وزارة التعليم العالي والبحث العلمي

BADJI MOKHTAR-ANNABA UNIVERSITY
UNIVERSITY BADJI MOKHTAR-ANNABA

جـامعة باجي مختار-عنـابة

Faculté : ***Sciences d'ingénieur*** **Année :** ***2003***

Département : ***Informatique***

MEMOIRE

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

Construction d'ontologie à partir de textes techniques – application à l'expansion de requête utilisateur

Option

Intelligence Artificielle

Par

Ladjailia Ammar

DIRECTEUR DE MEMOIRE : *LASKRI Mohamed Tayeb* *Dr* *Annaba*

DEVANT LE JURY

PRESIDENT :	*BOUFAIDA Mahmoud*	*Pr*	*Constantine*
EXAMNATEURS :	*BENSEBAA Tahar*	*Dr*	*Annaba*
	MEROUANI Hayette	*Dr*	*Annaba*

Remerciements

Au terme de ce travail, je tiens à remercier mon encadreur Mr MT. Laskri pour m'avoir proposé ce sujet, pour toutes ces remarques constructives, ses encouragements et son soutien scientifique et moral.

Je remercie Mr M. Boufaida qui m'a fait honneur en président mon jury, ma gratitude va à Mr T. Bensebaa ainsi qu'à Mme H. Merouani qui ont bien voulu examiner mon travail.

Je tiens à remercier tout particulièrement Mme Nathalie Aussenac-Gilles professeur à l'université de Toulouse pour la confiance et le soutien amical qu'ils m'ont constamment accordés. Ses conseils et ses qualités scientifiques ont été très précieux pour mener à bien cette mémoire.

Enfin, je tiens à témoigner de toute ma reconnaissance à mes amis et collègues pour leurs encouragements et leurs orientations tout au long de ma mémoire de magister.

Résumé

Notre mémoire traite de la problématique de la construction d'ontologies de domaine ou régionale, à partir de textes techniques. Nos proposons un modèle d'aide le cogniticien et les expert de domaine à la construction d'ontologie à partir de textes. Dans la première partie nous nous intéressons aux différents usages du terme «ontologie», leur classification, les méthodes de construction et leur spécification. Puis nous abordons la problématique de la construction d'ontologie à partir des textes techniques. A la repense de cette problématique, nous proposons, dans la première partie une méthode linguistique quantitative d'acquisition des termes à partir des textes, puis nous structurons les syntagmes sous forme d'un réseau, dans lequel chaque syntagme est relié à sa tête et à ses expansions. A partir de ce réseau, le module d'analyse distributionnelle construit pour chaque terme du réseau l'ensemble de ses contextes terminologique. Un autre module (CAH) rapproche ensuite les termes, pour construire les classes des termes, sur la base de mesures de proximité distributionnelle. L'ensemble de ces résultats est utilisé comme aide à la construction d'ontologie à partir de corpus spécialisés. Notre ontologie est destinée à l'expansion de la requête par suggestion à l'utilisateur de termes de plus ou moins liés à ceux de sa requête.

Mot-clé : Ontologie, Systèmes à Base de Connaissances, Web sémantique, ingénierie des connaissances, ingénierie ontologique, construction d'ontologie, analyse distributionnelle, terme, syntagme, recherche d'information, expansion de requête, classification automatique.

Abstract

Our milked of the problems of construction of ontologies of field or regional memory, starting from technical texts. Our let us propose to a model of assistance the knowledge engineer and the expert of field to the construction of ontology starting from texts. In the first part we are interested in the various uses of the term «ontology", their classification, the methods of construction and their specification. Then we approach the problems of the construction of ontology starting from the technical texts. With reconsider these problems, we propose, in the first part a quantitative linguistic method of acquisition of the terms starting from the texts, then we structure the syntagms in the form of a network, in which each syntagm is connected at its head and its expansions. From this network, the module of distributionnelle analysis builds for each term of the network the whole of its contexts terminological. Another module (CAH) brings closer then the terms, to build the classes of the terms, on the basis of measurement of distributional proximity. The whole of these results is used like contributes to the construction of ontology starting from specialized corpora. Our ontology is intended for the expansion of the request by suggestion to the user of terms of more or less related to those of its request.

Key word: Ontology, Systems containing Knowledge, semantic Web, engineering of knowledge, ontological engineering, construction of ontology, analyze distributional, term, syntagm, search for information, expansion of request, automatic classification.

Ce rapport présente, de façon non exhaustive, un état de l'art en matière d'ingénierie ontologique, en particulier les acquis du domaine et les nombreux problèmes restant à traiter. La place des ontologies au sein du processus de représentation des connaissances, les besoins auxquels répondent les ontologies et les éléments que l'on est amené à y intégrer sont les sujets de la première partie. La deuxième partie expose le processus de construction des ontologies et en détaille les différentes étapes. Finalement, quelques applications des ontologies sont évoquées, le projet de Web Sémantique étant plus particulièrement détaillé.

Mots-clés : Ontologie, Systèmes à Base de Connaissances, Web sémantique

---

Notre mémoire traite de la problématique de la construction d'ontologies de domaine ou régionale, à partir de textes techniques. Nos proposons un modèle d'aide le cogniticien et les expert de domaine à la construction d'ontologie à partir de textes. Dans la première partie nous nous intéressons aux différents usages du terme «ontologie», leur classification, les méthodes de construction et leur spécification. Puis nous abordons la problématique de la construction d'ontologie à partir des textes techniques. A la repense de cette problématique, nous proposons, dans la première partie une méthode linguistique quantitative d'acquisition des termes à partir des textes, puis nous structurons les syntagmes sous forme d'un réseau, dans lequel chaque syntagme est relié à sa tête et à ses expansions. A partir de ce réseau, le module d'analyse distributionnelle construit pour chaque terme du réseau l'ensemble de ses contextes terminologique. Un autre module (CAH) rapproche ensuite les termes, pour construire les classes des termes, sur la base de mesures de proximité distributionnelle. L'ensemble de ces résultats est utilisé comme aide à la construction d'ontologie à partir de corpus spécialisés. Notre ontologie est destinée à l'expansion de la requête par suggestion à l'utilisateur de termes de plus ou moins liés à ceux de sa requête.

La méthode proposée s'articule sur l'analyse distributionnelle proposée par Harris. Dans la première partie, nous proposons une méthode de l'acquisition de terme, puis nous construisons un réseau de mots et syntagmes, dans lequel chaque syntagme est relié à sa tête et à ses expansions. A partir de ce réseau, le module d'analyse distributionnelle construit pour chaque terme du réseau

l'ensemble de ses contextes terminologique. Un autre module rapproche ensuite les termes, pour construire les classes des termes, sur la base de mesures de proximité distributionnelle. L'ensemble de ces résultats est utilisé comme aide à la construction d'ontologie à partir de corpus spécialisés. Notre ontologie est destinée à l'expansion de la requête par suggestion à l'utilisateur de termes de plus ou moins liés ceux de sa requête.

Abstract

Our thesis treats the problems of the field ontology construction, starting from technical texts. We propose a model of assistance to the knowledge engineers and the field ontology construction experts starting from texts. In the first part of our thesis we are interested in the various uses of the term "ontology ", their classification, the methods of construction and their specification. In the second part we approach the problems of the construction of ontology starting from the technical texts. Lastly, like a response to these problems, we propose a method of assistance to the knowledge engineers and the experts of field for the construction of ontology, in the first part a quantitative linguistic method of acquisition of the terms starting from the texts is defined, then we structure the syntagms in the form of a network, in which each syntagm is connected at its head and its expansion. From this network, the module of distributional analysis builds for each term of the network the whole of its contexts terminological. Module (CAH) brings closer then the terms, to build the classes of the terms, on the basis of measurement of distributional proximity. Our ontology is intended for the expansion of the request by the suggestion with the user of the terms of more or less related to those of its request.

Table des matières

Chapitre 2 : Construction d'ontologie à partir des textes techniques

Chapitre 3 : Notre méthode de la construction d'ontologie

INTRODUCTION

L'objet du mémoire est de présenter une méthode de construction d'ontologie de domaine à partir des textes techniques, cette ontologie devant répondre à des besoins de l'expansion de la requête d'utilisateur.

Une ontologie est une structure de donnée opérationnelle qui rend compte des concepts d'un domaine et de leurs relations. Leur développement croissant en Intelligence Artificielle vient de leur intérêt pour associer du sens à des ressources textuelles, pour localiser et gérer des connaissances dans diverses applications.

Les ontologies sont à l'heure actuelle le coeur des travaux menés en Ingénierie des Connaissances (IC). 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. Les ontologies participent donc pleinement aux dimensions scientifique et technique de l'Intelligence Artificielle (IA) : scientifique comme étude des connaissances humaines et plus largement de l'esprit humain, ce qui rattache l'IA aux sciences humaines, et technique comme création d'artefacts possédant certaines propriétés et capacités en vue d'un certain usage.

Au fur et à mesure des expérimentations, des méthodologies de construction d'ontologies et des outils de développement adéquats sont apparues. Les ontologies apparaissent ainsi comme des composants logiciels s'insérant dans les systèmes d'information et leur apportant une dimension sémantique qui leur faisait défaut jusqu'ici.

Le champ d'application des ontologies ne cesse de s'élargir et couvre les systèmes conseillers (systèmes d'aide à la décision, systèmes d'enseignement assisté par ordinateur), les systèmes de résolution de problèmes ou les systèmes de gestion de connaissances. Un des plus grands projets basés sur l'utilisation d'ontologies consiste à ajouter au Web une véritable couche de connaissances permettant, dans un premier temps, des recherches d'informations au niveau sémantique et non plus simplement syntaxique. A terme, il est prévu que des applications Internet pourront mener des raisonnements utilisant les connaissances stockées sur la Toile [**Fürst 2002**].

L'enjeu de l'effort engagé est de rendre les machines suffisamment sophistiquées pour qu'elles puissent intégrer le sens des informations, qu'à l'heure actuelle, elles ne font que manipuler formellement. Mais en attendant que des ordinateurs « bourrés » d'ontologies et de

connaissances nous déchargent en partie du travail de plus en plus lourd de gestion des informations dont le flot a tendance à nous submerger, de nombreux problèmes théoriques et pratiques restent à résoudre.

Notre problématique est de proposer une méthode d'aide au cogniticien de construction d'ontologie à partir des textes techniques, pour cette raison, nous divisons notre mémoire en trois chapitres.

Chapitre 1 : 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 discipline. Il représente des définitions et un état de l'art sur les utilisations, la construction et les langages de représentation des ontologies. Finalement, quelques applications des ontologies sont exposées, le projet de Web sémantique étant plus particulièrement détaillé.

Chapitre 2 : Le but de ce chapitre est de présenter les éléments théoriques de la construction d'ontologie à partir des textes techniques, nous abordons principalement les travaux de BACHIMONT qui concernent les méthodologies suivies pour un engagement ontologique, puis nous exposons quelques principes de la construction d'ontologie. Finalement, nous présentons une méthodologie linguistiquement fondue s'appelle : TERMINAE.

Chapitre 3 : Ce chapitre représente la réponse à notre problématique, après l'exposé de quelques principes théorique que nous l'utilisons dans notre modèle, nous présentons notre méthode de la construction d'ontologie. Finalement nous appliquons notre ontologie dans le domaine de la recherche d'information, et avec précision nous utilisons celle-ci pour l'expansion de la requête d'utilisateur.

Chapitre 1

Définition, construction et utilisation des ontologies

1. INTORDUCTION

Notre but est de développer une méthodologie pour assister un cogniticien dans sa tâche de construction d'ontologie du domaine, à partir d'une documentation technique. Dans le présent chapitre, nous précisons les différents points de vue sur la notion de « l'ontologie » et plus particulièrement la notion d'ontologie en IC. Puis, nous citons les différentes classifications, axes de recherches concernés par la notion d'ontologie, les méthodes de la construction et les langages de spécifications.

2.DONNEE, INFORMATION ET CONNAISSANCE

Si une ***information*** est une ***donnée*** plus une ***signification***, alors une ***connaissance*** est une ***information*** plus du ***raisonnement.*** Une forme courante de connaissances, e.g. dans un programme Prolog, est une collection de faits et de règles à propos d'un sujet. Par exemple, une base de connaissances à propos d'une famille peut contenir le fait que Ahmed est le fils de Ali, le fait que Rafik est le fils de Ahmed, et la règle représentant que le fils du fils d'une personne est son petit-fils. A partir de cette connaissance, le programme peut inférer le fait que Rafik est un petit-fils de Ali.
La conceptualisation représente la collection des ***objets***, de ***concepts*** et des autres ***entités*** qui sont supposés exister dans un certain domaine d'intérêt, et les ***relations*** qui les relient. Une conceptualisation est une ***vue abstraite***, ***simplifiée*** du monde que l'on veut représenter. Par exemple, on peut conceptualiser une famille par un ensemble de noms, de sexes et de relations entre les membres de la famille. Le choix d'une conceptualisation est la *première étape de la* ***représentation de connaissances***. Chaque ***base de connaissances***, ***système à base de connaissances***, ou ***agent modélisé au niveau connaissance*** est, *explicitement* ou *implicitement*, relative à une certaine conceptualisation. Une représentation formelle de cette conceptualisation s'appelle une ***ontologie*** : "*an ontology is an explicit specification of a conceptualisation*" *(Gruber, 1993).*

3. QU'EST-CE QU'UNE ONTOLOGIE ?

Une ontologie est une structure de donnée opérationnelle qui rend compte des concepts d'un domaine et de leurs relations. Leur développement croissant en Intelligence Artificielle vient de leur intérêt pour associer du sens à des ressources textuelles, pour localiser et gérer des connaissances dans diverses applications. **[Aussenac-Gilles2002]**. L'étude des ontologies intéresse non seulement l'ingénierie des connaissances mais aussi la philosophie et la linguistique.

3.1 Le point de vue de l'ingénierie des connaissances.

L'Ingénierie des Connaissances (IC) est une branche de l'IA issue de l'étude des Systèmes Experts (SE) et en générale les Systèmes à Base de Connaissances (SBC). Les ontologies sont apparues en Ingénierie des connaissances et plus largement en

Intelligence artificielle, avec l'idée de construire mieux et plus rapidement des SBC en réutilisant le plus possible des composants génériques, que ce soit au niveau du raisonnement ou des connaissances du domaine. Ce qui nous a amené incidemment à une première définition simple selon **[Charlet 2003]**:

> ***Ontologie (déf. 1)*** *: Ensemble des objets reconnus comme existant dans le domaine. Construire une ontologie c'est aussi décider de la manière d'être et d'exister des objets.*

Il faut bien rappeler que les travaux sur les ontologies sont développés dans un contexte informatique où le but final est de spécifier un artefact informatique. Ce contexte est important pour comprendre les buts poursuivis par les concepteurs d'ontologie. En particulier, la question de la conceptualisation devient centrale dans le but de construire un artefact puisqu'on a besoin, dans ce contexte, de définir et spécifier les concepts à prendre en compte. On peut alors proposer la définition de Gruber (1993) selon **[Charlet 2003]**:

> ***Ontologie (déf. 2)*** *: Une ontologie est une spécification explicite d'une conceptualisation.*

La *spécification* en question est opérée au niveau des connaissances et non pas au niveau de symboles. Cette dernière inclut des définitions et une indication de la façon dont les concepts sont reliés entre eux, les liens imposant collectivement une structure sur le domaine et contraignant les interprétations possibles des termes. Une *conceptualisation* est une description d'une partie du monde en termes de classes de concepts et de relations entre ceux-ci [**Assadi 1998**] et [**Aussenac-Gilles 2002**].

Dans un sens large, on peut adopter pour la notion d'ontologie la caractérisation suivante [**Uschold, 98**] :

> ***Ontologie (déf. 3)*** *: «Une ontologie peut prendre différentes formes, mais elle inclura nécessairement un vocabulaire de termes et une spécification de leur signification. Cette dernière inclut des définitions et une indication de la façon dont les concepts sont reliés entre eux, les liens imposant collectivement une structure sur le domaine et contraignant les interprétations possibles des termes.»*

Cette caractérisation convient pour des objets divers tels des glossaires, des terminologies, des thesaurus et des ontologies (au sens strict), mis en œuvre par différents professionnels (ingénieurs de la connaissance, terminologues, bibliothécaires, traducteurs) et se distinguant suivant que l'accent est mis sur les termes ou leur signification**.**

Selon **[Assadi 1998]**, Guarino (1997) propose une définition qui ajoute une dimension intéressante à l'ontologie : celle du *consensus* (en anglais *Agreement*). En effet, il propose que l'ontologie soit vue, non pas comme la *spécification* d'une conceptualisation, mais comme un consensus (partiel) sur cette conceptualisation. Les ontologies sont partielles car une conceptualisation ne peut pas toujours être entièrement formalisée dans un cadre logique, du fait d'ambiguïtés ou du fait qu'aucune représentation de leur sémantique n'existe dans le langage de représentation d'ontologies choisi.

Selon **[Sébastien 2000]**, une ontologie concerne principalement l'intension des concepts. Cette dernière, correspond au sens de concept; elle représente un ensemble invariant de règles permettant, quelle que soit la situation, de reconnaître les objets constituant la réalisation, ou l'extension, du concept. Un concept est lui, une entité complexe composée d'un terme, d'une notion et d'un objet comme le montre la figure 1.

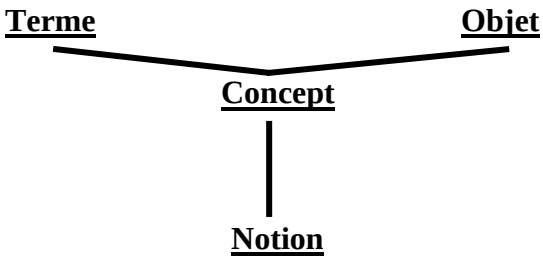


Fig 1 : Le concept

Le **terme** est un élément lexical qui permet d'exprimer le concept en langue naturelle. Il peut admettre un acronyme et des variations.

La **notion** est le sens du terme, la partie intensionnelle du concept. Elle exprime un point de vue sur un objet individuel ou un ensemble d'objets.

L'**objet** est une entité de référence pour la notion, la partie extensionnelle du concept.

Et selon **[Aussenac-Gilles2002]** et **[Sébastien 2000]**, Le terme «conceptualisation» situe les ontologies sur le versant sémantique, donc, une conceptualisation est un système différentiel d'intensions. Ce dernier consiste en système différentiel d'intensions prenant la forme d'une taxinomie de subsomption. Une taxinomie de subsomption peut être vue comme un "classement" dans lequel les éléments sont hiérarchisés en père et en fils hérite de toutes les propriétés relatives à son père, donc l'intension fils est donc plus spécifique et concerne moins d'objets que l'intension père. La plupart des ontologies sont structurées au moyen de la relation «est un» de *subsomption*, ou de *généralisation*, entre concepts. La relation Tout-Parties «est

composé de» est également utilisée. Un système différentiel d'intensions est un système dans lequel les concepts sont positionnés par rapport aux différences existant dans l'intension qu'ils expriment.

3.2 Le point de vue de l'ontologie

Selon **[Aussenac-Gilles2002] :**

> ***Ontologie (déf. 4)*** *: L'Ontologie est la branche de la philosophie qui traite de la nature et de l'organisation de la réalité.*

Elle côtoie l'Épistémologie qui traite de la nature et des origines de notre connaissance.

Selon (Le Petit Robert):

> ***Ontologie (déf. 5)*** *: la partie de la métaphysique qui s'intéresse à l'Être en tant qu'Être.*

Mais l'Ontologie est habituellement davantage comprise comme une science des étants que comme une science de l'Être en tant qu'Être, c'est-à-dire qu'elle s'intéresse davantage à ce qui existe (les étants ou existants) qu'aux principes de ce qui existe (l'Être). Cette science, l'Ontologie, produit des ensembles, les ontologies.

L'Ontologie d'Aristote*:*

Aristote s'intéresse à la nature des choses et aborde la question de l'essence et du genre spécifique.

> *Il (Aristote) définit l'essence des êtres qui n'existe à part des êtres individuels sensibles. Cette essence doit être discernée au sein des substances sensibles et dégagée au moyen d'une opération intellectuelle appelée abstraction.*
>
> *[. . .].*
>
> *Le travail d'abstraction discerne ce qu'il y a d'intelligible dans la substance et qui est dû à sa forme. Cet intelligible s'exprime au moyen de concepts, qui ont chacun une universalité (le concept de cheval vaut pour tous les chevaux). Ce discernement abstractif s'effectue au départ d'une observation de chaque substance et de la comparaison des substances entre elles. La comparaison implique un classement au moyen des concepts. Ceux-ci ont des extensions diverses et l'un peut recouvrir l'autre. Ainsi les concepts, comme les classes, ont entre eux des relations hiérarchiques de genre et d'espèce. La définition d'une essence désigne toujours une espèce qu'elle situe à l'intérieur d'un*

genre selon la formule : «espèce = genre prochain + différence » **[Charlet 2003]**

Ainsi, l'homme appartient au genre animal et se différencie spécifiquement par sa raison. L'homme est un animal raisonnable. Aristote propose ici la première méthode d'élaboration d'une ontologie, dans sons cas, l'Ontologie : les concepts se définissent par genre et différence au sein d'une relation de subsomption. Les taxinomies en sciences naturelles s'en sont inspirées.

3.3 Le point de vue de la linguistique

La Linguistique est concernée par la question des ontologies dans la mesure où les données dont on dispose pour élaborer les ontologies consistent en des expressions linguistiques de connaissances. La caractérisation du sens de ces expressions conduit à déterminer des *signifiés contextuels*, dépendants des contextes (documents) où les expressions apparaissent. Ces signifiés contextuels doivent alors être *normés*, ce qui revient à fixer une signification pour un contexte de référence, celui de la tâche (application) pour laquelle l'ontologie est élaborée **[Aussenac-Gilles2002]**. L'ontologie régionale (non universelle) que l'on obtient est ainsi une *spécification de signifiés normée*.

3.4 De point de vue du Sciences Naturelles et Taxinomie

La science a toujours eu pour premier but de repérer et classifier les objets du monde pour les comprendre, comprendre leur fonctionnement et leur genèse. La recherche s'est systématisée en sciences naturelles, d'abord en botanique et ensuite pour tout le règne animal. Les classifications ainsi construites sont des taxinomies. Elles comportent la classification elle-même et les critères d'icelle. Sa définition rend compte de cette nature.

Taxinomie (déf. 1) : *Étude théorique des bases, lois, règles, principes, d'une classification.*

Taxinomie (déf. 2) : *Classification d'éléments (Le Petit Robert*).

4. LES CARACTERISTIQUES DES ONTOLOGIES

Les ontologies à trois caractéristiques nous permettent de préciser ce qu'on peut représenter avec une ontologie :

4.1 Les propriétés. Une ontologie est non seulement le repérage et la classification des concepts mais c'est aussi des caractéristiques qui leur sont attachées et qu'on appelle ici des propriétés. Ces propriétés pouvant être values. Par exemple un patient a un âge qui a une certaine valeur ou est soigné par tel médecin.

4.2 Le type d'ontologie. Les méthodes en Ingénierie des connaissances ont répertorié plusieurs types d'ontologie liés à l'ensemble des objets conceptualisés et manipulés au sein d'un SBC. Nous allons en citer quelques-unes : on a (1) l'ontologie du domaine (cf. § 5.1), (2) l'ontologie générique ou qui se veut comme telle et qui repère et organise les concepts les plus abstraits du domaine ou autre (cf. § 5.2), (3) l'ontologie d'une méthode de résolution de problème où le rôle joué par chaque concept dans le raisonnement est rendu explicite (p. ex. signe ou syndrome dans le cadre du raisonnement médical) (cf. § 5.3), (4) l'ontologie d'application qui se veut une double spécialisation : d'une ontologie du domaine et d'une ontologie de méthode (cf. § 5.4), enfin (5) l'ontologie de représentation qui repère et organise les primitives de la théorie logique permettant de représenter l'ontologie (cf. § 5.5) (p. ex. la frame ontology d'ONTOLINGUA [**Gruber 1993**])).

4.3 Les liens reliant les concepts. La relation de subsomption is-a qui définit un lien de généralisation – i.e. hyperonymie – est la plus utilisées dans les ontologies et ce depuis Aristote. Mais ce n'est pas la seule possible et, surtout, pas la plus utile dans certains cas. On peut avoir besoin de relations de partie-tout ou méronymie. Ce type de conceptualisation est, par exemple, indispensable en anatomie médicale où il est nécessaire de décrire des organes ou des systèmes et ce qui les compose**.**

5. CLASSIFICATION DES ONTOLOGIES

5.1 L'ontologie du domaine : exprime des conceptualisations spécifiques à un domaine.

5.2 L'ontologie générique : exprime des conceptualisation valables dans différents domaines. Classiquement, de telles ontologies définissent des notions telles que état, processus, cause, composant, etc.

5.3 L'ontologie d'une méthode de résolution de problème où le rôle joué par chaque concept dans le raisonnement est rendu explicite (*p. ex. signe* ou *syndrome* dans le cadre du raisonnement médical),

5.4 L'ontologie d'application : une ontologie applicative contient les connaissances nécessaires à une application donnée. Elle est donc spécifique et non réutilisable.

5.5 L'ontologie de représentation : conceptualisation des primitives des langages de représentation de connaissances. Ainsi, les ontologies génériques ou du domaine peuvent être écrites en utilisant des primitives de telle ou telle ontologie de représentation.

6. À QUOI SERT UNE ONTOLOGIE ?

Le développement de l'ontologie croissant tente de répondre à des demandes appliquées : sus associer du sens à des ressources textuelles, localiser et gérer des connaissances dans diverses applications, faciliter la communication, mettre en commun des ressources ou mieux les communiquer **[Aussenac-Gilles2002]**. Toute activité humaine spécialisée développe son propre jargon (langue de spécialité) sous la forme d'une terminologie et d'une conceptualisation associée spécifique. L'existence de tels jargons entraîne des problèmes de compréhension et des difficultés à partager des connaissances entre les acteurs de l'entreprise, les services d'une entreprise, les entreprises d'une industrie, qui font des métiers différents. Fondamentalement, le rôle des ontologies est d'améliorer la communication entre humains, mais aussi entre humains et ordinateurs et finalement entre ordinateurs

6.1 Communication : Les ontologies peuvent intervenir dans la communication entre humains. Elles servent par exemple, à créer au sein d'un groupe ou d'une entreprise un « vocabulaire» standardisé. Pour cette raison, on utilise les ontologies informelles. L'existence de vocabulaires différents au sein d'une entreprise (ex : bureau d'études, bureau des méthodes) ou d'une industrie (ex : constructeur automobile, équipementier) constitue un frein à la collaboration et aux partenariats. Les enjeux touchent donc directement la compétitivité de l'entreprise. Pour l'entreprise, l'ontologie sert à :

- améliorer la compréhension entre les employés,

- favoriser la diffusion des informations et leur exploitation,
- Promouvoir une nouvelle approche de conception des systèmes d'information (réutilisation et interopérabilité de logiciels).

Pour ces besoins de standardisation du vocabulaire, une terminologie ou une ontologie informelle peuvent suffire.

6.2 L'aide à la spécification de systèmes : au niveau du système d'information, l'ontologie fournit une liste classifiée des objets que doit manipuler le système : c'est le référentiel du système d'information. Une volonté de *réutilisabilité*, présente déjà dans les SBC, sous-tend l'utilisation des ontologies dans le cadre des systèmes d'information dans les points suivantes :

Spécification ; Acquisition des connaissances : une ontologie peut aider à l'analyse des besoins et à définir les spécifications d'un SI.

Réutilisation ; Partage : une ontologie peut être, ou peut devenir suite à une traduction, un composant réutilisable et/ou partagé par plusieurs logiciels.

Fiabilité ; Maintenance : une ontologie peut servir à améliorer la documentation d'un logiciel et/ou à automatiser des vérifications de cohérence (SBCs), réduisant les coûts de maintenance.

Interopérabilité : en jouant le rôle d'un format d'échange, l'ontologie permet à des systèmes d'information, basés sur des paradigmes de modélisation et des langages d'implantation différents, de coopérer.

Recherche : une ontologie peut jouer le rôle de méta data servant d'index dans un répertoire d'information.

