Rint
Réseau international de néologie et de terminologie

19

Terminologie et intelligence artificielle

(actes du colloque de Nantes, 10-11 mai 1999)

Afrique centrale et de l'Est
Afrique de l'Ouest
Canada
Communauté française de Belgique
France
Haïti
Madagascar
Maroc
Québec
République centrafricaine
Suisse
Tunisie
Union latine

Revue semestrielle coéditée par l'Agence de la francophonie et la Communauté française de Belgique

N° 19
décembre 1998
juin 1999

Terminologies nouvelles

Sommaire

Terminologie et intelligence artificielle

actes du colloque TIA 99, Nantes

Éditeurs des actes

Anne Condamines, Équipe de recherche en syntaxe et sémantique, CNRS, Toulouse;
Chantal Enguehard, Institut de recherche en informatique de Nantes, Université de Nantes.

Comité de programme

De nouvelles avenues pour la terminologie

En s'associant au groupe *Terminologie et intelligence artificielle* pour la publication des actes du colloque tenu à Nantes les 10 et 11 mai 1999, le Rint confirme son engagement toujours plus important dans l'informatisation des langues et dans le développement et la mise en œuvre de la terminotique. Les thèmes abordés au cours du colloque *Terminologie et intelligence artificielle (TIA 99)*, correspondent manifestement aux tendances importantes que l'on peut observer dans de nombreux milieux de la terminologie à l'échelle planétaire. Ces tendances se situent à trois niveaux.

Multiplicité des approches

À l'approche classique de la terminologie wüstérienne se sont ajoutées depuis un certain nombre d'années d'autres façons de concevoir la démarche terminologique, notamment sous l'influence de l'école dite «aménagiste » de la terminologie. Il s'agit tout d'abord de l'approche de l'analyse du discours scientifique et technique illustrée par le texte de Didier Bourigault et de Monique Slodzian. Cette approche de la terminologie, qui se fonde sur l'analyse de texte comme point de départ de l'analyse conceptuelle, compte tenu du fait que les concepts naissent, se nomment et se définissent dans le discours, avait déjà été mise en relief notamment à l'occasion d'une table ronde intitulée «Terminologie, discours et textes spécialisés», tenue lors du XV[e] Congrès international des linguistes en 1992[1].

Cette approche s'est amplifiée avec le développement récent de la terminotique qui fournit les outils informatiques nécessaires à la mise en œuvre de cette approche. Le colloque TIA 99 apporte une excellente contribution à la poursuite de la réflexion sur les aspects théoriques et pratiques de l'analyse textuelle. Une autre filière pour l'étude de la terminologie, à laquelle ont fait allusion certains exposés du colloque TIA 99, est l'approche situationnelle qui se caractérise par la démarche socioterminologique et qui s'intéresse aux conditions dans lesquelles les textes scientifiques et techniques sont produits et aux circonstances de communication qui induisent la variation terminologique, qu'elle soit géo- ou sociotechnolectale.

Multiplicité des utilisateurs de la terminologie

Une autre caractéristique importante de la terminologie est la

(1) Kokourek (R.) et Rousseau (L.-J.), 1993: «Terminologie, discours et textes spécialisés», *Actes du XV[e] Congrès international des linguistes*, Québec, Presses de l'Université Laval. Les actes complets de la table ronde ont été publiés dans: ALFA, 1994-1995, volume 7/8.

multiplicité de ses utilisateurs. Aux usagers traditionnels que sont les traducteurs et les rédacteurs techniques s'ajoutent de nos jours de nombreuses catégories d'usagers humains ou d'usagers machines: systèmes experts, aides à la traduction, aides à la rédaction, gestion des données documentaires et textuelles, bases de connaissances, outils d'indexation, etc. Cette multiplicité d'usagers entraîne la conception de gammes variées de produits terminologiques.

Généralisation de l'usage des nouvelles technologies de l'information et de la communication (NTIC)

La généralisation de l'usage des technologies de l'information et de la communication entraîne des mutations profondes sur l'organisation et l'instrumentation du travail terminologique. Les caractéristiques du nouvel environnement du travail qui se dessine sont les suivantes:

• La production de la terminologie est largement informatisée à toutes les étapes du travail, qu'il s'agisse de la veille documentaire, de la phase du traitement ou de la phase de la diffusion;

• La terminologie devient un travail collectif et se caractérise par l'intervention simultanée de plusieurs intervenants, y compris les spécialistes et les usagers; le travail est marqué par l'interaction des acteurs;

• Il est maintenant possible de créer des réseaux informatisés de production et d'échange de données (par exemple, le réseau Balnéo créé par le Rint);

• Le nouvel environnement technologique introduit la possibilité du travail à distance: les intervenants peuvent être dispersés géographiquement et travailler collectivement en temps réel ou en différé;

• L'ordonnancement du travail terminologique peut varier dans le temps: on assiste à la fin de la linéarité de la production;

• Les bases de données terminologiques font l'objet d'une mise à jour permanente. Cette caractéristique fait en sorte qu'il faut définir de nouvelles formes de systémicité du travail terminologique, compte tenu, par exemple, de l'étalement dans le temps du traitement d'ensembles conceptuels. Il faut alors songer à développer, pour la définition de classes de concepts, des schémas définitionnels permettant de traiter celles-ci de façon systémique sans avoir à revenir sur les concepts déjà définis.

• Le poste de travail du terminologue est désormais constitué par la mise en relations de nombreux outils informatisés, ce qui permet l'adaptation du poste aux différentes situations de travail, mais cela pose le problème de l'intégration et de l'interfaçage des systèmes utilisés;

• Sur le plan humain, on assiste à l'appropriation par le terminologue des outils informatisés, ce qui n'est pas sans provoquer des problèmes d'adaptation au nouvel environnement technologique.

La nouvelle chaîne de production de la terminologie comporte les caractéristiques suivantes.

En amont du travail terminologique, on a déjà la possibilité de constituer des corpus de dépouillement à partir de banques de textes existantes (on parle de plus en plus de banques virtuelles de textes) ou à constituer par la numérisation de textes imprimés ou transcrits. Le contenu des corpus est évolutif. Il est de plus en plus courant d'utiliser des outils tels les moteurs de recherche sur Internet pour la veille documentaire, des logiciels d'analyse de textes, etc. Par ailleurs, le terminologue a accès à de nombreuses sources de terminologie (banques et dictionnaires en ligne sur Internet, etc.) ou à des sources de matériaux terminologiques (ex.: *Balnéo*, un outil de veille néologique sur Internet créé par le Rint pour permettre la collecte et l'échange de matériaux terminologiques pour la mise à jour des dictionnaires et des banques de terminologie). On dispose depuis quelques années de logiciels de dépouillement terminologique assisté par ordinateur. Plusieurs de ces catégories d'outils ont été présentées au cours du colloque TIA 99.

Il existe de nombreux logiciels de saisie, de gestion et de traitement des données terminologiques, et les grandes banques de terminologie possèdent leur propre logiciel. Ces logiciels permettent la rédaction des fiches et peuvent comporter des fonctions de comparaison, d'importation et d'exportation de données. Enfin, les banques de terminologie sont également munies d'un gestionnaire de documentation et d'un gestionnaire de thésaurus de domaines intégrés au poste de travail du terminologue.

En aval du travail terminologique, la diffusion peut se faire par tous les médias existants: diffusion sur support électronique (principalement sur cédérom), diffusion sur Internet, diffusion d'imprimés produits par un système de publication électronique. Comme nous l'avons déjà signalé, il y a une diversification importante des formes que prennent les produits ou extraits des bases de données terminologiques. Du point de vue de l'usager, il est de plus en plus question de la possibilité d'interroger simultanément plusieurs bases de données, indépendamment des différents formats d'origine.

De plus, cet environnement informatique n'est plus seulement le fait de quelques équipes

exceptionnelles de terminologues, mais se généralise sous toutes les latitudes grâce à la décroissance régulière des coûts de l'équipement et la disponibilité des logiciels adaptés au traitement de toutes les langues. Le texte de K.M. Brou en est une excellente illustration. De même, au cours des trois dernières années, le Rint et le Riofil(2) ont organisé des stages de formation sur l'informatisation des travaux terminologiques dans les pays du Sud et cette formation s'est concrétisée par la mise en place d'équipes utilisant toutes les catégories d'outils dont il vient d'être question pour le traitement terminologique des langues nationales.

Ces tendances irréversibles de la terminologie font naître de nouveaux champs pour la recherche fondamentale et pour la recherche-développement. Il nous faudra faire en sorte que la recherche s'inspire des besoins manifestés dans la pratique sur le terrain, de façon à produire des outils susceptibles d'améliorer la productivité des travaux terminologiques et d'accroître la qualité des résultats.

Louis-Jean Rousseau,
Secrétaire général du Rint.

(2) Réseau international francophone de l'inforoute et de l'informatisation des langues.

Introduction

Les troisièmes rencontres *Terminologie et intelligence artificielle*, tout comme les premières (1995) et les deuxièmes(1997), ont été organisées à l'initiative du groupe de travail *Terminologie et intelligence artificielle* (TIA). Ce groupe(1), constitué en 1994 par Didier Bourigault et Anne Condamines, réunit des chercheurs en linguistique, terminologie, intelligence artificielle et traitement automatique des langues. Les réunions bimensuelles de ces chercheurs les ont amenés à expliquer leurs objectifs, leurs méthodes et aussi l'histoire de leur discipline. La nécessité d'éclaircir ces éléments pour un public interdisciplinaire oblige à une réflexion distanciée qui a conduit à une évolution parallèle, à l'intérieur du groupe, de la réflexion en intelligence artificielle (IA) et en terminologie et à une remise en question des postulats de ces deux disciplines. Grâce à des rencontres plus larges, organisées tous les deux ans, le groupe s'est donné comme objectif de diffuser les résultats de ses réflexions, tout en créant un lieu de discussions interdisciplinaires fécond sur les possibilités de renouvellement des approches en IA et en terminologie.

Comme les précédentes, les rencontres de Nantes ont connu un vif succès. Plus de quatre-vingts personnes ont participé aux deux journées dont plus du tiers d'informaticiens et près du cinquième de professionnels (responsables de service en ingénierie linguistique ou en documentation, terminologues...). La présence de nombreux étudiants est le signe d'un intérêt croissant pour des problématiques émergentes. On note cependant une contradiction entre un public majoritairement linguiste ou/et terminologue alors que les intervenants sont quasiment tous informaticiens. Plusieurs éléments d'explication à ce phénomène. Tout d'abord, il se peut que les travaux de linguistes/terminologues soient moins avancés que ceux des informaticiens, sur un sujet où tout est à faire : renouveler la réflexion sur la terminologie et mettre en place une linguistique de corpus. Ensuite, le terme *intelligence artificielle* fait peur et éloigne des travaux essentiellement linguistiques qui auraient pourtant toute leur place dans ces rencontres. Enfin, le besoin d'information et de formation est patent pour les terminologues ou même les linguistes qui sont directement confrontés aux demandes des entreprises.

Sans atteindre celui des linguistes/terminologues, le nombre des chercheurs en IA a toutefois augmenté cette année, conséquence à la fois d'une prise de conscience de l'importance de la problématique

(1) Parrainé par l'Association française d'intelligence artificielle (AFIA), ce groupe a été subventionné par le Programme de recherche coordonné en intelligence artificielle (PRC IA) et l'est, à présent, par le Programme de recherche coordonné information, interaction, intelligence (PRC I3).

Introduction

pour la discipline (le nombre élevé de thèses qui devraient être soutenues prochainement est un autre indice de cette prise de conscience) et du fait que la conférence était organisée par une équipe d'informaticiens (l'équipe *Langage naturel* de l'Institut de recherche en informatique de Nantes). Les nombreuses questions qui ont suivi chaque exposé témoignent du dynamisme de la communauté et de son désir de participer aux débats.

Ces troisièmes rencontres ont reçu le soutien du Rint, qui a accepté de publier les actes dans sa revue *Terminologie nouvelles*. Ce parrainage pour la publication des actes, tout comme la présence lors des rencontres du secrétaire général, Louis-Jean Rousseau, constituent un encouragement à développer une réflexion alternative sur la terminologie examinée dans son fonctionnement textuel, réflexion qui garantisse à la fois une approche scientifique et une réponse aux besoins des entreprises.

Les deux jours ont été organisés de la manière suivante: un tutoriel présentant la terminologie textuelle, deux conférences invitées, des présentations sélectionnées par un Comité de programme et des démonstrations d'outils (six outils étaient présentés).

Le tutoriel détaillé et très argumenté de Didier Bourigault et Monique Slodzian, qui a débuté les rencontres, a permis de présenter l'évolution inéluctable de la terminologie telle qu'elle est perçue par le groupe TIA et de tracer les grandes lignes de ce qui apparaît comme une alternative à la terminologie traditionnelle, la terminologie textuelle.

La première conférence invitée a été assurée par Asuncion Gomez Perez, de l'Universidad Politecnica de Madrid. A. Gomez Perez a fait le tour des méthodes et des projets concernant les ontologies. Elle a ainsi dressé un état des travaux les plus importants en nombre, travaux qui, malheureusement, ne prennent que rarement en compte la dimension textuelle des connaissances qui sont représentées.

La deuxième conférence invitée a, elle, été assurée par Jennifer Pearson, de l'Université de Dublin. En écho au tutoriel, J. Pearson a insisté sur le rôle des corpus en terminologie et sur la nécessité de définir des critères de constructions et des méthodes d'exploration des données textuelles.

Dix communications avaient été sélectionnées par le Comité de programme. Elles peuvent être organisées en quatre rubriques.

1. Présentation d'outils de gestion ou/et d'aide à la constitution de terminologies: trois articles peuvent être rassemblés sous cette rubrique: celui de A. Hocine et K.M. Brou (« *CoDB-Web*: un système de représentation et de gestion hypertexte de concepts »), celui de C. Roche, J.-C. Marty et S. Lacroix (« Ontologie et terminologie: le modèle *OK* ») et celui de N. Aussenac (« *Gediterm*: un logiciel pour gérer les bases de connaissances terminologiques »).

2. Description de méthodes pour acquérir des données terminologiques: article de T. Hamon, D. Garcia, A. Nazarenko (« Détection de liens de synonymie: complémentarité des ressources générales et spécialisées ») et article de P. Séguéla (« Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés »).

3. Compte-rendus d'expériences sur la terminologie (remarquons qu'elles utilisaient toutes le logiciel Lexter, de Didier Bourigault): article de Y. Abbas et M.L. Piccard («Exemple de pratique terminographique en entreprise») , article de D. Bourigault, C. Chodkiewicz et John Humbley (« Construction d'un lexique bilingue des droits de l'homme à partir de l'analyse automatique d'un corpus aligné »), article de A. Chrobot (« Extraction terminologique en anglais basé sur des dictionnaires généraux et spécialisés »).

4. Utilisation de méthodes de terminologie textuelle pour la veille terminologique: deux articles sont concernés, celui de B. Goujon (« Extractions d'informations techniques par l'exploitation de notions indépendantes d'un domaine pour la veille ») et celui de T. Poibeau (« Le repérage des entités nommées: un enjeu pour la veille technologique »).

Ces différentes interventions sont données dans leur intégralité dans ce numéro de *Terminologies nouvelles*.

Anne Condamines,
Présidente du Comité de programme de TIA 99,
Équipe de recherche en syntaxe et sémantique,
CNRS,
Toulouse,
France.

Développements récents en matière de conception, de maintenance et d'utilisation des ontologies

Le présent article offre une synthèse des développements récents survenus dans le domaine de l'ingénierie ontologique: les bases théoriques, les ontologies les plus connues, les méthodologies et les environnements logiciels disponibles pour la création d'ontologies, ainsi que l'utilisation d'ontologies dans des applications à des fins commerciales et de recherche.

Terme-clé: ontologies

1 Introduction

En 1991, le *Knowledge Sharing Effort* de l'Arpa (Agence pour les projets de recherche avancés) (Neches *et al.* 1991) remet entièrement en question la méthode de conception de systèmes intelligents lorsqu'il déclare que «dans l'état actuel des choses, la construction de systèmes intelligents implique habituellement la mise au point, à partir de zéro, de bases de connaissances, alors que cela pourrait être réalisé en assemblant des composants réutilisables. Les constructeurs de systèmes devraient alors uniquement se consacrer à la conception de nouvelles connaissances et de systèmes de raisonnement spécifiques adaptés à la tâche de leur système en vue de leur faire exécuter une partie du raisonnement. Ainsi, la connaissance déclarative, les techniques de résolution de problèmes et les services de raisonnement constitueraient un fonds commun exploitable par tous les systèmes. En procédant de cette manière, il serait possible de mettre au point des systèmes plus performants et moins onéreux…»

Depuis lors, les bases conceptuelles de la construction de technologies qui permettent tant la réutilisation que la mise en commun de composants de connaissances ont fortement évolué. Des méthodes de résolution de problèmes (*Problem Solving Methods – PSM*) et des ontologies ont été mises au point dans le but de partager et de réutiliser les connaissances ainsi que les raisonnements de différents secteurs d'activité et de production. Les ontologies renvoient aux connaissances statiques d'un domaine alors que les méthodes de résolution de problème reflètent des connaissances dynamiques de raisonnement. L'intégration d'ontologies et de PSM constitue une solution envisageable au problème d'interaction (Bylander *et al.* 1998), à savoir que la représentation de connaissances à des fins de résolution de problèmes est fortement influencée par la nature du problème, d'une part, et par la stratégie de déduction à appliquer au problème, d'autre part. Les ontologies et les PSM offrent l'avantage de pouvoir expliciter cette interaction et de la prendre en compte.

Le présent document a pour objectif de répondre aux questions suivantes: qu'est-ce qu'une ontologie? Quels sont les principes à suivre pour construire une ontologie? Quels sont les composants d'une ontologie? Quels types d'ontologies existent déjà? Comment les ontologies sont-elles organisées en bibliothèques? Quelles méthodes dois-je suivre afin

de mettre au point ma propre ontologie? Quelles sont les techniques adéquates pour chaque étape? Comment les outils logiciels soutiennent-ils les processus de construction et d'utilisation des ontologies? Quelles sont les ontologies les plus connues? Quelles sont les utilisations des ontologies? Quels critères dois-je prendre en compte pour choisir l'ontologie appropriée à mon application? etc.

En vue de répondre à ces questions, l'article est constitué comme suit: présentation des bases théoriques de l'ingénierie ontologique, et ensuite, des ontologies les plus connues, des méthodologies, des outils et des langages de conception d'ontologies, et finalement quelques utilisations d'ontologies dans des applications.

2 Bases théoriques

2.1 Qu'est-ce qu'une ontologie?

Ce terme est issu du domaine de la philosophie, où il signifie «explication systématique de l'existence». Dans le cadre de l'intelligence artificielle, Neches et ses collègues (Neches *et al.* 1991) furent les premiers à en proposer une définition, à savoir: «une ontologie définit les termes et les relations de base du vocabulaire d'un domaine ainsi que les règles qui indiquent comment combiner les termes et les relations de façon à pouvoir étendre le vocabulaire». Cette définition nous dit comment élaborer une ontologie, en nous offrant des directives relativement floues: repérer les termes de base et les relations entre les termes, identifier les règles servant à les combiner, fournir des définitions de ces termes et de ces relations. Notons que d'après cette définition, une ontologie inclut non seulement les termes qui y sont explicitement définis, mais aussi les termes qui peuvent être créés par déduction en utilisant les règles. En 1993, Gruber (Gruber 1993) formule la définition suivante, à savoir «une ontologie est une spécification explicite d'une conceptualisation», qui deviendra célèbre et restera la définition la plus citée dans la littérature scientifique. En 1997, Borst (Borst 1997) apporte une légère modification à la définition de Gruber en précisant que «les ontologies se définissent comme une spécification formelle d'une conceptualisation commune». Studer et ses collègues (Studer *et al.* 1998) ont donné l'interprétation suivante de ces deux définitions: «la conceptualisation renvoie à un modèle abstrait d'un quelconque phénomène après en avoir relevé les concepts significatifs». Par *explicite*, il faut entendre que le type de concepts utilisés, ainsi que leurs contraintes d'utilisation, sont définies de façon explicite; quant à l'adjectif *formel*, il exprime le fait que l'ontologie doit être lisible par ordinateur. *Commun* renvoie à l'idée qu'une ontologie rend compte d'un savoir consensuel, c'est-à-dire qu'elle n'est pas l'objet d'un individu, mais qu'elle est reconnue par un groupe».

