UQÀM
Université du Québec à Montréal

DIC9410 Présentation du projet de recherche

Doctorat en Informatique Cognitive

Titre :

Maintenance d'ontologies de domaine à partir d'analyses textuelles

YASSINE GARGOURI
GARY31087407
gargouri.yassine@courrier.uqam.ca

Codirecteurs de Recherche :

Mr Bernard Lefebvre & **Mr Jean-Guy Meunier**
lefebvre.bernard@uqam.ca meunier.jean-guy@uqam.ca

Décembre, 2003

Table des matières

Résumé

Les ontologies sont des nouvelles formes de contrôle intelligent de l'information. Elles présentent un savoir préalable requis pour un traitement systématique de l'information à des fins de navigation, de rappel, de précision, etc. Toutefois, les ontologies sont confrontées de façon continue à un problème d'évolution. Étant donné la complexité des changements à apporter, un processus, du moins semi-automatique, de maintenance s'impose de plus en plus pour faciliter cette tâche et assurer sa fiabilité.

Nous mettons les textes au centre du processus d'ingénierie des connaissances et présentons une approche se démarquant des techniques formelles classiques en représentation de connaissances par son indépendance de la langue. Le modèle proposé, représentera une chaîne de traitement (ONTOLOGICO) au sein de la plate-forme SATIM. Cet outil vise à assister les experts de domaine dans leur tâche de maintenance des ontologies en se basant sur un processus itératif supporté par un ensemble de modules, en particulier ; un extracteur de termes simples, un extracteur de termes complexes, un lemmatiseur, un segmenteur, un classifieur, un thésaurus du domaine, un module de raffinement sémantique (basé sur l'Indexation Sémantique Latente) et un identificateur de termes reliés (basé sur le calcul de similarité sémantique entre les couples de vecteurs conceptuels).

La méthodologie proposée constitue une aide précieuse dans le domaine de la maintenance des ontologies. Elle assiste les terminologues chargés de naviguer à travers de vastes données textuelles pour extraire et normaliser la terminologie. Elle facilite également la tâche des ingénieurs en connaissances chargés de modéliser des domaines.

Ce projet de recherche se place au cœur des échanges entre terminologie et acquisition de connaissances. Il amène une réflexion sur les divers paliers à envisager dans une telle démarche de modélisation de connaissances textuelles pour des objectifs de maintenance d'une ontologie de domaine.

Mots clés : *Maintenance des ontologies, représentation de connaissances, traitement de données textuelles, Traitement du Langage Naturel, cooccurrence, repérage de termes reliés, psychologie cognitive, classification textuelle, Indexation Sémantique Latente, thésaurus, vecteurs conceptuels.*

1. Introduction

La maintenance des ontologies est un champs multidisciplinaire impliquant le traitement du langage naturel, la prospection de données, l'apprentissage machine et la représentation de connaissances. De ce fait, il est irréaliste de s'attendre de l'humain à ce qu'il comprenne la totalité de l'ontologie et de ses interdépendances internes. Il lui est difficile, voir même impossible, de repérer de nouvelles relations entre termes à partir de la simple lecture de données textuelles et d'évaluer leur pertinence par rapport à l'ontologie actuelle. Cette tâche cognitive est d'autant plus ardue que les nouveaux textes à analyser et l'ontologie existante sont de larges tailles. Ce problème se pose également lorsque la conceptualisation du domaine est ambiguë ou encore si l'utilisateur ne possède pas suffisamment d'expérience.

Au cours de la dernière décennie, de nombreuses recherches ont été réalisées dans le domaine de l'ingénierie des ontologies. La majorité de ces recherches se sont concentrées sur les problèmes de construction. Toutefois, il n'existe pas, jusqu'à date, de méthodes consensuelles et de lignes de conduite répondant à cette problématique. L'absence de telles méthodes consensuelles entrave l'extension d'une ontologie donnée à partir d'autres et sa réutilisation dans d'autres ontologies et dans des applications finales.

Notre objectif de recherche consiste à mettre en place une « passerelle » entre les documents (lexiques, réseaux sémantiques…) et l'ontologie courante. Ceci revient à proposer un modèle permettant, à partir de l'analyse de nouveaux textes, d'identifier les nouveaux concepts spécifiques à un domaine ainsi que leurs relations, et d'automatiser certaines tâches relatives à leur intégration au niveau de l'ontologie du domaine.

Le présent document est articulé autour de quatre principaux chapitres. En guise d'introduction, nous présentons notre problématique de recherche et les objectifs généraux de notre projet. Nous détaillons par la suite, notre problématique, en partant d'une revue de l'état de l'art sur la question de la maintenance des ontologies, les approches d'extraction de termes à partir de textes, ainsi que celles relatives au repérage de relations entre termes. Ensuite, nous identifions les aspects cognitif et informatique de cette problématique et formulons les principales hypothèses de notre modèle. Le troisième chapitre est consacré à la présentation de notre proposition de solution et la méthodologie que nous envisageons de suivre. Nous concluons enfin, par l'énumération des différentes contributions originales que

notre projet ambitionne d'apporter sur le plan scientifique, ainsi que les défis et les obstacles auxquels nous sommes confrontés.

1.1 Mise en contexte

Dans l'objectif de présenter le contexte général de nos recherches, nous commençons tout d'abord par exposer le problème de recherche d'information et mettre en évidence « l'imperfection » et l'insuffisance des outils classiques présentement utilisés, à savoir les moteurs de recherches. Nous introduisons par la suite la notion d' «*ontologie*» comme remède à cette problématique, et exposons le problème d'évolution qui touche les ontologies. Cet aspect d'évolution, impliquant la nécessité de maintenir les ontologies, constitue le centre d'intérêt de nos recherches.

Bien que les moteurs de recherche jouent un rôle important dans l'appariement de documents à des requêtes spécifiques, ils sont considérés comme des outils restreints et limités. En raison de l'ampleur et de la complexité de la documentation dans les domaines spécialisés, la pertinence de ces outils a été remise en cause. En effet, la recherche dans les documents s'effectue habituellement par le biais de mots clés, mais le seul critère des moteurs de recherche demeure la présence des mots dans le texte. En d'autres termes, l'ordre de présentation des résultats, dépend de la proximité des mots recherchés dans le texte trouvé. Moins l'écart entre les mots est grand, plus l'ordre de présentation sera élevé (présenté en premier). Les documents n'étant pas spécifiquement annotés pour identifier clairement leur contenu, les résultats de recherche retournent dans la plupart des cas une foule de documents n'ayant aucun rapport avec les besoins et les recherches sont systématiquement imprécises.

Dans un domaine en perpétuel changement comme les télécommunications, ce problème devient sérieux. Par conséquent, les annotations de documents se présentent comme une solution incontournable, améliorant ainsi les résultats de recherche et permettant à des applications de manipuler, extraire et réutiliser cette information.

L'annotation des documents revient à utiliser un langage (tel que DAML+OIL, OWL, RDF, etc.) pour rattacher des méta-données (des explications, des commentaires, etc.) à un document Web ou encore pour représenter son contenu sémantique en se basant sur l'ontologie du domaine.

Étant donné que le concept « ontologie » est un terme clé dans notre problématique, il est important de rappeler les principales définitions qui ont été présentées. L'ontologie a été définie chez la communauté de l'Ingénierie des Connaissances comme étant une compréhension commune et partagée d'un domaine qui peut être communiquée entre des personnes et des systèmes (Guarino, 1995). Chez la communauté de représentation des connaissances, la définition d'ontologie la plus utilisée et fortement citée est celle de Gruber (1993); «*une ontologie est une spécification formelle et explicite d'une conceptualisation partagée*». Cette spécification représente un modèle abstrait d'un phénomène du monde réel qui est défini par des concepts et des relations. Le principe général de l'ontologie est en fait analogue à celui des bases de données dans la mesure où elle regroupe le vocabulaire d'un domaine en différentes classes (termes) et relie ces classes par le biais de relations.

Partant de ces définitions, les connaissances intégrées dans les ontologies peuvent être formalisées en mettant en jeu cinq types de composants : les classes, les relations, les fonctions, les axiomes et les instances (Gruber, 1993).

- **Les classes** sont habituellement organisées en taxonomies. Elles réfèrent à des concepts, utilisés dans le sens large. Ces concepts peuvent être abstraits ou concrets, élémentaires (électron) ou composés (atome), réels ou fictifs.
- **Les relations** (R) représentent un type d'interaction entre les notions d'un domaine (C_i). Elles sont formellement définies comme tout sous-ensemble d'un produit de n ensembles, c'est-à-dire $R : C_2 \times C_2 \times \ldots \times C_n$. Par exemple, les relations binaires sont du type «*sous-classe-de*», «*connecté-à*», etc.
- **Les fonctions** sont des cas particuliers de relations dans lesquelles le nième élément de la relation est défini de manière unique à partir des n-1 premiers. Formellement, les fonctions (F) sont définies ainsi : $F : C_2 \times C_2 \times \ldots \times C_{n-1} \rightarrow C_n$.
- **Les axiomes** sont utilisés pour structurer des phrases qui sont toujours vraies.
- **Les instances** sont utilisées pour représenter des éléments selon un principe similaire à la relation classe/objet en UML.

1.2 Problématique de recherche

Les ontologies constituent certes, la solution à adopter pour assister la recherche d'information, leur utilisation pratique est néanmoins confrontée à un problème d'évolution. Les documents relatifs à un domaine particulier changent et évoluent de façon perpétuelle. Par conséquent, les ontologies sont souvent sujet à changement, parce que des incomplétudes

ou des erreurs se sont révélées dans les versions précédentes, une nouvelle façon de modélisation du domaine a été préférée ou encore le domaine a changé. Les ontologies doivent supporter ces révisions et faire en sorte que les nouveaux documents relatifs à une nouvelle version d'ontologie soient compatibles à ceux qui précèdent.

De plus, il est important de penser à une façon de faire pour automatiser ce processus ou du moins, minimiser le traitement manuel relatif à la maintenance. En effet, il est irréaliste de s'attendre de l'humain à ce qu'il comprenne la totalité de l'ontologie et de ses interdépendances internes (Tallis & al., 1999). Il lui est difficile, voir même impossible, de repérer de nouvelles relations entre termes à partir de la simple lecture de données textuelles et d'évaluer leur pertinence par rapport à l'ontologie actuelle. Cette tâche cognitive est d'autant plus ardue que les nouveaux textes à analyser et l'ontologie existante sont de larges tailles. Ce problème se pose également lorsque la conceptualisation du domaine est ambiguë ou encore si l'utilisateur ne possède pas suffisamment d'expérience.

Étant donné la difficulté de maintenir les ontologies disponibles, celles-ci sont difficilement réutilisables et partageables, même lorsqu'elles sont exprimées selon le même formalisme et couvrent le même domaine. Cette contrainte justifie d'ailleurs le recours quasi systématique des ingénieurs de connaissances à la construction d'une ontologie de domaine à partir de zéro. En effet, les efforts manuels nécessaires pour réutiliser une ontologie existante et la faire maintenir seraient beaucoup plus coûteux.

Au cours de la dernière décennie, beaucoup de recherches ont été réalisées dans le domaine de l'ingénierie des ontologies. La majorité de ces recherches se sont concentrées sur les problèmes de construction. Toutefois, la gestion des changements et les mécanismes de maintenance doivent être abordés différemment. Jusqu'à date, il n'existe pas de méthodes consensuelles et de lignes de conduite répondant à cette problématique.

La complexité de cette problématique se justifie principalement par le fait qu'en pratique, l'élaboration d'ontologies relève plus du savoir-faire que de l'ingénierie. C'est ainsi que, lors du processus de mise au point d'une ontologie, chaque équipe de développement suit habituellement ses propres principes, ses critères de conception et ses étapes d'élaboration. L'absence de méthodes consensuelles entrave, d'une part, le développement d'ontologies communes et acceptées par les équipes et entre elles, et d'autres part, l'extension d'une ontologie donnée à partir d'autres et sa réutilisation dans d'autres ontologies et dans des applications finales.

1.3 Objectifs généraux du projet

Nos objectifs de recherches consistent à mettre en place une « passerelle » entre les documents (lexiques, réseaux sémantiques…) et l'ontologie courante. Ceci revient à proposer un modèle permettant, à partir de l'analyse de nouveaux textes, d'identifier les nouveaux concepts spécifiques à un domaine ainsi que leurs relations, et d'automatiser certaines tâches relatives à leur intégration au niveau de l'ontologie du domaine. La maintenance est ici vue comme la mise à jour incrémentale de l'ontologie au fur et à mesure que de nouveaux concepts sont extraits de textes du domaine. En général, notre objectif consistera à assurer un raffinement continu de l'ontologie, maintenant dans l'esprit la consistance et la cohérence de celle-ci et de ses artefacts.

Plus spécifiquement, nos objectifs de recherches s'orienteront vers les points suivants :

- Identifier de nouveaux concepts clés, à partir d'analyses de textes.
- Découvrir des relations conceptuelles entre les concepts. Tant les nouveaux concepts que ceux déjà figurant dans l'ontologie courante sont visés par cet objectif.
- Préciser, sur la base d'un ensemble d'hypothèses, la place d'un nouveau concept dans l'ontologie.
- Enrichir l'ontologie par des termes particulièrement spécifiques au domaine. Cet objectif revient à éviter d'alourdir l'ontologie par des termes non pertinents par rapport au domaine.
- Enrichir les connaissances extraites à partir de données textuelles par d'autres sources, provenant notamment de données terminologiques (thésaurus, dictionnaire électronique, base terminologique, etc.), en vue de faire intégrer, le mieux que possible, les connaissances du domaine au sein de l'ontologie.
- Assurer une certaine continuité entre le processus de construction d'ontologie et celui de sa maintenance. La construction initiale d'une ontologie respecte des normes, des procédures d'élaboration, des règles et des outils de formalisation. Par conséquent, il est inconcevable de séparer ces deux processus (mise en place et maintenance). En effet, les actions de maintenance doivent être alignées vers les règles et normes régissant l'ontologie courante.
- En présence d'une action de maintenance, il existe généralement, plusieurs façons qui permettent de préserver la consistance d'une ontologie. Par exemple, si un concept à l'intérieur d'une hiérarchie est supprimé, tous les sous-concepts peuvent être, soit

supprimés, soit rattachés à d'autres concepts. Si les sous-concepts sont préservés, alors les propriétés du concept supprimé peuvent être propagées, leurs instances peuvent être distribuées,… Ainsi, pour chaque changement dans l'ontologie, il est possible de générer différents scénarios de changements additionnels, conduisant à divers état finaux cohérents. La plupart des systèmes actuels de développement d'ontologies fournissent une seule possibilité de changement qui est généralement la plus simple (Staab, 2001). Notre objectif à cet effet sera d'assister le cogniticien, chargé de la maintenance, dans la génération de différents états plausibles de changements et de l'orienter dans son choix.

- Préserver un certain de degré de flexibilité du modèle proposé pour supporter toute ontologie de domaine. Cet objectif vise en d'autres termes à éviter la dépendance du modèle par rapport à un domaine particulier.
- Gérer les erreurs potentielles dans l'ontologie existante ; Cette gestion constitue l'une des préoccupations de base lors du processus de maintenance. Une prolifération excessive de l'ontologie entraîne l'accroissement de la complexité de la gestion des changements et empêche par conséquent, l'expert du domaine de parcourir efficacement les différents concepts et de procéder manuellement aux corrections. Par conséquent, il est important d'explorer d'autres alternatives plus efficaces et plus efficientes pour assister l'utilisateur lors de sa gestion des erreurs.
- Garantir une performance raisonnable en terme de temps d'exécution du système assistant à la maintenance.

2. Problématique

Nous détaillons dans ce chapitre, notre problématique, en partant d'une revue de l'état de l'art sur la question de la maintenance des ontologies, les approches d'extraction de termes à partir de textes, ainsi que celles relatives au repérage de relations entre termes. Nous identifions par la suite, les aspects cognitif et informatique de cette problématique et formulons les principales hypothèses de notre modèle.