Intégration : dans une application «entrepôt de données», une ontologie peut jouer le rôle d'un schéma conceptuel commun reliant entre elles plusieurs sources d'information hétérogènes.

6.3 L'interopérabilité : L'interopérabilité est une spécialisation de la communication, dans ce cas vue entre deux ordinateurs. L'ontologie répertorie alors les concepts que des applications peuvent s'échanger même si elles sont distantes et développées sur des bases différentes. Cette interopérabilité est l'interopérabilité sémantique.

6.4 Interface Homme-Machine : la visualisation de l'ontologie permet à l'utilisateur de comprendre le vocabulaire utilisé par le SI et de mieux formuler ses requêtes.

6.5 L'indexation et la recherche d'information : Plus récemment, les travaux autour du Web sémantique (§ 6.7) ont réactivé la problématique et l'utilisation des ontologies : en plus d'un rôle de médiateur, les ontologies y sont utilisées pour l'indexation, fournissant les index conceptuels décrivant les ressources sur le Web. Ce type d'usage, ressortissant comme certains points précédents à la communication entre être humain et machines, pose la question de l'accès et la compréhension de l'ontologie. une ontologie linguistique peut permettre de comprendre les requêtes (représentation du contenu) de l'utilisateur formulé en langue naturelle.

6.6 Les ontologies dans les systèmes à base de connaissances : La première et originelle utilité d'une ontologie était liée à une volonté de réutilisation. Plus précisément, on peut dire qu'elle sert de squelette à la représentation des connaissances du domaine dans la mesure où elle décrit les objets, leurs propriétés et la façon dont elles peuvent se combiner pour constituer des connaissances du domaine complètes. La principale application des ontologies reste la gestion de données au niveau connaissance. De nombreux projets plus ou moins opérationnels existent dans différents domaines. On peut par exemple citer le projet MENELAS, et qui vise la gestion des rapports médicaux et leur analyse par un système utilisant le modèle des graphes conceptuels. Les graphes, qui représentent les connaissances médicales incluses dans les rapports, sont générés à partir des textes et stockés dans une structure *ad-hoc*. L'utilisation de mécanismes de raisonnement adaptés permet alors la consultation interactive de la connaissance, le système disposant des moyens d'aiguiller la recherche de l'utilisateur par des questions et/ou des propositions. D'autres projets, tournés vers

la gestion des *mémoires d'entreprise*, sont actuellement en cours. Le projet TOVE (TORONTO VIRTUAL ENTERPRISE) a pour but de créer *un modèle d'entreprise* exprimé dans une ontologie, permettant à un système utilisant cette ontologie de gérer les connaissances liées à l'organisation et aux activités des entreprises. Le projet COMMA [**Fabien, 2002**], vise également à permettre la gestion d'une mémoire partagée des connaissances à l'intérieur d'une entreprise. Les scénarios auxquels le système doit pouvoir s'appliquer sont l'apport d'information à un nouvel employé et le support au processus de veille technologique. L'utilisation d'ontologies au sein de systèmes offrant de réelles possibilités de raisonnement est encore peu développé, du fait que les *langages de représentation* sont encore peu outillés à ce niveau. Certains projets ont cependant été lancés, comme le projet GINA (Géométrie Interactive et NAturelle). Le but de ce projet est de développer un système de conception assistée par ordinateur qui soit interactif et dialogue avec l'utilisateur au niveau connaissance. Ce dialogue peut servir à l'analyse de la scène en cours de conception, le système pouvant répondre à des questions du type « y a t-il des droites parallèles à telle droite ? ». Le système doit également pouvoir détecter les erreurs de conception commises par l'utilisateur et lui suggérer des modifications. Le projet GINA nécessite donc la construction d'une ontologie de la géométrie, incluant les connaissances de raisonnement, c'est-à-dire les axiomes de la géométrie. L'ontologie de l'axiomatique de la géométrie projective a déjà été représentée à l'aide du modèle des graphes conceptuels et validée par son utilisation dans un système de preuve automatique de théorèmes [**Fürst, 2002**].

6.7 Le Web sémantique : Le Web constitue un terrain idéal d'application des ontologies considérées en tant que spécifications partagées de connaissances, les pages Web représentant une masse de connaissances aussi énormes que disparate. Cette masse augmente sans cesse ainsi que le nombre d'utilisateurs qui veulent pouvoir trouver facilement les informations qu'ils y recherchent. L'éventail des thèmes traités dans les différentes pages Web est tel qu'une recherche syntaxique par mot-clés retourne quasi systématiquement des pages qui ne portent pas toutes sur le même domaine de connaissances. L'exploitation efficace des ressources du Web suppose donc que les moteurs de recherche puissent accéder à la thématique de chaque page, et à son sens. De plus, la variété des sources d'information sur le

Web (textes, images, etc) plaide pour un traitement de l'information qui soit indépendant des formes sous lesquelles elle est stockée, c'est-à-dire pour un traitement au niveau conceptuel. Une partie des manipulations d'informations actuellement assurées par les utilisateurs pourra ainsi être prise en charge par les machines. « *The vision of the semantic Web is to provide computer interpretable markup of the Web's content and capability, thus enabling automation of many tasks currently performed by human beings* » [BERNERS-LEE 2001]. L'ajout d'une couche sémantique au dessus de la couche HTML, qui ne peut servir qu'à décrire formellement les pages Web, est donc nécessaire. Chaque page doit ainsi intégrer une représentation des connaissances qu'elle contient. De plus, les liens sémantiques entre chaque page doivent être spécifiés, ce que ne permet pas de faire l'hypertexte classique. Les différentes applications Internet (moteurs de recherche, services, etc.) pourront alors accéder à la sémantique des connaissances intégrées aux pages Web, du moins à une représentation de ces connaissances. Dans ce cadre, les ontologies vont servir à l'interprétation de ces connaissances en spécifiant la sémantique de la représentation utilisée. Bien évidemment, les ontologies développées pour le Web doivent utiliser le même formalisme, ou au moins être formellement compatibles. De plus, les sémantiques qu'elles expriment doivent aussi être compatibles. Si la compatibilité des formalismes peut être assurée par la définition d'un standard de représentation, celle des sémantiques peut difficilement être garantie a priori. Poser des règles sur la construction des ontologies et imposer des *méta-ontologies* standard pourrait constituer un moyen de limiter les incohérences. Dans le même ordre d'idée, les différentes ontologies développées doivent avoir des objectifs opérationnels identiques. Dans le cadre du Web, ces objectifs sont heureusement assez limités. Ils concernent la recherche d'information, la recherche de services Web adaptés aux requêtes des utilisateurs ou la gestion de ces services. Le W3C a adopté le langage RDF (Ressource Description Framework) comme formalisme standard de représentation. Utilisant la syntaxe XML (Extended Markup Language) 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 Ressource 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 spécifier que la page d'adresse « www.adresse-de-la-page.com » a été écrite par « auteur-de-la-page » qui a pour mail « mail-de-l'auteur », on écrira :

```
< ?xml version="1.0" ?>
< !– Référence à la page décrivant la syntaxe rdf –>
xmlns :rdf="http ://www.w3c.org/1999/02/22-rdf-syntax-ns#"
< !– Déclaration de deux espaces de noms –>
xmlns :s=http ://description.org/schema
xmlns :v=http ://description.org/identity
<rdf :RDF
<rdf :Description about="http ://www.adresse-de-la-page.com">
< !– La propriété CreatedBy appartient à l'espace de noms "s" –>
<s :CreatedBy>auteur-de-la-page</s :CreatedBy>
</rdf :Description>
<rdf :Description about="auteur-de-la-page.com">
< !– La propriété Email appartient à l'espace de noms "v" –>
<v :Email>mail-de-l'auteur</v :Email>
</rdf :Description>
</rdf :RDF>
```

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 SCHEMA (RDF(S)). La figure 2 montre les primitives incluses dans le RDF(S). Les concepts et relations sont déclarés dans un document RDF(S) comme instances de « Class » et de « Property ».

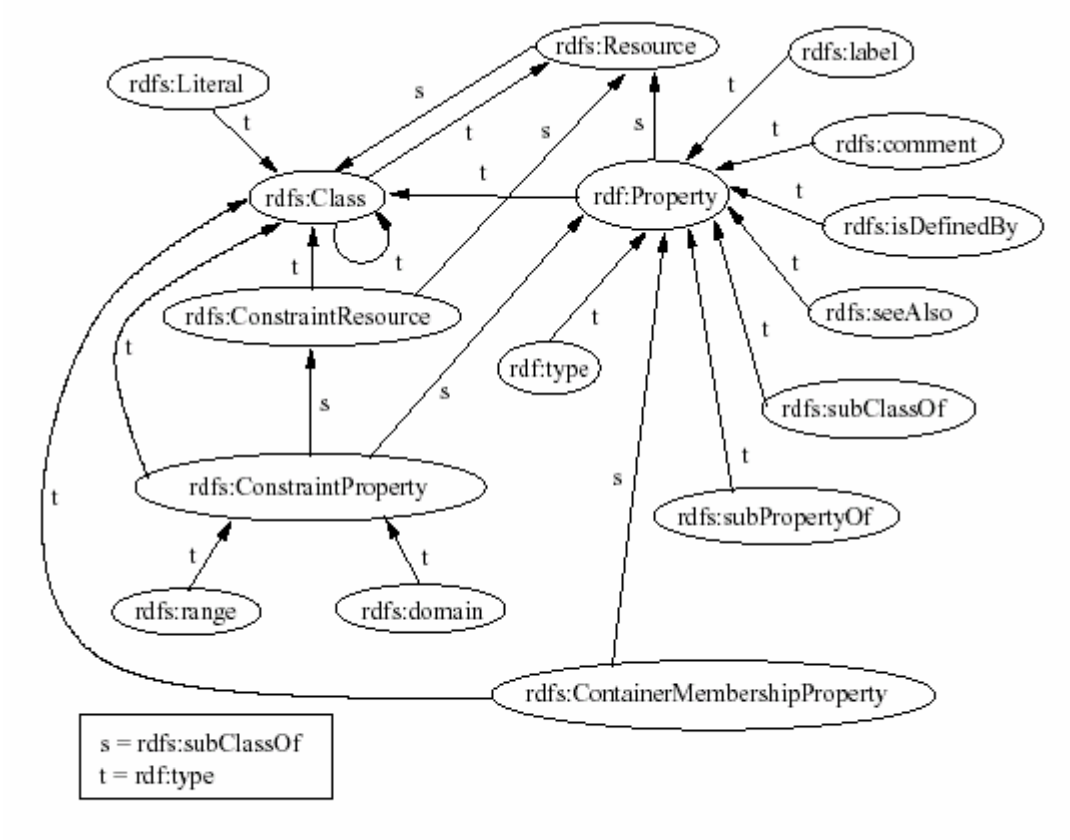


Fig 2 : Le schéma de RDF(s)

Le RDF(S) permet ainsi d'exprimer formellement des connaissances mais également de décrire des ontologies. Par exemple, on pourrait décrire une ontologie spécifiant les notions de pageWeb et d'auteur de la façon suivante :

< !– Définition de la classe document –>

<rdfs :Class rdf :ID='document'/>

< !– Définition de la classe page Web comme une sorte de document–>

<rdfs :Class rdf :ID='page-Web'>

<rdfs :subClassOf rdf :resource='#document'/>

</rdfs :Class>

< !– Définition de la classe auteur –>

<rdfs :Class rdf :ID='auteur'/>

< !– Définition de la classe mail –>

<rdfs :Class rdf :ID='mail'/>

< !– Définition de la relation est-auteur liant un auteur et une page Web –>

<rdfs :Property rdf :ID='est-auteur'>

<rdfs :domain rdf :resource='#auteur'/>

<rdfs :range rdf :resource='#page-Web'>

<rdfs :Property/>

Une fois ce schéma stocké sur le Web, les primitives qui y sont décrites peuvent être utilisées dans une page si on y inclut une référence à l'URI du schéma. Une application nécessitant l'accès à la sémantique de la page utilisera alors le schéma d'interprétation. Le RDF(S) n'est cependant pas un langage opérationnel de représentation, au sens où il est impossible de déclarer des connaissances de raisonnement. Des extensions ont été proposées afin de pallier cette lacune. L'ajout de contextes, permettant de disjoindre les descriptions et d'éviter de perdre de l'information a été proposé par A. DELTEIL Plusieurs propositions ont été faites tendant à représenter les axiomes (règles) et définitions de classes et de propriétés en introduisant d'autres classes et/ou propriétés. D.MCDERMOTT comment représenter une implication à l'aide d'une classe « if » et de deux propriétés « antecedent » et « consequent ». Aucun mécanisme inférentiel n'existant actuellement sur le Web, A.DELTEIL propose également de traduire au besoin les descriptions RDF(S) en graphes conceptuels puis d'utiliser les mécanismes inférentiels de ce modèle. 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) 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 contraignant la cardinalité ou en en restreignant la portée. Le langage OIL a été fusionné avec le langage DAML pour former le DAML+OIL. DAML (Darpa Agent Markup Language) 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. L'intégration de OIL rend possible les inférences compatibles avec les logiques de description, essentiellement les calculs des liens de subsomption. 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 cet « Ontology Web Language » (OWL) ont déjà été adoptées par le W3C. L'architecture en couches d'un tel langage est résumée dans le dessin de la figure 3, 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.

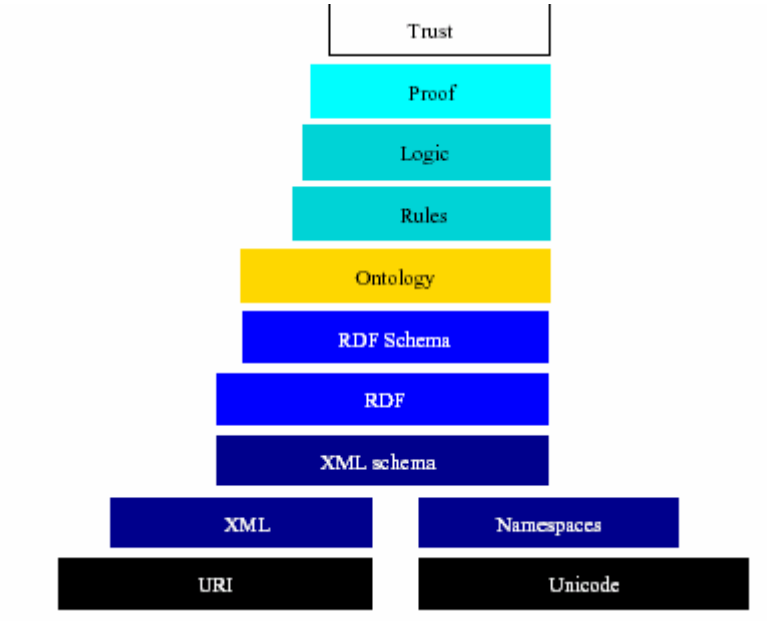


Fig 3 : les couches des Web sémantique

Parmi les applications utilisant le *Web sémantique*, le projet COMMONCV est un bon exemple de ce que peut apporter l'utilisation des connaissances.

Ce projet, a pour but de faire évoluer les outils de recrutement en ligne conformément aux attentes des demandeurs d'emploi et des recruteurs. En effet, les outils actuels de recrutement sont uniquement basés sur des recherches par mots-clés dans des banques de CV. Ce type de recherche permet au mieux d'identifier des diplômes et des postes, mais n'autorise pas la gestion des compétences professionnelles, qui sont devenues primordiales de par l'évolution du marché du travail et des profils professionnels. Ces compétences se déclinent en termes de connaissances techniques, de savoir-faire, de comportements et traits de caractère, et sont liées à des connaissances portant sur les secteurs de l'économie, les types d'entreprises, leur organisation et leurs activités. Or, les CV

déposés par milliers sur les sites de e-recrutement n'expriment pas explicitement les compétences des candidats et des liens doivent être établis entre la formation et les activités d'un demandeur d'emploi d'une part, et les exigences des recruteurs d'autre part. Etablir automatiquement ce lien nécessite une gestion informatisée des compétences professionnelles. Des ontologies doivent donc être construites pour spécifier les différentes connaissances explicitées précédemment, et seront exprimées en RDF(S)/DAML+OIL afin d'être implantées dans les outils de recrutement en ligne.

7. LA METHODOLOGIE DE LA CONSTRUCTION

7.1 Le cycle de vie des ontologies

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. Un cycle de vie inspiré du génie logiciel est proposé dans **[DIENG 2001]**. 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 1997].** 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 en *figure 4* [**Fabien, 2002**]. 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.

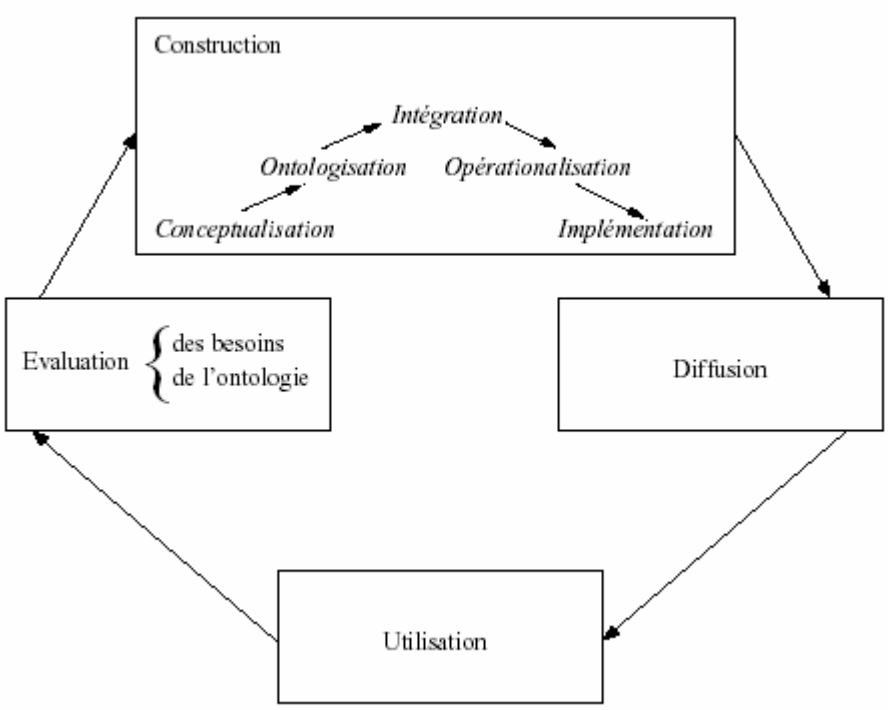


Fig 4. cycle de vie de d'ontologie

7.2 Les méthodologies de construction d'ontologies

Bien qu'aucune méthodologie générale n'ait pour l'instant réussi à s'imposer, de nombreux principes et critères de construction d'ontologies ont été proposés. Ces méthodologies peuvent porter sur l'ensemble du processus et guider l'ontologiste à toutes les étapes de la construction. C'est le cas de METHONTOLOGY, élaborée en 1998 par A. GOMEZ-PEREZ, qui couvre tout le cycle de vie d'une ontologie **[FERNANDEZ 1997]**. M. USCHOLD et M. KING ont également proposé une méthodologie générale, inspirée de leur expérience de construction d'ontologies dans le domaine de la gestion des entreprises. La méthodologie présentée par M. GRUNINGER et M. S. FOX est elle aussi issue d'une expérience de construction d'ontologie sur ce domaine.

D'autres méthodologies se focalisent sur une des étapes du processus. Celle présentée dans **[Aussenac-Gilles 2000]** insiste sur l'étape de conceptualisation par l'analyse d'un corpus textuel. La méthodologie ONTOSPEC de G. KASSEL constitue une aide à la structuration des hiérarchies de concepts et de relations durant la phase

d'ontologisation **[KASSEL 2002].** C'est également le cas des principes différentiels énoncés par B. BACHIMONT et des critères de classification des propriétés, proposés par N. GUARINO et C. WELTY.

Mais quelque soit la méthodologie adoptée, le processus de construction d'une ontologie est une collaboration qui réunit des experts du domaine de connaissance, des ingénieurs de la connaissance, voire les futurs utilisateurs de l'ontologie. Cette collaboration ne peut être fructueuse que si les objectifs du processus ont été clairement définis, ainsi que les besoins qui en découlent.

7.2.1 L'évaluation des besoins

Le but visé par la construction d'une ontologie se décline en 3 aspects :

L'objectif opérationnel : il est indispensable de bien préciser l'objectif opérationnel de l'ontologie, en particulier à travers des scénarios d'usage **[BIEBOW 1999];**

Le domaine de connaissance : il doit être délimité aussi précisément que possible, et découpé si besoin est en termes de connaissances du domaine, connaissances de raisonnement, connaissances de haut niveau (communes à plusieurs domaines) ;

Les utilisateurs : ils doivent être identifiés autant que faire se peut, ce qui permet de choisir, en accord avec l'objectif opérationnel, le degré de formalisme de l'ontologie, et sa granularité. Une fois le but défini, le processus de construction de l'ontologie peut démarrer, en commençant par la phase de conceptualisation.

7.2.2 La conceptualisation

La conceptualisation consiste à identifier, dans un corpus, les connaissances du domaine. Cette conceptualisation peut se décomposer en deux étapes. Tout d'abord, il faut faire le tri entre connaissances spécifiques au domaine et celles qui, bien que présentes dans le corpus, ne participent qu'à l'expression des connaissances du domaine. En outre, s'il est prévu d'intégrer d'autres ontologies, les connaissances spécifiées dans ces ontologies ne doivent pas être prises en compte. La nature conceptuelle (concepts, relations, propriétés des concepts et relations, règles, contraintes, etc.) des connaissances ainsi extraites du corpus doit ensuite être précisée. Des choix liés aux contextes d'usage de l'ontologie doivent donc être effectués dès cette étape. La découverte des connaissances d'un domaine peut s'appuyer à la fois sur l'analyse de documents et sur l'interview d'experts du domaine. Ces activités doivent être raffinées au fur et à mesure que la conceptualisation émerge. Les

interviews très ouvertes et les brain-stormings doivent alors laisser place à des questionnaires permettant, par exemple, de préciser la sémantique différentielle d'un concept mis en évidence lors d'une interview. De même, l'analyse informelle des textes doit être doublée par une analyse automatique qui permet de détecter les termes et structures sémantiques (définitions, règles) présentes dans le corpus **[Fernandez 1997]**. L'analyse de corpus ne peut suffire à elle seule à spécifier la sémantique du domaine. En effet, l'écrit n'est qu'un support, et les connaissances qui y sont représentées ne prennent sens que lorsqu'elles sont lues par un expert, le terme expert désignant justement ici une personne pour qui le corpus fait sens. De plus, la sémantique est partiellement contrainte, voire donnée par les scénarios d'usage des connaissances et par l'étude des réponses attendues du système dans des cas donnés. Dans cette optique, l'utilisation de questions de compétences, c'est-à-dire de questions auxquelles le système projeté est censé pouvoir répondre, est préconisée par M. GRUNINGER et M.S. FOX. Exprimées tout d'abord en langage naturel, ces questions servent à la conceptualisation, du fait qu'on en peut extraire le vocabulaire du domaine et sa sémantique. On peut ensuite les formaliser et les utiliser pour valider l'ontologie construite, une fois celle-ci intégrée dans un SBC.

L'analyse d'un ensemble de documents, ou l'interview de plusieurs experts peut facilement mener à des différences significatives voire à des contradictions au niveau du sens prêté aux concepts ou aux relations. En effet, les connaissances humaines sont essentiellement subjectives et ne seront pas exprimées de la même façon par tous les experts. Il faut également être conscient que seuls des symboles sont manipulés dans la machine, symboles que les utilisateurs interpréteront avec leur propre subjectivité. Il convient donc de restreindre au maximum les possibilités d'interprétation des termes et propriétés utilisés. Afin d'objectiver les connaissances, une normalisation sémantique est nécessaire, normalisation qui doit être le fruit d'un dialogue entre experts. De manière générale, l'échange entre experts et entre les experts et les ingénieurs de la connaissance est le meilleur moyen de faire émerger une sémantique claire et non ambiguë **[Fernandez 1997]**.

Certaines connaissances implicitement utilisées dans le domaine ne sont cependant jamais exprimées, ni dans le corpus, ni par les experts, car elles vont de soi pour tous. Un des points les plus délicats de la conceptualisation consiste donc à identifier ces connaissances. La mise en évidence de ces connaissances implicites ne peut a priorise faire que lors de l'utilisation de l'ontologie, lors d'une phase de test opérationnel et/ou

de test des questions de compétences. Ceci montre que le processus de construction des ontologies ne peut être séquentiel et que des aller-retours entre les différentes phases du processus sont à prévoir. Une fois les concepts et relations identifiés par leurs termes, il faut en décrire la sémantique en indiquant, à priori en langage naturel, leurs instances connues, les liens qu'ils entretiennent entre eux, leurs propriétés. La description d'une primitive conceptuelle doit contenir des liens vers les parties du corpus qui mettent en évidence sa sémantique, ce qui permet, au cas où une ambiguïté sémantique demeure, de revenir au corpus. Une fois les ressources cognitives passées au travers du tamis de la conceptualisation, il convient de formaliser, au cours de la phase d'ontologisation, le modèle conceptuel obtenu.

7.2.3 L'ontologisation

Une formalisation partielle, respectant l'intégrité du modèle conceptuel, va permettre, à cette étape, de construire une ontologie proprement dite. Afin de respecter les objectifs généraux des ontologies, T. GRUBER propose 5 critères permettant de guider le processus d'ontologisation [GRU 93] :

- la *clarté* et *l'objectivité* des définitions, qui doivent être indépendantes de tout choix d'implémentation ;
- la *cohérence* (consistance logique) des axiomes ;
- *l'extensibilité* d'une ontologie, c'est-à-dire la possibilité de l'étendre sans modification ;
- la *minimalité des postulats d'encodage*, ce qui assure une bonne portabilité ;
- la *minimalité du vocabulaire*, c'est-à-dire l'expressivité maximum de chaque terme.

De plus, il faut bien voir que l'ontologisation est une traduction dans un certain formalisme de connaissances exprimées a priori en langage naturel. Le respect de la sémantique du domaine doit être assuré par un *engagement ontologique*, notion proposée initialement par T. GRUBER comme un critère pour utiliser une spécification partagée d'un vocabulaire. « *We say that an agent commits to a knowledge-level specification if its observable actions are logically consistent with the specification* » **[GRUBER 93]**. Pour T. GRUBER, un *engagement ontologique* est une garantie de cohérence entre une ontologie et un domaine, mais pas une garantie de complétude de l'ontologie. N. GUARINO définit *l'engagement ontologique*

comme une relation entre un langage logique et un ensemble de structures sémantiques ; plus précisément, le sens d'un concept est donné par son extension dans l'univers d'interprétation du langage. Respecter l'engagement ontologique revient à donner à chaque concept son extension et à manipuler ce concept conformément au sens prescrit par cette extension **[GUARINO 1994]**. B. BACHIMONT, dans **[BACHIMONT 2000]**, distingue *l'engagement sémantique*, qui, à travers des principes différentiels, permet de préciser le sens des concepts (concepts sémantiques) de manière non ambiguë, et *l'engagement ontologique* qui associe des extensions à des concepts (concepts formels). Deux concepts sémantiques sont identiques si leurs interprétations, conformément aux principes différentiels utilisés, sont les mêmes. Deux concepts formels sont identiques s'ils ont même extension. Ces *engagements sémantiques et ontologiques* doivent être garantis par une structuration sémantique des connaissances, préalable à la formalisation proprement dite. Cette structuration est de plus nécessaire pour combler le fossé formel entre les connaissances conceptualisées et le formalisme utilisé pour les représenter en machine. Cette structuration va consister à préciser les liens sémantiques entre les différentes primitives conceptuelles, en particulier les liens de subsomption entre concepts et entre relations **[Fernandez 1997**]. L'ontologisation doit mener à la construction de hiérarchies de concepts, de relations, mais aussi d'attributs des concepts. USCHOLD préconise de bâtir ces hiérarchies de bas en haut, c'est-à-dire en identifiant les concepts de base puis en regroupant ces concepts à l'aide de concepts plus généraux **[USCHOLD 1996].** On donne ainsi la priorité aux concepts de bas niveau réellement utilisés dans le domaine, par rapport aux concepts qui ne sont souvent qu'ajouter artificiellement pour bâtir la hiérarchie. Par exemple, pour construire une ontologie de la géométrie, les concepts de « point » et « droite » vont, entre autres, être considérés, mais le concept « ensemble de points » peut être ajouté, concept qui subsume « point » et « droite », et ceci uniquement pour structurer l'ontologie. Cette hiérarchisation doit s'accorder avec les propriétés des concepts et des relations et être cohérente avec les intensions et extensions des concepts. B. BACHIMONT propose de commencer par bâtir une ontologie différentielle en organisant les concepts à l'aide des quatre principes suivant :

1. **Communauté avec le père ou principe de similarité** : un concept partage l'intension de son concept père ;

2. **Différence avec le père ou principe de différence** : l'intension d'un concept est différente de celle de son concept père, sinon il n'y aurait pas besoin de définir le concept fils ;
3. **Communauté avec les frères ou principe de sémantique unique** : une propriété est commune aux concepts frères issus du même concept père mais s'exprime différemment pour chaque frère. e.g. : les concepts « homme» et « femme » portent la propriété « sexe » héritée de leur concept père « humain », mais cette propriété vaut « masculin » chez « homme » et « féminin » chez « femme » ;
4. **Différence avec les frères ou principe d'opposition :** les frères doivent tous être incompatibles, sinon il n'y aurait pas besoin de tous les définir.