De nombreuses définitions ont été appliquées à l'ontologie après celle de Gruber. En 1995, Guarino et Giaretta (Guarino *et al.* 1995) ont recueilli sept définitions dans la littérature scientifique et en ont fourni des interprétations sémantiques. D'autres auteurs, cependant, offrent des définitions fondées sur l'approche qu'ils ont adoptée pour construire leurs ontologies. Selon Swartout et ses collègues (Swartout *et al.* 1997), «une ontologie est un ensemble de termes structurés de façon hiérarchique, conçu afin de décrire un domaine et qui peut servir de charpente à une base de connaissances». Cette définition se base sur le fait qu'ils construisent des ontologies de connaissances spécifiques à des domaines d'expertise en identifiant les termes significatifs d'un certain domaine de l'ontologie *Sensus* (qui inclut plus de 50 000 termes). Ils affinent ensuite cette dernière à l'aide d'une sorte d'heuristique. Bernaras et ses condisciples (Bernaras *et al.* 1996) construisent une ontologie différemment en partant d'une base de connaissances, qui sera raffinée et enrichie de nouvelles définitions dans le cas où de nouvelles applications sont créées. Nous proposerons en conséquence la définition suivante: «une ontologie fournit les moyens de décrire de façon explicite la conceptualisation des connaissances représentées dans une base de connaissances».

Pour conclure cette section, nous pouvons donc affirmer que les définitions du terme ontologie abondent dans la littérature scientifique. Les définitions, dans leur diversité, offrent des points de vue à la fois différents et complémentaires sur un même concept.

2.2 Principes à suivre dans le cadre de l'élaboration d'une ontologie

Nous résumons ici certains critères conceptuels et un ensemble de principes qui se sont avérés efficaces dans le domaine de la conception d'ontologies.

- Clarté et objectivité (Gruber 1993): l'ontologie devrait fournir le sens des termes définis en offrant des définitions objectives ainsi que de la documentation en langage naturel.
- Exhaustivité (Gruber 1993): une définition exprimée par une condition nécessaire et suffisante est préférable à une définition exprimée seulement par une condition nécessaire ou par une condition suffisante.

• Cohérence (Gruber 1993): afin de pouvoir formuler des inférences cohérentes avec les définitions.
• Extensibilité monotone maximale (Gruber 1993): les nouveaux termes, qu'ils relèvent de la langue générale ou d'une langue de spécialité, devraient être inclus dans l'ontologie sans entraîner de modifications dans les définitions existantes.
• Interventions ontologiques minimales (Gruber 1993): intervenir le moins possible sur le monde en phase de modélisation. L'ontologie devrait spécifier le moins possible le sens de ses termes, de façon à ce que les parties impliquées dans l'ontologie aient les mains libres pour spécialiser et instancier l'ontologie à leur guise.
• Principe de distinction ontologique (Borgo *et al.* 1996): les classes d'une ontologie doivent être séparées. Le critère d'identité sera utilisé afin d'isoler le noyau des propriétés jugées invariables pour une instance d'une classe.
• Diversification des hiérarchies afin d'optimaliser la puissance dérivant des mécanismes d'héritage multiple (Arpírez *et al.* 1998). Il est d'autant plus facile d'intégrer de nouveaux concepts (dans la mesure où ils peuvent être définis sur la base des concepts et des critères de classification préexistants) et d'hériter de propriétés de différents points de vue que le volume de connaissances inclus dans l'ontologie est suffisant et que l'éventail de critères de classification est large.
• Modularité (Bernaras *et al.* 1996): afin de minimiser le couplage entre modules.
• Minimisation de la distance sémantique entre des concepts frères (Aspírez *et al.* 1998): les concepts proches sont regroupés et représentés dans des sous-classes d'une même classe et devraient être définis en ayant recours aux mêmes primitives, alors que les concepts plus éloignés sont éclatés dans la hiérarchie.
• Normalisation systématique des noms dans la mesure du possible (Arpírez *et al.* 1998).

2.3 Composants des ontologies

Comme mentionné plus haut, les ontologies fournissent le vocabulaire commun d'un domaine et définissent, de façon plus ou moins formelle, le sens des termes et les relations entre ces derniers. Les connaissances intégrées dans les ontologies sont formalisées en mettant en jeu cinq types de composants: les classes, les relations, les fonctions, les axiomes et les instances (Gruber 1993). Les classes dans l'ontologie sont habituellement organisées en taxonomies. Il arrive que les définitions des ontologies aient été diluées, en ce sens que les taxonomies sont considérées comme des ontologies complètes (Studer *et al.* 1998).
• Les concepts sont utilisés dans leur sens large. Ils peuvent être abstraits ou concrets, élémentaires (électron) ou composés (atome), réels ou fictifs. En résumé, un concept peut être tout ce qui peut être évoqué et, partant, peut consister en la description d'une tâche, d'une fonction, d'une action, d'une stratégie ou d'un processus de raisonnement, etc.
• Les relations représentent un type d'interaction entre les notions d'un domaine. Elles sont formellement définies comme tout sous-ensemble d'un produit de n ensembles, c'est-à-dire R: C1 x C2 x … x Cn. Des exemples de relations binaires sont *sous-classe-de* ou encore *connecté-à*.
• 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 sont définies ainsi: F: C1 x C2 x … x Cn-1 ➜ Cn. Comme exemple de fonctions binaires, nous avons la fonction *mère de* et le carré, et comme exemple de fonction ternaire, le prix d'une voiture usagée sur lequel on peut se baser pour calculer le prix d'une voiture d'occasion en fonction de son modèle, de sa date de construction et de son kilométrage.
• On a recours aux axiomes pour structurer des phrases qui sont toujours vraies.
• Des instances sont utilisées pour représenter des éléments.

Maintenant que les principaux éléments des ontologies ont été présentés, la question qui se pose est de savoir à quoi ressemble une ontologie explicite. Uschold et Grüninger (Uschold *et al.* 1996) ont distingué quatre sortes d'ontologies en fonction du type de langage utilisé: les ontologies hautement informelles (écrites en langage naturel), les ontologies semi-formelles (exprimées dans un langage naturel structuré et limité, c'est-à-dire que des patrons ont été mis en œuvre), les ontologies semi-formelles (définies dans un langage défini artificiellement et formellement) et les ontologies rigoureusement formelles (définies dans un langage contenant une sémantique formelle, des théorèmes et des preuves de propriétés telles que la robustesse et l'exhaustivité).

2.4 Types d'ontologies mises au point

Cette section n'a pas l'ambition de fournir une typologie exhaustive des ontologies telle que celles de van Heijst (*et al.* 1997) et Mizoguchi (*et al.* 1995). Elle présente néanmoins les types d'ontologies les plus couramment utilisés afin de permettre au lecteur d'avoir une idée des connaissances à inclure dans chaque type d'ontologie. En gros, on identifie les catégories suivantes: les ontologies de représentation des connaissances, les méta-ontologies, les ontologies de domaine, les ontologies de tâches, les

ontologies de domaine-tâche, les ontologies d'application, les ontologies d'index, les ontologies interactives, etc.

• Les ontologies de représentation de connaissances (van Heijst *et al.* 1997) regroupent les primitives de représentation utilisées afin de formaliser les connaissances selon des paradigmes de représentation des connaissances. L'exemple le plus représentatif de ce type d'ontologie est la *Frame-Ontology* (Gruber 1993), qui rassemble les primitives de représentation (classes, instances, cases, facettes, etc.) utilisées dans les langages à base de *frames*.

• Les ontologies générales/communes (Mizoguchi *et al.* 1995) incluent le vocabulaire lié aux objets, aux événements, au temps, à l'espace, à la causalité, au comportement, à la fonction, etc.

• Les méta-ontologies, également appelées ontologies génériques ou noyaux d'ontologies, (van Heijst *et al.* 1997) sont réutilisables dans différents domaines. L'exemple le plus représentatif serait une ontologie méréologique (Borst 1997), qui inclurait le terme *partie de*.

• Les ontologies de domaine (Mizoguchi *et al.* 1995) (van Heijst *et al.* 1997) sont réutilisables dans un domaine donné. Elles fournissent le vocabulaire des concepts d'un domaine (par ex., scalpel, scanner dans un domaine médical) et les relations entre ces derniers, les activités de ce domaine (par ex., anesthésier, accoucher) ainsi que les théories et les principes de base de ce domaine.

• Les ontologies de tâche (Mizoguchi *et al.* 1995) fournissent un vocabulaire systématisé des termes utilisés pour résoudre les problèmes associés à des tâches qui peuvent appartenir ou non à un même domaine. Ces ontologies fournissent un ensemble de termes au moyen desquels on peut décrire au niveau générique comment résoudre un type de problème. Elles incluent des noms génériques (par ex., plan, objectif, contrainte), des verbes génériques (par ex., assigner, classer, sélectionner), des adjectifs génériques (par ex., assigné) et d'autres mots qui relèvent de l'établissement d'échéances.

• Les ontologies de domaine-tâche sont des ontologies de tâches réutilisables dans un domaine donné, mais pas dans différents domaines. Par exemple, une ontologie domaine-tâche dans le domaine médical pourrait inclure les termes liés au timing d'une intervention chirurgicale: planifier – intervention chirurgicale.

• Les ontologies d'application (van Heijst *et al.* 1997) contiennent suffisamment de connaissances pour structurer un domaine particulier.

Les méta-ontologies, les ontologies de domaine et les ontologies d'application saisissent les connaissances statiques indépendamment de la façon dont on résout les problèmes alors que les ontologies de PSM, les ontologies de tâches et les ontologies domaine-tâche sont axées sur les connaissances visant à résoudre des problèmes. Tous ces types d'ontologie peuvent être combinés de façon à construire une nouvelle ontologie. Si l'on applique le problème du compromis entre l'utilisabilité et la réutilisabilité (Klinker *et al.* 1991) au domaine de l'ontologie, on peut affirmer que plus une ontologie est réutilisable, moins elle est utilisable, et inversement.

3 Les ontologies les plus connues

Il est actuellement facile de recevoir des informations des organisations qui ont des ontologies sur le WWW. De nombreuses ontologies telles que les ontologies *Ontolingua* sur le serveur *Ontolingua*(1) (Farquhar *et al.* 1996) et *Wordnet*(2) (Miller 1990) à Princenton sont disponibles gratuitement sur la toile. D'autres ontologies, telles que les ontologies de *Cyc*(3) (Lenat *et al.* 1990) sont partiellement disponibles gratuitement sur le web. La majorité d'entre elles, cependant, ont été mises au point par des compagnies pour leur propre utilisation et ne sont donc pas disponibles. La *Ontology Page*(4) (également connue sous le nom de *Top*) et (*Onto*) *2Agent*(5) (Arpírez *et al.* 1998) (un moteur de recherche sur la toile s'appuyant sur une ontologie et qui aide à sélectionner des ontologies) peuvent aider à choisir des ontologies. Cette section introduit les ontologies les plus connues en prenant en compte la typologie d'ontologies énoncée ci-dessus.

L'exemple le plus représentatif des ontologies de représentation des connaissances est *la Frame Ontology* (Gruber 1993). Elle saisit les primitives de représentation utilisées dans les langages de *frame*, telles que les classes, les attributs des sous-classes, les partitions de classes, les relations et les axiomes. Elle permet de codifier d'autres ontologies en ayant recours aux conventions habituelles des frames. Elle est implémentée en *Kif 3.0* (Genesereth *et al.* 1992) et constitue le matériau de construction de base des traducteurs d'*Ontology Server*.

---

(1) http://www-ksl.stanford.edu:5915 et http://www-ksl-svc-lia.dia.fi.upm.es:5915

(21) http://www.tio.darpa.mil/Summaries95/B370-Princenton.html

(3) http://www.cyc.com/

(4) http://www.medg.lcs.mit.edu/doyle/top

(5) http://delicias.dia.fi.upm.es/OntoAgent/

Les ontologies de haut niveau fournissent des concepts généraux à partir desquels tous les termes des ontologies existantes peuvent être définis. Citons le treillis booléen de Sowa (Sowa 1997), le *Penman Upper Level* (Bateman *et al.* 1990), *Cyc* (Lenat *et al.* 1990), la proposition de très haut niveau de Guarino (Guarino 1997), etc. En outre, des travaux sur une sorte d'ontologie «normalisée» de haut niveau ont été entamés au sein de l'Ansi en 1996.

L'*Ontologie méréologique* (Borst 1997) pourrait être l'exemple typique d'une méta-ontologie. Cette ontologie définit la relation *partie-de* et ses propriétés. Cette relation permet d'exprimer que des instruments sont formés de composants, qui peuvent eux-mêmes être constitués d'éléments plus petits.

L'*Ontologie Cyc* (Lenat *et al.* 1990) est une ontologie de sens commun qui fournit une grande quantité de savoir humain élémentaire. Elle consiste en un ensemble de termes et d'affirmations liées à ces termes. Elle se décompose, par ailleurs, en différentes microthéories. Chaque microthéorie rend compte seulement d'un point de vue important d'un domaine de connaissances. Certains domaines peuvent traiter plusieurs microthéories, qui représentent différentes perspectives et affirmations, divers niveaux de granularité et de distinction. Les *Ontologies Cyc* sont implémentées dans le langage *CycL*.

Le *Generalized Upper model* (Bateman *et al.* 1995), *Wordnet* (Miller 1990), et *Sensus* (Swartout *et al.* 1997) représentent le mieux les ontologies linguistiques. Le *Generalized Upper model* (*Gum*)(6) est une ontologie linguistique générale, indépendante de tout domaine et de tout type de tâche. Afin de pouvoir la transférer dans différentes langues, il a été prévu que l'ontologie *Gum* inclue seulement les notions linguistiques principales et leur organisation dans toutes les langues; elle omet ainsi les détails qui différencient les langues. Cette philosophie a permis d'utiliser *Gum* pour créer des ontologies pour des langues spécifiques, telles que l'anglais, l'allemand, l'espagnol et l'italien en rajoutant les traits sémantiques propres à chaque langue. L'ontologie a été implémentée en *Loom*. *WordNet* est une base de données lexicale pour l'anglais fondée sur des principes psycholinguistiques. Ses informations sont ventilées en unités appelées «synsets» en anglais, qui sont des jeux de synonymes interchangeables dans un contexte particulier utilisés pour représenter différents sens. *WordNet* contient une série de paires (*w, m*) où *w* est une série de caractères Ascii et *m* un élément d'un ensemble de sens, ou *synset*. Les *synsets* sont accompagnés dans leur plus grand nombre de glossaires explicatifs, et ils sont organisés dans un réseau sur la base de relations sémantiques, au nombre desquelles: l'antonymie, l'hyponymie, la métonymie, l'implication. *WordNet* se compose de quatre réseaux qui représentent les catégories syntaxiques principales: noms, verbes, adjectifs et adverbes. *Sensus* est une ontologie basée sur le langage naturel qui a pour fonction de fournir une vaste structure conceptuelle aux travaux menés en matière de traduction automatique. Il a été mis au point en rassemblant et en extrayant des données de ressources électroniques telles que: *Penman Upper model*, *Ontos*, *WordNet* et des dictionnaires électroniques de langages naturels. Il compte plus de 50 000 notions.

Dans le domaine des ontologies d'ingénierie, les ontologies *EngMath* (Gruber *et al.* 1994) et *PhysSys* (Borst 1997) méritent une attention particulière. *EngMath* est une ontologie *Ontolingua* mise au point pour la modélisation mathématique en ingénierie. Elle inclut des bases conceptuelles pour des grandeurs scalaires, vectorielles et tensorielles, des dimensions physiques, des unités de mesure, des fonctions sur les quantités et des quantités de dimensions. *PhysSys* est une ontologie d'ingénierie destinée à modéliser, simuler et concevoir des systèmes physiques. Elle comporte trois ontologies d'ingénierie qui formalisent les trois points de vue sur les outils physiques: présentation du système, comportement de processus physique et relations mathématiques descriptives. Trois ontologies d'ingénierie formalisent chacun de ces trois points: une ontologie de composants, une ontologie de processus et l'ontologie *EngMath*. Les interdépendances entre ces ontologies sont formalisées comme des projections d'ontologies. Ces ontologies mettent en œuvre d'autres méta-ontologies: méréologie, topologie et théories des systèmes.

Les ontologies qui représentent le mieux les ontologies dédiées à la modélisation d'entreprises sont l'*Enterprise Ontology* (Uschold *et al.* 1996) et la *Tove Ontology* (Gruninger *et al.* 1995). L'*Enterprise Ontology* (7) est un ensemble de termes et de définitions pertinent pour les entreprises commerciales et inclut des connaissances sur les activités et les processus, les organisations, les stratégies, le marketing, etc. Les ontologies élaborées dans le cadre du projet *Tove* (Toronto Virtual Enterprise)(8) sont l'ontologie de conception d'entreprises, l'ontologie des projets, l'ontologie-agenda, ou encore l'ontologie des services.

L'ontologie *(KA)2*(9) (Benjamins *et al.* 1999) constitue un

(6) http://www.darmstadt.gmd.de/publish/komet/gen-um/newUM.html

(7) http://www.aiai.ed.ac.uk/project/enterprise

(8) http://www.ie.utoronto.ca/EIL

(9) http://www.aifb.uni-karlsruhe.de/WBS/broker/KA2.html

bon exemple d'ontologie dédiée à la gestion de connaissances, qui sera utilisée par le *Knowledge Annotation Initiative* de la communauté d'acquisition des connaissances. Cette ontologie servira de base pour annoter les documents sur internet de la communauté d'acquisition des connaissances de façon à fournir un accès intelligent à ces documents. Des spécialistes situés dans des zones géographiques différentes travaillent ensemble à la mise au point de cette ontologie.

4 Méthodologies

L'élaboration d'ontologies relève plus du savoir-faire que de l'ingénierie. 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 directives et de méthodes consensuelles entrave, d'une part, le développement d'ontologies communes et acceptées par les équipes et entre elles, et d'autre part, l'extension d'une ontologie donnée à partir d'autres, sa réutilisation dans d'autres ontologies et dans des applications finales.

Les constructeurs d'ontologies ont l'habitude de passer directement de l'acquisition de données à l'implémentation, ce qui n'est pas sans causer quelques problèmes : les modèles conceptuels des ontologies sont implicites dans les codes d'implémentation ; les interventions ontologiques et les critères d'élaboration sont à la fois implicites et explicites dans le code de l'ontologie ; il est impossible pour les experts et les utilisateurs finaux de décrypter les ontologies formelles codées dans des langages ontologiques ; comme c'est le cas pour les bases de connaissances traditionnelles. Le codage direct du résultat de l'acquisition de connaissances est trop abrupt, spécialement en ce qui concerne les ontologies complexes ; les préférences des concepteurs d'ontologies pour un certain langage conditionnent l'implémentation des connaissances acquises ; et les personnes qui mettent au point des ontologies (qui ne connaissent pas les langages dans lesquels les ontologies sont codées ou qui ne les maîtrisent pas bien) peuvent trouver difficile de comprendre les ontologies implémentées ou même d'en construire une nouvelle dans la mesure où les outils ontologiques traditionnels sont trop focalisés sur les problèmes d'implémentation et pas assez sur la conception.

4.1 Méthodes à suivre pour élaborer sa propre ontologie

Le processus d'élaboration d'ontologies se réfère aux tâches à accomplir pour construire des ontologies (Fernandez *et al.* 1997). Les activités se répartissent en trois groupes : la gestion du projet, le développement et les activités intégrales. La gestion du projet concerne le bon déroulement du processus, qui inclut des tâches de l'ordre de la planification, de la supervision et de l'assurance qualité. Le développement consiste à construire l'ontologie en travaillant à sa spécification, à sa conceptualisation, à sa formalisation, à son implémentation et à sa maintenance. Quant aux activités intégrales, elles servent à soutenir le développement et incluent la gestion de l'acquisition de connaissances, l'intégration, l'évaluation, la documentation et la gestion de configuration. Si les ontologies sont de taille réduite, il est possible de supprimer quelques tâches. Par contre, s'il s'agit de construire des ontologies à grande échelle correctes et exhaustives, il convient d'éviter des constructions anarchiques.