2.1 État de l'art

La problématique de la maintenance des ontologies de domaine implique deux sous-problèmes fondamentaux ; Le premier est relatif à l'extraction de termes spécifiques au domaine et pertinents à l'ontologie existante. Le second consiste à identifier des relations entre termes. Une variété de techniques de traitement du langage naturel, d'extraction

d'information, d'apprentissage machine et d'analyse textuelle sont utilisées pour extraire des termes à partir d'un corpus. Il s'agit de techniques suffisamment matures et présentant des résultats prometteurs pour le domaine de la construction d'ontologie. Cependant, l'extraction de relations entre termes est un problème plus complexe et difficile à résoudre. Ceci constitue d'ailleurs la problématique de base qui affecte directement la performance des techniques d'apprentissage d'ontologie.

Après un aperçu des approches de maintenance qui ont été proposées, et en particulier celles fondées sur des données textuelles, nous présentons l'état de l'art des techniques d'extraction de termes et de relations entre termes à partir de textes.

2.1.1 Différentes approches pour la maintenance

La maintenance des ontologies implique principalement un processus d'« *apprentissage d'ontologie* ». Ce processus met en commun plusieurs activités de recherche traitant certes, différents types d'intrants, mais visant un même objectif, celui de la conceptualisation du domaine. Il s'agit en effet, d'un champs multidisciplinaire impliquant le traitement du langage naturel, la prospection de données, l'apprentissage machine et la représentation de connaissances.

La plupart des recherches relatives à l'apprentissage des ontologies n'ont toujours pas atteint un stade de maturité satisfaisant et impliquent inévitablement une intervention plus ou moins importante d'un expert du domaine. Nous proposons dans cette section de passer en revue les approches les plus importantes présentées dans la littérature. En se basant sur le type de ressources (intrants) utilisées par l'apprentissage, Alexander Maedche et Steffen Staab (Maedche and Staab, 2001) proposent la classification suivante : apprentissage d'ontologie à partir de texte, de dictionnaire, de base de connaissances, de schémas semi-structurés et de schémas relationnels.

2.1.1.1 Apprentissage d'ontologie à partir de texte

Ce type d'apprentissage a été largement employé chez la communauté de l'ingénierie de connaissances. Il s'agit en particulier des travaux de : (Aguirre *et al.*, 2000), (Alfonseca E. and Manandhar S., 2002a, 2002b), (Aussenac-Gilles et al., 2000a, 2000b), (Bachimont et al., 2002), (Faatz and Steinmetz, 2002), (Gupta et al., 2002), (Hahn and Markó, 2001), (Hearst, 1998), (Hwang, 1999), (Khan and Luo, 2002), (Kietz et al., 2000), (Lonsdale et al., 2002), (Missikoff *et al*., 2002), (Moldovan and Girju, 2001), (Nobécourt, 2000), (Roux et al., 2000),

(Wagner, 2000) et (Xu et al., 2002). Toutefois, aucune méthodologie suffisamment détaillée n'a été présentée pour assister le processus d'apprentissage d'ontologie. En effet, la littérature se limite à la présentation de lignes de conduite plus ou moins générales.

Ces méthodes sont principalement fondées sur des techniques d'analyse du langage naturel. Elles utilisent un corpus à travers différentes étapes du processus. Seuls les travaux de Maedche (Maedche et al., 2001) utilisent tant des corpus généraux que ceux du domaine pour écarter les concepts non spécifiques au domaine de l'ontologie existante. Les autres travaux traitent uniquement des documents relatifs au domaine en vue d'apprendre de nouveaux concepts et de nouvelles relations.

Selon le point de vue technologique, les outils développés dans le cadre de telles approches peuvent être regroupés sous trois catégories principales, dépendamment de la technique d'apprentissage adoptée :

- les outils basés sur le « clustering » conceptuel, tels que ASIUM (Faure and Nedellec, 1999), MO'K (Bisson et al., 2000), SVETLAN (Chaelandar and Grau, 2000), TERMINAE (Biébow et al., 1999);

- les outils basés sur des approches statistiques, tels que LTG (Mikheev and Finch, 1997), Text-To-Onto (Maedche and Volz, 2001), TFIDF (Xu *et al*., 2002), WOLFIE (Thompson and Mooney, 1997), SubWordNet (Gupta et al., 2002), KEA (Jones and Paynter, 2002);

- les outils basés sur des approches linguistiques et/ou sémantiques, tels que Prométhée (Morin, 1998 and 1999), Corporum-Ontobuilder (Engels R, 2001a, 2001b), TextStorm (Pereira, 1998), Welkin (Alfonseca and Rodríguez, 2002), OntoLearn (Velardi et al., 2002), DOE (Bachimont B., 2000), SOAT (Wu and Hsu, 2002).

Aucun de ces outils n'est complètement automatique. Certains sont orientés vers l'assistance à l'acquisition de connaissances lexico-sémantiques, d'autres visent à repérer des concepts ou des relations à partir d'un corpus prétraité, avec l'aide de l'utilisateur, etc. Par ailleurs, compte tenu de la complexité du processus d'apprentissage, l'évaluation de l'exactitude de ces outils a toujours fait défaut. De plus, aucune comparaison de résultats obtenus en utilisant différentes techniques d'apprentissage n'a été proposée.

2.1.1.2 Apprentissage d'ontologie à partir de dictionnaire

L'apprentissage d'ontologie a également fait usage, dans certains travaux, de dictionnaires électroniques. La performance de ces méthodes ((Hearst, 1992) ; (Jannink and Wiederhold, 1999) et (Rigau G., 1998)) est basée sur l'utilisation d'analyses sémantiques et linguistiques pour extraire de nouveaux concepts ou des relations à partir de dictionnaires. La plupart de ces méthodes utilisent WordNet comme ontologie initiale pour l'enrichir avec de nouveaux concepts ou de nouvelles relations. Les outils, tels que SEISD (Rigau, 1998) et DODDLE (Yamaguchi, 1999), mettant en application de telles techniques, procèdent principalement à des analyses syntaxiques. Ils nécessitent également l'intervention de l'utilisateur pour valider leurs résultats.

2.1.1.3 Apprentissage d'ontologie à partir de bases de connaissances

Jusqu'à date, l'apprentissage d'ontologie à partir de bases de connaissances n'est pas suffisamment exploré par la communauté de construction d'ontologie. D'ailleurs, nous n'avons pas retrouvé des outils fondés sur une telle approche. Seuls Suryanto et Compton ont proposé une approche (Suryanto and Compton, 2001 and 2002) visant à générer une ontologie à partir de règles d'une base de connaissances.

2.1.1.4 Apprentissage d'ontologie à partir de schémas semi-structurés

Des connaissances ontologiques peuvent également être extraites à partir de ressources semi-structurées telles que (XML Schemas, RDF, DAML+OIL, OWL, etc.) en se basant sur des approches de « *clustering* » ou de « *reconnaissance de patterns* ».

(Deitel et al, 2001) ont proposé une approche permettant l'apprentissage d'ontologies à partir des annotations sémantiques RDF d'une base de documents du Web. La méthode consiste à apprendre une ontologie à partir de descriptions de ressources extraites du graphe RDF que constitue l'ensemble des annotations de ces ressources. Une hiérarchie de concepts est construite en générant systématiquement les généralisations les plus spécifiques de tous les regroupements possibles de ressources. Le résultat de l'apprentissage est évalué par un expert du domaine pour un objectif de validation.

Des outils, tels que « OntoBuilder » (Modica et al., 2001) ont été développés pour partir d'une ontologie existante et l'enrichir avec de nouveaux concepts repérés à partir de ressources semi-structrées.

2.1.1.5 Apprentissage d'ontologie à partir de schémas relationnels

Les schémas relationnels constituent également une source plausible pour extraire des connaissances ontologique et les inclure d'une façon manuelle dans une ontologie existante. En effet, certains travaux à l'instar de (Stojanovic et al., 2002), (Kashyap, 1999), (Rubin et al., 2002) et (Stojanovic et al., 2002) se basent sur l'hypothèse que les connaissances spécifiques à un domaine sont intégrés dans les données et les schémas de bases de données sélectionnées. Ils proposent de construire une ontologie à partir de schémas de bases de données relationnelles en suivant un processus d'appariement. Ce processus s'appuie sur un ensemble de règles pour migrer les éléments du modèle de la base de données (les relations, les attributs, les types d'attributs, les clés primaires, etc.) vers l'ontologie.

Toutefois, nous n'avons pas retrouvé dans la littérature des outils appropriés qui permettent de procéder à un tel apprentissage d'ontologies.

2.1.1.6 Conclusions

L'apprentissage d'ontologies est un processus fondamental au sein des activités de maintenance. Il facilite en effet, l'acquisition de connaissances pour enrichir une ontologie et réduit, par voie de conséquence, le temps nécessaire à cette tâche.

Cependant, il n'existe pas de solution intégrée qui permet de combiner différentes techniques d'apprentissage et des sources de connaissances hétérogènes.

2.1.2 Les approches d'extraction de termes à partir de textes

Différentes techniques sont aujourd'hui utilisées pour repérer des syntagmes susceptibles d'être des termes. Elles sont opérationnelles dans des logiciels comme *SATIM* (Biskri, I., Meunier, J.G. 2002), *NOMINO*[1] *(*David et Plante 1990*), LEXTER (Bourigault, 1996), etc.* Ces approches peuvent être regroupées en trois grandes catégories possibles: les approches structurelles basées sur l'utilisation de grammaires formelles; les approches non structurelles, telles que les approches statistiques et quantitatives qui sont de plus en plus utilisées grâce à la disponibilité de gros corpus en format électronique; et les approches mixtes associant analyses statistiques et méthodes structurelles.

[1] Nomino est appelé à l'origine Termino

2.1.2.1 Les approches structurelles

Il s'agit essentiellement des « *méthodes utilisant une grammaire* » ou encore des « *méthodes de surface* ».

- **Les méthodes utilisant une grammaire** : Les approches structurelles, utilisées pour les systèmes de traitement automatique de la langue, requièrent souvent des grammaires (par exemple des grammaires probabilistes, par règle, etc..) et parfois des lexiques ou des dictionnaires électroniques de la langue utilisée (Brill 1994). Ce type d'approches visent à dissocier les traitements à chaque niveau d'analyse et servent principalement de connaissances linguistiques pour chaque étape.

  L'approche classique utilisée pour l'analyse de textes consiste en une analyse morpho-lexicale suivie d'une analyse syntaxique (Sabah 1989). La première produit à partir de ressources lexicales exhaustives, une liste de «tokens» (des mots arbitraires). La seconde produit, pour chaque phrase, un ou plusieurs arbres syntaxiques.

  En raison de ses caractères sémiotique et linguistique, le traitement classique de l'information est habituellement linguistique. En effet, un texte est considéré comme étant une succession de phrases qui doivent faire l'objet d'analyseurs linguistiques. Cette approche semble être complètement naturelle dans la mesure où elle correspond, en théorie, au processus normal de lecture chez l'humain (Meunier, 1996). Cependant, un problème délicat concerne la théorie des textes. Les textes, sont-ils des phénomènes linguistiques ? La réponse dépend de la définition et de la compréhension privilégiées du concept « linguistique ». Si ce concept est strictement considéré en tant que « grammaire », alors un texte n'est pas un phénomène grammatical. Bien que certains auteurs le pensent (Pavel, 1976, Dijk, 1977), d'autres, tels que (Rastier and al., 1994; Meunier, 1996) refusent une telle vision.

  De point de vue technique, la difficulté majeure de ces méthodes est leur aspect combinatoire. En effet, chaque unité lexicale se voit affectée une étiquette et l'ensemble de ces étiquettes permettent d'attribuer à la phrase une structure syntaxique attendue, souvent codée sous forme d'une grammaire. De plus, les systèmes utilisant ces méthodes ne garantissent pas une information exacte pour des mots ou des séquences de mots inconnus (à moins de prévoir un traitement des exceptions). Par ailleurs, dans les

approches basées sur des grammaires, on ne reconnaît que les occurrences se trouvant sous une des formes explicitement attendues.

- **Les méthodes de surface** : D'autres approches structurelles reposent sur des méthodes dites « *de surface* ». Ces approches se caractérisent notamment par l'utilisation de patrons syntaxiques de reconnaissance. À titre d'exemple, Bourigault (1994) utilise des marqueurs[2] de frontières complétés par un étiquetage grammatical des mots du corpus afin d'acquérir des syntagmes susceptibles d'être des termes. Son outil d'extraction terminologique LEXTER utilise des bornes de syntagmes nominaux (SN) qui sont des mots appartenant, pour la plupart, aux catégories grammaticales suivantes : verbes, pronoms, déterminants et adverbes.

2.1.2.2 Les approches non structurelles

- **Les méthodes statistiques** :

Les méthodes statistiques sont souvent réalisées sur de gros corpus (étiquetés ou pas). Elles se caractérisent par l'utilisation de la notion de seuil. Cette notion est utilisée pour filtrer ou repérer les informations contenues dans le corpus, ce qui explique en fait la possibilité de perte d'information. Plusieurs méthodes statistiques peuvent être appliquées, parmi lesquelles on trouve celles utilisant la notion d'«*information mutuelle* » ou la notion de « *segments répétés* ».

Les méthodes utilisant la notion d'information mutuelle consistent à détecter des associations récurrentes de mots, par exemple des paires de mots, ayant une forte valeur d'association mutuelle dans une fenêtre de n mots. Parmi ces associations, on trouve entre autre, les termes composés d'un texte dont la fréquence est assez élevée pour qu'ils soient repérables.

La spécificité de telles méthodes est qu'elles utilisent des scores pour mesurer le poids d'association de deux mots. Ainsi, dans le cadre de l'extraction de ressources lexicales monolingues, Church et Hanks (1990) utilisent un score d'association fondé sur la notion d'information mutuelle. Il s'agit d'un score d'association de deux lemmes qui permet de

2 Un marqueur est une formule, opérationnelle ou non, d'éléments linguistiques qui est rattachée à une relation lexicale.

comparer la probabilité d'observer ces deux lemmes ensemble avec la probabilité de les observer séparément.

D'autres méthodes statistiques ont été proposées en utilisant **la notion de segments répétés**. Dans l'objectif de ne pas se limiter à l'extraction des récurrences de mots associés en paires, comme celles obtenues par l'information mutuelle, Lebart et Salem (1994) ont proposé d'étendre l'extraction à des suites de plus de deux mots et répétées dans le corpus. On parle alors d'extraction de *segments répétés* de textes. Cette méthode privilégie le voisinage des chaînes et met en évidence l'importance du contexte dans l'apparition d'une séquence de mots. Les données obtenues par cette méthode sont linguistiquement hétérogènes (elles contiennent par exemple des syntagmes nominaux, des syntagmes verbaux, des formes figées, etc..). Mais on trouve aussi, des morceaux de syntagmes nominaux plus au moins figés, ou simplement des fragments de texte. Ces données doivent donc être filtrées, voire retraitées afin d'obtenir des objets linguistiques homogènes.

- **Les méthodes quantitatives :**

Les «*méthodes quantitatives*» impliquent l'utilisation d'outils statistiques qui ne sont pas contraints par la taille d'un corpus ou par l'utilisation de la notion de seuil.

Tant les méthodes statistiques que quantitatives sont réalisées sur des données textuelles. Leur caractéristique principale est qu'elles requièrent moins de connaissances linguistiques initiales que les approches structurelles. De plus, les résultats obtenus par ces méthodes sont souvent hétérogènes et doivent donc être retraités pour isoler les objets homogènes sur le plan linguistique, tels que les noms composés, les groupes nominaux, etc..

Les méthodes statistiques et quantitatives sont désormais utilisées dans de nombreuses applications, par exemple la recherche d'associations récurrentes entres mots voisins (Zernik U. 1992), la recherche de patrons syntaxiques associés, utilisés pour un classement sémantique de lexèmes (Grefenstette G. 1992). Toutefois, ces méthodes se trouvent confrontées à certaines difficultés ; D'abord, elles sont contraintes par l'utilisation de gros corpus. D'ailleurs, Grefenstette montre qu'une taille minimale de corpus est à respecter et qu'il faut choisir la méthode statistique en fonction de la taille du corpus. Une seconde contrainte est relative aux seuils de fréquence; Si ces derniers sont bas, alors la liste risque de présenter des cooccurrents sans valeur générale, car liés à quelques spécificités du corpus. Si les seuils sont élevés, alors on risque de ne conserver que les cooccurrents les plus typiques,

et donc de perdre de l'information. Il faut donc augmenter ou abaisser les seuils en fonction des objectifs de l'application; mais est-on sûr que les seuils soient les mêmes quels que soient les mots à étudier ? Il n'y a donc probablement pas de seuil unique qui soit valable pour tous les mots d'un corpus.