Dans une telle ontologie, il ne peut y avoir d'héritage multiple, la hiérarchie des concepts différentiels ne peut donc qu'être un arbre et non un graphe. En sus d'en guider la construction, ces principes permettent d'autre part de vérifier la cohérence de la hiérarchie. Une fois bâtie l'ontologie différentielle, on ajoute l'ontologie référentielle, i.e. les extensions des concepts. Dans le cadre référentiel, une instance peut faire partie de l'extension de plusieurs concepts. Les arbres de concepts et de relations obtenus sont ensuite étiquetés avec les propriétés et attributs des primitives, ce qui permet de présenter de manière synthétique et structurée l'ensemble des connaissances identifiées **[GUARINO 2000].**

Une fois le modèle conceptuel structuré, il faut le traduire dans un langage semi-formel de représentation d'ontologies. Le travail de structuration peut d'ailleurs être mené en même temps, les langages en question offrant tous la possibilité de représenter des hiérarchies de concepts et de relations. La sémantique de la subsomption n'est cependant pas toujours conforme aux principes différentiels cités précédemment. Parmi les langages de représentation développés au niveau conceptuel, trois grands modèles sont distingués : *les langages à base de frame*, *les logiques de description et le modèle des graphes conceptuels.*

Introduit dès les années 70 en IA, le modèle des frames a depuis été adapté à d'autres problématiques puisqu'il a donné naissance au modèle objet, qui envahit peu à peu les différentes branches de l'informatique. Une frame représente n'importe quelle primitive conceptuelle et est dotée d'attributs (*slots*), qui peuvent porter différentes valeurs (*facets*), et d'instances. La sémantique de la subsomption est purement référentielle : une frame *F1* est plus spécifique qu'une frame *F2* si toute instance de *F1* est instance de *F2*. FLOGIC est l'exemple le plus connu de langage à base de

frames. L'OPEN KNOWLEDGE BASE CONNECTIVITY (OKBC), protocole et API de requête et d'interfaçage entre bases de connaissances, utilise également le modèle de frame.

Les logiques de description permettent de représenter les connaissances sous forme de concepts, de rôles et d'individus. Les rôles sont des relations binaires entre concepts et les individus sont les instances des concepts. Les propriétés des concepts, rôles et individus sont exprimées en logique des prédicats, en particulier les propriétés de subsomption. Au niveau terminologique sont définis les concepts, les rôles et leurs propriétés. Les faits portant sur des individus (types des individus et relations entre individus) sont exprimés au niveau factuel. LOOM et KL-ONE sont des exemples de systèmes implémentant ce modèle. Il est de plus utilisé dans le langage de représentation de connaissance OIL développé pour le Web.

Introduit par SOWA au début des années 80, le *modèle des graphes conceptuels* se décompose 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 et PROLOG+CG, une extension de PROLOG basée sur les graphes conceptuels.

Quelques uns de ces langages ou des langages utilisant ces modèles sont déjà opérationnels et les ontologies exprimées dans ces formalismes peuvent être directement utilisées en machine. Dans les autres cas, une opérationnalisation de l'ontologie est nécessaire.

7.2.4 L'opérationnalisation

L'opérationnalisation consiste à outiller une ontologie pour permettre à une machine, via cette ontologie, de manipuler des connaissances du domaine. La machine doit donc pouvoir utiliser des mécanismes opérant sur les représentations de l'ontologie. Or, si beaucoup de langages réifiant les modèles cités précédemment autorisent l'expression de connaissances inférentielles, peu sont outillés pour rendre possible la manipulation de ces connaissances. Le modèle des graphes conceptuels fait exception car la représentation des connaissances sous forme de graphes permet de mettre en

oeuvre des raisonnements par des opérations formelles sur les graphes (comparaison, fusion, etc.). La façon de mener ces opérations dépend cependant de l'objectif opérationnel du système envisagé.

Dans le cas où le langage d'ontologisation n'est pas opérationnel, il est nécessaire, soit d'outiller ce langage, dans la mesure du possible, soit de transcrire l'ontologie dans un langage opérationnel. Mais certains langages offrent des possibilités de raisonnements limités qui peuvent convenir à certaines applications limitées. Par exemple, les langages à base de frames et les logiques de description permettent de savoir si une connaissance donnée, ou une connaissance plus spécifique qu'une connaissance donnée, est présente dans une base de connaissances en utilisant la relation de subsomption. Dans le cas d'un simple système de stockage et de consultation de connaissances, de tels langages sont donc suffisants.

Finalement, l'ontologie opérationnalisée est intégrée en machine au sein d'un système manipulant le modèle de connaissances utilisé via le langage opérationnel choisi. Mais avant d'être livrée aux utilisateurs, l'ontologie doit bien sur être testée par rapport au contexte d'usage pour lequel elle a été bâtie.

7.3 L'évaluation et l'évolution d'une ontologie

L'évaluation 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. Si cette dernière répond aux attentes, un système qui l'implémente doit donner les réponses prévues aux questions de compétences. 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.

De plus, la validation de l'ontologie en amont de son opérationnalisation est souhaitable. Elle évite de propager des erreurs qui, si les réponses fournies par le système aux questions de compétences se révèlent fausses, peuvent être difficilement repérables. La validité des hiérarchies doit donc être testée dès la phase d'ontologisation, aussi bien du point de vue formel que du point de vue sémantique. La validation formelle consiste à vérifier s'il n'y a pas de cycle, c'est-à-dire de définition en boucle, s'il n'y a pas redondance de concepts ou de relations, si 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. Des vérifications liées aux choix de modélisation sont également à effectuer, par exemple la détection de l'héritage multiple.

La *validation sémantique* permet de contrôler que la structure des hiérarchies est correcte vis-à-vis des principes différentiels utilisés. En particulier, la cohérence d'une ontologie vis-à-vis des principes énoncés par B. BACHIMONT peuvent être facilement contrôlée. Les méta-propriétés proposées par C. WELTY et N. GUARINO permettent de vérifier la cohérence sémantique de l'ontologie puisque ces méta-propriétés imposent des contraintes sur les liens de subsomption.

Si l'évaluation de l'ontologie a échoué, il est nécessaire de la faire évoluer. Se pose alors le problème délicat consistant à modifier certaines parties de l'ontologie en s'assurant de ne pas provoquer de nouvelles erreurs. Il paraît raisonnable de remonter s'il le faut jusqu'à l'étape de conceptualisation afin d'éviter des modifications qui ne respectent pas la sémantique du domaine. C'est dans ce cadre que les documents produits lors des différentes phases du processus de construction de l'ontologie vont s'avérés précieux. Le deuxième cas nécessitant l'évolution d'une ontologie est celui où les objectifs changent, c'est-à-dire que le contexte d'usage est modifié, ou que le domaine de connaissance est élargit. Là aussi, l'ajout de nouvelles connaissances est un processus délicat, compliqué par le fait qu'une ontologie va croître par agrégation de connaissances dans toutes les directions, et non par ajout d'une couche de connaissances [**FERENDAZ 1997**]. On peut décider soit de construire une nouvelle ontologie avec les connaissances à ajouter et l'intégrer dans l'ontologie déjà constituée, soit d'agréger directement les nouvelles connaissances dans l'ontologie existante. Dans le premier cas, les problèmes auxquels on va devoir faire face sont ceux posés par la fusion d'ontologies, qui reste un thème de recherche encore peu exploré.

7.4 La fusion d'ontologies

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. L'alignement suffit dans le cas de l'utilisation d'ontologies portant sur des domaines de connaissance complémentaires, ou sur des domaines de niveaux sémantiques différents. Par exemple, l'utilisation, dans un même système, d'une ontologie de haut niveau et d'une ontologie de domaine ne va nécessiter qu'une compatibilité entre les deux. La compatibilité de deux ontologies est assurée par

l'utilisation des mêmes formalismes de représentation, ou l'utilisation de formalismes compatibles, mais également par la compatibilité des modèles de connaissance utilisés.

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. Le problème se ramène alors à l'intégration d'une ontologie dans une autre. La fusion de deux ontologies suppose la présence dans ces deux ontologies d'entités conceptuelles (concepts ou relations) communes. 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 différents critères qui peuvent alors être appliqués pour repérer les similarités entre entités conceptuelles sont :

- **La similarité des termes** désignant deux entités ;
- **La similarité des propriétés** portées par deux entités ;
- **La similarité des entités subsumant ou étant subsumées** par deux entités.

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. 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.

La fusion d'ontologies apparaît donc comme un processus délicat, qui suppose au minimum une compatibilité entre les formalismes de représentation et entre les modèles de connaissances utilisées. Pour le moment, la diversité prévaut dans ce domaine, comme le démontre la variété des outils de construction d'ontologies disponibles.

8. LES OUTILS DE CONSTRUCTION D'ONTOLOGIES

De nombreux outils de construction d'ontologies utilisant des formalismes variés et offrant différentes fonctionnalités ont été développés. Seuls les plus significatifs seront cités ici, essentiellement ceux qui constituent des implémentations de méthodologies. Tous ces outils offrent des supports pour le processus d'ontologisation, mais peu offrent une aide à la conceptualisation. On peut cependant citer TERMINAE, qui, à travers l'outil d'ingénierie linguistique LEXTER, permet d'extraire d'un corpus textuel les candidats-termes d'un domaine. Ces concepts doivent ensuite être triés par un expert et organisés hiérarchiquement, puis la sémantique du domaine est précisée à travers des axiomes.

Les outils d'aide à la construction d'ontologie sont plus ou moins indépendants des formalismes de représentation. DOE (DIFFERENTIAL ONTOLOGY EDITOR) offre la possibilité de construire les hiérarchies de concepts et relations en utilisant les principes différentiels énoncés par B. BACHIMONT, puis en ajoutant les concepts référentiels. La sémantique des relations est ensuite précisée par des contraintes. Ce n'est qu'une fois l'ontologie ainsi structurée qu'elle est formalisée en utilisant la syntaxe XML (une extension permettant la transcription en RDF(S)/DAML+OIL est prévue).

De même, l'outil ODE (ONTOLOGY DESIGN ENVIRONMENT) permet de construire des ontologies au niveau connaissance, comme le préconise la méthodologie METHONTOLOGY proposée par A. GOMEZ-PEREZ. L'utilisateur construit son ontologie dans un modèle de type frame, en spécifiant les concepts du domaine, les termes associés, les attributs et leurs valeurs, les relations de subsomption. L'ontologie opérationnelle est alors générée en utilisant les formalismes ONTOLINGUA ou FLOGIC. WEBODE est l'adaptation de ODE pour le Web.

ONTOEDIT (ONTOLOGY EDITOR) est également un environnement de construction d'ontologies indépendant de tout formalisme. Il permet l'édition des hiérarchies de concepts et de relations et l'expression d'axiomes algébriques portant sur les relations, et de propriétés telles que la généricité d'un concept. Des outils graphiques dédiés à la visualisation d'ontologies sont inclus dans l'environnement. ONTOEDIT intègre un serveur destiné à l'édition d'une ontologie par plusieurs utilisateurs. Un contrôle de la cohérence de l'ontologie est assuré à travers la gestion des ordres d'édition. Enfin, un plug-in nommé ONTOKICK offre la possibilité de

générer les spécifications de l'ontologie par l'intermédiaire de questions de compétences.

Parmi les outils non liés à des formalismes de représentation, citons pour finir PROTEGE2000 interface modulaire permettant l'édition, la visualisation, le contrôle (vérification des contraintes) d'ontologies, l'extraction d'ontologies à partir de sources textuelles, et la fusion semi-automatique d'ontologies. Le modèle de connaissances sous-jacent à PROTEGE2000 est issu du modèle des frames et contient des classes (*concepts*), des slots (*propriétés*) et des facets (valeurs des propriétés et contraintes), ainsi que des instances des classes et des propriétés. PROTEGE2000 se veut compatible avec le modèle OKBC mais en diffère sur certains points. Par exemple, une frame, qui est soit une classe, soit un slot, soit une facet, est instance d'une classe et une seule dans PROTEGE2000 ce qui n'est pas toujours le cas dans OKBC, où une même frame peut d'ailleurs représenter plusieurs objets conceptuels. PROTEGE2000 autorise la définition de méta-classes, dont les instances sont des classes, ce qui permet de créer son propre modèle de connaissances avant de bâtir une ontologie.

Contrairement aux outils précédemment cités, ONTOLINGUA est un serveur d'édition d'ontologies au niveau symbolique : une ontologie est directement exprimée dans un formalisme également nommé ONTOLINGUA, qui constitue en fait une extension du langage KIF (KNOWLEDGE INTERCHANGE FORMAT). Le langage ONTOLINGUA utilise des classes, des relations, des fonctions, des objets (instances) et des axiomes pour décrire une ontologie. Une relation (ou une classe) peut contenir des propriétés nécessaires (contraintes) ou nécessaires et suffisantes qui définissent la relation (ou la classe). ONTOLINGUA propose un outil permettant d'inclure une ontologie dans celle en cours de construction. L'inclusion consiste à ajouter à l'ontologie courante les axiomes de l'ontologie à inclure, après traduction des axiomes. La traduction consiste à établir une relation d'identité entre les termes des deux ontologies qui désignent les mêmes classes ou relations. Ces termes sont tous différents entre eux car préfixés par le nom de l'ontologie à laquelle ils appartiennent.

Moins ambitieux qu'ONTOLINGUA, OILED (OIL EDITOR) est un éditeur d'ontologies utilisant le formalisme OIL. Il est essentiellement dédié à la construction de petites ontologies dont on peut ensuite tester la cohérence à l'aide de FACT, un moteur d'inférences bâti sur OIL.

Après ce bref survol des outils de construction d'ontologies les plus significatifs, la présentation de quelques applications utilisant les ontologies va permettre de mieux cerner les enjeux de l'ingénierie ontologique.

8.1 Exemples

8.1.1 OCML : un langage facilitant l'opérationnalisation des ontologies

a- Exemple de représentations :

```
(def-class  père (parents, homme) ?p
    :iff-def
        (or  (and (parents ?p) (a_pour_sexe ?p «masculin»))
             (exists  ?e  (and  (homme ?p)
                                (a_pour_enfant ?p ?e)
                                (personne ?e)))))
(def-relation  a_pour_père (?p ?h)
            :constraint  (and  (personne ?p) (homme ?h)))
```

b- Caractéristiques principales du langage

OCML est développé au Knowledge Media Institute de l'Open University (Mylton Keynes, Angleterre). Il a été initialement défini dans le cadre du projet VITAL pour permettre de spécifier des modèles de résolution de problèmes au niveau connaissance, puis de les opérationnaliser.

La couche domaine du langage étant considérée comme équivalente aux connaissances visées par ONTOLINGUA, OCML est supposé opérationnaliser des ontologies spécifiées en ONTOLINGUA.

La principale caractéristique d'OCML est de supporter différents styles de spécification : informel, formel et opérationnel, l'opérationnel correspondant à du «prototypage au niveau connaissance».

8.1.2 DEFONTO : un langage permettant l'expression de méta-connaissances

a- Exemple de représentations :

```
(DefGenConcept  #père
    = [#parents] -> (MI#a_pour_sexe) -> «masculin»
    = [#homme] -> (ME#a_pour_enfant) -> [#personne] )
(DefRelation  #a_pour_père
    IsA  [#a_pour_parents]
    RelationProperties
     -> (#has_for_domain) -> [#personne]
     -> (#has_for_range) -> [#homme] )
```

b- Caractéristiques principales du langage

DefOnto est développé dans l'équipe Ingénierie des Connaissances du LaRIA à l'Université de Picardie Jules Verne, comme sous-langage du langage d'**opérationnalisation** de modèles de résolution de problèmes Def*.

DefOnto permet de représenter des méta-connaissances (ex : des propriétés de propriétés) et cette caractéristique est notamment importante pour rendre compte de concepts de résolution de problèmes.

Un objectif visé par la définition de DefOnto est de doter le langage de mécanismes de compilation modulaire pour faciliter le développement et la maintenance des ontologies formelles.

8.1.3 OIL : un langage pour échanger des ontologies sur le Web

a- Exemple de représentations

```
class-def defined père
    subclass-of
      ((parents and (slot-constraint a_pour_sexe
                          has-filler «masculin»))
          or (homme and (slot-constraint
                                  a_pour_enfant
                          has-value personne)))
slot-def a_pour_père
     subclass-of a_pour_parents
          domain personne
          range homme
```

b- Caractéristiques principales du langage

OIL résulte d'une initiative internationale rassemblant plusieurs équipes collaborant au développement de ce langage dans les projets IST : On-To-Knowledge et IBROW3 et dans le programme américain DAML : langage DAML-OIL : ***http://www.daml.org***

Les primitives de représentation le situent à mi-chemin entre langages de Frames et Logiques de description.Un effort particulier a été fait pour l'échange d'ontologies sur le Web. Sur le plan syntaxique, OIL est doté de deux notations compatibles Web, définies selon les deux standards respectifs que sont XML et RDF.

8.2 Bilan

Plusieurs langages sont actuellement en cours de définition, qu'il s'agisse de langages de modélisation (Mdos) ou de représentation opérationnels : DefOnto, OIL et PowerLOOM. Tous ces langages permettent de représenter un noyau commun de connaissances.

Plusieurs articles **[Corcho 2000] [Barry 2001]** présentent des études comparatives de langages de spécification d'ontologies, opérationnels ou non, les comparaisons portant sur la puissance d'expression et les capacités inférentielles. Les résultats de la 2° étude sont accessibles à l'adresse :

http://www.laria.u-picardie.fr/EQUIPES/ic/LangComp/

9. CONCLUSION

La variété des besoins et des champs de recherche concernés par le développement des ontologies explique la diversité des objets dénotés par le terme «ontologie». L'ingénierie Ontologique, au sein de l'Ingénierie des Connaissances, définit des concepts, méthodes et outils, pour rationaliser le développement des ontologies. Cette discipline est en plein essor, en témoigne le nombre important de projets en cours et de séminaires qui lui sont consacrés. Cependant, les propositions émanant des projets et les pratiques ne sont pas encore unifiées. Un effort de synthèse et de diffusion hors du cadre académique reste encore à réaliser.

L'essor récent des ontologies traduit les espoirs qu'elles véhiculent et la diversité des types d'application qui peuvent les intégrer sous des formes plus ou moins riches : recherche d'information sur le Web ou dans de grandes bases documentaires, gestion documentaire (indexation, classification, référencement), capitalisation de connaissances, etc. Or ces espoirs ne sont pas encore des certitudes, et ne le deviendront qu'au terme de nouvelles recherches et expérimentations. On ne peut pas considérer que les technologies associées à la construction des ontologies soient parvenues à maturité.

A ce jour, évaluer la pertinence des ontologies se heurte à plusieurs obstacles : le coût de leur construction, le passage à l'échelle du Web de travaux de recherche sur des cas d'école et la difficulté de mesurer le gain qu'apporte une ontologie à une application donnée.

Mettre en oeuvre un projet de construction d'ontologie aujourd'hui, avec l'état de l'art actuel, est donc à la fois une contribution à l'avancement des résultats dans ce domaine et un pari sur leur pertinence.

Chapitre 2

La construction d'ontologie à partir des textes techniques

1. INTRODUCTION

Dans ce chapitre, nous exposons les différentes méthodes de la construction d’ontologie. Nous nous intéressons aux méthodes qui utilisent les textes comme source d’information. Donc, la première étape de la construction d’ontologie est étroitement liée au domaine de TALN. Nous abordons principalement les travaux de BACHIMONT qui concernent les éléments théoriques de la construction d’ontologie à partir des textes techniques et les méthodologies suivies pour un engagement ontologique, puis nous exposons quelques principes de la construction d’ontologie. Finalement, nous présentons une méthodologie linguistiquement fondue qui s’appelle : TERMINAE.

2. METHODES DE LA CONSTRUCTION D'ONTOLOGIE

Jusqu'en 1996, les premières ontologies ont été développées de façon complètement artisanale, sans suivre de méthode prédéfinie.
De ces premiers projets (ex : Mikrokosmos, Enterprise Ontology, TOVE, MENELAS) sont issues des listes de recommandations constituant des ébauches de méthodes, ou cadres méthodologiques.
Depuis 1998, on assiste à la naissance de cadres méthodologiques plus élaborés inspirés des méthodes de l'Ingénierie des Connaissances (ex : METHONTOLOGY) et fondés, soit sur la linguistique (ex : TERMINAE), soit sur l'Ontologie (ex : principes proposés par N. Guarino).

2.1 Les anciens projets

La méthode « Enterprise Ontology » [Uschold 95] ***http://www.aiai.ed.ac.uk/project/enterprise***	La méthode, basée sur l'expérience du développement de l'ontologie *Enterprise Ontology*, repose sur l'identification de différentes étapes • Identification du POURQUOI de l'ontologie ; • Construction de l'ontologie (identification des concepts clef ; modélisation informelle ; formalisation) et intégration d'onto-logies existantes ; • Évaluation et documentation de l'ontologie. Cette méthode s'inspire du développement de SBCs. Les étapes et sous-tâches sont décrites de façon abstraite. Les techniques à utiliser pour les sous-tâches ne sont pas précisées (ex : comment identifier les concepts clef ? Quel langage de formalisation utiliser ?).
La méthode « TOVE » [Grüninger 1995] ***http://www.eil.utoronto.ca/tove/ontoTOC.html***	Cette méthode est basée sur l'expérience du développement de l'ontologie du projet TOVE (TOrento Virtual Enterprise). Elle aboutit à la construction d'un modèle logique de connaissance. L'ontologie est développée selon les étapes suivantes :

- Identification de scénarii (problèmes) dépendants d'une application ;
- Formulation de questions informelles (basées sur les scénarii) auxquelles l'ontologie doit permettre de répondre ;
- Spécification d'une terminologie à partir des termes apparaissant dans les questions.
- Spécification formelle (en KIF) des axiomes et des définitions pour les termes de la terminologie.
- Évaluation de la complétude de l'ontologie.

La méthode reste spécifiée de façon abstraite. Ni les différentes étapes ni les techniques ne sont décrites en détail.

2.2 Une méthode inspiré de l'IC.

METHONTOLOGY

[Fernandez-Lopez 1999]

Cette méthode est développée au Laboratoire d'Intelligence Artificielle de l'Université polytechnique de Madrid. Elle vise la construction d'ontologies au «niveau connaissance». Ce projet a été motivé par le constat suivant : l'absence de méthodes ou de guides structurés est un obstacle au développement d'ontologies partagées et consensuelles. Il est également un obstacle à l'extension d'une ontologie existante ou à sa réutilisation dans d'autres ontologies.

L'approche de METHONTOLOGY, en référence à des approches comme CommonKads, permet la construction d'ontologies au niveau connaissances en préconisant, comme étape intermédiaire pour le développement de SBCs, une modélisation conceptuelle au « niveau connaissance » faisant abstraction des contraintes liées aux langages de programmation.

La méthode permet la construction d'ontologies au « niveau connaissance ». Elle inclut : l'identification du processus de développement d'ontologie (PDO), un « cycle de vie ontologique » (CVO) basé sur l'évolution de prototypes.

- un processus de développement d'ontologies (PDO) identifie les activités qui doivent être appliquées lors de la construction d'une ontologie, comportant des activités de gestion de projet (planification,

assurance qualité), des activités orientées développement (spécification, conceptualisation, formalisation) et des activités de support (documentation) ; chacune des 12 activités est clairement définie dans le processus de développement, par exemple l'activité de spécification établit pourquoi l'ontologie est construite, quelle sont ses buts d'utilisation et qui sont les utilisateurs.

- un cycle de vie des ontologies (CVO) définit l'ensemble des « étape » à travers lesquelles l'ontologie passes durant son temps de vie et décrit quelles activités sont à effectuer dans chacune de ces étapes et comment elles sont liées entre elles (relation de précédence….etc.).

METHONTOLOGY s'inspire d'une méthode de développement de SBCs. Elle est spécifiée de façon très détaillée et a été utilisée pour construire plusieurs ontologies dont l'ontologie des ontologies : *Reference Ontology*. Elle est supportée par l'outil ODE.

2.3 Apport méthodologique de l'Ontologie

***Travaux de N. Guarino au LADSEB (Padoue, Italie*)**

N. Guarino et son équipe cherchent à évaluer l'apport des notions et principes de l'Ontologie pour la construction d'ontologies. Leur proposition prend la forme d'une ontologie générique et d'une ontologie de méta- propriétés **[Guarino 2000].**

L'ontologie générique rassemble un ensemble d'objets abstraits et leur définition (ex : objet physique, substance, système, état, processus, activité). La construction d'une ontologie d'application peut donc se faire par spécialisation de l'ontologie générique.

Les méta-propriétés (ex : type, rôle) sont fondées sur des notions de l'Ontologie (ex : identité, unité, rigidité) **[Guarino 2000]** et permettent de vérifier la cohérence d'une ontologie d'application qui s'en trouvera d'autant plus facilement réutilisable.

3. LA CONSTRUCTION D'ONTOLOGIE A PARTIR DES TEXTES

De nombreuses méthodes de construction d'ontologies sont orientées sur des problèmes de cycle de vie de l'ontologie vue comme un logiciel (*voir chapitre 1 § 7.1)*. Elles sont basées sur des bons principes mais ne proposent pas de réelle méthodologie. À l'inverse, la méthodologie proposée par B. BACHIMONT est linguistiquement et épistémologiquement fondée **[BACHIMONT, 2000]** et c'est elle que nous allons décrire dans cette partie. A titre d'information, cette méthodologie a été élaborée à la suite du projet MENELAS et de la construction de son ontologie. D'autres méthodes fondées sur des principes proches ont été élaborées au sein du groupe TIA. Ne voulant pas développer une comparaison de ces méthodes, nous renvoyons le lecteur à, par exemple, [Aussenac-Gilles 2000].

La première question qui se pose pour développer une méthodologie de construction d'ontologies, est le matériau de départ : nous avons développé au chap. 2, § 1 que l'Ingénierie des connaissances avait souvent recours aux textes comme matériel de base pour élaborer ses artefacts. Ensuite, il y a le matériau d'arrivée, ici une ontologie formelle qui doit servir dans un SBC. La question est alors de caractériser le passage d'une connaissance exprimée sous forme linguistique à une connaissance formalisée.

L'art et la manière sont proposés par B. BACHIMONT dans la méthodologie qui suit.

3.1 Constitution du corpus

A partir de la description des besoins, il s'agit de choisir des textes de façon à couvrir complètement le domaine requis par l'application. Le choix nécessite une bonne connaissance du domaine autant que des textes eux-mêmes, afin de caractériser leur type et d'évaluer leur couverture du domaine. Ce choix n'est pas le seul fait du cogniticien, qui doit s'appuyer sur les connaissances d'experts et d'utilisateurs. Un glossaire sur le domaine peut être utile pour déterminer les sous-domaines à explorer et vérifier qu'ils sont tous couverts [**Aussenac-Gilles2000**]. Le corpus est ensuite préparé pour être traité informatiquement si besoin. Une évaluation du contenu du corpus permet de mieux en juger la pertinence et peut conduire à le modifier.

