mesure l'ajout de nouveaux nœuds va modifier la structure arborescente. En effet, les « transcriptions » des notions différentielles en concepts formels peuvent admettre des extensions qui ont un sous-ensemble commun. L'héritage multiple devient donc possible, et la structure hiérarchique obtenue est celle d'un treillis.

On a vu que la sémantique référentielle nous permet de définir de nouveaux concepts en fonction d'autres concepts. Il faut signaler que l'on peut également exprimer de telles définitions pour les concepts formels directement issus de la transposition de l'ontologie différentielle. C'est notamment à cette étape que l'on va définir les relations par la donnée de leur arité et de leur domaine, ce qui les associera de fait à des produits cartésiens de références de concepts.

On peut aussi introduire des axiomes logiques, lois qui régissent les comportements des individus qui constituent l'extension des concepts formels. Par exemple, on peut spécifier que les sous-concepts d'un concept forment une partition disjointe de celui-ci. On peut enfin exprimer des lois concernant les relations, comme leur attacher des propriétés algébriques : réflexivité, transitivité, etc. Il est clair que l'ontologie référentielle doit expliciter toutes les lois logiques induites par la conceptualisation que l'on veut spécifier.

Étape 3 : l'opérationnalisation des connaissances

Dans cette troisième et dernière étape, on va chercher à munir les concepts présents dans l'ontologie référentielle d'une signification en termes d'opérations informatiques. On spécifie ainsi le comportement des objets informatiques présents dans le système qui utilise l'ontologie, pour aboutir à ce que l'on va appeler une *ontologie computationnelle.*

Les concepts computationnels se définissent par les inférences, les calculs que pourra effectuer un système à partir de la donnée des individus qu'ils couvrent dans le monde. Le système utilise alors un langage opérationnel de représentation des connaissances qui fait appel à des capacités d'inférence précises répondant aux besoins exprimés lors de la spécification du système. Pour un langage de représentation reposant sur les graphes conceptuels, il s'agira d'opérations de manipulation de graphes (jointure, projection, etc.). Pour un langage fondé sur le paradigme des logiques de description, il s'agira plutôt des tests de subsomption permettant la classification des individus introduits dans la base de connaissances. L'expression des primitives de l'ontologie référentielle dans un de ces langages assigne donc une véritable signification computationnelle aux concepts manipulés par le système.

Dans la section 5.7, nous passerons en revue les environnements d'édition d'ontologies mettant en œuvre ces différentes méthodologies. Mais avant cela, nous allons étudier dans la section suivante les formalismes de représentation des connaissances qui sont à l'origine des langages permettant d'exprimer des ontologies.

5.4. Formalisme de représentation des connaissances

En toute généralité, représenter des connaissances propres à un domaine particulier consiste à décrire et à coder les entités de ce domaine de manière à ce qu'une machine puisse les manipuler afin de raisonner ou de résoudre des problèmes [Kayser, 1997], [Euzenat *et al.,* 2000]. Cette définition met en évidence deux composantes complémentaires de la représentation des connaissances (RC), à savoir l'expression et la manipulation des connaissances. D'une part, les connaissances sont exprimées à l'aide d'un langage formel, dit de description des connaissances. Le langage est doté d'une *syntaxe,* précisant l'ensemble des expressions admissibles du langage et d'une *sémantique* qui permet de fournir un sens aux formules justifiant ainsi la validité des opérations effectuées. D'autre part, le but est de mécaniser un certain nombre de manipulations sur les connaissances exprimées. Ainsi, il sera

nécessaire de modifier, compléter, inférer de nouvelles connaissances. Ces manipulations possibles sont spécifiées sous forme de mécanismes respectant la sémantique et opérant sur les éléments de la représentation.

Les travaux en RC ont donné naissance à de nombreux formalismes. Parmi les formalismes de représentation développés au niveau conceptuel, trois grands modèles sont distingués: les logiques de description, les graphes conceptuels et les langages de frames.

5.4.1. Logiques de description

Les composants de base des logiques de description sont les concepts, les rôles et les individus qui correspondent respectivement aux classes, relations et instances des langages à objets [Kayser, 1997]. Un concept est une description des propriétés communes à une collection d'individus; un individu est une entité particulière, instance du concept; un rôle est une relation binaire entre deux individus.

Les DL offrent un langage terminologique (*T-Box*) permettant de décrire les concepts et les rôles, et un langage assertionnel (*A-Box*) permettant de décrire les règles et contraintes qui s'appliquent aux concepts ainsi que les instances des concepts. Les principaux mécanismes d'inférence offerts par les DL sont basés sur la relation de subsomption. **LOOM** [Macgregor, 1991] et **KL-ONE** [Brachman, 1985] sont des exemples de systèmes implémentant ce modèle.

5.4.2. Les graphes conceptuels

Les graphes conceptuels ont été initialement conçus pour l'analyse et la compréhension du langage naturel en s'inspirant des graphes existentiels et des réseaux sémantiques [Sowa, 2000]. Leur but est d'exprimer des connaissances sous une forme logique précise compréhensible par des humains et adapté à un traitement automatisé. Facilement interprétables en langage naturel, les graphes conceptuels peuvent servir d'intermédiaire pour traduire différents formalismes. Leur aspect graphique permet, quant à lui, une lecture facile des connaissances tout en assurant un cadre formel.

Les graphes conceptuels sont décomposés en deux niveaux : le niveau terminologique où sont décrit les concepts, les relations et les instances de concepts, ainsi que les liens de subsomption entre concepts et entre relations et le niveau assertionnel où sont représentés les faits, les règles et les contraintes sous forme de graphes où les sommets sont des instances de concepts et les arcs des relations.

Ce formalisme est implémenté, entre autres, dans **COGITANT**, une plateforme de développement de SBC utilisant les graphes conceptuels [Genest, 1998] et **PROLOG+CG**, une extension de **PROLOG** basée sur les graphes conceptuels [Kabbaj, 2000].

5.4.3. Langages des frames

Introduit dés les années 70 par Minsky [Minsky, 1975] comme une modélisation de base pour la représentation de connaissances dans le domaine d'Intelligence Artificielle (AI), le modèle des *frames* ou *schémas* a depuis été adapté à d'autre problématiques puisqu'il a donné naissance au modèle objet, qui envahit peu à peu les différentes branches de l'informatique.

L'idée de frame est très simple. Un « frame » est dans ce contexte un objet nommé, qui est utilisé pour représenter un certain concept dans un domaine. Une frame représente n'importe quelle primitive conceptuelle et est dotée d'attributs (*slots*), qui peuvent porter différentes valeurs (*facets*), et d'instances [Kiffer, 1995]. Parmi les langages de frame on peut citer **YAFOOL** (*Yet Another Frame Based Object-Oriented Language*), **OML** (*Ontology Markup Language*), **SHIRKA** et bien d'autres encore.

Le choix du formalisme de représentation des ontologies doit être guidé par la nature des connaissances à représenter et par le raisonnement à effectuer sur ces connaissances. Le formalisme utilisé doit permettre de représenter efficacement la connaissance, c'est-à-dire définir, structurer et classer les concepts relatifs au domaine, décrire les propriétés qui les caractérisent ainsi que les relations sémantiques qui existent entre eux. Il doit aussi permettre d'exploiter cette connaissance et de la faire évoluer. Les logiques de description et les graphes conceptuels sont les seuls à fournir cette capacité d'inférence.

5.5. Langages de spécification d'ontologies

Plusieurs langages de spécification d'ontologies (ou langage d'ontologies) ont été développés pendant les dernières années, et ils deviendront sûrement des langages d'ontologie dans le contexte du *Web sémantique*. Certains d'entre eux sont basés sur la syntaxe de **XML**, tels que **XOL** (*Ontology Exchange Language*), **OML** (*Ontology Markup Language*), **RDF** (*Resource Description Framework*), **RDF Schéma**. Les deux derniers sont des langages créés par des groupes de travail du *World Wide Web Consortium* (*W3C*). En conclusion, trois langages additionnels sont établis sur **RDF(S)** pour améliorer ses caractéristiques: **OIL** (*Ontology Inference Layer*), **DAML+OIL** et **OWL** (*Web Ontology Language*).

Ces langages marquent une première étape dans la volonté de rendre la sémantique des informations du Web accessible aux machines et agents informatiques et contribuent donc dans la vision d'un *Web sémantique* telle que la proposée T. Berners-Lee [Berners-Lee *et al*., 2001], Ceci passe par une structuration des données qu'il contient. Dans cet esprit, l'utilisation des ontologies paraît la solution la plus intuitive, mais la mise en place d'ontologie sur le Web nécessite des standards qui pourraient assurer une interopérabilité sémantique. Cependant pour des raisons pratiques, il est presque impossible de créer un langage unique qui permettrait de construire des ontologies sur le Web. L'approche adoptée est d'utiliser plusieurs langages en couches pour parvenir à créer des ontologies (voir Figure 5). Nous présentons dans les sous sections suivantes les plus important de ces langages.

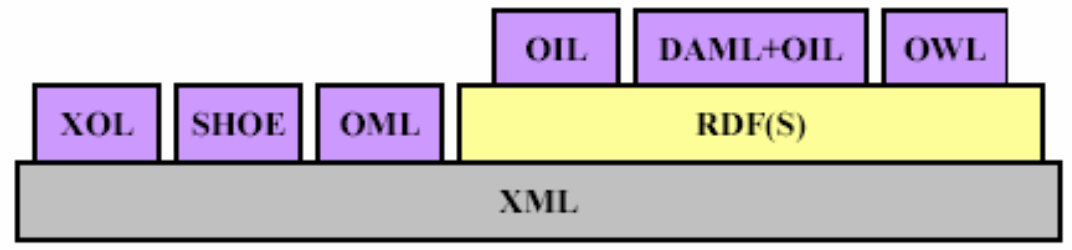


Figure 5 : Les langages d'ontologies

5.4.1 RDF et RDF Schéma

Le W3C a adopté le langage RDF (*Ressource Description Framework*) comme formalisme standard de représentation [RDF, 2002a]. Utilisant la syntaxe XML (*Extended Markup Language* [XML, 2002]) qui constitue déjà un standard, le RDF permet de décrire des ressources Web en termes de ressources, propriétés et valeurs. Une ressource peut être une page Web (identifiée par son **URI**, *United Resource Identifier*) ou une partie de page (identifiée par une balise). Les propriétés couvrent les notions d'attributs, relations ou aspect et servent à décrire une caractéristique d'une ressource en précisant sa valeur. Les valeurs peuvent être des ressources ou des chaînes de caractères.

Pour décrire n'importe quel type de connaissances à l'aide de ce formalisme, on doit d'abord décrire en RDF le modèle sémantique à utiliser. Par exemple, pour décrire des connaissances en terme de concepts et de relations hiérarchisés, l'introduction des types « concepts » et « relations » et des propriétés de subsomption et d'instanciation est nécessaire.

Un schéma de base incluant les primitives sémantiques généralement utilisées a ainsi été ajouté au RDF et constitue ce qu'on appelle le RDF Schéma (RDF(S) [RDF, 2002b]).

Dans l'optique d'une utilisation d'ontologies sur le Web, le langage RDF(S) a été enrichi par l'apport du langage OIL (*Ontology Interchange Language* [OIL, 2002]) qui permet d'exprimer une sémantique à travers le modèle des frames tout en utilisant la syntaxe de RDF(S). OIL offre de nouvelles primitives permettant de définir des classes à l'aide de mécanismes ensemblistes (intersection de classes, union de classes, complémentaire d'une classe). Il permet également d'affiner les propriétés de RDF(S) en en contraignant la cardinalité ou en en restreignant la portée [Fensel, 2000].

5.4.2. DAML

Le langage DAML (*DARPA Agent Markup Language* [DAML, 2002]) est développé comme une extension du XML et du RDF et il permettra de créer des ontologies et de baliser l'information pour la rendre lisible par des machines. Un des objectifs du programme DAML est de mettre en place des technologies permettant aux agents d'identifier et comprendre dynamiquement des sources d'information, et de communiquer au niveau sémantique.

5.4.3. DAML+OIL

Le langage OIL a été fusionné avec le langage DAML pour former le DAML+OIL. DAML est conçu pour permettre l'expression d'ontologies utilisées dans le cadre de systèmes multi-agents. Il offre les primitives usuelles d'une représentation à base de frames et utilise la syntaxe RDF [Hendler, 2001]. L'intégration de OIL rend possible les inférences compatibles avec les logiques de description, essentiellement les calculs des liens de subsomption.

5.4.4. OWL

La combinaison de RDF(S) et de DAML+OIL laisse augurer de l'émergence d'un langage standard et opérationnel de représentation de connaissances pour le Web. Les spécifications opérationnelles de OWL (*Ontology Web Language*) ont déjà été adoptées par le W3C [OWL, 2002]. L'architecture en couches d'un tel langage est résumée dans le dessin de la Figure 6 et originellement proposée par T. Berners-Lee, où la représentation des ontologies apparaît comme l'objectif immédiat d'un processus qui conduira à la construction d'un Web incluant non seulement une énorme quantité d'informations, mais également tous les mécanismes permettant d'accéder à cette information et de l'exploiter.

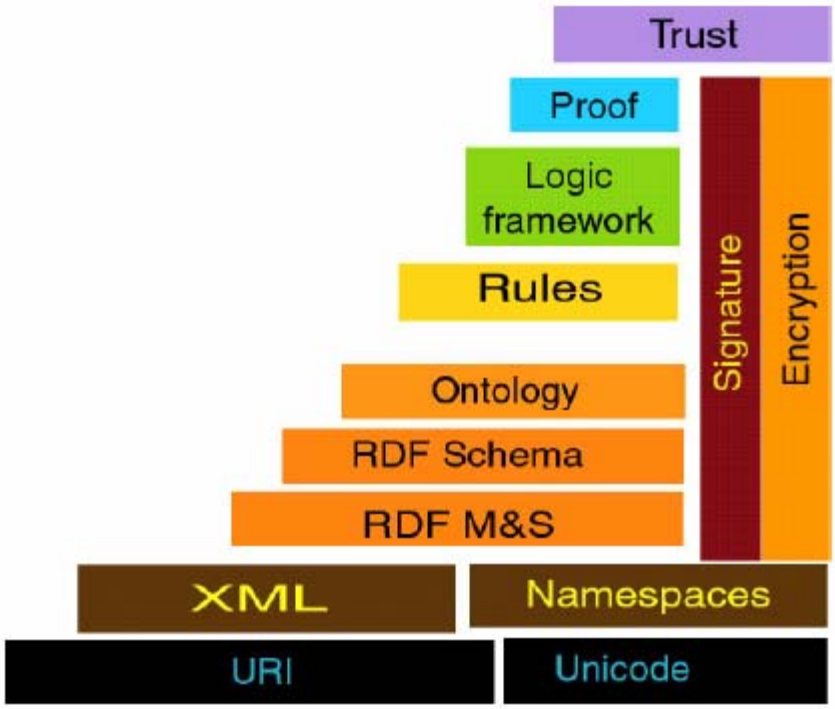


Figure 6 : Les couches du Web sémantique

5.6. Environnements d'édition

Après avoir étudié les différentes méthodologies de construction d'ontologies et détaillée les formalismes et langages de spécification, nous présentons maintenant les outils informatiques permettant de les créer. Ces outils peuvent se regrouper grossièrement en deux catégories. Dans la première, on trouve les plus anciens historiquement, qui permettent de spécifier les ontologies au niveau symbolique (voir, par exemple, le serveur **Ontolingua**[4] [Farquhar *et al.,* 1995]) : une grande partie des définitions des objets se fait directement dans un langage de représentation des connaissances donné (pour Ontolingua, il s'agit de **KIF**), auquel le créateur et l'utilisateur de l'ontologie doivent se plier. Dans la seconde catégorie, les outils prennent mieux en compte l'importance du niveau des connaissances : ils proposent à l'utilisateur de créer l'ontologie de manière relativement indépendante de tout langage implémenté et prennent ensuite automatiquement en charge l'opérationnalisation de l'ontologie, en la transposant dans divers langages. Cette évolution tend à rapprocher les ontologies de leur but original : il semble en effet naturel de chercher à s'abstraire - dans un premier temps - du niveau symbolique si on veut obtenir une ontologie permettant un réel

10.[4] http://www-ksl-svc.stanford.edu:5915/

partage d'une compréhension. Cette dernière catégorie regroupe les outils principalement utilisés aujourd'hui. Nous les détaillons ci-dessous[12].

Protégé2000[5] [Noy *et al.,* 2000] : est un environnement graphique de développement d'ontologies développé par le SMI de Stanford. Dans le modèle des connaissances de *Protégé,* les ontologies consistent en une hiérarchie de *classes* qui ont des attributs *(slots),* qui peuvent eux-mêmes avoir certaines propriétés *(facets).* L'édition des listes de ces trois types d'objets se fait par l'intermédiaire de l'interface graphique, sans avoir besoin d'exprimer ce que l'on a à spécifier dans un langage formel : il suffit juste de remplir les différents *formulaires* correspondant à ce que l'on veut spécifier. Ce modèle autorise d'ailleurs une liberté de conception assez importante, puisque le contenu des formulaires à remplir peut être modifié suivant les besoins *via* un système de *méta-classes,* qui constituent des sortes de « patrons » de connaissance. L'interface, très bien conçue, et l'architecture logicielle permettant l'insertion de *plug-ins* pouvant apporter de nouvelles fonctionnalités (par exemple, la possibilité d'importer et d'exporter les ontologies construites dans divers langages opérationnels de représentation tels que OWL ou encore la spécification d'axiomes) ont participé au succès de *Protégé2000,* qui regroupe une communauté d'utilisateurs très importante et constitue une référence pour beaucoup d'autres outils.

OILed [Bechhofer *et al.,* 2001], développé sous la responsabilité de l'université de Manchester, a été conçu pour éditer des ontologies dans le langage de représentation **OIL**, un des précurseurs du langage **OWL** (voir section 5.5). Officiellement, il n'a pas d'autre ambition que de construire des exemples montrant les vertus du langage pour lequel il a été créé. À ce titre, *OIled* est souvent considéré comme une simple interface de la logique de description. Néanmoins, il offre la plus grande partie de ce que l'on peut attendre d'un éditeur d'ontologies. On peut créer des hiérarchies de classes et spécialiser les rôles, et utiliser avec l'interface les types d'axiomes les plus courants. Cet éditeur offre également les services d'un raisonneur, ***FaCT*** [Horrocks, 1998], qui permet de tester la satisfaisabilité des définitions de classes et de découvrir des subsomptions restées implicites dans l'ontologie.

OntoEdit [Sure *et al.,* 2002] : Contrairement aux deux outils précédents, il n'est pas disponible gratuitement dans sa version complète[6]. Il présente les fonctionnalités essentielles communes aux autres éditeurs (hiérarchie de concepts, expression d'axiomes, export de

[5] http://protege.stanford.edu/index.shtml
[6] Une version de démonstration est disponible sur le site D'**ONTOPRISE**, la société qui le développe en collaboration avec L'**AIFB** de Karlsruhe

l'ontologie dans des langages divers) et a le mérite de s'appuyer sur une réflexion méthodologique significative. La modélisation des axiomes a fait l'objet de soins particuliers pour pouvoir être effectuée - en tout cas pour les types les plus répandus - indépendamment d'un formalisme privilégié, et cela pour faciliter la traduction d'un langage de représentation à un autre. Il propose également une gestion originale des *questionnaires de compétences.* Des questions pour les réponses desquelles l'ontologie doit fournir le matériel conceptuel, on peut extraire les termes appelés à intégrer l'ontologie. Un petit outil fait une comparaison lexicale des termes extraits des différentes questions pour en déduire automatiquement d'éventuelles subsomptions. Le procédé semble cependant loin d'être fiable, puisqu'il repose sur l'hypothèse que le nom d'un concept se retrouve parfois dans le nom de ses spécialisations.

WebODE[7] [Arpirez *et al.,* 2001], développé par le LAI de Madrid, est une plate-forme de conception d'ontologies fonctionnant en ligne. D'un point de vue méthodologique, l'outil fait suite à **ODE**, un éditeur qui assurait fidèlement le support de la méthodologie **METHONTOLOGY**. Il illustre bien l'évolution des outils de construction d'ontologies, puisque les nombreuses tables de son prédécesseur ont été remplacées par une interface très travaillée, réalisant un pas supplémentaire vers une conception au niveau des connaissances. On peut cependant regretter que cette évolution se soit faite au détriment de l'application des contraintes méthodologiques : les *représentations intermédiaires* utilisées dans le processus de conception sont désormais moins mises en avant, à tel point que le guide de l'utilisateur ne les signale que pour « *assurer la compatibilité conceptuelle avec* **ODE**». L'accent a plus été mis sur la possibilité d'un travail collaboratif ou sur la mise à disposition d'outils complémentaires, comme un moteur d'inférences ou le module **ODECLEAN**.

DOE[8] (*Differential Ontology Edito*) [Troncy and Isaac, 2002], [Bachimont *et al.,* 2002] : À l'instar des autres éditeurs, il offre une représentation graphique des arbres de concepts et de relations de l'ontologie et permet d'interagir avec les hiérarchies. L'outil assiste également la saisie des principes différentiels issus de la méthodologie de B. Bachimont en automatisant partiellement cette tâche. Le modèle de représentation de l'ontologie est finalement proche de celui du langage RDFS, à ceci près qu'il autorise la modélisation de relations n-aires. Au niveau formel, l'éditeur est capable de faire quelques inférences en vérifiant la consistance de l'ontologie. Il permet également d'ajouter des individus à l'ontologie. Finalement, le passage à une ontologie computationnelle s'effectue par un export

[7] http://delicias.dia.fi.upm.es/webODE/
[8] L'outil est disponible gratuitement à http://opales.ina.fr/public/.

de l'ontologie formelle dans un certain nombre de langages opérationnels. Cette traduction s'effectue grâce à des feuilles de style **XSLT** appliquées au format de sauvegarde XML de l'éditeur. De la même façon, DOE peut importer des ontologies modélisées dans d'autres outils grâce à des feuilles **XSLT** dédiées.

5.6.1. ONTOLOGOS

Si tous ces outils peuvent être considérés- en tout cas dans une première approche- comme satisfaisants en matière d'expressivité ou d'interface, on peut affirmer qu'il existe toujours un certain vide méthodologique. En effet, compte tenu de la tendance actuelle consacrant l'usage de corpus pour la construction automatique d'ontologies, et les gains, en termes de temps et coût de développement, que présente la réutilisation des ontologies, il s'avère que ces outils sont mal menés face à cette nouvelle donne. Ainsi nous avons décidé de développer note propre éditeur d'ontologies que nous avons appelé **OntoLogos**. Cet outil n'a pas pour ambition de concurrencer les grands environnements existants, mais plutôt de combler le gap entre l'acquisition des ontologies à partir de textes d'une part et d'une autre part la réutilisation des ontologies déjà développées. Cette synergie vise à tirer meilleur parti de ces deux *sources terminologiques.* Idéalement, OntoLogos sera développé sous forme d'un *plugiciel* à intégrer dans Protégé2000, dans le but de profiter des potentialités ingénierique de cet environnement, et ce conformément aux principes méthodologiques que nous avons introduits pour soutenir son développement.

OntoLogos et la méthodologie sous-jacente seront abordés en détail dans la **Partie II** de ce mémoire. A présent, nous allons clore le chapitre par la présentation et discussion de certains compléments théoriques relatifs à l'ingénierie ontologique.

6. COMPLEMENTS THEORIQUES

Dans cette section nous rassemblons quelques discussions portant notamment sur la fusion, la validation, les applications et l'acquisition automatiques à partir de corpus des ontologies. Ces points ne seront présentés que brièvement, pour les éléments ayants un rapport direct avec notre problématique, des explications plus poussées seront rapportées dans les chapitres suivants.

6.1. Fusion

La fusion d'ontologies présente beaucoup d'intérêts pour de multiples applications. Ceci va non seulement faciliter le maintien des ontologies existantes, mais également réduire le temps et coût nécessaires pour le développement de nouvelles ontologies. Cette utilisation conjointe de plusieurs ontologies peut passer d'un simple alignement à une véritable fusion. La fusion a pour objectif de produire à partir de deux (ou plus de deux) ontologies une nouvelle ontologie regroupant les ontologies sources. Une fusion manuelle utilisant des outils d'édition conventionnelles, serait difficile, demandant beaucoup d'effort et sujette aux erreurs. Pour cela, plusieurs systèmes et méthodologies automatisant cette tâche sont crées [Hovy, 1998], [Chalupsky, 2000], [Noy, 2000], [McGuinness *et al.*, 2000]. Ces outils sont basés sur des heuristiques d'appariement syntaxique et sémantique entre les concepts composants les ontologies à fusionner. FCA-MERGE [Stumme & Maedche, 2001] est une autre méthode de fusion qui emploie une approche ascendante pour effectuer l'opération de fusion. Poussés par une quête de meilleure qualité des ontologies fusionnées, les recherches dans ce domaine vont certainement proposer de nouvelles techniques et outils pour faciliter cette tâche. La fusion des ontologies constitue une partie importante dans notre système de construction d'ontologies, nous allons présenter dans la deuxième partie de ce mémoire l'approche que nous avons employé pour l'intégration des ontologies à réutiliser.

6.2. Validation

La validation d'une ontologie se fait *a priori* par des tests correspondants à l'objectif opérationnel de l'ontologie. Cette méthode est en particulier préconisée par M. GRUNINGER et M.S. FOX qui proposent d'utiliser des questions de compétences permettant de tester l'ontologie [Gruninger, 95]. Il est cependant difficile de traduire le but d'une application en quelques questions dont on sera certain qu'elles couvrent l'ensemble du contexte d'usage.

La validation manuelle est souvent difficile à effectuer, d'autant plus qu'on ne peut pas déterminer qui sera l'acteur de cette opération, est-ce l'expert du domaine, le cogniticien ou encore l'utilisateur finale de l'ontologie. La validation automatique, quant à elle, est beaucoup plus compliquée, en effet il serait difficile d'opérationnaliser les critères d'évaluation qui garantissent la validité de l'ontologie par rapport au domaine et à l'application cible. Par conséquent, cette question continue donner lieu à de multiples travaux visant l'automatisation de cette tâche du cycle de vie d'une ontologie. Nous n'avons pas ignoré cette tâche dans notre

méthodologie de construction, et quoique nous ayons utilisé des techniques rudimentaires, elles permettent de garantir la validité des hiérarchies construites.

6.3. Applications

Les ontologies, en tant que représentation partagée et consensuelle des concepts d'un domaine, sont un élément clé dans toute une gamme d'applications faisant appel à des connaissances. En effet, la gestion de connaissances, la traduction automatiques, le commerce électronique, la recherche d'information, sont des exemples de domaines où les ontologies sont utilisées pour faciliter la diffusion, le partage, la réutilisation et la conservation des connaissances. Le Web sémantique [Berners-Lee *et al.*, 2001] est sans doute le terrain idéal d'application des ontologies, elles y sont utilisées pour rendre compte du contenu sémantique des ressources disponibles. Ainsi beaucoup de langages ont été introduits pour faciliter le développement des ontologies dans le cadre du Web sémantique. Toutefois, cette utilisation des ontologies reste entravée par le temps et le coût consacrés au développement. Ce mémoire vient en réponse à cette problématique, en essayant de proposer un cadre unificateur pour l'acquisition automatique à partir de corpus et la réutilisation des ontologies dans le but d'accélérer le développement et améliorer la qualité des ontologies produites.

6.4. Acquisition automatique à partir de corpus

Afin de profiter de la masse importante de documents électroniques rédigés en langue naturelle pour surmonter le goulot d'étranglement auquel se heurte l'acquisition manuelle des ontologies, une nouvelle tendance se fait ressentir consacrant le texte comme source d'une acquisition automatique. Dans cette perspective, les outils d'extraction de termes et de relations entre termes seront utilisés

Une telle opération est possible si l'on automatise les tâches de génération de l'arborescence de concepts et des différents types de relations entre ces concepts trouvées dans les textes. A partir d'une telle ontologie partielle, un cogniticien peut décider par exemple, de compléter certains éléments du modèle obtenu en rajoutant des relations implicites ne figurant pas dans les textes traités ou encore des relations génériques non repérées par les outils d'acquisition de connaissances.

Dans le domaine de l'acquisition d'ontologies propres au domaine à partir de textes techniques, on peut citer les travaux de Assadi [Assadi, 1998] où l'auteur propose une

méthodologie et utilise des outils d'extraction et de classification de termes (LEXTER et LEXICLASS) afin de construire une ontologie qu'il appelle régionale, relative à un domaine donné, en adoptant des principes issus de la sémantique différentielle de F. Rastier. D'autre travaux plus récents ont abordé ce problème de construction notamment, les travaux de [Maedche & Staab, 2001], [Velardi *et al.*, 2002], [Faatz & Steinmetz, 2002], [Kietz *et al.*, 2000]. Ce domaine continue d'évoluer par le développement d'approches pluridisciplinaires mettant en œuvre des techniques issues de l'apprentissage automatique, du traitement automatique de la langue naturelle, de la linguistique computationnelle et bien d'autres disciplines, ceci dans le but d'améliorer la qualité de l'acquisition et de réduire le recours au facteur humain.

7. RESUME

Dans ce chapitre, nous avons montré comment construire des ontologies, et ce en explicitant de quoi elles sont composées, les principes régissant leur développement et les méthodologies adoptées à cet égard. Sur le constat qu'aucune de ces méthodologies ne pouvait satisfaire notre besoin d'intégrer l'acquisition à partir de corpus et la réutilisation des ontologies dans une seule approche, nous avons mis en avant les motifs du développement de notre outil ONTOLOGOS.

Le traitement de corpus textuel étant la première composante de notre méthodologie, nous allons présenter dans le chapitre suivant, le cadre théorique de cette tâche. Nous verrons que celle-ci se ramène essentiellement à l'extraction de termes et de relations entre termes, nous présentons alors les différentes techniques employées pour effectuer cette extraction.