La méthodologie d'Uschold et Kings (Uschold *et al.* 1995) se fonde sur l'expérience de la construction de l'*Enterprise Ontology*, qui inclut un ensemble d'ontologies pour la modélisation d'entreprises. Ils proposent les étapes suivantes : (1) identification du but et de l'étendue de l'ontologie, (2) construction de l'ontologie en consignant et en codant des connaissances, ainsi qu'en les intégrant à des ontologies existantes, (3) évaluation, (4) documentation, et (5) établissement de directives pour chaque étape.

Quant à la méthodologie de Grüninger et Fox (Grüninger *et al.* 1995), elle se base sur l'expérience de la construction d'une ontologie de modélisation d'entreprise dans le cadre du projet *Tove*. Il s'agit essentiellement de la construction d'un modèle logique des connaissances à inclure dans l'ontologie. Ce modèle n'est pas construit directement. Les spécifications que doit comprendre l'ontologie sont décrites de façon informelle en identifiant un ensemble de questions de compétence, et cette description est ensuite formalisée dans un langage basé sur la logique des prédicats. Les questions de compétence constituent l'élément-clé qui permet de caractériser de façon rigoureuse les connaissances que doit inclure une ontologie et elles spécifient le problème ainsi que ce qui constituerait une bonne solution au problème. Par un mécanisme de composition et de décomposition, on peut utiliser les questions de compétence et leurs réponses pour répondre à des questions de compétence plus complexes qui figurent dans d'autres ontologies, et permettre ainsi d'intégrer d'autres ontologies.

Le cadre *Methontology* (Gómez-Perez 1998 et Fernandez *et al.* 1999) permet de construire des ontologies

au niveau des connaissances. Il inclut l'identification du processus de développement ontologique, une proposition de cycle de vie et la méthodologie elle-même. Le processus de développement ontologique identifie les tâches à accomplir lorsque l'on construit une ontologie (planification, supervision, assurance qualité, spécification, acquisition de connaissances, conceptualisation, intégration, formalisation, implémentation, évaluation, maintenance, documentation et gestion de configuration). Le cycle de vie basé sur l'évolution des prototypes identifie les phases de l'évolution de l'ontologie. Finalement, la méthodologie elle-même spécifie les étapes à suivre pour réaliser chaque activité, les techniques utilisées, les produits à être mis au point ainsi que la façon de les évaluer. La phase de conceptualisation constitue la principale phase du processus d'élaboration de l'ontologie selon l'approche de la *Methontology*. Au cours des phases de spécification et de conceptualisation, on a achevé un processus d'intégration en utilisant des ontologies réalisées en interne ou à l'extérieur. Ce cadre est soutenu en partie par l'*Ontology Design Environment* (ODE) (Blázquez *et al.* 1998) (Fernandez *et al.* 1999), un environnement logiciel. Plusieurs ontologies ont été mises au point en ayant recours à cette méthodologie: *Chemicals*, une ontologie spécialisée dans les produits chimiques; les ontologies des polluants environnementaux (Gómez-Pérez *et al.* 1999) qui représentent les méthodes de détection des composantes de différents polluants de plusieurs environnements: eau, air, sol, etc., ainsi que les concentrations maximales autorisées de ces composants, en tenant compte de la législation en cours (réglementations de l'Union européenne, de l'Espagne, de l'Allemagne, des États-Unis, etc.); la *Reference-Ontology* (Arpírez *et al.* 1998), une ontologie qui constitue une sorte de pages jaunes des ontologies; ainsi que la version restructurée de l'ontologie *(KA)2* (Blásquez *et al.* 1998). La *Foundation for Intelligent Physical Agents (Fipa)*(10), qui promeut l'interopérabilité entre les applications programmées à l'aide d'agents, a proposé de recourir à cette méthodologie pour construire des ontologies.

L'identification du but recherché et le besoin d'acquisition de connaissances relatives à un domaine constituent le point de départ commun à toutes ces méthodologies. Néanmoins, après l'assimilation d'une quantité impressionnante de connaissances, la méthodologie d'Uschold propose de coder dans un langage formel et *Methontology* suggère d'exprimer l'idée sous la forme d'un ensemble de *représentations intermédiaires (RI)*. Ensuite, l'ontologie est produite par des traducteurs. Ces *RI* comblent l'espace entre la façon dont les gens perçoivent un domaine et les langages de formalisation des ontologies. Ces représentations intermédiaires offrent une approche conviviale tant pour l'acquisition de connaissances que pour l'évaluation effectuée par des informaticiens et des spécialistes qui ne sont pas des cogniticiens (Aguado *et al.* 1998).

Les trois méthodologies susmentionnées révèlent également la nécessité d'évaluer les ontologies (Gómez-Perez 1996). Si la méthodologie d'Uschold prévoit cette opération, elle ne précise pas la façon dont elle devrait être menée. Grüninger et Fox proposent d'identifier un ensemble de questions de compétence. Dès que l'ontologie a été exprimée de façon formelle, elle est comparée à cet ensemble de questions de compétence. Finalement, *Methontology* propose que l'évaluation soit effectuée tout au long du processus d'élaboration d'ontologies. La plus grande partie de l'évaluation est menée durant la phase de conceptualisation.

(10) http://www.fipa.org

5 Langages et environnements pour la construction d'ontologies

5.1 Langages les plus couramment utilisés pour construire des ontologies.

En gros, plusieurs systèmes de représentation ont été présentés ici pour formaliser des ontologies à l'aide d'une approche basée sur les *frames*, sur la logique des prédicats ou les deux. Les langages les plus représentatifs sont *Ontolingua* (Gruber 1993), *CycL* (Lenat *et al.* 1990), *Loom* (MacGregor 1991) et *Flogic* (Kifer *et al.* 1995).

Ontolingua est un langage basé sur *Kif* et sur la *Frame Ontology*; il est en outre le langage de construction d'ontologies utilisé par *l'Ontology Server. Kif (Knowledge Interchange Format)* est une *interlingua* qui intègre une sémantique déclarative, qui a une force expressive suffisante en général pour représenter la connaissance déclarative contenue dans la base de connaissances du système des applications, et une structure qui a rendu possible de procéder à des traductions semi-automatiques à partir de/ vers des langages de représentation classiques. Il s'agit d'une version préfixée de la logique des prédicats, qui comporte des extensions servant à optimaliser sa force expressive, au nombre desquelles figurent la définition des termes, la représentation de la connaissance sur la connaissance, la réification des fonctions et des relations, la spécification d'ensembles et le

raisonnement non monotone. La *Frame Ontology*, qui, comme mentionné plus haut, est une ontologie de représentation de connaissances destinée à modéliser des connaissances dans une approche s'appuyant sur des *frames*, a été construite sur la base de *Kif* et d'une série d'extensions à ce langage.

Le langage *Ontolingua* permet de construire des ontologies des trois façons suivantes : (1) emploi d'expressions *Kif*, (2) utilisation exclusive du vocabulaire de la *Frame Ontology* (impossibilité de représenter des axiomes), (3) recours aux deux langages simultanément, en fonction des préférences du constructeur d'ontologies. Quoi qu'il en soit, la définition d'*Ontolingua* se compose d'un titre, d'une définition informelle en langage naturel et d'une définition formelle écrite en *Kif* ou en utilisant le vocabulaire de l'ontologie des *Frames*. Une application GFP (Chaudhri *et al.* 1997) est nécessaire pour raisonner avec les ontologies *Ontolingua*.

CycL, qui n'est autre que le langage de représentation de connaissances de *Cyc*, est un langage déclaratif et expressif proche de la logique des prédicats et qui comprend des extensions qui permettent de traiter l'égalité, le raisonnement par défaut, l'application de la fonction de Skolem et quelques aspects de la logique du second ordre. *CycL* utilise une certaine forme de circonscription, inclut l'hypothèse du nom unique, et peut utiliser l'hypothèse du monde clos où cela s'avère nécessaire. Le moteur d'inférence de *Cyc* réalise des déductions logiques générales, fait appel à la stratégie du meilleur d'abord en ayant recours à un ensemble d'heuristiques exclusives, utilise des microthéories afin d'optimaliser les inférences dans des domaines restreints, et inclut plusieurs modules de déduction spécialisés afin de traiter des types d'inférences spécifiques.

Loom est un langage de programmation perfectionné basé sur la logique et l'environnement de premier ordre, qui appartient à la famille *KL-One*. Le langage *Loom* fournit un langage expressif et explicite de spécification de modèles déclaratifs ainsi qu'un support déductif puissant, qui inclut à la fois un raisonnement strict et un raisonnement par défaut, et une vérification automatique de la cohérence. Il offre en outre plusieurs paradigmes de programmation, qui constituent une sorte d'interface avec la spécification de modèle déclaratif, et des services à base de connaissances.

Flogic est une intégration de langages de frames et de logique des prédicats. Il inclut des objets (simples et complexes), le principe d'héritage, les types polymorphes, les méthodes de recherche et l'encapsulage. La logique des prédicats, d'une part, et l'héritage structurel et comportemental, d'autre part, forment la base de son système déductif.

5.2 Comment les outils logiciels soutiennent-ils le processus de conception et d'utilisation des ontologies ?

L'*Ontolingua Server* (Farqhuar *et al.* 1996), *Ontosaurus* (11) (Swartout *et al.* 1997), *Ode* (Blázquez *et al.* 1998 et Fernandez *et al.* 1999) ainsi que *Tadzebao* et *WebOnto* (Domingue 1998) constituent les outils principaux destinés à concevoir des ontologies.

L'*Ontolingua Server* est l'environnement le plus connu pour construire des ontologies dans le langage *Ontolingua*. Il s'agit d'un ensemble d'outils et de services qui assistent la conception d'ontologies communes à laquelle collaborent des groupes de travail opérant depuis des endroits différents. Il a été élaboré par le *Knowlege Systems Laboratory* dans le cadre du programme *Knowledge Sharing Effort* de l'*Arpa* à l'université de Stanford. L'architecture du serveur d'ontologie permet d'accéder à une bibliothèque d'ontologies, à des traducteurs de langages de programmation (*Prolog, Corba's IDL, Clips, Loom, Kif*) et à un éditeur qui permet de créer et de parcourir des ontologies. Trois types d'interaction sont possibles : il peut s'agir ainsi de collaborateurs qui souhaitent écrire et examiner des ontologies à distance, d'applications éloignées susceptibles de vouloir interroger et modifier des ontologies sur le serveur via l'Internet ou des applications locales. L'URL suivant permet d'accéder au serveur de l'ontologie : http://www-ksl-svc.stanford.edu:5915/.

L'*Ontosaurus*, créé à l'Information Sciences Institute de l'University of South California, s'articule en deux parties : un serveur d'ontologie qui utilise *Loom* comme système de représentation des connaissances et un serveur de navigation dans les ontologies qui crée dynamiquement des pages HTML (qui incluent de la documentation constituée tant d'images que de textes). Ce serveur présente la hiérarchie de l'ontologie et utilise les formulaires HTML pour permettre à l'utilisateur d'éditer l'ontologie. En outre, des traducteurs de *Loom* vers *Ontolingua, Kif, KRSS* et *C++* ont été élaborés.

Ode (Ontology Design Environment) est actuellement mis au point à la faculté d'informatique de l'Universidad Politecnica de Madrid. Au nombre des avantages que présente *Ode* figurent le module de conceptualisation destiné à construire des ontologies et le module servant à élaborer des modèles conceptuels *ad hoc*. Grâce au module de conceptualisation, l'ontologue peut

(1) http://www.indra.isi.edu:8000

développer l'ontologie au niveau de la connaissance en mettant en œuvre un jeu de représentations indépendantes du langage cible dans lequel l'ontologie sera implémentée. Une fois la conceptualisation terminée, des codes *Ode* sont créés automatiquement par des générateurs de code. Les générateurs les plus courants sont *Ontolingua*, *Flogic* et une base de données relationnelles. De cette façon, les personnes non-initiées aux langages d'implémentation des ontologies sont en mesure de spécifier et de valider des ontologies grâce à cet environnement. Le module pour construire des modèles conceptuels *ad hoc* inclut un langage appelé *Language for building intermediate representations (LBIR)*, qui donne la possibilité aux ontologues de spécifier le type de modèle adapté à leur ontologie.

Le *Knowledge Media Institute de l'Open University* a créé deux outils complémentaires, à savoir *Tadzebao* et *WebOnto*. *Tadzebao* permet aux cogniticiens de discuter en synchronisé ou en asynchronisé des ontologies; quant à *WebOnto*, il offre un support à la consultation, à la création et à l'édition conjointes d'ontologies.

6 Applications qui utilisent des ontologies

Si les ontologies peuvent être utilisées (Uschold *et al.* 1996) afin de communiquer entre des systèmes, des personnes et des organisations, d'interopérer entre des systèmes, de soutenir la conception et le développement de systèmes logiciels intelligents ou non, le nombre d'applications pour lesquelles on a recourt à des ontologies en vue d'en modéliser les connaissances est réduit. Cela signifie que bon nombre de ces ontologies ont été construites pour une application spécifique, sans intention de les réutiliser ou de les partager. Le fait que la réutilisation d'ontologies dans des applications soit une pratique encore peu répandue tient à différentes raisons(12) (Arpírez *et al.* 1998): les ontologies sont hébergées sur différents serveurs, la formalisation varie en fonction du serveur où se trouve l'ontologie, les ontologies placées sur un même serveur sont généralement décrites avec des degrés de précision variables et il n'existe aucun format commun pour présenter les informations sur les ontologies de façon à ce que les utilisateurs puissent choisir l'ontologie qui leur convient le mieux. Ces problèmes expliquent sans doute le nombre réduit d'applications dans les domaines de la gestion des connaissances, des moteurs de recherche s'appuyant sur des ontologies, de la génération de langues naturelles, de la modélisation d'entreprise, des systèmes intelligents et de l'interopérabilité entre les systèmes. Les applications qui utilisent des ontologies sont reprises dans leur plus grand nombre dans les actes de l'atelier sur les *Applications of ontologies and PSMs* organisé dans le cadre de l'*ECAI98* (voir http://delicias.dia.fi.upm.es/WORKSHOP/ECAI98/index.html).

Plusieurs applications utilisent des ontologies écrites en langage naturel. *Gum* est utilisé pour des applications spécialisées dans le traitement du langage naturel dans différentes langues: *Penman* (Bateman *et al.* 1990), un générateur de textes dans différents domaines; *Komet*, qui génère des textes en anglais, allemand et néerlandais; *TechDocb* (Rosner 1994), qui génère de textes techniques multilingues; *AlFresco* (Stock *et al.* 1993), un système de recherche documentaire spécialisé dans l'histoire de l'art italien; *Gist*, un système multilingue qui génère des textes administratifs en anglais, allemand et italien; *OntoGeneration* (Aguado *et al.* 1998), qui réutilise des ontologies spécialisées (produits chimiques); *Gum* et la technologie de génération de langage naturel (KPLM (Bateman *et al.* 1994)) pour la génération de textes espagnols dans le domaine des substances chimiques; et l'atelier d'utilisation de plusieurs ontologies dans une architecture *pipeline* de génération de langage naturel présenté dans Frohlich (*et al.* 1998). Hoenkamp (Hoenkamp 1998) utilise *WordNet* pour localiser les lacunes dans une ontologie qui représente les besoins en information des utilisateurs en analysant des documents consultés.

Dans le domaine de la modélisation d'entreprises, l'*Enterprise Toolset* constitue l'environnement le plus perfectionné qui inclut l'*Enterprise Ontology*. Il utilise une architecture constituée d'agents de façon à intégrer des outils prédéfinis selon un mode «plug-and-play». Les éléments constitutifs de l'*Enterprise Toolset* sont: un *Procedure Builder* servant à saisir les modèles de processus, un *Agent Toolkit* qui soutient le développement des agents, un *Task Manager* servant à intégrer, visualiser et assurer le déroulement du processus, ainsi qu'une *Enterprise Ontology* pour la communication (voir http://www.aiai.ed.ac.uk/project/enterprise pour plus d'informations). D'autres applications utilisent des ontologies *Tove*. L'*Enterprise Design Workbench* constitue un environnement d'étude qui permet à l'utilisateur d'explorer une série de structures d'entreprises. Il offre une étude comparée des différentes structures d'entreprises et oriente le concepteur. Dans le cadre de l'*Integrated Supply Chain Management Project*, un réseau d'agents intelligents coopératifs réalise une ou plusieurs fonctions de la

(12) http://delicias.dia.fi.upm.es/OntoAgent

chaîne d'approvisionnement, et chaque agent coordonne ses actions avec les autres agents. L'entreprise virtuelle *Tove* fournit le banc de test utilisé par les agents qui ont été conçus pour les fonctions de chaînes d'approvisionnement principales : logistique, transport, gestion, etc.

Les courtiers en WWW ont commencé récemment à utiliser les ontologies dans différents domaines. *Ontobroker*(13) (Fensel *et al.* 1998) est un service pour la gestion de connaissances qui est utilisé dans le contexte de la *Knowledge Annotation Initiative* de la communauté d'acquisition des connaissances. *(Onto)²Agent* (Arpírez *et al.* 1998) est un moteur basé sur une ontologie pour rechercher des ontologies sur la toile : il utilise la *Reference-Ontology* comme source de connaissances et il extrait des descriptions d'ontologies qui satisfont un nombre donné de contraintes. Il est disponible à l'URL suivant : http://delicias.dia.fi.upm.es/OntoAgent/). *Chemical OntoAgent* (Arpírez *et al.* 1998), un moteur de recherche sur la toile s'appuyant sur une ontologie et spécialisé dans l'enseignement de la chimie, permet aux étudiants d'apprendre la chimie et d'évaluer leurs connaissances dans ce domaine. *Chemicals* constitue sa source de connaissances.

Kactus (Schreiber *et al.* 1995) est un projet *Esprit* qui traite de la modélisation des connaissances sur les systèmes techniques complexes à emplois multiples et du rôle des ontologies dans cette modélisation. Des ontologies spécialisées dans les domaines de la conception et de l'évaluation des réseaux électriques, de l'extraction de pétrole en mer, et des navires ont été créées.

Plinius (van de Vet *et al.* 1995) est un système d'acquisition de connaissances semi-automatique spécialisé dans les textes en langage naturel traitant de la céramique, de ses propriétés et de sa production.

(13) http://www.aifb.uni-karlsruhe.de/WBS/broker

Conclusions

Dans cet article, nous avons examiné les développements récents survenus dans le domaine de l'ontologie. Dans la situation actuelle, l'on remarque une bonne compréhension globale de la nature et de la fonction des ontologies et il apparaît que les travaux réalisés se fondent désormais sur les acquis pour évoluer dans de nouvelles directions.

Dans le domaine de l'ontologie, on s'intéresse avant tout à l'intégration d'ontologies hétérogènes, à la caractérisation et à la consultation des ontologies sur la toile, à l'intégration d'ontologies et de méthodes de résolution de problèmes, ainsi qu'à l'utilisation d'ontologies dans le but d'analyser et de générer du langage naturel. Par ailleurs, des efforts sont fournis pour se rapprocher du monde orienté objet et des bases de données. Il est clair que les ontologies acquièrent une importance notoire dans un grand nombre de domaines.

Asunción Gómez-Pérez,
Faculté d'informatique,
Universidad Politécnica de Madrid,
Madrid,
Espagne.

Traduit de l'anglais par S. Descotte,
Centre de recherche Termisti,
Bruxelles.

Bibliographie

Aguado (G.), Bateman (J.), Bañón (A.), Bernardos (S.), Fernández (M.), Gómez-Pérez (A.), Nieto (E.), Olalla (A.), Plaza (R.), Sanchez (A.), 1998: «ONTOGENERATION: Reusing domain and linguistic ontologies for Spanish», dans *Workshop on Applications of Ontologies and PSMs*. Brighton, p. 1-10.

Arpirez (J.), Gómez-Pérez (A.), Lozano (A.), Pinto (S.), 1998: «(ONTO)2Agent: An ontology-based WWW broker to select ontologies», dans *Workshop on Applications of Ontologies and PSMs*, Brighton, p. 16-24.