3.1.1 Notion de corpus

> ***Définition** : En linguistique, un **corpus** désigne l'ensemble des énoncés de la langue qui sont pris en compte et analysés lors d'une étude donnée.*

Avec la mise sur support informatique des documents, on parle de plus en plus de corpus en *ingénierie documentaire*, en *extraction* ou en *recherche d'information*. Un corpus est alors un ensemble de documents exploités avec un objectif particulier. De ce fait, il est en général construit pour cet objectif.

Dans le cas de l'acquisition de connaissances à partir de texte, les corpus sont choisis de manière à couvrir le domaine d'application, à fournir des connaissances pertinentes pour l'objectif fixé et à avoir une taille adaptée à un traitement outillé mais en partie manuel.

3.1.2 Caractéristiques d'un corpus

La plupart des corpus utilisés pour la construction d'ontologies d'entreprises s'appuient sur la documentation technique de l'entreprise (institut, usine, etc.), sur le contenu de bases de données semi-structurées, sur des retranscriptions d'entretiens ou encore sur des rapports internes, des fiches de retour d'expérience ou autres documents produits par l'entreprise. Il peut aussi s'agir de textes didactiques, de documents de communication interne ou externe (commerciale), de spécifications techniques, de normes, de comptes rendus d'expériences, d'articles scientifiques…

Les documents d'un corpus sont caractérisés par leur auteur, leur date de production, leur style, leur taille, leur support, leurs destinataires, leur contenu, etc. Il est important de connaître ces éléments au moment de choisir d'ajouter ou non un document au corpus.

Un corpus est *homogène* lorsqu'il contient des documents ayant plusieurs de ces caractéristiques communes (même type de contenu ou produits lors de la même activité, etc.), hétérogène sinon [**Aussenac-Gilles 2000**].

3.1.3 Corpus / application / domaine

Finalement, constituer le corpus, c'est trouver un compromis entre des facteurs contradictoires

- Couverture la plus large possible du domaine
- Couverture la plus fine, précise du domaine
- Adéquation avec l'application
- Homogénéité et cohérence des documents (au moins par sous-ensembles)
- Volume raisonnable
- Adéquation aux traitements informatiques
- Disponibilité des documents.

3.1.4 La démarche de constitution

3.1.4.1 Tâches

La constitution du corpus suppose plusieurs tâches étroitement liées et effectuées de manière cyclique jusqu'à parvenir à un état stable et satisfaisant du corpus :

- choisir des documents représentatifs du domaine étudié et/ou adaptés à l'application ciblée ;
- les mettre au format informatique adéquat ;
- décider de la manière de les traiter ;
- Évaluer ces documents, leur qualité et leur apport potentiel au modèle à construire.

3.1.4.2 Choisir des documents

Il s'agit de rechercher parmi les documents disponibles, si possibles sur support informatique, les mieux adaptés à l'application. Il faut ensuite constituer à partir de là un ou plusieurs ensembles cohérents, assez homogènes et qui répondent au compromis entre représentativité (sujet, genre textuel) et taille.

Le choix de la langue des documents engage celui de la langue dans le modèle final. Aujourd'hui, il est coûteux de constituer d'un premier jet des ontologies multilingues. Il n'est pas commode de travailler des documents de langues différentes pour alimenter l'ontologie. On préfèrera alors procéder en plusieurs temps, langue par langue.

3.1.4.3 Décider de la manière de les traiter

Nous venons de souligner que, pour traiter correctement les documents et savoir quelle valeur donner aux connaissances qu'ils contiennent, comment les structurer, etc., il est fondamental d'identifier des groupes homogènes, par type de document, de sujet ou de production.

Mais alors se pose la question de savoir si chaque groupe de documents va être traité séparément, et fournir une sorte d'ontologie locale, ou bien si on va chercher uniquement les points communs aux différents groupes, pour ne faire qu'une seule ontologie commune. Cela dépend encore de l'objectif de l'ontologie, si elle doit servir de modèle unificateur, de vecteur de cohérence, ou bien si elle doit aussi rendre compte des divergences de points de vue et tracer des ponts entre eux.

3.1.4.4 Mettre des documents au format informatique adéquat

Cette phase, purement technique, peut s'avérer délicate. Il s'agit de scanner les documents sur papier, de récupérer des champs textes dans des bases de données et de toute manière de ramener des formats plus ou moins complexes à des formats ASCII, tout en se donnant les moyens de pouvoir retrouver le document dont est issue une phrase.

Cette phase suppose un travail minutieux de vérification de la qualité des résultats obtenus. Des caractères parasites peuvent avoir un effet très négatif sur les résultats des outils de TAL ou même les empêcher de bien fonctionner (par exemple, textes tout en majuscules).

3.1.4.5 Evaluer le corpus

Un des moyens de repérer des erreurs dans le corpus et d'en évaluer le contenu est d'observer les premiers résultats produits par les logiciels d'analyse.

Le corpus textuel construit représente la source privilégiée des connaissances qui permettra de caractériser les notions utiles à la modélisation ontologique. Pour ce faire, on utilise des outils terminologiques pour commencer à modéliser le domaine. Ces outils, pour la plupart, reposent 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 – *ex. contextes partagés* – pour des regroupements de notions. Ils font ce qu'on appelle de l'extraction terminologique et permettent d'obtenir des signifiés linguistiques avec une organisation plus ou moins structurées, souvent sous forme de réseaux.

3.2 Acquisition des termes

Pour répondre à certaines préoccupations de l'ingénierie des ontologies, cette partie présente une approche d'aide à l'acquisition de connaissances à partir de corpus d'un domaine donné. Plus précisément, il s'agit d'une approche qui permet la recherche de *termes* à partir de textes d'un domaine donné pour l'aide à l'acquisition de *concepts* liés à ce domaine. Le but est de proposer un état de l'art dans le domaine de l'acquisition des connaissances, plus particulier dans le domaine de l'acquisition des termes à partir des textes techniques.

Il existe deux méthodes pour l'acquisition des termes à partir des textes, soit l'acquisition se fait de manière manuelle ou se fait de manière automatique ou semi-automatique.

3.2.2 Acquisition manuelle des termes

Acquérir des connaissances manuellement à partir de corpus de texte s'avère une tâche compliquée et coûteuse en temps. Certain auteur **(** voir, **[OUESLATI 1999])** a tenté par exemple d'encoder des règles englobant à la fois des connaissances lexicales et sémantiques afin d'extraire manuellement des connaissances du domaine. Ces règles manuelles sont souvent incomplètes et s'appliquent difficilement à un domaine nouveau. De plus, la maintenance de ces règles devient plus difficile dès que leur nombre croit.

C'est en partie en réponse à ces inconvénients, et dans le but de rendre l'analyse de corpus le plus possible automatique, plusieurs méthodes ont été conçues.

3.2.3 Acquisition automatique des termes

La disponibilité croissante de données sous forme numérique, permet de traiter de grandes masses d'information textuelle. C'est pourquoi des outils de traitement de corpus sont de plus en plus disponibles. Ils permettent d'effectuer des tâches allant du simple repérage de contextes (ex. les *concordanciers* : Ces sont des outils qui aident à afficher la liste des contextes d'un mot ou d'un groupe de mots dans un corpus, en permettant souvent la prise en compte de formes fléchies et des opérateurs de restriction de la recherche de contextes.) jusqu'aux traitements les plus complexes comme par exemple *l'extraction de terminologies*.

Pour ce faire, certaines méthodes consistent par exemple en l'examen au moyen de calculs linguistiques ou statistiques des unités linguistiques constituants le corpus afin de les organiser et de les représenter sous une forme exploitable.

3.2.3.1 Modèles mécaniques

À l'origine, ces travaux ne portaient pas directement sur l'acquisition automatique des termes, mais plutôt sur l'identification des *collocations*[1]. Puisqu'ils ont été repris comme point de départ pour d'autres travaux dans ce domaine, nous jugeons essentiel de les décrire. Les travaux décrits ici, que nous qualifions de mécaniques, reposent entièrement sur des algorithmes utilisant la force brute des ordinateurs. L'utilisation du qualificatif mécanique a pour but de représenter l'aspect très systématique de cette approche entièrement dépourvue de connaissances linguistiques ou statistiques. Leur but est essentiellement de relever des segments de texte qui se répètent à l'intérieur d'un corpus. Ces segments peuvent donc

[1] 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.

contenir des termes ou une manifestation de tout autre phénomène linguistique répétitif, observable en discours.

3.2.3.2.1 Exemples de la méthode

- Les travaux de Choueka, Klein et Neuwtiz (1983) et Choueka (1988) portant sur la langue anglaise. L'objectif de ces travaux est de fournir au lexicographe un outil de repérage des *collocations*. Par ce terme, il désigne les séquences de deux ou plusieurs mots consécutifs. Le corpus utilisé est d'une taille importante. L'approche adoptée est entièrement mécanique et se contente d'identifier les chaînes de caractères qui se produisent côte à côte plus d'une fois à l'intérieur d'un corpus. L'auteur ne prévoit pas une utilisation directe des résultats bruts. Ceux-ci doivent être revus par un lexicographe possédant la connaissance nécessaire à l'utilisation des résultats et à leur extrapolation. L'algorithme est très lent et beaucoup de bruit, mais il a cependant l'avantage d'être indépendant des langues, d'être systématique et de repérer l'ensemble des formes répétées dans la mesure où celles-ci ne connaissent pas de variation orthographique.
- Les travaux de Salem (1987) et Lebart et Salem (1988, 1994) sur les *segments répétés* en français. Ces derniers sont définis comme des suites de formes graphiques non séparées par un caractère délimiteur de séquence, qui apparaissent plus d'une fois dans le corpus de textes. les chercheurs utilisent des textes ayant été préalablement lemmatisés. L'avantage principal des SR est de mettre en lumière des redondances observées en discours. L'hypothèse sous-jacente aux travaux sur les SR est que la f*réquence* des segments permet de mettre en évidence les unités les plus intéressantes au sein d'un corpus.
- Les travaux de Drouin et Ladouceur (1994) s'inspirent des techniques proposées par Choueka et al. (1983). Ces travaux portent sur la langue française et visent l'extraction automatique des unités nominales. Les chercheurs utilisent comme point de départ une analyse des segments répétés. Le résultat obtenu à partir d'un corpus technique est ensuite filtré selon divers critères afin d'isoler les unités nominales permettant de cerner le contenu du corpus. De nombreux indices sont utilisés par les chercheurs afin de réduire le nombre d'entrées retenues : la fréquence, la morphologie et une analyse de similarité entre les SR retenus. Les résultats de ces travaux de recherche sont intéressants d'un point de vue de l'indexation des textes. Les SR retenus peuvent faire l'objet d'une utilisation dans le cadre d'une démarche visant l'acquisition automatique des termes.

- Dans sa thèse, Rochdi Oueslati (1999) propose une méthode d'aide à l'acquisition des connaissances à partir d'un corpus en langue française. Afin d'y parvenir, l'auteur propose une technique qui fait appel aux travaux sur les segments répétés présentés dans les paragraphes précédents. Les étapes adoptées par l'auteur pour l'acquisition des termes sont : le prétraitement du corpus, l'extraction et le filtrage des segments répétés, la structuration des segments répétés sous forme d'arbres de termes, le filtrage par un linguiste et la constitution d'une liste de termes. Le système proposé par l'auteur procède ensuite à un repérage de nouveaux termes à partir des termes ayant été validés. Ce processus d'identification de nouveaux termes utilise des formalismes permettant de repérer les têtes et les expansions déjà relevées et de les appliquer à de nouvelles structures. Les travaux de Oueslati (1999) offrent cependant l'avantage de bien mettre en évidence l'apport non négligeable que l'analyse des segments répétés peut avoir pour le travail terminologique.

3.2.3.2.2 Conclusion :

Même si les approches mécaniques sont, d'un point de vue informatique, plutôt lentes, elles offrent un avantage indéniable sur le travail manuel de dépouillement d'un corpus. Cet avantage est nettement mis en évidence lorsque le corpus analysé est imposant. l'approche des SR exige la mise en place d'une stratégie de filtrage afin d'isoler avec succès le phénomène faisant l'objet de la recherche.

3.2.3.2 Modèles linguistiques

Les systèmes présentés ici sont qualifiés de linguistiques puisqu'ils font appel à des techniques d'analyse reposant sur les connaissances actuelles de la langue et de sa structure. On distingue principalement les systèmes utilisant des informations syntaxiques et ceux qui utilisent des informations lexicales ou morphologiques. Les premiers reposent sur une analyse complète de la phrase en ses constituants afin d'en dégager les syntagmes intéressants selon les objectifs de la recherche. Dans le second cas, des grammaires locales procèdent à une analyse de surface de la phrase à la recherche de syntagmes potentiels.

3.2.3.2.1 Exemples de la méthode

- *Les travaux de David et Plante (1990); Plante, Dumas et Plante (2000)* : Nomino compte parmi les systèmes d'acquisition automatique de termes. Il a été élaboré dans le cadre d'une collaboration entre l'Office de la langue française du Québec et une équipe

du Centre d'ATO de l'Université du Québec à Montréal. La première version de ce logiciel se nommait Termino; il a depuis été remplacé par un nouveau système nommé Nomino. Nomino procède à l'acquisition des termes en quatre grandes étapes. La *première* consiste en un découpage du document en lexèmes et en phrases. Lors de la *deuxième* étape de traitement, chaque lexème identifié est soumis à une analyse morphosyntaxique qui a pour but de lemmatiser la forme et de lui attribuer une catégorie grammaticale. Les formes ambiguës peuvent se voir attribuer plusieurs catégories grammaticales. Le *troisième* traitement est une analyse syntaxique en vue de désambiguïser, en contexte, les formes qui ont reçu plus d'une catégorie grammaticale à l'étape précédente. À la fin de cette étape, toutes les unités de la phrase ne possèdent qu'une seule catégorie grammaticale. L'étape de désambiguïsation terminée, le logiciel peut procéder à l'identification des unités nominales complexes.

Nomino est le doyen des logiciels d'acquisition automatique des termes. Ses performances initiales et l'intérêt qu'il a su susciter ont lancé les recherches dans le domaine. Elles ont aussi démontré l'importance de l'ordinateur dans le travail terminologique. Le recours à des analyses morphosyntaxique et syntaxique lors de traitement rend le système très dépendant de sources externes d'informations linguistiques. Le logiciel est donc étroitement lié à la langue à laquelle il est destiné. L'adaptation de l'approche proposée à une autre langue constitue donc un projet d'envergure non négligeable.

- *Les travaux de Bourigault (1992)* : Le logiciel LEXTER a été élaboré par Didier Bourigault de sa thèse de doctorat (1992, 1994). Il a pour but d'enrichir les thésaurus d'un système d'indexation automatique de textes de la société.

  Pour le repérage automatique des termes Bourigault utilise le concept *de frontière de terme*. Il adopte une approche s'articulant autour d'une analyse syntaxique locale à partir des frontières de termes. Avec LEXTER, l'acquisition automatique des termes s'effectue en trois grandes étapes principales : l'étiquetage des formes, le découpage des textes en CT par repérage de frontières et la décomposition des groupes nominaux obtenus. L'analyse locale permet l'introduction du concept de frontières de termes, point fort des travaux de Bourigault. Ces dernières rendent possible un recensement des termes avec un minimum de connaissances linguistiques. L'absence de recours à des connaissances sur les termes rend l'approche indépendante des domaines de spécialité.
- *Les travaux de Voutilainen* (1993) sur l'anglais ont donné naissance au logiciel nommé NPtool. Contrairement aux travaux présentés dans les pages précédentes, la

raison d'être de NPtool n'est pas l'identification des termes, mais l'identification des syntagmes nominaux.

Ces travaux reposent sur une analyse syntaxique complète qui utilise une grammaire à contraintes afin d'identifier les syntagmes nominaux (SN). Le logiciel NPtool procède à une identification des SN en deux grandes étapes. Le logiciel procède d'abord à une attribution de toutes les catégories grammaticales possibles pour chacun des éléments de la phrase. Les ambiguïtés sont ensuite levées automatiquement afin que tous les mots possèdent une étiquette unique. La seconde étape d'analyse consiste en une identification des syntagmes nominaux. Pour ce faire, NPtool a recours à deux grammaires qui possèdent des niveaux de contraintes très différents. Ces grammaires sont dites lâches ou restrictives selon le nombre de SN qu'elles permettent de recenser. Le premier type de grammaire permet de détecter un grand nombre de SN (SN approximatifs) alors que le second n'accepte que des SN qui répondent à des contraintes beaucoup plus fortes (SN restreints). Les résultats obtenus à l'aide de cette technique sont impressionnants. NPtool permet d'atteindre un taux de rappel variant de 98,5 % à 100 % avec une précision allant de 95 % à 98 %.

- Les travaux de Jacquemin (1997) ont débouché sur la création d'un logiciel nommé FASTER. Ce dernier peut être utilisé pour traiter des corpus en langue anglaise ou en langue française. L'auteur cherche à décrire les transformations possibles des groupes nominaux terminologiques. On peut regrouper les phénomènes de variation des termes sous trois grandes catégories : syntaxique, morphosyntaxique et sémantique. Dans le premier cas, la structure syntaxique de la réalisation textuelle du terme est totalement différente. De sont côté, la variation morphosyntaxique s'opère grâce aux règles de dérivation morphologique et à une modification de la structure syntaxique. Pour leur part, les cas de variations sémantiques sont moins courants, mais il s'agit principalement de remplacement d'un des éléments du terme par un hyperonyme ou un hyponyme. Les systèmes qui se fondent sur une analyse syntaxique complète et sur le regroupement en constituants, de même que ceux qui utilisent une analyse syntaxique de surface, ont le potentiel d'identifier ces variations non décrites préalablement.
- Les travaux de Bourigault & Fabre (2000) : pour la création de système SYNTEX, par rapport à LEXTER, la couverture de SYNTEX est bien plus large, puisque l'analyse vise à repérer les relations de dépendance syntaxique autour d'un verbe et le principe de l'apprentissage endogène a été repris et largement étendu. Il effectue une analyse syntaxique en dépendance des phrases du corpus, préalablement étiquetées par un

analyseur morpho-syntaxique. Les principale relations de dépendance sont : sujet de verbe, objet de verbe, complément prépositionnel de verbe ou nom ou adjectif et épithète de nom.

3.2.3.2.2 Conclusion

Même si les approches linguistiques permettent l'obtention de bons résultats, l'intérêt de ces dernières est pondéré par la dilution de l'information causée par un fort taux de bruit. Sur le plan de la reconnaissance des unités complexes, les principaux problèmes proviennent du fait que l'analyse syntaxique fait très rarement appel à la sémantique. Le couplage d'un module sémantique afin d'augmenter les performances de cette approche ne peut être une solution économiquement envisageable pour l'élaboration d'un système de reconnaissance des unités terminologiques complexes. En effet, la confection de dictionnaires électroniques ou la réutilisation de l'information sémantique «cachée» dans les grandes banques de terminologie sont encore trop coûteuses, en temps et en efforts, pour être facilement intégrées dans le cadre d'une démarche flexible.

3.2.3.3 Modèles statistiques

L'approche statistique offre, des avantages indéniables puisqu'elle permet de s'attaquer à des ensembles de données d'une taille imposante qu'il serait tout à fait impensable de traiter manuellement. Elle permet aussi de traiter des ensembles textuels pour lesquels des dictionnaires électroniques n'ont pas été élaborés en vue d'un traitement linguistique.

3.2.3.3.1 Exemples de méthode

- Church et Hanks (1989) font office de pionniers dans le domaine du traitement statistique des données linguistiques. Leurs travaux ont pour but de repérer automatiquement l'ensemble des collocations contenues dans un ensemble de données textuelles.

  Church et Hanks présentent une mesure théorique, *l'information mutuelle*, qui rend possible l'évaluation du ratio d'association entre deux formes contenues dans un corpus. Si ces dernières, x et y, ont des probabilités d'occurrence P(x) et P(y), alors que L'information mutuelle tente de comparer la probabilité d'observer x et y ensemble par rapport à leur probabilité d'occurrence indépendante. Ces travaux sont devenues le point de départ de nombreuses recherches en acquisition automatique de la terminologie. Même si leurs travaux ont été effectués sur la langue anglaise, l'information mutuelle est purement statistique et elle est indépendante des langues. À l'analyse de leurs

résultats, les auteurs constatent que les formes dont les valeurs de l'IM sont plus élevées sont très intéressantes, alors que les couples dont les valeurs de l'IM tendent vers 0 sont moins intéressants pour le lexicographe.

- Les travaux de Enguehard et al. (1992) proposent le logiciel ANA, élaboré dans le cadre de travaux sur l'indexation automatique de textes en langue française. ANA contient les éléments suivants : les connaissances procédurales, le Boostrap et les connaissances déclaratives. Malgré l'affirmation des auteurs sur l'absence de nécessité du système ANA d'avoir recours à des ressources extérieures au corpus analysé, la description du logiciel faite par les auteurs (Enguehard et al. 1992) souligne l'importance de l'utilisation de listes de mots (les connaissances déclaratives). Malgré l'absence de définition, d'information morphologique, etc. Ces listes correspondent à des dictionnaires électroniques.

- Les travaux d'Ahmad (1996), effectués sur la langue anglaise, ne visent pas l'identification automatique des termes, mais plutôt de mettre à la disposition des terminologues des listes de mots ainsi que des concordances utiles pour l'identification de la terminologie. Ce travail représente le premier à présenter une approche tirant profit de corpus non spécialisés afin d'isoler des particularités lexicales dans des corpus spécialisés.

3.2.3.2 Conclusion

Si on les oppose aux techniques mécaniques et linguistiques, les techniques statistiques sont plus rapides, car elles permettent de cibler les sous-ensembles de données intéressants. Les ressources logicielles et matérielles nécessaires sont aussi moins imposantes puisque de telles approches ne requièrent pas de recours à des données linguistiques extérieures au corpus. En effet, elles peuvent fort bien effectuer leur travail en l'absence de dictionnaires et de grammaires. Il s'agit d'un avantage indéniable, car ces dernières ressources sont bien souvent les plus coûteuses à élaborer puisqu'elles sont habituellement le fruit d'un travail manuel. La capacité des approches statistiques de travailler sans avoir recours à des connaissances linguistiques les rend indépendantes des domaines abordés dans les corpus. En effet, les techniques statistiques ne reposent que sur les corpus eux-mêmes. Malgré tous ces avantages, on note aussi des désavantages. Les résultats obtenus par les méthodes statistiques sont intimement reliés aux corpus utilisés et ne peuvent être interprétés en dehors de ce

contexte. On doit aussi s'assurer que les corpus analysés possèdent une taille suffisamment grande pour que les résultats soient significatifs.

3.2.3.4 Modèles hybrides

Les modèles hybrides sont, comme leur nom l'indique, une combinaison entre les modèles linguistiques et les modèles statistiques. En effet, certains auteurs préfèrent de commencer par l'analyse linguistique dont les résultats sont filtrés à l'aide d'analyse statistique alors que d'autres procèdent à l'inverse.

3.2.3.4.1 Exemples de méthode

- Daille (1993) s'intéresse uniquement à l'acquisition automatique des termes complexes et elle n'aborde pas la problématique de l'acquisition des termes simples. le corpus analysé est préalablement étiqueté. La seconde étape de traitement fait appel à une description des groupes nominaux à l'aide de matrices syntagmatiques. Ces matrices sont identifiées dans les corpus à l'aide d'une grammaire. Les candidats termes trouver seront soumis à des tests statistiques pour affiner les résultats de l'analyse linguistique. C'est à partir des résultats de ces tests que seront éliminés ou conservés les candidats termes. Ces travaux constituent un premier pas vers une intégration des statistiques aux techniques linguistiques.
- Justeson et Katz (1993) ont élaboré le logiciel TERMS qui a pour objet l'acquisition automatique des termes en langue anglaise. Les auteurs concentrent leurs efforts sur l'acquisition des termes complexes et laissent entièrement de côté la problématique des unités simples. l'acquisition automatique des termes repose sur deux grandes contraintes. La première contrainte que doivent satisfaire les CT[2] est un seuil minimal de fréquence supérieur ou égal à deux. La seconde contrainte imposée lors de l'identification automatique des termes en est une de conformité à des matrices syntagmatiques. Le processus d'acquisition des termes est effectué en deux grandes étapes. La première consiste en l'attribution de catégories grammaticales aux formes d'un corpus à l'aide de dictionnaires électroniques. Lorsqu'une forme est soit un substantif, soit un adjectif, soit une préposition, elle est alors conservée pour l'étape suivante. Le logiciel relève ensuite les enchaînements qui sont conformes à la grammaire présentée précédemment, dans la mesure où ils satisfont la contrainte de

[2] CT : candidat terme

fréquence. Selon [Drouin 2002], les performances du logiciel diminuent avec l'augmentation de la taille du corpus analysé. Une explication potentielle de cette contre-performance réside dans le fait que le nombre de CT dont la fréquence est égale à 1 augmente avec la taille du corpus et que ces CT posent des difficultés à l'algorithme présenté par les auteurs.

- Smadja (1993) conçu le système XTRACT dans le domaine du repérage d'information et de l'indexation automatique en langue anglaise. Le système XTRACT exploite une approche hybride qui combine des techniques statistiques et linguistiques. L'application des techniques statistiques précède celle des techniques linguistiques. XTRACT fonctionne en trois grandes étapes : extraction de couples de mots (Bigrammes) présentant une information mutuelle importante selon la technique de Church et Hanks (1989); analyse contextuelle des Bigrams pour le repérage d'enchaînements plus longs (n-grams) et, finalement, filtrage des collocations obtenues aux étapes précédentes à l'aide d'information syntaxique.
- Lauer (1994) procède à une analyse qui a pour but de repérer les syntagmes nominaux de type N1 N2 N3 en langue anglaise. L'auteur propose aussi une approche pour la désambiguïsation automatique de ces structures. La première étape de traitement consiste en un étiquetage grammatical du corpus effectué à l'aide de dictionnaires. L'algorithme examine chacune des formes délimitées par des blancs typographiques. Lorsque les dictionnaires permettent de l'identifier de façon catégorique comme étant une unité nominale, elle est étiquetée et retenue pour la prochaine étape du traitement. Un calcul statistique évalue ensuite le degré d'association conceptuelle entre les éléments du segment de texte retenu. Afin d'évaluer cette association, le logiciel a recours à un thésaurus. Une analyse syntaxique prend ensuite en considération le poids de l'association conceptuelle entre les divers éléments afin de construire un arbre syntaxique et de trancher en cas d'ambiguïté. Ainsi, si l'association des deux premiers éléments reçoit une valeur plus grande que l'association des deux derniers, l'arbre de ce segment composé de trois éléments sera [[N1 N2] N3]. Dans le cas contraire, l'algorithme propose l'arbre [N1 [N2 N3]].
- Frantzi et Ananiadou (1997) présentent une technique visant l'acquisition automatique des termes en anglais. La démarche de ces auteurs ne vise cependant pas uniquement l'acquisition des termes, mais aussi l'élaboration d'un indice permettant de cerner le caractère terminologique d'un CT. L'acquisition automatique des termes s'effectue à l'aide d'une grammaire. Cette dernière identifie les CT qui

correspondent à des matrices de formation syntagmatiques dans un corpus qui a préalablement été étiqueté. Pour ce faire, les auteures ont recours à l'étiqueteur d'Éric Brill (1994, 1995). Les séquences retenues par l'algorithme de Frantzi et Ananiadou (1997 : 1) correspondent à la grammaire suivante : $(Nom|Adjectif)^{+}$ Nom

3.2.3.4.2 Conclusion

Les approches hybrides constituent un compromis entre les deux grandes tendances de base et s'en approprient donc les avantages et les inconvénients. En effet, leur puissance de traitement, reposant principalement sur l'adoption de modèles traitant de l'information sous forme numérique plutôt que linguistique, permet de s'attaquer plus facilement à des corpus de taille imposante. Cette caractéristique les sert bien puisque, de façon à obtenir des résultats de qualité et à minimiser le niveau de bruit obtenu, ces algorithmes doivent avoir accès à un volume de données important.

L'approche hybride exploite aussi la systématicité, la rapidité et l'indépendance par rapport au domaine des algorithmes statistiques. Cette indépendance se manifeste aussi par l'absence de besoin de dictionnaires et de grammaires spécialisées. Il s'agit là d'un avantage indéniable étant donné que ces techniques ont généralement pour but d'assister l'humain dans l'élaboration de dictionnaires.