CHAPITRE 2

TERMINOLOGIE ET EXTRACTION DE CONNAISSANCES À PARTIR DE CORPUS

« L'accumulation des connaissances n'est pas la connaissance. »

ALBERTO MANGUEL

Notre problématique est d'acquérir à partir de corpus de documents textuels l'ensemble de connaissances utiles pour l'élaboration (semi) automatique d'ontologies de domaines. Nous avons défini deux objectifs majeurs pour l'acquisition de connaissances dans des corpus textuels. Le premier objectif consiste à acquérir des termes significatifs et représentatifs du domaine. Le deuxième objectif est d'acquérir des relations entre ces termes.

Ces objectifs ne sont pas très éloignés des objectifs d'extraction de terminologie. En effet, la plupart des méthodes d'extraction de terminologie essayent de capturer la notion de concept à l'aide de classes, contenant des termes, utilisées pour préciser le concept. La tâche de construction de ressources terminologiques est en quelque sorte une activité d'interprétation de textes au cours de laquelle un analyste construit une description des termes existants dans le texte. Le terme est alors une unité de description résultant d'un travail d'interprétation et de modélisation mené à partir de l'analyse d'un corpus de référence et s'intégrant dans une ressource terminologique cible.

Puisque l'acquisition d'ontologies à partir de corpus est en premier lieu un travail de terminologie, ce chapitre sera dédié à un tour d'horizon des travaux existants dans le domaine de l'acquisition d'informations sémantiques lexicales sur corpus (relevant du domaine de la terminologie). Nous présentons dans un premier temps les enjeux applicatifs de ce domaine de recherche, après cela nous introduisons quelques concepts liés à cette tâche d'acquisition, nous essayons par la suite de situer les origines de ce domaine de recherche. Nous nous intéressons ensuite plus spécifiquement à l'acquisition de termes puis à celle de relations

lexicales sémantiques; nous nous attachons, à travers un panorama des travaux existants dans chacun de ces domaines, à faire ressortir les grandes familles de techniques utilisées. En particulier, l'opposition des méthodes exploitant l'aspect numérique des données ou leur aspect structurel est mise en avant, et les avantages et inconvénients de chacune sont détaillés

1. DES ENJEUX APPLICATIFS IMPORTANTS

Le présent chapitre porte sur les méthodes d'aide à la construction de ressources terminologiques à partir de corpus. Les recherches sur ce thème ont été suscitées d'abord par les exigences d'efficacité des systèmes informatiques de gestion de l'information dans les entreprises industrielles ou dans les institutions. Suite à l'utilisation généralisée des outils de bureautique, à l'internationalisation des échanges et au développement d'Internet, la production de documents sous forme électronique s'accélère sans cesse. Or pour produire, diffuser, rechercher, exploiter et traduire ces documents, les outils de gestion de l'information ont besoin de ressources terminologiques. La gamme des produits à base terminologique nécessaires pour répondre à ces besoins s'élargit considérablement. A côté des bases de données terminologiques multilingues classiques pour l'aide à la traduction, on voit apparaître de nouveaux types de ressources terminologiques adaptées aux nouvelles applications de la terminologie en entreprise : thesaurus pour les systèmes d'indexation automatique, index structurés pour les documentations techniques hypertextuelles, terminologies de référence pour les systèmes d'aide à la rédaction, référentiels terminologiques pour les systèmes de gestion de données techniques, ontologies pour les mémoires d'entreprise, pour les systèmes d'aide à la décision ou pour les systèmes d'extraction d'information, glossaires de référence et liste de termes pour les outils de communication interne et externe, etc.

Au moment de la mise en place d'une application de gestion de l'information dans une entreprise, les ressources terminologiques nécessaires pour garantir l'efficacité du système sont rarement déjà disponibles sous la forme adéquate. Se posent alors les problèmes de leur construction, ou de leur mise à jour et de leur recyclage si elles existent déjà sous des formes inappropriées. Il s'avère d'emblée que le gisement essentiel pour l'acquisition de ressources terminologiques est constitué par les documents produits ou manipulés par l'entreprise. Il faut donc disposer d'outils informatiques d'analyse de textes pour la construction de ressources terminologiques. Dans les années quatre-vingt-dix, cette pression des applications a rencontré un contexte favorable du côté des recherches en traitement automatique des langues : d'une part, les travaux en analyse statistique de la langue ont connu un renouveau certain, et, d'autre part, on a conçu des analyseurs à grande échelle qu'ils soient partiels, c'est-à-dire ne traitant qu'une partie des textes, ou peu profonds, c'est-à-dire ne fournissant que des informations incomplètes sur les données traitées. C'est de la rencontre entre ces besoins importants en milieu industriel et les recherches menées en traitement de corpus textuels que s'est

constituée la problématique de recherche sur l'acquisition de ressources terminologiques à partir de textes.

2. DEFINITIONS

Dans cette section nous présentons quelques notions liées à la tâche de traitement de corpus textuel, nous les aborderons à travers des définitions succinctes :

Le nombre d'occurrences d'un mot dans un texte : Le nombre d'occurrences d'un mot ou d'un lexème, désigne l'apparition de ce mot ou de ce lexème dans le texte (on utilisera désormais le mot « *fréquence* » pour désigner le nombre d'occurrences d'un mot). Les occurrences de mots qui constituent un texte s'appellent « *token* ».

Contexte linguistique : Le contexte linguistique d'un mot (ou d'un groupe de mots) est le contexte qui représente l'entourage linguistique du mot (ou du groupe de mots). Par exemple le contexte linguistique du mot « à » dans « turbine à gaz » est composé d'un contexte gauche « turbine » et d'un contexte droit « gaz ».

Les cooccurrences : La cooccurrence est la présence mutuelle de deux unités lexicales dans une fenêtre de texte de taille définie. Ainsi, par exemple dans la phrase ci-dessous :

Le rôle d'une station de compression est de réduire

L'unité lexicale « rôle » et l'unité lexicale « est »sont des cooccurrents dans la fenêtre « le rôle d'une station de compression est de ».

Les collocations : Mounin **[Mounin, 1974]** a défini la collocation comme étant « *l'association habituelle d'une unité lexicale avec d'autres unités* ». C'est donc une cooccurrence particulière puisque les unités lexicales sont en plus ici juxtaposées. Par exemple, « station de compression » est un exemple de collocation. Dans Smadja [Smadja, 1993] une collocation est une cooccurrence de mots ayant une forme syntaxique précise : N N, ADJ N, etc.

La notion de syntagme : Un syntagme est un groupe de mots formant une unité à l'intérieur de la phrase. Dans le cadre de l'analyse syntaxique d'une phrase, on parle de segmentation en unités fonctionnelles appelées syntagmes. Par exemple on peut citer les types de syntagmes suivants : syntagme nominal, syntagme verbal, syntagme prépositionnel, etc. Exemples de syntagmes nominaux :

Gaz pauvres (syntagme nominal composé de deux noms)

Station de compression (syntagme nominal composé d'un nom suivi d'un adjectif)

Corpus et domaine : Dans le domaine linguistique, le corpus est défini comme un ensemble d'énoncés rassemblés pour une étude linguistique spécifique. Dans le cadre d'une analyse statistique, le corpus est défini comme un échantillon fini représentatif de la langue ou des aspects de la langue qu'on veut étudier. Cet échantillon de textes doit être sélectionné selon des critères précis, puis intégrés dans des fichiers.

Les corpus sont souvent associés à des domaines, des sous-domaines voire à des micro-domaines spécialisés. C'est à l'intérieur d'un domaine bien défini, que l'on procède à l'étude d'un corpus pour par exemple l'acquisition de connaissances.

3. ORIGINES DE L'ACQUISITION AUTOMATIQUE

Comme nous l'avons mentionné dans la première section de ce chapitre, les nécessités industrielles sont en grande partie à l'origine de l'acquisition de ressources sémantiques [Bourigault & Jacquemin, 2000]. De nombreux outils d'extraction (entre autres **ANA**, **ACABIT**, **LEXTER**, etc.) ont ainsi été développés dans ce contexte pour répondre à des attentes précises telles que la traduction automatique, la gestion des connaissances, l'indexation automatique. Du fait de cette naissance en milieu industriel, ces besoins d'acquisition se sont exprimés pour des langages spécialisés et non pas pour la langue « générale ». Cela a conduit les travaux de sémantique lexicale à se porter sur l'étude du domaine connexe de la terminologie.

Cependant, d'autres domaines de recherche ont également contribué aux développements de tels outils pour satisfaire leurs propres besoins. C'est notamment le cas de la recherche d'information dans laquelle la recherche d'une représentation adaptée des documents a conduit certains chercheurs à s'intéresser à la sémantique lexicale et plus particulièrement à la terminologie. Le domaine de l'ingénierie de connaissances a également contribué, à travers la construction d'ontologies, au développement de l'acquisition de terminologie, le terme étant généralement considéré comme porteur d'un concept et donc de la connaissance.

L'acquisition d'informations lexicales sémantiques sur corpus peut se découper artificiellement en deux types de travaux. Les premiers, que nous présentons dans la section

suivante porte sur l'extraction des unités sémantiques, les seconds, que nous verrons en section 5, se concentrent sur les relations entre ces unités.

4. UNITES DES LEXIQUES SEMANTIQUES

Beaucoup de travaux en acquisition d'unités sémantiques relèvent de l'extraction de terminologie, cela s'explique par le fait que ces travaux se sont souvent focalisés sur des domaines spécialisés, et donc des langages de spécialité. Or, dans ce contexte, les unités sémantiques porteuses de sens sont principalement des termes.

Cependant, les techniques utilisées dans ce domaine ne précisent pas toujours le statut linguistique exact des unités acquises, leur utilisation à l'acquisition de termes étant principalement le fait de leur application aux textes spécialisés. On peut donc utiliser ces mêmes techniques pour acquérir d'autres unités linguistiques ; nous les présentons donc comme des méthodes génériques d'acquisition d'unités de lexiques sémantiques.

Après une brève présentation de ce que recouvre, traditionnellement et de nos jours, la notion de terme, nous passons en revue quelques-uns des nombreux travaux effectués dans le domaine. Pour ce faire, nous les regroupons en trois grandes familles suivant les différents aspects du texte que ces techniques exploitent : fréquentiel, structurel ou les deux ensemble.

4.1. Qu'est-ce qu'un terme

Pour répondre à la question de ce qu'est un terme, la terminologie traditionnelle se heurte de nos jours à la linguistique de corpus. Les fondements théoriques classiques se révélant peu adaptés à cette pratique se trouvent revus. Dans la section suivante, nous montrons en effet que la définition classique du terme a été récemment critiquée et des définitions plus pragmatiques proposées. Nous nous intéressons ensuite aux différents modes de création des termes et terminons en rappelant quelques variantes de forme reconnues de ces unités sémantiques.

4.1.1. Définitions

La terminologie en tant que discipline est définie par l'ISO (Organisation international de normalisation) comme l'« étude scientifique des notions et des termes en usage dans les langues de spécialités ». Ainsi au sein de ces langages de spécialités, le terme est généralement défini comme un objet linguistique à part entière [Lerat, 1995] utilisé

principalement dans la littérature technique et scientifique, visant à faire référence à des concepts de façon consensuelle.

Plus précisément, selon E. Wüster [Wüster, 1981], fondateur de la théorie générale de la terminologie, le terme désigne un concept scientifique, lui-même lié à d'autres concepts dans une organisation principalement taxinomique. Il est ainsi postulé que le terme est univoque et mono-référentiel (il y a donc une correspondance un à un entre termes et concepts) et universellement accepté comme tel parmi les utilisateurs de la langue de spécialité. La terminologie est donc la représentation parfaite du système conceptuel sous-jacent au domaine de connaissance.

L'approche wüsterienne, avec sa notion de label unique, est largement compatible avec le modèle aristotélicien représenté par le triangle sémiotique signe/concept/objet [Lerat, 1995], [Rastier, 1995]. D'autres linguistes, comme Rondeau, considère le terme comme un signe linguistique au sens saussurien [Saussure, 1916], avec un signifiant (appelé encore *dénomination*) et un *signifié* (ou notion). Contrairement à la définition de Wüster, un terme désigne donc en même temps la dénomination et la notion. Quel que soit le point de vue adopté, la fonction de dénomination du terme semble impliquer naturellement l'utilisation de noms ou de syntagmes nominaux comme base du terme (voir [Rastier, 1995] pour une discussion sur ce point).

Cependant, en marge de cette définition traditionnelle, certains phénomènes inattendus, comme la variabilité de la terminologie même au sein d'un domaine, ont été constatés lors de travaux sur de grandes masses de données et contredisent la nécessaire fixité posée par la théorie. Il est notamment fait le constat qu'il n'y a pas *une* terminologie, qui représenterait *le* savoir sur le domaine, mais autant de terminologies que d'applications dans lesquelles ces terminologies sont utilisées [Bourigault & Slodzian, 1999]. De plus, ces applications opérant le plus souvent sur des documents textuels (à des fins de traduction, d'indexation, etc.), et les connaissances d'un domaine s'exprimant généralement sous forme de documents écrits, la constitution de terminologies semble se définir désormais comme une tâche d'analyse de corpus textuels. Ces considérations contredisent donc la notion référentielle et unique du terme en l'inscrivant dans un cadre beaucoup plus pragmatique qu'est celui de l'utilité pour une application donnée. Par voie de conséquence, le nom, privilégié dans le cadre classique pour porter le terme, n'est plus forcément le seul élément de texte à prendre en compte dans cette nouvelle terminologie qualifiée de textuelle.

4.1.2. Différentes formes de termes

Distinction terme simple et terme complexe

On fait généralement la distinction entre termes simples et termes complexes. Les premiers sont composés d'un unique mot plein, comme un nom par exemple. Ils sont de ce fait plus susceptible d'être ambigus, mais en revanche ont un comportement syntaxique plus simple à modéliser, permettant de les acquérir et de les interpréter plus facilement. Les termes complexes sont quant à eux constitués d'au moins deux unités lexicales pleines. Ils ont des caractéristiques opposées à celles des termes simples : ils sont peu ambigus mais requièrent une analyse syntaxique plus fine pour être modélisés. Ils sont également plus difficiles à repérer, leurs différents constituants pouvant être séparés au sein de la phrase, et plus difficiles à interpréter. Les termes complexes, rendant mieux compte de la technicité d'un texte, sont plus couramment rencontrés dans un corpus spécialisé [Bourigault, 1992].

Variations de termes

La variation terminologique est un phénomène connu et très présent dans les textes et plutôt le fait des termes complexes. Les variantes d'un terme doivent bien sûr en partager la sémantique et donc renvoyer au même réfèrent pour être valide. B. Daille [Daille, 2002] propose une typologie des variations directement dérivée de ses travaux en acquisition de termes :

- la variation graphique concernant principalement les changements de graphie (casse, absence ou présence d'un trait d'union, etc.) ;
- la variation flexionnelle (mise au pluriel d'un ou de plusieurs constituants d'un terme complexe) ;
- la variation syntaxique faible, affectant les mots grammaticaux (comme *fixation d'azote* pour *fixation de l'azote)* ;
- la variation syntaxique forte, modifiant la structure interne du terme (telle que *lait cru de brebis* pour *lait de brebis)* ;
- la variation morphosyntaxique, modifiant la structure du terme et les mots qui le compose (comme *acidité du sang* et *acidité sanguine)* ;
- la variation paradigmatique, échangeant un mot par un synonyme sans modification de la structure morphosyntaxique *(épuisement du combustible* pour *appauvrissement du combustible)* ;

- la variation anaphorique, faisant référence à une mention préalable dans le texte *(procédé alimentaire* pour *procédé de conservation alimentaire).*

La prise en compte de la variation de termes, et plus généralement d'unités lexicales sémantiques, est un enjeu important de l'acquisition automatique. Il est en effet essentiel de pouvoir distinguer les variantes d'un même élément pour pouvoir, le cas échéant, les regrouper comme une seule unité.

4.1.3. Candidats-termes

La question qui se pose à la suite des définitions qui précèdent est, sans contredit : peut-on vraiment isoler automatiquement un terme en contexte sans craindre de se tromper? Les multiples travaux en acquisition automatique de termes ont démontré que ce n'est pas possible et que les listes de termes retenus dans un corpus contiennent des unités qui ne possèdent pas de statut terminologique. La présente section aborde le concept de *candidat-terme* mis en place pour pallier cette impossibilité qu'a la machine de déterminer le statut terminologique d'une unité lexicale.

L'objectif des recherches en acquisition automatique des termes est d'établir des listes exhaustives de termes contenus dans des corpus textuels spécialisés. Cependant, l'identification automatique des termes pose, de par la nature référentielle particulière des termes et leur réalisation morphosyntaxique tout a fait comparable a celle des syntagmes de discours, des problèmes majeurs. Par exemple, les syntagmes nominaux dans les phrases *il parle à cette* ***fille de programmation*** (non-terme) et il *utilise un* ***langage de programmation*** (terme) ont une structure syntagmatique identique, mais un statut terminologique très différent.

Cette réalisation en discours identique de groupes nominaux de nature terminologique et des syntagmes nominaux non spécialisés ne fait que compliquer la donne pour les ordinateurs. Etant donné la très grande difficulté qu'ont ces derniers à gérer les informations de type sémantique, il est donc impossible, pour le moment, de penser a distinguer automatiquement les termes des non-termes sans une intervention humaine non négligeable.

Afin d'adopter une terminologie représentative des résultats d'un système d'acquisition automatique, nous devons considérer que le système fournit à l'humain une liste de candidats-termes et non une liste de termes. L'adoption d'une telle terminologie peut sembler trop

prudente a certains, mais nous croyons que les systèmes d'acquisition automatique des termes s'inscrivent dans une chaîne de travail interactive ou le terminologue est celui qui porte un jugement sur la valeur terminologique d'une proposition faite par 1'outil. Ainsi, il incombera au terminologue de distinguer, dans 1'ensemble des résultats proposés par le système, les termes, ou les candidats-termes valides, des propositions sans intérêt terminologique.

L'adoption de la notion de candidat-terme ne remet cependant pas en cause 1'objectif des systèmes, qui est de repérer les termes. Pour le moment, nous désirons, comme la majorité des auteurs s'intéressant à la reconnaissance automatique des termes, nuancer nos propos. La terminologie que nous adoptons rejoint celle mise de 1'avant par Bourigault [Bourigault, 1994a] et qui semble reprise par la majorité des chercheurs oeuvrant dans le domaine : [Habert *et al.,* 1997], [Jacquemin, 1997], [Bouveret, 1998], et [Daille, 1999].

Mesure d'efficacité

Comme nous 1'avons décrit plus tôt, une liste de candidats-termes est composée à la fois de termes et d'unités qui ne sont pas des termes. Ces dernières, malgré qu'elles n'aient pas de statut terminologique bien défini, sont tout de même intéressantes pour le terminologue dans la majorité des cas. Nous ne pouvons cependant pas en tenir compte dans un cadre d'acquisition automatique de termes. Ces formes sont cependant généralement conservées pour 1'evaluation et la validation des résultats.

C'est pour cette raison que certains indices ont été mis en place afin d'évaluer la performance des systèmes et leur capacité à isoler le *pertinent* du *non-pertinent.* Afin de permettre de comparer les résultats obtenus par les divers systèmes d'acquisition automatique de termes au travail d'un terminologue, on peut envisager le scénario présenté par la Figure 7.

L'ensemble TR représente une liste de termes de référence compilée par un terminologue ou un spécialiste; c'est cette liste qui sera utilisée pour 1'evaluation des performances du logiciel. La liste des termes de référence contient la section TR^{-} qui correspond aux termes de la liste de référence qui n'ont pas été identifiés par 1'outil.

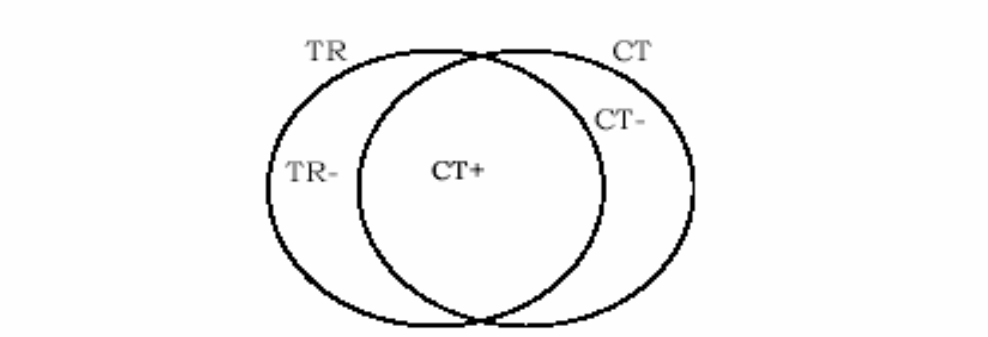


Figure 7 : Mesure d'efficacité des logiciels

L'ensemble des candidats-termes identifiés par le logiciel se nomme CT. Les candidats-termes dont le statut terminologique est confirmé par un terminologue ou un spécialiste sont regroupés dans le sous-ensemble CT^+ alors que les candidats-termes ayant été retenus par erreur par le logiciel se trouvent dans 1'aire étiquetée CT^-.

Parmi les indices les plus fréquents retenus pour évaluer les performances à 1'aide des divers sous-ensembles illustrés par la Figure 7, on a souvent recours aux notions de *rappel,* de *précision,* de *silence* et de *bruit.* Ces indices nous viennent du domaine de 1'indexation et de la recherche documentaire ou ils sont utilisés depuis de nombreuses années.

Le *rappel (R)* correspond à la proportion des réponses pertinentes extraites par un système donné par rapport à 1'ensemble des réponses pertinentes possibles. Ainsi, si un corpus contient 100 termes attestés par un terminologue (TR) et qu'un système identifie 150 candidats-termes, dont 95 qui apparaissent dans la liste validée par le terminologue (CT^+), on dira que le taux de rappel est de 0,95 (95/100). Cet indice nous permet donc de nous concentrer uniquement sur 1'intersection entre la liste qui nous sert de référence et la liste générée par le logiciel.

$$R = \frac{CT^+}{TR}$$

A notre avis, la majorité des logiciels obtiennent une bonne performance lorsque l'on examine leurs résultats selon l'angle du rappel. La majorité des outils présentement disponibles sur le marché ont un taux de rappel supérieur à 90 %. On peut donc considérer que les systèmes sont à même de repérer les termes dans les textes et d'en dresser une liste.

Par contre, cette liste devient moins intéressante dés que les termes sont dilués dans une liste de candidats-termes erronés. Le terminologue se doit alors de procéder au dépouillement

terminologique de la liste des candidats-termes extraits et 1'attrait pour un système automatique en est grandement diminué. Le gain de productivité que le système automatisé laisse miroiter est ainsi mis en veilleuse par le temps nécessaire au nettoyage des données.

La *précision (P)* correspond au nombre de réponses pertinentes (CT^+) identifiées par un système donné par rapport à 1'ensemble des réponses (pertinentes ou non) identifiées par le même système (CT). Reprenons 1'exemple cité plus haut où un corpus contient 100 termes. Le logiciel a, à la suite de son analyse, recensé 150 candidats-termes, dont 95 CT^+. On dira alors que la précision est de 0,63, ce qui correspond au nombre de termes identifiés (95) sur 1'ensemble des candidats-termes identifiés (150).

$$P = \frac{CT^+}{CT}$$

Pour sa part, **le *silence (S)*** compare le nombre de termes d'une liste de référence établie par un terminologue ou un spécialiste qui n'ont pas été identifiés (TR^-) par un logiciel avec le nombre total de termes dans cette même liste de référence (TR). Ainsi, en reprenant 1'exemple cité plus haut, si 5 termes qui apparaissent dans la liste de référence n'ont pas été relevés par le logiciel, on peut parler d'un silence de 5/100 (0,05).

$$S = \frac{TR^-}{TR}$$

Malheureusement, le silence n'est généralement pas pris en considération dans le cadre des recherches en acquisition automatique des termes. II s'agit cependant d'un indice très pertinent et il serait intéressant d'entreprendre des travaux visant à évaluer les impacts de la recherche d'une précision accrue sur le silence. Une telle démarche permettrait de vérifier combien de bons candidats-termes sont mis de côté au profit d'une augmentation de la qualité des résultats.

Enfin, **le *bruit (B)*** évalue le nombre de candidats-termes extraits par le logiciel qui sont absents de la liste de référence (CT^-) par rapport au nombre total de candidats-termes extraits par le logiciel (CT). En reprenant notre exemple, on obtient une valeur de 55 candidats-termes erronés (CT^-) alors que le total de candidats-termes proposés est fixe à 150; le bruit est donc de 55/150 (0,37).

$$B = \frac{CT^-}{CT}$$

Cet indice recouvre donc la portion négative des candidats-termes et peut aussi être exprimé sous la forme *B = 1-P* à partir moment où la précision est une valeur connue.

Après avoir présenté les différentes définitions de la notion de terme et introduit les mesure de d'efficacité utilisées pour la validation des outils d'acquisition automatique de termes, à présent, nous allons aborder les différentes techniques qu'utilise ces systèmes d'acquisition pour leur identification automatique.

4.2. Acquisition de termes

La plupart des définitions du terme données ci-dessus, même pragmatiques, sont non opératoires. Elles ne permettent donc pas de dériver une technique d'acquisition qui serait universelle. Cela explique sans doute la diversité des travaux effectués et des approches utilisées, ainsi que des communautés scientifiques s'intéressant à cette tâche d'acquisition (recherche d'information, ingénierie des connaissances, etc.).

Nous tentons dans ce qui suit de définir un cadre unificateur à ces différentes approches. Nous examinons en particulier les présupposés communs à toutes ces techniques et définissons quelques notations utilisées par la suite. Nous présentons ensuite, sous un cadre formel commun, les grandes familles d'outils existants, en différenciant les approches numériques, structurelles et celles manipulant ces deux types d'information.

La tâche d'acquisition d'unités sémantiques lexicales (par exemple les termes d'un domaine) peut se représenter formellement par une fonction *f* telle que $f: D \rightarrow T_D$ où *D* est le domaine étudié et T_D l'ensemble des unités sémantiques lexicales de *D.* Cette fonction *f* peut représenter le travail d'un expert du domaine, un processus automatique, ou bien encore un processus semi-automatique allié à un expert humain.

Dans ces deux derniers cas, la fonction *f* est en réalité utilisée sur un corpus C_D représentatif du domaine. Même dans le cas où *f* représente une expertise humaine, celle-ci ne se faisant généralement pas *ex nihilo,* l'expert s'appuie sur un corpus. Il a en effet été constaté que l'hypothèse selon laquelle l'expert d'un domaine est le seul dépositaire d'un système conceptuel qu'il suffit de mettre au jour est non productive [Bourigault & Slodzian, 1999]. Par

ailleurs, la variabilité de terminologie au sein d'un même domaine évoquée ci-dessus semble contredire l'existence même d'une terminologie d'un domaine qui ne soit pas directement reliée à un corpus de ce domaine. Le problème de l'acquisition se récrit donc en $f: C_D \rightarrow T_{C_D}$ et, si la représentativité du domaine par le corpus est avérée, on espère avoir $T_{C_D} = T_D$.

Enfin, la fonction *f* parfaite est en réalité impossible à atteindre, et ce même dans le cas où l'on fait appel à un expert humain. On approxime donc en pratique cette fonction d'acquisition par $\bar{f}$. Finalement, la problématique de l'acquisition est donc : $\bar{f}: C_D \rightarrow \overline{T}_{C_D}$ avec, on l'espère encore, $\overline{T}_{C_D} \approx T_{C_D} \approx T_D$.

Cette pseudo-égalité montre les deux sources de problèmes auxquels on se heurte lorsque l'on souhaite construire une base d'informations sémantiques d'un domaine. Il faut dans un premier temps constituer un corpus représentatif du domaine, c'est-à-dire couvrant exhaustivement le domaine ciblé et seulement ce domaine. Selon la méthode employée pour construire $\bar{f}$, le corpus C_D peut nécessiter des caractéristiques supplémentaires comme par exemple la redondance (la même information est présente plusieurs fois dans le corpus sous des formes proches). Dans un deuxième temps, pour acquérir les informations voulues sur ce corpus, il faut trouver une méthode fiable, c'est-à-dire extrayant toutes les informations ciblées mais seulement celles-ci. Nous examinons ci-dessous plusieurs des diverses approches utilisées pour la construction de $\bar{f}$. La phase finale de la tâche d'acquisition est néanmoins commune à toutes ces approches : il s'agit de l'examen des propositions de $\bar{f}$, appelés candidats-termes, par un expert qui les valide ou non en tant que termes.

On peut imaginer plusieurs façons de classer les différentes techniques d'acquisition d'informations, aucune d'elles n'offrant un découpage parfait. Nous choisissons pour notre part de considérer la nature des informations exploitées par les techniques pour les regrouper. Ainsi, nous examinons dans un premier temps les approches que l'on peut appeler numériques dans le sens où elles exploitent la nature fréquentielle des objets à acquérir et utilisent le plus souvent pour ce faire des techniques statistiques. Nous proposons ensuite un tour d'horizon des approches utilisant au contraire des informations structurelles ou symboliques pour acquérir les objets ciblés. Enfin nous présentons les techniques combinant explicitement ces deux approches.

Bien entendu, ce découpage du panorama des approches de l'acquisition, entièrement artificiel, ne saurait rendre compte de manière parfaite du caractère continu et complexe de ce spectre, certaines méthodes mélangeant de manière indissociable différentes méthodes et natures d'informations. Par ailleurs, cette section n'a pas vocation à présenter une liste exhaustive de tous les outils développés dans ce domaine, mais tente d'illustrer par des représentants pionniers ou célèbres les différentes techniques considérées.

4.2.1 Approche numérique

Les approches numériques d'acquisition de termes sur corpus ont été largement utilisées depuis de nombreuses années et elles continuent de connaître un grand succès. Elles sont aidées en cela par leur grande robustesse et par le fait que les documents informatisés sont de plus en plus facilement disponibles, rendant en cela la constitution de corpus volumineux (point de départ obligatoire de ces techniques purement quantitatives) plus aisée. Dans ces approches, $\bar{f}$ est en effet construite en exploitant la redondance de l'information terminologique. L'aspect fréquentiel utilisé est donc d'autant plus fiable et performant que le corpus est volumineux. Nous présentons ci-dessous deux familles de techniques très proches tirant parti de cet aspect fréquentiel : l'approche par cooccurrences et l'approche par segments répétés.