Bateman (J.A.), Magnini (B.), Fabris (G.), 1995: «The Generalized Upper Model Knowledge Bare: Organization and Use», dans *Towards Very Large Knowledge Bases*, IOS Press, p. 60-72.

Bateman (J. A.), 1994: *KPML: The KOMET-Penman (Multilingual) Development Environment.* Technical Report, GMD/IPSI, Darmstadt.

Bateman (J. A.), Kasper (R. T.), Moore (J. D.) and Whitney (R. A.), 1990 : *A General Organization of Knowledge for Natural Language Processing : the Penman Upper Model*. Technical Report, USC/ISI, Marina del Rey, California.

Benjamins (R.), Fensel (D.). Decker, Gómez-Pérez (A.), 1999: «(KA)2 Building Ontologies for the Internet : A mid term report». À paraître dans l'International Journal of Human Computer Studies.

Bernaras (A.), Laresgoiti (I.) and Corera (J.), 1996: *Building and Reusing Ontologies for Electrical network Applications. Proceedings of the 12th ECAI*, p. 298-302.

Blázquez (M.), Fernández (M.), García-Pinar (J. M.), Gómez-Pérez (A.), 1998: *Building Ontologies at the Knowledge Level using the Ontology Design Envirnoment. Proceedings of the Eleventh Knowledge Acquisition Workshop*, KAW98, Banff.

Borgo, (S.); Guarino, (N.); Masolo, (C.), 1996: «Stratified Ontologies: the case of physical objects», dans *Proceedings of the Workshop on Ontological*

Engineering. Held in conjunction with ECAI96, Budapest, p. 5-15.

Borst (W. N.), 1997: *Construction of Engineering Ontologies*. University of Tweenty. Enschede, Centre for Telematica and Information Technology.

Bylander (T.), Chandrasekaran (B.), 1998: «Generic Tasks in Knowledge-based reasoning: The right level of abstraction for Knowledge Acquisition», dans B. Gaines and J. Boose Editors, *Knowledge Acquisition of Knowledge Based Systems*. Volume 1, Academic Press London, p. 65-77.

Chaudhri Vinay (K.), Farquhar (A.), Fikes (R.), Karp (P. D.), Rice (J, P.), 1997: *The Generic Frame Protocol 2.0*, Technical Report, Stanford.

Domingue (J.), 1998: «Tadzebao and Webonto: Discussing, Browsing and editing ontologies on the web», dans *Proceedings of the Eleventh Knowledge Acquisition Workshop*, KAW98, Banff.

Farquhar (A.), Fikes (R.), Rice (J.), 1996: «The Ontolingua Server: A Tool for Collaborative Ontology Construction», Proceedings of the 10th Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff, Alberta, Canada, p. 44.1-44.19, 1996.

Fensel (D.), Decker (S.), Erdman (M.) Studer (R.), 1998: «Ontobroker: The Very High Idea», dans *Proceedings of the 11th International Flairs Conference* (FLAIRS-98), Sanibal Island.

Fernández (M.); Gómez-Pérez (A.), Pazos (J.); Pazos (A.), 1999: *Building a Chemical Ontology using methontology and the ontology desing environment. IEEE Intelligent Systems and their applications*. 14 (1):37-45, 1999.

Fernández (M.), Gómez-Pérez (A.) Juristo (N.), 1997: METHONTOLOGY: *From Ontological Art Toward Ontological Engineering*. Spring Symposium Series on Ontological Engineering, AAAI97, Stanford, USA.

Frohlich (M.); van de Riet (P.), 1998: *Using Multiple ontologies in a framework for Natural language generation. Workshop on Applications of Ontologies and PSMs*. Brighton, England, p. 67-77.

Genesereth (M.), Fikes (R.), 1992: *Knowledge Interchange Format. Technical Report*, Computer Science Department, Stanford University, Logic-92-1.

Gómez-Pérez (A.); Rojas-Amaya (M.D.), Ontological Reengineering for Reuse. Knowledge Acquisition Modeling and Management. 11th European Workshop, EKAW'99. Dagstuhl Castle, Germany, May 26-29, 1999, p. 139-156.

Gómez-Pérez (A.), 1998: *Knowledge Sharing and ReuseThe Handbook of Applied Expert Systems*. Edited by J. Liebowitz, CRC Press

Gómez-Pérez (A.), 1996: *A Framework to Verify Knowledge Sharing Technology. Expert Systems with Applicatio*n. Vol. 11, N. 4., p. 519-529.

Gruber (T.) and Olsen (R.), 1994: *An Ontology for Engineering Mathematics. Technical Report KSL-94-18*, Knowledge Systems Laboratory, Stanford University, CA.

Gruber, (T.), 1993: *A translation Approach to portable ontology specification. Knowledge Acquisition*, 5: 199-220

Gruber (T.), 1995: *Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human Computer Studies*, 43:907-928

Gruninger (M.), Fox (M.), 1995: *Methodology for the Design and Evaluation of Ontologies, Proceedings of IJCAI95's Workshop on Basic Ontological Issues in Knowledge Sharing*.

Guarino (N.), 1997: «Some Organizing Principles for a unified top-level ontology», dans *Spring Symposium Series on Ontological Engineering*. Stanford, p. 57-63.

Guarino (N.), Giaretta (P.), 1995: «Ontologies and Knowledge Bases: Towards a Terminological Clarification», dans Mars (N.J.I.), *Towards Very Large Knowledge Bases: Knowledge Building & Knowledge Sharing*. IOS Press, p. 25-32.

van Heist (G.), Schreiber (A. Th.), Wielinga (B. J.), 1997: *Using explicit ontologies in KBS development, International Journal of Human-Computer Studies*, 45, p. 183-292.

Hoenkamp (E.), 1998: «Spotting Ontological Lacunae through spectrum analysis of retrieved documents», *dans Workshop on Applications of Ontologies and PSMs*, Brighton, England, p. 73-77.

Kifer (M.), Lausen (G.), Wu (J.), 1995: *Logical Foundations of Object-Oriented and Frame-Based Languages, Journal of the ACM*.

Klinder (M.); Lausen (G.); Wu (J.), 1995: *Logical Foundations of Object oriented and frame-based languages. Journal of ACM*.

Lenat (D.B.), Guha (R V.), 1990: *Building Large Knowledge-based systems. Representation and Inference in the Cyc project*. Addison-Wesley, Reading, Massachusetts.

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

Miller (G. A.), 1990: *WordNet: An Online Lexical Database, International Journal of Lexicography* 3, 4: 235- 312

Mizoguchi (R.); Vanwelkenhuysen, (J.); Ikeda (M.), 1995: «Task Ontology for reuse of problem solving knowledge», dans Mars (N.J.I), *Towards Very Large Knowledge Bases: Knowledge Building & Knowledge Sharing*, IOS Press, p. 46-57.

Neches (R.), Fikes (R.E.), Finin (T.), Gruber (T.R.), Senator (T.), and Swartout (W.R.), 1991: *Enabling technology for knowledge sharing. AI Magazine*, 12(3):36-56.

Rösner (D.), 1994: *Generating Multilingual Documents from a Knowledge Base: The TECHDOC Project. Technical Report FAW Ulm*, Ulm (Germany).

Sowa (J. F.), 1997*: Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Boston, MA, PWS Publishing Company. À paraître.

Stock (O.), Carenini (G.), Cecconi (F.), Franconi (E.), Lavelli (A.), Magnini (B.), Pianesi (F.), Ponzi (M.), Samek-Lodovici (V.) and Strapparava (C.), 1993: «ALFRESCO: Enjoying the Combination of Natural Language Processing and Hypermedia for Information Exploration», dans Mark T. Maybury, editor, *Intelligent Multimedia Interfaces*, The MIT Press, p. 197-224,

chapter 9: *Extended and revised version of a paper previously published at IJCAI-91.*

Studer (R.), Benjamins (V.R.), Fensel (D.), 1998: *Knowledge Engineering: Principles and Methods. Data & Knowledge Engineering*. 25: 161-197.

Swartout (B.), Patil (R.), Knight (K.) and Russ (T.), 1997: *Towards Distributed Use of Large-Scale Ontologies. Spring Symposium Series on Ontological Engineering*, Stanford University, CA, p. 138-148.

Uschold (M.), Grüninger (M.), 1996: *ONTOLOGIES: Principles, Methods and 16 applications, Knowledge Engineering Review*, Vol. 11, N. 2

Van der Vet (P.E.), Speel (P.-H.), Mars (N. J. I.), 1994: *The Plinius ontology of ceramic materials. Proceedings of ECAI94's Workshop on Comparison of Implemented Ontologies*, Amsterdam.

Comment accéder aux éléments définitoires dans les textes spécialisés ?

Dans cet article, je montre l'intérêt d'utiliser des corpus de textes spécialisés dans un travail terminographique (identification de termes et constructions de définitions terminologiques). J'essaie d'identifier les types de textes spécialisés qui conviennent à ce travail et je propose un certain nombre de critères pour sélectionner des textes appropriés. Je pars du principe que dans beaucoup de cas de communications spécialisées, les auteurs vont, de façon explicite ou implicite, expliquer certains des termes qu'ils utilisent. En utilisant trois corpus différents, je proposerai une méthodologie pour accéder à ces éléments définitoires.

Termes-clés :
corpus ; extraction ; définition ; communication spécialisée.

1 Introduction

La terminographie exige des investissements importants en temps, en argent et en énergie. La disponibilité de textes électroniques devrait nous permettre de construire des corpus de textes importants qui se prêteraient à une exploitation terminographique semi-automatique. J'ai l'intention de parler de la conception et de la construction de corpus qui sont adaptés à ces activités et de proposer une méthodologie pour extraire des informations terminologiques.

Le travail sur corpus n'est pas une activité récente. Depuis déjà une trentaine d'années, des chercheurs construisent des corpus de plus en plus grands et les exploitent dans des buts différents, pour construire des dictionnaires, pour étudier la grammaire d'une langue, pour observer l'évolution d'une langue. Ce qui caractérise les grands corpus les plus connus (en langue anglaise tout au moins), c'est qu'ils contiennent des textes très variés. Ces corpus ont été construits pour être représentatifs de la langue telle qu'elle est utilisée tous les jours, l'intention étant de créer une ressource qui reflète les éléments communs d'une langue. Si ces corpus contiennent des textes spécialisés, ceux-ci ne représentent qu'une toute petite partie de l'ensemble du corpus. Par conséquent, ces corpus sont moins intéressants pour ceux qui cherchent à étudier la langue de domaines spécialisés. Le travail sur des corpus spécialisés est beaucoup plus récent et pendant très longtemps, les quelques chercheurs qui voulaient exploiter des corpus de textes spécialisés ont dû créer leur propre corpus. Pour différentes raisons, ces corpus ont été difficiles à construire : il y avait un manque de textes électroniques, les permissions étaient difficiles à obtenir et, de la part des chercheurs, il y avait un manque d'enthousiasme pour partager ou prêter des ressources existantes. Par conséquent, pendant très longtemps, le nombre et la taille des corpus de textes spécialisés sont restés relativement faibles. La situation a évolué de façon notable ces dernières années. Les éditeurs sont plus prêts à donner leur permission, le volume de textes disponibles sous forme électronique ne cesse de croître. De nos jours, celui qui veut construire un corpus de textes spécialisés a l'embarras du choix, ce qui amène d'autres problèmes sur lesquels je reviendrai plus tard.

J'ai l'intention de me poser plusieurs questions qui me semblent importantes pour ceux qui s'intéressent à un travail sur corpus de textes spécialisés. Je me demanderai d'abord ce qu'est un corpus pour établir ce qui distingue un corpus d'autres collections de textes, telles que des archives par exemple. Je vais aussi essayer de caractériser ce qui distingue une approche à base de corpus d'une approche plus traditionnelle. Ensuite, j'essayerai d'établir ce que nous voulons dire quand nous utilisons le terme *texte spécialisé*. Ce terme est utilisé par des communautés différentes pour décrire un grand nombre de types de textes qui ne se prêtent pas tous à un travail

terminographique. Je tâcherai donc, dans la deuxième partie de mon article, et de définir le terme *texte spécialisé*, et de caractériser le genre de texte spécialisé qui conviendrait à une exploitation terminographique précise, c'est-à-dire l'extraction d'éléments définitoires qui serviront de base pour la formulation de définitions terminologiques. Quand nous aurons caractérisé le genre de texte qui nous intéresse, il nous restera à décider quels critères il faut appliquer pour sélectionner les textes les plus appropriés à notre activité. Dans la troisième partie, je proposerai donc un ensemble de critères que j'utilise dans mon propre travail terminographique.

La plus grande partie de mon article consistera à décrire la méthodologie que j'ai adoptée pour réaliser mon travail. J'ai conçu deux approches différentes, l'une qui consiste à exploiter la présence de certaines structures grammaticales et l'autre qui consiste à utiliser les termes eux-mêmes comme point de départ. Dans cet article, l'accent sera sur la première de ces deux approches. Il est important de noter que ce que je propose n'est pas une solution parfaite; elle ne mène pas à une solution automatique. L'expert aura seulement les éléments avec lesquels il pourra travailler ensuite.

2 Qu'est-ce qu'un corpus?

D'après Sinclair (1994: 2), un corpus est une «*collection of pieces of language that are selected and ordered according to explicit linguistic criteria to be used as a sample of the language*». D'après Atkins, Clear et Ostler (1992:1) un corpus est «*a subset of an ETL (electronic text library) built according to explicit design criteria for a specific purpose, e.g. the Corpus Révolutionnaire (Bibliothèque Beaubourg, Paris), the Cobuild Corpus, the Longman/Lancaster corpus, the Oxford Pilot Corpus*». Puisque ces définitions sont très proches des définitions proposées par d'autres linguistes, nous pouvons les considérer comme étant représentatives. Ces définitions sont révélatrices et nous permettent de voir en quoi un corpus est différent d'autres collections de textes. Tout d'abord, les textes que l'on trouve dans un corpus ne s'y trouvent pas par hasard. Il ne s'agit pas de prendre n'importe quel texte tout simplement parce qu'il est disponible. Chaque texte est sélectionné selon des critères linguistiques précis. Deuxièmement, si l'on crée un corpus, on le fait dans un but précis; on sait pourquoi on le crée, on sait d'avance à quoi il va servir. Troisièmement, les textes que l'on trouve dans un corpus sont des textes authentiques et naturels, c'est-à-dire qu'ils n'ont pas été édités. Bien qu'un corpus contienne des textes qui ont été rédigés pour être lus, il ne sera jamais utilisé de cette façon-là. S'il nous arrive d'en lire des bouts, ces bouts viendront sans doute de textes différents et même quand ils viennent du même texte, ils ne se seront pas suivis dans le texte d'origine. On utilise des outils informatiques (tels que *WordSmith Tools*) pour accéder au corpus; ces outils nous permettent de voir des mots dans leur contexte même si, d'une certaine façon, on les voit aussi hors de leur contexte. Le fait de lire plusieurs occurrences à la fois nous permet d'identifier des tendances que l'on n'aurait peut-être pas observées avec une approche plus conventionnelle. En résumé, un corpus est un produit artificiel qui contient des textes authentiques qui ont été sélectionnés selon des critères précis pour être utilisés comme un échantillon de la langue. En ce qui concerne notre travail, nos corpus contiendront des textes spécialisés, ce qui nous mène à la question suivante, à savoir, qu'est-ce qui caractérise les textes spécialisés?

3 Caractérisation du texte spécialisé

Quand nous utilisons le terme *texte spécialisé*, nous le faisons habituellement pour différencier les textes dits *de langue générale* et les textes dits *de langue de spécialité*. On peut utiliser des critères différents pour caractériser les textes spécialisés et les critères que l'on utilisera dépendront en grande partie de la raison pour laquelle on étudie ces textes. Certains chercheurs s'intéressent aux aspects lexicaux, d'autres aux aspects grammaticaux et d'autres encore aux aspects de typologie du texte. Ceux qui s'intéressent aux aspects lexicaux, tels que les terminologues, diront qu'un texte spécialisé se caractérise par son vocabulaire spécialisé, c'est-à-dire par la fréquence de termes techniques utilisés dans le texte. Ceux qui s'intéressent aux aspects grammaticaux, tels les chercheurs qui se consacrent aux sous-langages, diront que les textes spécialisés se caractérisent non seulement par le vocabulaire utilisé mais aussi et peut-être plus par leurs structures grammaticales. C'est-à-dire que souvent, la grammaire de textes spécialisés est différente de la grammaire pour la langue générale, différente parce que limitée en ce qui concerne les structures utilisées, différente parce que les structures grammaticales qu'on y trouve n'obéissent pas aux règles de grammaire normales (utilisation d'ellipses, construction de termes composés longs et compliqués). Ceux qui s'intéressent à la typologie des textes s'occupent principalement de la fonction et de l'organisation des textes. Si, par exemple, ils étudient la langue de la médecine, ils vont plutôt essayer de cerner les différences entre les différents *types* de documents, que ce soit des articles de recherche, des notices d'emploi pour des

médicaments, des rapports de laboratoire.

Quand on veut construire un corpus de textes spécialisés qui va servir à un travail terminographique, il se trouve qu'aucune des façons de caractériser les textes spécialisés citées ci-dessus ne convient à nos besoins. Contrairement à ce que l'on pourrait penser, tout texte spécialisé ne convient pas à une exploitation terminographique. Rappelons que notre but principal est de sélectionner des textes qui contiennent des éléments définitoires, c'est-à-dire des textes où certains des termes utilisés sont définis ou de façon implicite, ou de façon explicite. Si nous utilisons uniquement le critère de la fréquence des termes pour sélectionner nos textes, nous nous trouverons avec beaucoup de textes qui sont peu riches en éléments définitoires. Il en est le même du critère des structures grammaticales inhabituelles. Il semblerait que le critère de typologie puisse convenir le mieux, mais ce critère pose aussi des problèmes. Il me semble que toute discussion de ce que c'est qu'un texte spécialisé doit tenir compte d'un élément très important qui est extra-textuel, c'est-à-dire le rapport entre l'auteur et le lecteur.

Cet aspect est primordial quand nous voulons décider si un texte doit faire partie de notre corpus. Nous cherchons à construire un corpus qui nous permette non seulement de repérer des termes mais aussi de repérer des éléments définitoires. Si nous pouvons nous attendre à trouver un nombre considérable de termes techniques dans tout texte spécialisé, il n'en est pas le même des éléments définitoires. Alors qu'il y a beaucoup de situations où des auteurs expliqueront de façon explicite certains des termes qu'ils utilisent, il y en a autant, sinon plus, où ils ne le font pas. Par exemple, on ne s'attend pas à en trouver dans des bulletins de météo, ni dans des recettes de cuisine, ni dans des notices d'emploi de médicaments par exemple. Par contre, on s'attend à en trouver dans des livres de formation, dans des normes industrielles, et même quelquefois dans des articles de recherche s'il s'agit de créer de nouvelles notions ou de redéfinir des notions existantes.

Les rapports entre l'auteur et le lecteur détermineront combien d'explications seront fournies dans un texte donné et ceci est valable pour toute sorte de texte. Je crois qu'il y a trois types de rapports auteur-lecteur qui sont susceptibles de nous intéresser. Le premier concerne la communication entre experts. Les textes rédigés par des experts pour des experts vont avoir une très haute fréquence de termes techniques. Considérons par exemple des articles de recherche publiés dans des revues spécialisées. Dans ces textes, on trouvera une très haute densité de termes mais probablement très peu d'éléments définitoires. L'explication est simple : le lecteur est censé connaître et comprendre les termes utilisés.

Le deuxième type de rapports concerne la communication entre des experts et des gens qui ont déjà une certaine compétence dans le domaine en question. Je pense ici à des gens qui travaillent dans le même domaine mais qui n'ont pas le même niveau de formation. Cela pourrait être des communications entre ingénieurs et techniciens, entre médecins spécialisés et médecins généralistes. Ce qui distingue cette situation de la précédente est le niveau d'expertise. Les auteurs vont définir ou expliquer certains des termes qu'ils utilisent quand ils estiment que ces termes ne sont pas connus par leurs lecteurs.