3.3 Normalisation sémantique

L'étape précédente nous a fournit des candidats-termes **[Bourigault, 1994]** dont les libellés ont un sens pour le lecteur, souvent spécialiste du domaine. Mais rien n'assure que ce sens soit unique : au contraire, nous sommes dans un fonctionnement linguistique où les significations sont ambiguës, les définitions circulaires et dépendant en particulier du contexte interprétatif des locuteurs. Or, dans la modélisation ontologique, on cherche à construire des primitives dont le sens ne dépend pas des autres primitives et est surtout non contextuel. Il est nécessaire, pour prendre le chemin du formel, de *normaliser* les significations des termes pour ne retenir, pour chacun d'eux, qu'une seule signification, qu'une seule interprétation possible par un être humaine.

C'est ce que propose la *sémantique différentielle* de B. BACHIMONT.

Les résultats de l'extraction terminologique permettent de poser des relations entre des unités linguistiques fonctionnant comme des termes selon la morphologie et la syntaxe. Ces relations sont par conséquent intralinguistiques. Pour entreprendre une modélisation

sémantique, il paraît donc idoine de retenir une sémantique reposant sur ce type de relation. La sémantique différentielle est une telle sémantique : elle permet d'expliciter le sens (en contexte) d'une unité linguistique ou sa signification (hors contexte, quand le linguiste veut la considérer pour elle-même) par d'autres termes de la langue, sans faire appel à des notions ou entités extralinguistiques. Dans cette perspective, les unités linguistiques ont du sens dans la mesure où leur usage les distingue les unes des autres, si bien que le sens d'une unité se constitue des identités et des différences qu'elle entretient avec les autres unités de la langue. Nous adoptons cette sémantique et proposons de caractériser chaque unité par les unités avec lesquelles elle est identique sous certains aspects et celles avec lesquelles elle est différente sous d'autres [Bachimont, 2000].

On dégage ainsi un ensemble de termes en construisant un système de différences entre ces termes. Dans ce contexte, la structure construite est, comme pour Aristote un arbre. Cette structure arborescente est une affirmation forte de cette méthodologie et les arguments n'ont rien à voir avec une implémentation dans un langage informatique ou un autre qui ne permettrait pas d'exprimer des héritages multiples comme dans les treillis : on est simplement en train de préparer l'élaboration des primitives non formelles d'une théorie formelle et on ne peut se permettre que leur signification soit « instable», héritée soit d'un *père* soit d'un autre dans le treillis. Enfin, les principes de construction de cet arbre découlent directement des fondements épistémologiques de cette méthodologie : sans les détailler, ils reprennent les principes aristotéliciens de genre proche et différence spécifique additionnés de principes liés au paradigme différentiel choisi [Charlet 2003].

Construction et lecture de l'arbre ontologique sont duales et l'arbre de libellés linguistiques ainsi construit pourra être lu et interprété de façon unique.

Les principes différentiels constituent donc une grille de lecture et sont des prescriptions interprétatives qu'il faut suivre pour savoir comment interpréter le libellé. C'est donc le respect de ces principes qui permet de considérer ce libellé non pas comme une unité linguistique dont le sens varie selon le contexte de son utilisation, mais comme une primitive au sens invariable. [. . .] On obtient un réseau dans lequel la position d'un nœud conditionne sa signification. La signification définie par la position dans l'arbre est invariable selon les contextes. [. . .] En respectant les principes différentiels, en s'engageant à suivre la sémantique qu'ils prescrivent, les nœuds de l'arbre ontologique correspondent à des concepts pouvant être utilisés comme des primitives de modélisation et de formalisation. Nous venons donc de définir l'engagement sémantique à la base de l'ontologie : ensemble des

prescriptions interprétatives qu'il faut respecter pour que le libellé fonctionne comme une primitive [Bachimont, 2000].

À la fin de cette étape, l'ontologie construite n'est pas formelle : c'est un arbre de signifiés linguistiques normés – ou concepts linguistiques – par les principes différentiels appliqués. Ayant fixé le contexte d'interprétations de ces signifiés linguistiques, on a fabriqué une ontologie qui n'est valable que pour un contexte particulier, c'est-à-dire localement, c'est une *ontologie régionale* [Bachimont, 2000].

3.4 L'engagement ontologique

À cette étape, nous avons un arbre de primitives qui vont pouvoir être modélisées de façon formelle en définissant une sémantique formelle. Celle-ci va permettre de créer des concepts formels à partir et par opposition aux concepts linguistiques. Cette sémantique ne considère plus des notions sémantiques mais des extensions, c'est-à-dire l'ensemble des objets qui vérifient des propriétés définies en intension dans l'étape précédente, propriétés ayant une définition formelle à ce niveau. La structure de l'ontologie formelle doit alors être abordée.

Premièrement, les concepts formels vérifient les relations d'identité unissant les concepts sémantiques. En effet, les concepts sémantiques s'interdéfinissent par identités et différences. L'identité correspond au fait qu'une notion est comprise dans une autre : la notion de bistouri comprend celle d'instrument d'incision. Par conséquent, tout objet qui est un bistouri est un instrument d'incision. [. . .] Deuxièmement, la différence entre concepts sémantiques, où deux notions s'excluent mutuellement, ne se répercute pas directement sur les concepts formels. Prenons un exemple avec les notions d'acteur et d'être humain. Ces notions s'excluent : la notion d'acteur correspond à un rôle, celle d'être humain à une entité biologique. Or, une entité biologique n'est pas un rôle [. . .]. Mais, dès lors que l'on adopte une sémantique formelle, on ne considère plus des notions, mais des extensions d'objet. Par conséquent, un acteur, c'est l'ensemble des objets qui sont des acteurs, c'est-à-dire qui jouent un rôle. Les êtres humains sont également un ensemble d'objets. [. . .] à l'évidence, un objet qui est un homme peut être un objet qui est un acteur, même si la notion d'acteur n'est pas celle d'homme. Cela implique que l'on retrouve dans les relations unissant les concepts formels les relations d'héritage des concepts sémantiques, mais pas les exclusions. La structure des concepts formels n'est plus obligatoirement un arbre, mais, plus généralement, une structure de treillis. Cela se comprend d'ailleurs facilement pour la raison suivante : si la sémantique des concepts est référentielle, les relations entre les concepts sont des relations

entre ensembles. La structure des concepts formels doit correspondre à la structure algébrique des ensembles, c'est-à-dire un treillis [Bachimont, 2000].

Au sein de cette ontologie formelle, le treillis des concepts doit être compris comme la possibilité de créer des concepts dits définis en combinant les concepts primitifs : par exemple, une *personne* qui a pour rôle social d'être un *médecin*, cet « objet » défini en extension héritant des caractéristiques des personnes et des médecins alors qu'au niveau précédent, l'intension des personnes et des médecins étaient irréductibles.

Pour finir, il est important de noter que nous sommes bien dans le contexte de *l'engagement ontologique* de N. Guarino : la formalisation proposée est une spécification formelle donc extensionnelle de l'ontologie ainsi définie et le sens des concepts est dans les objets définis en extension. Comme rappelé précédemment les contraintes de cette formalisation font qu'elle ne peut rendre compte exactement du sens visé au niveau sémantique, celui de l'ontologie régionale, ce qui justifie qu'elle ne puisse en rendre compte que « partiellement».

3.5 L'opérationnalisation

Dernière étape de la méthodologie, l'opérationnalisation consiste en la représentation de l'ontologie dans un langage de représentation des connaissances permettant de surcroît des services inférentiels de type classification des concepts ou généralisation, etc. Selon les langages considérés, les calculs possibles et donc les services inférentiels ne sont pas identiques et, à ce niveau aussi, il y a un engagement qui est pris avec de nouvelles contraintes et possibilités, justifiant l'existence d'une *ontologie computationnelle*.

Les langages de représentation les plus connus sont, comme noté auparavant, les graphes conceptuels et les logiques de description (*cf.* chap. 1, § 8), permettant, l'un comme l'autre, d'effectuer un certain nombre d'opérations sur des ontologies : inférences propres aux structures de graphes comme la jointure ou la projection pour les graphes conceptuels, classifications dans des structures arborescentes pour les logiques de description. Ceci étant, même s'ils ne percent pas réellement, d'autres langages permettent d'intéressants débats sur les niveaux de représentation et tentent de satisfaire, en navigant entre les deux extrêmes, les contraintes d'expression d'une ontologie régionale, formelle et computationnelle et les nécessaires inférences à faire dessus.

3.6 Les relations

Arrivés à la fin de la méthodologie, il reste à définir plus précisément les relations dans l'ontologie.

[...] elles ne se définissent pas de la même manière que les concepts, car, unissant des concepts, elles se caractérisent à partir d'eux. Si l'on ne retient que des relations binaires, les relations se définissent de la manière suivante : (1) Une relation se définit par les concepts qu'elle relie : par exemple, être animé et action ; ces concepts constituent la signature sémantique de la relation. (2) Une relation se définit en outre par un contenu sémantique intrinsèque articulant les deux concepts : par exemple, le fait que l'être animé est l'agent de l'action. La sémantique intrinsèque de la relation est spécifiée vis-à-vis des autres relations possédant la même signature sémantique selon les principes différentiels vus plus haut. Par exemple, la relation patient entre être animé et action se définit par identité et différence avec la relation agent. L'identité, c'est le fait d'avoir la même signature, la différence, c'est le fait de subir l'action plutôt que de l'exercer. L'identité n'est pas forcément réduite au fait d'avoir la même signature : par exemple, la relation agent volontaire et agent involontaire possède comme identité, outre la même signature, le fait d'avoir un père commun, la relation agent. Autrement dit, chaque signature sémantique est potentiellement la racine d'un arbre différentiel de relations possédant la même signature et spécifiées selon les principes différentiels. Les signatures sémantiques constituent également un arbre : on a donc un arbre de relations venant compléter l'arbre des concepts [Bachimont, 2000].

Avec les relations, on complète les briques conceptuelles – non formelles – nous permettant de construire toutes les représentations formelles valides du domaine. Il faut remarquer que les concepts et relations de l'ontologies sont dual l'un par rapport à l'autre. Un concept primitif pourrait être un concept défini, une relation pourrait se retrouver implicitement définie au sein d'un concept primitif. Ce sont les choix de la deuxième étape qui auront permis de décider de ce qui est essentiel – et donc primitif – ou non.

L'ensemble des étapes et la nature des objets élaborés est résumé dans les 4 points :

- **Analyse de corpus**. description linguistique du corpus caractérisant le domaine ; reflète les normes sociales, techniques, pratiques. . .
  Corpus ➔ Signifié
- **Normalisation sémantique** du signifié linguistique pour dégager les objets du domaine et leur type
  Signifié ➔ Signifié normé (ou Concept linguistique)
- **Engagement ontologique**
  Concept linguistique ➔ Concept formel
- **Opérationnalisation** dans un langage de représentation des connaissances

3.7 Quelques bons principes

Passés les questions de méthodologies fondamentales, un certain nombre de travaux proposent des principes de construction d'ontologies. D'après [Charlet 2003] R. Gruber (1993) et M. Fernàndez *et al.* (1999) proposent un certain nombre de principes à respecter pour construire une ontologie :

- **Clarté.** Les ambiguïtés doivent être réduites, quand une définition peut être axiomatisée, elle doit l'être. Dans tous les cas, des définitions en langage naturel doivent être fournies.
- **Cohérence.** Une ontologie doit être cohérente. Les axiomes doivent être consistants. La cohérence des définitions en langage naturelle doit être vérifiée autant que faire se peut.
- **Extensibilité.** L'ontologie doit être construite de telle manière que l'on puisse l'étendre facilement, sans remettre en cause ce qui a déjà été fait.
- **Biais d'encodage minimal.** L'ontologie doit être conceptualisée indépendamment de tout langage d'implémentation. Le but étant de permettre le partage des connaissances (de l'ontologie) entre différentes applications utilisant des langages de représentation différents.
- **Engagement ontologique minimal.** Une ontologie doit faire un minimum d'hypothèses sur le monde : elle doit contenir un vocabulaire partagé mais ne doit pas être une base de connaissances comportant des connaissances supplémentaires sur le monde à modéliser.

Toujours, d'après l'auteur [Charlet 2003] d'autres principes du même type sont proposés par d'autres auteurs (voir l'article de A. Gómez-Pérez (2000)).

4. LE PROJET TERMINAE

Nous proposons ici une méthode linguistiquement fondée, cette méthode est développée depuis 1997 au sein du laboratoire d'informatique de Paris Nord (Université Paris 13) par B. Biébow et S. Szulman.

La motivation de ce projet est venue d'un constat : la plupart des objets utilisés pour modéliser un domaine sont dénotés par des termes et son « repérables » à l'intérieur des textes portant sur le domaine. Aussi il est possible d'avoir, dans le cadre d'une modélisation d'un domaine, une approche linguistique, terminologique, pour élaborer des ontologies.

Biébow et Szulman proposent une telle approche terminologique, pour construire une ontologie d'un domaine. Cette approche mène à une typologie de concepts qui permet d'améliorer la lisibilité de l'ontologie et sa maintenance. Cette typologie catégorise les concepts selon la manière avec laquelle ils ont été élaborés. Donc, TERMINAE est outil d'aide à la construction d'ontologie à partir de texte en suivant quatre principales étapes :

- *Constitution d'un corpus* (documents techniques, comptes rendus, livres de cours, etc.), à partir d'une analyse des besoins de l'application visée,

- *Étude linguistique*, pour identifier des termes et des relations lexicales, en utilisant des outils de traitement de la langue naturelle (extracteurs de termes, outils d'analyse distributionnelle, outils d'aide au repérage de relations par des patrons linguistiques, etc),

- *Normalisation sémantique*, conduisant à des concepts et des relations sémantiques définis dans un langage semi-formel. La structuration des concepts s'appuie sur les résultats du dépouillement des textes tout en tenant compte de l'objectif d'utilisation de l'ontologie.

- *Formalisation en logique* de description et intégration des concepts au sein d'une ontologie formelle.

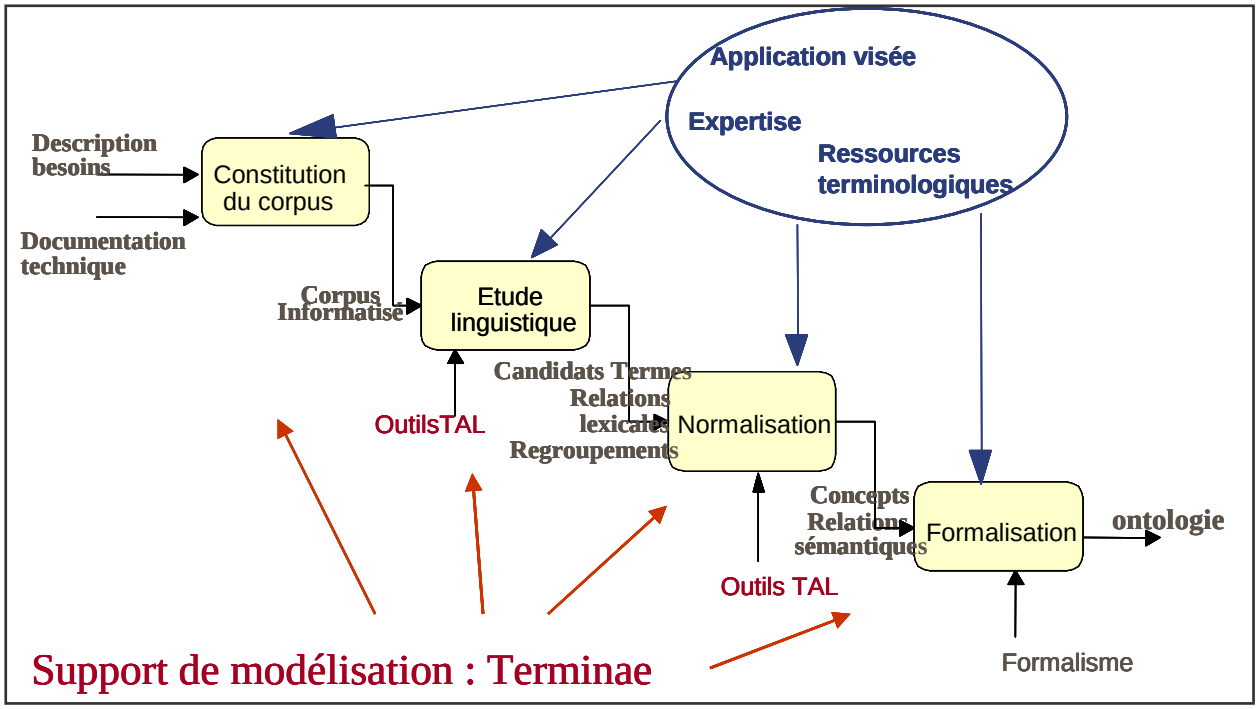


Fig 5. Le projet Terminae

Le logiciel TERMINAE associé à la méthode fournit des aides pour toutes les étapes de l'analyse des textes à la formalisation. Plus qu'un éditeur, il constitue un support méthodologique qui permet d'évoluer progressivement et en conservant des liens du niveau linguistique au niveau conceptuel puis à la représentation formelle.

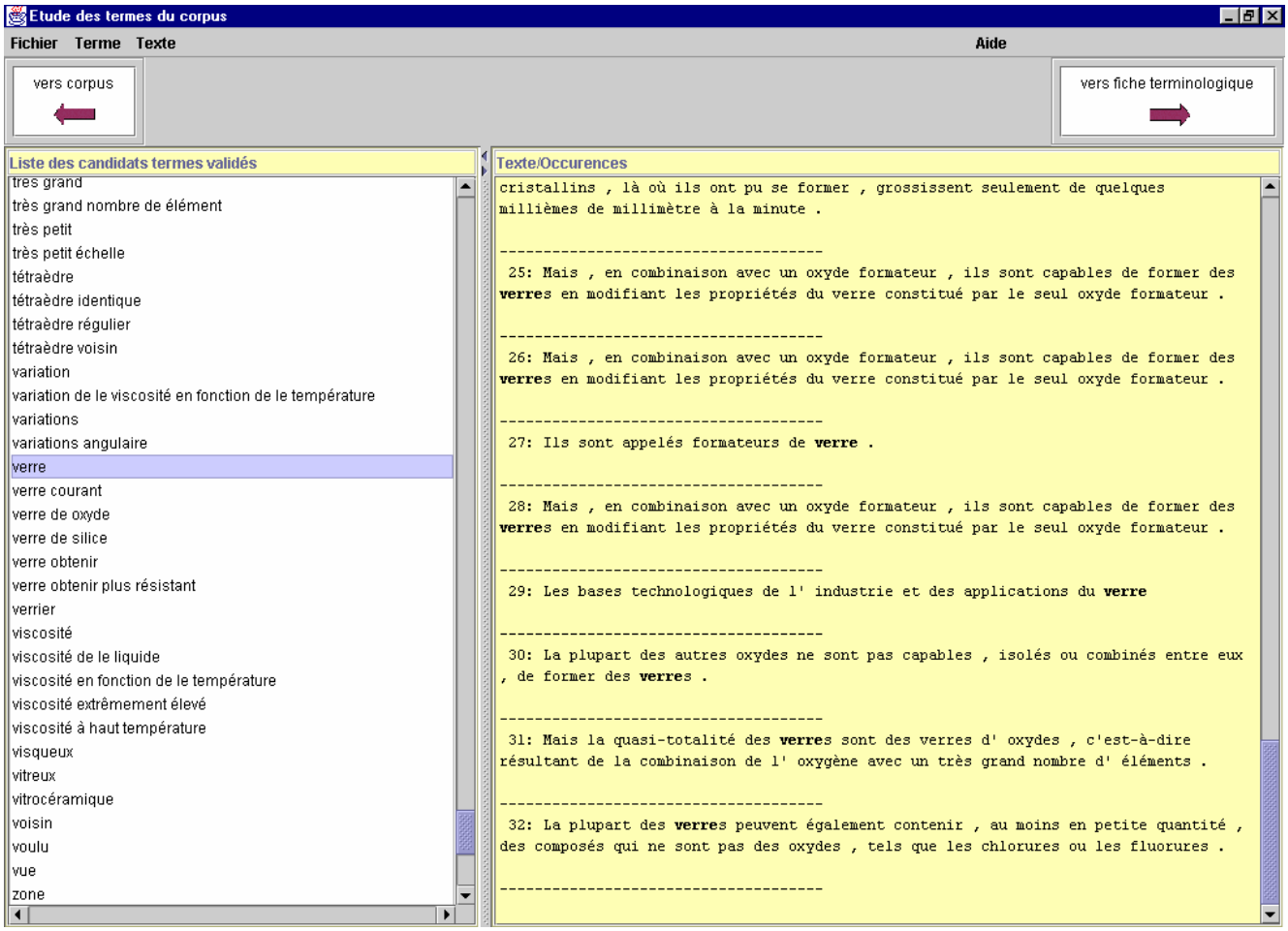
Cette approche a été expérimentée pour décrire une ontologie des outils d'ingénierie des connaissances dans le cadre d'un projet, Th(IC)2, mené au sein du groupe TIA. Les documents sources de connaissances étaient des articles scientifiques de ce domaine de recherche ainsi que des présentations courtes de travaux de laboratoires.

TERMINAE : un logiciel et une méthode pour construire des ontologies à partir de textes [Aussenac-Gilles 2000].

La méthode associée à TERMINAE, s'appuie sur des principes linguistiques pour repérer des concepts à partir de l'étude de leurs traces lexicales. Il s'agit d'observer l'usage des mots dans des documents. Les fonctionnalités de TERMINAE en font un support méthodologique plus qu'un éditeur de terminologie. Son interface permet de conduire le dépouillement de données tirées de textes, de mener une analyse terminologique et de modéliser un réseau conceptuel avant de formaliser l'ontologie ainsi spécifiée.

TERMINAE : (1) repérage de connaissances dans des textes

L'utilisateur peut accéder au contenu des textes via les résultats de logiciels d'extraction de termes (comme LEXTER [Assadi & Bourigault, 00] ou SYNTEX [Bourigault, 2002]).

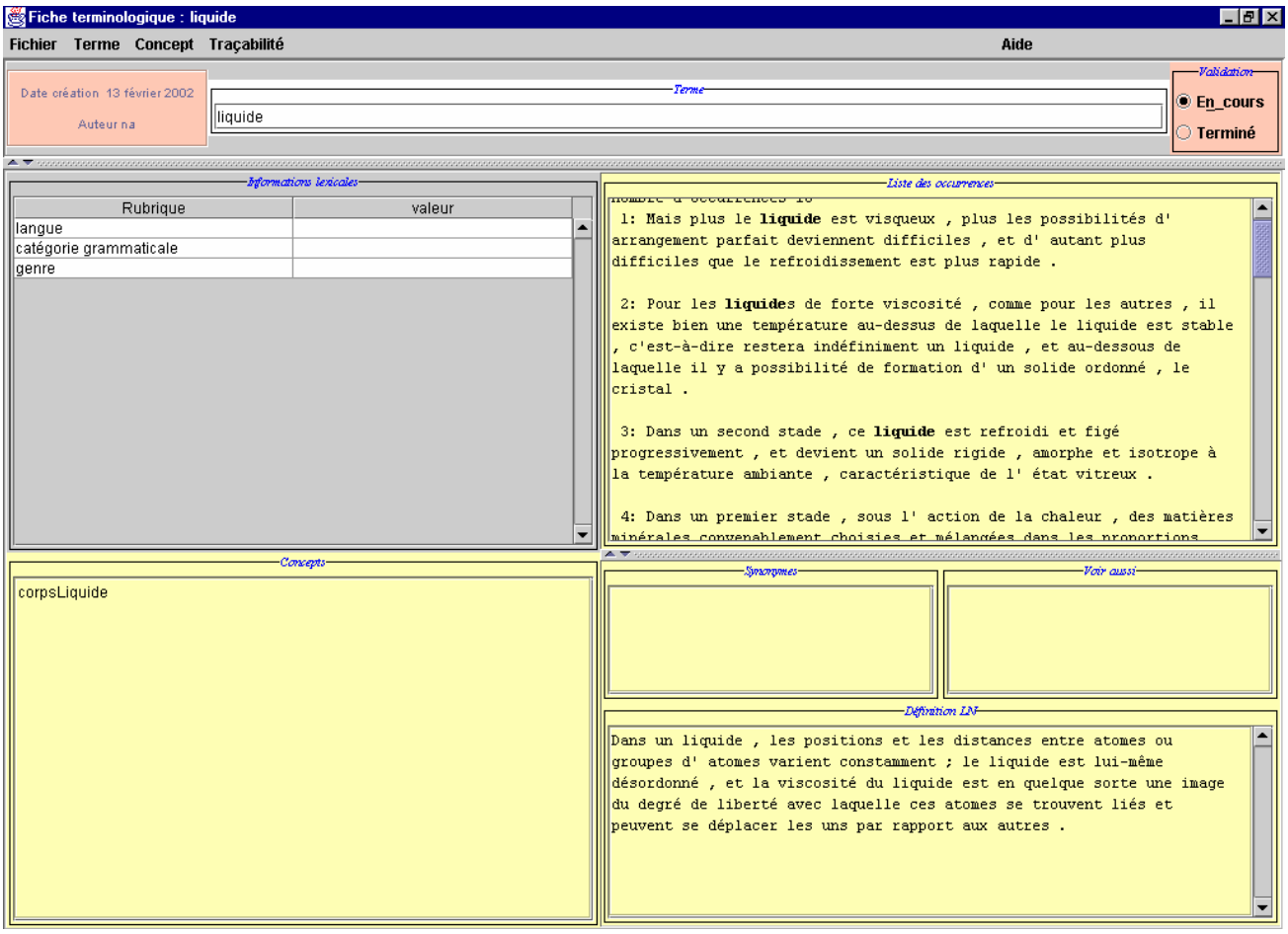


La figure ci-dessus illustre un type de dépouillement possible : cette interface permet de lister des mots et groupes de mots issus des analyses du corpus par LEXTER. Pour un mot donné, on peut consulter toutes ses occurrences puis décider de définir un terme (fiche terminologique).

TERMINAE : (2) analyse terminologique

L'utilisateur peut enregistrer et structurer des données terminologiques (termes retenus, leurs synonymes ou équivalents, leurs concepts associés et leurs occurrences) sous forme de fiches terminologiques.