Approche par cooccurrences

De nombreux travaux cherchent à associer les mots apparaissant ensemble dans un texte de manière statistiquement significative. Les associations binaires ont notamment fait l'objet de travaux dans des buts lexicographiques [Church & Hanks, 1989 ; 1990], [Church *et al.,* 1991]. Les techniques utilisées reposent pour la plupart sur l'évaluation de la probabilité que les deux mots étudiés apparaissent ensemble dans une certaine fenêtre de texte plus souvent que le hasard ne l'aurait permis. Les entités trouvées par ce type de méthode sont d'ailleurs souvent dénommées *collocations p*our mettre en relief cette notion d'apparition conjointe des composants dans le texte.

L'hypothèse sur laquelle repose cette approche est que le contexte d'un mot apporte des informations sur le sens du mot, ce que J. R. Firth, cité dans [Church & Hanks, 1989], exprime par : « *You shall know a word by the company it keeps* ». Le principe opératoire de ces techniques est relativement simple :

1. il faut calculer pour chaque couple de mots un indice statistique (un score) mesurant la force du lien unissant ces deux mots ;
2. les couples finalement retenus sont ceux dont le score dépasse un seuil fixé.

Si la méthode est simple, les indices statistiques peuvent en revanche être extrêmement sophistiqués et variés. Parmi ceux utilisés dans ce cadre, un des plus célèbres est certainement celui de l'information mutuelle, adapté au besoin de l'acquisition de collocations par [Church & Hanks, 1989].

En reprenant les notations définies ci-dessus, ces techniques peuvent donc se formaliser sous la forme : $\bar{f}(C_D) = \overline{T}_{C_D} = \{(xy) \in C_D \,|\, AS(P(x), P(y), P(x,y)...) > seuil\}$, où AS est un indice statistique mesurant l'association entre les constituants du candidat-terme binaire (*xy*) et les *P(x)*, *P(y)*, *P(x,y)* sont les probabilités d'apparition de *x*, de *y* et de *x* et *y* ensemble[9] dans une certaine fenêtre sur C_D. Ces probabilités d'apparition étant a priori inconnues, elles sont généralement estimées à partir des fréquences relatives et conjointes des mots sur C_D ou un extrait de C_D.

Même si tous les candidats extraits sont effectivement de bons termes ($\bar{f}$ a alors une précision parfaite) on a néanmoins l'inclusion suivante : $\overline{T}_{C_D} \subseteq \overline{T}_{C_D}$ (et plus certainement $\overline{T}_{C_D} \subset \overline{T}_{C_D}$) puisqu' aucun des termes du domaine composés d'un mot ou de plus de deux mots ne peut être repéré par ce type de technique. Par ailleurs, des associations lexicales autres que les termes composés sont trouvés par cette approche, notamment les séquences répétitives telles que les locutions adverbiales ou prépositionnelles. Deux paramètres sont particulièrement influents dans ces techniques : le seuil à partir duquel un couple sera considéré pertinent, et la taille de la fenêtre choisie. Le premier paramètre est important puisqu'il détermine la qualité des résultats : on reproche souvent à ces techniques de ne pas réussir à détecter les phénomènes rares (associations pertinentes mais dont le nombre d'occurrences est trop faible pour dépasser le seuil du bruit). La taille de la fenêtre est également importante puisqu'il est en effet montré [Brown et al., 1992] qu'une fenêtre petite (2 à 5 mots) favorise la détection de composés alors qu'une fenêtre plus grande (supérieure à 5 mots) permet d'observer des associations d'ordre paradigmatique ou sémantique entre les

[9] Suivant les cas, *P(x, y)* représentera la probabilité d'apparition des deux mots ensemble quel que soit leur ordre, ou bien la probabilité d'apparition de *x* suivi de *y* au sein d'une fenêtre de texte.

deux constituants. Ce dernier résultat explique que ce genre de méthodes soit également utilisé pour l'acquisition de relations sémantiques sur corpus.

Approche par segments répétés

Les techniques précédentes ont le défaut de n'extraire que des candidats-termes binaires. Pour contourner ce problème, l'approche dite des segments répétés [Lebart &: Salem, 1994], développée en premier lieu dans un contexte lexicométrique, peut être utilisée. Son fonctionnement consiste à identifier dans le texte toute suite d'unités textuelles (segments répétés) reproduite sans variations à plusieurs endroits d'un corpus. Ainsi, cette approche se formalise de la manière suivante :

$$\bar{f}(C_D) = \overline{T}_{C_D} = \left\{ (x_1 x_2 ... x_n) \in C_D \,|\, freq((x_1, ..., x_n)) > seuil \right\}$$

Où $(x_1 x_2 ... x_n)$ est un segment répété de *n* composants, *freq(L)* est la fonction indiquant la fréquence de la séquence *L* dans le corpus C_D, et *seuil* est une valeur numérique choisie par l'utilisateur. La forte similitude avec la formalisation des approches par cooccurrences souligne la parenté évidente de ces deux familles ; on notera cependant que la notion de séquence, et donc d'ordre des mots, est explicitement formulée dans l'utilisation des segments répétés alors qu'elle n'est pas forcément prise en compte pour les cooccurrences.

Sans autre raffinement, comme notamment des restrictions sur les catégories de mots pouvant appartenir à un segment, cette méthode permet de repérer des objets linguistiques très hétérogènes comme des morceaux de syntagmes nominaux plus ou moins figés ou des fragments de texte récurrents mais peu intéressants (par exemple, le fragment *est un)* [Habert &: Jacquemin, 1993]. Les résultats sont donc trop bruités pour être utilisés directement dans un cadre de détections de termes, mais peuvent fournir un point de départ à d'autres techniques.

4.2.2. Approche symbolique

Nous l'avons vu, les définitions traditionnelles ou plus actuelles des termes sont non opératoires et ne peuvent donc directement être utilisées pour acquérir des candidats-termes. Néanmoins, un certain nombre de travaux s'appuient pour mener l'acquisition sur des indices structurels. Ces indices portent soit sur les constituants du terme, soit sur leur contexte, et

peuvent être de nature différente : informations lexicales, morphologiques, syntaxiques ou sémantiques.

Les techniques d'acquisition structurelles exploitent principalement deux sources d'obtention de définitions des structures porteuses des termes. La première, la plus commune, est l'expertise linguistique; des définitions opérationnelles des termes, ou d'objets linguistiques proches, sont établies par des linguistes puis utilisées pour trouver les candidats-termes répondant à ces définitions. La seconde source de structures est moins utilisée; il s'agit de techniques d'apprentissage artificiel, qui proposent, en se basant souvent sur l'analyse d'exemples, des patrons d'extraction manipulant divers indices structurels.

Par expertise linguistique

TERMINO [David & Plante, 1990], [David & Plante, 1991] est un outil pionnier de l'acquisition sur corpus de terminologie. Il est basé sur les travaux de É. Benveniste sur les synapsies, structures composées binaires (ou récursivement binaires) constituées d'un déterminé (tête) et d'un déterminant (expansion) qui se définissent selon un ensemble de traits [Benveniste, 1974] :

- la nature syntaxique (non morphologique) de la liaison entre les membres composant la synapsie ;
- l'emploi de joncteurs à cet effet, notamment *de* et à ;
- l'ordre déterminé et déterminant des membres ;
- leur forme pleine, et le choix libre de tout substantif ou adjectif;
- l'absence d'article devant le déterminant ;
- la possibilité d'expansion pour l'un ou l'autre membre ;
- le caractère unique et constant du signifié.

Ainsi, à la différence de *garde-malade,* qui est un composé, *gardien d'asile* est une synapsie, ainsi que *gardien d'asile de nuit* (dont la décomposition synaptique en arbre binaire est ambiguë).

TERMINO acquiert dans un premier temps tous les syntagmes nominaux d'un texte à l'aide d'une analyse syntaxique des phrases. Ces syntagmes sont ensuite examinés pour en extraire les synapsies avec une grammaire dédiée opérant sur les catégories morphosyntaxiques des mots et sur les informations syntaxiques fournies (notamment les dépendances entre

déterminés et déterminants). Enfin, un deuxième jeu d'heuristiques est utilisé pour supprimer, le cas échéant, certains compléments non pertinents au sein de ces synapsies.

Dans le prototype **TERMS**, J. Justeson et S. Katz [Justeson & Katz, 1995] utilisent une technique symbolique proche pour l'anglais. Ils proposent d'extraire les composés à partir d'une expression régulière sur les étiquettes catégorielles des mots. Cette même approche de patrons catégoriels a également été utilisée par I. Dagan et K. Church pour construire le module d'extraction de candidats-termes de leur outil **TERMIGHT** [Dagan & Church, 1997]. L'expression caractérisant les termes peut également parfois porter sur d'autres types d'informations que les catégories. C'est le cas par exemple dans les travaux de [Heid *et al.*, 1996 ; 2000], qui proposent à l'utilisateur de spécifier lui-même l'expression recherchée grâce à leur système **CQP**, ou de [Voutilainen, 1993]. L'outil d'acquisition mis au point par ce dernier, **NPTOOL**, utilise une approche à bases de règles pour extraire des syntagmes nominaux. Ces règles s'appuient sur l'analyse morphologique et la description syntaxique des phrases obtenues grâce à une technique à base de grammaires à contraintes (écrites à la main). Les règles sont des expressions régulières portant sur ces informations, les syntagmes obtenus sont les séquences maximales répondant à ces règles.

LEXTER [Bourigault, 1992 ; 1994] extrait lui aussi des candidats-termes sur des corpus étiquetés morphosyntaxiquement (à chaque mot est assigné sa catégorie : nom, verbe, adjectif, *etc.)*. Il utilise pour ce faire une approche duale des précédentes puisque les termes sont définis *en négatif,* c'est-à-dire en spécifiant les catégories de mots ne pouvant pas entrer dans la composition d'un terme. On a donc là encore une approche à base de règles dont la plupart sont fixées sur des considérations linguistiques mais aussi, pour quelques-unes, générées au besoin à partir du corpus. C'est une approche similaire qui est employée dans **SYNTEX** [Bourigault & Fabre, 2000] qui étend la couverture des dépendances prises en compte par **LEXTER** aux syntagmes verbaux et adjectivaux. Ces travaux sur **LEXTER** sont aussi à rapprocher de ceux de J. Royauté [Royauté *et al.,* 1992] développés dans une perspective d'extraction de descripteurs textuels pour l'indexation de documents.

Ces différentes approches ont pour point commun d'exploiter une sorte de langage, pour caractériser ce qu'est ou n'est pas un terme; ce langage (noté L_C) est la plupart du temps défini par un ensemble de règles *G.* Elles peuvent donc se formaliser par des formules du type : $\bar{f}(C_D) = \overline{T}_{C_D} = \{(x_1 \ldots x_n) \in C_D \,|\, \inf((x_1 \ldots x_n)) \in L_C\}$, où *inf(S)* est la fonction donnant les

informations exploitées dans *G* (par exemple, les catégories des mots dans **TERMS**) d'une suite de mots *S*.

Par apprentissage

Dans ses travaux de thèse, É. Naulleau [Naulleau, 1997], [Naulleau, 1999] propose une approche originale pour extraire automatiquement d'un texte des syntagmes nominaux pertinents pour l'indexation de documents. Le principe de sa méthode est de généraliser par une technique d'apprentissage artificiel des *filtres* positifs (exemples de syntagmes intéressants) et négatifs (exemples de syntagmes non pertinents) fournis par l'utilisateur. Ces filtres et leurs généralisations exploitent des informations lexicales, morphologiques, catégorielles et sémantiques ajoutées au texte, et appliqués au texte, doivent ainsi permettre de proposer des syntagmes conformes aux filtres positifs et ne répondant pas au filtres négatifs.

La technique de généralisation des filtres est assez rudimentaire et très contrainte pour limiter les problèmes combinatoires. Elle se situe en effet dans un cadre propositionnel (les exemples sont décrits par des ensembles d'attributs-valeurs), et l'expressivité des généralisations est restreinte. L'absence de formalisation de l'espace de recherche des généralisations ainsi défini conduit à certaines redondances et à un coût calculatoire heureusement réduit par de fortes contraintes sur la forme de ces généralisations.

Cette relative faiblesse de la technique d'apprentissage employée se traduit par des résultats seulement moyens alors que le nombre d'exemples (filtres) positifs et négatifs utilisés est gigantesque : certaines expériences comptent en effet près de 20 000 filtres positifs et autant de négatifs. Ces derniers chiffres semblent donc interdire toute portabilité aisée de cette approche.

Ce type d'approche par apprentissage artificiel se formalise de la même façon que les autres approches symboliques : $\bar{f}(C_D) = \overline{T}_{C_D} = \{(x_1 \ldots x_n) \in C_D \,|\, \inf((x_1 \ldots x_n)) \in L_C\}$: à ceci près que les règles *G* définissant le langage L_C sont apprises à partir d'exemples (des mots de L_C) et non plus définies manuellement par un expert.

4.2.3. Approche mixte

Certains outils combinent les deux approches présentées précédemment pour en conjuguer les avantages respectifs. Cette combinaison peut être réalisée par une simple juxtaposition des deux techniques (soit structurelle puis numérique, soit l'inverse) ou bien par un couplage plus intime.

Ces techniques sont de ce fait plus difficiles à modéliser simplement à l'aide de nos notations. On peut néanmoins considérer que dans le premier cas, on se trouve dans un phénomène de conjonction de deux fonctions d'acquisition f_1 et f_2, soit : $\bar{f} = \bar{f}_1 \wedge \bar{f}_2$ avec $\bar{f}_1 : C_D \rightarrow \overline{T}_{C_D}^{f_1}$, $\bar{f}_2 : C_D \rightarrow \overline{T}_{C_D}^{f_2}$ et $\overline{T}_{C_D} = \overline{T}_{C_D}^{f_1} \cap \overline{T}_{C_D}^{f_2}$. Dans le second cas, il s'agit plutôt d'une composition des fonctions : $\bar{f} = \bar{f}_1 \circ \bar{f}_2$ avec $\bar{f}_2 : C_D \rightarrow E$, $\bar{f}_1 : E \rightarrow \overline{T}_{C_D}$ et $\bar{f}_1 \circ \bar{f}_2(x) = \bar{f}_1(\bar{f}_2(x))$, où *E* est un espace de représentation intermédiaire.

Approche structurelle suivie d'une approche statistique

ACABIT de B. Daille [Daille, 1994] est un outil d'acquisition terminologique sur corpus dont le processus se décompose en deux étapes :

- un repérage linguistique des termes à l'aide de règles simples (par exemple, la règle *N(Prep(Det)*)*N, ou N à Vinf)* appliquées par des transducteurs au corpus étiqueté ; des mécanismes de variation permettent aussi d'extraire des variantes *(cf.* section 2.1.2) des termes [Daille, 1999], [Daille, 2001)];
- un filtrage statistique des candidats-termes retenus à l'étape précédente.

Plusieurs indices statistiques usuels ont été testés pour cette deuxième phase, et leurs performances ont été comparées et rapportées dans [Daille, 1994]. Formellement, en utilisant les notations définies précédemment, l'approche se modélise de la façon suivante :

$$\bar{f}(C_D) = \overline{T}_{C_D} = \left\{ (x_1 \ldots x_n) \in C_D \middle| (cat((x_1 \ldots x_n))) \in L_C \wedge freq((x_1, \ldots, x_n)) > Seuil \right\},$$

où *cat(S)* indique la séquence d'étiquettes catégorielles correspondant à la séquence de mots *S.*

ANA (pour *Acquisition Naturelle Automatique*) est un système d'extraction de candidats-termes [Enguehard, 1992], [Enguehard & Pantera, 1995] reposant ouvertement sur une procédure excluant toute analyse linguistique. Il n'utilise en effet ni lexique, ni informations

syntaxiques, et se veut donc indépendant de la langue. La reconnaissance en corpus des termes est effectuée à l'aide d'une observation de répétitions de patrons, identifiés au départ à partir d'un petit ensemble de termes complexes, et d'un calcul d'égalités souples (basé sur la distance d'édition ou distance de Levenshtein) entre mots permettant ainsi de se passer de lemmatisation.

MANTEX [Oueslati, 1999] s'inscrit également dans une utilisation d'une approche numérique suivie d'une technique structurelle. Plus précisément, il emploie la méthode d'extraction des segments répétés présentée précédemment mais seules les séquences de mots ne contenant pas de mots grammaticaux, de ponctuations ou de verbes sont retenues. Il s'agit donc là d'un filtrage préalable sur des informations structurelles des segments candidats qui sont ensuite répertoriés, avec leur fréquence d'apparition, dans une liste proposée à un expert du domaine si cette fréquence dépasse 2. Cette légère amélioration de l'approche de L. Lebart et A. Salem a malheureusement le même défaut de reposer en grande partie sur la fréquence d'apparition, ce qui élimine toute possibilité de découverte de phénomènes rares dans le corpus.

Approche statistique suivie d'une approche structurelle

L'outil d'extraction **XTRACT** [Smadja, 1993a ; 1993b] emploie une technique inverse à celles que nous venons de voir. Il effectue en effet dans un premier temps un repérage statistique de mots cooccurrents puis un étiquetage de ces collocations grâce aux informations syntaxiques fournies par l'étiqueteur **CASS** [Abney, 1990]. Plus précisément, le processus comporte les trois étapes suivantes :

1. Extraction des collocations par une technique semblable à celle exposée en 2.2.1 ;
2. Expansion des collocations en répétant itérativement l'étape précédente avec les collocations déjà trouvées ;
3. Etiquetage des collocations repérées à l'aide des informations syntaxiques données par l'analyseur **CASS** et de patrons spécifiés par l'utilisateur (du type verbe-objet). Un couple dont les occurrences sont majoritairement dans une certaine relation (ce seuil est fixé à 80% dans **XTRACT**) est retenu comme représentant de cette relation.

Bien que **XTRACT** soit principalement un extracteur de collocations et non pas seulement de termes, son fonctionnement est prototypique des approches mêlant, dans cet ordre, les techniques statistiques aux techniques symboliques.

Imbrication complexe des approches structurelles et numériques

Développé en premier lieu dans un but d'indexation de documents, le système **CLARIT** [Evans & Zhai, 1996] cherche à acquérir des séquences de mots décrivant au mieux le contenu d'un document. À ce titre, il tente d'extraire des unités telles que des termes complexes et présente de nombreux points communs avec des techniques plus spécifiquement dédiées à l'acquisition de terminologie. Le principe de **CLARIT** est le suivant :

1. Extraction de tous les syntagmes nominaux et étiquetage catégoriel des constituants ;
2. Analyse en dépendances des syntagmes en s'appuyant sur les catégories des mots et sur les formes attestées trouvées par ailleurs dans le corpus ;
3. Génération des termes possibles à partir de l'analyse des syntagmes.

Durant l'étape 1, un processus itératif permet de repérer ce que les auteurs nomment des atomes lexicaux (comme *hot dog, part of* speech*)* en comparant les fréquences d'apparition de leurs constituants. De plus, seules certaines successions de catégories sont autorisées dans ces atomes lexicaux (*nom-nom*, *adjectif-nom*, *atome lexical-nom*, etc.). Dès qu'une paire est détectée comme atomique, elle est ensuite considérée comme un seul mot et le processus est relancé; cela permet de trouver des atomes lexicaux de taille de plus en plus importante. L'analyse en dépendances de l'étape 2 a pour but de regrouper deux à deux les mots adjacents d'un syntagme pour trouver la configuration la plus restrictive et informative au vu du corpus. Enfin, l'étape 3 génère, en se basant sur l'analyse du syntagme, les termes d'indexation répondant à certains schémas jugés intéressants par les auteurs dans leur cadre de recherche d'information. Les résultats obtenus sont de bonne qualité et évalués directement par rapport à leur besoin applicatif; ils montrent ainsi que les performances de leur système sont améliorés par l'emploi de ces termes d'indexation complexes.

Ce mélange entre nature structurelle et numérique des données textuelles se retrouve dans les travaux de K. Church [Church, 1988]. Ce dernier décrit une méthode permettant de trouver les frontières des syntagmes nominaux à partir de textes en utilisant une technique similaire à celle employée pour construire des étiqueteurs morphosyntaxiques stochastiques. Ces travaux sont à rapprocher de ceux de L. Ramshaw et M. Marcus [Ramshaw & Marcus 1995], qui se proposent également de faire le repérage de syntagmes nominaux en le ramenant à un exercice d'étiquetage. Cependant, la technique utilisée dans leur cas s'appuie sur l'utilisation des règles de transformation comme cela est fait dans l'étiqueteur de E. Brill

[Brill, 1992 ; 1994]. Les travaux de K. Frantzi *et al.* [Frantzi *et al.,* 1996 ; 2000] s'inscrivent aussi dans ce type d'approche hybride puisque leur outil d'acquisition de termes exploite à la fois des informations structurelles et numériques sur le corpus.

5. RELATIONS SEMANTIQUES

Un lexique sémantique est composé d'unités lexicales, dont nous venons de présenter les principales approches d'extraction, mais aussi de relations entre ces unités. Celles-ci structurent le lexique [Czap & Nedobity, 1990], [Skuce & Meyer, 1991] et exhibent les liens sémantiques entre les différentes unités lexicales. Ces relations sémantiques permettent donc également d'accéder au sens d'une unité lexicale en la comparant, à travers ces relations, à d'autres unités [Cruse, 1986].

Nous proposons ci-dessous de définir les relations sémantiques et leurs propriétés, linguistiquement dans un premier temps, puis plus formellement. Nous examinons ensuite les différentes approches existantes utilisées pour l'aide à l'acquisition ou l'acquisition automatique de ces relations, en nous attachant notamment à mettre en valeur leurs avantages et défauts.

5.1. Définition des relations sémantiques

Les relations sémantiques, et plus particulièrement certaines d'entre elles, ont été largement étudiées, tant d'un point de vue formel que dans une optique de linguistique de corpus. Nous en proposons ci-dessous une définition usuelle, mettant en particulier en exergue la différence entre relations syntagmatiques et paradigmatiques [Cruse, 1986]. Nous présentons ensuite, à travers la notion mathématique de relations entre ensembles, le formalisme permettant de modéliser les relations sémantiques et leurs propriétés.

5.1.1. Types de relations sémantiques

II existe plusieurs types de relations sémantiques, offrant ainsi différents liens pour structurer les bases de connaissances lexicales. On distingue en particulier les relations sémantiques portées dans le texte par un prédicat (celui-ci pouvant être explicite ou implicite) et celles dont le prédicat n'est pas exprimé [Cruse, 1986]. Le premier type de lien, reposant sur les propriétés syntaxiques des constituants du couple en relation ou de leur contexte, est qualifié de syntagmatique. Dans le second cas, on parlera plutôt de relations paradigmatiques.

Relations syntagmatiques

Au sein des textes, certaines formes syntaxiques suggèrent l'existence d'un lien sémantique entre deux mots. Ainsi, les verbes accompagnés des éléments de leurs structures argumentales peuvent être considérés comme des prédicats dotés d'arguments dénotant donc d'une relation particulière. D'autres structures syntaxiques sont également l'indice de relations syntagmatiques. Par exemple, une phrase contenant *l'effet de X sur Y* indique clairement une relation entre les entités *X* et *Y*. On retrouve ce même lien dans des variantes syntaxiques impliquant des verbes support *(X a un effet sur Y)*, ou d'autres schémas syntaxiques. Il y a ainsi plusieurs patrons équivalents pour dénoter une même relation prédicative entre mots.

Dans cette optique, H. Robison [Robison, 1970] a étudié près de 8 000 relations syntagmatiques explicites (qu'il appelle patrons primaires) et leurs variantes (patrons secondaires). Les patrons primaires indiquent les prédicats reliant les mots considérés mais peuvent être ambigus ; l'étude du patron secondaire sert alors à lever l'ambiguïté sur le sens de ce prédicat.

Le prédicat reliant plusieurs mots n'apparaît pas toujours aussi clairement dans un texte. Une relation sémantique peut par exemple s'établir entre les constituants d'un composé multinominal. L'explicitation du prédicat requiert alors une analyse fine des constituants du composé.

C'est dans ce cadre que se placent les travaux de C. Fabre [Fabre, 1996], [Fabre & Sébillot, 1999] qui a développé une technique permettant l'interprétation automatique des composés de la forme *nom-nom* pour l'anglais et *nom-préposition-nom* pour le français. Plus précisément, l'analyse du composé doit fournir le prédicat et les rôles sémantiques de chacun des constituants. Dans le cas le plus simple, le prédicat apparaît sous forme d'un déverbal comme dans *tondeuse à gazon, filtre à air* ou *détecteur de choc*. Pour ce dernier exemple, l'analyse produit alors une représentation de la relation de ce type : détecter (instrument : *détecteur,* objet : *choc).* Si le prédicat n'est pas directement accessible, il faut alors disposer d'une connaissance sémantique supplémentaire pour pouvoir interpréter le composé. Considérons par exemple le composé *bread knife* ; l'interprétation naturelle que l'on voudrait lui associer est cut (instrument : *knife,* objet : *bread).* Cela n'est possible que si l'on sait que la fonction typique de *knife* est *to cut.* C. Fabre propose d'utiliser les connaissances sémantiques

codées dans le Lexique génératif [Pustejovsky, 1995] pour avoir accès à ces informations prédicatives implicites.

Relations paradigmatiques

Certaines relations sémantiques entre mots n'apparaissent pas sous forme de lien syntaxique standard au sein des textes, mais relèvent d'une association paradigmatique [Grefenstette, 1994a] (comme la *synonymie* par exemple). Dans ces relations, la notion précédente de prédicat n'est pas pertinente ; elles sont donc parfois également appelées relations non prédicatives [Morin, 1999].

A. Cruse [Cruse, 1986] s'appuie sur une interprétation ensembliste pour définir des relations primaires qu'il nomme *congruences*, permettant de caractériser certaines relations paradigmatiques. Ainsi, la *synonymie* correspond à la relation d'identité entre ensembles de mots ; *l'hyponymie* correspond quant à elle à une relation d'inclusion.

Il existe bien sûr beaucoup de relations paradigmatiques, opérant essentiellement sur des unités de même catégorie (de *nom* à *nom* par exemple). Le modèle de A. Cruse permet de les représenter en combinant les opérations ensemblistes ; il les appelle *variantes de congruence.* Parmi ces dernières, notons que les *quasi-relations* permettent de modéliser des relations entre unités lexicales de catégories différentes. Ainsi, le participe passé *coloré* est *quasi hyperonyme* de *rouge* et *jaune.*

Les relations paradigmatiques induisent le plus souvent une structure particulière sur l'espace des unités sémantiques [Cruse, 1986]. Ainsi, une relation hiérarchique telle que *l'hyperonymie*, *l'hyponymie*, la *méronymie* (relation partie-tout) sur des termes impose une représentation sous forme de taxinomie des unités lexicales. On parle parfois pour ce type de relation de liens verticaux. Ces relations hiérarchiques sont d'ailleurs parfois considérées comme les plus intéressantes [Condamines & Amsili, 1993]. Elles permettent en effet, grâce à l'organisation arborescente qu'elles induisent, de refléter l'organisation hiérarchique des concepts portés par les unités sémantiques que l'on retrouve dans les ontologies [Rastier, 1995]. D'autres relations sont au contraire *symétriques* (ou *quasi-symétriques*), comme par exemple la *synonymie* et *l'antonymie*. Les liens qu'elles permettent d'établir dans l'espace des unités sémantiques sont alors qualifiés de *transversaux* ou *d'horizontaux*.

L'acquisition de ce type de relations est difficile : il faut non seulement les détecter, mais aussi les identifier, c'est-à-dire préciser le lien sémantique existant entre les divers constituants. Comme nous le montrons ci-après en section 3.2, un tel travail peut soit exploiter les similitudes de contextes de certains mots, soit recourir à des patrons lexicaux, morphosyntaxiques ou sémantiques.

5.1.2. Représentation formelle d'une relation

On peut représenter la notion de relation, notamment de relation sémantique, par un formalisme issu des mathématiques. Nous proposons ci-dessous une définition abstraite de cet objet mathématique.

Définition 1 : *Soient A et B deux ensembles. On appelle relation sur* $A \times B$ *tout prédicat défini sur* $A \times B$. *Si* $(x, y) \in A \times B$ *et si R est un prédicat à valeur dans* $A \times B$, *on notera R(x, y) ou plus simplement xRy pour signifier que x est en relation R, avec y.*

Une relation définit donc un sous-ensemble de $A \times B$. Dans le cas où l'on s'intéresse aux relations entre termes d'un domaine, les ensembles *A* et *B* sont égaux à l'ensemble des termes *T*. La relation *R*, se définit alors sur T^2 ; une telle relation sur T^2 est dite binaire.

Une relation sur $A \times B$ définit un graphe inclus dans $A \times B$, formellement présenté dans la définition 2.

Définition 2 : *On appelle graphe d'une relation R le sous-ensemble de* $A \times B$, *déterminé par :*

$$G_R = \{(x, y) \in A \times B | xRy\}$$

C'est-à-dire l'ensemble des couples dont les composantes sont en relation.

Il est important de noter que se donner un sous-ensemble de $A \times B$ revient exactement à définir une relation. Ainsi, dans le cadre des relations sémantiques, se fixer le graphe d'une relation correspond à définir cette relation en extension (c'est-à-dire uniquement par l'exemple) sans avoir besoin d'expliciter le lien ciblé (ce qui correspondrait à la définir en intention).

On définit ci-dessous quelques propriétés classiques des relations appliquées à notre cadre de relations sémantiques.

Propriété 1 : *Soient R une relation entre unités sémantiques et T l'ensemble des unités sémantiques ; on définit les propriétés suivantes :*

- *R est symétrique si* $\forall(x, y) \in T^2, xRy \Rightarrow yRx$.
- *R est antisymétrique si* $\forall(x, y) \in T^2, xRy \wedge yRx \Rightarrow x = y$.
- *R est réflexive si* $\forall x \in T, xRx$.
- *R est antiréflexive si* $\forall x \in T, \neg(xRx)$.
- *R est transitive si* $\forall(x, y, z) \in T^3, xRy \wedge yRz \Rightarrow xRz$.