Le troisième type de rapports concerne la communication entre les experts et les gens qui n'ont aucune formation dans le domaine en question mais qui ont besoin de connaître et de comprendre la terminologie du domaine. Dans ces communications, la densité de termes va être nettement moins élevée que dans les deux catégories précédentes mais on peut s'attendre à trouver une forte densité d'éléments définitoires. Je pense par exemple à des livres d'école ou à des livres de formation professionnelle. Ce qui est intéressant ici est que ce genre de texte est rarement considéré comme étant intéressant dans les discussions de textes spécialisés, tout simplement parce que la fréquence de termes n'est pas suffisamment élevée. Comme nous le verrons plus tard, ce sont précisément ces textes-là qui sont les plus riches en éléments définitoires et les communications entre experts les plus pauvres. Ce qui est important à retenir dans le cas de chacun de ces rapports auteurs-lecteurs est que le cadre de communication dans lequel on trouvera des textes qui correspondent à ces rapports est un cadre professionnel ou éducatif.

Il y a un quatrième type de rapports auteur-lecteur que l'on aurait pu envisager mais qui s'est avéré ne pas convenir à nos besoins. C'est le cas des rapports entre des experts et des lecteurs occasionnels ; là, il s'agit de communications spécialisées qui paraissent dans la littérature dite *d'intérêt populaire*, dans des revues comme *Scientific American* aux États-Unis ou *New Scientist* en Angleterre, par exemple. Le cadre n'est ni professionnel ni éducatif dans le sens où les lecteurs qui lisent ces textes ne le font pas dans le cadre de leur profession ou de leur formation.

4 Critères de sélection

Maintenant que nous avons décidé que nous voulons sélectionner des textes qui correspondent à un des cadres de communication énoncés précédemment, il nous reste à établir un jeu de critères pour sélectionner les textes que nous voulons inclure dans nos corpus. Il ne suffit pas de

dire que tel ou tel texte correspond à un de nos cadres de communications. Il y a d'autres critères qui entrent en jeu. Les critères les plus importants sont les suivants:

1. Les textes doivent avoir été publiés. Cela veut dire, d'après la définition de Biber, que «they are printed in multiple copies for distribution, they are copyright registered or recorded by a major indexing service» (1993: 245). Nous envisageons donc d'analyser des textes qui ont été rédigés pour être lus.

2. Les textes doivent être des textes entiers. Traditionnellement, les corpus contiennent des échantillons d'une longueur précise de textes différents.

3. L'auteur peut être un individu, un groupe d'individus, une association ou une fédération. Dans tous les cas, l'auteur doit être compétent pour écrire dans le domaine en question. Sa formation professionnelle doit correspondre au sujet dont il traite.

4. Le cadre doit être professionnel ou éducatif. Les textes sont destinés à être utilisés dans un cadre professionnel ou dans un cadre d'enseignement.

5. La date de publication d'un texte peut être importante, surtout dans des domaines en pleine évolution, tels que l'informatique.

6. La fonction du texte doit être informative, didactique ou normative.

5 Les corpus

Pour élaborer ma méthodologie, j'ai choisi d'utiliser trois corpus différents. Le premier, le corpus *Nature*, est une collection d'articles publiés dans la revue *Nature*. *Nature* est une revue spécialisée anglaise qui contient des articles rédigés par des experts pour des experts. Le cadre de communication correspond donc à notre première catégorie, la communication entre experts. J'ai choisi d'étudier ce corpus parce que mon intuition me suggérait que bien que ces textes sont on ne peut plus spécialisés, il y avait de fortes chances pour qu'il y ait très peu d'éléments définitoires, ce qui s'est avéré être le cas. Le corpus *Nature* contient environ 230 000 mots.

Le deuxième corpus, le corpus UIT, rédigé en anglais, français et espagnol, a été publié par l'Union internationale des télécommunications et est disponible sur CD-Rom. Le cadre de communication correspond à notre deuxième catégorie; il s'agit de communication entre des experts et des gens qui travaillent dans le même domaine mais qui n'ont pas nécessairement le même niveau de compétence que les auteurs. Le corpus UIT comprend des textes normatifs et contient environ 4,7 millions de mots.

Le troisième corpus contient des livres d'enseignement pour des matières sur le programme d'études GCSE en Angleterre. Les élèves passent l'examen GCSE deux ans avant de passer l'équivalent du baccalauréat français. Le cadre de communication correspond à notre troisième catégorie, la communication entre des experts et des gens qui n'ont aucune ou très peu de connaissances dans le domaine en question, mais qui ont besoin de connaître et comprendre les notions du domaine. Le corpus comprend des textes didactiques et contient environ un million de mots.

6 Des éléments définitoires

Maintenant que nous avons défini ce que nous voulons dire par *texte spécialisé* et, plus particulièrement, ce que nous entendons *par un corpus spécialisé à but terminographique*, je vais diriger mon attention vers la notion d'éléments définitoires. Je pars du principe que dans les cadres de communication mentionnés précédemment, les textes vont contenir des éléments définitoires. Les auteurs vont nous fournir des explications de certains des termes qu'ils utilisent.

Je crois qu'il y a plusieurs moyens d'accéder à ces éléments définitoires: par l'intermédiaire du terme, par les structures grammaticales, etc. Je propose d'abord d'élaborer une description des éléments définitoires du point de vue de leur structure grammaticale. Je vais partir de structures très formelles pour arriver ensuite à des structures moins formelles et donc plus difficiles à identifier. Je propose d'utiliser les descriptions élaborées par Trimble en 1985 et développées ensuite par Flowerdew en 1992, deux chercheurs qui ont analysé des textes réels. D'après ces chercheurs, il y a trois catégories de définition: la définition formelle, la définition semi-formelle et la définition non formelle. D'après Trimble, la définition formelle est «*of course, the well-known equation-like* "Species = Genus + Differentia", *usually called "formal" because of its rigidity of form*» (Trimble 1985: 75-76). La définition formelle a une structure rigide qui est généralement représentée par la formule suivante: *X = Y + caractéristique.* Dans cette formule, *x* représente le terme, *y* représente un hyperonyme et la caractéristique sert à distinguer *x* de tous les autres termes de la même catégorie. Voici quelques exemples que l'on trouve dans nos corpus:

• *A control circuit is a telephone-type circuit between the point of origin of the programme and the point where it terminates.* (UIT)

• *An ISDN connection is a connection established between ISDN reference points* (UIT)

• *A spore is a single cell which by itself can grow into a new organism* (GCSE)

En ce qui concerne la définition semi-formelle, Trimble dit «*By definition, a semi-formal definition contains only two of the three basic defining elements: the term being defined and the statement of differences* » (Trimble 1985: 77). La définition semi-formelle est généralement représentée par la formule: *x = caractéristique*, comme dans les exemples suivants:

• *Geometric elements are used to construct drawings of various types by a succession of overlay of points, straight lines, arcs etc.* (UIT)
• *Photographic elements are used to render an image by the transmission and display of an array of picture elements (pixels)* (UIT)
• *Arachnids have four legs and their bodies are made of two parts* (GCSE)
• *Amylase breaks up starch into sugar* (GCSE).

La troisième définition est une définition non formelle, décrite par Trimble comme suit «*The function of a non-formal definition is to define in a general sense so that a reader can see the familiar element in whatever the new term may be… Most non-formal definitions are found in the form of synonyms* » (Trimble 1985: 78). Elle fournit au lecteur le nom correct du terme et une autre unité lexicale qui a plus ou moins le même sens que le terme mais qui n'est pas obligatoirement un terme. Flowerdew (1992: 211) appelle cette définition *une définition par substitution* et suggère que la définition par substitution peut être exprimée de trois façons différentes: par un synonyme, par paraphrase ou par dérivation. Voici quelques exemples que l'on trouve dans nos corpus:

• *a signal of limited duration, known as a «measuring signal»* (UIT)
• *the aggregate of time during which the speech in question is present (called the active time)* (UIT)
• *A bulge called a pseudopodium…* (GCSE)
• *Peas have small shoots called tendrils* (GCSE)
• *…an extended region of homology called the POU domain* (*Nature*).

Ce qui est intéressant ici est que l'on trouve, outre des paraphrases, non seulement des rapports de quasi-synonymie mais aussi des rapports hyperonyme-hyponyme.

6.1 Les définitions formelles

6.1.1 Les définitions formelles simples

En étudiant les structures grammaticales utilisées dans les corpus, j'ai pu constater qu'il y en a certaines qui correspondent généralement aux différents types de définitions énoncés précédemment. J'ai donc conçu un ensemble de conditions qu'une structure grammaticale doit remplir pour être considérée comme un élément définitoire. Regardons d'abord les conditions que doivent remplir les définitions formelles simples, c'est-à-dire les définitions qui correspondent à la structure *x = y + caractéristique* et qui apparaissent à l'intérieur d'une seule phrase.

1. *X* doit être un terme. Pour être considéré comme terme, il doit avoir une structure qui correspond aux structures terminologiques identifiées pour le corpus en question. Si *x* apparaît au début de la phrase, il doit être précédé ou de l'article indéfini ou ne pas être précédé d'article du tout. Ceci est très important parce que, en anglais, si le terme est précédé de l'article défini (quand il est en début de phrase), il n'a pas nécessairement un statut générique. Si le terme est précédé de l'article défini ou d'un adjectif démonstratif, cela peut vouloir dire deux choses: ou que la phrase est une continuation d'une définition déjà élaborée en partie dans la phrase précédente, ce qui est le cas pour les définitions complexes qui ne nous concernent pas ici, ou que la phrase concerne un cas précis qui ne peut pas être considéré comme étant valable pour le terme en général. Par contre, si *x* apparaît à la fin de la phrase, il peut être précédé soit de l'article défini, soit de l'article indéfini ou il peut apparaître sans aucun article.

2. *Y* doit être un terme ou doit être un hyperonyme générique tel que *technic, method, process, function,* etc. Le tableau 1 indique les hyperonymes génériques que l'on trouve dans les trois corpus.

Hyperonyme générique	UIT	GCSE	*Nature*
technic	341	7	9
method	2,374	149	32
process	1,143	201	63
function	3,105	54	133
property	39	71	9
system	7,396	712	135
class	1,598	416	81
device	746	9	4

Tableau 1:
liste d'hyperonymes génériques

3. Il y a un certain nombre de verbes ou de syntagmes verbaux qui peuvent lier *x* et *y*. On les appelle *hinges* ou *connective verbs*. Ils servent de lien entre les deux éléments de la phrase. Quelques exemples de liens: *is/are, is/are called is/are known as, is/are defined as, denote(s)*. Le verbe ou la phrase verbale qui sert de lien entre *x* et *y* doit être au présent de l'indicatif et ne peut pas être modifié par un verbe modal. Le verbe ne doit être modifié par aucune particule négative, telle que *not, never.* La phrase définitoire ne doit pas être modifiée par un «adverbe de focus» (*focussing adverb* en anglais). Quelques exemples de ces adverbes: *commonly, frequently, usually, specifically, generally, mainly,*

primarily. Le tableau 2 indique les adverbes de focus que l'on trouve dans les trois corpus.

Adverbe	UIT	GCSE	*Nature*
Chiefly	5	9	1
commonly	91	19	4
especially	176	142	21
exceptionally	88	1	4
exclusively	61	3	6
frequently	93	59	10
generally	496	140	35
mainly	98	125	20
mostly	11	85	18
occasionally	18	13	1
often	213	574	37
only	4,504	1,457	385
on the whole	2	13	1
predominantly	12	5	6
primarily	118	4	19
principally	9	0	2
purely	61	10	3
rarely	22	31	1
solely	93	5	5
sometimes	111	252	11
specifically	175	2	27
usually	354	408	17

Tableau 2 :
Adverbes de focus dans les trois corpus

Bien que l'on puisse penser que ces adverbes sont souvent utilisés pour accentuer la validité d'une phrase définitoire, ils ont souvent l'effet opposé. Quand un auteur dit qu'un terme est *souvent* défini de telle ou telle façon, on ne peut pas conclure que c'est *toujours* le cas.

4. Il ne suffit pas de préciser quels mots peuvent se trouver à la place de *x*, *y* et =. Il faut aussi regarder et la forme de la phrase, et la situation de l'élément définitoire dans la phrase en général. Tout d'abord, la phrase définitoire doit constituer la proposition principale de la phrase et ne doit pas être précédée de phrases subordonnées. Si la phrase définitoire est précédée d'une phrase subordonnée, celle-ci peut atténuer la force et la validité de la phrase définitoire. Par contre, la phrase définitoire peut être suivie d'une autre phrase à condition que ces deux phrases soient liées par la conjonction *and*.

5. *Y* doit être suivi directement par la caractéristique qui distingue *x* de tous les autres membres de la même catégorie. La caractéristique peut être introduite par une proposition, un participe passé ou un pronom relatif. *Y* ne doit pas être immédiatement suivi de la conjonction *and* parce que cela signale généralement la fin de l'élément définitoire.

Voici quelques exemples de définitions formelles simples :

• *A robot is a machine that tries to copy one or more human functions (GCSE).*
• *A videotex service centre is a computer used by the videotex service provider to authorize access to a videotex service (ITU).*
• *INTERLEUKIN-1 (IL-1) is a cytokine produced by mononuclear phagocytes (Nature*).

6.1.2 Les définitions formelles complexes

Bien qu'il y ait un nombre considérable de phrases définitoires qui correspondent à une définition formelle simple, c'est-à-dire une définition formelle exprimée dans une seule phrase, il y en a beaucoup d'autres qui sont exprimées dans deux phrases. Ce sont ce que Trimble appelle *des définitions complexes*. En général, une définition formelle complexe est exprimée d'une des deux façons suivantes : a) le terme est introduit à la fin d'une phrase et expliqué au début de la phrase suivante, ou b) le terme est nommé au début d'une phrase et expliqué dans la phrase précédente. Voici quelques exemples, d'abord du premier cas, et ensuite du deuxième :

Premier cas (le terme apparaît à la fin de la phrase précédente) :

• *…mean holding time. This is the total holding time divided by the total number of seizures and can be calculated on a circuit group basis or for switching equipment. (UIT)*
• *…digital speech interpolation (DSI). This is a technique whereby advantage can be taken of the inactive periods during a conversation, creating extra channel capacity. (UIT)*
• *…a ring mains system. This is a loop of cable that runs from the consumer unit round the house and back to the unit. (GCSE)*
• *…a device called a proboscis. This is a hollow tube which is normally tightly coiled beneath the head of the butterfly when it is not feeding. (GCSE)*
• *an exoskeleton. This is a hard outer protective covering made of chitin.*

Deuxième cas (définition dans une première phrase, terme nommé dans la phrase suivante) :

• *There are millions of compounds containing just hydrogen and carbon. These are called hydrocarbons.*
• *Some bacteria get their food from the dead bodies of plants and animals. These are called saprophytic bacteria.*

Ces définitions n'ont pas la même structure formelle que les définitions formelles simples mais tous les éléments des définitions formelles simples sont présents. Ces définitions sont relativement faciles à repérer. Dans le cas a) cité ci-dessus, il suffit de préciser que les mots que l'on trouve à la fin de la première phrase doivent correspondre à une structure terminologique déjà précisée et que la phrase définitoire doit commencer avec l'expression *This is…* ou *These are…* La phrase définitoire doit remplir les mêmes conditions précisées pour les définitions formelles simples à l'exception de *x* qui est remplacé par un pronom (*this* ou *these*). Dans le cas b) cité ci-dessus, il faut préciser que les mots qui suivent l'expression *These are* ou *This is* doivent être un terme. Les définitions complexes sont plus difficiles à identifier que les définitions simples, mais, dans le

corpus UIT en particulier, elles sont beaucoup plus fréquentes.

6.2 Les définitions semi-formelles

Plus fréquentes encore sont les définitions semi-formelles. Ce sont là des structures qui correspondent à la formule : *x = caractéristique.* Dans une définition semi-formelle, l'hyperonyme est absent de la phrase. En réalité, quand on analyse les définitions semi-formelles que l'on trouve dans le corpus, on découvre que l'hyperonyme est souvent précisé dans la phrase précédente et que la définition semi-formelle sert à fournir des informations supplémentaires. Les conditions pour les définitions semi-formelles sont un peu différentes des conditions précisées pour les conditions formelles. *X* doit être un terme qui peut être précédé de l'article défini ou de l'article indéfini. Le verbe qui lie *x* avec la caractéristique doit être au présent de l'indicatif ; la phrase définitoire semi-formelle doit constituer la proposition principale de la phrase. Voici quelques exemples de nos corpus :

• *Expanded polystyrene is made by blowing a gas (such as carbon dioxide) into the liquid polymer. (GCSE)*
• *One of the most important applications of neutralization is in making fertilizers like ammonium sulphate (((NH4)2SO4) and ammonium nitrate (NH4NO3). Ammonium sulphate is manufactured by neutralizing sulphuric acid (H2SO4) with ammonia solution (NH3(aq)). (GCSE)*
• *Whenever a change in the status of a signalling link, route or point occurs, the three different signalling network management functions (i.e., signalling traffic management, link management and route management) are activated. The signalling traffic management function is used to divert signalling traffic from a link or route to one or more different links or routes, to restart a signalling point, or to temporarily slow down signalling traffic in the case of congestion at a signalling point…* (UIT).

6.3 Les définitions non formelles

Les définitions non formelles sont peut-être les plus intéressantes à étudier. Rappelons que d'après Flowerdew, ces définitions, qu'il appelle *définition par substitution*, peuvent être exprimées de trois façons différentes : par un synonyme, par paraphrase ou par dérivation. En cas de substitution, l'auteur peut fournir un synonyme pour un terme (ce cas est assez rare) ou il va introduire un terme en utilisant un mot équivalent de langue générale qui est déjà connu du lecteur. En cas de paraphrase, l'auteur peut expliquer un terme en utilisant une paraphrase. Dans ce cas, il fournira les mêmes éléments que l'on trouve dans les définitions formelles et semi-formelles mais emploiera d'autres méthodes pour le faire.

Il y a un certain nombre d'indicateurs qui nous signalent la présence de définitions non formelles. Ils comprennent entre autres l'utilisation de parenthèses, l'utilisation d'expressions telles que *i.e., e.g., called, known as.* En plus des fonctions élaborées ci-dessus, ces indicateurs peuvent aussi être utilisés pour indiquer des rapports hyperonyme-hyponyme. On trouvera ci-dessous quelques exemples de chacun des indicateurs qui révèlent combien les indicateurs sont multifonctionnels. On ne peut pas dire d'avance que tel indicateur va fournir telle information.

• *…but all microphones contain a disc called a diaphragm. (GCSE)*
• *…threads of pure cellulose known as rayon. (GCSE)*
• *Modulation with two-phase conditions, called bi-phase modulation (2-PSK)…* (UIT)
• *…a signal of limited duration known as a "measuring signal"…* (UIT)
• *Row crops (e.g. strawberries) can be protected from…* (GCSE)
• *cell types, e.g. root-hair cell, egg cell (ovum) sperm cell, muscle cell…* (GCSE)
• *some are steady state impairments (e.g. loss, noise, quantization, distortion, phase jitter…) (UIT)*
• *But how can liquids change to solids (i.e. melt) or liquids to gases (i.e. boil or evaporate)? (GCSE)*
• *an idle signalling terminal, i.e. a signalling terminal not connected to a signalling data link (UIT).*

7 Conclusion

La méthodologie proposée ici est une approche qui cherche à exploiter l'utilisation de certaines structures dans des circonstances précises. L'identification de ces structures nous permet de repérer des éléments définitoires qui peuvent être considérés dans la formulation de définitions terminologiques. On peut aussi aborder le problème d'une autre façon, en prenant les termes eux-mêmes comme point de départ. Dans ce cas, on produit des concordances pour chaque terme pour voir si on peut cerner des éléments définitoires dans les lignes de concordance. Cette approche est très proche de celle qui est utilisée par l'équipe lexicographique *Cobuild.* Bien que cette approche soit aussi très productive, elle demande beaucoup plus d'analyse manuelle. Je suis persuadée qu'en associant une approche grammaticale telle que celle décrite dans cet article avec une approche qui prend les termes comme point de départ pour l'analyse de textes spécialisés appropriés, nous

réussirons à réduire les coûts élevés de travail terminographique.

Jennifer Pearson,
Salis,
Dublin City University,
Irlande.