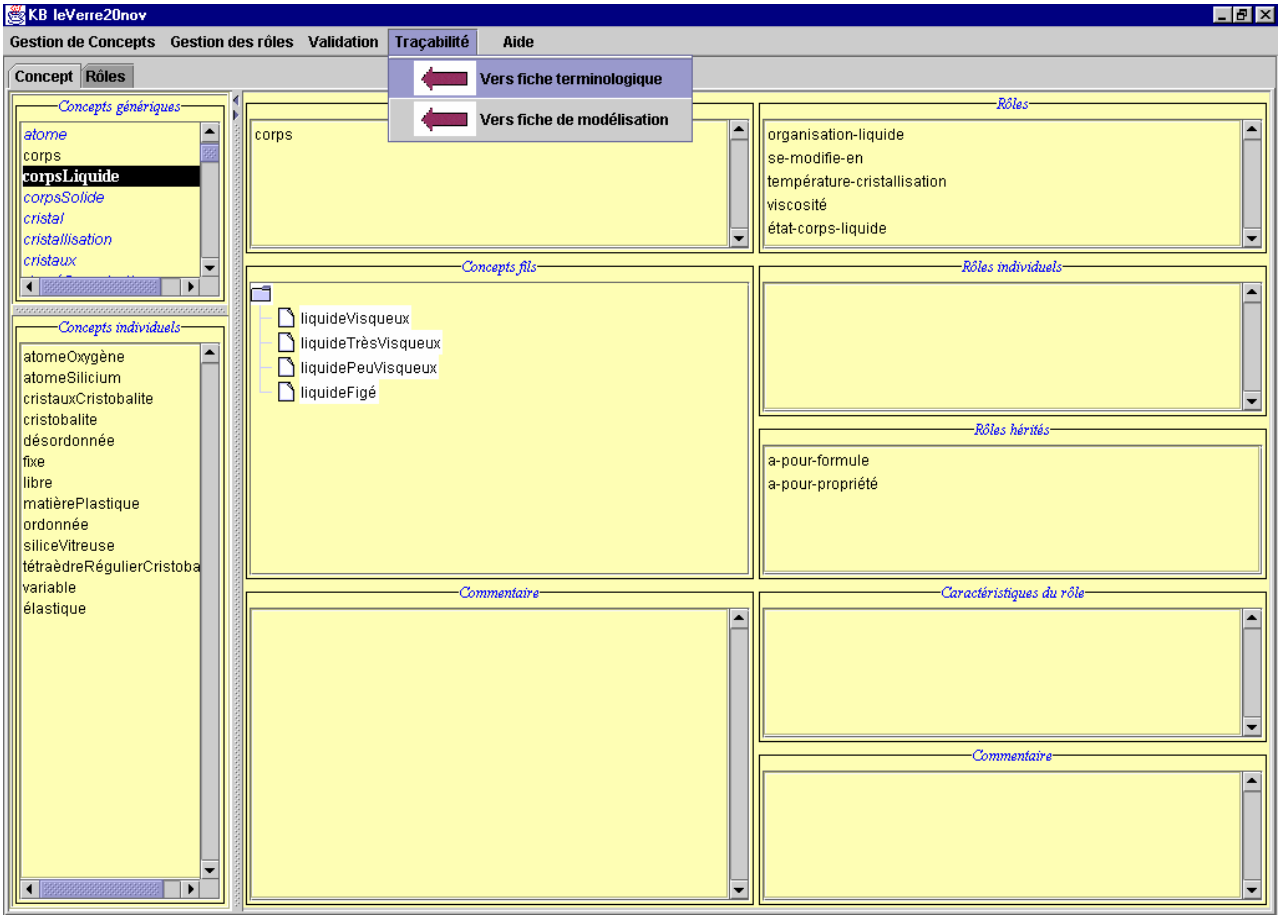
La figure ci-dessous présente une fiche pour le terme liquide, à partir duquel le concept CorpsLiquide a été défini. Les occurrences du terme sont visibles dans la partie droite. Un sous-ensemble de ces occurrences, jugées plus pertinentes pour ce terme, peut être sélectionné et associé au concept. Plusieurs concepts peuvent être associés à un terme s'il a plusieurs sens. Pour chacun de ces sens, des termes synonymes peuvent être indiqués.



TERMINAE : (3) structuration conceptuelle, normalisation

L'utilisateur peut ensuite définir des concepts (des classes), des instances de concepts (objets particuliers) et d'établir des relations entre eux (au moyen de rôles typés). Les classes sont organisées dans une hiérarchie et décrites à l'aide de fiches conceptuelles. Les rôles sont hérités ou non d'une classe vers ses sous-classes.

Sur la figure ci-dessous, on se focalise sur le concept corpsLiquide, dont le concept père est corps, dont les fils sont présentés au centre. Les rôles de ce concept sont affichés à droite en distinguant les rôles propres au concept (en haut) des rôles hérités de ses pères (a-pour-formule et a-pour-propriété).

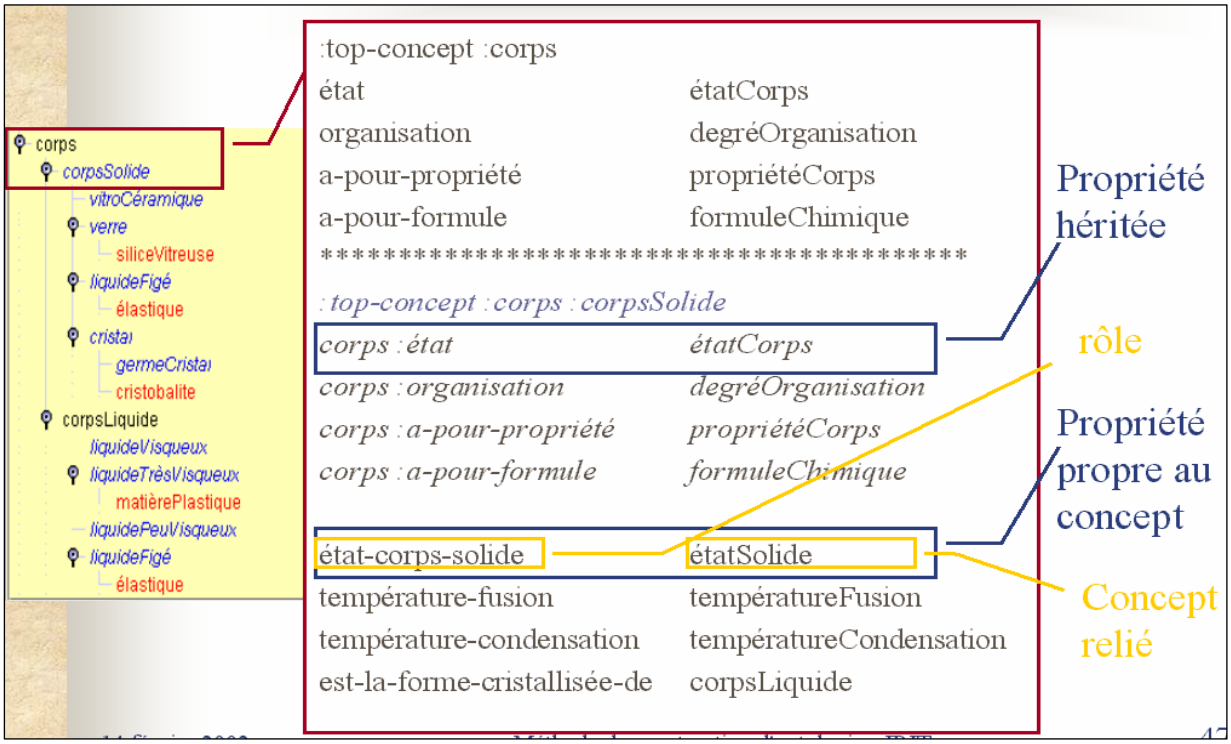


Cette fenêtre permet de retourner vers le niveau des fiches terminologiques : on peut revenir aux termes désignant le concept en cours (ici CorpsLiquide).

TERMINAE : (4) formalisation

La formalisation permet de vérifier la bonne définition syntaxique des concepts et la cohérence du modèle .

La figure qui suit présente une vue détaillée de la représentation d'un concept en fin de modélisation. C'est cette représentation qui est ensuite formalisée. TERMINAE permet de générer une ontologie dans des langages standards et compatibles XML : OIL et RDFs



5. CONCLUSION

La motivation de la construction d’ontologie à partir des textes représente une alternative au recours aux entretiens menés auprès d’experts. Outre un gain de temps et d’efficacité, l’exploitation des textes présente plusieurs avantages pour la construction des ontologies : Les textes peuvent garantir une meilleure utilisabilité de l’ontologie, les ontologies construites à partir de textes peuvent être plus riches ; elles peuvent rendre compte des différents termes associés aux concepts et de leurs usages ; aussi, elles peuvent maintenir un lien vers les textes pour justifier les choix de modélisation.

De ce fait, ces ontologies conviennent particulièrement pour des applications supposant une interprétation humaine, pour des applications relatives à des documents, comme la recherche d’information ou le classement de documents, etc.

Chapitre 3

Notre méthode de la construction d'ontologie

1. INTRODUCTION

Nous présentons dans cette partie notre méthode de la construction d'ontologie à partir de textes techniques. En effet, un texte technique sert, d'une part comme source pour la construction d'ontologie et fait l'objet, d'autre part, de l'application finale (la réponse sur la requête de l'utilisateur). Notre méthode s'applique directement à des corpus de textes, sans que ceux-ci nécessitent un étiquetage préalable, et sans utilisation de gros dictionnaires de langue ou du domaine. En effet, l'analyse distributionnelle que nous appliquons aux corpus ne nécessite pas de connaître les catégories des mots. Nous appliquons notre ontologie dans le domaine de la recherche d'information, et plus précisément pour l'expansion de la requête utilisateur.

2. LES ASPECTS THEORIQUES DE NOTRE MODELES

Avant de détailler notre méthode d'acquisition terminologique et d'aide à la construction d'ontologie à partir de corpus, il convient d'abord de résumer les aspects théoriques sur lesquels est basé notre modèle la construction.
Notre modèle de la construction n'est pas une théorie linguistique mais il s'inspire, d'une part

des modèles linguistiques basés sur la *notion de répétition* dans les textes (Lebart & Salem 1994) et des *règles de déduction contextuelle* **[Vergne, 1999]** ainsi que des travaux de *l'analyse distributionnelle* de Z. Harris (1954, 1991), d'autre part de certaines propriétés et caractéristiques des langues de spécialité et l'analyse de donnée comme par exemple les méthodes de la *classification automatique* pour regrouper les concepts les plus proches dans une même classe. Il permet par exemple de classer des éléments du texte appartenant à une même catégorie linguistique en utilisant les propriétés de l'analyse distributionnelle.
Les aspects pratiques de notre modèle concernent les algorithmes conçus pour l'acquisition et la structuration des termes.

2.1 Le textes scientifiques ou techniques

Premièrement, la documentation technique sert, d'une part comme source pour la construction de l'ontologie et fait l'objet d'autre part, de l'application finale. Dans sa thèse [**Assadi, 1998**] illustre un type de démarche empirique que l'on peut avoir pour dégager les caractéristiques linguistiques pouvant intervenir comme paramètres pour une typologie des textes. Selon l'auteur (Copeck et Alii, 1997) s'intéressent à la notion de "*technicalité du texte*" (*text technicality*), plutôt qu'à celle de "*texte technique*" (notion binaire, un texte est technique ou ne l'est pas). Le but de leur démarche est d'exhiber des paramètres linguistiques décrivant les textes et ayant un pouvoir discriminant quant à la technicalité du texte.
Nous présentons ici quelques caractéristiques pertinentes pour évaluer la technicalité de texte, certain auteur (Copeck & al. 1997) selon toujours [**Assadi, 1998**] dégager plus de 32 critères linguistiques qui sont des paramètres pertinents pour évaluer la technicalité d'un texte. Même si ce travail est critiquable sous certains aspects, il a le mérite de dégager un ensemble de paramètres linguistiques que l'on peut observer assez facilement sur un texte et rende comte, à divers degrés, de la technicalité. Nous donnons des exemples dans le tableau ci-dessous.

Critères lexico- syntaxiques	Verbes au présent, nominalisation,…
Critères sémantiques	Terminologie spécifique, terme vague,…
Critères structuraux	Table des matières, index,…

Tableau : exemples de critères linguistiques dégagés par Copeck et alii.

Nous citons dans la partie de l'extraction de termes d'autre exemples (voir § 3.2.2), que nous les appuyons pour l'extraction des termes (syntagmes nominaux). Mais le point important dans les textes techniques, que la communication dans les textes est précise et objective : La phrase scientifique tend à être précise et objective et le sujet humain n'intervient que sous forme personnelle (absence d'expression liée à un sentiment, etc.). Cette précision de l'information scientifique a donné entre autre naissance aux unités lexicales complexes (par exemple les syntagmes nominaux, les noms composés, etc.).

Nous donnons ici notre définition de la documentation technique.

> ***Définition :*** *Nous appelons documentation technique un ensemble de documents relatifs à un domaine donnée comme instrument de travail pour une activité bien déterminée.*

2.2 L'analyse distributionnelle

Le but ici est l'étude de la distribution des unités linguistiques dans les corpus de textes afin de repérer les différents contextes linguistiques d'un mot. La distribution d'une unité linguistique peut être définie comme la somme des environnements où cette unité se rencontre dans le corpus.
Considérons les représentations formelles suivantes des premières phrases d'un corpus (les unités linguistiques sont représentées par les symboles Ti) :
T1 T2
T1 T4 T3
T5 T2 T3
T3 T2 T5
T1 T4 T5
T3 T4
T5 T4 T1
T3 T2 T1

D'après ces phrases la distribution de l'unité T2 par exemple sera :

On remarque également que l'unité T4 a la même distribution que l'unité T2. Par conséquent on peut supposer déjà que T2 et T4 font partie d'une même catégorie d'unités. Pour confirmer cette hypothèse il faut étudier la distribution de ces unités dans le corpus complet.

> ***Définition :*** *L'analyse distributionnelle permet donc de regrouper des unités distinctes dans une même catégorie d'unités.*

2.2.1 Les travaux de Harris

Un point de vue linguistique qui renforce notre modèle, qui se veut attaché à la notion de contextes des unités linguistiques du texte est celui de Harris. Harris (1951) a formalisé une méthode distributionnelle inspirée de l'étude des formes linguistiques proposée par son professeur Bloomfield (1933). Cette méthode qui est issue du courant *structuraliste américain* (la linguistique structurale refuse de poser le problème du sens; elle considère que la forme (syntaxe) est indépendante du sens et que la sémantique ne peut pas faire partie de la description linguistique), se caractérise par l'observation dans un corpus de formes linguistiques et de leur distribution. La notion de sens étant mal définie, Harris a donc eu recours à cette notion de distribution qui est une notion plus stable que la notion de sens. La distribution d'un élément est en effet l'ensemble des contextes de cet élément dans le corpus. Harris (1954) propose un premier fait distributionnel qui est la possibilité de segmenter toute chaîne du texte en parties de manière à découvrir certaines régularités d'occurrence de l'une de ces parties, relativement à d'autres parties de la chaîne. Dans notre cas on applique cette notion à des séquences plus grandes que celles des séquences de morphèmes étudiées par Harris, ce sont les syntagmes nominaux.

2.3 Un terme

Le rôle principal de la terminologie est d'une part, d'établir la liste et la définition des notions liées à des termes (termes simples, termes complexes) d'un domaine donné, d'autres parts d'établir les relations qui existent entre ces termes (hyperonymie, inclusion, etc.). Les termes et les relations sont souvent représentés soit sous forme arborescente [**Oueslati, 99**] soit sous forme de réseaux terminologiques [**Bourigault, 94**].
Une des caractéristiques du la langue scientifique et technique est que les termes qui le

composent sont souvent moins ambigus que dans la langue générale. Un avantage est qu'il est donc plus facile de faire la distinction entre un terme et les autres termes qui lui sont apparentés dans un ensemble terminologique. La terminologie distingue deux aspects du terme : un aspect linguistique et un aspect conceptuel.

2.3.1 Un aspect linguistique : L'aspect linguistique du terme consiste à considérer le terme comme une entité linguistique à part entière (Lerat 1995). Les termes servent souvent à la dénomination des notions appartenant à un ou plusieurs domaines spécialisés.

2.3.2 L'aspect sémantique et conceptuel du terme : En terminologie, le domaine est défini comme un ensemble de notions qui sont des représentations mentales; le but de la terminologie est donc d'établir et de représenter ce système de notions. Dans ses travaux sur la terminologie, [**Rastier 1995**], précise que "*c'est le travail terminologique qui transforme la notion en concept*". Le concept dans ce cas est défini comme une structure construite pour décrire une entité du domaine.
Pour bien marquer l'interdépendance entre termes et concepts, certains linguistes (Rastier 1990), (de Bessé 1990) se réfèrent couramment à un modèle triadique, décrit par le triangle sémiotique de Ogden et Richards :

signe
chose/objet
(terme (nom), symbole, signifiant)
(réalité, référent)
concept
(représentation mentale, sens, signifié)

Fig 6 : Le triangle sémiotique

Ce triangle, résume les relations suivantes : 1) le signe représente l'objet 2) le signe symbolise le concept 3) le concept se rapporte à l'objet. Les mots entre parenthèses sont des mots équivalents.

D'autre notion, nous présenterons dans la suite de ce chapitre.

3. NOTRE METHODE d'AIDE DE LA CONSTRUCTION D'ONTOLOGIE

Notre modèle proposé en dessous (*fig. 7*) est un modèle complet : à partir des textes représente comme étant un réservoir de connaissances de domaine jusqu'à la modélisation de domaine se forme d'une ontologie de domaine. En générale, les autres méthodes comme TERMINAE de [**Aussenac-Gilles, Biébow & Szulman, 2000**] et les travaux de Assadi et Bourigault s'appuient sur les résultats des outils qui sont leurs buts initiaux diffèrent de but de la construction d'ontologie. Le grand travail des chercheurs est de trouver des stratégies pour la combinaison entres les outils pour atteindre l'objectif de la construction. Cet objectif n'est pas simple à atteindre. En générale, pour chaque étape on utilise un ou plusieurs outils et le passage d'une autre étape provoque l'utilisation des résultats de l'étape précédente, un but qui n'est pas toujours facile. Le passage de résultat d'une étape à l'étape suivante se fait en générale manuellement; par conséquent, ce travail représente un grand gaspillage de temps pour des raisons très éloigner à l'objectif de la construction. Pour remédier ces problèmes nous proposons une méthode et par conséquent un outil dédier essentiellement à la construction d'ontologie. Le point de départ de cet outil est les textes ou le corpus jusqu'à la sortie qui représente l'ontologie de domaine.

Normalisation

Acquisition des termes

TEXTES b

Corpus

Candidat termes
Structuration, concepts et relations
ONTOLOGIE

Fig 7. Un modèle pour la construction d'ontologie conceptuelle

3.1 Constitution du corpus

À partir de description des besoins, on peut choisir des textes de façon à couvrir le domaine requis par l'application. Le choix est nécessite une bonne connaissances de domaine. Le choix des textes du corpus doit s'appuyer sur les informations données par les experts du domaine. Dans notre cas où l'application visée par notre projet est de regroupé les supports de cours des enseignants informatique dans un site. La public visé sera en premier temps les étudiants en informatique. Donc les experts de domaine sont pour notre cas les enseignants de l'informatique. Ces sont qui choisir les documents les plus appropries. Le fait de considérer le corpus comme une source privilégiée de connaissances lui donne un poids énorme sur le contenu du modèle obtenu. Le modèle sera finalement un reflet du corpus. Un mauvais choix du corpus conduira à un modèle inadapté ou incomplet. Donc, la qualité d'ontologie dépend de la qualité de corpus choisi.
La constitution du corpus suppose plusieurs tâches étroitement liées et effectuées de manière cyclique jusqu'à parvenir à un état stable et satisfaisant du corpus :

- choisir des documents représentatifs du domaine étudié et/ou adaptés à l'application ciblée ;

- les mettre au format informatique adéquat ;
- décider de la manière de les traiter ;
- évaluer ces documents, leur qualité et leur apport potentiel au modèle à construire.

Le corpus est ensuite préparé pour être traité informatiquement. Nous proposons de transformer tous les textes dans les différents formats (Word, PS, TXT, etc.) en format unique, codé par exemple en ASCII.
Les données de départ de cette phase sont constituées par des documentations techniques : soient sous forme de papier ou sous forme électronique. Le but est de constitution un corpus du domaine.

Données en entrée	Documents sous forme papier ou électronique (DOC, HTM..)
Données en sortie	Corpus informatisé (ASCII)
Intervenants	Cogniticien + expert

2.2 Extraction de termes

Nous proposons ici une méthode d'aide à l'acquisition de termes à partir de corpus de textes basée sur une approche linguistique quantitative. Leur but est essentiellement de relever des segments de texte qui se répètent à l'intérieur d'un corpus. Cette approche privilégie l'étude des séquences récurrentes de mots dans des textes d'un domaine spécialisé susceptibles d'exprimer des termes de ce domaine. Ces termes sont souvent des syntagmes nominaux qui servent à dénoter des concepts du domaine de manière la moins ambigu possible. Notre approche traite directement des documents textes sans utiliser aucune analyse syntaxique et utilise une analyse linguistique basée sur l'analyse distributionnelle de Z. Harris.
La figure ci-dessous résume les grandes étapes d'acquisition de termes. Ces étapes sont automatiques, elles sont toutefois terminées par une étape de filtrage et de validation effectuée par un linguiste.

Textes

Prétraitement de phrases

Structuration

Extraction de termes

Fig. 8 : Les étapes de l'acquisition de termes

Filtres

Validation

3.2.1 Le prétraitement de corpus

Le prétraitement est une étape importante puisqu'elle permet de constituer les données lexicales. Il s'agit de découper le texte en une suite de mots en utilisant les caractères séparateurs et les caractères d'espacement (il s'agit essentiellement des caractères non imprimables comme le blanc).

3.2.1.1 L'apostrophe et le trait d'union

Les chaînes de caractères comme "rendez-vous", "peut-on" sont segmentées, elle sont représentées par deux occurrences (rendez et vous, etc.). Le trait d'union n'est donc pas considéré comme partie intégrante d'un mot. De même pour l'apostrophe qui est considéré comme un séparateur; ainsi par exemple la chaîne de caractères "l'homme" représente deux occurrences de mots : "l" et "homme".

3.2.1.2 Les séparateurs et espacements

Les séparateurs et les espacements sont utilisés pour la segmentation du texte en phrases et dans le calcul des termes :

ESPACES

- Tabulation
- Nouvelle ligne
- Blanc

SEPARATEURS

- les espaces
- guillemets, tiret, parenthèses, accolade, crochets " ( ) [ ] { }
- virgule, point-virgule , ;
- points ? ! . :

3.2.1.3 Segmentation du texte en phrases

Cette segmentation a pour but de constituer une liste de phrases du corpus qui seront aussi

numérotées. Une fin de phrase est indiquée par les séparateurs de fin de phrase suivants : (; : ! . et ?) qui sont considérés comme des mots qui terminent la phrase. Un retour chariot est toujours interprété comme une fin de phrase, même en l'absence de ponctuation. Cela permet de considérer les titres comme des phrases.

Données en entrée	Corpus informatisé
Données en sortie	Phrases + Mots
Intervenants	Cogniticien + expert

3.2.2 Extraction de terme

L'extraction de termes s'appuie sur les présupposés suivants (ou sur une partie d'entre eux selon la stratégie adoptée).

a) Un terme significatif sera utilisé à plusieurs reprises dans un texte spécialisé.

b) La très grande majorité des termes sont de nature nominale.

c) La plupart de ces termes sont complexes, c'est-à-dire qu'il sont composée de plusieurs mots par ailleurs utilisés isolément (ex. *pression artificielle; intelligence artificielle*)

d) Les termes complexes se construisent au moyen d'un nombre fini de séquences de catégories grammaticales. En effet, la plupart des termes complexes français se composent d'un nom modifier par :

- Un adjectif : ex. *intelligence artificielle, haute tension;*
- Un syntagme prépositionnel contenant un nom : ex. *robinet de commande, traitement de la langue.*
- Un syntagme prépositionnel contenant un verbe : ex. *machine à coudre;*
- Un autre nom : ex. *imprimante laser, page Web.*
- N'importe quelle combinaison des séquences ci-dessus : ex. *temps de conduction auriculaire.*

Pour extraire les termes à partir des phrases du corpus de manière automatique, on calcule la liste des séquences de mots répétées, ne contenant pas de signes de ponctuations; on les appelle les *segments répétés*; puisqu'elles apparaissent au moins deux fois dans le corpus et ne contiennent pas de signes de ponctuation.

3.2.2.1 Notion de segment répété

D'après [**Oueslati, 99**] et [**Drouin, 2002**] :

> ***Définition** : Lebart et Salem (1994) définissent un **segment répété** comme une suite d'occurrences (une séquence de mots) qui apparaît au moins deux fois dans le texte, et qui ne comprend pas de signe de ponctuation.*

La séquence de mots est calculée au sein d'une fenêtre étroite (par exemple de 2 à 10 mots). D'après [**Oueslati, 99**] Cette notion de segments a été également utilisée par Justeson et Katz (1995) dans le cadre du système TERMS pour des segments de longueur 2. Dans notre cas les termes sont variés de 2 à une valeur maximale définie par l'expert de domaine.

Notre méthode d'acquisition des termes est donc basée sur le calcul de segments répétés. Le but est de repérer des termes qui sont des syntagmes nominaux et de les structurer.

Pour filtrer les segments répétés, on utilise les filtres ou les *frontières de terme* selon la terminologie de Bourigault [**Drouin, 2002**].Cette méthode de calcul et de filtrage des segments répétés est donc facilement transportable dans d'autres langues (par exemple l'anglais) puisqu'il suffit d'utiliser le filtre et les règles de découpage de segments dans la nouvelle langue sans modifier les algorithmes de traitement.

3.2.2.2 Les étapes de constructions de termes

Afin d'obtenir de tels segments répétés, on utilise un filtre composé des mots grammaticaux de la langue (conjonction, prépositions, etc.) et *les règles de déductions contextuelles* qui portent sur un mot et ces proches voisins. Connaissant la nature d'un mot on peut déduire, à l'aide de ces règles, celle des ces mots qui l'entourent. Plus des ces règles, *le traitement morphologique* permet de connaître le finale de mots. Nous utilisons les deux dernières méthodes pour la détection des formes fléchies des verbes. Ces filtres servent à délimiter les segments répétés.

La dernière étape de l'acquisition consiste à valider les segments répétés par un linguiste (familiarisé avec le domaine) afin de constituer une liste de termes.

- ……..
- …intelligence artificielle……
- ……

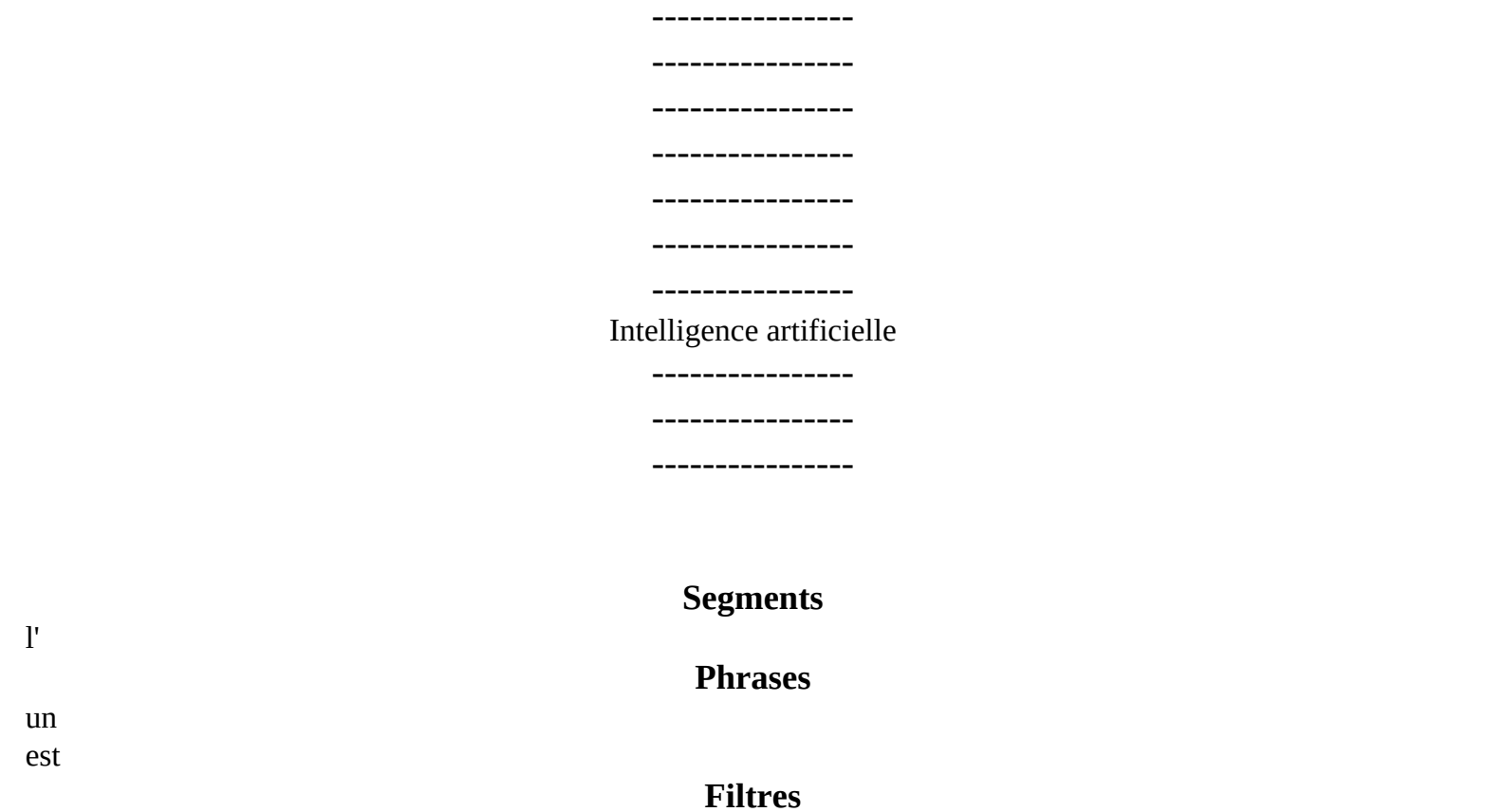


Fig 9 : Processus de construction des syntagmes répétés

ALGORITHME : Acquisition des segments répétés

DONNEE

- soient *FILgram, FILponc, et FILverb* les filtres utilisés pour le filtrage des segments répétés.
- soit *Table (Seg, NBocc)* : la table des séquences de mots calculées avec *Seg* une séquence de mots, et *NBocc* son nombre d'occurrences.
- Initialement : *Seg* (segment répété) est vide, *Cat* (catégorie de mot) est vide.