Parmi la variété des relations possibles, deux types de relations particulières sont très utilisés :

- *R* est une relation *d'équivalence* si elle est *réflexive*, *symétrique* et *transitive*;
- *R*. est une relation *d'ordre*, si elle est réflexive, *antisymétrique* et *transitive*.

Lorsque *R*, est une relation *d'équivalence*, on définit par ailleurs les *classes d'équivalence* de *x modulo R* par: $\bar{x} = \{x \in T | xRy\} = \{y \in T | yRx\}$.Par exemple, si la relation *R* représente la synonymie, $\bar{x}$ est l'ensemble des synonymes d'un terme *x*.

On définit également, pour une relation *binaire d'équivalence*, les *ensembles quotients T/R*, avec *P*(*T*) l'ensemble des parties de *T* par: $T / R = \{C \in P(T) | \exists x \in T telque C = \bar{x}\}$. Si *R* désigne comme dans l'exemple précédent la synonymie, les *ensembles quotients* de *R* représentent une partition de *T* en classes (*clusters*) de termes synonymes.

Les relations hiérarchiques comme *l'hyperonymie* ou la *méronymie* correspondent dans ce cadre plutôt à des relations *d'ordre* (*partiel*) sur *T*. En effet, si *R* est une relation *d'hyperonymie*, en supposant que l'on ait *végétal R arbre* (un arbre est une sorte de végétal) et *arbre R chêne* (un chêne est une sorte d'arbre), alors on sait que *végétal R chêne* (*transitivité*), que l'on n'a pas *arbre R végétal* (*antisymétrie*) et que *arbre R arbre* (*réflexivité*).

5.2. Acquisition de relations sémantiques

Nous présentons dans cette section les différentes familles de travaux portant sur l'acquisition de relations sémantiques sur corpus, en nous efforçant comme précédemment d'utiliser un cadre formel commun à leur description. Nous reprenons également l'approche utilisée en section 2.2, en distinguant les méthodes exploitant l'aspect fréquentiel du corpus et celles s'appuyant sur des indices structurels pour détecter les relations sémantiques.

5.2.1. Approche numérique

Beaucoup de travaux exploitant des techniques statistiques ont été menés en acquisition de relations sémantiques (pour une vue de domaine, voir [Grefenstette, 1994b], [Pichon & Sébillot, 1997], [Habert *et al.,* 1997]). L'idée sur laquelle reposent ces méthodes est de détecter les associations statistiquement significatives, c'est-à-dire plus fréquentes que du fait du hasard. Ces associations permettent soit de mettre au jour directement des mots en relation sémantique (c'est l'approche statistique décrite ci-dessous), soit d'étudier les mots partageant les mêmes associations (cette approche est alors qualifiée de distributionnelle).

Statistique

Comme nous l'avons souligné en section 4.2.1, certains indices statistiques d'association sont parfois utilisés pour la détection non plus de termes complexes mais de relations sémantiques. Le principe méthodologique est le même que celui précédemment exposé, à ceci près que les fenêtres utilisées pour calculer les cooccurrences sont de tailles souvent plus importantes. Les relations mises au jour par ce type de méthodes sont généralement syntagmatiques ; on les note R_S ci-après.

Par ce type d'approche, on estime donc la relation R_S par $\overline{R}_S$ que l'on définit par $x\overline{R}_S y$ si $f(p(x), p(y), p(x,y),...) > Seuil$; ces probabilités sont elles-mêmes estimées à l'aide des fréquences d'apparition dans le corpus :*p(x) = freq(x)*. On fait ensuite le postulat que $\forall (x,y) \in T^2, x\overline{R}_S y \Rightarrow xR_S y$. Il faut noter que suivant les indices statistiques utilisés, la relation R_S sera symétrique ou non. Ce type d'approche produit généralement des résultats bruités et hétérogènes puisqu'il ne permet pas de typer les relations obtenues.

Analyse distributionnelle

A partir de l'hypothèse harrissienne [Harris *et al,* 1989] selon laquelle une analyse distributionnelle des *propriétés contextuelles* (nous précisons ci-après ce que ces propriétés peuvent être) des mots fait apparaître des classes de concepts (regroupant les mots partageant les mêmes propriétés) et des relations entre elles, beaucoup de travaux ont vu le jour (voir [Grefenstette, 1994b]). Ces travaux s'attachent donc à faire ressortir des textes des relations dites paradigmatiques (que l'on note R_P par opposition aux relations syntagmatiques précédentes). La plupart de ces méthodes s'appuient sur une procédure en trois temps :

1. recherche des propriétés contextuelles de chaque mot du corpus ;
2. mise en relation deux à deux de mots partageant les mêmes propriétés contextuelles;
3. construction de classes à partir des relations découvertes à l'étape 2.

Cette définition volontairement large ne précise ni les propriétés contextuelles étudiées, ni la façon dont elles sont recherchées, ni la mise en relation des mots partageant les mêmes propriétés, ni enfin ce qui est précisément entendu par la construction de classes.

Concernant les deux premiers points, les propriétés considérées varient selon les travaux. Ce peut être le contexte syntaxique des mots [Faure & Nédellec, 1999] ou les relations de dépendance *tête-expansion* au sein de syntagmes nominaux [Bouaud *et al.,* 1997], [Habert &: Fabre, 1999] ou de tout syntagme [Bourigault, 2002], ou d'une relation quelconque spécifiée par l'utilisateur [Grefenstette, 1992]. Cela peut également être les mots cooccurrant dans une certaine fenêtre [Grefenstette, 1994b], [Pichon & Sébillot, 1999], [Rossignol & Sébillot, 2002] ou les segments répétés [Rousselot *et al.,* 1996] ou encore les mots d'un même domaine sémantique [Chalendar & Grau, 2000], [Chalendar, 2001].

L'appariement deux à deux des mots suivant leurs propriétés partagées (et aussi celles non partagées) est une phase complexe influençant ensuite la phase de construction de classes homogènes. En effet, si l'on note *Prop(x)* l'ensemble des propriétés considérées du mot *x* (par exemple l'ensemble des dépendances syntaxiques dans lesquelles il apparaît), alors on définit la relation R_P sur base d'analyse distributionnelle par la formule: $xR_P y \Leftrightarrow \left(prop(x) = prop(y)\right)$ Ainsi définie, la relation trouvée par analyse distributionnelle serait une relation d'équivalence (c'est-à-dire symétrique, réflexive et transitive). La phase 3 consiste donc à construire les classes d'équivalence et à produire ainsi l'ensemble quotient censé représenter la partition de l'espace des termes en classes conceptuelles.

Malheureusement, la définition précédente, trop contraignante sur le partage complet des propriétés contextuelles, n'est pas celle utilisée en pratique car elle aurait pour effet de n'assembler que très peu de termes. Généralement, on emploie une définition plus *lâche* indiquant qu'une *majorité* des propriétés contextuelles doit être partagée pour que deux mots soient déclarés en relation ; cela peut se traduire plus formellement par $xR_P y \Leftrightarrow \left|prop(x) \cap prop(y)\right| > Seuil$ avec $|...|$ indiquant le cardinal d'un ensemble. Cette définition plus souple a bien sûr pour effet de mettre en relation plus de mots mais la

transitivité de la relation est perdue *(i.e.* on peut avoir $xR_P y$ et $xR_P z$ et pas $xR_P z$*)*. Le regroupement en classes ne peut donc plus se faire par *classes d'équivalence* ; c'est pourquoi ont été proposées des structures plus faibles que ces dernières pour modéliser les *classes conceptuelles* attendues (*cliques*, *composantes connexes*, *etc.)*.

Chacune de ces formes de regroupements semble faire ressortir des informations de nature différente [Bouaud *et al,* 1997], [Bourigault, 2002] mais aucune ne permet de n'obtenir que des classes homogènes ni d'isoler un type de relation sémantique fixé. L'interprétation des classes obtenues est dans la plupart des cas laissée au soin de l'utilisateur. Notons enfin que les différents regroupements obtenus peuvent également se faire sous forme d'arbres ou de pyramides à l'aide de techniques de classification hiérarchique [Assadi, 1998], [Rossignol & Sébillot, 2002], [Agarwal, 1995], [Faure & Nédellec, 1999] : la racine contient une unique classe conceptuelle (certainement très hétérogène) et à l'inverse les feuilles sont des classes très spécialisées.

5.2.2. Approche symbolique

Par opposition aux techniques numériques précédentes, nous présentons ci-dessous des approches symboliques d'acquisition de relations sémantiques. Ces approches peuvent elles-mêmes se classer en deux grandes familles : les approches linguistiques, où les indices structurels exploités sont donnés *a priori* (par une analyse linguistique par exemple) et les approches basées sur une notion d'apprentissage (artificiel ou non). Il faut noter, comme pour l'acquisition d'éléments terminologiques, que ce type d'approche travaille le plus souvent au niveau des occurrences des couples de mots en relation. Chaque occurrence est donc individuellement classée comme porteuse d'une relation; on note $\mathrm{occ}_i(x, y)$ la i^e occurrence d'un couple (*x*,*y)* dans un corpus.

Approche linguistique

Le système d'acquisition **SEEK** [Jouis, 1995] est prototypique des systèmes fondés sur une expertise linguistique. Il fonctionne à partir d'un corpus lemmatisé et nécessite une grande intervention humaine. À l'aide de règles dites *d'exploration contextuelle* [Jouis, 1993], le système détecte des couples de mots en relation binaire. Ces relations sont variées (plus d'une vingtaine au total), elles représentent par exemple la notion d'inclusion, d'identification, d'appartenance, de localisation ou de tout à partie, mais statiques. Les couples détectés sont ensuite proposés à l'utilisateur. Ce dernier les rejette ou les valide comme représentant de la

relation proposée ; dans ce dernier cas, il doit manuellement indiquer quels sont les arguments de cette relation. Le système propose enfin un graphe représentant l'ensemble des relations ainsi acquises.

Les règles *d'exploration contextuelles* reposent sur l'identification de marqueurs linguistiques (principalement lexicaux) et sont construites manuellement. Ce sont ainsi plus de 220 règles de la forme SI <*condition de co-présence de marqueurs linguistiques*> ALORS <*actions*> OU <*conclusions*>, manipulant plus de 3 300 marqueurs linguistiques qui sont utilisées par **SEEK**.

Plus récemment, D. Garcia propose à travers son système **COATIS** [Garcia, 1998], [Garcia *et al.,* 2000] une approche similaire mais en se focalisant sur la relation de causalité. Avec **ATERM**, R. Oueslati [Oueslati, 1999] propose également une technique proche pour la détection de couples en relations sémantiques. Seul le degré d'interactivité de son outil le différencie de **SEEK** ou **COATIS**. En effet, **ATERM** est une interface permettant au linguiste de spécifier à tout moment les patrons qu'il souhaite voir utilisés. Ces patrons sont exprimés dans un langage dédié, **LEXICA**, manipulant les lemmes, catégories ou formes fléchies des mots du corpus. Enfin, les travaux de C. Jacquemin [Jacquemin, 1996 ; 1997 ; 2001] sur **FASTER** peuvent également s'inscrire dans cette approche. Ce dernier acquiert des variantes morphosyntaxiques de termes (la variation peut donc être vue comme une relation d'équivalence) à l'aide de plusieurs niveaux de règles (les règles opérant sur d'autres règles sont appelées méta-règles).

Dans l'ensemble de ces travaux, on estime la relation R par $\hat{R}$ définie comme un ensemble de règles données par un expert. Plus généralement, $\hat{R}$ est souvent la réunion de n règles $R_1,...,R_n$, définissant chacune un aspect de la relation et manipulant des indices pouvant être de différente nature (marqueurs lexicaux, catégories morphosyntaxiques, etc.). Ainsi, un couple de mots (x,y) est considéré en relation si une (ou un nombre minimal) de ses occurrences répond à une des règles définies. En notant Desc(o) la description d'une occurrence o (par exemple la séquence d'étiquettes catégorielles de cette occurrence), cela peut se transcrire par :

$$x\hat{R}y \Leftrightarrow \exists i Desc\left(occ_i(x,y)\right) \in L_R \, avec L_R = \bigcup_{1 \le l \le n} L_{R_l}$$

II faut remarquer par ailleurs que suivant la relation étudiée, *R*, n'est pas forcément symétrique.

Le postulat de base de ces outils est de supposer que les relations statiques décrites sont suffisamment génériques pour ne pas dépendre d'un domaine en particulier. Malheureusement, les expériences rapportées dans [Jouis *et al.*, 1997] montrent que les résultats d'extraction sur un plus gros volume de textes sont entachés de bruit (relations détectées et non pertinentes). Il semble donc que certains de ces travaux soient en réalité difficilement portables d'un domaine à un autre sans opérer de lourdes modifications manuelles dans la base de règles. De la même façon, l'ajout d'un nouveau type de relation nécessite de découvrir et d'insérer dans ce type de système complexe de nouvelles règles contextuelles décrivant cette relation. La portabilité et l'utilisation à grande échelle sur des textes variés de ces techniques semblent donc très problématiques et coûteuses. Enfin, dans ces travaux, aucune discussion n'est faite de l'expressivité des règles à base de marqueurs lexicaux et de leur pouvoir de représentation des relations.

Approche par apprentissage symbolique

L'acquisition de relations sémantiques par apprentissage de patrons d'extraction lexico-syntaxiques (ou LSPE en anglais pour *lexico-syntactic pattern extraction)* est l'une des techniques les plus connues. L'idée principale de cette approche est d'identifier dans un corpus les marqueurs ou indices d'une relation sémantique sur un petit ensemble d'exemples pour ensuite les réutiliser pour extraire de nouvelles unités en relation. Cette approche a été initiée par M. Hearst [Hearst, 1992 ; 1998] et formalisée en cinq étapes :

1. choisir une relation cible *R* ;
2. réunir une liste de paires en relation *R*. (par exemple les extraire d'un thésaurus, d'une base de connaissances) ;
3. retrouver les phrases du corpus contenant ces paires et enregistrer leurs contextes lexical et syntaxique ;
4. trouver les points communs entre ces contextes et supposer que cela forme un schéma lexico-syntaxique de *R* ;
5. appliquer les schémas pour obtenir de nouvelles paires et retourner en 3.

Elle se différencie de la méthode exposée précédemment par le fait que les marqueurs de la relation (ici les informations lexicales et catégorielles) sont issus d'une analyse d'exemples et non plus d'une connaissance linguistique *a priori.*

La relation *R*. n'est donc pas définie explicitement au départ mais par la donnée d'exemples. Cela permet de s'intéresser à des relations connues partiellement en extension (par des instances) mais pas en intention (non formalisée par une règle). L'algorithme, et en particulier la phase 4 qui reste la plus vague, propose d'estimer une définition $\overline{R}$ explicite de *R*. en généralisant les exemples. On fait ensuite le postulat que $\forall(x, y) \in T^2, x\overline{R}y \Rightarrow xRy$.

Cette technique a été utilisée avec succès pour la relation d'hyponymie. Pour cela, M. Hearst s'est servi de **WORDNET** pour générer une liste de paires candidates en relation d'hyponymie (étape 2). En revanche, l'étude d'autres types de relations (comme la méronymie) semble avoir donné de moins bons résultats du fait de l'obtention de patrons trop généraux.

La phase 4 est entièrement manuelle dans [Hearst, 1992] ; la généralisation des structures lexico-syntaxiques des phrases en patrons d'extraction est donc faite par l'utilisateur. M. Hearst suggère une automatisation de cette étape [Hearst, 1998] par différentes techniques dont l'apprentissage artificiel mais qui ne sont pas mises en œuvre. C'est cette phase de généralisation que E. Morin se propose d'automatiser [Morin, 1999 ; 1997] avec son système **PROMETHEE**. Pour ce faire, il s'appuie sur un calcul de similarité [Morin, 1998] deux à deux entre les contextes lexico-syntaxiques de deux occurrences de paires.

Dans son logiciel **CAMELEON** [Séguéla & Aussenac-Gilles, 1999], [Séguéla, 2001], P. Séguéla propose une approche à l'intersection de celles de M. Hearst et C. Jouis puisqu'il suggère de réutiliser, en les adaptant si besoin, certains marqueurs de relations dits génériques, et d'acquérir d'autres marqueurs spécifiques au corpus étudié. Même si l'idée de la réutilisabilité des patrons et de leur adaptation est intéressante, ce travail souffre du même manque d'automaticité que la méthode de M. Hearst. L'utilisateur intervient en effet de façon centrale dans le processus : c'est lui qui examine la pertinence des marqueurs génériques pour le corpus, et au besoin c'est lui qui les adapte ; enfin, c'est également lui qui doit construire l'ensemble des marqueurs spécifiques. Ces travaux s'inscrivent de ce fait plutôt dans une approche d'aide à l'extraction de relations que d'une extraction entièrement automatique.

Il faut noter que le découpage en cinq étapes proposé par M. Hearst sur lequel repose tous ces travaux est en réalité le processus habituel d'un apprentissage supervisé *(i.e.* à partir d'exemples) : trouver un concept à apprendre, trouver un ensemble d'exemples répondant à ce concept, décrire les exemples par des attributs, inférer une définition du concept à l'aide des exemples et de leur attributs, trouver d'autres objets répondant à la définition du concept. La relation *R,* à apprendre est alors estimée par un classifieur. Ce dernier est généré par inférence à partir d'exemples. Dans le cas de M. Hearst ou de **CAMELEON** cette inférence est manuelle alors qu'elle est automatisée (quoique rudimentaire) chez E. Morin.

Ce type d'approche par apprentissage se formalise donc de la même manière que précédemment : chaque occurrence de couple (et son contexte) est examinée pour constater si elle répond ou non à l'une des règles. On a donc encore pour un couple *(x, y)* et avec les mêmes notations que précédemment:

$$x\hat{R}y \Leftrightarrow \exists i Desc(occ_i(x,y)) \in L_R \, avec L_R = \bigcup_{1 \le l \le n} L_{R_l} \,,.$$

La différence avec l'approche précédente est que ces règles sont inconnues *à priori* (ou du moins certaines) et dérivées d'exemples de couples dont les constituants sont en relation. Cette technique d'acquisition permet donc d'apprendre de nouvelles règles pour tout nouveau corpus qui sont pertinentes car adaptées au corpus.

6. SYNTHESE

Nous l'avons vu, de nombreuses techniques d'acquisition d'informations lexicales sémantiques existent et reposent sur des approches très diverses. Certaines de ces approches sont d'ailleurs également utilisées pour extraire sur corpus d'autres sortes d'informations que les unités sémantiques et les relations les structurant. Ainsi, le repérage d'entités nommées, de dates, de lieux, etc., fait aussi l'objet de travaux d'acquisition relevant du domaine de l'extraction d'informations.

Les travaux d'acquisition que nous avons présentés, aussi bien de relations sémantiques que d'unités lexicales, peuvent se grouper (approximativement) selon deux familles : les méthodes exploitant l'aspect numérique des données textuelles et celles exploitant leur aspect structurel.

Les avantages des approches numériques sont principalement leur automaticité et leur portabilité : elles sont relativement faciles à mettre en place car elles ne nécessitent souvent aucune donnée autre que le corpus. Elles sont de ce fait adaptées aux traitements de nouveaux textes, et les résultats produits sont propres au domaine étudié. À l'inverse, les techniques d'acquisition symboliques nécessitent des connaissances supplémentaires pour fonctionner. Elles doivent par exemple disposer en plus du corpus d'un ensemble de patrons d'extraction, ou bien, si ces patrons sont appris automatiquement, d'exemples d'apprentissage. La constitution de ces données supplémentaires demande donc un investissement humain qu'il est le plus souvent nécessaire de reproduire à chaque nouveau domaine que l'on souhaite traiter.

En retour, les méthodes numériques souffrent d'un manque d'interprétabilité. Il est en effet souvent difficile de comprendre pourquoi un certain élément sémantique a été retenu et pas un autre, le seul indice fourni à ce sujet étant généralement un score statistique. La détection se passe au niveau du corpus et non pas de l'occurrence; il n'est donc pas possible de « revenir » au texte pour l'expliquer. C'est également pour cette raison que ces méthodes produisent parfois des résultats très hétérogènes, la seule nature fréquentielle des objets linguistiques n'étant pas toujours à même de les différencier. Enfin, toujours à cause du manque d'interprétabilité du processus, ces méthodes n'offrent aucun retour sur la définition de l'information recherchée. Elles ne permettent pas d'en avoir une compréhension plus précise ou d'en donner une définition opérationnelle. Les approches structurelles sont en revanche plus facilement interprétables à plusieurs titres. La granularité de détection des éléments sémantiques intéressants est très fine puisqu'elle s'opère au niveau de l'occurrence. Cela permet de comprendre plus naturellement pourquoi une information sémantique a été retenue ou non. Par ailleurs, les règles ou patrons employés au sein de ces techniques permettent également de définir de manière très pragmatique l'information recherchée. Cette définition peut s'avérer intéressante lorsque les éléments que l'on tente d'acquérir, et plus précisément leurs réalisations en corpus, sont mal connus. Cette dernière propriété d'interprétabilité est encore plus intéressante lorsque les patrons sont appris automatiquement à partir d'exemples dans le corpus, comme dans l'approche par apprentissage présentée précédemment.

7. RESUME

Dans ce chapitre, nous avons présenté les différents travaux réalisés dans le domaine de l'acquisition de connaissances textuelles sur corpus. Cette tâche se décline principalement en

deux étapes; l'extraction de termes et de relations entre termes. Nous avons ainsi présenté séparément les méthodes se rapportant à l'extraction de termes, ensuite celles concernées par l'extraction de relations qu'ils maintiennent. Nous avons vu que ces méthodes émanent de deux courants théoriques principaux; le courant statistique mettant en œuvre des données numérique pour l'extraction et le courant linguistique se basant sur des connaissances linguistiques pour effectuer l'extraction. A la fin du chapitre nous avons exposé une comparaison entre ces deux théorie à travers de laquelle nous avons pesé les avantages et inconvénients de chacune d'elles.

La partie suivante présente plus avant les cadres méthodologiques, techniques et applicatifs dans lequel se place le développement de notre outil. Les différents choix présidant au fonctionnement D'**ONTOLOGOS** seront notamment explicités et motivés.

PARTIE II

NOTRE MODELE DE CONSTRUCTION

« Ne craignez pas la perfection, vous n'y parviendrez jamais »
SALVADOR DALI

CHAPITRE 3

METHODOLOGIE DE CONSTRUCTION

« Il n'y a pas une méthode unique pour étudier les choses. ».

ARISTOTE

Conçues comme une réponse à des problèmes posés par l'intégration de connaissances au sein des systèmes informatiques, les ontologies apparaissent désormais comme une clé pour la manipulation automatique de l'information. Cette omniprésence des ontologies est entravée par le coût, en termes de temps et efforts, très élevé de leur développement, un développement qui se fait le plus souvent manuellement, et ce malgré la présence de nombreux principes et critères pour guider la construction.

Par conséquent, le besoin pour des outils et méthodes de construction automatique d'ontologies se fait de plus en plus ressentir. Le texte apparaît comme une source privilégiée pour démarrer cette construction. Ainsi plusieurs tentatives de construction (semi) automatique ont été initiées. Dans ces travaux, la construction se limitait à acquérir les termes du domaine et les relations qui les réunissent afin de les structurer sous forme d'une hiérarchie. En parallèle à ces efforts, d'autres travaux se sont concentrés sur la réutilisation des ontologies en essayant de développer des méthodes adéquates pour l'alignement et la fusion des ontologies à réutiliser.

Par l'analyse de ces différents travaux, il nous a apparu qu'il existe un vide méthodologique dans la construction de ces ontologies. En effet, malgré l'intérêt que représente la réutilisation pour de démarrer ou raffiner le processus de d'acquisition à partir de textes, aucun effort n'ait été fait pour intégrer ces deux approches de construction. Ainsi notre travail vient pour combler le gap entre l'acquisition à partir de corpus d'une part et la réutilisation d'ontologies d'une autre part, en proposant une méthodologie de construction (semi) automatique qui s'effectue en deux passes : acquisition et réutilisation.

L’objectif de ce chapitre est de présenter en détail les choix et les techniques présidant au développent de notre méthodologies de construction. Nous détaillons dans la première partie grands axes de notre méthode de construction et nous expliquons l’architecture fonctionnelle d’un système développé à la lumière de cette méthodologie. La deuxième partie sera consacrée à la présentation des techniques et algorithmes que nous avons développé pour soutenir ce travail de construction.

1. METHODOLOGIE DE CONSTRUCTION

La construction d'ontologies est une tâche difficile, complexe et nécessitant, à l'instar de tout autre travail de conception, la présence d'une méthodologie claire et bien définie pour guider le concepteur durant le processus de construction. La méthodologie que nous proposons, s'articule sur l'utilisation conjointe d'ontologies prédéfinies (nous utilisons le terme ontologie par abus de langage pour désigner toute structure terminologique susceptible de démarrer le processus de construction) et de corpus textuels et plaçons le concepteur au centre du processus de construction selon des transactions de coopération bien établies [Allalga & Sellami, 2005]. On dissocie la tâche de conception en quatre phases (voir Figure 8) :

- Importation et fusionnement d'ontologies : la réutilisation de ressources existantes est une technique qui permet d'éviter la construction *ex nihilo* des ontologies. Cela est fait à travers une opération d'importation puis de fusionnement des ontologies sélectionnées.
- Traitement de corpus : le traitement de corpus permet d'obtenir une représentation des connaissances qu'il contient, sous la forme d'une ontologie de domaine.
- Intégration des structures taxonomiques extraites du corpus dans la taxonomie établie dans la première phase : cela dans le but d'aboutir une structure commune regroupant les deux sources d'ontologies utilisées.
- Evaluation de l'ontologie finale: l'évaluation consiste à vérifier s'il n'y a pas de cycle, s'il n'y a pas redondance de concepts ou de relations et si chaque hiérarchie est bien connexe.

Il est à noter que ces phases de construction constituent un cycle qui peut être réitéré pour des fins de maintenance de l'ontologie, ce qui est essentiel pour sa validité vis-à-vis des objectifs de constructions et pour tenir compte des changements éventuels dans son domaine.

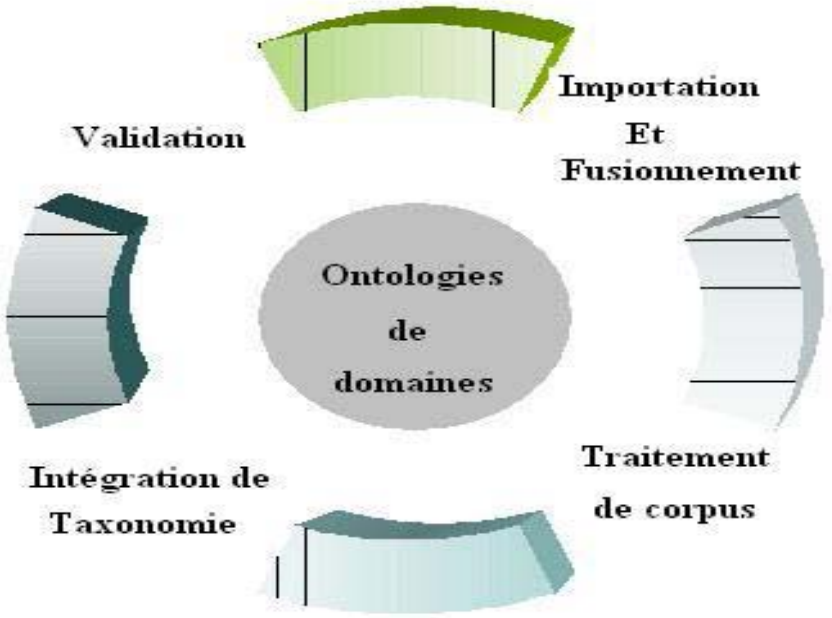


Figure 8 : Méthodologie de construction

2. ARCHITECTURE FONCTIONNELLE DU SYSTEME

A la lumière de la méthodologie décrite ci-dessus, nous développons un système d'assistance à la conception d'ontologies, appelé OntoLogos qui vise à aider le concepteur tout au long du processus de construction, en mettant à sa disposition les outils nécessaires pour la création de l'ontologie à partir des diverses ressources envisagées. Du point de vue génie logiciel, OntoLogos se distingue par son degré d'extensibilité ce qui offre la possibilité d'adapter ou d'intégrer de nouveaux outils pour tenir compte de la diversité de ressources ontologiques potentielles. OntoLogos compte les entités fonctionnelles suivantes (voir Figure 9) :

- Un élément d'importation et de fusion d'ontologies : qui aide le concepteur à choisir dans une première étape, des structures ontologiques existantes, et se charge de l'adaptation de ces ressources si besoin est. Dans la seconde étape, se fait l'intégration des différentes ressources importées pour aboutir à une structure ontologique commune ;
- un élément de traitement de corpus : prend en entrée un corpus de domaine et un corpus de comparaison, qui est lui même composé des éléments de traitement suivants;

- élément d'extraction de candidats-termes : à partir d'un corpus du domaine, cet élément extrait l'ensemble des candidats-termes ayant une signification pour le domaine considéré ;
- élément de filtrage des candidats-termes : sur l'ensemble des candidats-termes extraits, une opération de filtrage est nécessaires pour ne garder que les termes pertinents, cet élément prend en entrée le corpus du domaine et un corpus de comparaison et s'appuie sur des calculs statistique pour élaguer les termes inutiles ;
- élément de construction de la structure taxonomique : en s'appuyant sur la relation d'hyperonymie, cet élément se chargent de la construction de l'hiérarchie taxonomique qui est considérée comme le squelette de toute ontologie ;
- élément d'extraction de relations conceptuelles : en utilisant le corpus du domaine et en mettant en œuvre les fréquences de cooccurrences de termes, cet élément se charge de l'extraction des différentes relations conceptuelles ;

- un élément d'intégration : pour aboutir à une structure taxonomique unique, vient l'élément d'intégration qui prend en entrée l'hiérarchie issue du travail effectué sur le corpus textuel et celle issue de l'élément d'importation et de fusionnement pour proposer la structure finale de l'ontologie ;

Etant conscient de la difficulté inhérente à l'intégration des relations conceptuelles (relations non taxonomiques) dans la taxonomie, un dernier élément qui vient en aide au concepteur pour effectuer cette tâche a été ajouté. Cet élément garantis que l'insertion de ces relations dans l'hiérarchie soit sémantiquement correcte, en aidant le concepteur à trouver l'endroit idéal pour l'incorporation de chaque relation. Remarquez que certaines étapes de ce processus de construction sont indépendantes et peuvent ainsi se faire en parallèle. Dans la suite des sections, nous allons présenter les fondements théoriques de chaque opération effectuée par le système OntoLogos.