2.1.2.3 Les approches mixtes

Les approches mixtes tentent de tirer profil, tant des méthodes statistiques qui sont multilingues et pouvant traiter de larges données textuelles, que des approches linguistiques qui sont capables de rendre compte de certains néologismes dans des domaines spécifiques.

Dans le cadre de cette approche, Daille (Daille 1994) propose une méthode mixte statistico-syntaxique qui consiste à repérer des candidats-termes à partir de schémas syntaxiques, puis à les filtrer à l'aide de méthodes statistiques. Elle utilise une mesure (appelée MI) de la variation des distances entre les mots simples d'un nom composé. Ce dernier a été obtenu au préalable par des automates décrivant des patrons syntaxiques de noms composés et opérant sur un texte déjà catégorisé et désambiguïsé. D'après l'auteur, cette mesure de la variation des distances entre les mots permet de rapprocher des formes syntaxiques différentes, composées des mêmes mots. Cette méthode statistique permet enfin de confirmer ou d'infirmer les séquences de mots étudiées en fonction des seuils choisis.

Daille procède à l'inverse de Smadja (Smadja, 1993) qui applique des contraintes syntaxiques à des cooccurrences repérées statistiquement ; Son outil, appelé ***Xtract***, utilise des étiquettes grammaticales et syntaxiques (obtenues suite à un étiquetage préalable) pour valider ou écarter des collocations.

Justeson et Katz (1995) ont également proposé un prototype, appelé ***TERMS***. Le principe qui régit consiste à utiliser des contraintes syntaxiques sous forme de schémas syntaxiques pour repérer des syntagmes à partir de textes étiquetés. Le critère de répétition est utilisé pour ne retenir que des schémas de composés, répétés et de longueur deux et trois.

2.1.2.4 Conclusion

Si l'ensemble de ces techniques s'assurent d'une reconnaissance effective de syntagmes susceptibles d'être des termes, elles extraient également de nombreux syntagmes qui ne correspondent pas à ce qu'il convient d'appeler un terme. En effet, ce dernier n'est pas défini par rapport à sa forme mais plutôt à sa fonction dans le domaine. Par contre, les principes

opérationnels structurels ou statistiques sélectionnent de nombreux candidats au statut de terme que l'expertise humaine n'aurait pas retenu (Béguin et al, 1997).

En fait, il n'existe pas de modèle structurel ou statistique capable de décider qu'un syntagme est un terme. Par conséquent, une validation manuelle par élagage de la liste de candidats termes extraits est une étape incontournable.

2.1.3 Les approches d'extraction de relations à partir de textes

Notre but est maintenant d'étudier différents travaux ayant envisagé la reconnaissance de relations entre des termes dans des corpus. Il s'agit donc de présenter et évaluer différentes méthodes dont les résultats permettent de relever des informations sémantiques explicitées dans des textes.

Les relations sémantiques entre termes sont essentiellement du type *généralisation - spécialisation*. Toutefois, d'autres types de relations peuvent faire associer des termes, telles que la *composition*, la *dépendance* et la *disjonction*. De telles relations véhiculent ainsi une sémantique plus riche pour décrire un domaine.

Pour des systèmes disposant d'un outil d'extraction de termes, l'acquisition de relations sémantiques entre termes se situe en aval de l'acquisition de termes. Cependant, ces systèmes d'extraction de terminologie ne se sont pas vraiment intéressés au problème de l'acquisition de relations sémantiques entre termes. En effet, la sémantique est encore une discipline complexe et il est difficile de pouvoir modéliser les mécanismes linguistiques et cognitifs auxquels elle fait appel. Il existe tout de même quelques systèmes qui traitent de ce problème qu'on peut classer en deux catégories :

- les systèmes qui effectuent le repérage de relations sémantiques à partir de règles préétablies et qui utilisent une liste de marqueurs morphosyntaxiques (verbes, prépositions, etc..).
- les systèmes qui, au contraire, effectuent le repérage de ces marqueurs morphosyntaxiques directement à partir de textes en utilisant des algorithmes de repérage de séquences de mots et en exploitant la liste des termes du domaine.

Un autre principe consiste à fouiller par des méthodes statistiques la distribution de classes de mots en corpus afin de proposer des relations entre ces mots, sans se soucier de la nature sémantique caractérisant ces relations.

Nous présentons brièvement dans ce qui suit ces trois types d'approches.

2.1.3.1 Analyse par règles pour l'extraction de connaissances

Cette approche consiste à utiliser une analyse du texte basée sur des règles décrivant des schémas de relations à rechercher dans le corpus. L'objectif est de repérer les contextes de ces relations dont les contraintes morphosyntaxiques sont décrites dans des règles déclaratives du type : (SI conditions ALORS conclusion). Cette démarche a été appliquée au système SEEK (Jouis, 1995) en se basant sur la grammaire applicative et cognitive de Desclés (1990).

En appliquant des règles morphosyntaxiques préétablies, il est possible de repérer des relations à partir de textes connaissant des indices et des marqueurs linguistiques. Cependant, (Jouis et al., 1997) ont montré que l'analyse d'un nouveau domaine, utilisant ces mêmes règles prédéfinies, peut se heurter à des difficultés, dans la mesure où l'expression des relations peut être décrite par d'autres indices ou d'autres types de marqueurs. En effet, pour décrire les concepts d'un domaine, la langue utilise des moyens d'expressions très variés et très riches tels que la synonymie, la métaphore, la paraphrase, l'introduction de néologismes, etc... Ces moyens d'expression peuvent contenir des ambiguïtés et leur évolution incessante ne permet pas de fixer des règles préétablies pour les décrire.

2.1.3.2 Extraction d'information en utilisant des «templates»

Parmi ces systèmes, on peut citer ***PALKA*** : «Parallel Automatic Linguistic knowledge Acquisition» (Kim & Moldovan 1995), dont le but est de faciliter la construction d'une base de connaissances de schémas syntactico-sémantiques, relatifs à des structures de relations entre termes.

Pour extraire ces structures de schémas, il faut d'abord remplir manuellement des structures prédéfinies appelées «*templates*» à partir du texte. Pour construire les structures, le système effectue à partir des templates des mises en correspondance entres les phrases et les « *frames* » existants.

2.1.3.3 Approches statistiques

La base théorique de ces approches repose sur l'hypothèse formulée par Zellig Harris (1968) ; On peut classer les divers sens d'un terme en fonction des constructions auxquelles

il participe. Des termes qui ont des distributions comparables ont souvent un élément de sens commun. Partant de cette hypothèse, une première série de travaux a étudié la distribution lexicale en corpus afin de proposer des hypothèses de relations entre ces mots. Pour l'anglais, certains travaux comme ceux de (Smadja, 1993) ont étudié les fréquences de cooccurrences de mots pour proposer des relations entre ces mots. Pour le français, certains auteurs, comme (Toussaint et al 97), ont étudié ces phénomènes de cooccurrences afin de former des regroupements de termes. Présentés à l'expert, ces regroupements permettent le repérage de relations de synonymie, d'hyponymie[3] ou encore de méronymie[4].

Bien que les approches statistiques sont suivies d'une interprétation humaine systématique, elles sont reconnues comme des méthodes robustes et ne sollicitant pas des connaissances préalables sur le domaine. Nous pensons qu'elles sont très pertinentes pour distinguer des classes d'usage de termes dans l'espoir de les organiser en systèmes structurés reflétant une organisation conceptuelle. Parmi les outils utilisant de telles approches, nous citons : ***TFIDF*** (Xu et al., 2002), ***SVETLAN*** (Chaelandar and Grau, 2000), ***Text-To-Onto*** (Maedche and Volz, 2001), ***WOLFIE*** (Thompson and Mooney, 1997), ***TERMINAE*** (Biébow and Szulman, 1999).

2.2 Composante cognitive

Ce projet de recherche s'appuie sur l'avancement des connaissances dans plusieurs disciplines inter-reliées; principalement l'apprentissage machine, l'intelligence artificielle, le traitement automatique du langage naturel, la psychologie cognitive, etc. Ces disciplines s'inscrivent à des degrés divers dans les sciences cognitives et informatiques.

Sur le plan cognitif, l'élaboration de systèmes d'assistance à la maintenance des ontologies nécessite l'intégration de processus de raisonnement et d'apprentissage, principalement dédiés à la découverte de relations conceptuelles entre termes à partir de données textuelles. Ainsi, nous ambitionnons proposer un modèle s'inspirant entre autres, des caractéristiques du cerveau humain en reproduisant certaines de ses fonctions.

La présente section sera consacrée à des discussions d'ordre cognitif, car nous pensons que certaines des difficultés du TALN (Traitement Automatique du Langage Naturel) ne sont pas

[3] L'hyponyme est le nom de l'élément d'un tout dont le sens est inclu dans le sens du nom du tout (exemple : *Chien, chat, âne...* sont des hyponymes de *animal)*

[4] Méronymie : Relation hiérarchique existant entre deux concepts ou deux signes linguistiques, dans laquelle le premier est une partie d'un tout que constitue le second.

dues à des causes contingentes comme la taille de la mémoire, la puissance des microprocesseurs ou la performance des algorithmes, mais bien à des conceptions théoriques sur le traitement de données textuelles. Par conséquent, notre approche sera articulée principalement sur des hypothèses et des fondements cognitifs que nous décrivons dans ce qui suit.

2.2.1 Analyse Sémantique :

L'extraction de connaissances à partir de textes ne peut généralement se passer d'une discipline telle que la sémantique qui a pour objet la description des significations propres aux langues et leur organisation théorique (Tamba-Mecz, 1994). Même si nous affirmons que l'analyse sémantique joue un rôle primordial en ingénierie de connaissances, notre objectif n'est pas la compréhension automatique et complète de la documentation. Il s'agit plutôt d'analyser la documentation dans le cadre d'une tâche bien déterminée, celle de l'identification de relations conceptuelles entre termes. Ainsi les outils de TAL sont vus comme des « *outils d'aide à l'analyse des textes* ». Ceci nous amène à privilégier les outils et techniques effectuant des analyses partielles, mais robustes et fonctionnant sur des données textuelles réelles.

Sur le plan sémiotique, aucune, parmi les techniques proposées dans la littérature, ne semble capturer toute la richesse sémantique encapsulée dans les données textuelles, du moins, pas aussi efficacement que font les processus cognitifs de l'être humain. En effet, les architectures supportant ces techniques, ont rarement été élaborées sur la base de comportements humains, affectant ainsi, d'une certaine manière, les hypothèses relatives à la représentation, l'organisation, l'utilisation et l'acquisition de connaissances à partir de textes. Ce constat nous a amené à explorer les fondements de la « *Psychologie Cognitive* » en vue de mettre en évidence son efficacité à résoudre le problème de la sémantique.

2.2.2 Psychologie Cognitive

La cognition est un ensemble de processus intellectuels, à travers lesquels l'information est obtenue, transformée, stockée, retrouvée et utilisée. L'approche à adopter pour la maintenance des ontologies devrait s'inspirer de façon étroite des mécanismes intellectuels de compréhension, de production et d'apprentissage chez l'être humain (psycholinguistique).

Les recherches en « *Psychologie Cognitive* » montrent que la plupart des mots sont assimilés par la lecture (Landauer et S.T. Dumais, 1997). Étant exposé à des textes, un apprenant tente,

tout le long de son processus de lecture, de raffiner graduellement la signification du mot en utilisant les occurrences conjointes de ce mot avec d'autres. Par exemple, en absence d'une définition explicite du mot « *micro-processus* », l'apprenant est en mesure, à travers la lecture de textes, d'acquérir la signification du mot parce que celle-ci est confirmée dans le contexte dans lequel ce mot apparaît avec d'autres, tels que « *carte* », « *ordinateur* », « *électronique* », « *matériel* », « *Unité Centrale de Traitement* », etc. Toutefois, une simple cooccurrence répétée d'un mot avec d'autres semble être insuffisante pour l'assimilation de sa signification. En effet, toutes les cooccurrences de tous les mots à travers le texte sont plutôt nécessaires.

En se basant sur ce fondement cognitif[5], nous supportons l'idée de l'application de la technique de classification de documents pour identifier des groupes de termes qui apparaissent ensemble et qui ont des relations sémantiques, ou du moins, des similarités sémantiques lorsque utilisés dans des contextes comparables. (Nous détaillons davantage les apports de la classification dans la section suivante).

2.2.3 Architecture Cognitive :

Notre orientation vers une architecture cognitive est motivée par un objectif de mise en place d'un système intelligent, supportant les potentialités de l'humain. Cet objectif se rapporte, d'une certaine manière, au paradigme fondateur du test de Turing[6]. Les architectures cognitives supportent également l'objectif central de l'intelligence artificielle et des sciences cognitives ; à savoir la création et la compréhension de comportements intelligents mettant en application les potentialités de l'humain.

Par ailleurs, un problème central, auquel les concepteurs d'architectures cognitives se trouvent confrontés, est de faire en sorte que les agents[7] puissent accéder à différentes sources de connaissances ; Par exemple, les connaissances relatives à l'environnement sont accaparées à travers la perception, les connaissances relatives aux implications de la situation courante proviennent sont saisies à travers le planning, le raisonnement et la prédiction, les connaissances produites par les agents sont échangées via la communication et les connaissances induites à partir des expériences passées sont acquises à travers la souvenance et l'apprentissage. Plus ces caractéristiques sont supportées par l'architecture, plus les

[5] Ce fondement cognitif est relatif à l'hypothèse (H7)
[6] Ce test consiste à faire dialoguer un humain et le système et déterminer si l'humain peut déceler si le système n'est pas humain.
[7] Ici on parle d'agents. Pour notre cas, il s'agit plutôt de « modules », mais le raisonnement est toujours valable.

sources de connaissances sont mises au profil de cette architecture en vue d'influencer son comportement.

Partant de ce principe, nous avons opté pour une architecture qui se base sur différentes sources de connaissances, à savoir les textes relatifs au domaine, les requêtes formulées par les utilisateurs lors de recherches documentaires, un thésaurus spécifique au domaine en question et bien entendu l'ontologie courante qui intègre progressivement les connaissances du domaine, au fur et à mesure que les actions de maintenance sont appliquées.

2.3 Composante informatique

Notre projet de recherche sera appuyé par une réalisation informatique mettant en application le modèle proposé. Cette réalisation constituera une réponse complète et opérationnelle à la problématique posée et se présentera comme un « *prototype d'atelier* » d'ingénierie des connaissances, regroupant les modules, déjà existants dans la plate-forme SATIM ainsi que d'autres modules à développer, et permettant une mise en œuvre efficace de la méthodologie.

Il est à noté que nos travaux de recherche rentrent dans le cadre du projet GDST (*Gestion et Diffusion de Savoir en Télécommunication*)[8]. L'objectif principal de ce projet est de mettre en place une gestion informatisée de documents pour une entreprise qui puisse permettre le développement des compétences des ressources humaines pour les besoins de l'entreprise. Le domaine pris en compte est celui des télécommunications et plus spécifiquement celui des télécommunications sans fil relatives aux réseaux informatiques. Ce domaine est récent, en évolution rapide et les documents qui se rattachent sont, pour l'essentiel, sous une forme directement exploitable électroniquement.

Ainsi, le but ultime du projet GDST est de concevoir un prototype d'application permettant à l'entreprise[9] de résoudre le problème de transmission et d'acquisition sur mesure de l'information relative aux progrès technologiques qui touchent son département de réseaux sans fils. Celui-ci devra permettre l'annotation automatique de documents (internes et externes à l'entreprise) et de rechercher intelligemment à travers la base documentaire (composée des documents annotés) de l'information en fonction du contexte d'utilisation d'un employé.

[8] Projet réalisé par une équipe de chercheurs à l'UQAM (Mr Bernard Lefebvre étant le coordinateur du projet).
[9] Bell Canada

La réalisation de cet objectif passe par une étape primordiale; celle de la construction d'une forme de représentation du contenu informationnel de documents pour une entreprise au moyen d'ontologies. La problématique de la maintenance des ontologies constitue donc une partie cruciale du projet. Ce dernier permettra d'offrir un cadre pratique très utile à la validation des résultats de recherches.