Références

Atkins (S.), Clear (J.) & Ostler (N.), 1992 : « Corpus Design Criteria », dans *Literary and Linguistic Computing*, Oxford, Oxford University Press, vol. 7,1, p. 1-16.

Biber (D.), 1993 : « Representativeness in corpus design », dans *Literary and Linguistic Computing*, Oxford, Oxford University Press, vol. 8 (4), p. 243-257.

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

Flowerdew (J), 1992 : « Definitions in Science Lectures », dans *Applied Linguistics*, vol.13 (2), p. 202-221.

Jacquemin (C.), Royaute (J.), 1994 : « Retrieving Terms and their Variants in a Lexicalized Unification-Based Framework », dans *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, New York, Heidelberg : Springer-Verlag, p. 132-141.

Nkwenti-Azeh (Bl.), 1992 : *Positional and Combinational Characteristics of Satellite Communications Terms*, Final Report, Eurotra Project, UK-CCL-UMIST.

Pearson (J.), 1998 : *Terms in Context*, Amsterdam, John Benjamins Publishing Company

Sinclair (J.) 1994 : « Corpus Typology : A Framework for Classification » EAGLES document. 1-18, disponible maintenant dans Sinclair (J.), 1995 : « Corpus Typology : A Framework for Classification », dans Melchers (G.) & B. Warren (éd.) *Studies in Anglistics*. Stockhom, Almquist and Wiksell International, p. 17-34.

Trimble (L.), 1985. *English for Science and Technology : A Discourse Approach*, Cambridge, Cambridge University Press.

Pour une terminologie textuelle

Ce texte reprend de façon synthétique le tutoriel donné à l'ouverture des troisièmes journées *Terminologie et intelligence artificielle*. Il présente l'analyse, faite par le groupe TIA, des nouveaux enjeux pratiques, théoriques et méthodologiques de la terminologie. Sur le plan pratique, l'accroissement des besoins en terminologie dans les entreprises et dans les institutions s'accompagne d'un élargissement qualitatif considérable de la gamme des produits à base terminologique nécessaires pour répondre à ces besoins. Ces évolutions entraînent des changements en profondeur de la pratique terminologique : l'activité de construction d'une terminologie est désormais essentiellement une tâche d'analyse de corpus textuels. Ils appellent du même coup à un renouvellement théorique de la terminologie : c'est dans le cadre d'une linguistique textuelle que doivent être posées les bases théoriques de la terminologie.

Termes-clés : terminologie ; théorie de la terminologie ; terminologie textuelle ; linguistique de corpus ; ingénierie des connaissances.

1 Introduction

Ce texte constitue un résumé du tutoriel donné à l'ouverture des troisièmes journées *Terminologie et intelligence artificielle*. Les points de vue présentés sont le résultat de discussions et réflexions menées depuis plusieurs années au sein du groupe TIA, qui, compte tenu de la complémentarité disciplinaire de ses membres et de leur expérience conjuguée dans le champ de la terminologie, est en mesure d'offrir une analyse suffisamment complète et équilibrée des nouveaux enjeux, pratiques, théoriques et méthodologiques, de la terminologie. Notons cependant qu'étant donné la nouveauté et la richesse qui sous-tendent cette nouvelle thématique, un certain nombre de questions sont toujours ouvertes et continuent d'être débattues au sein du groupe.

2 Applications de la terminologie : état des lieux

Les besoins en terminologie dans les entreprises et dans les institutions se multiplient. Suite à l'utilisation généralisée des outils de bureautique, à l'internationalisation des échanges, au développement d'Internet, la production de documents sous forme électronique s'accélère sans cesse. Or pour produire, diffuser, rechercher et exploiter ces documents, les outils de gestion de l'information ont besoin de ressources terminologiques. Outre l'accroissement quantitatif de la demande, l'un des impacts essentiels de ces évolutions sur la pratique terminologique est l'élargissement qualitatif considérable de la gamme des produits à base terminologique nécessaires pour répondre à ces besoins. À côté des bases de données terminologiques multilingues classiques pour l'aide à la traduction, on voit apparaître de nouvelles productions terminologiques adaptées aux nouvelles applications de la terminologie en entreprise :

- Thésaurus pour les systèmes d'indexation automatique ;
- Index structurés pour les documentations électroniques ;
- Terminologies de référence pour les systèmes d'aide à la rédaction ;
- Référentiels terminologiques pour les systèmes de gestion de données techniques ;
- Ontologie pour les mémoires d'entreprise ou pour les systèmes d'aide à la décision ;
- Réseaux lexicaux spécialisés pour les moteurs de recherche thématique sur le *Web* ;
- Glossaires de référence et liste de termes pour les outils de communication interne et externe ;
- Bases de connaissances terminologiques pour la description de corpus de référence ;
- ...

L'essor de ces applications conduit à traiter, à l'aide d'outils, des quantités de documents considérables. Ce changement d'échelle met en évidence des phénomènes largement sous-estimés jusqu'ici.

C'est ainsi que s'impose le constat de la variabilité des terminologies : étant donné un domaine d'activité, il n'y a pas UNE terminologie, qui représenterait LE savoir sur le domaine, mais autant de terminologies que d'applications dans lesquelles ces terminologies ont été utilisées. Ces terminologies diffèrent quant aux unités retenues et à leur description selon l'application visée. Par ailleurs, la croissance terminologique, induite par la prolifération en tous sens de connaissances, entraîne la nécessité de mises à jour permanentes si l'on veut répondre aux besoins des utilisateurs.

Ce constat sur la variabilité remet en cause le principe de l'universalité des terminologies. L'expérience montre en effet qu'une terminologie élaborée pour une application à un moment donné n'est jamais identique à celle construite pour une application différente. Ces limites fortes à la réutilisabilité n'excluent pas des relations d'inclusion ou de chevauchements partiels entre terminologies dédiées à des applications différentes dans un même domaine d'activité.

L'ensemble de ces constats empiriques entraîne des changements en profondeur de la pratique terminologique : l'activité de construction d'une terminologie est désormais essentiellement une tâche d'analyse de corpus textuels. Ils appellent du même coup à un renouvellement théorique de la terminologie : c'est dans le cadre d'une linguistique textuelle que doivent être posées les bases théoriques de la terminologie.

3 Nouvelles pratiques terminologiques

L'activité de construction d'une terminologie devient avant tout une tâche d'analyse de corpus textuels. Il y a à cela deux raisons essentielles :

- Les applications de la terminologie sont le plus souvent des applications textuelles (traduction, indexation, aide à la rédaction) ; la terminologie doit « venir » des textes pour mieux y « retourner ». C'est parce qu'elle n'est jamais déliée du texte qu'on parle de « terminologie textuelle ».
- C'est dans les textes produits ou utilisés par une communauté d'experts, que sont exprimées, et donc accessibles, une bonne partie des connaissances partagées de cette communauté, c'est donc par là qu'il faut commencer l'analyse.

L'expérience montre que l'hypothèse selon laquelle l'expert d'un domaine serait le dépositaire d'un système conceptuel qu'il suffirait de mettre au jour est non productive. La tâche d'analyse terminologique vise alors avant tout la construction d'une description des structures lexicales à l'œuvre dans un corpus textuel à partir d'une analyse réglée de ce corpus.

Cette tâche ne peut être menée à bien par les experts ; la médiation d'un analyste (linguiste terminologue, cogniticien) est nécessaire, en premier lieu parce qu'on colle trop à ses propres usages langagiers ; c'est le médiateur qui garantit la distance nécessaire à l'analyse. En second lieu, la pluralité des pratiques à l'intérieur de ce que l'on a coutume d'appeler « domaine » induit des points de vue différents sur le lexique (préférences, rejets, désaccords sur la définition) qu'il faut arbitrer. La division du travail linguistique à l'intérieur d'une entreprise requiert donc un médiateur, qui a en charge l'application.

Pour chaque unité choisie, l'analyste construit une signification (type) à partir des sens (occurrences) attestés dans le corpus. Dans cette tâche, il est guidé en amont par le corpus (spécificités lexicales) et en aval par l'application (utilisation des descriptions).

L'expert doit être considéré comme un partenaire du linguiste terminologue, dans un travail de collaboration ; il est sollicité pour valider les descriptions construites par celui-ci.

Le domaine doit être lié à une pratique, maîtrisée par une communauté d'experts. Comme action (instrumentalisation du savoir propre à la technique), la pratique ne procède pas de connaissances statiques, liées à des expressions linguistiques bien stabilisées.

Avant la tâche de description lexicale, la constitution du corpus de référence est une étape essentielle, prise en charge par le linguiste terminologue. Il s'agit pour lui de collecter et de caractériser un ensemble de textes jugés pertinents pour l'application visée.

Devant la masse des données à analyser et les délais imposés, la tâche d'analyse de corpus ne peut être envisagée qu'avec l'utilisation des outils de la terminologie textuelle (concordanciers, extracteurs de candidats termes, extracteurs de relations candidates, classifieurs, etc.). L'utilisation de ces différents outils doit être encadrée par une méthodologie précisant à quel stade du processus et selon quelles modalités il convient de les utiliser.

4 Renouvellement théorique

Ces changements en profondeur de la pratique terminologique appellent un renouvellement théorique.

Les propositions théoriques et méthodologiques qui suivent ont des bases empiriques ; elles sont issues d'une analyse des nouvelles pratiques de la terminologie, et elles ont pour ambition de les améliorer. Il ne s'agit

donc pas de fonder un nouveau dogme, mais de susciter des courants de recherche variés dans le champ de la linguistique, dont chacun pourra contribuer à cet objectif.

Proposition 1: objet empirique d'une linguistique textuelle, le texte est le point de départ de la description lexicale à construire. On va du texte vers le terme. Les bases théoriques de la terminologie doivent être ancrées dans une linguistique textuelle.

Proposition 2: le terme est un construit. Il est le produit d'un travail d'analyse, mené par le linguiste terminologue, dont les choix sont guidés par une double contrainte de pertinence:

• Pertinence vis-à-vis du corpus. Il s'agit de retenir et de décrire des structures lexicales qui présentent des caractéristiques à la fois spécifiques et stables. C'est à ce stade qu'intervient la validation par l'expert.

• Pertinence vis-à-vis de l'application. Les unités finalement retenues doivent l'être en fonction de leur utilité dans l'application visée, qui s'exprime en termes d'économie et d'efficacité. La validation est à chercher du côté des utilisateurs de l'application.

La tâche de description lexicale est un travail de fixation, de stabilisation, d'homogénéisation d'une signification, dont le résultat est le terme. Il s'agit de construire un type (une signification stable) à partir des occurrences manifestées en texte. C'est ainsi qu'on parle de *normalisation*, non plus au sens que la planification terminologique donne au mot, mais au sens où la communauté d'experts «entérine» des signifiés comme des termes du domaine.

Le résultat de la description peut se présenter sous des formes diverses: réseau, liste, glossaire, etc. Il n'existe pas de format canonique. Les noms ne sont pas les seules unités lexicales à décrire. En attribuant au terme la fonction de dénommer les concepts, la terminologie classique privilégie les noms. En s'éloignant de cette approche référentielle très limitative, on est en mesure d'accueillir les autres catégories du discours (verbes, adjectifs, adverbes, prépositions, conjonctions), ainsi que des unités linguistiques plus ouvertes (syntagmes nominaux, verbaux, adjectivaux…).

5 Fin de la doctrine

Le virage méthodologique, rendu nécessaire par le travail sur corpus, crée une onde de choc qui ébranle les fondements de la doctrine wüsterienne, fortement référentielle (le mot comme étiquette du concept) et taxinomique (primauté de la relation générique/spécifique).

Il est illusoire de chercher à aménager la doctrine: le postulat d'une signification conçue comme discrète ou discrétisable, objectivante et permanente qui caractériserait le terme *a priori* est antinomique avec une terminologie textuelle. Les reformulations théoriques superficielles qui ont apparu ces dernières années sont vaines: la notion de «phraséologie», en particulier, ne peut sauver le postulat doctrinal du «terme» dans la mesure où elle est un biais pragmatique pour détourner la question du contexte et de l'unité terminologique.

Les termes ne sont pas des «unités de connaissances» qui viendraient «habiter la langue». La tâche d'analyse terminologique n'est donc pas un exercice de redécouverte d'un système notionnel préexistant qui caractériserait le domaine.

Les notions n'ont pas d'antériorité ou de priorité sur les mots: la terminologisation est un processus parallèle à l'élaboration conceptuelle.

La terminologie doit sortir d'une sémiotique du signe fondée sur la triade terme/concept/référent qui la rend inapte à aborder le texte. Cette critique du réductionnisme référentiel est à l'ordre du jour en philosophie du langage. Les appels à desserrer l'étreinte des postulats logicistes nous viennent de plusieurs côtés (Putnam, Auroux, Eco…), le positivisme logique qui a nourri la doctrine ayant été remis en cause dès la fin des années 60.

On peut constater par ailleurs que les connaissances nouvelles sont plutôt éphémères et partagées par des communautés restreintes au-delà desquelles elles ne circulent pas. On est loin de la conception idéalisée du domaine comme fragment de connaissances bien structurées, permanentes et clairement circonscrites.

On ne peut plus dire que la signification du terme est définie par la position du concept dans le système conceptuel correspondant dès lors que l'on met en doute la représentation métaphysique d'un système conceptuel préexistant représentable par l'arbre du domaine.

Il est aussi illusoire de se soumettre au référent, y compris dans les domaines techniques qui manipulent des artefacts. La description d'un objet technique est elle-même tributaire du point de vue imposé par la spécialité de l'expert. C'est en bout de chaîne, en normalisant le terme, qu'on lui prescrit une référence.

Dès que l'on abandonne l'approche logiciste du terme, étroitement liée à une sémantique véri-conditionnelle, on reconsidère le statut de la définition qui cesse d'être le résultat d'une procédure logique, métalinguistique. La définition doit être cohérente avec les sens contextuels (avérés en corpus) et pertinente vis-à-vis de l'application (comme elle s'inscrit dans une application, elle participe aux

objectifs communicationnnels, elle doit être «localisée»).

Antinomique d'une approche étroitement onomasiologique, peau de chagrin du linguistique, l'approche textuelle ouvre largement les portes à tous les acquis de l'analyse linguistique et textuelle (on dépasse ainsi la vision étroite des LSP).

L'approche textuelle est descriptive (on analyse le fonctionnement d'unités lexicales en corpus) et non plus normative: les enjeux de la planification linguistique, si légitimes soient-ils, sont dissociés du travail terminologique proprement dit. L'objectif premier de la terminologie classique était la normalisation des langages techniques via la fixation *a priori* de la signification des mots. Les textes réels qui prolifèrent et circulent en tous sens, bousculant les frontières de domaines, remettent en cause ce projet de mise en ordre des termes *a priori*. Un tel programme de régulation prescriptive est contredit par le caractère fondamentalement ouvert des textes et de leurs signes. Le constat de la plasticité du donné linguistique conduit à refonder une «bonne pratique terminologique» sur le descriptif.

Pour conclure

L'actualité de la question terminologique au travers des changements intervenus en termes d'échelle et de rythme de production, ainsi que l'ampleur des besoins, appellent un renouveau théorique et méthodologique. En permettant d'aborder systématiquement l'étude des pratiques textuelles réelles, la linguistique de corpus, avec ses techniques et ses outils, donne accès aux expressions linguistiques concrètes d'où il sera possible de faire émerger, puis de normaliser les termes pertinents. C'est une formidable ouverture pour la réflexion théorique et méthodologique. Il va sans dire que la question des procédures linguistiques présupposées par cette approche est à peine défrichée. Le groupe TIA entend participer au débat en pleine conscience de la complexité des enjeux théoriques et pratiques.

Il va également de soi que la linguistique ne peut couvrir à elle seule le processus complet de modélisation des connaissances; en fournissant la terminologie adéquate à l'application, le linguiste prépare le travail de représentation conceptuelle, mais il ne prend pas en charge la tâche de modélisation des connaissances qui aboutira à la construction d'une ontologie. Le relais est pris par l'ingénierie des connaissances. Le groupe TIA s'inscrit dans la nécessaire coopération interdisciplinaire entre linguistes et ingénieurs de la connaissance.

Didier Bourigault,
Équipe de recherche en syntaxe et sémantique,
CNRS,
Université de Toulouse Le Mirail,

Monique Slodzian,
Centre de recherche en ingénierie multilingue,
Institut national des langues et civilisations orientales,
Paris.

Extraction d'informations techniques pour la veille par l'exploitation de notions indépendantes d'un domaine

Dans cet article nous présentons une méthode originale d'analyse de textes techniques (résumés de brevets en anglais) pour l'aide à la veille technologique. Cette méthode exploite des notions indépendantes du domaine, telles que l'/amélioration/ ou l'/utilisation/, qui permettent d'identifier des informations potentiellement intéressantes. Le système *Vigitext* manipule actuellement huit notions, couplées à des règles d'exploration contextuelle, qui mettent en valeur des extraits textuels que le veilleur peut consulter pour s'informer sur le contenu du corpus.

Termes-clés :
Linguistique sur corpus spécialisés ; outils et applications.

Introduction

Cet article propose une nouvelle exploitation des sources textuelles dans une démarche de veille technologique, basée sur l'identification de notions indépendantes d'un domaine. Cette méthode permet d'obtenir des extraits de textes sans avoir à manipuler des lexiques techniques ou des calculs de fréquence basés sur les mots. Les extraits obtenus, organisés selon les notions identifiées, peuvent étonner le veilleur ou l'amener à identifier des informations fréquentes. L'objectif est d'extraire dans un premier temps des informations indépendamment de connaissances sur le domaine, afin de permettre au veilleur, dans un deuxième temps, d'utiliser ses propres connaissances pour organiser les informations obtenues.

Dans une première partie nous décrivons quelques caractéristiques de la veille technologique. Puis nous décrivons l'approche qui nous a permis d'identifier certaines notions indépendantes d'un domaine, par l'analyse d'un corpus d'abrégés descriptifs de brevets. Ensuite nous expliquons l'utilisation de l'exploration contextuelle pour l'identification d'extraits pertinents. Et enfin, nous décrivons le premier prototype *Vigitext* qui a été réalisé.

1 Les caractéristiques de la veille technologique

1.1 Généralités

Selon Henri Dou et François Jakobiak (1995), la veille technologique est définie comme étant l'observation et l'analyse de l'environnement scientifique, technique, technologique, suivies de la diffusion bien ciblée, aux responsables, des informations sélectionnées et utiles à la prise de décision stratégique. Cette définition décrit principalement des comportements humains (observation, analyse, sélection de la part du veilleur, prise de décision stratégique pour les dirigeants) et l'organisation générale (diffusion bien ciblée) qui doivent être mis en place dans une entreprise. Pour l'analyse des données électroniques, les veilleurs doivent de plus en plus utiliser des outils informatiques performants.

Parmi les sources d'informations nécessaires dans une démarche de veille technologique, l'un des principaux types de documents exploités est l'abrégé descriptif de brevet obtenu par l'interrogation de banques de données spécialisées. En effet, selon Daniel Rouach (1996), ces sources constituent la forme principale d'accès à l'information technique, à la fois sur le plan français, européen ou mondial. Elles renseignent sur les sujets d'études ou l'orientation de la compétition. Par ailleurs elles sont adaptées à l'utilisation d'outils bibliométriques qui permettent des comptages croisés

inter-champs et intra-champs. Par exemple, *Tétralogie*, mis au point à l'Irit (Dkaki *et al*., 1997), permet l'identification de réseaux de collaborations.

1.2 Notre contexte de veille

Nous avons basé nos travaux sur l'observation d'une démarche de veille technologique réalisée dans notre entreprise et en collaboration avec l'Inist sur le sujet des plantes transgéniques. Cette étude a nécessité comme ressources textuelles environ 2000 abrégés descriptifs de brevets, et 2000 résumés d'articles obtenus à partir de banques de données spécialisées. La plate-forme linguistique et infométrique ILC, développée à l'Inist par l'équipe de Xavier Polanco (1998), a été utilisée par les veilleurs pour analyser les corpus.

Notre objectif a été défini à partir de ces données: il faut réaliser un logiciel basé sur des analyses textuelles adaptées pour l'exploitation d'un corpus contenant environ 2000 abrégés descriptifs de brevets, car ces sources sont particulières et leur contenu textuel n'a apparemment jamais fait l'objet de recherches spécifiques. Nous détaillons plus loin les particularités de ces documents.