TRAITEMENT

REPETER

LIRE (chaîne**).**

/* la chaîne soit un mot ou une chaîne de caractère spéciaux ou une ponctuation. */

SI chaîne ∈ FILponc **ALORS** è **SI** NBMot (Seg)>=2 **ALORS**

Insert (Table (Seg, NBocc**)).**

è Seg : = ϕ et Cat : = ϕ.

SINON /* la chaîne est un mot */

SI chaîne ∈ FILgram **ALORS**

è Cat : = Cat + catégorie (chaîne).

è **SI** Seg ≠ ϕ **ALORS** Seg : = Seg + chaîne

SINON SI chaîne ∈ FILverb ou adverbe **ALORS**

è **SI** NBMot (Seg)>=2 **ALORS**

Insert (Table (Seg, NBocc**)).**

è Seg : = ϕ et Cat : = ϕ.

SINON è Seg : = Seg + chaîne

è Cat : = ϕ.

JUSQU'A la fin de texte.

On note que la procédure **Insert** permettre de l'insertion le segment s'il n'existe pas dans la liste sinon on incrémente le nombre d'occurrence de 1.

RESULTAT

Le résultat de traitement est une liste des segments avec leur fréquence. Avant d'éliminer les segments qui ont leurs fréquences égales à 1, il faut bien normaliser l'écriture des segments. C'est le but de la lemmatisation des mots.

Filtre grammatical

Filtre de verbe et adverbe

Non

Non

Oui

Un nouveau mot

Recherche le mot dans la liste

Mot dans la liste

Recherche le mot dans la liste

Mot dans la liste

Cat < Cat + CatMot

Si Seg ≠ ϕ **Alors**

Seg < Seg + Mot

Oui

Cat < =ϕ.

Si NB(Seg) ≥2 **Alors**

Insert (Seg).

Seg <= ϕ.

)

Cat < =ϕ.

Seg <= Seg + Mot

Fig. 9 : Algorithme de calcule de segment répété

3.2.2.3 Lemmatisation

Définition *: On appelle lemmatisation le regroupement des formes morphologiques d'une même unité linguistique en en seule unité appelé lemme.*

On remarque qu'il y a une beaucoup plus grande économie des marques en oral qu'en écrit. Pour remédier au problème des redondances dans les marques morphologiques, il serait intéressant de regrouper par exemple les formes singulier/pluriel d'un même SN sous forme unique, ce que on appel *lemmatisation*.

Au cours de l'étape de lemmatisation, les ressources du corpus sont exploitées au maximum puisque l'information qu'il contient sera utilisée pour prendre des décisions relatives au statut de formes tirées elles aussi du corpus. L'analyse de la liste des formes identifiées laisse prévoir qu'à l'aide de quelques règles simples, il est possible de procéder à une lemmatisation automatique.

A la différence des systèmes de TALN qui utilisent des analyseurs morphologiques et des dictionnaires, nous avons choisi de procéder à une *lemmatisation sommaire*, qui consiste à regrouper sous un même lemme uniquement les termes au singulier et au pluriel. L'avantage est que ce regroupement est facile à implémenter est plus simple à adapter à d'autres langues que le français. Dans notre regroupement, on considère que le lemme est :

- la forme infinitive pour les verbes, et

- pour les autres catégories, la forme du masculin singulier du lexème.

En générale, les segments répétés sont des syntagmes nominaux, donc pour la lemmatisation nous intéressons de lemmatiser les syntagmes nominaux.

Les étapes de lemmatisation

- Trie la liste des segments répétés dans un ordre alphabétique.
- Si la différence entre deux mots successifs appartient dans l'ensemble {l, s, x , ux} on prend la forme singulier et on incrémente à 1 le nombre d'occurrence. On ajoute la forme plurielle à la table des segments répétés dans la colonne de lexème et on le retire de la liste des segments répétés.

On se résume cette étape par le tableau suivant :

Données en entrée	Phrases + Mots
Données en sortie	Segment répété
Intervenants	Cogniticien

3.2.2.3 Traitements d'ambiguïtés (*coordination*)

Les différents éléments les plus fréquents de la coordination sont : ou, et. L'utilisation de ces éléments peut entraîner l'absence de quelques unités (mots) dans la proposition.

Dans l'exemple « *l'informatique est un outil rapide et précis* », en analysant la portion « *et précis* » le lecteur sait que le mot « outil » est distribué sur les différents membres de la coordination, par contre la machine ne sait pas. De ce fait, on doit restaurer les unités syntaxiques effacées.

Pour résoudre le problème de la coordination, la méthode suivie est de chercher une structure syntaxique identique des deux cotés au sein d'un syntagme nominal. A titre d'exemple :

Cas 1 : Mot + COOR + Mot + PREP + Mot

Ex. Acquisition et représentation des connaissances.

- è Acquisition des connaissances.
- è Représentation des connaissances.

Cas 2 : Mot + Mot + COOR + Mot

Ex. Structuration sémantique et statistique

è Structuration sémantique

è Structuration statistique

3.2.3 Décomposition et structuration des termes

Les syntagmes identifies sont le plus souvent des syntagmes nominaux complexes. La fonction du module de décomposition est d'effectuer une décomposition binaire des ces syntagmes, en une *tête* et une *expansion*, de façon, d'une part, à extraire à partir des syntagmes nominaux complexe des sous- groupes constituants susceptibles d'être eux aussi de bon candidats termes et d'autres part pour permettre l'organisation ultérieure de l'ensemble des candidats termes sous la forme d'un réseau.

A partir de considérations empiriques et théoriques, deux types de décomposition binaire ont été retenus :

(1) Dans le cas où la tête et l'expansion sont connectées par une séquence « préposition + déterminent » (c'est-à-dire *du, de la, des, d'un, sur le, etc.*), les postions tête et expansion sont alors T' et E'.

Exemple :

Acquisition des connaissances

è **T'** : acquisition

è E' : connaissances

(2) dans tous les autres cas, les position têtes et expansion sont alors notées T et E. *Exemple :*

Structuration sémantique

è **T :** Structuration

è **E :** sémantique

Le module de décomposition indique quelles sont les sous structures qui correspondent à la tête (notée T ou T', selon les cas) et à l'expansion (notée E ou E', selon les cas).

En cas d'ambiguïté, c'est-à-dire, lorsqu'il existe plusieurs décompositions possibles pour un candidat terme complexe (plus de trois mots), nous proposons quelques règles qui permettent de lever l'ambiguïté et de donner une seule décomposition.

A l'issue de la phase de décomposition, un module de structuration exploite les décompositions en tête et expansion des candidats termes pour construire un réseau, dit *réseau*

terminologique : chaque candidat terme est relié, d'une part, à sa tête et à son expansion, et, d'autre part, à tous les candidats termes dont il est lui-même tête ou expansion. Un exemple de réseau simple est donné dans la figure ci-dessous.

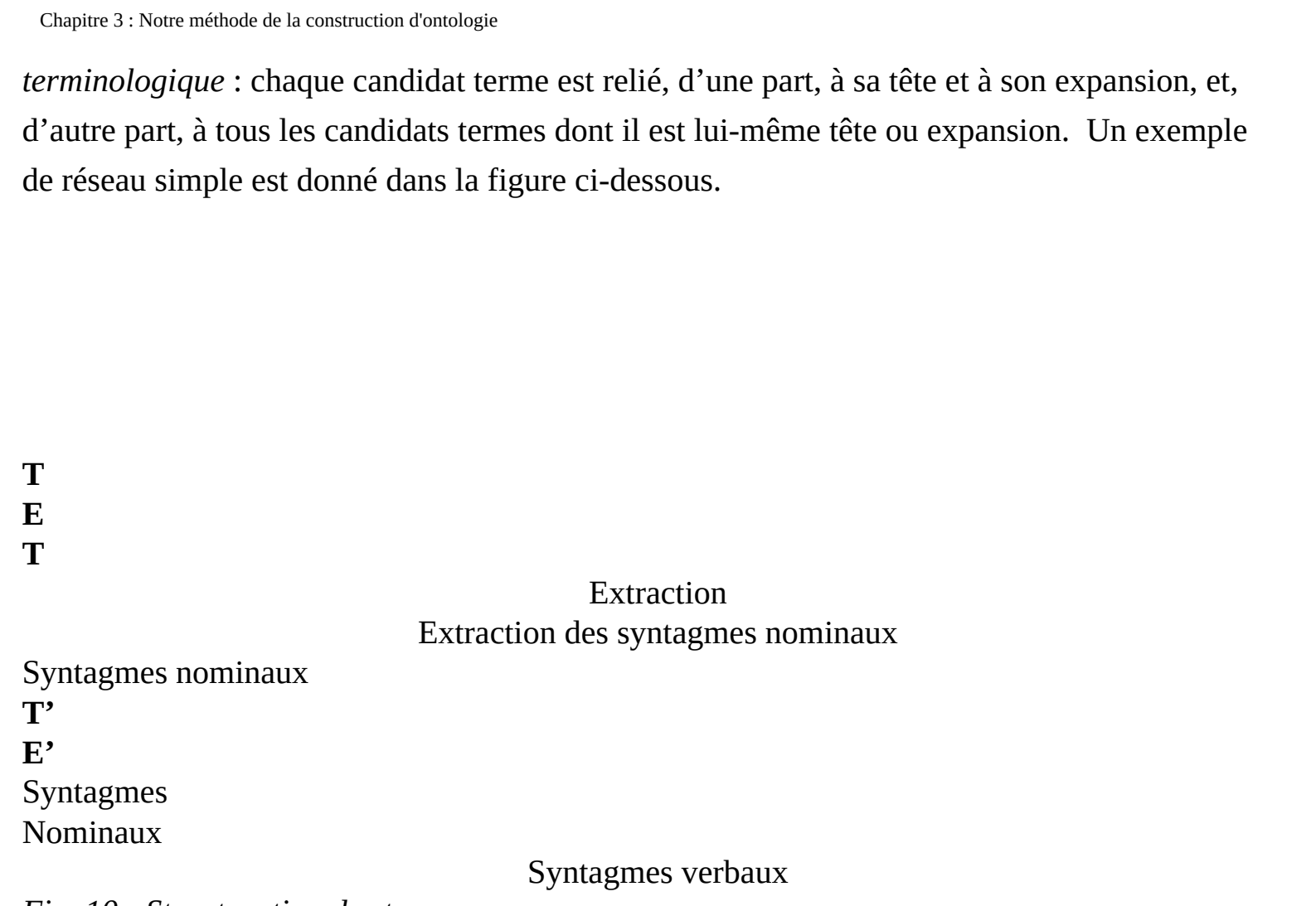


Fig. 10 : Structuration des termes.

3.2.3.1 Traitement d'ambiguïté

Un problème posé par les syntagmes nominaux comportant plus de deux mots, est que l'on ne sait pas retrouver la façon dont ils ont été formés. Ainsi, quand on rencontre un syntagme nominal de forme N1 N2 N3, doit-on l'interpréter comme [N1 N2] N3 ou N1 [N2 N3] ? bien que cette structure soit ambiguë. Mais le premier cas se rencontre plus souvent que le second.

Cas 1 : **N1 N2 N3 è N1 [N2 N3]**

Ex. Système nerveux central è [Système nerveux] central

è Système [nerveux central] ?

Cas 2 : **N1 P N2 N3 è N1 P [N2 N3]**

Ex. Acquisition d'informations lexicales è Acquisition d'[informations lexicales] _

Cas 3 : **N1 P N2 P N3 è N1 P [N2 P N3]**

Ex. logiciel d'extraction de terminologie è logiciel d'[extraction de terminologie]

Cas 4 : **N1 N2 P N3 è [N1 N2] P N3**

Ex. Indexation automatique de document è [Indexation automatique] de document

D'après les exemples illustrés, nous remarquons que généralement une préposition relie deux syntagmes nominaux. Toutefois, on peut rappeler que, même pour un expert, il est souvent difficile de donner un résultat sans information sur le contexte d'utilisation du syntagme.

Données en entrée	Segment répété
Données en sortie	Réseau terminologique
Intervenants	Cogniticien

3.2.4 Validation

Cette tâche manuelle n'est pas obligatoire ; elle est destinée au « nettoyage » de la liste des segments répétés par élimination des erreurs éventuelles et des candidats termes jugés non pertinents par rapport au domaine. Cet élagage permet d'avoir des données moins « bruitées » pour les tâches suivantes, et notamment pour la classification automatique. Cette tâche permet également au cogniticien de commencer à se familiariser avec la terminologie du domaine.
En général, l'élagage correspond à plusieurs cas :

- erreurs de catégorisation morpho-syntaxique (certains verbes, participes passés, etc.).
- candidats termes rejetés car jugés trop généraux par rapport au domaine considéré (par exemple, CAS, SUPPLEMENTAIRE, EVENTUEL).
- Candidats termes rejetés car toutes leurs occurrences sont contenues dans des candidats termes plus complexes.

L'interface de l'extracteur des termes doit permettre la navigation dans les réseaux terminologiques des candidats termes. Cette interface doit comporte toutes informations concernant le candidats termes sélectionné : les information numérique, les variantes et les synonymes du candidat terme, liste des termes permet l'accès direct au texte et les descendants

d’un candidat terme.

On peut résumé cette étape dans les points suivantes :

Données en entrée	Segment répété bruité
Données en sortie	Segment répété moins bruité
Intervenants	Cogniticien + expert

3.2.5 Mesure d’efficacité

La liste des candidats termes est composée à la fois de termes et d’unités qui ne sont pas des termes. Ces dernières [**Drouin 2002**], malgré qu’elles n’aient pas de statut terminologique bien défini, sont tout de même intéressantes pour le terminologue dans la majorité des cas. Ces formes sont cependant généralement conservées pour l’évaluation 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 ci-dessous.

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 l’évaluation 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 l’outil.

CT

TR

TR^{-}

CT^{+} **CT^{-}**

Fig. 11 : Mesure d’efficacité des logiciels

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

Parmi les indices les plus fréquents retenus pour évaluer les performances à l'aide de divers sous-ensembles illustrés par la figure ci-dessus, on a souvent recours aux notions de rappel, de précision, de silence et de bruit. Ces indices nous viennent du domaine de l'indexation et de la recherche documentaire où 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 l'ensemble des réponses pertinentes possibles. Ainsi, si un corpus contient 100 termes attestés par un terminologue (TR) et qu'un système identifie 150 CT, 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 l'intersection entre la liste qui nous sert de référence et la liste générée par le logiciel.

D'après [**Drouin 2002**], 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 %.

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

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 l'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). Malheureusement, le silence n'est généralement pas pris en considération dans le cadre

des recherches en acquisition automatique des termes. Il 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 CT sont mis de côté au profit d'une augmentation de la qualité des résultats.

Enfin, le *bruit (B)* évalue le nombre de CT extraits par le logiciel qui sont absents de la liste de référence (CT^-) par rapport au nombre total de CT extraits par le logiciel (CT). En reprenant notre exemple, on obtient une valeur de 55 CT erronés (CT^-) alors que le total de CT proposés est fixe à 150; le bruit est donc de 55/150 (0,37).
Cet indice recouvre donc la portion négative des CT et peut aussi être exprimé sous la forme $B = 1-P$ à partir du moment où la précision est une valeur connue.

3.3 Normalisation

3.3.1 Principe générale

L'idée de base de la réalisation de la normalisation est d'inspiration Harrissienne (Harris, 68) : *rapprocher les syntagmes nominaux ayant des contextes syntaxiques similaires et constituer ainsi des classe des candidates termes, dont certaines seront interprétables sémantiquement et constitueront l'amorce d'une organisation de l'ontologie en champs conceptuels*. Plus précisément, si l'on décrit chaque candidat termes par son contexte terminologique, c'est-à-dire, en simplifiant, par l'ensemble des termes auxquels il est relié dans les réseaux terminologique on peut utiliser des méthodes statistiques de classification ascendante hiérarchique (CAH) pour obtenir des classes des candidats termes.
Lors de cette étape, le cogniticien et l'expert de domaine doivent se familiariser avec le domaine, dégager les champs conceptuels importants et entamer la validation des candidats termes et la constitution de l'ontologie de domaine : c'est la phase d'amorçage de l'ontologie. Donc, l'interprétation des résultats de la classification par les cogniticiens conduit à définir des champs conceptuels. Les candidats termes apparaissent dans un champs conceptuel représentent les étiquettes de concept. Cette étape est l'occasion de constituer des ensembles de synonymes : les termes jugés équivalents sont groupés et l'un des termes est choisi comme étiquette de concept. Les autres valeurs sont alors les valeurs génériques du concept. Le processus se répète jusqu'à trouver une ontologie stable (voir la figure ci-dessous).

Fig. 11 : L'étape de normalisation des concepts

Pour atteindre l'objectif de normalisation, nous proposons de suivi les points suivants :

3.3.2 La classification

Le but de classification est d'appréhender le comportement des syntagmes sur la globalité du corpus. En effet, si le cogniticien se contente de lire la documentation technique de manière linéaire et de parcourir la liste, des candidats termes, il ne sera pas en mesure d'avoir une idée du comportement global des termes du domaine. La mise en place d'une ontologie revient un problème très difficile ou impossible.
Notre idée est de construire un outil d'aide qui offre au cogniticien un outil qui fournit des classes de syntagmes pouvant être interprétés en termes conceptuels. Un tel outil peut être vu comme un outil d'aide à l'analyse sémantique. Une classe de candidats termes est pertinente lorsque son interprétation par le cogniticien conduit celui-ci à construire *un champ conceptuel*.

> ***Définition :*** *on définit un champ conceptuel comme étant un ensemble de candidats termes auquel le cogniticien a associé un label conceptuel. Les candidats termes d'u champ conceptuel peuvent entretenir entre eux différentes relations : synonymie, méronymie, hyperonymie.*

a- Principe de la conception

Les dépendances syntaxiques donnent des indices sur la proximité sémantique. Par exemple, il est possible de construire des classes de noms en étudiants leurs contextes phrastiques et en regroupant ceux qui sont sujet des mêmes verbes. Cette idée est due à Z. Harris (1968).

b- Le contexte terminologique

Pour effectuer la classification des candidats termes nous utilisons les relations de dépendance entre candidats termes. Les candidats termes sont considérés comme des individus à classer en

fonction d'une certaine description (les variables). Cette description représente des relations syntaxiques que ces individus entretiennent au sein du réseau terminologique. Pour cette raison nous introduisons le terme de *contexte terminologique* pour désigner cette description. Donc, les candidats termes ou les individus sont décrits par leur contexte terminologique.
Par exemple le contexte terminologique en position E du terme *système* est l'ensemble des termes en position E et comme tête le terme *système*.

Candidats termes	CTE (système)
- système d'information - système exploitation - système expert - système de gestion de base de donnée - système en temps réels	- information - exploitation - expert - gestion de base de donnée - temps réels

Définition : *on peut définir le contexte terminologique expansion d'un candidat terme comme étant l'ensemble des candidats termes apparaissant en expansion de candidats termes plus complexes ayant le dit candidat terme en tête.*

On peut donnée la même définition de contexte terminologique dans les différents d'autre position (T, E', T').

c- Interprétation de contexte terminologique

Les termes construits par ajout d'expansions représentent des spécialisations du terme tête en spécifiant un attribut de celui-ci, les termes décrits par leurs *« contextes terminologiques expansion* » seront rapprochés lorsqu'ils partagent un certain nombre d'attributs en commun. Les classifications faites à partir de cette similarité pourraient alors conduire à mettre en évidence des rapprochements sémantiques entre candidats termes. Ces rapprochements conduisent le cogniticien à construire des champs conceptuels.
Pour le contexte terminologique en position T', les candidats termes apparaissant en position T' par rapport à un candidat terme donné peuvent exprimer des propriétés ou des actions concernant le concept ayant le dit candidat terme comme étiquette. Par exemple {*capacité de la mémoire, extension de la mémoire*}, « *capacité* » représente l'un des propriétés de la « *mémoire* » et «*extension* » est une action pouvant être appliquée à la « *mémoire* ».

On remarque que la notion de contexte ici est différente par rapport aux notions classiques utilisées en recherche documentaire où le contexte d'un terme est souvent obtenu en considérant ses voisins dans une fenêtre de mot de taille fixée au préalable. Pour cette raison nous proposons les différentes combinaisons possibles de contexte terminologique pour qu'il soit utilisé ultérieurement dans la méthode de la classification. Voici deux exemples de combinaisons possibles :

- Les individus dans notre cas sont décrits par leur contexte terminologique E et T'.

l'idée est de rapproché les individus ayant le même contexte terminologique E et T', c'est-à-dire les individus qui partagent les attributs, les propriétés ou les actions communs.

- Les variables sont décrits par leur contexte terminologique T et E'

d- Les étapes de la constitution des classes

d.1- la sélection des individus

En général, le nombre de candidats termes dépend de la taille de corpus, et pour des raisons liées à la capacité de validation manuelle de résultats de la classification par le cogniticien nous proposons de travailler sur une population ou une partie de population réduite.
Pour sélectionner cette population parmi les milliers de candidats termes nous calculons un critère numérique, nommé « *la connectivité* ». Le cogniticien choisit un seuil, et retient alors tous les candidats termes dont la connectivité est supérieure à ce seuil.

Définition : *soient deux candidats termes t_1 et t_2. Soit R • {T, E, T', E'}. On définit la connectivité de t_1 et t_2 selon la relation R de la manière suivante :*

$$Connec\_R(t_1, t_2) = card(ct\_R(t_1) \cap ct\_R(t_2))$$

On définit également la connectivité selon plusieurs relations par addition des connectivités selon chacune des relations prises séparément.
Pour sélectionner la population à classer, le cogniticien donne un seuil. Tous les couples de candidats termes ayant leur connectivité selon les relations choisies, supérieures à ce seuil sont intégrés dans la population.
Ensuite, il faut sélectionner les variables qui contribuent à la description de la population d'individus sélectionnée.
L'utilisation du critère de la connectivité nous assure d'obtenir une population partagent un nombre minimal de variables, nous évitons ainsi une trop grande dispersion des données.

d.2- La classification

L'idée est de rapprocher les candidats termes qui ont des contextes terminologiques similaires. On peut ainsi constituer des classes de candidats termes ayant le même contexte syntaxique et espérer d'obtenir une interprétation conceptuelle à partir de la lecture de ces classes. Nous avons utilisé des méthodes de classification automatique et plus précisément une classification ascendante hiérarchique (CAH). En effet, pour nous, la classification automatique est un outil et non un objet de recherche (un exemple donnée en annexe pour la méthode CAH).

d.3- Méthodes de classifications

> ***Définition** : Le but des méthodes de classification est de construire une partition ou une suite de partitions emboîtées, d'un ensemble d'objets dont on connaît les distances deux à deux. Les classes formées doivent être le plus homogène possible.*

Une classification est dite *hiérarchique* lorsqu'elle abouti à un arbre de classification : on obtient ainsi des partitions en classes d moins en moins fines obtenues par regroupement successifs de parties. Enfin, une classification hiérarchique est dite ascendante lorsqu'on commence par regrouper les deux individus les plus proches pour former un nouvel objet ; on réitère alors le processus jusqu'à regroupement complet. Une CAH exige la définition d'une distance (ou une *mesure de similarité*) entre individus et entre classes et l'adoption d'une stratégie pour *l'agrégation* des parties au cours de la classification.

d.4- Le tableau individu- variable

Les colonnes sont étiquetées par les expansions E ou les têtes T' (les variables). Chaque ligne représente un individu (la tête T ou l'expansion E') : il y a un '1' lorsque le candidat terme constitué par un individu et variable existe et '0' sinon.

A partir du tableau individu- variable, (voir l'exemple ci-dessous extrait de la thèse [**Assadi, 1998**]) une mesure de similarité entre les individus est calculée. Ensuite, une méthode de classification hiérarchique ascendante est utilisée avec en entrée, le tableau de dissimilarité. Cette méthode produit un arbre de classification qui est coupé à un niveau donné pour former des classes.

	Admissible	Nominale	HT	THT	63kv
Câble	1	0	0	1	1
Ligne	1	1	0	0	1

Tension	0	1	1	0	1

Tableau. Format des données utilisées pour la classification

d.5- Les mesures de similarité

Le tableau individus- variables est un tableau binaire : une présence de la variable v chez l'individu i est signalée par un '1' sinon par un '0'.

Pour cette raison nous travaillons sur de mesures de similarité adaptées a des données binaires. Nous donnons ici deux exemples de mesures de similarité adaptée aux données binaires.

On suppose que :

1) soit i et j deux individus du tableau individus – variables

2) On définit les grandeurs suivantes :

a(i, j) = nombre de variables présentes à la fois chez i et chez j.

b(i, j) = nombre de variables présentes chez i et absente chez j.

c(i, j) = nombre de variables présentes chez j et absentes chez i.

d(i, j) = nombre de variables absentes à la fois chez i et chez j.

3) La mesure de *similarité de Jaccard*

La mesure de *similarité d'Anderberge modifiée* :

Tel que α est une paramètre, compris entre 0 et 1, qui permet d'ajuster la contribution de chacun des deux termes de *sim_A*. En effet, le premier terme de la formule représente la contribution des variables présentes chez i te j à la mesure de similarité alors que la deuxième terme représente la contribution des variables absentes des deux individus.

Par rapport au notion de *contexte terminologique*, le paramètre *a(i, j)* donne directement le nombre de contextes partagés par les individus i et j. La somme *(a(i, j) + b(i, j) + c(i, j))* donne le nombre total de variables présentes chez l'individu i ou l'individu j. Donc, la mesure de similarité *sim_J* est la proportion des variables communes aux individus i et j par rapport à l'ensemble des variables présentes chez i ou chez j. On peut obtenir *des mesures de dissimilarité* par la complémentation à 1.

d.6- Les stratégies d'agrégation des classes

L'algorithme de CAH prend en entrée une matrice de dissimilarité entre les individus à classer. La première itération de l'algorithme, les deux individus les plus proches (qui réalisent le minimum dans la matrice de dissimilarité) sont regroupés, on obtient ainsi un nouvel objet. Le problème qui se pose est celui de calculer la dissimilarité entre un tel objet et un individu ou entre deux objets dans les différents itérations de l'algorithme.
Pour ce faire, plusieurs méthodes peuvent être mises en œuvre, on parle de *stratégie d'agrégation sur dissimilarité.* Parmi ces stratégies, les plus classiques sont :

- *Le saut minimum* : la distance entre deux objets est la plus petite, autrement dit, il s'agit du minimum de la distance entre éléments des deux parties.
- *Le diamètre (ou saut maximum)* : on prend ici le maximum de la distance entre éléments des deux parties.
- *Saut moyen* : on prend la moyenne des distances entre éléments des deux parties.
- *Saut de Ward* : cette stratégie est valable uniquement pour les distances euclidiennes. Dans cette stratégie on cherche à chaque itération un minimum local de l'inertie intraclasse ou un maximum de l'inertie interclasse.

d.7- La constitution des classes par coupure de l'arbre

Coupure de l'arbre :
Le résultat d'une CAH est un arbre de classification. Pour obtenir des classes, la méthode la plus usuelle est de couper l'arbre à niveau (ou seuil) donné. Il existe des critères numériques, dépend de la méthode de classification, qui indiquent le(s) seuil(s) optimal (optimaux) pour couper l'arbre. Le niveau de coupure s'appel *le niveau d'agrégation.*

Allocation des variables aux classes
Une fois les classes constituées, il faut allouer pour chaque classe par des variables qui la caractérisent au mieux, on obtient ainsi le contexte de la classe. Nous présentons ici une méthode inspirée de (Blanchard & Augendre 1994). Soit C la classe en question, la réallocation des variables se déroule en trois étapes :

- d'abord, nous retenons l'ensemble des variables V qui apparaissent chez au moins un individu de la classe.
- Ensuite, nous calculons le *pouvoir discriminant* de chaque variable j de V.