Notre projet de thèse sera accompagné d'une réalisation informatique, qui consiste en la réalisation d'une chaîne de traitement sur les textes, appelée « ONTOLOGICO » au sein de la plate forme SATIM. Certains modules font d'ores et déjà, partie de cette plate forme, d'autres, plus spécifiques à notre modèle, seront implémentés au courant de notre projet.

2.4 Les hypothèses de recherche

Nous avons formulé certaines hypothèses[10] sur les objectifs que nous nous sommes fixés, en particulier :

(H1) : Application de l'ISL sur des classes de termes :

Notre hypothèse principale repose sur l'idée que l'application de la technique d'Indexation Sémantique Latente sur les groupes de termes identifiés par la classification, plutôt que sur la totalité des termes du corpus, possède l'avantage de réduire la matrice de cooccurrence de termes dans les documents à une dimension raisonnable. Nous présumons que ce choix méthodologique constitue une remède à la difficulté d'identifier dans la théorie, une dimension adéquate et précise de cette matrice.

(H2) : Ontologie monolingue :

Nous avons opté pour la proposition d'un modèle de maintenance d'ontologies qui sont monolingues. Nous n'aborderons pas la complexité impliquée par l'utilisation d'une ontologie multilingue. En effet, celle-ci suscite des réflexions cruciales, principalement attribuées au domaine de la traduction automatique, laquelle problématique s'éloigne des objectifs que nous avons établis.

(H3) : L'ontologie est une conceptualisation partagée :

[10] D'autres hypothèses plus spécifiques pourront être rajoutées au fur et à mesure de l'élaboration du modèle.

Partons des définitions retenues pour le concept d'ontologie, il est important de souligner la présence, quasiment systématique, de l'aspect «*conceptualisation partagée* » dans ces définitions. Cet aspect a influencé notre orientation vers la sélection de textes, propres à un domaine particulier (ou à une entreprise, etc.), comme source principale pour enrichir le modèle de l'ontologie.

(H4) : Indépendance du modèle par rapport à un domaine particulier

Bien que l'ontologie soit spécifique à un domaine particulier, notre objectif est de proposer un modèle de maintenance qui soit applicable pour d'autres domaines, par le biais de simples procédures d'adaptation. Par conséquent, la solution proposée ne doit pas être excessivement dépendante d'un domaine particulier, empêchant ainsi l'option d'adaptation, et rendant le modèle d'une application réduite. Toutefois, nous estimons que l'hypothèse d'indépendance, par rapport à un domaine particulier, ne serait pas remise en cause par l'utilisation d'un thésaurus spécialisé (par domaine), dans la mesure où le modèle reste valable suite à la substitution d'un thésaurus par un autre.

(H5) : Indépendance du modèle par rapport à la langue

L'analyse textuelle visée n'est pas spécifique à une langue particulière et par conséquent, ne doit prendre avantage des connaissances linguistiques, à l'exception de certaines procédures simples (telles que la lemmatisation, le filtrage de termes, etc.). Cette hypothèse est fondée sur les critiques (énoncées dans la section État de l'art), se rattachant à l'utilisation de la grammaire, tant pour l'extraction de termes à partir de textes que pour le repérage de relations entre termes.

(H6): Importance de la richesse des sources de connaissances

Notre modèle est fondé sur l'hypothèse que l'extraction de connaissances à partir de textes ne peut se contenter d'un traitement statistique (ni même linguistique) de données textuelles pour accaparer toute leur richesse sémantique. En effet, certaines connaissances implicites, spécifiques au domaine, ne peuvent être extraites à partir du corpus. Cette hypothèse est en fait, formulée sur la base de fondements cognitifs cohérents ; Lors d'un processus de lecture de textes par un expert du domaine, ce dernier fait souvent usage de certaines de ses propres connaissances du domaine (qu'on ne retrouve nécessairement pas dans les textes), pour repérer des relations d'associations conceptuelles entre termes. Il fait également recours à un

dictionnaire ou un thésaurus pour compléter ses connaissances par certaines informations, telles que les définitions, les synonymies, etc.

Ainsi, nous supportons l'idée que de telles connaissances sont accessibles à travers d'autres sources, à savoir ; un thésaurus spécifique au domaine en question, l'ontologie courante qui intègre progressivement les connaissances du domaine, au fur et à mesure que les actions de maintenance sont appliquées, ainsi que les requêtes formulées par les utilisateurs lors de recherches documentaires.

(H7) : La cooccurrence est un critère de choix pour le repérage de relations entre termes

La proximité sémantique entre les termes repose principalement sur la cooccurrence d'un ensemble de termes à travers différents segments d'un corpus. La base théorique de cette hypothèse repose sur celle formulée par Harris (1968) ; On peut classer les divers sens d'un terme en fonction des constructions auxquelles ce dernier participe. Des termes qui ont des distributions comparables ont souvent un élément de sens commun. Partant de cette hypothèse, nous avons préviligié l'emploi de la classification textuelle pour repérer, dans un premier temps, des regroupements de termes sémantiquement reliés. Nous argumentons davantage ce choix dans la section « Modèle proposé ».

(H8) : La cooccurrence résout, dans une certaine mesure, le problème d'ambiguïté lexicale

L'application de la technique de classification permet d'identifier des groupes de termes qui apparaissent ensemble et qui ont des relations sémantiques ou, du moins, des similarités sémantiques lorsque utilisés dans des contextes comparables.

La cooccurrence de termes à l'intérieur de différentes parties de textes implique une couverture fort probable d'un même thème. Cette hypothèse permet d'identifier le contexte qu'un terme possède dans le texte, et par conséquent, préciser son environnement sémantique correspondant dans lequel sa signification est employée, pour résoudre, dans une certaine mesure, le problème d'ambiguïté lexicale. Par exemple, le sens du terme « *souris* » est directement mis au clair, lorsque ce terme est présenté avec des cooccurrents tels que « *cliquer* », « *pointeur* », « *clavier* », etc., et est, par conséquent, distingué de la souris « *animal* » ou du verbe sourire.

(H9) : Les significations de termes contribuent activement au repérage de relations entre termes

Les définitions des termes constituent une source d'information importante, permettant de contribuer activement au repérage de relations entre termes. En effet, en présence d'un regroupement de termes qui sont potentiellement reliés (identifiés à l'aide de la classification par exemple), les définitions de chacun de ces termes peuvent enrichir le contenu informationnel de cette classe et concourir, par conséquent, à la découverte de relations conceptuelles plus fortes entre des couples de termes. Cette contribution consiste en fait, en un rapprochement entre les groupes de termes clés, faisant partie de chacune des définitions des termes en question. Cette hypothèse met en évidence l'importance de l'utilisation d'une ressource terminologique (dictionnaire électronique, thésaurus, etc.).

Dans le cadre de la représentation de connaissances et de la signification de lexique, « *les vecteurs conceptuels* » ont prouvé leur efficacité pour construire des taxonomies hiérarchiques et mettre en évidence des relations entre termes. Les vecteurs conceptuels sont généralement utilisés avec la distance thématique pour prendre des décisions par rapport à la qualité d'association entre les termes. Ils sont d'ailleurs, largement utilisés en recherche d'information (Salton et MacGill, 1983) ainsi qu'en représentation de signification par le modèle ISL (Deerwester et al., 1990), relatif aux études d'Analyse Sémantique Latente (ASL) en psycholinguistique.

(H10) : L'intervention d'un expert est une opération incontournable

Notre objectif consiste à construire une méthodologie supportant l'utilisateur lors de sa découverte de relations entre termes qui sont potentiellement utiles pour la maintenance de l'ontologie. Il est relativement facile de remettre en cause des systèmes automatiques prétendant accomplir cette tâche sans biais ou imperfections. Il semble plus raisonnable, et plus réaliste en terme de faisabilité, de suivre un processus plutôt semi-automatique, impliquant une simple intervention d'un expert du domaine, à travers certaines étapes, et spécialement pour la validation des résultats.

(H11) : Importance de la complétude des hypothèses de relations entre termes :

Les méthodes linguistiques de génération d'hypothèses de relations entre termes génèrent souvent moins d'hypothèses de relations que les méthodes statistiques de regroupement. Elles proposent cependant des relations étiquetées, directement vérifiables en contexte. Selon

notre point de vue, pour un cogniticien en phase de modélisation de connaissances, la complétude des hypothèses de relations (à proposer au cogniticien pour validation) est un objectif prioritaire par rapport à l'étiquetage. En effet, repérer une relation entre deux termes à partir d'un corpus nécessite un effort cognitif beaucoup plus important que pour étiqueter une relation préalablement identifiée. Cette hypothèse nous conduit à assigner plus d'importance au taux de rappel qu'à celui de précision en termes de relations candidates, présentées au cogniticien pour validation.

3 Proposition de solution et méthodologie

3.1 Modèle proposé

3.1.1 Description générale du modèle

Maintenir une ontologie de domaine consiste principalement à extraire à partir de documents des termes et des relations entre termes qui sont pertinents par rapport à l'ontologie courante. Notre objectif principal est de fournir de l'assistance à l'utilisateur pour assurer un raffinement continuel de l'ontologie tout en assurant la consistance et la cohérence de celle-ci et de ses artefacts. Notre approche, que nous proposons d'appeler « *Raffinement Conceptuel par Analyse Vectorielle* » (RCAV), donne une place centrale aux données textuelles et s'appuie sur la classification de textes et la technique d'Indexation Sémantique Latente, pour peaufiner graduellement les relations entre termes et extraire des couples de termes fortement reliés. Notre approche intègre efficacement une ressource terminologique (thésaurus) et fait appel à une intervention humaine (expert du domaine).

À travers un ensemble de traitements sur les textes, nous visons l'extraction de relations importantes entre les termes. Dans un premier temps, une méthode numérique (classification) est employée pour sélectionner rapidement des groupes de termes, qui sont potentiellement reliés, et qui nécessitent un processus de peaufinage pour extraire des couples de termes fortement reliés. Cette tâche est accomplie en utilisant la technique de l'Indexation Sémantique Latente (ISL) (Deerwester and al. 1990, Srivastava and al. 2002), qui aide à identifier, à l'intérieur de chaque classe de termes, ceux les plus corrélés. Toutes les relations entre ces termes doivent être vérifiées et analysées par un expert pour confirmer leur pertinence par rapport à la mise à jour de l'ontologie existante.

Pour procéder à la classification numérique, nous utilisons principalement le réseau de neurones ART (Adaptive Resonance Theory) (Grossberg, 1988), incorporé à l'intérieur de **GRAMEXCO**, qui est une instance d'une séquence de modules construite à partir de la plate forme **SATIM**[11] (Biskri, I., Meunier, J.G. 2002). En tant que système de traitement d'information, cette plate forme permet l'exploration et l'expérimentation de différents types d'analyses grâce à sa modularité, ses diverses fonctions d'analyse et sa capacité d'adaptation par rapport à la croissance des données textuelles. En particulier, **GRAMEXCO** permet l'exécution d'une séquence de traitements sur des textes pour classifier les segments en se basant sur l'approche des N-grams (Damashek 1989).

À ce stade, nous ne sommes pas rendus à une étape suffisamment avancée dans notre recherche pour prétendre décrire d'une façon exhaustive les composantes du modèle que nous proposons. Toutefois, nous avons déjà établi l'architecture générale (figure 1) du modèle et détaillé une bonne partie de ses composantes. Nous proposons dans ce qui suit, d'exposer notre processus itératif d'ingénierie, organisé sous forme de huit principales étapes.

3.1.2 Processus itératif d'ingénierie

Étape 1 : Extraction de N-grams et filtrage de termes

En utilisant **GRAMEXCO**, la première étape consiste à extraire les N-grams de caractères à partir de textes et à identifier les segments (parties de documents). Ces N-grams peuvent être définis comme une séquence de N caractères (par exemple, les séquences de trois caractères sont appelées des tri-grams). Ces deux objets forment la matrice qui sera utilisée par le classifieur. En d'autres termes, les segments seront comparés et classifiés sur la base de la cooccurrence de N-grams.

L'analyse de texte en termes de N-grams constitue une approche précieuse pour un texte écrit en toute langue basée sur un alphabet et la concaténation d'opérateurs de construction de textes. Il s'agit évidemment d'un avantage considérable répondant à la problématique : qu'est-ce qu'un mot? Par ailleurs, l'utilisation des N-grams de caractères à la place des mots, offre un autre avantage important : elle permet de contrôler la taille du lexique utilisé par le processeur, tel qu'illustré dans (Lelu and Halleb, 1998).

[11] SATIM peut accepter d'autres types d'information que les textes, tels que : les images, le son, etc.

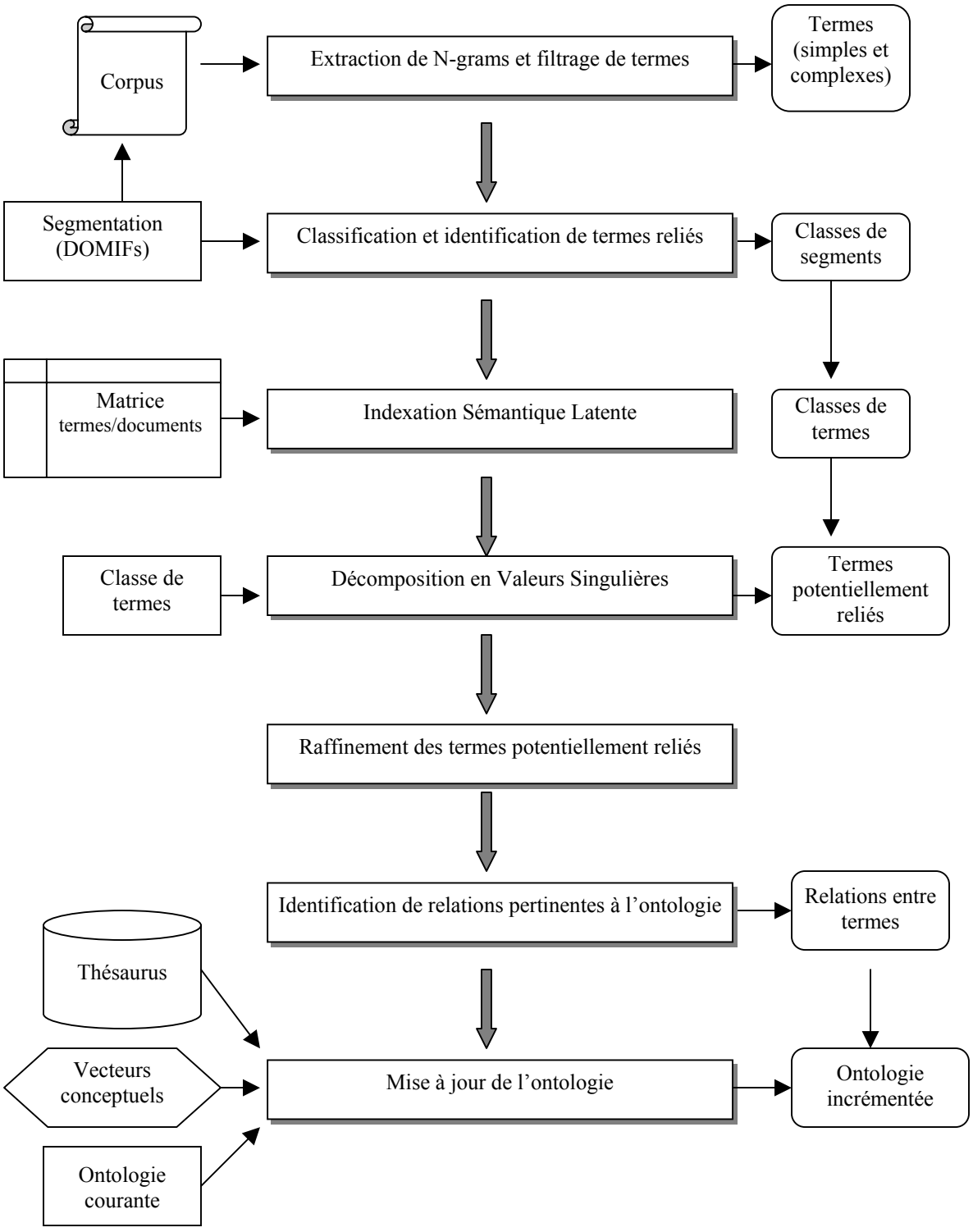

Figure 1 : Architecture de la chaîne de traitement ONTOLOGICO

L' « *extracteur de termes* » (un module de **GRAMEXCO**) est utilisé pour identifier le lexique (ensemble de lexèmes) à partir d'un corpus. Si les N-grams servent uniquement à la classification, le lexique joue un rôle plus actif dans les étapes suivantes. Par conséquent, avant de traiter ce lexique et extraire les N-grams, des opérations de filtrage doivent être réalisées pour garantir des résultats plus fiables. En d'autres termes, un processus de

lemmatisation automatique est employé pour remplacer les termes par leur lemmes correspondants.