1.3 Les différents besoins des veilleurs

En observant des veilleurs lors de cette démarche de veille sur le thème des plantes transgéniques, nous avons défini deux types de besoins pour l'analyse des documents: d'une part les veilleurs veulent exploiter des analyses automatiques ne faisant appel à aucune connaissance spécifique du domaine, afin d'obtenir des résultats bruts qui peuvent étonner ou du moins informer sur le contenu global des documents sans aucun a priori terminologique. D'autre part, les veilleurs ont besoin de faire intervenir leurs propres connaissances, soit en réalisant une requête pour rechercher des informations précises, soit en regroupant des mots ou des informations afin de classer les documents par thèmes.

Cette identification de deux types de besoins est différente de celle proposée par Françoise Rousseau-Hans (1998) qui distingue trois besoins: le besoin d'exploration, le besoin de structuration et/ou de positionnement, et le besoin de prospective. Ces deux classifications mettent en valeur le fait qu'un seul outil d'analyse des données ne peut être suffisant dans une démarche de veille, puisque la diversité des besoins nécessite l'utilisation de méthodes automatiques complémentaires. Donc, pour la réalisation d'un logiciel d'aide à la veille, nous ne pouvons pas viser la résolution de tous les problèmes, mais nous devons viser la résolution d'une partie des problèmes. Ainsi, après analyse des sources, l'extraction automatique d'informations pouvant étonner, et la classification interactive des documents nous ont semblé être des besoins intéressants à résoudre.

À partir d'une étude sur les logiciels existants d'aide à la veille et d'extraction automatique d'informations, nous avons défini trois caractéristiques de notre approche: le système que l'on va réaliser n'exploitera pas de lexique technique, pour permettre l'identification de concepts nouveaux et pour être opérationnel sur n'importe quel sujet; il ne va pas évaluer l'intérêt d'une information en fonction de sa fréquence pour permettre l'identification de signaux faibles; et il va proposer des résultats simples à interpréter et à exploiter pour un veilleur.

2 Analyse d'un corpus d'abrégés descriptifs de brevets et résultats

Nous avons étudié un corpus de 30 documents de type abrégé descriptif de brevets en anglais sur le thème des plantes transgéniques. Ce corpus a été obtenu en sélectionnant les 30 documents les plus récents du corpus global (tous enregistrés en 1997). Nous avons par ailleurs observé un corpus de 105 documents de type abrégé descriptif de brevets en français sur la chimie minérale afin de vérifier l'intérêt de nos résultats dans un autre domaine.

2.1 Contenu textuel des abrégés descriptifs de brevets

Voici trois extraits d'abrégés descriptifs de brevets, qui illustrent la spécificité de ces informations textuelles très techniques:
[1] « A method for killing insect larvae which are susceptible to a lectin from Artocarpus intergritolia (jacalin), Bauhinia purpurea alba (camels foot tree) (BPS) Codium fragile (CFL), sambucus nigra (elderberry) bark (EBL), Griffonia siplicifolia, lectin II (GSL), phytolacca americana (PAL), Maclura pomifera (osage orange), (MPL), Triticum vulgare (wheat germ agglutinin, WGA), Vicia villosa (VVL), Cicer arietinum, Cystis scoparius, Helix aspersa... »
[2] « Phenylglycinamides I (R = alkyl, alkenyl, cycloalkyl; R1 = H, halo, perfluoroalkyl; R2 = CH2OH, alkoxymethyl, CHO, CO2H, carbalkoxy; R3 = H, halo, nitro, OH, etc.; R4 = cycloalkyl, alkyl, cycloalkylalkyl; R5, R6 = H, alkyl; R7 = Ph or substituted phenyl; R8, R8 = H, pyridyl, cycloalkyl, alkyl or substituted alkyl) were prepd. as angiotensin II antagonists. Thus, 2-[2-[4-[(2-butyl-4-chloro-5- formyl-1-

imidazolyl) methyl]phenyl]-2-cyclopentylacetamido]-2-phenylacetamide was prepd. via acylation of phenylglycinamide.»
[3] « A new A/T rich gene promoter enhancer (I) has more than 50% A and T bases in the nucleotide sequence. Also new are: (1) a chimeric gene comprising at least one copy of (I), a gene promoter, a coding or non- coding sequence and a terminator sequence; (2) a plant having increased expression of one or more genes, by virtue of using (I); (3) propagules of a plant as in (2); and (4) a cell harbouring a gene having increased expression as in (1). »

Ces documents, rédigés par des indexeurs spécialistes du domaine, contiennent six types de vocabulaires:
• Des abréviations générales: contg., prodn., prods, prepn.,...;
• Des informations générales: *A method for, Also new are:, having increased,...;*
• Des mots liés au vocabulaire des brevets: *claimed, specification*;
• Des données chiffrées: *3.5 wt.%, in 5' to 3' order, 1-23315 of,...*;
• Des informations très spécifiques: *S-adenosylmethionine hydrolase (SAMase), 5'-T-G-A-C-G-(T/C)-A-A..., pKS-OS-KB3.0, cDNA,...*;
• Et des notations de renvois internes: *(a), (b),..., (I), (II),...*

Les trois extraits de brevets précédents montrent bien ces caractéristiques. Il semble par ailleurs que des consignes de rédaction particulières soient données aux indexeurs, mais il est difficile d'en tenir compte car elles dépendent des fournisseurs.

2.2 Identification de notions indépendantes d'un domaine

Après avoir analysé quelques documents de type abrégé descriptif de brevets, nous avons fait les remarques suivantes: certaines notions, comme le /changement/, l'/utilisation/, l'/amélioration/, reviennent fréquemment dans les textes. En effet, ces éléments sont nécessaires pour décrire une innovation, puisqu'il faut qu'une méthode ou un élément breveté apporte quelque chose de différent par rapport à ce qui existait, ou alors il doit avoir une utilisation nouvelle par rapport aux utilisations initiales, ou alors il est amélioré par rapport à ce qui existait.

Nous nous sommes donc concentrées sur l'exploitation de ces notions, qui sont indépendantes d'un domaine, et qui semblent exprimer ou introduire des éléments informatifs.

2.3 Autres particularités de ces sources d'informations

Comme nous l'avons identifié précédemment, les abrégés descriptifs de brevets contiennent des abréviations, que l'on peut regrouper en trois catégories: l'abréviation de mots fréquents non techniques, qui consiste à réduire ou supprimer la fin des mots comme *prod.* ou *redn*, l'abréviation d'expressions techniques, qui consiste à ne conserver que quelques lettres d'une expression complexe comme *untranslated region (UTR)*, et le renvoi, ou abréviation de propositions ou définitions, qui consiste à marquer d'un chiffre ou d'une lettre une expression longue comme *An isolated nucleic acid molecule (I) encoding...* pour éviter les répétitions lourdes.

Ces abréviations, qui sont compréhensibles pour un lecteur même non spécialiste, faussent les analyses automatiques basées sur la fréquence des chaînes de caractères, puisqu'elles expriment une même information sous deux formes différentes. Nous pensons que ces notations doivent être étudiées car elles semblent faire appel à certaines régularités, et elles peuvent faciliter la compréhension automatique de documents techniques. Ainsi, une même abréviation peut permettre d'associer des orthographes (ou utilisation de traits d'union) variables: *5- enolpyruvylshikimate 3-phosphate synthase (EPSPS), 5-enol-pyruvylshikimate-3-phosphate synthase (EPSPS).* D'autre part, une variation majuscule – minuscule peut n'avoir aucun sens, comme dans *455 amino acids (aa), an amino acid (AA) sequence*. Cependant, trois lettres semblent abréger parfois quatre mots *a functional acetolactate synthase enzyme (ALS)*, parfois deux mots *adenosine deaminase (ADA)*. Enfin, un même concept peut être décrit et abrégé avec des notations différentes comme *cauliflower mosaic virus-derived 35S RN=A gene (CaMV35S), (partic. The 35S component of cauliflower mosaic virus, CaMV)*. Au cours de nos recherches, nous n'avons pas eu besoin d'approfondir l'étude de ces notations, puisque notre approche n'est pas statistique, mais il nous a semblé nécessaire de décrire cette spécificité des documents techniques observés.

3 Les notions indépendantes d'un domaine

3.1 Détails sur les notions exploitées

Nous avons actuellement défini huit notions: /amélioration/, /détérioration/, /augmentation/, /diminution/, /changement/, /production/, /résistance/, /utilisation/. Ces notions sont organisées en deux ensembles: d'une part un ensemble cohérent de notions exprimant un changement (/changement/, /amélioration/, /détérioration/, /augmentation/, /diminution/), d'autre part un ensemble de notions diverses

(/production/, /résistance/, /utilisation/). À chaque notion est associé un ensemble d'indicateurs linguistiques et un ensemble de règles qui vont permettre d'identifier les occurrences de ces notions dans les textes. Les indicateurs ont été définis à partir du dictionnaire Longman, et enrichis avec des verbes de la classification proposée par Beth Levin (1993) (ce qui a permis d'associer entre autre *manufacture* et *design* à /production/).

Ces huit notions ont été sélectionnées car elles sont indépendantes d'un domaine, elles sont assez fréquentes pour être recherchées, et elles semblent porter une information intéressante pour un veilleur, puisqu'elles répondent à des questions comme: «Qu'est-ce qui est modifié?», «Quelles sont les applications proposées?», etc. Nous avons cependant identifié d'autres notions qui seront peut-être intéressantes à exploiter par la suite: l'/appartenance/, le /contrôle/, la /nouveauté/.

L'identification des occurrences d'une notion, qui peut amener à savoir que 12 documents sur 30 abordent la notion d'/amélioration/, n'est pas forcément un résultat suffisant. Ainsi, après avoir identifié une notion dans un texte, nous allons chercher à extraire des éléments textuels du contexte pour obtenir une information intéressante. En effet, l'extraction de *altered* est moins informative que l'extraction de *synthesis of starch is altered*. Donc, pour chaque indicateur de notion, on va rechercher en plus dans le texte un ou plusieurs complément(s) d'information (détaillé par la suite) pour former un résultat.

3.2 Hiérarchisation de la notion de /changement/

En analysant les documents, nous avons remarqué que la notion de /changement/ apparaît sous différentes spécifications: parfois cette notion est plutôt qualitative, avec des valeurs comme l'/amélioration/ ou la /détérioration/, parfois cette notion est plutôt quantitative, avec des valeurs comme l'/augmentation/ ou la /diminution/, parfois cette notion reste assez générale. Nous avons donc choisi de hiérarchiser cette notion de /changement/, en lui attribuant deux sous-concepts théoriques /changement qualitatif/ et /changement quantitatif/, auxquels sont associés respectivement les sous-concepts /amélioration/, /détérioration/ et /augmentation/, /diminution/. Cette hiérarchisation a pour but d'éclaircir au maximum notre approche et les informations que l'on exploite. Elle vise aussi à faciliter l'évolution des notions: il sera peut-être jugé pertinent par les veilleurs de rajouter des sous-concepts de changement comme /vieillissement/, /rajeunissement/, /réchauffement/, /refroidissement/, /accélération/, /ralentissement/.

Pour la notion de /changement/ et les sous-notions, l'identification du complément d'information va consister à repérer l'expression de l'élément qui subit le /changement/. En effet, nous pensons que c'est ce qui est le plus informatif dans un objectif de veille technologique. Ainsi, dans l'extrait suivant: *Increasing (I) activity will also modify fruit texture and processing properties*, le fait de récupérer le sujet de *modify* ne nous semble pas très intéressant dans un objectif de mise en valeur de résultats obtenus.

3.3 Qu'est-ce qu'une information pertinente dans un texte?

Pour définir le plus précisément possible ce que nous allons considérer ici comme une information pertinente dans un texte, nous avons étudié les points de vue présentés dans d'autres travaux.

Ainsi, les systèmes qui se basent sur des termes «simples» pour indexer un document, tels que les outils classiques de gestion électronique de documents, considèrent qu'une information pertinente est un mot isolé de son contexte textuel initial. Par exemple, une analyse de cet article mettrait en valeur «notion» et «information». Nous ne sommes pas sûr que cela soit suffisant pour en décrire le contenu car ces mots isolés ne sont pas assez précis.

Dans une autre approche proposée par F. Ibekwe et G. Lallich (1995), les unités d'information linguistiquement pertinentes sont des syntagmes nominaux terminologiques, capables de représenter les concepts et objets du domaine hors du texte. L'information pertinente est alors plus riche que précédemment, car elle ne se limite pas à un terme, mais il n'est pas évident que «*plant cell*» décrive suffisamment le contenu informatif d'un texte.

Dans un système d'extraction de connaissances, comme *Coatis* réalisé par Daniela Garcia (1998), l'objectif est de mettre en valeur des relations cause-effet, chaque cause et effet correspondant à un syntagme nominal. Là encore l'information pertinente est plus complexe, car elle combine deux syntagmes nominaux et une relation particulière sous forme structurée.

Enfin, dans un système de filtrage de textes comme *Safir* présenté par Berri *et al.* (1996:141), l'élément du texte initial qui est

manipulé et étiqueté est la phrase. L'information est donc ici encore plus complète, mais moins structurée que précédemment.

Donc la notion d'information pertinente est variée. Il faut cependant remarquer que plus l'information considérée est courte (mot), plus elle est pertinente à pondérer et à regrouper en fonction de la fréquence et de cooccurrences. C'est pourquoi la plupart des systèmes opérationnels se basent sur cette information. Au contraire, les informations longues qui sont plus précises ne peuvent être pondérées, et obligent l'utilisateur à consulter chaque information.

Dans notre approche, une information pertinente est constituée d'une notion et d'un complément d'information, ce qui correspond à une relation prédicat-argument, comme par exemple: «*enhances protein import*». Nous ne souhaitons pas nous limiter à un mot, car le contenu informatif d'un tel élément nous semble trop faible, et nous ne pouvons pas non plus nous baser sur l'extraction de phrases dans des documents mal rédigés qui contiennent parfois deux phrases très longues. Les informations pertinentes que nous allons identifier vont être regroupées en fonction des notions qu'elles expriment.

3.4 Résultats possibles et intérêt pour le veilleur

Nous avons identifié plusieurs questions qui correspondent à des interrogations de veilleurs, quelle que soit leur spécialité, et qui font le lien avec l'approche que nous présentons ici. Les questions sont accompagnées d'exemples tirés d'un corpus en français de documents de chimie minérale qui se prêtent bien à une analyse ne tenant pas compte du domaine:

• Qu'est-ce qui est modifié? «transformation du phosphate», «nickel-aluminium multiphase modifié »;
• Qu'est-ce qui est amélioré? «améliorer certaines propriétés de la chevelure», «améliorant le rendement du dépôt d'ions métalliques lourds »;
• Qu'est-ce qui est produit ou créé? «produire un oxyde de nickel-lithium», «produire une électrode de batterie», «production d'une électrode positive frittée »;
• Contre quels éléments a-t-on une résistance? «résistance à la corrosion», «film de chromate résistant au noircissement», «protection de la peau contre les rayonnements ultraviolets »;
• Quelles sont les applications ou utilisations qui sont décrites? «utilisées pour des revêtements», «utiliser pour empêcher que les cheveux soient abîmés», «utile comme conditionneur de rinçage».

Ces questions, intéressantes pour le veilleur, ne sont pas formulables avec un outil classique de recherche d'information, même avec ceux qui acceptent les requêtes en langage naturel. En effet, ces outils n'exploitent que les termes de la requête, et s'intéressent uniquement à «modifié » dans la question «Qu'est-ce qui est modifié?», les autres mots étant considérés comme «vides». Ils n'exploitent pas la construction de la requête, qui permettrait de rechercher «X est modifié», ou «modification de X».

Et nous pouvons constater, avec ces exemples, que les résultats peuvent être très variés, donc inattendus parce qu'ils ne se basent pas sur des connaissances prédéfinies spécifiques au domaine abordé.

4 La méthode de l'exploration contextuelle appliquée à l'identification d'extraits pertinents

4.1 Description du complément d'information principal pour chaque notion

Dans notre objectif d'aide à la veille, nous souhaitons identifier des informations liées à des résultats obtenus, et non pas liées à des descriptions.

Dans l'extrait suivant obtenu automatiquement: «**enhance* rubber production in plants*», le complément d'information désigne la chaîne textuelle «*rubber production in plants*». Concrètement, si l'on a identifié un indicateur linguistique lié à la notion d'/amélioration/, alors on va rechercher au moins l'expression ou le mot qui décrit l'élément qui subit l'/amélioration/. Dans l'extrait ci-dessus l'élément qui subit une amélioration est exprimé par «*rubber production*». Nous avons dans un premier temps choisi de ne pas nous limiter à cet élément car si un syntagme nominal est présent juste à la suite de l'élément qui subit l'amélioration, c'est peut-être parce qu'il apporte une information supplémentaire enrichissant l'extrait initial (ici «*in...* » précise une localisation, on aurait pu avoir «*by...*» pour préciser un agent, etc.). C'est pourquoi nous avons distingué le complément d'information principal, qui désigne l'élément qui subit, et le complément d'information, qui correspond à l'ensemble des informations dans un extrait hormis le déclencheur.

Voici, pour chaque notion définie, ce que l'on va chercher à identifier principalement, c'est-à-dire le complément d'information principal:

• /changement/ (et notions dérivées): expression de ce qui subit le changement;
• /utilisation/: expression de l'application, du résultat de l'utilisation. Le complément d'information va contenir en plus quand c'est possible l'expression de l'élément qui est utilisé (pour apporter une information supplémentaire);
• /production/: expression de ce qui est produit;
• /résistance/: expression de l'élément qui a été contré par une résistance. Le complément d'information va contenir en plus, quand c'est possible, l'expression de l'élément qui a résisté.

L'extrait textuel qui est considéré comme résultat contient l'indicateur d'une notion et le complément d'information.

Dans la version en cours de réalisation de notre système, le complément d'information principal va permettre de regrouper certains extraits comme «**transformation* of cells with high efficiency* » et «**transformed* cells* » qui, même s'ils ne contiennent pas réellement des informations identiques, concernent un même sujet.

4.2 La méthode d'exploration contextuelle dans notre approche

Pour l'implémentation de notre approche, nous avons utilisé la méthode d'exploration contextuelle, mise au point par Jean-Pierre Desclés et Jean-Luc Minel (1994), qui a déjà été utilisée avec succès pour d'autres tâches (résumé automatique avec *Seraphin*, repérage d'actions dans les textes avec *Coatis*,...) pour le français.

Cette méthode doit nous permettre d'identifier dans les textes les occurrences des notions prédéfinies en s'appuyant sur:

• Des indicateurs linguistiques tels que les formes *increase*, *useful*, *transform*, associées respectivement aux notions /augmentation/, /utilisation/ et /changement/, et des indices linguistiques tels que certaines prépositions;
• Un ensemble de décisions à prendre. Nous manipulons trois sortes de décisions: l'identification ou non de l'expression d'une notion intéressante; la localisation partielle ou complète du complément d'information principal, et la construction d'un résultat, qui comporte la délimitation complète de l'extrait;
• Un ensemble de règles qui mettent en relation des indicateurs, en présence de certains indices, avec des décisions à prendre. Par exemple, si l'on repère l'indicateur *useful* suivi de l'indice *for*, alors on a repéré l'expression d'une /utilisation/, et le complément d'information principal se trouve dans le contexte droit des éléments ci-dessus.

Dans notre approche, seul le contexte local d'un indicateur est exploité (c'est-à-dire que le système peut rechercher des indices dans les mots précédents et suivants). Nous n'utilisons pas la position d'un indicateur par rapport au texte entier (début, milieu ou fin), car même si les sources qui nous intéressent semblent structurées (avec au début la description des revendications, et à la fin des applications), la limite entre ces deux informations n'est pas toujours très marquée, et notre objectif est d'éviter le silence. Enfin, les sources que nous avons analysées n'ont pas de mises en forme ou d'organisation en paragraphes. Donc nous n'exploitons pas non plus d'informations relatives à la structure du texte.