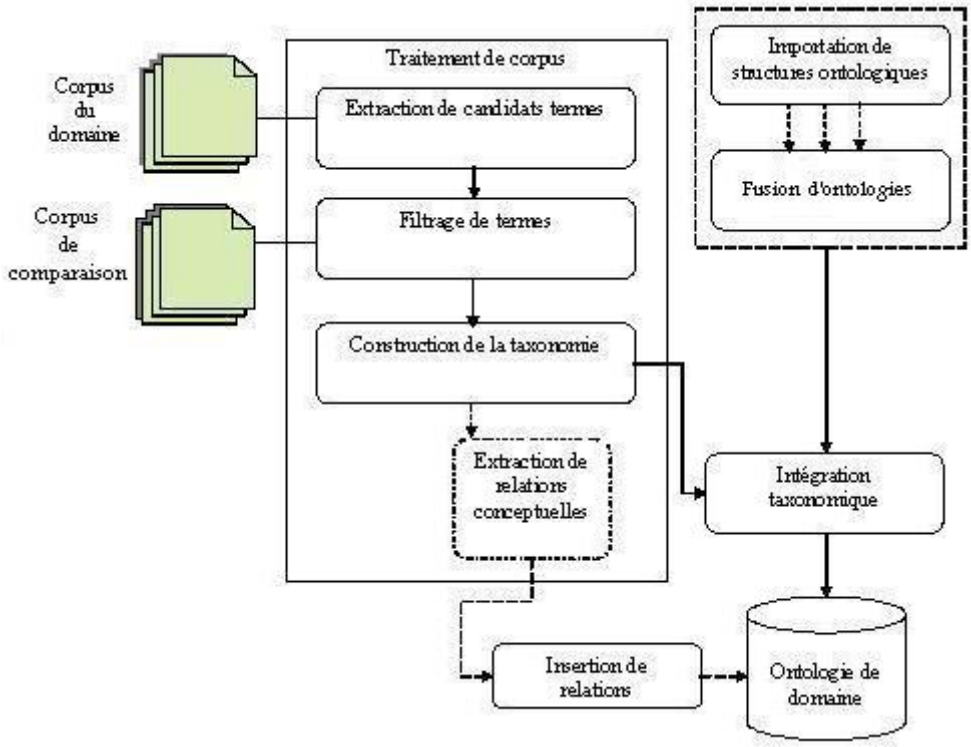


Figure 9 : Architecture fonctionnelle du système

Avoir après présenté l'architecture fonctionnelle de notre système nous détaillons dans section suivante les différentes méthodes ayant servies à effectuer la tâche de construction. Ce processus de construction est divisé principalement en deux étapes ; l'étape de l'acquisition à partir de corpus textuel et la deuxième étape mettant en œuvre la réutilisation des ontologies existantes.

3. ETAPES DE LA CONSTRUCTION

3.1. Extraction de candidats-termes

La première étape dans notre processus de construction d'ontologie est l'extraction de candidats-termes. La Figure 10 résume les grandes étapes d'acquisition. Ces étapes sont automatiques, mais la constitution des listes filtres se fait manuellement. Certaines listes de filtres peuvent être établies indépendamment du corpus. Ces différentes étapes sont :

1. Le prétraitement du corpus
2. la constitution des listes-filtres
3. L'extraction des syntagmes nominaux

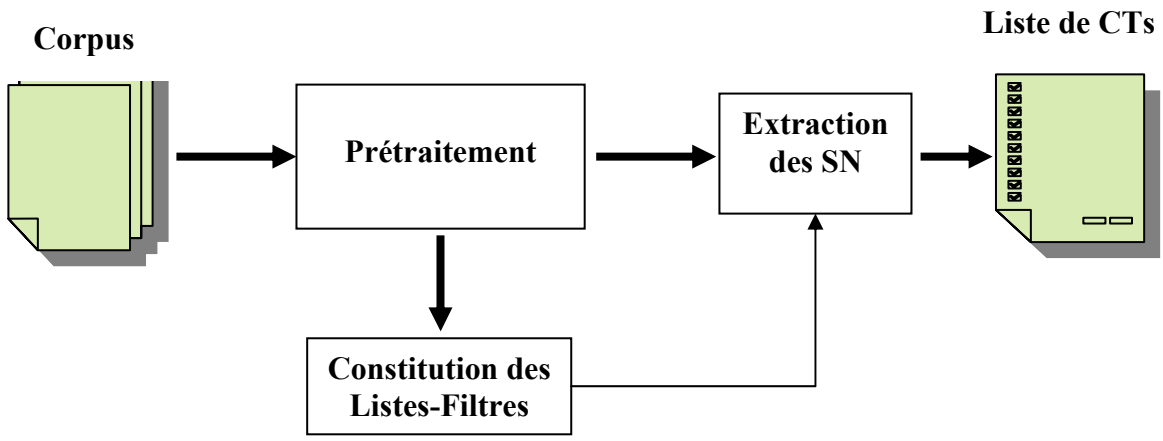


Figure 10 : Les étapes d'acquisition de candidats-termes

3.1.1. Le prétraitement du corpus

Le prétraitement est une étape importante puisqu'elle permet d'éliminer du corpus les informations sans valeur pour le travail d'extraction. Le prétraitement se résume principalement en le découpage du texte en phrases avec un numéro d'ordre d'apparition (la phrase étant notre unité d'analyse), la suppression des figures, des schémas et des tableaux et enfin l'élimination des éventuelles mises en forme pour produire un fichier au format « *.txt* ».

3.1.2. Acquisition des syntagmes nominaux

Dans les textes scientifiques et techniques, les termes et leurs variantes ont une expression linguistique privilégiée sous forme de syntagmes nominaux restreints dont la structure syntaxique est apparentée à celle des noms composés. Il donc possible de les détecter et de les extraire en exploitant cette forme syntaxique privilégiée des termes, c'est-à-dire le groupe nominal, et en prenant en compte sa variabilité.

Par conséquent, le processus d'acquisition sera basé sur le repérage de syntagmes nominaux qui sont les seules structures susceptibles de produire des termes. Pour ce faire, nous nous appuyons sur l'hypothèse que l'ensemble des structures linguistiques délimitant les syntagmes nominaux peut être borné. Le repérage du syntagme peut donc se faire par la détection du premier mot à gauche et à droite qui ne peut pas faire partie du syntagme nominal. Par exemple, dans la phrase :

```
Une turbine à deux arbres est une machine à mouvement rotatif.
```

Le syntagme `nominal turbine à deux arbres` est borné à gauche par la phrase précédente et à droite par un auxiliaire (`est`).

Ainsi, nous établissons quatre listes de tels mots qu'on appelle filtres et qui seront utilisées pour déterminer les frontières gauches et droites de chaque syntagme nominal présent dans le corpus. Ces quatre listes répertorient des filtres de natures différentes :

- **Liste de filtres grammaticaux** : cette première liste-filtre a été établie à partir de la littérature grammaticale classique. Elle contient les mots de la langue appartenant aux catégories grammaticales suivantes :

  - les déterminants : de nature démonstrative (ce, cette, cela, etc.), possessive (sa, ma, son, etc.) et les articles (la, le, les)
  - conjonctions : telles que `Car, Puisque, Alors`, etc.
  - les propositions : telles que `Pour`, `Sur`. Néanmoins, nous avons exclu de cette listes les prépositions à et `De`, qui sont très fréquentes dans la constitution des termes.
  - certains adverbes : telles que `Très, Trop, Beaucoup`, etc.

Elle contient, en outre les lettres de l'alphabet, les dix chiffres et nombres romains ce qui permet par exemple d'éliminer les numéros de paragraphes ;

- **Liste de filtres verbaux** : l'établissement de cette liste présente certaines difficultés; les verbes ne constituent pas une classe fermée et peu nombreuse comme les prépositions ou les articles. Les néologismes verbaux peuvent donc influencer la qualité de l'extraction. Pour remédier à ce problème, cette liste sera créée d'une façon manuelle à partir du corpus à analyser et contiendra tous les verbes présents dans le corpus sous les différentes formes (les variations selon le temps, le genre et le nombre).
- **Liste de filtres spécifiques au corpus** : cette liste a été établie après une analyse partielle du corpus pour en extraire des mots ou expressions qui appartient à la langue des documents, mais qui ne sont pas spécifiques à la terminologie du domaine.
- **Liste de signes de ponctuations** : elle regroupe l'ensemble de signes de ponctuations telles que le point, la virgule, le point-virgule, etc.

Algorithme d'extraction

Soit L_{phrase}, la liste de toutes les phrases du corpus.
Soit Fil_{verb}, Fil_{gram}, Fil_{ponc} et Fil_{dom} les filtres utilisées
Soit L_{CT} la liste des candidats termes {initialement vide}
Soit C la chaîne de caractère en cours.
Soit CT le candidat-terme en cours de traitement.
Début
    Tant que L_{phrase} n'est pas vide faire
    Début
        Si C $\notin (Fil_{verb} \cup Fil_{gram} \cup Fil_{ponc} \cup Fil_{dom})$ alors
        Début
            CT←CT+C.
            Avancer vers la chaîne suivante.
            Tant que C$\notin (Fil_{verb} \cup Fil_{gram} \cup Fil_{ponc} \cup Fil_{dom})$faire
            Début
                CT←CT+C.
                Avancer vers la chaîne suivante.
            Fin
            Si CT$\notin L_{CT}$ alors ajouter CT à L_{CT}.
            CT←la chaîne vide.
        Fin
        Avancer vers la chaîne suivante.
Fin.

Notons que dans le déroulement de cet algorithme, une occurrence d'un syntagme nominal n'est ajouté à la liste des candidats-termes qu'après une vérification qu'il n'existe pas déjà une occurrence du même syntagme dans la liste, sinon elle sera simplement ignorée.

Cette méthode d'extraction a le mérite d'être indépendante de la langue; pour l'appliquer à une autre langue il suffit de créer de nouvelles listes-filtres. En outre, aucun étiquetage grammatical n'est nécessaire pour démarrer le repérage. Par ailleurs, elle permet de collecter la quasi-totalité des syntagmes nominaux du corpus, en effet le taux de rappel est presque

100% : presque tous les syntagmes nominaux d'un corpus sont repérés et même les termes simples (composé d'un seul mot) sont rapportés.

3.2. Filtrage de candidats-termes

Aussi efficace qu'elle soit, l'approche d'extraction de termes adoptée ne peut garantir à elle seule la pertinence de tous les syntagmes nominaux repérés. Par conséquent, pour éviter de travailler sur un ensemble bruité ou trop volumineux de termes, on doit recourir à une étape de filtrage afin d'éliminer les syntagmes redondants ou superflus. Dans cet esprit, dans la méthode que nous proposons l'opération de filtrage se déroule en deux phases :

- **Elimination des redondances** : puisque les termes ne se présentent pas toujours dans le corpus sous la même forme, une phase d'identification et de regroupement de variantes est nécessaire pour éliminer les redondances dues à des variations, syntaxiques ou morpho-syntaxiques ;
- **Filtrage statistique** : pour trancher sur la pertinence des termes restants vis-à-vis du domaine considéré, nous nous appuyons sur une hypothèse qui stipule que les termes du domaine sont plus fréquents dans un corpus représentant ce domaine que dans d'autres corpus. Ainsi, pour effectuer un filtrage à la lumière de cette supposition nous utilisons un deuxième corpus que nous appelons corpus de référence et qui regroupera des documents issus de domaines variés. Le filtrage est essentiellement basé sur une comparaison entre les fréquences d'apparition des candidats-termes dans le corpus du domaine et le corpus de référence, et on élimine tout candidat-terme ayant une fréquence d'apparition dans le corpus de référence supérieure à sa fréquence dans le corpus du domaine.

3.3. Hiérarchisation taxonomique

La construction de la taxonomie est fondée sur l'utilisation de la relation générique/spécifique, cette relation peut être identifiée par un ensemble restreint de marqueurs [Meyer, 2000] qui ont l'avantage d'être généralement indépendant d'un domaine particulier. Par conséquent, le processus de la hiérarchisation taxonomique se fait en deux étapes :

- Extraction des relations d'hyperonymie (hyponymie).
- Construction de la hiérarchie.

Dans ce qui suit nous allons aborder séparément ces deux points. Notons que nous utilisons le terme relations taxonomiques pour faire référence aux relations d'hyperonymie/hyponymie[10]. Ces deux types de relations sont réciproques, ainsi il suffit de trouver, par exemple une relation d'hyperonymie entre les termes T_1 et T_2 pour en déduire la présence de la relation hyponymie entre T_2 et T_1 dans cet ordre.

3.3.1. Extraction des relations taxonomiques

En linguistique, il est admis de considérer que la langue répond à des règles que l'on peut expliciter. La notion de marqueur traduit cette hypothèse relativement au repérage de relations sémantiques (du moins celles de nature taxonomique). On suppose que l'observation de certaines formes linguistiques entre deux ou plusieurs éléments du lexique peut révéler un rapport de sens entre ces éléments. Il s'agit alors d'élaborer des marqueurs pour rendre compte précisément de fonctionnement lexicaux et les associer à une interprétation sémantique systématique.

Un *marqueur* correspond à une *formule linguistique* dont l'interprétation définit régulièrement le même rapport de sens entre des termes. Les marqueurs sont composés de mots de la langue. Par exemple, la formule linguistique suivante est un des marqueurs de la relation d'hyponymie « `Est un` ». Dans la phrase :

`Une turbine à deux arbres` **`est une`** `machine motrice`

A l'aide du marquer lexical `Est une`, on peut extraire la relation d'hyponymie entre le `terme turbine à deux arbres` et le terme `machine motrice`.

Par conséquent, l'utilisation d'une liste de marqueurs préétablis (marqueurs d'hyperonymie), nous permet de trouver toutes les relations d'hyperonymie présentent dans le corpus. Ces marqueurs peuvent être acquis automatiquement à partir de corpus [Aussenac-Gilles & Séguéla, 2000], dans notre cas les marqueurs sont définis manuellement et regroupés dans une liste de marqueurs. Cette liste peut être enrichie au fur et à mesure de l'acquisition des relations taxonomiques pour y inclure d'autres marqueurs. Les marqueurs lexicaux exprimant des liens taxonomiques entre termes sont de deux types :

[10] D'autres relations paradigmatiques peuvent être utilisées pour la construction de la hiérarchie, mais la relation d'hyperonymie est la plus utilisée

1. Marqueurs d'hyperonymie, comme « `En particulier, surtout. etc.` »
2. Marqueur d'hyponymie, comme « Est un »

Les couples de termes reliés par une relation taxonomique seront organisés selon l'ordre suivant :

`Hyperonymie(Terme`$_1$`, Terme`$_2$`)`

Donc, si on reprend l'exemple précédent on aura le couple :

`Hyperonymie(Machine motrice, Turbine à deux arbres)`

Les formes linguistiques reflétant une relation d'hyperonymie sont de plusieurs types, selon l'étude menée par [Borillo, 1996] sur cette relation, les différentes structures mises en évidence sont :

- **Structure prédicative attributive** : La structure prédicative attributive « `Terme`$_1$ `être un Terme`$_2$ » est surtout caractéristique des dictionnaires et des textes de nature didactique.
- **Structures coordonnées comportant un marqueur de spécification** : La coordination entre le terme hyponyme et le terme hyperonyme peut se faire par la structure adjectivale « `et d'autres` » comme dans le schéma « Terme$_1$ {et/ou} de autre Terme$_2$ ». La coordination peut aussi s'établir par un adverbe de spécification comme dans « Terme$_1$ `et notamment Terme`$_2$ ».
- **Structures appositives :** Les structures mettant en jeu une apposition sont relativement difficiles à relever dans la mesure où elles ne s'accompagnent pas toujours de marqueurs syntaxiques comme dans le schéma « Terme$_1$, Terme$_2$ ». On notera que l'élément apposé ne doit pas être précédé d'un déterminant. Une structure appositive est aussi utilisée dans le schéma « Terme$_1$, particulièrement Terme$_2$, ».
- **Structures d'exemplification et d'énumération** : Les schémas reflétant des exemplifications sont les plus productifs et les plus fiables dans la mesure où le terme hyperonyme est précédé d'un adjectif indéfini comme *certains, quelques, plusieurs...* La présence de l'adverbe *comme* ou des adjectifs *tel* et *tel que* sert de marqueur à une exemplification. Une énumération peut être détectée par des schémas de la forme « Terme$_1$: Terme$_2$,..,Terme$_i$ », « Terme$_1$ (Terme$_2$,..., Terme$_i$).

Le tableau suivant donne quelques marqueurs qu'on a définis

Marqueur/schéma
Etre un
En particulier
Surtout
Notamment
Comme
Tel que
Et d'autres
T_1, T_2
T_1 (T_2,…,T_i)
Tous les T_1 sauf T_2
T_1 ainsi que d'autres T_2

Tableau 1 : Marqueurs et schémas d'hyperonymie

Algorithme d'extraction

Notre méthode d'extraction consiste en premier lieu à définir une liste de marqueurs lexicaux L_M pouvant désigner une relation d'hyperonymie, ensuite nous recherchons dans le corpus les marqueurs de la liste L_M. Pour chaque occurrence d'un marqueur nous relevons deux à deux les termes qui sont autours de ce marqueur, et nous construisons ainsi des couples de termes hyperonymes. La figure suivante montre les étapes de l'acquisition des couples hyperonymes.

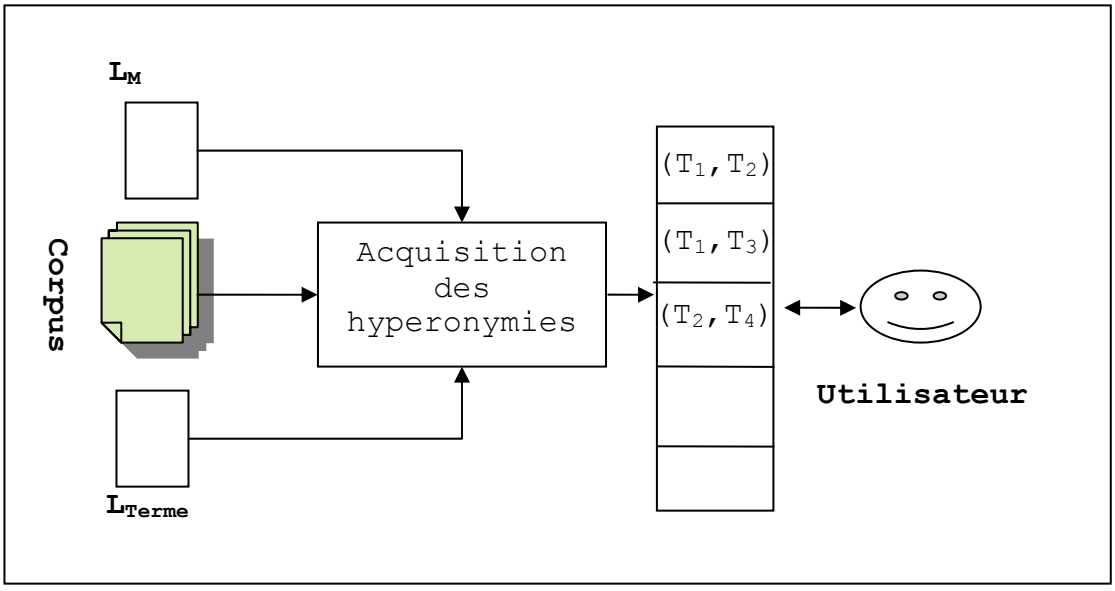


Figure 11 : Acquisition des Hyperonymies

Ainsi, l'algorithme d'acquisition peut être formulé comme suit :

```
Soit L_phrase, la liste de toutes les phrases du corpus.
Soit L_M la liste de marqueurs.
Soit L_Terme la liste des termes.
Soit M le marqueur en cours traitement.
Soit Terme_g et Terme_d les termes à gauche et à droite du marqueur.
Soit L_couple la liste des couples hyperonymes.
Début
Pour chaque verbe M de L_M faire
Tant que L_phrase n'est pas vide faire
    Début
        Trouver(M) ;
        RécupérerCouple(T_g,T_d) ;
        Insérer((T_g,T_d),L_couple) ;
        Fin
Fin.
{
```

Vu que certaines phrases du corpus peuvent présenter des ambiguïtés ce qui peut provoquer des erreurs lors de l'acquisition, nous avons décidé de soumettre l'ensemble de couples extraits à l'utilisateur pour les valider. Pour chaque couple de termes nous pouvons aussi faire référence au contexte dans lequel ils apparaissent afin de faciliter la tâche de l'utilisateur lors de la validation.

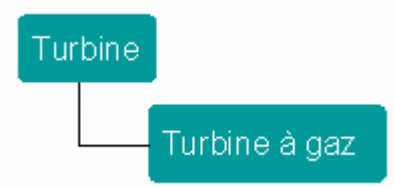
Maintenant que l'ensemble de couples des termes hyperonymes a été extraits, nous pouvons passer à la deuxième étape de la hiérarchisation taxonomique.

3.3.2. Construction de la hiérarchie

La hiérarchie de termes à construire sera une arborescence puisque un terme (désignant un concept donné) peut hériter de plus d'un terme. Pour la construction de la hiérarchie nous adoptons une approche descendante, c'est-à-dire que la construction sera faite de la racine vers les feuilles. La racine de la hiérarchie est un concept artefact qu'on a introduit pour réunir tous les termes sans parents sous un père commun, nous avons ainsi choisi d'utiliser le

concept **THING** utilisé dans Protégé2000 comme racine de l'arborescence. La hiérarchisation est fait à deux niveaux :

1. **Hiérarchisation endotermes** : il s'agit structurer les termes complexes en tête et expansion. Par exemple pour le terme turbine à gaz on aura la structure :



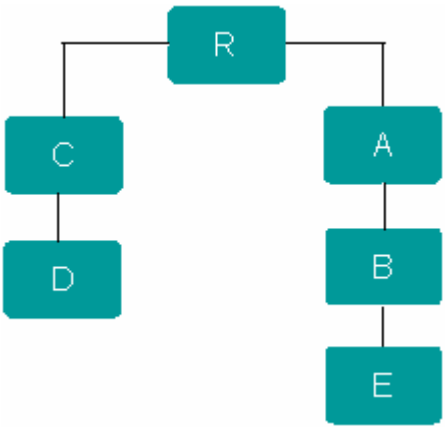
2. **Hiérarchisation exotermes :** qui va utiliser la liste de couples établie précédemment pour structurer les termes entre eux. A ce niveau le processus de structuration sera fait selon l'algorithme suivant :

En utilisant la liste de couples établie dans l'étape précédente, le processus de construction se déroulera comme suit :

1. récupérer les termes racines (ne figurant pas dans la partie droite d'un couple)
2. Relier ces termes à la racine artificielle R.
3. Chercher les termes qui n'ont que les termes déjà classés comme parents
4. Connecter ces termes aux termes parents selon les couples où ils apparaissent.
5. Supprimer les couples de termes traités.
6. Aller à 3.

Ce processus est réitéré jusqu'à ce que la liste de couples soit vide.

Ainsi, si on démarre de l'ensemble de couple {(a,b), (b,e), (c,d)}, le processus de construction va engendrer la hiérarchie suivante :



3.4. Extraction de relations conceptuelles (relations non-taxonomiques)

Beaucoup de travaux se sont intéressés à l'extraction de relations sémantiques entre termes et la plupart d'eux s'inscrivent dans la lignée de [Harris, 1968], tels que les travaux de [Morin, 1999], [Séguéla, 1999]. Il serait maladroit d'utiliser des marqueurs linguistiques comme un moyen de repérage des relations conceptuelles, car il n'est pas possible de prévoir la nature de toutes les relations que peuvent entretenir les termes d'un corpus donné.

Dans l'approche que nous proposons, qui est inspirée des travaux de [Oueslati, 1999] et s'inscrivant dans un contexte distributionnel, nous mettons en œuvre une variante de la synthèse automatique de contextes proposée par R. Oueslati en s'appuyant sur des données sur les cooccurrences de termes.

3.4.1. Phrase comme prédicat

Pour l'acquisition de relations entre termes, on considère que la phrase est un segment prédicatif (*i.e.* disant quelque chose à propos de quelque chose) et donc pour nous elle marque une limite dans l'analyse linguistique du niveau termes/relations.

Dans ce type de phrases prédicat-arguments, la relation sémantique entre termes est souvent exprimée par l'intermédiaire d'expressions pivots morphosyntaxiques. Ces expressions sont souvent composées de prédicats (verbes), ou de formules morpho syntaxiques semi-figées. Le schéma de ces relations est donc du type :

$$\text{Terme}_1 \ \textbf{EXP} \ \text{Terme}_2$$

3.4.2. Méthode d'extraction

Les expressions qui nous intéressent dans notre étude, sont celles exprimées par le moyen d'un verbe. En effet, nous considérons que la coprésence de deux termes dans le contexte d'un verbe donné, est une information suffisante pour conclure de la présence d'une relation conceptuelle entre ces deux termes et qui sera marqué par le verbe pivot. Par exemple dans la phrase suivante :

```
Turbine haute tension, entraîne un compresseur axial ainsi que des
accessoires (pompe à huile de lubrification,...
```

La cooccurrence des termes `Turbine Haute Tension` et `Compresseur Axial` autour du verbe `Entraîne`, indique la présence d'une relation sémantique entre ces deux termes, relation désignée par le verbe `Entraîne`.

Dans cette phrase, on constate que le terme `Turbine Haute Tension` maintient la même relation (`Entraîne`) avec le terme `Pompe à Huile de Lubrification`, qui se situe à la distance (qu'on appellera désormais, *Fenêtre*) de quatre chaînes du verbe pivot. De cette constatation, le schéma des relations à extraire peut se réécrire comme suit :

$$\{Terme_i\}\ Fenêtre_{gauche}\ \mathbf{Verbe}\ Fenêtre_{droite}\ \{Terme_j\}$$

Qui signifie que tous les termes $Terme_i$ maintiennent la relation **`Verbe`** avec les termes $Terme_j$. La méthode d'extraction va se limiter à rechercher, dans le corpus de départ, les verbes de la liste Fil_{verbe}, établie précédemment, et pour chaque verbe trouvé, recenser les termes se situant à une fenêtre $Fenêtre_{gauche}$ à gauche du verbe et les termes d'une distance $Fenêtre_{droite}$ à droite du verbe ; la recherche est effectuée dans les limites d'une phrase. Cette opération de recensement va donner lieu à une matrice de correspondances entre termes.

Matrice de correspondances

Dans cette matrice, est répertorié pour deux termes donnés l'ensemble de verbes réunissant les deux termes. Schématiquement, la matrice de correspondances sera représenté comme suit :

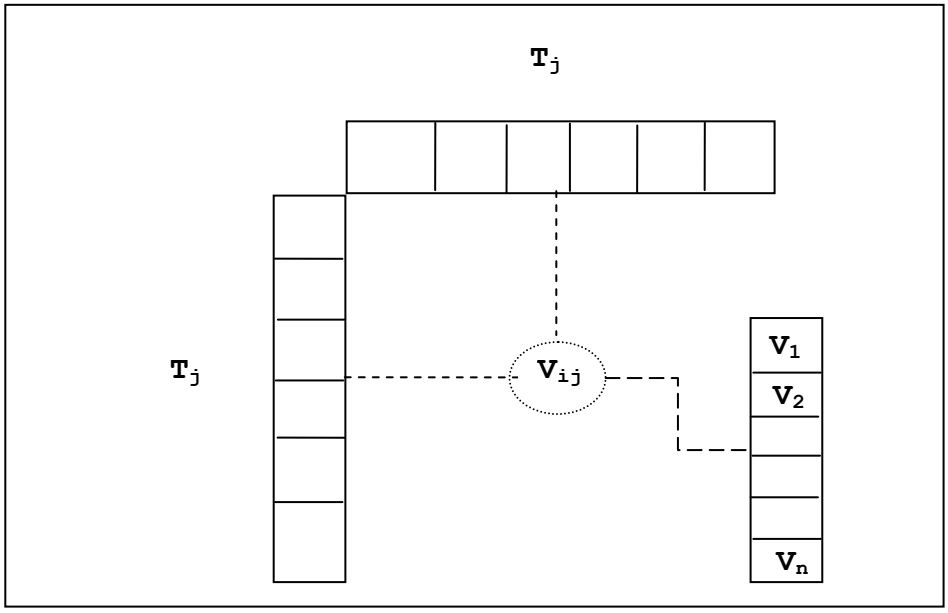


Figure 12 : représentation d'une matrice de correspondances

Où V_{ij} est un tableau regroupant les verbes réunissant les Termes T_i et T_j.

Algorithme d'acquisition

Pour faciliter la tâche de l'utilisateur lors de la validation et l'étiquetage des relations extraites, nous allons procéder à un regroupement des Verbes synonymes d'un tableau donné Vij, pour lui présenter un tableau de groupes de verbes synonymes à la place d'un simple tableau de verbes. Ceci nécessite, bien entendu, le recours à une ressource externe, qui sera par exemple un dictionnaire électronique de synonymes. La figure suivante illustre les étapes de l'acquisition des relations conceptuelles entre termes.