En effet, des termes tels que {*informe, information, informant, etc.*} réfèrent au même concept, et doivent, par conséquent, être analysés en tant qu'un terme unique dans les étapes suivantes. Par ailleurs, un processus de filtrage permet d'éliminer les termes fonctionnels tels que {*le, la, dans, à, etc.*}, ainsi que les termes sémantiquement insignifiants. En effet, les termes trop fréquents ou les hapax ne jouent pas un rôle significatif dans la discrimination de segments (bien qu'ils peuvent être important pour la maintenance de l'ontologie. Malgré le fait de conserver ces termes n'affecte pas véritablement le processus de classification, ceci engendre toutefois, du bruit au cours des processus suivants. Il est clair ici que d'importants choix doivent être prix par l'utilisateur. Pour cette raison, **GRAMEXCO** offre une flexibilité et une convivialité pour supporter ses tâches.

Par ailleurs, l'identification complète et exacte de termes appartenant à un domaine spécifique est considérée comme un pré-traitement d'une grande importance pour la production de résultats adéquats et fiables. Devant les limites des techniques d'extraction automatique de tels candidats termes, il est plutôt important, à notre avis, de se focaliser sur les informations à présenter à un utilisateur pour effectuer une validation efficace. La véritable question est alors de définir les informations à visualiser pour décider de fixer comment le recours aux occurrences du terme est envisagé et dans quelle mesure les réseaux de termes peuvent affecter son jugement.

Étape 2 : Classification et identification de termes reliés

L'objectif de la classification est d'extraire certains types de « *régularités sémantiques* » entre les segments du texte (Manning, and Schütze, 1999; Sebastiani, 2002; Gelbukh and al., 1999). Ces segments contiennent un type d'information similaire et servent, par conséquent, à détecter des indices précieux aux associations entre termes. En tant que méthode de data mining textuel, ce processus, souvent moins détaillée que les approches linguistiques et conceptuelles, permet une première exploration générale et rapide du corpus. Elle identifie les classes de segments et groupes de lexèmes ayant des associations connues sous le nom de cooccurrence, et détecte par conséquent leurs réseaux sémantiques (Church and al. 1989, Lebart and Salem 1988, Salton 1988). Elle est habituellement exécutée en utilisant un classifieur numérique tel que exploré dans (Meunier & al., 1997) (Memmi et al 1998) (Benhadid & al., 1998) (Biskri & Delisle, 1999). En ce qui concerne **GRAMEXCO**, un

réseau de neurones ART (Adaptive Resonance Theory) (Grossberg, 1988) a confirmé son efficacité pour calculer les similarités entre les segments et produire les classes.

La classification utilise en tant qu'entrée, un modèle vectoriel, qui considère le texte dans sa totalité et vise à inférer, à partir des textes, une structure sémantique implicite (Salton & McGill 1983). Ce modèle traduit un texte sous forme d'espace matriciel qui associe les segments de textes avec les termes (ici les N-grams) et produit par conséquent, des réseaux de termes correspondant à des thèmes traités dans le texte (Memmi, 2000). En effet, la cooccurrence des termes à l'intérieur de différentes parties de texte implique une couverture fort probable d'un même thème. En général, le modèle vectoriel suit des étapes classiques : vectorisation, classification, interprétation et utilisation. L'étape de vectorisation, qui est une représentation du texte en termes de vecteurs, nécessite un choix décisif des éléments sélectionnés et des caractéristiques représentatives. Ce choix est évidemment dépendant des tâches subséquentes qui doivent être réalisées.

Après la classification de segments, nous extrayons le lexique à partir de chaque classe, qui représente l'intersection des termes appartenant à toutes les classes de segments. La cooccurrence de termes dans certains segments particuliers constitue un indice précieux pour de fortes associations entre termes. En d'autres termes, à l'intérieur de chaque classe, les termes ont une forte chance d'être reliés. Les étapes suivantes vise à déterminer des associations plus précises entre des couples de ces termes.

Étape 3 : Indexation Sémantique Latente

À ce niveau, nous espérons extraire, à partir de ces classes de termes, ceux représentant un niveau élevé de corrélation. L'Indexation Sémantique Latente (ISL) utilise une telle technique (Deerwester and al. 1990, Srivastava and al. 2002). Elle a été employée depuis les années 90 pour la recherche d'information sémantique à partir des textes, bien que les premiers travaux sur la cooccurrence ont commencé depuis les années 70. Cette technique a été utilisée, spécialement pour sa simplicité et sa justification par des fondements mathématiques assez précis.

Cette technique d'ISL utilise en tant qu'entrée, une matrice termes-documents correspondant aux poids des termes. Nous précisons ici que le corpus est considéré comme une collection de documents. Alors, nous calculons un poids pour chaque terme en utilisant son occurrence dans le document. Sa valeur est donnée par la formule suivante :

$$w_{i,k} = \frac{C_{i,k}}{\sum_{j=1}^{n_k} C_{j,k}}$$

où $w_{i,k}$ est le poids du terme T_i dans le document D_k. $C_{i,k}$ est le nombre d'occurrence du terme T_i dans le document D_k et n_k est le nombre total de termes dans le document. Ces termes sont limités à ceux filtrés et sectionnés pour la classification.

Les statistiques pour chaque document individuel sont combinées en vue de produire une analyse statistique pour la totalité de la collection. Un standard de normalisation de la longueur de documents , telle que expliquée dans (Greengrass, 1997), est utilisé pour éviter le fait qu'un terme peut avoir un poids élevé, simplement parce que le document, dans lequel il apparaît, est court, plutôt qu'en raison de sa fréquence élevée à travers la collection de documents. Les poids des termes normalisés deviennent :

$$W_{i,k} = \frac{w_{i,k}}{\sqrt{\sum_{j=1}^{n_k} w_{j,k}^2}}$$

Pour chaque classe de termes c, les poids correspondants forment la matrice W_c. Cette matrice de termes-documents est constituée de lignes représentant la collection des m documents (D_k), et de colonnes représentant n_c termes (T_i^c) appartenant à la classe c.

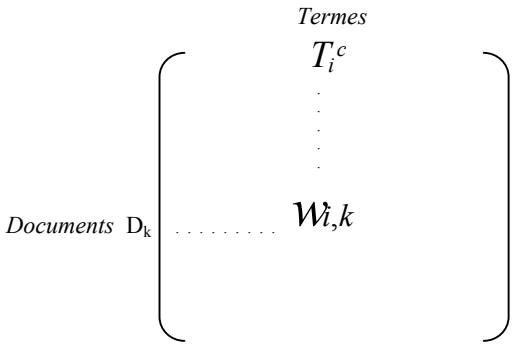

Figure 2 : La matrice terme-document W_c

Étape 5 : Décomposition en Valeurs Singulières

À l'intérieur de chaque classe de termes, nous visons la détermination des couples de termes reliés. La méthode ISL relie des termes sous forme d'une structure sémantique intéressante, tel que détaillé dans (Berry et al., 1995). L'ISL représente les documents par des concepts qui sont réellement et statistiquement indépendants de telle sorte que les termes ne le sont pas. Un concept est considéré ici en tant qu'un ensemble de termes reliés. L'ISL implique principalement la décomposition de la matrice W_c en utilisant la Décomposition en Valeurs Singulières (DVS) (Golub et al., 1969), qui est un type de régression linéaire. Alors, W_c peut être décomposée comme suit :

$$W_c = U \Sigma V^T$$

où U est une matrice de terme (m × r), V est une matrice de document (r × n_c) et ∑ est une matrice (r × r), avec r est le rang de W_c. ∑ est une matrice diagonale contenant les valeurs singulières de W_c. Dans cette décomposition, la valeur singulière σ_i correspond au vecteur u_i (la $j^{ème}$ colonne de U), et v_i (la $i^{ème}$ ligne de V). Les colonnes de U, les lignes de V et les valeurs diagonales de ∑ ont été arrangées de sorte que les valeurs singulières sont dans un ordre décroissant, en descendant la diagonale. Cette transformation de formule n'entraîne aucune perte de généralité.

Étape 6 : Raffinement des termes potentiellement reliés

Dans cette étape, nous visons l'extraction, à partir de chaque classe de termes, ceux les plus reliés. En effet, seuls ces termes et leurs relations doivent être considérés lors de la maintenance de l'ontologie.

Tel que discuté par (Deerwester and al. 1990; Nicholas and al. 1998), et en utilisant ISL, nous éliminons toutes les valeurs singulières de ∑ inférieures à un seuil de pourcentage de la valeur singulière la plus large, σ_1. Par conséquent, W_c représente une approximation de W_c^S dont l'exactitude s'accroît au fur et à mesure que *s* s'approche de *r* :

$$W_c^s = U^s \Sigma^s V^{sT}$$

où Σ^s est dérivée de ∑ en éliminant toutes les valeurs sauf les *s* valeurs singulières les plus larges, U^s est dérivée de U en éliminant toutes les valeurs sauf les *s* colonnes correspondant

aux valeurs singulières les plus larges, et V^s est dérivée de V en éliminant toutes les valeurs sauf les *s* lignes correspondant, où $s \leq r$.

U^s semble être le composant le plus important pour nous. En effet, la matrice ($m \times s$) représente les corrélations entre les termes dans la collection de documents et appartenant à la classe *c*. Chaque colonne de la matrice, u_i, est un vecteur que nous considérons représenter un concept. Les éléments de u_i indiquent la corrélation des termes par rapport au concept. Le fait de mettre à zéro tous les éléments dans u_i qui sont inférieurs à un seuil du pourcentage des termes les plus corrélés dans u_i, permet d'éliminer les termes qui sont faiblement associés (voir figure 3).

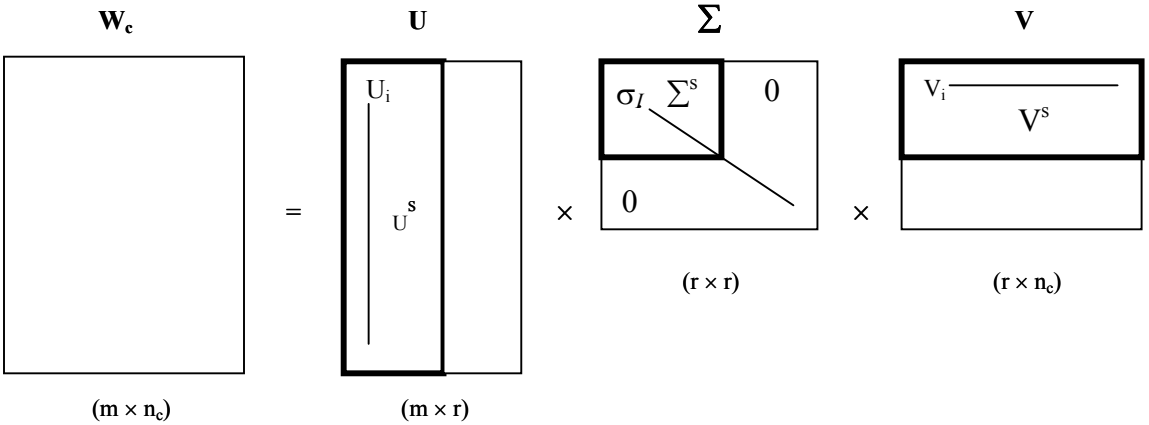

Figure 3 : Décomposition en Valeurs Singulières de la matrice terme/document W_c

L'application de la technique de ISL sur les groupes de termes identifiés par la classification, plutôt que sur la totalité des termes du texte, possède l'avantage de réduire la matrice de cooccurrence de termes dans les documents à une dimension raisonnable. En effet, malgré la difficulté d'identifier dans la théorie une dimension adéquate et précise de cette matrice, il est facile de prouver qu'une dimension immense pourrait empêcher l'émergence de suffisamment de relations sémantiques entre les termes, et aussi, une dimension trop réduite pourrait entraîner une grande perte d'information (Deerwester and al., 1990).

À ce niveau du processus, tous les couples possibles parmi les termes restants, sont présumés être potentiellement reliés. Il convient toutefois, de procéder à un filtrage manuel de ces termes pour éliminer ceux jugés non spécifiques au domaine, dans la mesure où seules les relations entre termes propres au domaine sont pertinentes pour l'ontologie. À cet effet, le bon sens de l'expert du domaine est retenu comme critère principal de filtrage, orientant son jugement.

Étape 7 : Identification de relations entre termes pertinentes à l'ontologie :

Dans l'objectif de découvrir des relations d'association parmi les termes retenus (τ_i), lors de l'étape précédente, nous proposons une méthode fondée sur un mariage entre, d'une part, l'approche *des réseaux sémantiques* (exp. : *WordNet*) associée au domaine de la « représentation des connaissances », et d'autre part, *l'approche vectorielle* issue des « représentations saltoniennes » (Salton, 1968) et de la recherche d'information.

Les connaissances disponibles dans les corpus sont habituellement explicites. Elles nécessitent des connaissances implicites, dans la mesure où en intelligence artificielle, les connaissances doivent être déclarées pour supporter le traitement inductif. Les thésauri sont spécialement utiles pour offrir des réseaux lexicaux et de l'information additionnelle reliée à la signification de termes (utilisation, définition, synonymie, etc.).

Ainsi, nous représentons les termes par des vecteurs conceptuels. Ces vecteurs sont construits à partir des items lexicaux associés à chacun de ces termes. Un ensemble de mesures est mis à la disposition de l'utilisateur (telle que la « *distance thématique* ») en vue de repérer différents types de proximité sémantique entre vecteurs conceptuels.

Étape 8 : Mise à jour de l'ontologie

Finalement, les nouveaux termes, ne figurant pas dans l'ontologie courante, ainsi que leurs relations, sont intégrés à cette ontologie. Un ensemble de règles assurant la cohérence du modèle global doit être respecté.

Un expert doit spécifier les types de relations sémantiques entre des termes jugés corrélés. par l'ISL. À cet effet, il est facile d'imaginer une interface où cette technique propose un réseau que l'expert procède graduellement à son étiquetage.

Les itérations de 5 à 8 doivent être reprises pour chacune des classes de termes identifiées lors de l'étape 2.

3.2 Méthodologie de recherche

Dans cette section, nous décrivons la méthodologie, que nous adoptons, pour le développement de notre modèle de maintenance des ontologies.

3.2.1 Justification du modèle proposé

La méthodologie de recherche que nous proposons suit un cheminement relativement classique chez la communauté scientifique. Comme toute problématique, il est primordial, dans un premier temps d'explorer les approches et les outils qui se rattachent d'une façon directe ou indirecte au problème posé. L'analyse critique des approches existantes, de leurs points forts et points faibles nous a permis de formuler un ensemble d'hypothèses (détaillées à la section 2.4) sur le modèle à proposer.

D'abord, la capacité des techniques statistiques, à traiter de larges données textuelles, a été le critère de base qui a nous a orienté vers ces approches plutôt que celles fondées sur la linguistique. En effet, bien que plus précises, ces dernières se trouvent confrontées à une difficulté majeure ; celle de l'aspect combinatoire.

À l'intérieur des méthodes statistiques de repérage de relations conceptuelles entre termes, la technique de l'Indexation Sémantique Latente a spécialement montré sa fiabilité. Cette technique est entre autres, privilégiée pour sa simplicité et sa justification par des fondements mathématiques précis et solides. Nous pensons toutefois, que cette technique a tout l'intérêt d'être associée avec la classification textuelle pour plusieurs raisons. Cette hypothèse a été argumentée en grande partie dans nos travaux antérieurs (Gargouri et al., 2003).

En particulier, l'application de la technique de ISL sur les groupes de termes identifiés par la classification, plutôt que sur la totalité des termes du texte, constitue un choix original possèdent l'avantage de réduire la matrice de cooccurrence de termes dans les documents à une dimension raisonnable. En effet, il est facile de prouver qu'une dimension immense pourrait empêcher l'émergence de suffisamment de relations sémantiques entre les termes, et aussi, une dimension trop réduite pourrait entraîner une grande perte d'information.