I désigne l'ensemble de tous les individus de la population et M la matrice individus –

variables. D_j mesure la proportion des individus de I possédant la variable j et faisant partie de la classe C est caractéristique de la classe.

Nous retenons alors les variables dont le pouvoir discriminant dépasse un certain seuil S, ce qui correspond à l'ensemble V' défini par :

§ Enfin, une troisième phase consiste à trier les variables de l'ensemble V' en fonction de leur *taux de recouvrement* R_j.

Plus la grandeur R_j est élevée, plus la variable j contribue à rapprocher les individus de la classe. Ainsi, l'ensemble V' trié par R_j décroissant permet d'avoir, pour chaque classe, l'ensemble des variables les plus représentatives et d'avoir en tête de cette liste les variables qui contribuent le plus au rapprochement des individus de la classe.

On peut résumé l'étape de la classification dans le tableau suivant :

Données en entrée	Réseau terminologique
Données en sortie	Plusieurs classifications
Intervenants	Cogniticien

E. L'exploitation des résultats

L'exploitation des résultats se fait par un expert du domaine et un cogniticien. Avant, le cogniticien doit essayer plusieurs combinaisons des paramètres ci-dessus avant de retenir une ou plusieurs classifications qui vont servir pour la tâche suivante. Nous préconisons la démarche suivante :

D'abord, il faut commencer par sélectionner une population relativement réduite pour effecteur la première classification. Ainsi, en travaillant sur des données de taille réduite le cogniticien peut aborder la terminologie du domaine en minimisant la « surcharge cognitive ». La première classification à faire est donc celle des termes, avec les contextes terminologiques T' et E et une connectivité assez forte (pour permettre d'obtenir une population de taille réduite).

Ensuite, une fois que le cogniticien a mis en évidence les premiers champs conceptuels, il peut effectuer d'autres classifications plus importantes, pour étendre les champs conceptuels existants ou en découvrir d'autres. Par exemples, il peut diminuer le seuil de connectivité.

Le cogniticien examine les résultats des différentes classifications effectuées à l'étape précédente les premiers champs conceptuels du domaine. Le cogniticien soumet alors à l'expert les champs conceptuels pour validation. Le cogniticien peut également demander à l'expert de lui donner des indications sur des classes qu'il n'est pas arrivé à interpréter et dont

il attend des informations intéressantes.

Données en entrée	Les classifications
Données en sortie	Principaux champs conceptuels
Intervenants	Cogniticien + expert

Lors de l'étape précédente, plusieurs classifications ont été effectuées, avec des populations de taille croissante et décrites par des cotextes terminologiques de plus en plus larges. Le processus d'interprétation de la classification de la première population est essentiellement inductif : il s'agit de « découverte » par le cogniticien de champs conceptuels dans les classes. Ensuite, le cogniticien aborde l'interprétation des classifications des populations suivantes avec des connaissances issues de la première interprétation. Dans ce cas, le processus n'est pas purement inductif. Il s'agit d'abord de retrouver les noyaux conceptuels déjà identifiés ; si ces noyaux sont présents dans des classes plus larges, c'est-à-dire contenant des termes absents de la première classification, alors le cogniticien est amené à étudier la possibilité d'élargir les champs conceptuels déjà existants, de les modifier ou d'en créer d'autres.
On peut résumé que pour chaque classe, il y a trois décisions possibles :

1. rejeter la classe car le regroupement effectué est non pertinent.
2. retenir la classe et lui donnant un label conceptuel ; il y a alors deux cas, soit la classe correspond à des synonymes, soit elle regroupe des termes faisant partie du même champ conceptuel et ayant entre eux des liens autres que la synonymie (hyperonymie, méronymie). C'est le cas d'un champ conceptuel homogène.
3. retenir la classe en observant qu'elle correspond à plusieurs champs conceptuels hétérogènes.

Lorsque l'expert retient une classe, il peut éliminer un certain nombre de termes de celle-ci en considérant qu'ils ne sont pas pertinents par rapport au champ conceptuel qu'il a défini à partir de la classe.

3.3.3 Construction de la première version de l'ontologie

Données en entrée	Les classifications + champs conceptuels
Données en sortie	Première version d'ontologie

Intervenants	Expert + cogniticien

Lors de l'étape précédente, le cogniticien a abouti à une première organisation de la terminologie de domaine. A présent, il met en place une structure plus élaborée : l'ontologie de domaine. Tout d'abord, le cogniticien construit un concept à partir de chaque champ conceptuel identifie à l'étape précédente. Ensuite, il construit un concept à partir de chaque ensemble de synonymes appartenant à un champ conceptuel, crée un concept relié à cet ensemble de synonymes. Enfin, il établit un lien *sort_de* entre chacun des concepts provenant d'un groupe de synonymes (éventuellement réduit à un seul terme) d'un champ conceptuel et le concept correspondant à ce champ conceptuel.
L'enrichissement des concepts se fait par l'ajout des descendants par exemple le concept *Intelligence Artificiel* est descendant de *intelligence*. Donc, Le but est de synthétiser les termes avec une même tête (T) ou une même expansion (E') dans une seule structure d'arbre.

Ligne
Champ conceptuel : ouvrage
Type : objet
Etiquettes : ligne, ligne électrique, ligne aérienne

[1]

Fig. 12 : description partielle du concept ligne .

3.3.4 Raffinement itératif de l'ontologie

Données en entrée	Première version d'ontologie
Données en sortie	version finale de l'ontologie
Intervenants	Expert + cogniticien

Lors de cette étape, il s'agit de compléter la description des concepts centraux. Le cogniticien s'intéresse, pris un à un, aux concepts qu'il juge centraux et essaie de les décrire de manière plus exhaustive. Cette description s'appuie sur l'analyse des candidats termes autour des étiquettes de concept. A la fin de cette étape, la description de chaque concept central sera complétée par des liens avec d'autres concepts et par des attributs en partant d'une description minimale comme celle de la figure.
Nous illustrons ici avec exemple la première itération de raffinement. D'abord le cogniticien

sélectionne un concept. Il peut alors accéder aux termes contenant toutes les étiquettes de celui-ci. Par exemple, à partir du concept *ligne,* il peut accéder aux termes contenant *ligne, ligne électrique, etc.* Ensuite, il précise, au sein de la liste T, c'est-à-dire la liste des autres termes le contenant en position tête, à quel patron syntaxique il s'intéresse. Par exemple, <SN de SN>. Le système propose alors des paradigmes en projetant les champs conceptuels sur les candidats termes. L'interprétation et la validation de ces propositions par le cogniticien conduisent celui-ci, d'une part à décrire avec précision les concepts centraux, d'autre part à compléter les champs conceptuels déjà repérés à la première étape ou en créer d'autres.

A l fin de cette étape, la description des concepts centraux est étendue. Ensuite, un cycle de raffinement de l'ontologie est entrepris, le système fait d'autres propositions en utilisant les informations données par les cogniticien à l'itération précédente.

4. APPLICATION A L'EXPANSION DE LA REQUETE UTILISATEUR

Ontologie
Requête
Information

Moteur
de recherche

AO
UA

Notre projet a pour but de modéliser une vision par exemple de l'informatique, adaptée à nous besoins et de regrouper au sein d'un site un ensemble de cours informatique. Le public visé sera en premier temps les élèves de l'informatique (UA : utilisateur d'application). Les termes de la requête saisis par l'utilisateur dans l'interface d'interrogation d'un moteur de recherche aidée par ontologie (voir la figure ci-dessous) sont analysés par un programme de moteur de recherche. Si un ou plusieurs termes de la requêtes sont connus dans l'ontologie, leurs termes associés sont proposés. L'utilisateur peut alors choisir de lancer une requête en combinant termes de sa requête avec ceux suggérés.

Fig. 12 : Recherche d'informations assistée par les concepts

L'ontologie est développée et maintenue par (AO : administrateur de l’ontologie) (*figure. 12*). Le rôle principal de l'ontologie est l'expansion de requête par suggestion à l'utilisateur de termes plus ou moins liés à ceux de sa requête. Notre ontologie est donc un outil complémentaire d'un moteur de recherche. Notre principal travail est de proposé une méthode de construction de cette ontologie à partir des documentations techniques.

4.1 Mise en œuvre de logiciel

L'interface de notre ontologie documentaire doit être comporte deux aspects : la visualisation et navigation dans le réseau de termes du domaine qu'elle constitue, et son interaction avec les requêtes d'utilisateurs d'un moteur de recherche. Le fait de permettre à l'utilisateur de visualiser et parcourir le graphe des termes qu'est notre ontologie est à notre avis, fondamental. Cela permet d'appréhender le champ des possibles, les termes de domaine et leurs liens. Cette visualisation donne un point de vue sur le domaine et peut conférer un aspect pédagogique à l'ontologie. On propose par exemple que la navigation dans l'ontologie se fait par le biais de liens hypertexte. Chaque terme est cliquable, laissant apparaître la liste des termes qui lui sont associés.
La principale tâche à laquelle est destinée notre ontologie est l'expansion de requête par suggestion à l'utilisateur de termes plus ou moins liés à ceux de sa requête et même, elle est même utilisée comme index thématique au sein duquel l’utilisateur peut naviguer pour définir ou préciser sa requête. Notre ontologie est un adjoint de moteur de recherche. Ce procédé permet à l'utilisateur soit de préciser sa requête par un terme plus précis soit de la généraliser à un terme plus général. La liste des termes suggérer peut également permettre à l'utilisateur de combiner dans une même requête plusieurs termes proches.

4.2 Validation

L’étape de l’évaluation et de validation d’une ressource terminologique ne peut se concevoir, que par l’usage. La validation de notre ontologie doit donc se faire au regard des taches qui lui sont attribuées, principalement l’aide à la reformulation de requête.

[1] On prend cet exemple à partir de la thèse de doctorat de H. Assadi (1998).

CONCLUSION ET PERSPECTIVES

Nous avons abordé dans ce mémoire la problématique de la construction d'ontologie à partir des textes techniques. Les ontologies constituent l'un des ingrédients essentiels de la gestion des connaissances. Elles sont utilisées dans un grand nombre d'applications informatiques pour faciliter la diffusion, le partage, la réutilisation et la conservation des connaissances. Les ontologies sont parfois aussi utilisées dans des systèmes de recherche d'information. Elle est le résultat d'un consensus entre les personnes ayant participé à sa construction. Le rôle de l'ontologie n'est pas de normaliser un domaine, mais de donner une représentation de l'existant. Les concepts de l'ontologie sont organisés en une hiérarchie générique/spécifique avec héritage simple, donc avec une structure arborescente. La principale différence entre la terminologie et l'ontologie réside dans le fait que, dans une terminologie, les définitions des concepts sont rédigées en langue naturelle, tandis que dans une ontologie, les définitions sont le plus souvent formalisées de manière à permettre un certain nombre de traitement automatique.

Notre méthode de construction d'ontologie représente une modélisation d'un domaine sous forme d'une ontologie. Cette méthode s'applique directement à des corpus de textes, sans que ceux-ci nécessitent un étiquetage préalable, et sans utilisation de gros dictionnaires de langue ou du domaine. La partie de l'extraction et la classification des termes peut apporter une aide précieuse à un terminologue en lui permettant d'obtenir de manière automatique et semi-automatique une partie non négligeable de la terminologie du domaine. Les ressources linguistiques utilisées sont limitées à des filtres linguistiques et des règles de déduction contextuelle, l'utilisation de ces règles a l'avantage de fournir des segments relativement homogènes et la structuration des termes permet d'inférer d'autre termes et même des termes simples. Mais le problème reste toujours de la nécessité de trouver une méthodologie pour la détection des termes simples.

La partie de la normalisation représente la mise en œuvre d'un principe d'analyse distributionnelle de Harris. Elle permet de regrouper les termes ayant le même contexte syntaxique partagé.

Notre ontologie a été utilisée pour l'expansion de la requête utilisateur. Le public visé par l'application sera en premier temps les personnes de domaines. L'aspect pédagogique de notre ontologie permet aux novices d'appréhender un nouveau domaine. Le corpus dans notre cas est la principale source de connaissances.

En ce qui concerne nos perspectives, nous prévoyons :

- d'implémenter notre système et dérouler un cas réel avec un corpus complet.
- L'un des problèmes posés est de trouver une méthodologie pour l'extraction des termes simples, pour cette raison nous proposons la règle suivante : *le sujet qui se répète peut représenter un terme de domaine.*
- Pour l'étude de similarité entre les termes simples nous proposons la règle suivante : *les mots simples qui partagent les mêmes verbes peuvent être un indice qui indique l'existence des proximités sémantiques entre ces termes.*
- On peut généraliser la règle précédente sur les termes complexes.
- Repérage des termes nouveaux : cette méthode s'appelle ***SEARCH-EXPANSIONS***. On suppose que T(A,B) désigne un terme qui commence par un lemme A et se termine par un lemme B. on procède comme suit: en partant de mot A, on cherche à trouver dans une fenêtre variable (maximum 10 mots) la plus longue expansion X du terme AX. La recherche est effectuée de gauche à droite en utilisant les contraintes suivantes : la chaîne X ne se termine pas par un mot des filtres utilisés pour le calcul des segments répétés (filtre des mots grammaticaux, etc.). Mais X peut contenir des mots du filtre de dérivation (à, au, avec, de, des, du, et, d', pour). Le processus est la même pour B sauf que la recherche est effectuer de droite à gauche. La liste des nouveaux termes est traitée à son tour de la même façon que l'ancienne, afin de rechercher d'autres termes nouveaux, et ainsi de suite. Le traitement s'arrête quand la liste nouvelle est vide. Par conséquent, on peut trouver des termes de domaine avec le nombre d'occurrence égale à 1.

-

Bibliographie

[ASSADI 1998] H.ASSADI Construction d'ontologie à partir de textes techniques- application aux systèmes documentaires. *Thèse de doctorat de l'université de Paris 6. octobre 1998.*

[ASSADI 2000] H.ASSADI et D. BOURIGAULT analyse syntaxique et statistique pour la construction d'ontologies à partir de textes. *Ingénierie des connaissances évolutions récentes et nouveaux défis. Edition Eyrolles.*

[Alda 2000] ALDA MARI, PATRICK SAINT-DIZIER. Nature et formation de classes sémantiques de verbes pour l'extraction de connaissances dans des textes: Esquisse d'une approche statistico-symbolique. *JADT 2000: 5es Journées Internationales d'Analyse Statistique des Données Textuelles.*

[Antipolis 2001] SOPHIA ANTIPOLIS. Projet ACACIA Acquisition des Connaissances pour l'Assistance à la Conception par Interaction entre Agents. *Rapport d'activité IRIA 2001.*

[Aussenac-Gilles2000] N. Aussenac-Gilles, B. Biebow & S. Szulman Revisiting ontology design : a method based on corpus analysis, *in Proceedings of the conference EKAW'2000, Springer LNCS 1937, pages 172-188, 2000.*

[Aussenac-Gilles2001] Nathalie Aussenac-Gilles et Jean Charlet. Ingénierie des connaissances

[Aussenac-Gilles2002] Nathalie Aussenac-Gilles et Agnès Busnel. Méthode de construction à partir du texte d'une ontologie du domaine de l'industrie de la fibre de verre. *Rapport Interne IRIT/2002-11-R. Avril 2002.*

[BACHIMONT 2000] BACHIMONT B., Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances, in CHARLET J., ZACKLAD M., KASSEL G. & BOURIGAULT D., eds., *Ingénierie des connaissances : évolutions récentes et nouveaux défis*, Eyrolles, pages 305-323, 2000.

[Barry 2001] Barry, C., Cormier, C., Kassel, G. & Nobécourt, J. Evaluation de langages opérationnels de représentation d'ontologies. Actes des Journées Ingénierie des Connaissances : IC'2001, Grenoble.

[**BIEBOW 1999**] M. Biebow & S. Szulman, TERMINAE : a method and a tool to build of a domain ontology, in *Proceedings of the 11th European Knowledge Acquisition Workshop (EKAW'99)*, Springer, 1999.

[**Berland** 2002] SOPHIE BERLAND, NATALIA GRABAR. Assistance automatique pour l'homogénéisation d'un corpus Web de spécialité. *JADT 2002 : 6es Journées internationales d'Analyse statistique des Données Textuelles.*

[BERNERS-LEE 2001] BERNERS-LEE T., HENDLER J. & LASSILA O., The semantic Web, in *Scientific American*, mai 2001.

[BOURIGAULT 2000] D. BOURIGAULT, C. JACQUEMIN Construction de ressources terminologiques. *(Ingénierie des langues, Jean-Marie Pierrel (ed.), Hermès, 2000, pp 215-233).*

[Bourigault 2002] Didier Bourigault. UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *Equipe de Recherche en Syntaxe et Sémantique CNRS – Université Toulouse le Mirail Maison de la Recherche 5, allées Antonio Machado 31058 Toulouse Cedex 1. didier.bourigault@univ-tlse2.fr*

[Belaïd 1998] A. BELAÏD ET Y. TOUSSAINT. Une méthode d'étiquetage morpho-syntaxique pour la reconnaissance de tables de matières. *LORIA-CNRS, Campus Scientifique B.P. 239, 54506 Vandoeuvre-Lès-Nancy France {abelaid,toussaint}@loria.f*

[Charlet 2003] JEAN CHARLET. L'ingénierie des connaissances: développements, résultats et perspectives pour la gestion des connaissances médicales. *Mémoire d'habilitation à diriger des recherches présentées à l'Université Pierre et Marie Curie. Version finale du 28 janvier 2003.*

[Côté 1998] MARC COTE ET NADER TROUDI. NetSA : Une architecture multiagent pour la recherche sur Internet. *Université Laval Département d'informatique Pavillon Pouliot Ste-Foy, G1K 7P4, Canada. {mcote, troudi}@iad.ift.ulaval.ca*

[Corcho 2000] Corcho, O. & Gomez-Perez, A. Evaluating Knowledge Representation and Reasoning Capabilities of Ontology Specification Languages. Proceedings of the *ECAI-00 Workshop on Applications of Ontologies and Problem-Solving Methods*, Berlin, pp. 3/1-3/9.

[Daille 1994] BEARICE DAILLE. approche mixte pour l'extraction automatique de terminologie: statistique lexical et filtres linguistiques. *Thèse doctorat, janvier* 1994.

[Drouin 2002] Patrick Drouin. Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés. *Thèse présentée à la Faculté des études supérieures en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.) en linguistique. Mai 2002.*

[DELTEIL 2001] A. DELTEIL & C. FARON-ZUCKER, Extending RDF(S) with contextual and definitional knowledge, *in Proceedings of the Semantic Web Working Symposium, 2001*.

[DIENG 2001] DIENG R., CORBY O., GANDON F., GIBOIN A., GOLEBIOWSKA J., MATTA N. & RIBIÈRE M., Méthodes et outils pour la gestion des connaissances : une

approche pluridisciplinaire du knowledge managemene 2ième édition, *Dunod Edition Informatiques Séries Systèmes d'Information, 2001*

[Felber 1987] Felber H. L., Manuel De Terminologie, Unesco, Paris, 1987

[Falquet 2001] Gilles Falquet, Claire-Lise Mottaz Jiang ; Navigation hypertexte dans une ontologie multi-points de vue. *Centre universitaire d'informatique. Université de Genève. Rue duGénéral-Dufour 24. 1211 Genève 4, Suise.*

[Fen 2000] d. Fensel, I. Horrocks, f. Van harmelen, s. Decker, m. Erdmann & m. Klein, OIL in a nutshell, in Proceedings of European Knowledge Acquisition Workshop (EKAW'2000), *Springer-Verlag LNAI 1937, pages 1-16, 2000.*

[Fernandez 1997] M. Fernandez, A. Gomez-Perez & N. Juristo, METHONTOLOGY: from ontological art towards ontological engineering, *in Proceedings of the Spring Symposium Series on Ontological Engineering (AAAI'97), AAAI Press , 1997.*

[Fernandez-Lopez 1999] Fernandez-Lopez, M. Overview Of Methodologies For Building Ontologies. Proceedings of the IJCAI'99 *Workshop on Ontologies and Problem-Solving Methods*, Stockholm (Suède), pp. 4/1,4/13.

[Frédéric 2002] Frédéric Fürst. L'ingénierie ontologique. Rapport de recherche N° 02-07 Octobre 2002.

[Fabien 2001] Fabien Gandon, Rose Dieng-Kuntz. Ontologie pour un système multi-agents dédié à une mémoire d'entreprise. *INRIA - Projet ACACIA, 2004, route des Lucioles B.P. 93 - 06902 Sophia Antipolis Cedex {Fabien.Gandon|Rose.Dieng}@sophia.inria.fr*

[Fabien 2002] FABIEN GANDON, ROSE DIENG-KUNTZ, OLIVIER CORBY, ALAIN GIBOIN Web Sémantique et Approche Multi-Agents pour la Gestion d'une Mémoire Organisationnelle Distribuée. *INRIA - Projet ACACIA, 2004, route des Lucioles, B.P. 93, 06902 Sophia Antipolis, France.*

[FÜRST 2002] FREDERIC FÜRST, MICHEL LECLERE, FRANCKY TRICHET. Construction d'une ontologie opérationnelle : un retour d'expérience. *Institut de Recherche en Informatique de Nantes (IRIN)2, rue de la Houssinière - BP 92208 44322 Nantes Cedex 3*

[Fabre 2002] Cécile Fabre et Cécile Frérot. Groupes prépositionnels arguments ou circonstants : vers un repérage automatique en corpus. *ERSS, UMR 5610, Université Toulouse-Le Mirail cfabre@univ-tlse2.fr, cessfrerot@aol.com*

[Gaspar 2000] Gaspar. Linguistique et informatique. *Rapport de recherche.*

[GRUBER 1993] GRUBER T., A translation approach to portable ontology specifications, *Knowledge Acquisition* 5(2), pages 199-220, 1993.

[GRÜNINGER 1995] Grüninger, M. & Fox, M.S. Methodology for the Design and Evaluation of Ontologies. Proceedings of the IJCAI-95 Workshop on *Basic Ontological Issues in Knowledge Sharing*.

[GUARINO 1994] GUARINO N., CARRARA M. & GIARETTA P., Formalizing ontological commitments, in *Proceedings of the AAAI conférence (AAAI'94)*, 1994.

[GUARINO 2000] GUARINO N. & WELTY C., A Formal Ontology of Properties, in DIENG R. & CORBY O., eds., *Knowledge Engineering and Knowledge Management : Methods, Models and Tools. International ConferenceEKAW'2000*, Springer-Verlag, pages 97-112, 2000.

[Kassel 2002] G. Kassel, OntoSpec : une méthode de spécification semi-informelle d'ontologies, *in Actes des journées francophones d'Ingénierie des Connaissances (IC'2002), pages 75-87, 2002*.

[Ladjailia 2003] Ladjailia Ammar. Une méthode d'aide de la construction d'ontologie à partir de texte technique : application à l'expansion de la raquette d'utilisateur. *Journées des sciences technologies avancées 24-25/5/03 JSTA'2003 Guelma- Algérie.*

[Lemay] CHANTAL LEMAY. Évaluation de l'étiqueteur WinBrill sur des textes médicaux anglais et français. *Département de linguistique et de traduction Université de Montréal.*

[Linga 1996] Gérard Dimanche Linga s.a.r.l. Extraction Automatique de Terminologie: L'Outil d'Analyse de Corpus CORONA.

[Leclère 2002] Michel Leclère, Francky Trichet et Frederic Fürst. Construction d'une ontologie du domaine de la géométrie projective. *IRIN 2, rue de la Houssinière - BP 92208 44322 Nantes*

[Marie-Claude, 2002] MARIE-CLAUDE L'HOMME. Exploitation des corpus en terminologie.

[Marie-Claude, 2002] Marie-Claude L'Homme. Techniques d'extraction des données terminologiques et leur impact sur le travail du terminographe. *Université de Montréal.*

[MARTIN, 1996] Philippe MARTIN, exploitation de graphes conceptuels et de documents structurés et hypertextes pour l'acquisition de connaissances et la recherche d'informations. *Thèse doctorat soutenue le 14.10.96* .

[Muller, 1997] Chantal Muller, Xavier Polanco, Jean Royauté, Yannick Toussaint. Acquisition et structuration des connaissances en corpus : éléments méthodologiques. *Rapport de recherche N°= 3198, juin 1997.*

[OUESLATI 1999] Rochdi OUESLATI. Aide à l'acquisition de connaissances à partir de

corpus. *Thèse de doctorat de L'université Louis Pasteur. Le 7 juillet 1999.*

[Pierra 2002] GUY PIERRA. Un modèle formel d'ontologie pour l'ingénierie, le commerce électronique et le Web sémantique:Le modèle de dictionnaire sémantique PLIB. *Laboratoire d'Informatique Scientifique et Industrielle, EA 1232 E.N.S.M.A. 86961 Futuroscope Cedex.*

[Rennes 2000] RENNES. Modélisation et Apprentissage pour l'Interprétation de Données et l'Aide à la décision. *Rapport d'activité INRIA 2000.*

[Rajman 2002] MARTIN RAJMAN ET JEAN-CEDRIC CHAPPELIER. Traitement Informatique des Données Textuelles: étiquetage morpho-syntaxique *Martin.Rajman@epfl.ch et Jean-Cedric.Chappelier@epfl.ch Laboratoire d'Intelligence Artificielle.*

[Rastier 1995] François Rastier **,** le terme : entre ontologie et linguistique, *Centre de Linguistique Française Université de Paris-Sorbonne. (Texte publié dans La banque des mots , n°7, pp. 35-65, 1995)*

[Séguéla 2001] M. Patrick Séguéla. Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques. *Thèse doctorat, 22 mars 2001.*

[Sébastien 2000] Sébastien Perpette. Définition d'une méthode de construction d'ontologies: application à la gestion des connaissances d'une équipe de recherche. *Rapport DEA, janvier 2000.*

[Sereno 2000] Bertrand Sereno, Annie Corbel et Jean-Jacques Girardot ; Le projet COSI : recherche d'information assistée par les concepts. *Département RIM, Ecole Nationale Supérieure des Mines de Sainte-Etienne, 158, cours Fauriel, F-42023 Sainte-Etienne cedex 2. {sereno, corbel, girardot}@emse.fr*

[Valli 1999] ANDRE VALLI & JEAN VERONIS. Étiquetage grammatical des corpus de parole: problèmes et perspectives. *Université de Provence.*

[Vergne 2002] JACQUES VERGNE. Entre arbre de dépendance et ordre linéaire, les deux processus de transformation : linéarisation, puis reconstruction de l'arbre. *Cahiers de Grammaire nº 23. Université de Caen, F-14032 Caen cedex courrier électronique : Jacques.Vergne@info.unicaen.fr*

[Vergne 1998] Jacques Vergne et Emmanuel Giguet. Regards Théoriques sur le "Tagging". *Jacques.Vergne@info.unicaen.fr, Emmanuel.Giguet@info.unicaen.fr*
GREYC - CNRS UPRESA 6072 - Université de Caen F-14032 Caen cedex France

[Vergne 1999] Jacques Vergne. Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur: Analyse syntaxique automatique non combinatoire; Synthèse et Résultats. *Dossier de synthèse des activités de recherches en vue de l'obtention du diplôme d'Habilitation à Diriger des Recherches présenté le 29 septembre 1999.*

[USCHOLD 1995] Uschold, M. & King, M. Towards a Methodology for Building Ontologies. Proceedings of the IJCAI-95 Worshop on *Basic Ontological Issues in Knowledge Sharing*.

[USCHOLD 1996] USCHOLD M., Building ontologies : towards a unified methodolgy, *in Proceedings of the 16th conference of the British Computer Society Specialist Group on Expert Systems, 1996.*

[Zweigenbaum 2000] Pierre Zweigenbaum et Natalia Grabar. Liens morphologiques et structuration de terminologie. *Service d'Informatique Médicale, Direction des Systèmes d'Information, Assistance Publique - Hôpitaux de Paris & UPRES EA 1528, Département de Biomathématiques, Université Paris 6.*