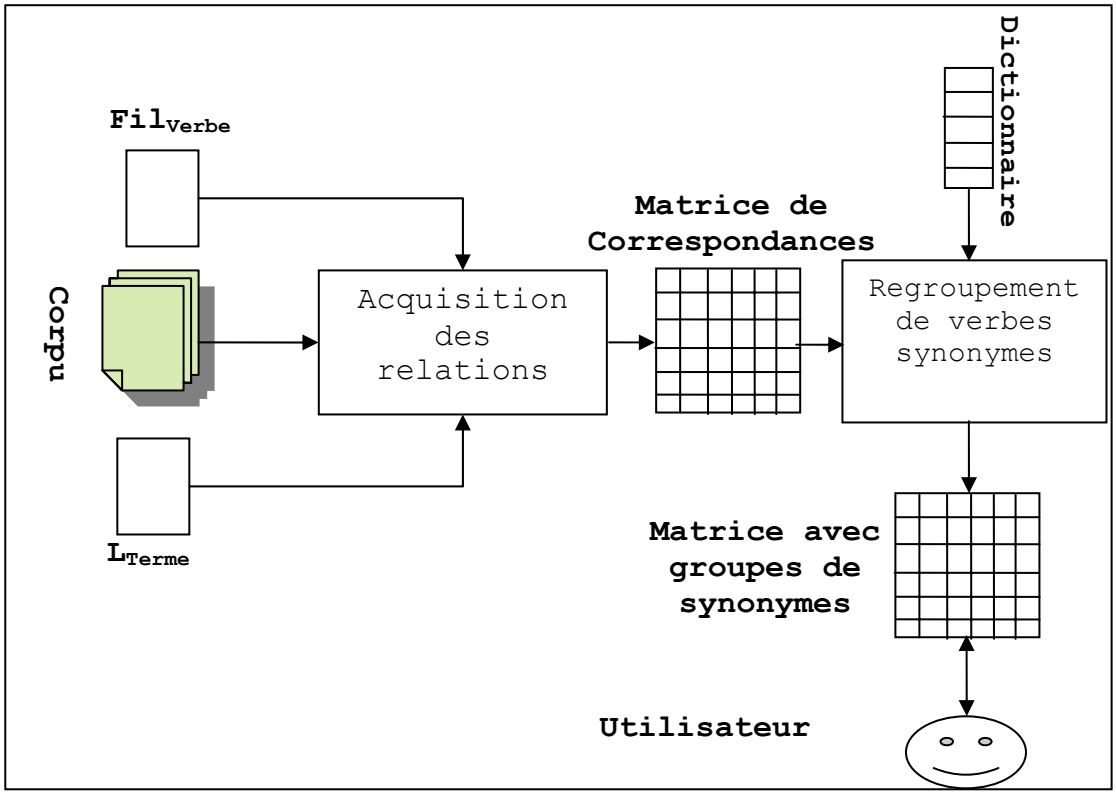


Figure 13 : Etapes de l'acquisition de relations

L'acquisition s'opère sur le corpus du domaine, en utilisant la liste $\mathbf{L_{Terme}}$ trouvée après l'étape de filtrage de termes et la liste $\mathbf{Fil_{Verbe}}$ établie auparavant. Le résultat final va être présenté à l'utilisateur pour choisir l'étiquette de chaque relation qu'il juge pertinente et éliminer les autres relations. L'algorithme générale d'acquisition peut alors être écrit comme suit :

```
Soit Lphrase, la liste de toutes les phrases du corpus.
Soit Filverb la liste des verbes.
Soit LTerme la liste des termes.
Soit V le verbe en cours traitement.
Soit {Terme}g et {Terme}d les termes à gauche et à droite du
verbe V.
Soit Mv la matrice de correspondances
Début
Pour chaque verbe V de Filverbe faire
Tant que Lphrase n'est pas vide faire
    Début
        Trouver(V) ;
        Récupérer({Terme}g, Fg) ;
        Récupérer({Terme}d, Fd) ;
        Pour chaque Tg de {Terme}g et Td de {Terme}d faire
        Mv[Tg,Td]←V ;
        Fin
Fin.
{
 *Trouver(V) : fonction qui se charge de trouver V.
 *Récupérer({Terme}g, Fg} : récupère la liste de termes Tg dont le premier élément
est à une distance 0 ≤ D ≤ Fg de V. La récupération continue jusqu'à la fin de la
phrase ou la rencontre d'un autre verbe dans la même phrase.
.*Récupérer({Terme}d, Fd} : Idem.
```

Cet algorithme est développé selon l'hypothèse de la cooccurrence de termes et ne nécessite aucune connaissance linguistique préalable pour son déroulement. Toutefois, l'utilisation d'une telle heuristique, ne fournit pas beaucoup d'explications à l'utilisateur sur la nature des relations extraites, pour cela, nous avons choisi de garder trace du contexte de chaque relation, c'est-à-dire une lien vers le phrase d'où elle a été extraite, ce qui va aider l'utilisateur lors de l'étiquetage des relations. Ainsi Les relations obtenues ont besoin d'être validées puis étiquetées (compte tenu de leurs contextes) par l'utilisateur pour les insérer ensuite dans la taxonomie finale sous la forme de propriétés des classes ou comme des classes indépendantes (selon le paradigme des frames employé par Protégé2000).

Remarquons que dans cette étape d'acquisition l'utilisateur joue un rôle central. En effet, l'extraction des relations conceptuelles est très difficile et une acquisition complètement automatique produirait des résultats trop bruités. Par ailleurs, l'utilisateur est le seul à détenir suffisamment de connaissances sémantiques qu'ils lui permettent d'étiqueter les relations.

3.5. Importation et fusion d'ontologies

Certains points sont à considérer lors de l'importation des ontologies :

- ✓ L'importation doit permettre à l'utilisateur de réutiliser des ontologies sans se soucier des formalismes et langages dans lesquels elles sont spécifiées.
- ✓ Les ressources à réutiliser peuvent être des ontologies, des réseaux sémantiques ou des documents semi-structurés. Donc on doit fournir à l'utilisateur les outils adaptés à chaque source de données qu'il veut exploiter.
- ✓ C'est à l'utilisateur de vérifier la pertinence de ces différentes sources par rapport aux objectifs de l'ontologie à construire.
- ✓ Ces différentes sources sont ramenées à de simples hiérarchies de termes.
- ✓ Les hiérarchies de termes qu'on a construites dans la phase d'importation doivent être fusionnées pour donner lieu à une seule structure taxonomique qu'on va intégrer dans l'ontologie construite à partir du corpus du domaine

Dans la méthode que nous proposons, c'est à l'utilisateur de localiser et choisir les ressources à importer, de plus il doit choisir l'outil adapté au type de ressource qu'il veut réutiliser.

Une fois l'étape d'importation réalisée, on passe à l'intégration des différentes ressources importées. L'utilisation conjointe de deux (ou plus de deux) ontologies peut nécessiter soit un simple alignement dans le cas où aucune partie n'est commune aux ontologies, soit une véritable fusion [Noy, 2002]. L'alignement d'ontologies consiste à trouver des correspondances entre les connaissances spécifiées dans les deux ontologies, de manière à pouvoir les exploiter conjointement dans le même système. En pratique, il s'agit d'identifier des concepts (ou des relations) de la première ontologie avec des concepts (ou des relations) de la seconde, ou de trouver des liens conceptuels (subsomption,...) entre eux. Contrairement à l'alignement où les deux ontologies de départ reste intactes, la *fusion d'ontologies* consiste, à partir de deux ontologies, à en créer une troisième qui intègre les connaissances spécifiées dans les deux premières.

Dans les deux cas, la connexité des deux domaines de connaissance modélisés par les ontologies est requise, sans quoi aucun lien ne peut être établi entre concepts. De plus, les formalismes de représentation d'ontologie utilisés doivent être au moins compatibles, ainsi que les paradigmes conceptuels [Maedche *et al.*, 2002]. En pratique, l'alignement de deux ontologies se fera dans le cadre du même langage de représentation et pourra donc nécessiter des transcriptions préalables d'un langage à l'autre.

L'uniformisation des modèles et formalismes de représentation sont également nécessaires à la fusion d'ontologies. Préalablement à la fusion, il convient de déterminer quelle est l'ontologie la plus générale, ou celle qui est la plus étendue, c'est-à-dire celle qui ne sera pas modifiée. Les autres devront être alignées sémantiquement et syntaxiquement sur l'ontologie la plus générale [Noy & Musen, 1999b]. Le problème se ramène alors à l'intégration d'une ontologie dans une autre. Une fois les deux ontologies exprimées dans le même formalisme et à travers le même modèle cognitif, ces entités communes aux deux ontologies doivent être identifiées. Les méthodes appliquées pour repérer les similarités entre concepts et/ou relations sont [Euzenat & Valtchev, 2004] :

- les méthodes terminologiques qui comparent les *labels* désignant deux concepts ou deux relations ;
- les méthodes qui comparent les **propriétés internes** des concepts et relations (attributs des concepts, portée d'une relation, etc.) ;
- les méthodes qui comparent les **propriétés externes** des concepts et relations (subsomptions, relations entre concepts, etc.) ;
- les méthodes qui comparent les **extensions** des concepts et relations, c'est-à-dire leurs instances ;
- les méthodes qui comparent la **sémantique** des concepts et relations.

Ces méthodes peuvent bien entendu être combinées entre-elles. Elles peuvent parfois recourir à des ressources extérieures aux ontologies à aligner. Ainsi, les méthodes terminologiques peuvent faire appel à des ressources linguistiques (tables de synonymes ou/et d'hyperonymes) pour faciliter l'identification de liens entre concepts (c'est le cas dans l'outil de fusion d'ontologie **ODEMERGE** intégré à l'éditeur d'ontologie **ODE** [OntoWeb, 2002]). Les méthodes comparant les extensions des concepts et relations peuvent utiliser des corpus où apparaissent les instances recherchées (c'est le cas de l'outil **FCA-MERGE** [Stumme & Maedche, 2001]).

Les correspondances ainsi établies entre entités conceptuelles ne sont pas forcément bijectives. Des conflits peuvent naître lors de cette « traduction au niveau sémantique », qui ne peuvent être résolus automatiquement. Si le degré de similarité ne permet pas de trancher entre deux correspondances possibles, l'intervention humaine est indispensable [Noy & Musen, 1999a]. D'autre part, si certaines entités de l'ontologie à intégrer n'offrent de similarité avec aucune entité de l'ontologie cible, il est tout de même nécessaire de leur trouver une entité subsumante dans l'ontologie cible. La différence de granularité entre les deux ontologies peut de plus entraîner la suppression de certaines entités, ou plus précisément leur agrégation au sein d'une même entité cible.

Plusieurs outils existants se basent sur une approche terminologique pour la mise en évidence d'analogies entre primitives conceptuelles, analogies qui sont ensuite raffinées par la comparaison de certaines propriétés structurant l'ontologie. C'est par exemple le cas des outils **SMART** et **PROMPT**, développés pour l'atelier **PROTEGE2000**, et qui comparent les attributs des concepts (slots et facets) pour affiner les analogies [Noy et Musen 1999b ; 2000]. Les analogies ainsi découvertes sont soumises à l'utilisateur qui tranche entre les différentes possibilités. **CHIMAERA** est un outil d'alignement lié à l'éditeur d'ontologie **ONTOLINGUA** [McGuiness *et al.*, 2000]. **CHIMAERA** utilise à la fois les termes des concepts et certaines de leurs propriétés pour proposer à l'utilisateur des analogies que celui-ci doit valider ou rejeter.

Une étude détaillée des différents outils d'alignement d'ontologie montre cependant qu'aucun outil ne propose pour le moment une méthode d'alignement principalement basée sur les axiomes des ontologies et comparant la sémantique des concepts et relations pour déterminer des analogies entre eux [Kalfoglou et al., 2003].

L'importation et la fusion d'ontologies sont deux problèmes qui suscitent beaucoup d'interrogations, à cause notamment de la diversité des formalismes et langages de spécifications. Vu la difficulté inhérente à la fusion d'ontologies, on n'a pas essayé de l'automatiser et s'est contenté dans ce mémoire de présenter le problème sous un angle purement théorique. Cette automatisation fera sans doute l'objet de travaux ultérieurs.

3.6. L'évaluation de l'ontologie finale

Deux niveaux peuvent être distingués dans l'évaluation d'une ontologie [Gomez-Perez *et al.*, 1999] :

- **La vérification**, qui consiste à s'assurer que l'ontologie est conforme à un modèle formel de représentation de connaissances. Cette vérification porte sur des propriétés formelles qui ne peuvent être violées par l'ontologie, sous peine de perdre son expressivité.
- **La validation**, qui consiste à s'assurer de la conformité sémantique de l'ontologie à un domaine de connaissance, c'est-à-dire que la sémantique exprimée dans l'ontologie doit être celle du domaine considéré [Gomez-Perez *et al.*, 2003].

En d'autres termes, la vérification correspond à l'exigence *« building the system right »*, relativement à un modèle formel, et la validation correspond à l'exigence *«building the right system »*, relativement au domaine de connaissance modélisé [O'Keefe et al., 1987]. De plus, l'évaluation de l'ontologie en amont de son opérationnalisation est souhaitable pour éviter de propager des erreurs, même si l'opérationnalisation peut être nécessaire pour mener certaines activités d'évaluation. Ainsi, utiliser l'ontologie pour répondre à des questions de compétence nécessite d'avoir opérationnalisé l'ontologie; le test de la cohérence d'une ontologie peut nécessiter des déductions pour mettre en évidence des contradictions logiques entre axiomes. Cependant, la validité des hiérarchies de concepts et/ou de relations doit être testée dès la phase d'ontologisation, aussi bien du point de vue formel que du point de vue sémantique.

3.6.1. La vérification

La vérification consiste à tester si l'ontologie est correctement construite du point de vue du modèle de représentation de connaissances adopté. La vérification d'une ontologie est donc un processus formel dépendant du modèle et non du domaine. Nous considérons que la vérification d'une ontologie repose sur le test de trois grands types de propriétés : la *conformité*, la *cohérence* et la *minimalité*.

La **conformité** d'une ontologie à un modèle de représentation exprime le fait que les représentations de connaissance incluses dans l'ontologie sont bien conformes au modèle utilisé. Par exemple, si on utilise un modèle de type Entité-Relation [Chen, 1976], il est nécessaire de préciser la signature de chaque relation, sinon la représentation n'est pas conforme au modèle. Comme son nom l'indique, la conformité implique que la forme de l'ontologie soit bien celle prévue par le modèle, qu'elle soit conforme à la syntaxe imposée par celui-ci. Elle est donc indépendante du domaine de connaissance. Bien évidemment, un défaut de conformité révèle souvent un défaut de modélisation et donc un manque

d'adéquation entre les connaissances du domaine et l'ontologie. Cependant, les tests de conformité d'une ontologie vis-à-vis d'un modèle de représentation ne dépendent pas du domaine et seront les mêmes pour toutes les ontologies construites dans ce modèle. C'est en ce sens que nous considérons la conformité comme indépendante du domaine.

La **cohérence** d'une ontologie est déterminée par l'absence de contradictions logiques entre les représentations qu'elle contient. Elle fait appel à la sémantique formelle du modèle de représentation, là où la conformité n'utilise que la syntaxe. Par exemple, deux axiomes de l'ontologie, syntaxiquement corrects, peuvent être logiquement contradictoires, ou un ensemble d'axiomes peut permettre une déduction en contradiction avec un autre axiome. La vérification de la cohérence d'une base de connaissances est un problème complexe ayant fait l'objet de nombreux travaux [Djelouah *et al.* 2002]. La cohérence est une propriété qui repose uniquement sur les représentations contenues dans l'ontologie et les tests de cohérence ne dépendent pas du domaine de connaissance.

La **minimalité** d'une ontologie désigne le fait qu'elle ne contient pas de connaissances superflues, c'est-à-dire des connaissances spécifiées deux fois ou plus ou des connaissances qu'on puisse facilement déduire du reste de l'ontologie. Par exemple, une ontologie où apparaît deux fois le même concept (c'est-à-dire deux concepts ayant le même label et les mêmes propriétés) n'est pas minimale. De même une ontologie contenant un axiome qu'on peut déduire d'une combinaison d'autres axiomes de l'ontologie n'est pas minimale. De plus, des connaissances inutiles peuvent être spécifiées, comme par exemple des axiomes indéclenchables du fait de la présence de connaissances contradictoires dans leur hypothèse.

En pratique, la vérification d'une ontologie consiste souvent, entre autres, à vérifier qu'il n'y a pas de cycle dans les hiérarchies, c'est-à-dire de définitions en boucle, qu'il n'y a pas redondance de concepts ou de relations, que chaque hiérarchie est bien connexe, c'est-à-dire s'il n'y a pas de concept ou de relation isolé des autres et donc sans aucun sens, le sémantique d'un concept reposant sur les liens conceptuels qu'il entretient avec les autres concepts [Gomez-Perez *et al.*, 1996]. D'autres tests ont été proposés pour contrôler formellement la cohérence des hiérarchies de concepts, de manière indépendante du domaine de connaissance. Ces tests reposent sur l'utilisation de critères de construction des hiérarchies qui contraignent les choix de modélisation et impose d'utiliser des propriétés conceptuelles qui pourront par la suite servir de base à la vérification. C'est le cas des principes différentiels énoncés par B. Bachimont qui peuvent être facilement contrôlés formellement de manière à tester la

cohérence générale de la hiérarchie des concepts d'une ontologie [Bouaud et al., 1994]. Les méta-propriétés proposées par C. Welty et N. Guarino permettent également de vérifier la cohérence sémantique de l'ontologie puisque ces métapropriétés imposent des contraintes sur les liens de subsomption [Welty & Guarino, 2001].

3.6.2. La validation

La validation d'une ontologie consiste à tester sa fidélité à la sémantique du domaine de connaissance considéré [Gomez-Perez, 1999]. La validation permet de tester la *complétude de l'ontologie* et la *conformité de l'ontologie* par rapport au domaine.

- La **complétude** de l'ontologie par rapport au domaine est assurée si toutes les connaissances du domaine sont présentes dans l'ontologie ;
- La **conformité** de l'ontologie par rapport au domaine est assurée si les connaissances représentées dans l'ontologie correspondent exactement à la sémantique du domaine.

La validation repose sur l'utilisation de **spécifications** [Laurent, 1992]. Ces spécifications peuvent être celles du comportement attendu du système, mais aussi du comportement interdit (spécifications d'anomalies) [Haouche-Gingins & Charlet, 1995] :

- Les **spécifications du système** explicitent le comportement attendu du système et, dans le cas des ontologies, ce que l'ontologie permet de déduire. Il s'agit de tester la complétude de l'ontologie par rapport au domaine.
- Les **spécifications d'anomalies** explicitent les comportements interdits au système. Il s'agit de tester la conformité de l'ontologie par rapport au domaine en contrôlant que les déductions permises par l'ontologie ne vont pas à l'encontre de la sémantique du domaine.

M. Gruninger propose dans [Gruninger & Fox, 1995] d'utiliser un ensemble de **questions de compétence** auxquelles l'ontologie doit permettre de répondre. Une question de compétence est constituée par un fait initial à partir duquel un autre fait donné au départ doit pouvoir être déduit en utilisant les connaissances représentées dans l'ontologie. L'impossibilité de répondre à une telle question, ou l'obtention d'une réponse différente de celle attendue, implique que certaines connaissances du domaine indispensables à la résolution de la question manquent dans l'ontologie où qu'elles y sont mal représentées. La validation des ontologies va ainsi reposer sur des spécifications externes à l'ontologie et des

spécifications internes constituées par les schémas d'axiome et axiomes décrivant la sémantique du domaine [Haouche-Gingins & Charlet, 1995].

La validation d'une ontologie ne peut être menée qu'à travers son utilisation opérationnelle puisque tester formellement la sémantique d'une ontologie (au sens de tester la forme de l'ontologie) suppose de pouvoir se référer à un modèle formel préexistant des connaissances du domaine. Or bâtir une ontologie a justement pour but de créer ce modèle formel, à partir d'un corpus informel de connaissances. On ne peut donc valider une ontologie à travers sa forme mais uniquement à travers son utilisation opérationnelle.

La complétude de l'ontologie peut être testée à deux niveaux :

- Le test de la **complétude du niveau terminologique** vise à contrôler si toutes les primitives conceptuelles du domaine sont bien présentes dans l'ontologie. Ce test peut être mené en s'assurant que les spécifications (questions de compétence) de l'ontologie peuvent être représentées formellement à l'aide du vocabulaire conceptuel présent dans l'ontologie. L'impossibilité de représenter une question de compétence implique une incomplétude du niveau terminologique de l'ontologie.
- Le test de la **complétude du niveau sémantique** vise à contrôler si les axiomes de l'ontologie représentent bien toutes les connaissances du domaine. Ce test peut être mené en s'assurant que les axiomes permettent de répondre aux questions de compétence.

La conformité de l'ontologie est testée en s'assurant que les schémas d'axiome et axiomes de l'ontologie permettent de déduire les réponses correctes des questions de compétence. En pratique, les tests de complétude et de conformité au domaine ne sont pas séparés et consistent à poser les questions de compétence dans un système opérationnel implémentant une version opérationnelle de l'ontologie. L'impossibilité de représenter une question, de générer une réponse ou la génération d'une réponse incorrecte met en évidence un problème de modélisation des connaissances du domaine.

Le test d'une ontologie doit bien entendu précéder son utilisation, mais est également utile à tous les niveaux de l'évolution d'une ontologie. En particulier, si on désire augmenter la taille du domaine de connaissance modélisé, donc ajouter de nouvelles connaissances à l'ontologie, il faut s'assurer que les représentations déjà construites sont valides, d'autant plus

qu'une ontologie va croître par agrégation de connaissances dans toutes les directions, et non par ajout d'une couche de connaissances [Fernandez *et al.*, 1997].

L'évaluation est une étape très importante dans le processus globale de construction des ontologies, néanmoins compte tenu des exigences liées à cette tâche, son automatisation reste un problème ouvert, et les réponses ne seront pas pour demain.

4. RESUME

Seule une construction automatique permet de répondre aux besoins en termes de ressources ontologiques des différentes applications. Nourri de cette conviction nous avons développé notre méthode de construction qui réunis au sein d'un même processus l'acquisition à partir de texte et la réutilisation des ontologies.

Au travers des différentes sections de ce chapitre nous avons détaillé les techniques employées pour automatiser le développement d'ontologies. Nous avons vu que certaines tâches de ce processus de construction nécessitent un prérequis sémantique énormes pour être accomplies, à l'heure actuelle l'expert de domaines est le seul à posséder cette sémantique, ainsi dans notre modèle de construction, il incombe à expert du domaine d'effectuer ces étapes du processus de construction.

Dans le chapitre qui suit nous étendons cette construction au développement d'ontologies bilingues. Notre étude va se limiter à présenter les grandes lignes d'un processus d'acquisition mettant en jeu plus d'une langue.

CHAPITRE 4

VERS DES ONTOLOGIES BILINGUES

« Un homme vaut autant d'hommes qu'il connaît de langues »

La recherche d'informations multilingues et la traduction automatique sont deux exemples de domaines où les ontologies bilingues peuvent être utilisées. Un nombre important d'ontologies bilingues sera ainsi utilisé pour améliorer la qualité de ces systèmes. Seule une construction automatique pourra répondre à cette demande.

A l'image des ontologies monolingues, la construction des ontologies multilingues est une tâche très difficile. Cette construction est beaucoup plus difficile lorsqu'il s'agit d'ontologies multilingues d'autant plus que la construction ici comprend une phase supplémentaire qu'est la phase d'appariement des termes entre les différentes langues de l'ontologies à construire.

Dans ce chapitre nous ramenons l'acquisition d'ontologies multilingues à une simple extraction de lexique, et nous abordons les différents types de corpus exploités pour effectuer une extraction bilingue. Nous présentons notamment, dans ce chapitre les hypothèses que nous pouvons mettre en œuvre pour automatiser la construction d'ontologies bilingues. Il ne s'agit pas ici de donner une méthode complète d'acquisition, mais seulement de quelques réflexions que nous jugeons utiles pour l'automatisation de l'extraction de lexique bilingue.

1. ACQUISITION TERMINOLOGIQUE BILINGUE

L'objectif de l'extraction de lexique bilingue à partir de corpus consiste à identifier et extraire les termes et leurs traductions. Il s'agit plus précisément de repérer les termes des textes sources et des textes cibles, puis de les mettre en correspondance. Les traitements effectués peuvent ainsi se décomposer en deux étapes. La première correspond à celle de l'extraction de terminologie monolingue et la seconde à celle de l'alignement ou plutôt de l'appariement.

Selon Véronis [Véronis, 2000a], ces deux étapes ne peuvent pas être totalement modularisées dans la pratique : « la détermination des unités dans la langue source est dépendante de la langue cible (par exemple, il faut aligner d'un bloc *demande de brevet* et *patentanmeldung* alors que l'alignement peut se fractionner avec *demanda di brevetto*). ».

L'utilité des corpus pour l'acquisition automatique de données terminologiques bilingues a été mentionnée et validée depuis un certain temps [Atkins, 1990] et les projets de dictionnaires comme *Oxford-Hachette French Dictionary* [Grundy, 1996] ou *Dictionnaire Canadien Bilingue* [Roberts & Montgomery, 1996] ont fait appel aux corpus bilingues.

1.2. Corpus parallèles et corpus comparables

Deux types de corpus bilingues sont exploités pour l'extraction de lexique bilingue : le corpus parallèle et le corpus comparable. Les corpus parallèles sont constitués de textes sources et de leurs traductions. Un corpus comparable désigne un ensemble de textes de langues différentes rassemblés selon des critères similaires, en ce qui concerne le domaine, le genre, la date de publication, etc. Si l'extraction de lexique bilingue à partir de corpus parallèles est maintenant beaucoup étudiée et développée [van der Eijk, 1993], [Smadja *et al,* 1996], [Dagan & Church, 1997], [Resnik & Melamed, 1997], [Fung & McKeown, 1997], [Hiemstra *et al.,* 1997], [Hiemstra, 1998], [Gaussier, 1998], l'extraction de lexique bilingue à partir de corpus comparables est plus récente [Fung & Yee, 1998], [Rapp, 1999], [Déjean & Gaussier, 2002], [Chiao & Zweigenbaum, 2002a].

La performance de l'extraction à partir de corpus parallèles dépend principalement de la qualité de l'alignement entre les textes. Les expériences en extraction de lexique bilingue à partir de corpus parallèles généralement alignés au niveau de phrases ont montré des résultats satisfaisants. Néanmoins l'utilisation de corpus parallèles présente quelques contraintes : outre

la qualité de l'alignement, les couples de langues concernés et la taille pour l'instant modeste des corpus parallèles par rapport aux corpus monolingues, leur représentativité est généralement limitée aux domaines de spécialité (articles scientifiques, techniques, etc.).

En revanche, il est plus facile d'accéder à un corpus comparable dans un domaine donné qu'à un corpus parallèle de bonne qualité [Fung & Yee, 1998]. De plus du point de vue terminologique, les usages réels des termes dans les deux parties monolingues de corpus comparables sont bien conservés puisqu'ils n'ont pas subi de transformation due à la traduction : le vocabulaire de la langue source influence lors d'une traduction le choix d'équivalents en langue cible du traducteur. Les recherches récentes se sont ainsi intéressées à l'exploitation de corpus comparables.

Néanmoins à cause de certaines caractéristiques propres aux corpus comparables, il est plus difficile de leur appliquer les méthodes statistiques utilisées pour l'alignement de corpus parallèles [Fung, 1998]. Par exemple les traductions d'un terme du corpus de langue source peuvent être absentes dans le corpus de langue cible, les fréquences et les positions d'occurrences ne sont pas homogènes et comparables dans les corpus comparables.

1.3. Acquisition de lexique bilingue en corpus parallèles

Les techniques d'alignement au niveau des termes que nous décrivons par la suite sont développées pour une application sur des corpus parallèles alignés au niveau des phrases. Il nous semble plus intéressant de les présenter pour certaines raisons pratiques. Tout d'abord, l'une des sources importantes de textes parallèles est la mémoire de traduction, constituée souvent des phrases qui sont traductions l'une de l'autre, stockées dans des systèmes de traduction automatique. La plupart des systèmes d'alignement de textes travaillent à ce niveau et le développement de techniques d'alignement de textes au niveau de phrases semblent parvenu à maturité : les systèmes atteignent plus de 98.5% d'efficacité [Véronis, 2000a].

En général, les techniques d'alignement exploitent différents types d'information : longueur des phrases, dictionnaires bilingues, distributions lexicales et cognats, c'est à dire les occurrences qui sont identiques ou se ressemblent graphiquement, etc. Nous trouvons un état de l'art sur les systèmes d'alignement dans la [Kraif, 2001a] et l'évaluation des systèmes existants effectuée dans le cadre du projet **ARCADE** [Véronis & Langlais, 2000].

1.3.1. Problématique de l'alignement

Aux alentours de 1990, une équipe de recherche chez IBM s'est penchée sur des modèles purement statistiques de traduction automatique basée sur l'apprentissage sur corpus parallèles [Brown *et al.,* 1990], [Brown & Mercer, 1994]. Les autres recherches utilisent des textes parallèles pour extraire des lexiques bilingues, à l'aide de dictionnaires bilingues en combinaison avec l'analyse statistique [Catizone *et al.,* 1989] ou l'analyse syntaxique [Klavans & Tzoukermann, 1990].

Certaines études s'intéressent à l'alignement au niveau des mots simples par des méthodes statistiques [Dagan *et al.,* 1993], [Wu & Xia, 1994], [Resnik & Melamed, 1997]. Cependant, ces techniques qui mettent en jeu des occurrences isolées ne prennent pas en compte des phénomènes courants présents dans un texte : termes complexes, collocations, expressions, etc. La plupart des causes d'erreur d'alignement concernent les variations linguistiques, *i.e.,* les flexions, les substitutions pronominales, les mots composés, les expressions semi-figées avec insertion ou suppression d'adjectif et d'adverbe, la passivation, etc.

Dans le cas de l'alignement d'unités complexes, le traitement s'effectue de deux façons : soit en les segmentant en mots simples soit les considérant comme une seule unité. On distingue en général trois méthodes principales d'extraction de terminologie bilingue à partir des corpus parallèles :

1. **Extraction parallèle des termes dans les langues source et cible et alignement :** cette approche suit le schéma classique en effectuant d'abord l'extraction des syntagmes nominaux des textes de chaque langue. Les termes extraits de chacun des corpus sont alignés sur la base de cooccurrences de termes dans des phrases alignées.
2. **Extraction des termes dans une langue et alignement avec des séquences dans une autre langue** : Cette approche est plus souple du point de vue pratique. Elle peut être plus facilement appliquée dans le cas où l'extraction de termes monolingues n'est disponible ou pertinente que pour une des deux langues (la langue source ou cible). Par exemple pour le couple de langues anglais-français, les termes sont plus faciles à identifier en anglais qu'en français. Un autre intérêt est que cela permet de prendre en compte le fait que certaines unités complexes sont figées dans une langue et ne peuvent pas être traduites mot à mot dans une autre langue. L'application de cette approche donne en général un taux de rappel plus élevé que la méthode classique.