Par ailleurs, partons de l'hypothèse que les données textuelles ne peuvent, toutes seules, supporter la modélisation d'un domaine, du moins à cause des problèmes reliés à l'ambiguïté sémantique, nous avons opté pour l'utilisation d'un thésaurus, afin d'enrichir les connaissances extraites à partir de données textuelles et faire intégrer, le mieux que possible, les connaissances du domaine au sein de l'ontologie.

Dans l'objectif de raffiner davantage le processus de repérage de relations entre termes, nous proposons la représentation de termes par des vecteurs conceptuels et de mettre en place un ensemble de mesures de similarité entre vecteurs, à l'instar de « *la distance thématique* ».

3.2.2 Constitution de corpus

Les textes, objets de notre expérimentation, sont sélectionnés à partir d'une documentation technique relative au domaine ; celui des télécommunications sans fils pour notre cas. Cette sélection est élaborée en faisant recours à un expert du domaine.

Par ailleurs, dans la mesure où pour un système de fouille documentaire, l'historique des requêtes formulés par les utilisateurs constituent une source d'information d'une richesse considérable, nous considérons important l'exploitation de ces données textuelles et leur intégration au niveau du corpus à traiter. Il est à noter que, lors de la classification textuelle, la segmentation privilégiée est celle relative aux phrases (représentant chacune des requêtes), plutôt qu'aux paragraphes ou à un certain nombre de lignes. En effet, la discrimination entre segments, en tant que critère de classification, est plus fiable pour des phrases relatives à des requêtes particulières.

3.2.3 Expérimentation et développement

Le modèle proposé sera implémenté sous forme d'une chaîne de traitements (ONTOLOGICO) au sein de la plate-forme SATIM. En tant que système de traitement d'information, SATIM permet l'exploration et l'expérimentation de différents types d'analyses grâce à sa modularité, sa flexibilité, ses diverses fonctions d'analyse et sa capacité d'adaptation par rapport à la croissance des données textuelles.

Notre chaîne de traitements ONTOLOGICO vise à assister les experts de domaine dans leur tâche de maintenance des ontologies en se basant sur un processus itératif supporté par un ensemble de modules, en particulier ; un extracteur de termes simples, un extracteur de termes complexes, un lemmatiseur, un segmenteur, un classifieur, un thésaurus du domaine, un module de raffinement sémantique (basé sur l'Indexation Sémantique Latente) et un identificateur de termes reliés (basé sur le calcul de similarité sémantique entre les couples de vecteurs conceptuels).

Le modèle que nous proposons vise principalement à fournir de l'assistance aux ingénieurs d'ontologies. Par conséquent, les choix élaborés par ces derniers sont cruciaux par rapport au résultat final du processus de maintenance. Partons de cette hypothèse, l'évaluation de notre modèle porte particulièrement sur l'importance de l'assistance fournie par notre système aux utilisateurs. L'évaluation consiste à comparer la version incrémentée de l'ontologie suite à sa mise à jour, par rapport à une ontologie de référence (conçue manuellement). Les principales

mesures que nous utiliserons pour cette évaluation sont détaillées dans la section suivante (3.3).

Il est à noter que les choix, pris par les experts du domaine sur la base de cette assistance, ne remettent pas en cause notre technique de validation, dans la mesure où l'ontologie de référence est élaborée par ces mêmes experts. Par conséquent, les choix d'ordre sémantique et conceptuel s'accordent bien avec ceux de l'ontologie de référence.

L'outil ONTOLOGICO sera programmée en langage C++. Le choix de ce langage de programmation se justifie tout simplement par des objectifs de continuité et d'extensibilité future, au niveau de la plate forme SATIM, développée avec ce même langage. L'aspect modulaire de cet outil constitue un choix pratique particulièrement puissant, permettant entre autres, la possibilité de réutilisation de certains de ces modules pour d'autres objectifs d'analyses textuelles, ainsi que l'éventualité d'intégration de nouveaux modules répondant à des fonctionnalités complémentaires.

3.3 Méthode de validation des résultats

Dans cette section, nous présentons la méthodologie de validation que nous adopterons pour évaluer notre modèle de maintenance des ontologies.

Évaluer une technique d'apprentissage d'ontologie revient à mesurer la similarité entre une ontologie, manuellement conçue, considérée en tant que « *standard de référence* » et une ontologie générée en utilisant cette technique. Le degré de réussite de cette technique est d'autant plus important que la mesure de similarité est élevée.

Étant donnée la complexité du processus d'évaluation de la maintenance, celui-ci a été souvent exécuté en faisant recours à un expert du domaine, à l'instar des travaux de (Bachimont et al., 2002), (Faatz and Steinmetz, 2002), (Gupta et al., 2002), (Hearst, 1998), (Hwang, 1999), (Khan and Luo, 2002), (Kietz et al., 2000), (Missikoff et al., 2002), (Moldovan and Girju, 2001), etc.

Nous proposons de procéder à l'évaluation en trois étapes; d'abord, nous mesurons la performance de l'apprentissage de l'ontologie dans sa totalité (les termes et les relations entre termes), en utilisant les mesures communément employés en recherche d'information ; la *précision* et le *rappel*. Par la suite, nous évaluons l'ontologie construite en distinguant les niveaux lexical et conceptuel. L'objectif de cette distinction est d'analyser les taux de rappel

et de précision pour mieux comprendre les raisons d'une bonne performance par exemple, en se basant sur celles des niveaux lexical et conceptuel.

3.3.1 Rappel et précision

La *précision* est une mesure standard de la qualité ; C'est la mesure de la proportion des éléments corrects sélectionnés par le système, c'est-à-dire la proportion d'éléments figurant, tant dans l'ontologie de référence (*Réf*) que dans l'ontologie à comparer (*Comp*), par rapport à ceux faisant partie de cette dernière :

$$\text{Précision} = \frac{Comp \text{ I } Réf}{Comp}$$

Les éléments, dont nous faisons référence, sont les termes et les relations entre termes.

Le *rappel* est une mesure standard de la quantité d'éléments collectés. Sa formule est la suivante :

$$\text{Rappel} = \frac{Comp \text{ I } Réf}{Réf}$$

3.3.2 Évaluation du niveau lexical :

Sur le plan lexical, l'évaluation de l'ontologie est conduite en se basant sur « *la distance d'édition* » (ed) formulée par Levenshtein (Levenshtein, 1996). Cette technique consiste à mesurer le nombre minimum d'insertions, de suppressions et de substitutions de caractères nécessaires pour transformer une chaîne de caractères en une autre. La distance d'édition permet de calculer une « *mesure de similarité lexicale* » (SM) (comprise entre 0 et 1) entre deux unités lexicales (l_i et l_j). Plus SM est proche de 1, plus l_i et l_j sont similaires.

$$SM(l_i,l_j) = \max(0, \frac{\min(|l_i|,|l_j|) - ed(l_i,l_j)}{\min(|l_i|,l_j|)}) \in [0,1]$$

Dans l'objectif de comparer les niveaux lexicaux de deux ontologies dans leur totalité, nous comparons leurs lexiques correspondants (L_1 et L_2) en calculant la mesure moyenne de «*l'appariement entre chaînes de caractères* » ($\overline{SM}$);

$$\overline{SM}(L_1,L_2) = \frac{1}{|L_1|} \sum_{l_i \in L_1} \max_{l_j \in L_2} SM(l_i,l_j)$$

Cette mesure détermine jusqu'à quel point le lexique de l'ontologie de référence (L_1) est couvert par celui de l'ontologie à évaluer (L_2).

3.3.3 Évaluation du niveau conceptuel :

Sur le plan conceptuel, Maedche et Staab (2002) ont proposé un cadre d'évaluation particulièrement adéquat, permettant la comparaison de structures sémantiques d'ontologies O_1 et O_2. Cette évaluation est principalement basée sur une « *mesure moyenne similarité* » ($\overline{TO}$) entre deux taxonomies (H_1^c et H_2^c) ;

$$\overline{TO}(O_1,O_2)=\frac{1}{|L_1^c|}\sum_{L\in L_1^c} TO(L,O_1,O_2)$$

avec
$$TO(L,O_1,O_2)=\begin{cases} TO'(L,O_1,O_2) & \text{Si } L\in L_2^c \\ TO''(L,O_1,O_2) & \text{Si } L\notin L_2^c \end{cases}$$

$$TO'(L',O_1,O_2)=\frac{|F_1^{-1}(SC(F(\{L'\}),H_1^c))\,\mathrm{I}\,F_2^{-1}(SC(F(\{L'\}),H_2^c))|}{|F_1^{-1}(SC(F(\{L'\}),H_1^c))\,\mathrm{Y}\,F_2^{-1}(SC(F(\{L'\}),H_2^c))|}$$

$$TO''(L'',O_1,O_2)=\max_{c\in c2}\frac{|F_1^{-1}(SC(F(\{L''\}),H_1^c))\,\mathrm{I}\,F_2^{-1}(SC(C),H_2^c)|}{|F_1^{-1}(SC(F(\{L''\}),H_1^c))\,\mathrm{Y}\,F_2^{-1}(SC(C),H_2^c)|}$$

$SC(C_i,H^c)$; représente la mesure de «*Conceptual Cotopy* » du terme C_i , c'est-à-dire l'ensemble de tous les termes aux niveaux supérieur et inférieur à C_i :

$$SC(C_i,H^c)=\{C_j\in C\,|\,H^c(C_i,C_j)\,\mathrm{Y}\,H^c(C_j,C_i)\,\mathrm{Y}\,C_i=C_j\}$$

La taxonomie étendue (TO) entre H1 et H2 est considérée comme l'ensemble des termes, référés par L et identifiés à l'aide de F_1^{-1} et F_2^{-1} à partir du lexique commun.

(Nous détaillons davantage ce cadre d'évaluation lors de notre présentation)

3.4 Plan sommaire de la thèse

Le plan de la thèse sera articulé autour de 5 chapitres principaux

Chapitre 1 : Introduction

Nous présenterons à titre d'introduction, une description générale de la problématique de la maintenance des ontologies et des objectifs généraux de notre projet de recherche. Nous parcourons les champs de recherche pouvant contribuer à une réponse à notre problématique. Nous donnons un aperçu de l'état de l'art dans chacun de ces champs par rapport à nos préoccupations.

Chapitre 2 : Problématique

Ce chapitre est consacré à une revue de l'état de l'art sur la problématique de la maintenance des ontologies, les approches d'extraction de termes à partir de textes et les techniques de repérage de relations entre termes. Par la suite, et dans l'objectif d'introduire la solution que nous proposons, nous décrivons la composante cognitive qui se rattache à notre projet en discutons de l'analyse sémantique, la psychologie cognitive et les fondements d'une architecture cognitive. Nous terminons le chapitre par la spécification des hypothèses qui orienteront notre modèle.

Chapitre 3 : Solution proposée et méthodologie

Nous identifions nos options théoriques s'appuyant spécialement sur la psychologie cognitive, la cooccurrence, la classification textuelle, l'Indexation Sémantique Latente et les vecteurs conceptuels. Nous exposons notre réponse à la problématique posée au niveau méthodologique.

Ce chapitre comprend le modèle proposé, les outils utilisés et la méthodologie. Ceci s'explique par le fait que les deux aspects (outils et méthodologie) sont liés : les outils permettent de rendre la méthodologie opérationnelle et la méthodologie sert à la fois comme aide à la spécification des outils d'assistance à la maintenance et comme guide pour leur utilisation.

Chapitre 4 : Implémentation et évaluation

Nous présentons les spécifications relatives à notre chaîne de traitement « ONTOLOGICO ». Le chapitre rapporte également la validation de notre approche sur la base d'un ensemble de méthodes d'évaluation.

Chapitre 5 : Conclusions et perspectives

Nous présentons dans ce chapitre une synthèse du travail réalisé dans ce projet de recherche et les contributions originales apportées. Nous discutons également des éventuelles critiques et limites touchant la solution que nous avons apportée à la problématique de recherche. Nous terminerons, ce chapitre, par la présentation des perspectives de cette contribution dans le domaine de l'ingénierie des connaissances.

3.5 État d'avancement des travaux

Les travaux de recherches pour aborder la problématique et développer la méthodologie requise ont progressé de façon significative. Certains de ces résultats ont déjà fait l'objet de publications (Gargouri et al., 2003 a et b).

Ainsi, nous avons commencé par explorer les modèles d'extraction et de représentation de connaissances et les applications qui ont été réalisées en intelligence artificielle pour extraire des connaissances à partir de données textuelles. En particulier, nos recherches se sont orientées vers l'étude des outils et des méthodologies disponibles permettant la conception d'ontologie, ainsi que les techniques, tant statistiques que linguistiques, relatives au repérage de termes reliés, de collocations, de cooccurrences et de termes complexes.

Nous avons également proposé, dans des travaux antérieurs, un modèle d'extraction de termes fortement corrélés à partir d'un corpus. Le modèle consiste à opérer une classification numérique de textes (à l'aide de réseaux de neurones), et à appliquer la technique d'Indexation Sémantique Latente (ISL) associée avec la Décomposition en Valeurs Singulières (DVS), en vue de peaufiner davantage les résultats de la classification. Nous avons par la suite complété notre modèle en intégrant l'ontologie courante et un thésaurus en vue de supporter la tâche d'assistance à la maintenance.

Le repérage de proximité sémantique entre vecteurs conceptuels, que nous proposons au niveau de la dernière étape de notre processus itératif, se limite pour l'instant à l'utilisation de la distance thématique. Nous explorons actuellement, d'autres alternatives de mesures à mettre à la disposition de l'utilisateur permettant de supporter davantage le repérage de relations entre termes et de l'orienter dans sa tâche d'étiquetage de ces relations.

Notre cheminement méthodologique se poursuivra par la spécification des modules à développer au sein d'ONTOLOGICO. Il s'agit en particulier de construire les modèles procéduraux et les cas d'utilisation.

4. Conclusions

4.1 Contributions originales du projet

Nos principales contributions peuvent être résumées en sept points :

- L'application de la technique de ISL sur les groupes de termes identifiés par la classification, plutôt que sur la totalité des termes du texte, constitue un choix original possèdent l'avantage de réduire la matrice de cooccurrence de termes dans les documents à une dimension raisonnable (Gargouri et al., 2003). En effet, il est facile de prouver qu'une dimension immense pourrait empêcher l'émergence de suffisamment de relations sémantiques entre les termes, et aussi, une dimension trop réduite pourrait entraîner une grande perte d'information.
- Nous introduisons la notion de « *Raffinement Conceptuel par Analyse Vectorielle* » (RCAV), qui consiste en un processus cohérent de raffinement graduel de relations conceptuelles entre termes.
- Nous présentons une *méthodologie* et des *outils* pour maintenir une ontologie à partir d'analyses textuelles. Dans la mesure où les outils d'édition des ontologies, disponibles actuellement, n'intègrent pas cette fonctionnalité, il s'agit de l'aspect le plus important de notre contribution, remédiant spécialement aux lacunes méthodologiques.
- Les réflexions, qui seront menées à l'occasion du repérage de relations conceptuels entre termes, constitueront une contribution aux travaux dans le domaine de l'extraction de connaissances à partir d'analyses statistiques de textes.
- Cette recherche se place au cœur des échanges entre terminologie et acquisition de connaissances. Elle amène par conséquent, une réflexion sur les divers paliers à envisager dans une telle démarche de modélisation de connaissances textuelles pour des objectifs de maintenance d'une ontologie.
- Bien que notre approche soit fondée, au départ, sur un corpus textuel, le processus méthodologique de la maintenance d'ontologie demeure indépendant, tant de la langue que du corpus d'expérimentation et reste valable à l'intérieur du domaine en question, ainsi que dans d'autres domaines. Ceci n'est pas le cas pour des approches linguistiques, basées par exemple, sur l'utilisation de marqueurs.

- Enfin, le modèle proposé assiste les terminologues chargés de naviguer à travers de vastes données textuelles pour extraire et normaliser la terminologie. Il facilite également la tâche des ingénieurs en connaissances chargés de modéliser des domaines.