3. **Extraction et alignement simultanés** : Le modèle *Inversion Transduction Grammars* (ITG) expérimenté dans les travaux de Wu [Wu, 1997; 2000] génère simultanément des paires de structures syntaxiques bilingues. Le principe des ITG est d'extraire des patrons syntaxiques bilingues dans lesquels l'ordre des constituants peut être inversé en fonction du couple des langues traitées. Ces patrons sont ensuite utilisés pour trouver des équivalences de traduction au niveau des mots ou des syntagmes.

1.4. Acquisition de lexique bilingue en corpus comparable

Nous avons mentionné au début de la section 1.2 la distinction entre corpus parallèles et corpus comparables. Contrairement aux textes parallèles qui sont des traductions l'un de l'autre, les corpus comparables sont un ensemble de textes liés, sans qu'ils soient des traductions réciproques, par une relation d'identité. La notion d'identité ici est floue, mais elle rend bien compte du continuum qui existe entre des corpus parallèles bruités [Véronis & Langlais, 2000], des corpus comparables et des corpus non reliés. La relation d'identité peut être temporelle, par exemple des textes rédigés et publiés pendant une même période, ou encore liée au vocabulaire pour des corpus traitant des mêmes sujets ou domaines.

Nous avons aussi expliqué les raisons pratiques et théoriques pour lesquelles les travaux en acquisition lexicale commencent à s'intéresser à l'extraction de traductions à partir de corpus comparables. Pourtant, l'identification d'équivalents de traduction dans ce genre de textes présente de toute évidence une tâche plus complexe et ambitieuse que l'exploitation des textes parallèles. Elle reste pour l'instant du domaine de la recherche.

Tous les travaux réalisés sont fondés sur la même hypothèse, celle de la sémantique distributionnelle. Cette hypothèse suppose que le sens d'un mot peut être décrit par la distribution de ses occurrences dans un ensemble de contextes [Rajman & Bonnet, 1992].

Dans un contexte multilingue, cette hypothèse peut être reformulée ainsi : un mot de la langue A dont la distribution est similaire à celle d'un mot de la langue B est, avec une forte probabilité, traduction de ce mot. La mise en correspondance entre deux mots de langues différentes est ainsi réalisée au niveau sémantique, et le corpus bilingue est considéré comme un objet d'acquisition de connaissances et de mise à jour de ressources lexicales existantes [Déjean & Gaussier, 2002].

Partant du principe que les mots qui ont une distribution similaire sont des traductions réciproques, l'approche suivie dans les travaux réalisés jusqu'à aujourd'hui [Fung & Yee, 1998], [Rapp, 1999], [Chiao & Zweigenbaum, 2002a] et [Déjean & Gaussier, 2002] consiste à d'abord établir les distributions des mots de la langue source et de la langue cible à partir de corpus comparables. Une distribution est définie par un vecteur de contexte constitué de mots qui cooccurrent avec le mot étudié. L'empan de cooccurrence est souvent une fenêtre graphique, i.e., n mots à gauche et à droite.

Les vecteurs de contexte sont ensuite traduits d'une langue à l'autre à l'aide de lexiques bilingues partiels [Fung & Yee, 1998], [Rapp, 1999], [Chiao & Zweigenbaum, 2002a]. Des méthodes statistiques (en général des mesures de similarité) sont utilisés pour calculer la ressemblance entre les vecteurs transférés d'une langue et originaux de l'autre langue. Les candidats à la traduction sont les termes dont les vecteurs de contexte ont les meilleurs scores de similarité.

Nous décrivons dans la section sui suit les différents paramètres et les mesures de similarité qui peuvent être utilisés.

2. ACQUISITION DE LEXIQUE BILINGUE EN CORPUS COMPARABLES : FONDEMENTS THEORIQUES

Dans la pratique, l'extraction automatique de lexique bilingue à partir de corpus se compose de deux tâches : une étape de segmentation qui aboutit à la détermination des unités à traduire (mots sources) ainsi que des unités candidates à la traduction (mots cibles), et une étape d'appariement de ces unités. Après avoir présenté le problème de la segmentation ou plus précisément celui de l'extraction des termes du point de vue terminologique dans le chapitre 3, nous nous intéressons dans ce chapitre à l'appariement des mots sources et cibles.

Nous avons montré dans la section 1.4 que les travaux d'acquisition automatique de lexique bilingue à partir de corpus comparables reposent sur l'idée que la relation de traduction entre langues (comme les relations paradigmatiques au sein d'une langue, *i.e.,* les mots voisins qui sont sémantiquement proches) peut être mise en évidence par la comparaison des distributions des mots dans les corpus. L'hypothèse sous-jacente est que le sens d'un mot peut être déterminé en contexte et donc, en simplifiant, par l'ensemble des mots qui figurent dans ses contextes [Habert *et al.,* 1997].

Cette idée peut alors être exploitée pour dériver automatiquement la sémantique d'un mot à partir de l'ensemble de ses contextes dans un corpus. Partant de ce principe, la tâche consiste alors à définir d'abord un ensemble de contextes en fonction desquels la distribution de chaque mot sera calculée, puis à appliquer une mesure de similarité entre distributions. Notons qu'un élément essentiel de cette démarche est l'utilisation d'un lexique bilingue partiel et préexistant qui joue le rôle de pont entre les langues et notamment entre les distributions. La démarche est en général décomposée en trois étapes [Habert *et al.*, 1997].

1. **Définition du contexte d'un mot** en fonction du corpus exploité et des relations sémantiques recherchées.
2. **Représentation des mots** par leur lien d'association, calculé en fonction du contexte défini.
3. **Choix de la mesure de similarité** entre les représentations des mots afin de construire des classes de mots en fonction de l'application visée.

Lorsque la relation sémantique que l'on cherche est celle d'équivalence traductionnelle à travers des langues, l'espace de recherche des unités lexicales à mettre en correspondance est réduite à la phrase dans le cas des corpus parallèles. Dans le cas des corpus comparables, aucune contrainte d'alignement ne réduit l'espace de recherche. Tous les segments possibles d'une langue peuvent être mis en correspondance avec n'importe quelle unité d'une autre langue. Pour pallier cette difficulté on exploite l'idée que les contextes dans lesquels apparaît le mot *b* traduction du mot *a* doivent être similaires à ceux dans lesquels apparaît le mot *a.* Dans ce cas, les contextes d'une langue doivent être traduits dans une autre langue afin de pouvoir reconstituer l'espace de recherche. Le passage d'une langue à une autre se fait par l'intermédiaire des ressources lexicales bilingues, c'est-à-dire, dictionnaires, thésaurus, etc.

Nous abordons maintenant les différents paramètres qui permettent de mettre en évidence la relation de traduction entre deux mots : contexte, pondération, similarité, etc.

2.1. Mise en évidence de la relation de traduction à partir des contextes

2.1.1. Contexte de cooccurrence

Le choix du contexte de cooccurrence dépend de ce qu'il est censé mettre en évidence et aussi de la nature du corpus traité. Ainsi, la cooccurrence entre deux unités définie dans un texte n'a pas la même interprétation que celle définie dans une phrase.

Le contexte le plus grand est le document lui-même. Il pose en soi un certain nombre de problèmes. Dans un corpus, la taille des documents peut être extrêmement variable : du court résumé au document de plusieurs pages. Les textes longs affichent des cooccurrences peu signifiantes et les textes courts contiennent peu de cooccurrences.

Le paragraphe, repéré par un passage à la ligne, est intéressant pour deux raisons. Sa taille est en général homogène : de une à quelques phrases. Le découpage en paragraphes est aussi l'expression d'une homogénéité de contenu. Un paragraphe constitue donc souvent une unité de sens homogène, ce qui laisse présager des liaisons fortes entre les mots.

La phrase est un bon contexte de cooccurrence. Elle est le lieu idéal où l'auteur met en rapport, notamment syntagmatique, les unités lexicales. La cooccurrence de deux mots dans la phrase peut être l'expression d'une relation syntagmatique stable comme d'une relation paradigmatique. Le contexte peut y être syntaxique, c'est-à-dire contraint par les relations syntaxiques (Nom prép. Nom...). Toutefois, l'utilisation de contextes syntaxiques pose des problèmes lorsque la relation sémantique recherchée n'est pas limitée à un cadre monolingue. C'est le cas notamment de la traduction pour laquelle un mot d'une catégorie grammaticale peut être traduit par un mot d'une autre catégorie grammaticale (par exemple l'adjectif *cardiaque* dans *crise cardiaque* et traduit par le nom *heart* dans *heart attack*). Notons également que la segmentation automatique en phrases pose problème. Les algorithmes de segmentation ajoutent en général du bruit à cause de l'ambiguïté des marqueurs typographiques de séparation : ponctuation, majuscule. En particulier, la suite point-espace-majuscule apparaît aussi bien en fin de phrase que dans des abréviations.

Enfin, le contexte peut être réduit à une fenêtre de quelques mots ou même à deux mots. Dans ce dernier cas, la cooccurrence est utilisée pour la mise en évidence de bigrammes. L'ordre d'apparition ou la position peuvent aussi être utilisés : on parle alors de contexte droit ou gauche.

2.1.2. Systèmes de pondération

Nous avons mentionné précédemment que l'approche générale adoptée pour l'extraction de lexique bilingue à partir de corpus comparables est de tout d'abord définir un ensemble de contextes dans lesquels la distribution de chaque mot dans chacune des langues est calculée ; de traduire ensuite chaque élément de chaque contexte dans une des deux langues à l'aide des ressources lexicales bilingues ; puis, de comparer les mots sur la base de leurs distributions.

Le calcul de la distribution contextuelle d'un mot fait appel à une pondération. En effet on ne peut se contenter de compter simplement les cooccurrences entre un mot et un mot de contexte. Des valeurs similaires de cooccurrences ne mettent pas toujours en évidence le même lien. En particulier pour deux mots de contexte ayant des fréquences très différentes dans un corpus (l'un de faible fréquence et l'autre de forte fréquence), des mesures de cooccurrences similaires entre un mot et ces deux mots de contextes ne sont pas comparables. Le nombre de cooccurrences observé pourra par exemple dans un cas être plus élevé que le nombre de cooccurrences estimé pour une distribution aléatoire, et dans l'autre cas moins élevé. Finalement deux nombres de cooccurrences similaires seront dans ce cas le reflet pour l'un d'une dépendance entre les deux mots et pour l'autre d'une absence de dépendance. Le simple dénombrement des cooccurrences ne suffisant pas, plusieurs systèmes de pondération ont été avancés reposant sur des mesures statistiques que nous examinons ici.

Information Mutuelle

L'information mutuelle s'inscrit dans le cadre de la théorie de l'information [Shannon, 1948]. La quantité d'information apportée par un mot *j* sur la présence d'un autre mot *i* est l'information mutuelle *IM(i,j).* Cette valeur s'exprime par le rapport de la probabilité d'observer *i* sachant que l'on a observé *j* sur la probabilité de *i,* soit :

$$IM(i,j) = \log_2 \frac{p(i/j)}{p(i)}$$

Avec la formule de Bayes *P(i\j)* s'écrit *P(i,j)/P(j) où P(i,j)* est la probabilité d'observer *i* et *j* simultanément. L'information mutuelle devient alors :

$$IM(i,j) = \log_2 \frac{p(i,j)}{p(i) \times p(j)}$$

La probabilité d'occurrence d'un mot *i* est calculée à partir de sa fréquence dans un corpus donné *f*(i), normalisée par le nombre total de mots dans le corpus soit *N.* La probabilité de cooccurrences de deux mots *i* et *j* est calculée à partir de leur cooccurrences dans le corpus *f(i,j),* divisée par le nombre total de mots dans le corpus :

$$IM(i,j) = \log_2 N \times \frac{f(i,j)}{f(i) \times f(j)}$$

Cet indicateur est utilisé pour mesurer l'association entre des mots simples composant des collocations [Church & Hanks, 1990] ou des termes complexes de type nominal [Rapp, 1995]

ainsi que pour les mettre en relation à partir de corpus comparables bilingues [Iram *et al.*, 1999].

Notons que l'information mutuelle est l'expression de liaisons récurrentes et exclusives de mots. Plus deux mots sont dans les mêmes contextes et rien que dans les mêmes contextes, plus l'information mutuelle a une valeur importante. Or beaucoup de mots cooccurrent avec des mots et ne sont pas liés de manière exclusive avec ces mots. L'information mutuelle a donc tendance à favoriser ces liaisons exclusives et notamment l'association entre les mots ayant de faibles occurrences.

Mesure du Chi 2

La statistique du χ^2 (Chi 2) mesure la dépendance entre deux mots à partir de la table de contingence du tableau :

	j	$\neg j$	
i	a	c	$i_1 = a + c$
$\neg i$	b	d	$i_0 = b + d$
	$j_1 = a + b$	$j_0 = c + d$	$N = a + b + c + d$

Tableau 2 : Table de contingence pour la dépendance de deux unités i et j.

Dans cette table, a est le nombre de contextes dans lesquels i et j apparaissent tous les deux, b est le nombre de contextes où j est présent mais i est absent, etc. La mesure d'association χ^2 est alors définie comme suit :

$$\chi^2(i,j) = \frac{N(ad - cb)^2}{j_1 i_1 i_0 j_0}$$

La distance du χ^2 ainsi définie mesure le degré de dépendance des mots. Plus les deux mots i et j apparaissent ensemble dans les mêmes contextes, plus grande est la valeur ad. Plus ils sont absents tous les deux des mêmes contextes, plus la valeur cb est petite. Ainsi, deux mots indépendants ont un χ^2 nul. Quand le nombre des contextes est trop faible, l'emploi du Chi 2 rencontre ses limites [Muller, 1997].

Rapport de vraisemblance

D'après plusieurs auteurs [Dunning, 1993], [Baayen, 2001], [Habert & Jardino, 2003], les modèles basés sur l'hypothèse d'une distribution normale des occurrences ne sont pas adaptés

à l'étude des événements rares. En effet, dans ces modèles, comme avec l'information mutuelle, la cooccurrence de deux hapax devient hautement improbable. Il propose alors d'associer une probabilité p constante d'obtenir un mot donné en choisissant une occurrence au hasard. L'apparition de k occurrences d'un mot est alors considérée comme issue de tirages indépendants de probabilité p :

$$\begin{aligned} loglike(i,j) &= \textstyle\sum_{ij} \log \frac{k_{ij}N}{C_i R_j} \\ &= k_{11} \log \frac{k_{11}N}{C_1 R_1} + k_{12} \log \frac{k_{12}N}{C_1 R_2} + k_{21} \log \frac{k_{21}N}{C_2 R_1} + k_{22} \log \frac{k_{22}N}{C_2 R_2} \\ C_1 &= k_{11} + k_{12} \\ C_2 &= k_{21} + k_{22} \\ R_1 &= k_{11} + k_{21} \\ R_2 &= k_{12} + k_{22} \\ N &= k_{11} + k_{12} + k_{21} + k_{22} \end{aligned}$$

K_{11} correspond aux cooccurrences des deux mots i et j

K_{12} est la différence entre le nombre d'occurrences de i et K11

K_{21} est la différence entre le nombre d'occurrences de j et K_{11}

K_{22} est le nombre total d'occurrences dans le corpus K_{12}-K_{21}+K_{11}

Cette mesure est couramment utilisée pour calculer la distribution des mots dans les corpus bilingues comparables [Fung & McKeown, 1997], [Rapp, 1999], [Déjean & Gaussier, 2002].

tf.idf

La valeur *tf.idf (term frequency* x *inverse document frequency)* est un indice classique utilisé en recherche d'information pour la pondération des termes. En général, $tf_{i,j}$ désigne la fréquence d'un terme *i* dans un document *j,* normalisée par la fréquence maximale des termes dans *j :*

$$tf_{i,j} = \frac{f_{i,j}}{\max_{l,j}} \bullet$$

Pourtant, la fréquence ne peut pas être le seul critère utilisé, puisque deux mots peuvent avoir la même fréquence dans un document mais l'un peut apporter plus d'informations sur le document que l'autre s'il est plus rare dans le corpus. Pour exploiter ce phénomène, l'indice *idf* est pris en compte et met en évidence le pouvoir de discrimination d'un terme *i :*

$$idf_i = \log \frac{N}{n_i}$$

N est le nombre total de documents dans le corpus et n_i est le nombre de documents dans lesquels se trouve le terme *i.*

Finalement le poids d'un terme *i* dans un document *j* est une combinaison de sa fréquence dans le document et de son pouvoir de discrimination dans le corpus :

$$p_{i,j} = tf_{i,j} \log \frac{N}{ni}$$

Lorsque l'on applique ce système de pondération au cas des distributions de mots dans des contextes [Fung & Yee, 1998], [Chiao & Zweigenbaum, 2002a], le document *j* est remplacé par le terme dont on calcule la distribution. Nous pouvons ainsi reformuler le *tfi,j* par la fréquence de cooccurrence d'un terme *j* dans le contexte d'un autre terme *i* (soit le nombre de cooccurrences de *i* et *j : $cooc_{i,j}$*) normalisée par la fréquence maximale de cooccurrence du terme *j* sur l'ensemble des contextes *(max_{coocj}) :*

$$tf_{i,j} = \frac{cooc_{i,j}}{\max_{cooc_j}}$$

Le *idf* devient :

$$idf_i = 1 + \log \frac{\max_{cooc}}{cooc_i}$$

Où cooci est le nombre total de mots pour lesquels i est un mot de contexte et max_{cooc} est le nombre maximal de mots de contexte dans le corpus. Pour un terme j, le poids d'un mot de contexte i est ainsi obtenu par :

$$p_{i,j} = tf_{i,j} \times idf_i$$

2.1.3. Modèle vectoriel et similarité entre vecteurs de contexte

Afin de représenter l'ensemble des contextes d'un mot, une idée est de considérer la matrice des associations mot/mot, en associant à chaque mot l'ensemble des mots de ses contextes. Cette idée introduit la notion de vecteur de contexte généralisant la notion de contexte simple constitué de cooccurrences. Le modèle vectoriel développé par Salton [Salton *et al,.* 1974] dans le cadre de la recherche d'information peut ainsi être exploité. Pour un corpus donné, il s'agit de calculer la matrice des cooccurrences entre les mots (matrice carrée et symétrique). Un mot est donc considéré comme un vecteur dans l'espace des mots. La coordonnée d'un mot sur une dimension est la force d'association entre ce mot et le mot

associé à la dimension en question (information mutuelle, *tf.idf...)*. Notons que cette matrice est creuse (remplie essentiellement de zéros).

La relation entre deux mots est alors calculée par une mesure de similarité (cosinus, Jaccard) ou par une mesure de distance (distance de Manhattan, distance euclidienne) entre deux vecteurs de l'espace des mots.

Un des problèmes de l'approche vectorielle est sa complexité élevée. En particulier, la comparaison des vecteurs entre eux nécessite la comparaison des mots deux à deux. La complexité est égale au nombre total des mots du corpus au carré (en $O(n^2)$). Sur des textes de taille importante le temps et la mémoire nécessaires deviennent rapidement rédhibitoires. Il est donc raisonnable de limiter la taille de la matrice pour réduire l'espace de recherche. La réduction du nombre de dimensions (qui est égal au nombre de mots dans le corpus) est possible en éliminant les mots vide ou grammaticaux qui, même s'ils ne sont pas nombreux cooccurrent souvent avec les autres mots.

Une fois les contextes des mots d'une langue construits et pondérés, ils sont traduits à l'aide des ressources bilingues disponibles afin de permettre la comparaison avec les contextes des mots de l'autre langue. Cette comparaison est effectuée classiquement par des mesures de similarité ou des distances présentées ici. Pour comparer deux vecteurs v, w de longueur n, deux mesures de similarité sont très utilisées :

Cosinus

$$\cos(v,w) = \frac{\sum_{k=1}^{k=n} v_k w_k}{\sqrt{\sum_{k=1}^{k=n} v_k^2} \times \sqrt{\sum_{k=1}^{k=n} w_l^2}}$$

Coefficient de Jaccard

$$Jaccard(v,w) = \frac{\sum_{k=1}^{k=n} v_k w_k}{\sum_{k=1}^{k=n} v_k^2 + \sum_{k=1}^{k=n} w_l^2 - \sum_{k=1}^{k=n} v_m w_m}$$

Distance de Minkowski Au contraire des mesures de similarité, les distances accordent une valeur maximale à deux objets complètement différents et minimale (0) à deux objets identiques. Les mesures les plus utilisées sont la distance euclidienne ou celle de Manhattan, qui ne sont en fait que des cas particuliers de la mesure de Minkowski :

$$D_p(v,w) = \left(\sum_{k=1}^{k=n} |v_k - w_k|^p \right)^{\frac{1}{p}}$$

Où $p = 1$ donne la distance de Manhattan et $p = 2$ la distance euclidienne.

2.1.4. Similarité interlangue

L'hypothèse sur laquelle sont fondées les méthodes exposées jusqu'ici est que la ressemblance des distributions contextuelles de deux mots est un signe qu'une relation de traduction les lie. Une hypothèse supplémentaire peut être avancée dans un cadre bilingue. Si un mot de la langue a est proche de mots au sens d'une similarité de leurs distributions dans le corpus de langue a, qu'un mot de la langue *b* est proche de ces mêmes mots (à la traduction près par une ressources bilingue pré-existante) au sens d'une similarité de leurs distributions dans le corpus de langue b, alors ces deux mots ont de fortes chances d'être traductions l'un de l'autre. Par exemple, si *médecin* a une distribution similaire à celle de *infirmière* de la même façon que *doctor* a une distribution proche de celle de *nurse* et que *infirmière* est la traduction de *nurse* alors *médecin* a de grandes chances d'être la traduction de *doctor*.

Cette hypothèse a été mise en œuvre et expérimentée dans [Déjean & Gaussier, 2002]. Nous avons vu que pour une même langue et un même corpus la similarité des distributions entre les mots peut être exploitée pour le regroupement ou le rapprochement de ceux-ci. Le modèle vectoriel et la comparaison des distributions à l'aide d'une mesure de similarité sont utilisés dans ce modèle dans un cadre monolingue pour chacun des corpus. Ainsi, la première étape est la même que celle de l'approche classique et consiste à définir et à constituer les contextes de cooccurrences pour calculer les proximités sémantiques au sein d'une même langue. Il s'agit de rapprocher pour un même corpus d'une langue donnée les mots qui partagent les mêmes contextes de cooccurrence.

Les similarités résultantes dans les deux langues sont ensuite utilisées pour établir que deux mots sont traductions l'un de l'autre. Ici aussi les ressources lexicales bilingues participent au rapprochement. Les similarités calculées entre d'une part les deux mots de langues différentes et d'autre part les entrées de la ressource bilingue servent à estimer la probabilité que ces deux mots sont traductions l'un de l'autre.

En effet l'idée développée dans ce modèle est que si deux mots de langues différentes ont les mêmes (en passant par une ressource bilingue) mots proches (au sens d'une mesure de

similarité dans leur corpus d'origine) alors ils ont toutes les chances d'être traductions l'un de l'autre.

3. RESUME

Dans le cadre d'une acquisition lexicale bilingue, l'acquisition consiste à extraire des termes et leurs traductions. On distingue deux types de corpus utilisés pour l'extraction lexicale bilingue : le corpus parallèle et le corpus comparable. Les corpus parallèles contiennent des textes qui sont traductions mutuelles. Des textes comparables sont des textes de langues différentes regroupés selon des critères similaires concernant le domaine, le genre, la date de publication, etc.

Une comparaison entre les méthodes exploitant des corpus parallèles et celles exploitant des corpus comparables a été présentée. L'acquisition lexicale bilingue en corpus comparables étant la tendance actuelle, on a abordé, dans ce chapitre les différentes hypothèses qui peuvent nous servir dans cette tâche, on a notamment présenté les différentes mesures et paramètres susceptibles de nous aider à acquérir un lexique bilingue

CONCLUSION

« En toute chose, c'est la fin qui est essentiel »

ARISTOTE

Nous avons présenté, dans ce mémoire, une méthodologie dédiée à la conception des ontologies de domaines et détaillé l'architecture fonctionnelle du système OntoLogos à développer à la lumière de cette méthodologie. Notre contribution s'inscrit dans la lignée des efforts menés pour combler les lacunes entre l'acquisition d'ontologies à partir de corpus textuels d'une part et d'autre part la réutilisation de ressources ontologiques existantes. Dans le but d'une acquisition ontologique bilingue, nous avons abordé les techniques qui peuvent être utilisées pour produire des ontologies bilingues

Aux travers des différentes chapitres de ce mémoire, nous avons essayé de poser les fondements théoriques de l'approche développée, néanmoins, beaucoup de choses restent à faire, notamment le développement de nouvelles méthodes adaptées à l'importation des différentes sources de connaissances, telles que les XML schémas, DTD, les Réseaux sémantiques, etc. En outre un effort supplémentaire doit être fait pour élaguer les termes présents dans les ressources importées et ne faisant pas partie de la terminologie du domaine considéré. En ce qui concerne l'acquisition d'ontologies bilingues, une méthode respectant les principes introduits dans ce mémoire va être développée afin de permettre une construction bilingue à partir de corpus comparables. Ces perspectives théoriques nécessitent d'être complétées par des perspectives applicatives permettant de développer à terme des outils implémentant les différentes méthodes de construction que nous avons détaillé.

BIBLIOGRAPHIE.

ABNEY S. (1990). Rapid Incremental Parsing with Repair. In *Proceedings of the 6th New OED Conference: Electronic Text Research,* Waterloo, Ontario, Canada.

AGARWAL R. (1995). *Semantic Feature Extraction from Technical Texts with Limited Human Intervention*. PhD thesis, Mississippi State University, Etats-Unis.

AGUIRRE, E., ANSA, O., HOVY, E., AND MARTINEZ, D. (2000). Enriching very large ontologies using the WWW. In Proceedings of the Workshop on Ontology Construction of the European Conference of AI (ECAI-00).

ALFONSECA E. AND MANANDHAR S. (2002A). An unsupervised method for general named entity recognition and automated concept discovery. In Proceedings of the 1st International Conference on General WordNet, Mysore, India.

ALFONSECA E. AND MANANDHAR S. (2002B). Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures, EKAW-2002, Siguenza, Spain. Published in Lecture Notes in Artificial Intelligence 2473 (Springer Verlag).

ALLALGA W. & SELLAMI M. (2005): OntoLogos : vers un outil d'assistance à la conception incrémentale coopérative des ontologies. In the 2th Mediterranean Seminar on Engineering Education (2MSEE), Algiers, Algeria.

ARPIREZ J., GÓMEZ-PEREZ A., LOZANO A. ET PINTO S. (1998) : (ONTO)2Agent : An ontology-based WWW broker to select ontologies. Paper presented at the Workshop on Applications of Ontologies and PSMs, Brighton, England.

ARPIREZ J., CORCHO O., FERNANDEZ-LOPEZ M., AND GOMEZ-PEREZ A. (2001). WebODE : a Workbench for Ontological Engineering. In First international Conférence on Knowledge Capture (K-CAP'Ol), pages 6-13, Victoria, Canada, 21-23 Octobre 2001. ACM.

AUROUX W. (1984). Nouveau vocabulaire des études philosophiques (Hachette ed.). Paris, Hachette.

ASSADI H., BOURIGAULT D. (1996). Acquisition de connaissances à partir de textes : Outils informatiques et éléments méthodologiques. in Actes du dixième congrès Reconnaissance de Formes et Intelligence Artificielle (RFIA'96), Rennes, pp. 505-514.

ASSADI H. (1998) : Construction d'ontologies à partir de textes techniques : Application aux systèmes documentaires. Thèse de doctorat, Paris VII ème

AUSSENAC-GILLES, N, BIÉBOW B, SZULMAN S. (2000A). Corpus Analysis For Conceptual Modelling.

AUSSENAC-GILLES N, BIÉBOW B, SZULMAN S (2000B). Revisiting Ontology Design: A Methodology Based on Corpus Analysis. In Dieng R., Corby O. (editors) 12th International Conference in Knowledge Engineering and Knowledge Management (EKAW'00). Juan-Les-

N. Aussenac-Gilles, B. Biebow, and S. Szulman. D'une méthode à un guide pratique de modélisation des connaissances à partir de textes. In 5es journées Terminologie et Intelligence Artificielle, pages 41-53, Strasbourg, France, 2003.

BACHIMONT B. (1996). Herméneutique matérielle et artéfacture : critique du formalisme en intelligence artificielle. Des machines qui pensent aux machines qui donnent à penser. PhD thesis, Ecole Polytechnique.

BACHIMONT B. (2000a). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. In Z. M. Charlet J., Kassel G., Bourigault D., (Ed.), Ingénierie des connaissances. Évolution Récentes et nouveaux défis Paris: Eyrolles, 305-

323.

BACHIMONT B., ISAAC A., ET TRONCY R. (2002) : Semantic Commitment for Designing Ontologies : A Proposal. In A. Gomez-Pérez and V.R. Benjamins, editors, 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02), volume (2473) of Lecture Notes in Artificial Intelligence, pages 114-121, Sigûenza, Espagne, 1-4 Octobre 2002. Springer Verlag.

BECHHOFER S., HORROCKS I., GOBLE C., ET STEVENS R. (2001) : OilEd: a Reasonable Ontology Editor for the Semantic Web. In Joint German/Austrian conférence on Artificial Intelligence (KI'Ol), volume (2174) of Lecture Notes in Artificial Intelligence, pages 396-408, Vienne, Autriche, 2001. Springer Verlag.

BENVENISTE E. (1974). *Problèmes de linguistique générale,* volume 2. Gallimard.

BERNARAS A., LARESGOITI I. ET CORERA J. (1996). Building and Reusing Ontologies for Electrical Network Applications. Paper presented at the Procedings of the 12th ECAI96.