4.2 Obstacles à franchir

La réalisation de notre projet de recherche se trouve bien entendu, confrontée à quelques obstacles, que nous ambitionnons de résoudre ;

- En pratique, les méthodologies les plus utilisées pour la construction et la maintenance des ontologies sont généralement celles basées sur une approche descendante, c'est-à-dire dirigées par un modèle. Les méthodes ascendantes (du texte vers le modèle) sont beaucoup plus rares. Nous estimons que les textes, en tant que source principale des connaissances sur le domaine, sont actuellement sous-exploités. Un de nos principaux défis serait d'affronter la complexité du traitement des données textuelles en vue d'exploiter la richesse implicite de cette source de connaissances.

- Nous sommes confrontés à la complexité de la terminologie, qui se trouve d'ailleurs au cœur des préoccupations dans diverses applications reliées à la représentation des connaissances, la construction d'ontologies, la gestion des connaissances, la veille scientifique, la traduction automatique, la classification ou la fouille de textes. Le problème majeur est celui de l'automatisation partielle ou totale des processus de repérage, de structuration des termes et de représentation des connaissances dans un domaine en s'appuyant sur la terminologie textuelle.

- Pour la maintenance des ontologies, l'identification complète et exacte de termes appartenant à un domaine spécifique est considérée comme un pré-traitement d'une grande importance pour la production de résultats adéquats et fiables. Le modèle que nous proposons se contente d'un filtrage manuel. Nous pensons toutefois, des techniques spécifiques doivent être employées pour évaluer la pertinence de ces termes, comparés avec d'autres, en se basant sur l'ontologie existante et aussi sur d'autres sources d'information telles que les thésauri.

- Les données textuelles que nous traitons sont de larges tailles. Nous sommes par conséquent confrontés à une contrainte de temps d'exécution, qui est croissante par rapport à la taille du corpus.

- Le problème d'ambiguïté sémantique se pose au niveau de plusieurs étapes de notre processus itératif de maintenance ; Par exemple, lors de l'appariement de termes extraits à partir des textes avec ceux de l'ontologie courante, les homonymes constituent un cas d'ambiguïté que nous allons considérer dans notre modèle.

- Dans la mesure où notre modèle suit un processus semi-automatique, l'intervention d'un expert du domaine pour la validation des résultats au niveau de différentes étapes de notre processus, se controuve confronté à un problème de subjectivité. Un compromis devrait être considéré entre différentes validations d'experts.

5. Bibliographie

D. Fensel, (2001) « *Ontologies, a Silver Bullet for Knowledge Management & Electronic Commerce*».

U. Hahn and K. Schnattinger, (1998) « *Towards text knowledge engineering*», in AAAI'98 -- Proc. 15th National Conf. on Artificial Intelligence,.

S.Luke, L.Spector, D.Rager, J.Hendler, (1997) « *Ontology based web agents* », in « Proceedings of 1st international conference on autonomous agents ».

J.R.Ambroziak, « *Conceptual assisted web browsing* », Sun Technical Report 61, 1997.

B.Chabbat, J.M. Pinon, M. Ou-Halima, (1995) « *Hypertexte sémantique pour l'aide à la décision*», in « Ingénierie des systèmes d'informations », Volume 3.

B.R. Gaines, (1995) « *Class library implementation of an open architecture knowledge support system* », in « International journal of human computer studies », volume 41.

Amar Kheirbek, Yves Chiaramella, (1995) « *Integrating hypermedia and information retrieval with conceptual graphs* », in HIM95, Konstanz, Germany, Avril 1995.

John F. Sowa, (1984) « *Conceptual structures : Information processing in mind and machine* », Addison-Wesley.

Gérard Sabah, (1998) « *L'IA et le langage : Représentation des connaissances* », Editions Hermès.

W.A.Woods, (1975) « *What's in a link : Foundations for semantic networks* », in «Representation and understanding : Studies in cognitive science », D.G.Bobrow-A.M.Collins Eds, Academic Press New York.

W.A.Woods, (1998) « *Conceptual indexing : a better way to organize knowledge* », Sun Technical Report.

Mitsuru Ikeda, Yusuke Hayashi, Jin Lai, Weiqin Chen, Jacqueline Bourdeau, Kazuhisa Seta, Riichiro Mizoguchi, «*An ontology more than a shared vocabulary*» AI-ED 99 «Workshop on Ontologies for Intelligent Educational Systems», Le Mans, France, July 18-19, 1999.

Ikeda, M, K. Seta, and R.Mizoguchi. «*Task Ontology Makes It Easier To Use Authoring Tools*», Proc.of IJCAI-97, Nagoya, Japan, pp.342-347, 1997.

Jin, L., Hayashi,Y., Chen, W., Ikeda, M., Riichiro, M., «*An Ontology aware authoring tool*», AIED99, Le Mans, France, 1999.

Mizoguchi, R. et. al. «*Task Ontology design for intelligent educational/training systems*», Workshop on Architectures and methods for Designing Cost-effective and Reusable ITSs, ITS96, Montoreal, 1996.

Mizoguchi, R., Ikeda, M., and K. Sinitsa. «*Roles of Shared Ontology in AI-ED Research*», Intelligence, Conceptualization, Standardization, and Reusability --. Proc. of AIED-97, Kobe, Japan, pp. 537-544, 1997.

Murray, T. «*Special Purpose Ontologies and the Representation of Pedagogical Knowledge*», ICLS96, 1996.

Shimizu, H., Seta, K., Hayashi,S. Motomatsu, M., Ikeda, M. and Mizoguchi, R. «*A basic consideration on design patterns for ontology building*», Proc. of the 58th National Conference of Information Processing Soc. Of Japan, 3U-9, 1999.

Sylvie Ranwez1, Michel Crampes1 and Torsten Leidig, «*Description and Construction of Pedagogical Material using an Ontology based DTD*», Laboratoire de Génie Informatique et Ingénierie de Production (LGI2P) EERIE-EMA, France, 1999.

J. Arpírez, A. Gómez-Pérez, A. Lozano, S. Pinto: « *An ontology-based WWW broker to select ontologies*». In Proceedings of the ECAI-98 Workshop on Applications of Ontologies and PSMs, Brighton. England. pp. 16-24, August 1998.

V. R. Benjamins, D. Fensel, S. Decker and A. Gomez Perez, « *Building Ontologies for the Internet*»: a Mid Term Report, International Journal of Human-Computer Studies, 51:687-712, 1999.

W. N. Borst and J. M. Akkermans: « *Engineering Ontologies* », International Journal of Human-Computer Studies, 46 (2/3):365-406, 1997.

R. Brachman and J. Schmolze: « *An overview of the KLONE Knowledge Representation System* », Cognitive Science, 9(2):171-216, 1985.

S. Decker, M. Erdmann, D. Fensel und R. Studer, « *Ontobroker: Ontology based Access to Distributed and Semi-Structured Information* ». In R. Meersman et al. (eds.), Semantic Issues in Multimedia Systems, Kluwer Academic Publisher, Boston, 1999.

M. Erdmann and R. Studer, « *Ontologies as Conceptual Models for XML Documents* », research report, Institute AIFB, University of Karlsruhe, 1999.

H. Eriksson, A. R. Puerta, and M. A. Musen, « *Generation of knowledge-acquisition tools from domain ontologies* », International Journal of Human Computer Studies (IJHCS), 41:425-453, 1994.

H. Eriksson, R. W. Fergerson, Y. Shahar, and M. A. Musen, « *Automated Generation of Ontology Editors* ». In Proceedings of the Twelfth Workshop on Knowledge Acquisition, Modeling and Management (KAW99), Banff, Alberta, Canada, October 16-21, 1999.

D. Fensel, «*Understanding, Developing and Reusing Problem-Solving Methods*», Lecture Notes of Artificial Intelligence (LNAI), Springer-Verlag, Berlin, 2000.

Biskri, I., Meunier, J.G., Mars 2002, «*SATIM : Système d'Analyse et de Traitement de l'Information Multidimensionnelle*». JADT 2002, St-Malo, France.

Meunier, Jean-Guy (2002) « *La représentation et les Sciences cognitives* », RSSI, 2002 et cahiers du LANCI 2001.

Chomsky, N. (1957) «*Syntactic structures*». The Hague, Mouton & co.

Chomsky N. (1965) «*Aspects of the theory of syntax*». MIT-Press.

Fodor, J. A. and Pylyshyn, Z. W. (1988) «*Connectionism and Cognitive Architecture: critical analysis*» O. Cognition 28(1-2), 3-71.

Johonson-Laird, P. N. (1988) «*The computer and the mind*» Harvard UP. 1988.

Pylyshyn, Z. (1984) «*Cognition and computation*» Cambridge: MIT Press.

Clark, A. (1997) «*Being there : Putting the Brain Body and World together again*». Cambridge UP.

Sowa, John F. (2000) «*Ontology, Metadata, and Semiotic*» In Conceptual Structures: Logical, Linguistic, and Computaional Issues (sous la direction de B. Ganter & G. W. Mineau), Berlin: Spring-Verlag, pp. 55-81.

Peirce, Charles Sanders (1958) «*Collected papers of C.S Peirce*» ed. By C. Hartshrne, P. Weiss, & A. Burks, 8 vols., Harvard University Press, Cambridge, MA, 1931-1958.

Saussure, Ferdinand de (1916) «*Cours de linguistique Générale*», translated by W. Baskin as Course in General Linguistics, Philosophical Library, New York, 1959.

M. Tallis, Y. Gill (1999) «*Designing scripts to guide users in modifying knowledge-based systems*», AAAI/IAAI 1999 : 242-249.

S. Staab, H.-P. Schnurr, R. Studer et Y. Sure (2001) «*Knowledge Processes and Ontologies*» IEEE Intelligent Systems. 16(1), jan./fev. 2001. Special Issue on Knowledge Management 2001.

Nicolas Guarino, (1995) "*Formal Ontology, conceptual analysis and knowledge representation"*, International Journal of human-computer studies, special issue on the role of formal ontology in the information technology,1995.

T.R. Gruber, (1993) '*A translation approach to portable ontology specifications"*, Knowledge Acquisition, 199-200, 1993

Eco, Umberto (1984) «*Semiotics and the philosophy of language*», Bloomington, Indiana University Press.

Eco, U., (1985) «*Lector in fabula ou la Coopération interprétative dans les textes narratifs*». Paris: Grasset.

Eco, Umberto (1992) «*Interprétation et surinterprétation*», Cambridge ; New York. Cambridge University Press.

Miller, G.A. (1995) «*WORDNET: A lexical database for English*». Communications of

ACM (11), 39-41.

Ying Ding (2001), «*Ontology Research and Development, Part 1 – A Review of Ontology Generation*», Division of Information Studies, School of Computer Engineering Nanyang Technological University, Singapore.

Jannink, J. & Wiederhold, G. (1999) «*Ontology Maintenance with an Algebraic Methodology: a Case Study*», in Proceedings of 1999 AAAI workshop on Ontology Management, Orlando FL.

Stumme, G., Studer, R. & Sure, Y. (2000) «*Towards an order-theoretical foundation for maintaining and merging ontologies*». In. Proc. Referenzmodellierung 2000, Siegen, Germany, October 12-13, 2000.

Ganter, B. & Wille, R. (1999) «*Formal concept analysis: Mathematical Foundations*». Springer, Berlin-Heidelberg.

A. Faatz, T. Kamps, R. Steinmetz (2000), «*Background Knowledge, Indexing and Matching Interdependencies of Document Management and Ontology-Maintenance*».

A. Todirascu, F. de Beuvron, D. Galea, F. Rousselot (2000), «*Using Description Logics for Ontology Extraction*». In. Proceedings of the First Workshop on Ontology Learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence (ECAI 2000). August, Berlin, Germany.

Guarino, N., Masolo, C., & Vetere, G. (1999) «*OntoSeek: Content-based access to the Web*». IEEE Intelligent Systems, May/June, 70-80.

Adam Farquhar Richard Fikes James Rice (1996) «T*he Ontolingua Server: a Tool for Collaborative Ontology Construction*», Knowledge Sustems Laboratory, Stanford University, Stanford, CA, Proceedings of Tenth Knowledge Acquisition for Knowledge-Based Systems Workshop KAW'96.

Uschold, M. (2000) «*Creating, integrating and maintaining local and global ontologies*». In. Proceedings of the First Workshop on Ontology Learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence (ECAI 2000). August, Berlin, Germany.

Brill E. (1994) «*Some Advances in Transformation-Based part-of-Speech Tagging*». Proceedings of the AAAI.

Sabah G. (1989) «*L'intelligence artificielle et le langage, tome 2: processus de compréhension*». Hermès, Paris.

Bourigault D. (1994) «*LEXTER, un logiciel d'Extraction de TERminologie, Application à l'acquisition des connaissances à partir de textes*». Ecole des Hautes Etudes en Sciences sociales, Paris.

Zernik U. (1992) «*Closed Yesterday and Closed Minds: Asking the Right Questions of the Corpus To Distinguish Thematic from Sentential Relations*». Actes, 14th International Conference on Computational Linguistics (COLLING'92), pp. 1305-1311. Nantes.

Grefenstette G. (1992) «*Use of syntactic context to produce term association lists for text retrieval*». Proceedings of the 15th International Conference on Research and Development in Information Retrieval (SIGIR'92), pp. 89-97. Copenhagen.

Daille B. (1994) «*Approche Mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques*». Paris 7.

Jouis C. (1995) «*SEEK, un logiciel d'acquisition des connaissances utilisant un savoir linguistique sans employer de connaissances sur le monde externe*». Actes de JAVA 95. Grenoble.

Desclés, J.-P. (1990). «*Langages Applicatifs, Langues Naturelles et Cognition*». Hermès, Paris.

Kim J.-T. & Moldovan D. I. (1995) «A*cquisition of linguistic Patterns for Knowledge-Based Information Extraction*». IEEE Transactions on Knowledge and Data Engineering.

Dowding J., Moore R., Andy F. & Moran D. (1994) *Interleaving syntax and semantics in an efficient bottom-up parser*.

Sowa J. F. (1984) «*Conceptual Structures, Information processing in mind and machine*». Addison-Wesley, Reading, Mass.

Binet J., Dierickx J. & Funck-Brentano J. (1987) « *Le français, langue des sciences et des techniques* ». RTL-Edition, Luxembourg.

Baylon C. & Fabre P. (1975) *Initiation à la linguistique*. Nathan, Paris.

Rastier f. (1990) «*La triade sémiotique, le trivium et la sémantique linguistique*». Nouveaux Actes sémiotiques (9), pp. 59-68. PULIM, Université de Limoges.

Rastier, F, (1994) «*Tropes et sémantique linguistique*» in Langue Française No 101, Février 1994, Larousse, Paris.

Rastier F. (1995) «*Le terme : entre ontologie et linguistique*». La banque des mots. 7- 1995.

Rastier, F., Cavazza, M., and Abeillé, A. (1994). "*Sémantique pour l'analyse, de la linguistique à l'informatique*". Paris,: Masson.

Bessé B. D. (1990) « *La définition terminologique* ». Larousse, Paris.

Aussenac-Gilles N., Bourigault D., Condamines A. & Gros C. (1995) « *How can knowledge acquisition benefit from terminologie?* » Proceedings of the 9th Knowledge Acquisition for Knowledge Based Systems Workshop. Banff.

Gadamer, H. G., (1976) «*Vérité et méthode* ». Édition du Seuil, Paris.

Papen, R. 1990 « *La sémantique* » chap. In : Introduction à la linguistique, Université du Québec à Montréal.

Salton G. (1989) «*Automatic Text Processing*». Addison-Wesley.

Salton G. & M. McGill (1983), «*Introduction to Modern Information Retrieval*», McGraw-Hill.

Salton G. (1968) "*Automatic Information Organisation and Retrieval"* McGraw-Hill, New York.

Bourigault, D. (2002) «*Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus*», Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002), Nancy.

Rebeyrolle J. (2000c). «*Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes*», In Actes Journées Francophones d'Ingénierie des Connaissances, IC'2000, Toulouse, mai 2000.

Wray, R et al. "*a survey of cognitive and agent architectures*", University of Michigan, HTML.

Morin E (1999) *Automatic acquisition of semantic relations between terms from technical corpora.* Proc. Of the Fifth Int. Congress on Terminology and Knowledge Engineering (TKE-99), TermNet-Verlag, Vienna

Hearst M.A. (1992) *Automatic acquisition of Hyponyms from large text corpora.* In Proceedings of the Fourteenth International Conference on Computational Linguistic, Nantes, France, July 1992.