BERNERS-LEE T. , HENDLER J. ET LASSILA O. (2001). The Semantic Web. Scientific American, mai 2001. (Scientific American Feature Article The Semantic Web May 2001.htm).

BLAZQUEZ M., FERNANDEZ M., GARCIA-PINAR J. M. ET GOMEZ-PEREZ A. (1998). Building Ontologies at the Knowledge Level using the Ontology Design Environment. Paper presented at the Proc. of the 11th KAW, Banff, Canada.

BOUAUD J., HABERT B., NAZARENKO A. & ZWEIGENBAUM P. (1997). Regroupements issus de dépendances syntaxiques en corpus : categorisation et confrontation avec deux modelisations conceptuelles. In *Actes de Ingenierie de la Connaissance,* Roscoff, France.

BRACHMAN R.J. AND SCHMOLZE J.G. (1985) : An Overview of the KL-ONE Knowledge Representation System. Cognitive Science, 9(2) : 171-216.

BRILL E. (1992). A Simple Rule-Based Part of Speech Tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing, ANLP92,* Trento, Italie.

BRILL E. (1994). Some Advances in Transformational-Based Part of Speech Tagging. In *Proceedings of the 12th National Conference on Artificial Intelligence, AAAI'94,* Seattle, Washington, Etats-Unis.

BORGO S., GUARINO N. ET MASOLO C. (1996). Stratified Ontologies: the case of physical objects. Paper presented at the ECAI96. Workshop on Ontological Engineering, Budapest.

BORST W. N. (1997). Construction of Engineering Ontologies. Center for Telematica and Information Technology, University of Tweenty, Enschede, NL.

BOURIGAULT D. (1992). *lexter,* un logiciel d'extraction de terminologie. In *Actes du 2eme Collogue International de TermNet,* Avignon, France.

BOURIGAULT D. (1994) : Extraction et structuration automatiques de terminologie pour l'aide à l'acquisition des connaissances à partir de textes. PhD thesis, Ecole des Hautes Etudes en Sciences Sociales, Paris.

BOURIGAULT D. (2002). Analyse distributionnelle etendue pour la construction d'ontologies a partir de corpus. In *Actes de la conference Traitement automatique des langues naturelles, TALN'02,* Nancy, France.

BOURIGAULT D & FABRE C. (2000). Approche linguistique pour 1'analyse syntaxique de corpus. *Cahiers de Grammaires,* 25, 131-151.

BOURIGAULT D & JACQUEMIN C. (2000). Construction de ressources terminologiques. In J.-M. PlERREL, Ed., *Ingenierie des langues,* chapitre 9, p. 215-223. Hermes.

BOURIGAULT D & SLODZIAN M. (1999). Pour une terminologie textuelle. *Terminologies nouvelles,* 19, 29-32.

BROWN P. F., DELLA PLETRA V. J., DESOUZA P. V., LAI J. C. & MERCER R. L. (1992). Class-based n-gram Models of Natural Language. *Computational Linguistics,* 18, 467-479.

BUCHANAN G., WILKINS C. (1993). Readings in Knowledge Acquisition and Learning. Automating the Constructionand Improvement of Expert Systems. Morgan Kaufmann, San Mateo, CA.

DE CHALENDAR G. (2001). *SVETLAN', un systeme de structuration du lexique guide par la determination automatique du contexte thematique.* These de doctorat, Universite de Paris XI, Prance.

DE CHALENDAR G. & GRAU B. (2000). SVETLAN' ou comment classer les mots en fonction de leur contexte. In *Actes de la conference Traitement Automatique des Langues Naturelles, TALN'OO,* Lausanne, Suisse.

CHARLET J. (2003). L'ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales. Mémoire d'habilitation à diriger des recherches, Université Pierre et Marie Curie, 2003.

P.P. CHEN. *The Entity-Relationship Model - Toward a Unified View of Data. ACM Trans. Data-base Syst., l(l):9-36, 1976.*

CHURCH K. W. (1988). A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of the 2nd Conference on Applied Natural Language Processing, ANLP'88,* Austin, Texas, Etats-Unis.

CHURCH K. W. & HANKS P. (1989). Word Association Norms, Mutual Information, and Lexicography. In *Proceeding of the 27th Annual Meeting of the Association for Computational Linguistics, ACL'89,* Vancouver, Canada.

CHURCH K. W. & HANKS P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics,* 16(1), 22-29.

CHURCH K. W., GALE W., HANKS P. & KINDLE D. (1991). Using Statistics in Lexical Analysis. In U. ZERNICK, Ed., *Lexical Acquisition: Using On-line Resources to Build a Lexicon,* p. 115-164. Laurence Erlbaum.

CONDAMINES A. & AMSLI P. (1993). Terminology between Language and Knowledge: an Example of Terminological Knowledge Base. In *Proceedings of the 3rd International Congress on Terminology and Knowledge Engineering,* Cologne, Allemagne : Indeks Verlag.

CONDAMINES A., REBEYROLLE J. (2000). Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode. J. Charlet, M. Zacklad, G. Kassel & D. Bourigault, (editors). : Ingénierie des Connaissances, évolutions récentes et nouveaux défis. Paris : Eyrolles.

COUTURAT L. (1903). Opuscules et fragments inédits de Leibniz. Paris.

CRUSES D. A. (1986). *Lexical Semantics.* Textbooks in Linguistics. Cambridge University Press.

H. CZAP & W. NEDOBITY, Eds. (1990). *Proceedings of the 2nd International Congress on Terminology and Knowledge Engineering,* Trier, Allemagne. Indeks Verlag.

DAGAN I. & CHURCH K. (1997). *TERMIGHT:* Coordinating Man and Machine in Bilingual Terminology Acquisition. *Machine Translation,* 12(1-2), 89-107.

DAILLE B. (1999). Identification des adjectifs relationnels en corpus. In *Actes de la conférence Traitement Automatique des Langues Naturelles, TALN'99,* Cargese, France.

DAILLE B. (2001). Identification des adjectifs relationnels. *TAL (Traitement automatique des langues),* 42(3), 815-832

DAILLE B. (2002). Découvertes linguistiques en corpus. Habilitation à diriger des recherches, Universite de Nantes, Prance.

DAML. (2002) : Darpa Agent Markup Language. http://www.daml.org/2000/10/daml-ont.html.

DAVID S. & PLANTE P. (1990). De la nécessite d'une approche morpho-syntaxique dans 1'analyse de textes. *Intelligence Artificielle et Sciences Cognitives au Quebec,* 3(3), 140-154

DAVID S. & PLANTE P. (1991). Le progiciel *TERMING :* de la necessite d'une approche morpho-syntaxique pour le depouillement terminologique de textes. In *Actes du Colloque sur les industries de la langue,* Quebec, Canada.

DE SAUSSURE F. (1916). *Cours de linguistique générale.* Editions Payot et Rivages, edition de 1996.

DIENG R., CORBY O., GANDON F., GIBOIN A., GOLEBIOWSKA J., MATTA N. et RIBIERE M. (2001) Méthodes et outils pour la gestion des connaissances : une approche pluridisciplinaire du knowledge management (2nd édition). Dunod Edition, Informatique, Séries Systèmes d'Information.

R. DJELOUAH, B. DUVAL et S. LOISEAU. Validation and Réparation of Knowledge Bases. In *Proceedings ofthe International Symposium on Méthodologies for Intelligent Systems (ISMIS'02),* volume 2366, pages 312-320. Springer-Verlag LNAI, 2002. [41] P. DOBREV, A.

ENCYCLOPAEDIA UNIVERSALISE. (2000). Dictionnaire de la philosophie. Paris: A. Michel, Encyclopaedia Universalise 2000.

ENGUEHARD C. (1992). *Acquisition naturelle automatique d'un réseau sémantique.* Thèse de doctorat, Université de Technologie de Compiegne, France.

ENGUEHARD C. & PANTERA L. (1995). Automatic Natural Acquisition of a termino logy. *Journal of Quantitative Linguistics,* 2(1), 27-32.

EUZENAT J., NAPOLI A., ET DUCOURNAU R. (2000) : Les représentations des connaissances par objets. Technique et science informatiques, 19(1).

EUZENAT J. (2002). Eight Questions about Semantic Web Annotations. IEEE Intelligent Systems, 17(2) :55-62, Mars-Avril 2002.

EVANS D. A. & ZHAI C. (1996). Noun-Phrase Analysis in Unrestricted Text for Information Retrieval. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, ACL'96,* Santa-Cruz, Etats-Unis.

FAATZ A. AND STEINMETZ R. (2002). Ontology enrichment with texts from the WWW. Semantic Web Mining 2nd Workshop at ECML/PKDD-2002, 20th August 2002, Helsinki, Finland.

FABRE C. (1996). *Interprétation automatique des séquences binominales en anglais et en français. Application a la recherche d'informations.* Thèse de doctorat, Université de Rennes 1, France.

FABRE C. & SEBILLOT P. (1999). Semantic Interpretation of Binominal Sequences and Information Retrieval. In *Proceedings of the International ICSC Congress on Computational*

Intelligence: Methods and Applications, CIMA '99, Symposium on Advances in Intelligent Data Analysis, AID A '99, Rochester, Etats-Unis.

FARQUHAR A., FIKES R., PRATT W., ET RICE J. (1995) : Collaborative Ontology Construction for Information Integration. Rapport de recherche KSL-95-63, Knowledge Systems Laboratory, Department of Computer Science.

FAURE D. & NEDELLEC C. (1999). Knowledge Acquisition of Predicate Argument Structures from Technical Texts using Machine Learning: the System ASIUM. In D. F. R. STUDER, Ed., *Proceedings of the llth European Workshop EKAW'99,* Dag-stuhl, Allemagne : Springer-Verlag.

FENSEL D., HORROSCKS I., VAN HARMELEN F., DECKER S., ERDMAN M. ET KLEIN M. (2000) : OIL in a nutshell, in Proceedings of European Knowledge Acquisition Workshop (EKAW'2000), Springer-Verlag LNAI 1937, pages 1-16.

FERNANDEZ-LOPEZ M., GÓMEZ-PEREZ A. ET JURISTO N. (1997). Methontology: From Ontological Art Toward Ontological Engineering. Paper presented at the Spring Symposium Series on Ontological Engineering. AAAI97, Stanford, USA.

FRANTZI K. & ANANIADOU S. (1996). Extracting Nested Collocations. In *Proceedings of the International Conference on Computational Linguistics, COLING'96,* Copenhague, Danemark.

FRANTZI K., ANANIADOU S. & MIMA H. (2000). Automatic Recognition of Multi-Word Terms: the C-Value/NC-Value Method. *Journal of Digital Library,* 3(2), 115-130.

FÜRST F. (2002). L'ingénierie Ontologique. Nantes: Institut de Recherche en Informatique de Nantes.

GANDON F. (2002). Distributed Artificial Intelligence and Knowledge Management: ontologies and multi-agent Systems for a corporate semantic Web. Thèse de Doctorat, INRIA - University of Nice-Sophia Antipolis.

GARCIA D (1998). Exploitation, pour 1'elaboration de requetes de filtrage de textes, des connaissances causales detectees par COATIS. In *Actes de la Rencontre Internationale sur le Filtrage et le Resume automatique, RIFRA '98,* Sfax, Tunisie.

GARCIA D., AUSSENAC-GILLES N. & COURCELLE A. (2000). Exploitation, pour la modelisation, des connaissances causales reperees par COATIS dans les textes. In J. CHARLET, M. ZACKLAD, G. KASSEL & D. BOURIGAULT, Eds., *Ingenierie des connaissances : evolutions recentes et nouveaux defis.* Eyrolles.

GENEST D. et SAUVAT E. (1998) :. A Platform Allowing Typed Nested Graphs: How CoGITo Became CoGITaNT. In Proceedings of the International Conférence on Conceptual Structures (ICCS'98), volume 1453, pages 154-161. Springer-Verlag LNAI, 1998.

GOMEZ-PEREZ A. (1999a). Tutorial on Ontological Engineering. Paper presented at the Proc. IJCAI99.

GRAF B. (1996). Lexique de philosophie (Éditions du Seuil ed.). Paris: Éditions du Seuil.

GRUBER T.R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition, 199-200.

GREFENSTETTE G. (1992). *SEXTANT:* Exploring Unexplored Contexts for Semantic Extraction from Syntactic Analysis. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, ACL'92,* Newark, Delaware, Etats-Unis.

GREFENSTETTE G. (1994a). Corpus-Derived First, Second and Third-Order Word Affinities. In

Proceedings of EURALEX'94, Amsterdam, Pays-Bas.

GREFENSTETTE G. (1994b). *Explorations in Automatic Thesaurus Discovery*. Dordrecht: Kluwer Academic Publishers.

GRUNINGER M. et Fox M. S. (1995): Methodology for the design and evaluation of ontologies. In Proceedings of the Workshop on Basic Ontological Issues on Knowledge Sharing of the IJCAI'95 conference, 1995.

GUARINO N. (1997a) : Some organizing principles for a unified top-level ontology. AAAI Spring Symposium on Ontological Engineering, 57-63.

GUARINO N. (1997b). Understanding, building and using ontologies. International J. Human-Computer Studies, 46, 293-310.

GUARINO N. et WELTY C. (2000a) : A Format Ontology of Properties. In R. DIENG et O. CORBY, reds., Knowledge Engineering and Knowledge Management: Methods, Models and Tools. Proceedings of EKAW'2000, pages 97-112. Springer-Verlag.

GUARINO N. et WELTY C. (2000b): Identity, Unity, and Individuality: Towards a Format Toolkit for Ontological Analysis. In Proceedings of the 14th European Conférence on Artificial Intelligence (ECAI'2000), pages 219-223. IOS Press.

GUPTA, K.M., AHA, D.W., MARSH, E., AND MANEY, T. (2002). An architecture for engineering sublanguage WordNets. In Proceedings of the First International Conference On Global WordNet (pp. 207-215). Mysore, India: Central Institute of Indian Languages.

HABERT B. & JACQUEMIN C. (1993). Noms composes, termes, dénominations complexes : problématiques linguistiques et traitements automatiques. *TAL (Traitement automatique des langues),* 34(2).

HABERT B., NAZARENKO A. & SALEM A. (1997). *Les linguistiques de corpus.* Armand Collin/Masson, Paris

HABERT B. & FABRE C. (1999). Elementary Dependency Trees for Identifying Corpus-Specific Semantic Classes. *Computer and the Humanities,* 33(3), 207-219.

HARRIS Z.S. (1968). Mathematical structures of language. Wiley, New York.

HARRIS Z., GOTTFRIED M., RYCKMAN T., MATTICK (JR) P., DALADIER A., HARRIS T. N. & HARRIS S. (1989). The Form of Information in Science, Analysis of Immunology Sublanguage. *Boston Studies in the Philosophy of Science,* 104.

HAHN U., AND MARKÓ K. (2001). Joint knowledge capture for grammars and ontologies. Proceedings of the First International Conference on Knowledge Capture K-CAP 2001: Victoria, BC, Canada.

HEARST M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th International Conference on Computational Linguistics, COLING'92,* Nantes, France.

HEARST M. A. (1998). Automated Discovery of WordNet Relations. In Christiane Fellbaum (Ed.) WordNet: An Electronic Lexical Database, MIT Press, pp. 132--152.

HEGEL G. W. F. (1812). Science de la Logique. Premier tome, premier livre : L'Être (P.-J. L. e. G. Jarczyk, Trans. 1812 ed.). Paris: Aubier Montaigne.

HEID U., JAUSS S., KRUGER K. & HOHMANN A. (1996). Term Extraction with Standard Tools for Corpus Extraction. Experience from German. In *Proceedings of the 4th*

International Congress on Terminology and Knowledge Engineering, TKE'96, Vienne, Autriche.

HENDLER J. ET MCGUINNESS D. (2001) : The Darpa Agent Markup Language, in IEEE Intelligent System, http ://www.daml.org.

HORROCKS I. (1998) : Using an Expressive Description Logic: FaCT or Fiction? In 6th International Conference on Principles of Knowledge Representation and Reasoning (KR'98), pages 636-649, Trento, Italie, 2-5 Juin 1998. Morgan Kaufmann.

HWANG, C. H. (1999). Incompletely and imprecisely speaking: Using dynamic ontologies for representing and retrieving information. In. Proceedings of the 6th International Workshop on Knowledge Representation meets Databases (KRDB'99), Linköping, Sweden, July 29-30, 1999.

ISAAC A. (2001): Vers la mise en œuvre informatique d'une méthode de conception d'ontologies. Master's thesis, Institut des Sciences Humaines Appliquées, Université Paris IV, Paris, France.

JACQUEMIN C. (1997). *Variation terminologique : reconnaissance et acquisition automatique de termes et de leurs variantes en corpus.* Habilitation à diriger des recherches, Universite de Nantes, France.

JOUIS C (1993). *Contributions a la conceptualisation et a la modelisation des connaissances a partir d'une analyse linguisique de textes.* These de doctorat, Universite de

JOUIS C.. (1995). *SEEK,* un logiciel d'acquisition de connaissances utilisant un savoir linguistique sans employer de connaissances sur le monde externe. In *Actes des 6es Journees Acquisition, Validation, JAVA'95,* Grenoble, France.

JOUIS C, BISKRI I., DESCLES J.-P., PRIO F. L., MEUNIER J.-G., MUSTAPHA W. & NAULT G. (1997). Vers 1'integration d'une approche semantique linguistique et d'une approche numerique pour un outil d'aide a la construction de bases terminologiques. In *Actes des 1ere Journees scientifiques et techniques du reseau francophone de I'ingenierie de la langue de I'AUPELF-UREF,* Avignon, France.

JUSTESON J. S. & KATZ S. M. (1995). Technical Terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering,* 1(1), 9-27.

KABBAJ A. (2000) :From Prolog++ to Prolog+CG : a CG object-oriented logic programming language. In Proceedings of the International Conference on Conceptual Structures (ICCS'OO), volume 1867, pages 540-554. Springer-Verlag LNAI.

KANT E. (1781). Critique de la raison pur. Paris: coll. Garnier- Flammarion.

KAYSER D. (1997) : La représentation des connaissances. Hermès, 1997.

KIETZ JU., MAEDCHE A, VOLZ R. (2000). A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet. In: Aussenac-Gilles N., Biébow B., Szulman S. (editors) EKAW'00 Workshop on Ontologies and Texts. Juan-Les-Pins, France. CEUR Workshop Proceedings 51:4.1–4.14. Amsterdam, The Netherlands.

KIFER M., LAUSEN G. et Wu J.(1995): Logical Foundations of Object-Oriented and Frame-Based Languages. Journal of the ACM, 42(4): 741-843.

LEBART L. & SALEM A. (1994). *Statistique textuelle.* Dunod.

LERAT P. (1995). Les langues spécialisées. PUF.

LINDBERG D., HUMPHREYS B. A.T. M. (1993). The Unified Medical System. Methods of

information in Medecine.

MACGREGOR R. (1991) : Inside the LOOM classifier. SIGART bulletin, 2(3):70-76.

MCGUINESS D., FIKES R., RICE J. ET WILDER S. (2000) : An Environment for Merging and Testing Large Ontologies. In Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR'2000).

MAEDCHE A., STAAB S. (2001). Ontology learning for the Semantic Web. In IEEE intelligent Systems, vol. 16, no. 2, pp. 72-79

MAEDCHE A. (2002). Ontology Learning for the Semantic Web. Boston: Kluwer Academic Publishers.

CONSORTIUM MENELAS (1994) : MENELAS : an access System for medical records using natural language. Computer Methods and Programs in Biomedicine, (45).

MEYER I. (2000). Extracting Knowledge-rich Contexts for Terminography : A Conceptual and methodological Framework. In D. Bourigault, M.-C. L'homme & C. Jacquemin (eds), Recent Advances in Computational Terminology, John Benjamins.

MILLER A. (1990). WordNet : An On-line Lexical Resource. J lexicography, vol. 3, no. 4.

MINSKY M. (1975) : A framework for representing knowledge. In P. Henry Winston, editor, The Psychology of Computer Vision. McGraw-Hill, New York, USA.

MIZOGUCHI R. ET IKEDA M. (1996). Towards Ontological Engineering (AI-TR-96-1,). Osaka: ISIR, Osaka University.

Mizoguchi R. (1998). A Step Towards Ontological Engineering. Paper presented at the 12th National Conference on AI of JSAI. June, 1998

MIZOGUCHI R. ET BOURDEAU J. (2000). Using Ontological Engineering to Overcome Common AI-ED Problems. International Journal of Artificial Intelligence and Education, vol.11 (Special Issue on AIED 2010), 107-121.

MORIN E. (1999). Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique. TAL (Traitement Automatique des Langues), vol.40, n°1, Paris : Université Paris VII, pp.143-166.

NAULLEAU E. (1997). *Apprentissage et filtrage syntaxico-semantique de syntagmes nominaux pertinents pour la recherche documentaire.* Thèse de doctorat, Université de Paris XIII, France.

NAULLEAU E. (1999). Profile-Guided Terminology Extraction. In *Proceedings of Terminology and Knowledge Extraction, TKE'99,* Innsbruck, Autriche.

NEECHES R., FININ T., GRUBER T. R., SENATOR T., AND SWARTOUT W. R. (1991). Enabling technology for knowledge sharing. AI Magazine, 12, 35-56.

NOBÉCOURT J (2000). A method to build formal ontologies from text. In: EKAW-2000 Workshop on ontologies and text, Juan-Les-Pins, France.

NOY N.F. ET HAFNER C.D.(1997). The State of the Art in Ontology Design: A Survey and Comparative Review. AI Magazine, 18 (3) :53-74.

NOY N.F., FERGERSON R.W., AND MUSEN M.A. (2000). The knowledge model of Protégé2000 : Combining interoperability and flexibility. In R. Dieng and O. Corby, editors, 12th International Conférence on Knowledge Engineering and Knowledge Managment (EKAW'OO), volume 1937)

of Lecture Notes in Artificial Intelligence, pages 17-32, Juan-les-Pins, France, 2-6 Octobre 2000. Springer Verlag.

NOY N. ET MUSEN M. A. (2000) :.ROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. In Proceedings of the 17th International Conference on Artificial Intelligence (AAAI'OO), pages 450-455. AAAI Press.

OIL.(2002): Ontology Interchange Language. http ://www.cs.colorado.edu/ eliuser /elionline4.3 /oil_5.htm, 2002

OUESLATI R. (1999). Aide à l'acquisition de connaissances à partir de corpus, thèse de doctorat.

OUESLATI R., FRATH P. (1996). Extracting concepts and relations from corpora. Procedings of the 12th European Conference on Artificial Intelligence (ECAI'96), workshop on Corpus-Oriented Semantic Analysis, Budapes.

ONTOWEB CONSORTIUM (2002) : Deliverable 1.3: A Survey on Ontology Tools. Technical report IST-2000-29243, IST.

OWL. (2002) : OWL Home Page, http ://www.w3c.org/TR/webont-req.

PICHON R. & SEBILLOT P. (1997). *Acquisition automatique d'informations lexicales a partir de corpus : un bilan.* Rapport de recherche n°3321, INRIA, Rennes, France.

PICHON R. & SEBILLOT P. (1999). Differencier les sens des mots a 1'aide du theme et du contexte de leurs occurrences : une experience. In *Actes de la conference Traitement automatique des langues naturelles, TALN'99,* Cargese, France.

PUSTEJOVSKY J. (1995). *The Generative Lexicon.* Cambridge, Massachusetts, Etats-Unis : The MIT Press.

RASTIER F. (1995). Le terme : entre ontologie et linguistique. La banque des mots, 7, 35-65.

RASTIER F., CAVAZZA M., ET ABEILLE A. (1994). Sémantique pour l'analyse. Masson, Paris, France.

RAMSHAW L. A. & MARCUS M. P. (1995). Text Chunking using Transformation-Based Learning. In *Proceedings of the 3rd Workshop on Very Large Corpora.*

RDF (2002A): RESOURCE DESCRIPTION FRAMEWORK SPECIFICATION, HTTP://WWW.W3.ORG/TR/REC-RDF-SYNTAX/.

RDFS (2002b) : Resource Description Framework Schema Specification, http://www.w3.org/TR/rdf-schema/, 2002.

ROBINSON J. A. (1965). A Machine-Oriented Logic Based on the Resolution Principle. *Journal of Association for Computing Machinery,* 12(1), 23-41.

ROSSIGNOL M. & SEBILLOT P. (2002). Automatic Generation of Sets of Keywords for Theme Characterization and Detection. In A. MORIN &; P. SEBILLOT, Eds., *Actes des 6es Journees internationales d'analyse statistique des donnees textuelles, JADT'02,* Saint-Malo, France.

ROUSSELOT F., FRATH P. & OUESLATI R. (1996). Extracting Concepts and Relations from Corpora. In *Proceedings of the Corpus-Oriented Semantic Analysis ECAI'96 Workshop,* Budapest, Hongrie.

ROUX C., PROUX D., RECHERMANN F., AND JULLIARD L. (2000). An ontology enrichment method for a pragmatic information extraction system gathering data on genetic interactions. Position paper in Proceedings of the ECAI2000 Workshop on Ontology Learning (OL2000), Berlin, Germany.

August 2000.

ROYAUTE J., SCHMITT L. & OLIVETAN E. (1992). Les expériences d'indexation a 1'INIST. In *Proceedings of the 15th International Conference on Computational Linguistics, COLING'92,* Nantes, France.

SEGUELA, P. (1999) : Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés. In Actes de TIA'99 (Terminologie et Intelligence Artificielle), Terminologies Nouvelles 19, pp. 52-60.

SEGUELA P. & AUSSENAC-GILLES N. (1999). Extraction de relations semantiques et enrichissement de modeles du domaine. In *Actes de la conference sur I'lngenierie des Connaissances, 1C'99,* Palaiseau, France.

SMADJA F. (1993a) : Retrieving Collocations from Text: Xtract. Computational Linguistics, 19(1), 143-178.

SMADJA F. (1993b). Xtract: an Overview. *Computer and the Humanities,* 26, 399-413.

SOWA J. (1995a) : Distinction, combination, and constraints. Proc. IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing.

SOWA J. (1995b) : Top-level ontological categories. International Journal of Human and Computer Studies, 43, 669-685.

SOWA J.F. (2004). KR Ontology. http://www.jfsowa.com/ontology/.

SOWA J.F (2000) Ontology, Metadata and Semiotics. In Proceedings of the 8th International Conference on Conceptual Structures (ICCS'2000), volume 1867, pages 55-81. Springer-Verlag LNCS.

STUDER R., BENJAMINS R. ET FENSEL D. (1998). Knowledge Engineering: Principles and Methods. Data Knowledge Engineering.

STUMME G. et MAEDCHE A. (2001) : Ontology Merging for Federated Ontologies on the Semantic Web. In Proceedings of the International Workshop for Foundations of Models for Information Integration (FMII'2001).

SKUCE D. R. & MEYER I. (1991). Terminology and Knowledge Acquisition: Exploring a Symbiotic Relationship. In *Proceedings of the 6th Knowledge Acquisition for Knowledge Based Systems Workshop,* Banff, Canada.

SURE Y., ERDMANN M., ANGELE J., STAAB S., STUDER R., AND WENKE D. (2002). OntoEdit : Collaborative Ontology Engineering for the Semantic Web. In I. Horrocks and J. Hendler, editors, First International Semantic Web Conférence (ISWC'02), volume (2342) of Lecture Notes in Computer Science, pages 221-235, Chia, Sardaigne, Italie, 9-12 Juin 2002. Springer Verlag.

SWARTOUT B., PATIL R., KNIGHT K. ET RUSS T. (1997). Towards Distributed Use of Large-Scale Ontologies. Spring Symposium Series on Ontological Engineering, pp.138-148.

TRONCY R. ET ISAAC A. (2002) : DOE : une mise en oeuvre d'une méthode de structuration différentielle pour les ontologies. In 13th Journées Francophones d'Ingénierie des Connaissances (IC'02), pages 63-74, Rouen, France, 28-30.

USCHOLD M. ET GRÜNINGER M. (1996b). Ontologies: Principles, Methods and Applications. J. of Knowledge Engineering Review, 11 (2).

VAN HEIJST G., SCHREIBER A. ET WIELINGA B. J. (1997). Using Explicit Ontologies in KBS Development. International Journal of Human and Computer Studies /Knowledge Acquisition, 46

(2/3), 183-292.

VANWELKENHUYSEN J. ET MIZOGUCHI R. (1994). Maintaining the workplace context in a knowledge level analysis,. Paper presented at the Proc. of JKAW'94, Hatoyama, Japan.

VANWELKENHUYSEN J. ET MIZOGUCHI R. (1995). Workplace-Adapted Behaviors: Lessons Learned for Knowledge Reuse. Paper presented at the KB&KS '95.

TALEB N., SELLAMI M. (2003). Approche hybride LSTAT d'acquisition des termes formants une ontologie du domaine. In 6th international symposium ISPS' 2003, Algiers, Algeria.

VELARDI P., NAVIGLI R., AND MISSIKOFF M. (2002). Integrated approach for Web ontology learning and engineering. IEEE Computer - November 2002.

VELARDI P., NAVIGLI R., AND MISSIKOFF M. (2003). Ontology Learning and its application to Automated Terminology Translation. IEEE Computer - November 2003.

VOUTILAINEN A. (1993). *NPTOOL,* a Detector of English Noun Phrases. *In Proceedings of the Workshop on Very Large Corpora,* Ohio State University, Ohio, Etats-Unis.

WIELINGA B. ET SCHREIBER A. (1993). Reusable and sharable knowledge bases: A European perspective. Paper presented at the KB & KS'93, Tokyo.

WINSTON, E.M., CHAFFIN R., HERRMANN D.J. (1987): Taxonomy of Part-Whole Relations. COGNITIVE SCIENCES 11, 417-444.

WOLFF C. (1729). Philosophia Prima sive Ontologia. Unpublished manuscript.

WUSTER E. (1981) : « L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses », Guy Rondeau et Helmut Felber (réd.), Textes choisis de terminologie. Fondements théoriques de la terminologie, Québec, GIRSTERM, p. 55-114

XML (2002): eXtended Markup Language Specification, http ://www.w3.org/TR/REC- xml.