Agrawal R, Imielinski T, Swami A (1993) *Mining association rules between sets of items in large databases.* In Proc. Of the ACM SIGMOD Conference on Management of Data, 207-216.

Adriaans P, Zantinge D. (1996) *Data Mining.* Addisson-Wesley, 1996.

Maedche A, Staab S. (2001) *Ontology Learning for the Semantic Web.* IEEE Intelligent Systems, Special Issue on the Semantic Web, 16(2)

Maedche, A. and Staab, S. (2000) *Discovering Conceptual Relations from Text*. In: W.Horn (ed.): ECAI 2000. Proceedings of the 14th European Conference on Artificial Intelligence, Berlin, August 21-25, 2000. IOS Press, Amsterdam, 2000.

Faure D, Poibeau T (2000) *First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX.* In: S. Staab, A. Maedche, C. Nedellec, P. Wiemer-Hastings (eds.), Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence ECAI'00, Berlin, Germany

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

Hahn U, Schulz S (2000) *Towards Very Large Terminological Knowledge Bases: A Case Study from Medicine.* In Canadian Conference on AI 2000: 176-186.

Agirre, E., Ansa, O., Hovy, E., and Martinez, D. (2000). *Enriching very large ontologies using the WWW*. In Proceedings of the Workshop on Ontology Construction of the European Conference of AI (ECAI-00).

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

Alfonseca E. and Manandhar S. (2002b). *Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures*, EKAW-2002, Siguenza, Spain. Published in Lecture Notes in Artificial Intelligence 2473 (Springer Verlag).

Aussenac-Gilles, N, Biébow B, Szulman S. (2000a) *Corpus Analysis For Conceptual Modelling.*

Workshop on Ontologies and Text, Knowledge Engineering and Knowledge Management: Methods, Models and Tools, 12th International Conference EKAW'2000, Juan-les-pins, France, Springer-Verlag.

Aussenac-Gilles N, Biébow B, Szulman S (2000b) *Revisiting Ontology Design: A Methodology Based on Corpus Analysis*. In: Dieng R, Corby O (eds) 12th International Conference in Knowledge Engineering and Knowledge Management (EKAW'00). Juan-Les-Pins, France. Springer-Verlag, Lecture Notes in Artificial Intelligence (LNAI) 1937, Berlin, Germany, pp 172–188.

Bachimont B., Isaac A., and Troncy R. (2002). *Semantic commitment for designing ontologies: a proposal.* In A. Gomez-Perez and V.R. Benjamins (Eds.): EKAW 2002, LNAI 2473, pp. 114–121, 2002. Springer-Verlag Berlin Heidelberg 2002.

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

Gupta, K.M., Aha, D.W., Marsh, E., and Maney, T. (2002). *An architecture for engineering sublanguage WordNets*. In Proceedings of the First International Conference On Global WordNet (pp. 207-215). Mysore, India: Central Institute of Indian Languages.

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

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

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

Khan L., and Luo F. (2002) *Ontology Construction for Information Selection* In Proc. of 14th IEEE International Conference on Tools with Artificial Intelligence, pp. 122-127, Washington DC, November 2002.

Lonsdale D, Ding Y, Embley D.W, and Melby A. (2002) *Peppering Knowledge Sources with SALT; Boosting Conceptual Content for Ontology Generation*. Proceedings of the AAAI Workshop on Semantic Web Meets Language Resources, Edmonton, Alberta, Canada, July 2002.

Missikoff M., Navigli R., and Velardi P. (2002). *The Usable Ontology: An Environment for Building and Assessing a Domain Ontology* Research paper at International Semantic Web Conference (ISWC) 2002, June 9-12th, 2002 Sardinia, Italia

Moldovan, D. I.; Girju, R. C. (2001). *An interactive tool for the rapid development of knowledge Bases*. In International Journal on Artificial Intelligence Tools (IJAIT), vol 10., no. 1-2, March 2001.

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

Roux C., Proux D., Rechermann F., and Julliard L. (2000). *An ontology enrichment method for a pragmatic information extraction system gathering data on genetic interactions*. Position paper in Proceedings of the ECAI2000 Workshop on Ontology Learning(OL2000), Berlin, Germany. August 2000.

Wagner, A. (2000). *Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis*. In Proceedings of the ECAI-2000 Workshop on Ontology Learning Berlin, August 2000, 37-42.

Xu F., Kurz D., Piskorski J., and Schmeier S. (2002). *A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping*. In Proceedings of LREC 2002, the third international conference on language resources and evaluation, Las Palmas, Canary island, Spain, May 2002.

Faure D, Nédellec C. (1999) *Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system ASIUM*. In D. Fensel and R. Studer editors, Proc. Of the 11th European Workshop (EKAW'99), LNAI 1621, pages 329-334. Springer-Verlag.

Engels R (2001a) *CORPORUM-OntoExtract. Ontology Extraction Tool*. Deliverable 6 Ontoknowledge. http://www.ontonowledge.org/del.shtml

Engels R (2001b) *CORPORUM-OntoWrapper. Extraction of structured information from web based resources*. Deliverable 7 – Ontoknowledge. http://www.ontonowledge.org/del.shtml

Bachimont B. (2000). *Engagement sémantique et engagement ontologique: conception et réalisation d'ontologies en ingénierie des connaissances*. In Ingénierie des Connaissances : Evolutions récentes et nouveaux défis, Eyrolles, 2000.

Jones, S. and Paynter, G.W. (2002) *Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications*. Journal of the American Society for Information Science and Technology (JASIST).

Mikheev, A. Finch, S.(1997) *A Workbench for Finding Structure in Texts*. Proceedings of ANLP-97 (Washington D.C.). ACL March 1997. pp 8.

Bisson G, Nedellec C, Cañamero D. (2000) *Designing Clustering Methods for Ontology Building. The Mo'K Workbench*. In S. Staab, A. Maedche, C. Nedellec, P. WiemerHasting (eds.), Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence, ECAI'00, Berlin, Germany, August 20-25.

Velardi P., Navigli R., and Missikoff M. (2002). *Integrated approach for Web ontology learning and engineering.* IEEE Computer - November 2002.

Morin E. (1999) *Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique.* TAL (Traitement Automatique des Langues). 40/1: 143-166, 1999 (Prométhée).

Morin E. (1998) *Prométhée un outil d'aide a l'acquisition de relations semantiques entre temes.* 5ème Conférence Nationale sur le Traitement Automatique des Langues Naturelles (TALN'98), pages 172-181, Paris, France, June 1998

Wu S.H, Hsu W.L. (2002). *SOAT: A Semi-Automatic Domain Ontology Acquisition Tool from Chinese Corpus.* In the 19th International Conference on Computational Linguistics, Howard International House and Academia Sinica, Taipei, Taiwan

Chaelandar G, Grau B. (2000) *SVETLAN'- A System to Classify Words in Context*. In S. Staab, A. Maedche, C. Nedellec, P. Wiemer-Hastings (eds.) Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence ECAI'00, Berlin, Germany, August 20-25.

Biébow B, Szulman S. (1999) *TERMINAE: a linguistic-based tool for the building of a domain ontology*. In EKAW'99 Proceedings of the 11th European Workshop on Knowledge Acquisition, Modelling and management. Dagstuhl, Germany, LCNS, pages 49-66, Berlin, 1999. Springer-Verlag.

Maedche, A. and Volz, R. (2001) *The Text-To-Onto Ontology Extraction and Maintenance Environment.* To appear in Proceedings of the ICDM Workshop on integrating data mining and knowledge management, San Jose, California, USA.

Pereira, F. C. (1998). *Modeling Divergent Production: a multi domain Approach*. European Conference of Artificial Intelligence, ECAI'98, Brighton, UK, 1998.

Thompson, C.A. & Mooney, R. J. (1997). *Semantic Lexicon Acquisition for Learning Parsers*. Technical Note. January 1997.

Alfonseca E., and Rodríguez P. (2002), *Automatically Generating Hypermedia Documents depending on User Goals*, Workshop on Document Compression and Synthesis in Adaptive Hypermedia Systems, AH-2002, Málaga, Spain.

Rigau G., Rodríguez H. and Agirre E (1998). *Building Accurate Semantic Taxonomies from Monolingual MRDs*. Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics COLING-ACL'98. Montreal, Canada. 1998.

Rigau G. (1998). *Automatic Acquisition of Lexical Knowledge from MRDs*, Ph.D. Thesis, Comptur Systems and linguistics Department. Polytechnic University of Catalunya.

Suryanto H, Compton P. (2001) *Discovery of Ontologies from Knowledge Bases.* Proceedings of the First International Conference on Knowledge Capture, Eds. *Yolanda Gil; Mark Musen; Jude Shavlik*, Victoria, British Columbia Canada, 21-23 Oct. 2001, The Association for Computing Machinery, New York, USA, pp171-178

Suryanto H, Compton P. (2000) *Learning classification taxonomies from a classification knowledge based system.* Proceeding of the First Workshop on Ontology Learning in conjunction with ECAI-2000, Eds. Steffen Staab, Alexander Maedche, Claire Neddellec, Peter Wierner-Hastings, Berlin Germany, 22 Aug. 2000, Berlin, pp1-6

Deitel A., Faron C., and Dieng R. (2001) *Learning ontologies from RDF annotations.* In Proceedings of the IJCAI Workshop in Ontology Learning, Seattle, 2001.

Modica, G., Gal, A. and Jamil, H. M. (2001); *The Use of Machine-Generated Ontologies in Dynamic Information Seeking.* In the Proceedings of the Sixth International Conference on Cooperative Information Systems (CoopIS 2001), Springer-Verlag LNCS series, September 5-7, 2001, Trento, Italy.

Kashyap, V. (1999). Design and Creation of Ontologies for Environmental Information Retrieval. Twelfth Workshop on Knowledge Acquisition, Modelling and Management Voyager Inn, Banff, Alberta, Canada. October, 1999.

Rubin D.L., Hewett M., Oliver D.E., Klein T.E., and Altman R.B. (2002). *Automatic data acquisition into ontologies from pharmacogenetics relational data sources using declarative object definitions and XML.* In: *Proceedings of the Pacific Symposium on Biology*, Lihue, HI, 2002 (Eds. R.B. Altman, A.K. Dunker, L. Hunter, K. Lauderdale and T.E. Klein).

Stojanovic, L.; Stojanovic, N.; Volz R. (2002). Migrating data-intensive Web Sites into the Semantic Web. Proceedings of the 17th ACM symposium on applied computing (SAC), ACM Press, 2002, pp. 1100-1107.

TAMBA-MECZ I., 1994. *La sémantique*. « Que sais-je ? » n° 655. Presse Universitaire de France.

DAVID S., PLANTE P., 1990. *Termino version 1.0*. Rapport de recherche du Centre d'Analyse de Textes par Ordinateurs, Université du Québec à Montréal.

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

SMADJA F., 1993. « Retrieving Collocations from Text : Xtract ». In *Computational Linguistics,* n° 19(1), pp 143-178.

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

TOUSSAINT Y., ROYAUTE J., MULLER C., POLANCO X., 1997. « Analyse linguistique et infométrique pour l'acquisition et la structuration des connaissances ». *Actes des deuxièmes rencontres Terminologie et Intelligence Artificielle (TIA'97)*, pp 27-46. Toulouse.

JOUIS C., BISKRI I., DESCLÈS J-P., LE PRIOL F., MEUNIER J.P., MUSTAFA W., NAULT G., 1997. « *Vers l'intégration d'une approche sémantique linguistique et d'une approche numérique pour un outil d'aide à la construction de bases terminologiques* ». Actes de la première Journée Scientifique et Technique (JST) du réseau Francophone de l'ingénierie de langue de l'AUPELF-UREF, pp 427-432. Avignon.

(Justeson and Katz, 1995) J.J. Justeson et M. Katz. "*Technical terminology: some linguistic properties and an algorithm for identification in text*." Natural Language Engineering Vol.1(1), (1995) pp.9-27.

Church, K., Gale, W., Hanks, P., Hindle, D., (1989). "W*ord Associations and Typical Predicate-Argument Relations*", Proceedings of the 1st International Workshop on Parsing technologies, Carnegie Mellon University.

Langley, P., & Laird, J. E. (2002). "*Cognitive architectures: Research issues and challenges*" (Technical Report). Institute for the Study of Learning and Expertise, Palo Alto, CA.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990) "*Indexing by latent semantic analysis*". JASIS, 41(6), 391-407.

Meunier, J. G. (1996) "*La théorie cognitive: son impact sur le traitement de l'information textuelle*". In V. Rialle et Fisette, D (Ed.), Penser l'Esprit, Des sciences de la cognition à une philosophie cognitive. (pp. 289-305). Grenoble: Presses de L'Université de Grenoble.

Pavel T., (1976) "*Possible worlds in Literary Semantics*" The Journal of Esthetics and Art Criticism, 34: 2.

Dijk, T.V. (1977) "*Text and context*". London: Longman.

Lelu A., M. Halleb & B. Delprat (1998). "*Recherche d'information et cartographie dans des corpus textuels à partir des fréquences de n-grams*", Proceedings of JADT-98, Nice, France.

Manning, C.D., Schütze, H., (1999), "*Foundations of Statistical Natural Language Processing*", MIT Press.

Sebastiani, F. (2002) "Machine learning in automated text categorization". ACM Computing Surveys, 34(1):1-47.

Gelbukh, A., Sidorov, G. and Guzmán-Arenas, A. (1999) "*Text categorization using a hierarchical topic dictionary*". Proc. Text Mining workshop at 16th International Joint Conference on Artificial Intelligence (IJCAI'99), Stockholm, Sweden, July 31 – August 6, 1999, pp. 34-35.

Lebart, L., Salem, A. (1988), "*Analyse statistique des données textuelles*", Paris: Dunod.

Meunier, J.G., Biskri, I., Nault, G., Nyongwa, M. (1997), "*Aladin et le Traitement Connexionniste de l'Analyse Terminologique*", Actes de RIAO-97, Montréal, Canada, 661-664.

Memmi, D., Meunier, JG. and Gabi, K. (1998) "*Dynamical Knowledge extraction from texts by Art Networks*". Proceedings of Neurap. Marseille. 1998. p. 205-210.

Memmi, D. (2000) "*Le modèle vectoriel pour le traitement de documents*", Cahiers Leibniz n° 2000-14.

Benhadid, I., Meunier, J.G., Hamidi, S., Remaki, Z. and Nyongwa, M. (1998), "*Étude Expérimentale Comparative des Méthodes Statistiques pour la Classification des Données Textuelles*", Actes de JADT-98, Nice, France.

Biskri, I., Delisle, S. (1999) "*Un modèle hybride pour le textual data mining : un mariage de raison entre le numérique et le linguistique*" . TALN 99, France, pages 55-64.

Grossberg, S. (1988) "*Neural Network and Natural Intelligence*". Cambridge: MIT Press, 1988.

Srivastava S., Gil De Ladadrid, J. and Elvadapu C.S. (2002) "*Document Ontology: A Statistical Approach*". SSGRR'2002, L'Aquila, Italy.

Harris, Z.S. (1968). *Mathematical Structures of Language*, Wiley & Sons, New York, USA.

Golub, G. H., Reinsch, C. (1969) "*Singular value decomposition and least squares solutions*". Handbook for Automatic Computation, Springer-Varlag, New York, 134-151.

V. Levenshtein. (1996) "*Binary codes capable of correcting deletions, insertions and reversals*".Cybernetics, and Control Theory, 10 (8):707—710, 1996.

A. Maedche and S. Staab. (2002) "*Measuring Similarity between Ontologies*". In: Proc. Of the European Conference on Knowledge Acquisition and Management - EKAW-2002. Madrid, Spain, October 1-4, 2002. LNCS/LNAI 2473, Springer, 2002, pp. 251-263.

Gargouri, Y., Lefebvre, B. et Meunier, J.G. (2003a) «*Maintenance des ontologies à partir d'analyses textuelles*» ACFS'2003, Rimouski, Québec, 21 Mai 2003.

Gargouri, Y., Lefebvre, B. and Meunier, J.G. (2003b) « *Ontology Maintenance using Textual Analysis* », SCI'2003 : The 7th World Multiconference Systems, Cybernetics and Informatics, Orlando, Florida, July 2003 - Systems, Cybernetics and Informatics Journal.