L'exploration contextuelle va être utilisée pour repérer les compléments d'informations, et pour l'analyse sémantique des indicateurs et indices linguistiques identifiés. Par exemple, si l'on a «*is produced*», alors le complément d'information principal que l'on va rechercher va correspondre au sujet, qui se trouve dans le contexte gauche de l'indicateur. Pour l'analyse sémantique, nous avons observé peu d'ambiguïté des indicateurs linguistiques que nous avons définis. Ainsi, «*change* » ou «*reduce* » sont toujours associés respectivement aux notions /changement/ et /diminution/ dans les textes techniques analysés. Cependant, quelques mots sont ambigus, comme «*raise* » qui peut désigner un mouvement, une /amélioration/ ou une /augmentation/.

4.3 Exemples de règles d'exploration contextuelle formalisées

Nous avons formalisé nos règles, afin de les rendre compréhensibles, et pour permettre une réutilisabilité de nos travaux. Pour cela, nous nous sommes inspirés du formalisme de G. Crispino (1998) adapté aux règles d'exploration contextuelle manipulées dans la plate-forme *Context*. Cette plate-forme, en cours de réalisation, regroupe différentes applications développées dans l'équipe Langage, logique informatique et cognition (Lalic) du Centre d'analyse et de mathématiques sociales (Cams): *Seek*, *Seraphin*, etc.

Chaque corps de règle contient des conditions et des actions. Nous utilisons «decl » pour désigner l'élément déclencheur, et «~x » désigne une variable qui ne doit pas être présente dans le contexte. Voici les opérations liées aux conditions, que nous avons défini dans notre propre formalisme:
• DistanceEnMots (Mot1, Mot2) < N + 1: N mots peuvent séparer Mot1 de Mot2 dans le texte.
• Position(Mot): renvoie la position de Mot dans le texte.

Voici quelques unes des différentes actions possibles:
• IdentifierDebutCIP(Mot): identification du premier mot du Complément d'Information Principal.
• CreerExtraitAvant (Mot): création d'un extrait à partir de Mot et dans son contexte gauche.
• CreerExtraitApres (Mot): création d'un extrait à partir de Mot et dans son contexte droit.
• CreerExtraitAvantApres (Mot): création d'un extrait à partir de Mot, dans son contexte droit et dans son contexte gauche.

La création d'un extrait entraîne l'application de règles de délimitation d'un segment textuel. Ainsi, lorsque la règle CreerExtraitApres(Mot) est appelée, elle va récupérer le segment textuel qui débute à Mot, et qui se termine avant une ponctuation ou une conjonction (il y a des cas particuliers).

Voici un exemple de règle qui concerne la notion d'/utilisation/:

La règle ci-dessus est toujours pertinente, ce qui n'est pas le cas de toutes les règles. Ainsi, la règle suivante concerne la notion de /résistance/, et plus particulièrement le verbe *treat*.

Règle R_RES_V_3
Exemples: « , *or treats diseases.* »,
« *for use in treating plants infected with leaf scald disease and/or reducing ...* »,
« *are useful for gene therapy to treat various non-inherited or inherited genetic or epigenetic diseases or disorders such as...* »
Conditions L1:= { *treat, treats, treating* }, L2:= { *of* }
Cond:= (decl (L1,!((~x (L2) / DistanceEnMots(decl, ~x) = 1,
Position(decl) < Position(~x)
Actions IdentifierDebutCIP(MotSuivant(decl))
CreerExtraitApres(decl)

Cette règle permet de ne pas repérer le contexte suivant: «*e.g. to treat of diabetes*», mais elle repère l'extrait suivant qui n'est pas pertinent: «*(b) treating samples of the culture*». Nous avons ignoré les autres sens de *treat*: «*treat someone well* », «*treat with someone* », «*to give someone a treat* », car ils ne semblent pas faire partie du type de discours scientifiques que nous étudions.

Nous insistons sur le fait que deux tâches différentes sont effectuées: d'une part, chaque règle localise le complément d'information principal quand un contexte est identifié; d'autre part, un extrait est créé en fonction de différents paramètres: recherche éventuelle du sujet, récupération possible d'un complément de localisation, longueur de l'extrait adaptée à l'interface de visualisation.

Règle R_UTIL_5
Exemples: «*They may be used to treat arterial hypertension and atherosclerosis (claimed) as well as coronary heart disease, cardiac insufficiency,*»,
«*The plants are esp. useful for the management of turf grass for golf course, sport field etc.*»,
«*Such reductions are useful e.g. to degreen fruits, seeds, floral parts or other edible plant parts.* »
Conditions L1:= { *is, are, was, were, be, being* }, L2:= { *used, useful* }, L3:= { *to, for* }
Cond:=(x (L1, (decl (L2, (y (L3 / DistanceEnMots(x,decl) < 2 et DistanceEnMots(decl,y) < 4 et Position(x) < Position(decl) < Position(y).
Actions IdentifierDebutCIP(MotSuivant(y))
CreerExtraitAvantApres(decl)

4.4 Exemples de résultats obtenus pour la notion d'/utilisation/

Cette notion exploite trois indicateurs: «*use*», «*useful* » et «*application*». Treize règles d'exploration contextuelle permettent d'obtenir les occurrences recherchées de ces indicateurs. Ainsi, on ne va pas considérer «*used as a reporter* » comme un contexte pertinent, tandis que «*used to modify...*» sera un contexte pertinent. Actuellement «*used in* » est aussi considéré comme introduisant une application, mais pas «*used as* » ni «*use sth*». Le tableau ci-dessous montre les résultats tels qu'ils sont présentés à l'utilisateur: une première colonne contient la référence au document source, une deuxième colonne contient la référence au champ source, et une troisième colonne contient les extraits textuels.

REFDOC	CHAMP	CONTENU
3	AB	The plants are *useful* for the management of turf grass for golf course
4	AB	reductions are *useful* e.g. to degreen fruits
23	TI	*used* to modify the sensitivity of a plant to light
28	AB	The peptides can be *used* to inhibit digestion and egg development in blood-sucking insects
10	AB	*used* to reduce the time for germination of seeds
18	AB	*used* to engineer
13	AB	The products can be *used* to produce plants
25	AB	The method is particularly *useful* for the transfection of antisense 1- aminocyclopropane-1-carboylate (ACC) synthase and ACC oxidase sequences
10	AB	A further *application* is the treatment of Varroa-Milben disease in bee colonies

Les exemples ci-dessus permettent de mettre en valeur le type de résultats que l'on peut obtenir, et montrent aussi quelques limites de l'approche. Ainsi l'extrait « **used* to engineer* », issu du contexte suivant : « *It can also be used to engineer, e.g. herbicide,...* » ne contient pas assez d'informations pour être pertinent. De même, « *The products can be *used* to produce plants* », extrait du contexte suivant : « *The products can be used to produce plants which are resistant to...* » n'est pas assez précis. Par contre, l'extrait suivant : « *The method is particularly *useful* for the transfection of antisense 1- aminocyclopropane-1- carboylate (ACC) synthase and ACC oxidase sequences* », issu de : « *The method is particularly useful for the transfection of antisense 1- aminocyclopropane-1- carboylate (ACC) synthase and ACC oxidase sequences which in turn prevent...* » est trop long, ce qui va être difficile à gérer au niveau de l'interface de visualisation des extraits.

Ces résultats montrent que l'approche présentée ici peut réellement permettre l'identification d'informations qui peuvent être étonnantes. Cependant, seul le veilleur peut estimer l'intérêt des extraits, en fonction de ses connaissances et ses attentes. Il est d'autre part important de noter que cette classification par notion des extraits n'est pas une classification thématique telle que peuvent le proposer certains outils.

4.5 Cas particuliers

Actuellement nous n'avons pas pris en compte la négation. En effet, le corpus initial ne contient que deux négations, qui ne s'appliquent pas à des informations jugées pertinentes. D'autre part, nous ne savons pas comment prendre en compte l'occurrence d'une notion concernée par une négation. De plus, certains cas peuvent être complexes, comme dans le cas suivant : « *which does not normally produce Sgp.* ». Dans le programme actuel, « *not produce Sgp* » est associé à la notion de /production/.

D'autre part, nous avons choisi de ne pas exploiter dans un premier temps la transitivité des verbes prédéfinis. En effet, les verbes prédéfinis sont en majorité transitifs. Et pour les quelques verbes prédéfinis qui peuvent être à la fois transitifs et intransitifs (comme *change*), nous n'avons observé dans le corpus initial que des utilisations à la forme transitive. Il sera peut-être nécessaire de modifier ce choix suivant les résultats que l'on va obtenir par la suite.

5. Le prototype opérationnel *Vigitext*

Un premier prototype a été réalisé fin 1998 avec le langage de programmation orienté objet Java. Ce prototype analyse des documents définis dans une base de données, et permet à un utilisateur de naviguer dans le corpus par l'intermédiaire d'une interface de visualisation des extraits identifiés.

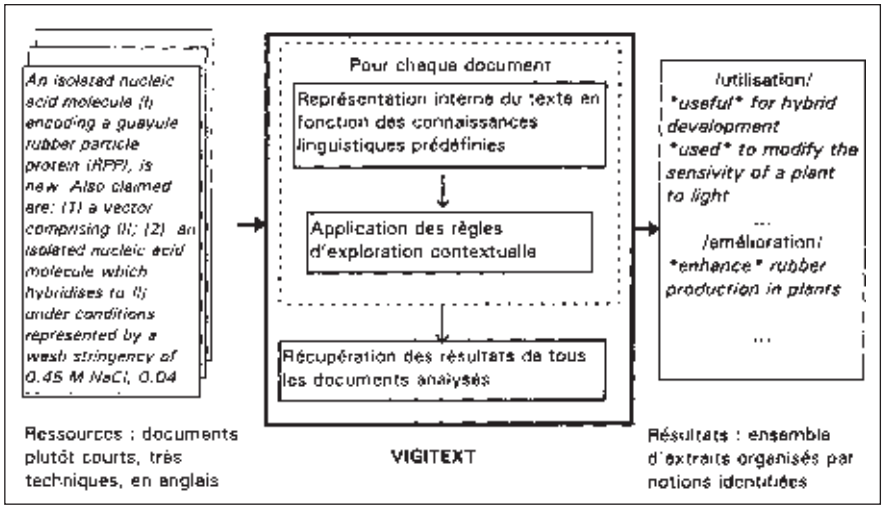
Nous ne présentons pas dans cet article de comparaison avec d'autres outils, car nous n'avons actuellement repéré aucun système présentant une approche équivalente, c'est-à-dire qui analyse le texte des résumés de brevets pour mettre en valeur des extraits informatifs, et qui est utilisable par des veilleurs.

5.1 Étapes de traitement et caractéristiques du prototype

Voici les différentes étapes de notre système : (voir page suivante).

Les documents en entrée ne subissent aucun traitement préalable. Ainsi, pour le prototype, nous avons utilisé en entrée une base de données documentaire contenant trois champs : numéro du document, titre, résumé. Par la suite il est possible d'imaginer l'utilisation de données textuelles ayant d'autres formats plus riches (textes structurés, documents XML...).

De plus, comme la méthode est basée sur l'exploitation de notions



indépendantes du domaine scientifique et technique, *Vigitext* est opérationnel pour traiter des bases de documents abordant d'autres thèmes que le sujet des plantes transgéniques.

Actuellement notre système s'appuie sur plus de 250 formes prédéfinies, dont environ 150 correspondent à 46 indicateurs liés aux notions (principalement des verbes, mais aussi quelques noms et adjectifs), et une centaine correspond à des indices linguistiques (prépositions, articles, formes de l'auxiliaire être, etc.).

À partir d'un corpus de 30 résumés de brevets en anglais, nous avons défini environ 67 contextes qui sont reconnus à l'aide des règles d'exploration contextuelle.

5.2 Module de regroupement interactif des extraits

Parmi les besoins exprimés par les veilleurs, nous avons identifié le besoin de créer des ensembles d'informations en fonction de critères personnels. En effet, dans une démarche de veille ce sont uniquement les connaissances du veilleur qui vont lui permettre d'identifier les informations étonnantes. Pour cela, nous allons ajouter dans la prochaine version de *Vigitext* un module de regroupement des extraits. La prise en compte du complément d'information principal va permettre de faciliter ce regroupement interactif, en créant automatiquement par exemple des groupes d'extraits liés à /changement/ + «*cell*». Nous envisageons aussi d'ajouter un groupe «supprimé» qui va contenir tous les extraits jugés inintéressants par le veilleur.

Avec Vigitext, les documents sources sont réduits en ensembles d'extraits, ce qui simplifie l'approche du corpus pour le veilleur. Mais cela ne vise pas à remplacer le veilleur en fournissant une analyse complète des données. En effet, les notions prédéfinies ne sont pas des concepts du domaine, elles ne fournissent donc pas une organisation thématique des documents de la base : nous pensons que seul le veilleur est apte à obtenir un tel résultat. C'est pourquoi nous pensons qu'une utilisation de notre système doit combiner la lecture et le tri des extraits.

5.3 Évaluation du prototype

Nous avons réalisé une évaluation de notre méthode sur un corpus nouveau de 30 documents, obtenu à partir du corpus global sur les plantes transgéniques, contenant plus de 2 000 abrégés descriptifs de brevets. Ce nouveau corpus correspond aux 30 derniers documents insérés dans la base source en 1994. Pour l'évaluation, un extrait pertinent bien formé doit contenir l'expression d'une notion et le complément d'information principal tel qu'il a été défini précédemment.

L'analyse de ces 30 documents a permis d'obtenir 131 extraits. Sur ces 131 extraits, 8 contiennent des informations déjà exprimées (même notion, même document), 11 sont intéressants mais mal délimités (exemple : «*e.g. caterpillars and is *used* to control* »). D'autre part, nous avons identifié 9 informations exprimant l'une des notions prédéfinies qui n'ont pas été repérées.

Le résultat de l'évaluation est le suivant : 112 extraits pertinents bien formés ont été repérés, et 9 extraits pertinents n'ont pas été repérés. En considérant que les 11 extraits mal délimités sont potentiellement pertinents, les taux de rappel et de pertinence (ou le taux d'extraction pertinente) sont de 0,85.

Un essai a d'autre part été réalisé sur un corpus de 10 résumés d'articles techniques en anglais concernant la santé, et nous avons obtenu 27 extraits. Cela prouve que notre approche donne des résultats sur un autre domaine, et pour un autre type de document proche.

6 Conclusion et perspectives

Notre méthode remplit les trois conditions suivantes : pas de terminologie spécifique, pas de calculs

statistiques, extraction d'informations quel que soit le domaine. Les extraits obtenus sont organisés selon des notions générales permettant l'identification d'informations. Ces résultats sont obtenus à partir des sources textuelles de type abrégés descriptifs de brevets en anglais.

Le veilleur peut consulter les extraits selon les notions qui l'intéressent, et identifier des travaux qui l'étonnent, ou qui sont fréquents, ou qui sont rares et très prometteurs, ou qui sont hors sujet. Il va ensuite pouvoir organiser les extraits obtenus selon ses connaissances personnelles.

6.1 Enrichissement du lexique et affinement des règles

Du point de vue linguistique, il nous reste à affiner les lexiques associés aux notions prédéfinies, et à ajouter de nouvelles notions selon les besoins exprimés par les veilleurs. Nous avons commencé à prendre en compte la notion de /détection-identification/, qui a été repérée par un veilleur, et nous adaptons actuellement une partie du lexique de la notion de /causalité/ définie sur le français par Daniela Garcia (1998). Une évolution future pourrait être d'adapter l'approche à des documents français, mais il est difficile de prévoir le coût d'adaptation du lexique et des règles pour le français, En effet, contrairement à un système comme *Ana*, développé par Chantal Enguehard, qui n'utilise ni lexique, ni grammaire, et qui est plus ou moins indépendant de la langue, notre approche nécessite un travail linguistique pour transposer les règles et le lexique à une autre langue.

6.2 Amélioration du système

Nous allons d'autre part réaliser le module de regroupement interactif d'informations qui doit permettre au veilleur d'organiser les extraits en groupes pertinents, et de supprimer des extraits inintéressants selon lui.

Bénédicte Goujon
Équipe Langage, logique informatique et cognition,
Centre d'analyse et de mathématiques sociales
et Bureau Van Dijk ingénieurs conseils Paris,
France.

Bibliographie

Berri (J.), Cartier (E.), Desclés (J.-P.), Jackiewicz (A.), Minel (J.-L.), 1996: *Safir, système automatique de filtrage de textes*, actes de TALN-96, p. 140-149.

Crispino (G.), 1998: *Éléments pour la manipulation de textes dans la plate-forme Context*, Rapport interne du Cams, UMR CNRS - EHESS, Université Paris-Sorbonne, août 1998.

Desclés (J.-P.), Minel (J.-L.), 1994: «L'exploration contextuelle», dans *Le résumé par exploration contextuelle*, rapport interne du Cams n°95/1, recueil des communications effectuées aux rencontres Cognisciences-Est, 25 novembre 1994, Nancy, p. 3-17.

Dkaki (T.), Dousset (B.), Mothe (J.), 1997: *Mining information in order to extract hidden and strategic information*, RIAO'97, p. 32-51.

Dou (H.), Jakobiak (F.), 1995: «De l'information documentaire à la veille technologique pour l'entreprise: enjeux, aspects généraux et définitions», dans *Veille technologique et compétitivité*, Dunod, p. 3.

Enguehard (C.), informations sur le système *Ana* à l'adresse: http://www.sciences.univ-nantes.fr/irin/ln/termino/home.html.

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

Ibekwe (F.), Lallich(G.), 1995: *L'analyse linguistique automatique comme point de départ pour la recherche de tendances thématiques dans les publications scientifiques*, dans colloque Île Rousse 1995, p. 39.

Levin (B.), 1993: *English Verb Classes And Alternations*, The University of Chicago Press, 1993.

Polanco (X.), François (C.), Royauté (J.), Grivel (L.), Besagni (D.), Dejean (M.), Oto (C.), 1998: *Organisation et gestion des connaissances en veille scientifique et technologique*, VSST'98, p. 328-337.

Rouach (D.), 1996: *La veille technologique et l'intelligence économique*, PUF, p. 37.

Rousseau-Hans (F.), 1998: «L'analyse de corpus d'information comme support de la veille stratégique», dans *Document numérique* Vol. 2, n°2/1998, p. 189.

Polanco (X.), François (C.), Royauté (J.), Grivel (L.), Besagni (D.), Dejean (M.), Oto (C.), 1998: «Organisation et gestion des connaissances en veille scientifique et technologique», dans VSST'98, p. 328-337.

Repérage des entités nommées : un enjeu pour les systèmes de veille(1)

Cet article présente un système de repérage et d'acquisition d'entités nommées pour le français, c'est-à-dire de repérage des noms de personnes, de géographie, de société, les organisations et les dates. Le système utilise des techniques originales pour l'acquisition d'entités nouvelles et de règles servant à reconnaître ces entités. Il permet de développer rapidement des ressources pour un nouveau domaine et atteindre ainsi des performances quasi comparables à celles de systèmes nécessitant davantage de connaissances *a priori*.

Termes clés : Linguistique sur corpus spécialisés ; outils et applications.

1 Introduction

Le développement de la documentation électronique a révélé le besoin de nouveaux accès au texte. C'est dans ce cadre que l'on a assisté récemment au développement d'analyseurs terminologiques permettant de fournir des bases de connaissances et des index à partir de textes techniques ou spécialisés (Bourigault *et al.*, 1996). De son côté, la veille économique ou stratégique a besoin de pouvoir repérer rapidement les noms d'entreprises et de dirigeants figurant dans les textes. Cette tâche est accomplie par le repérage de ce qui est souvent qualifié d'« entités nommées », c'est-à-dire les noms de personnes, de géographie, de société, les organisations et les dates (Appelt *et al.* 1995) (Gaizauskas *et al.* 1995) (Mani *et al.* 1996) (Wacholder 1996). On voit ainsi apparaître des systèmes d'aide à la décision, permettant aux opérateurs de faire des choix en fonction d'éléments mis en évidence à même le texte, de manière automatique.

Cet article présente un système de repérage et d'acquisition d'entités nommées pour le français. Le système utilise pour partie des techniques classiques pour la reconnaissance d'entités, techniques issues des systèmes d'extraction d'information et d'acquisition de terminologie. Il met également en jeu des techniques d'analyse particulières pour repérer et typer les mots inconnus puis proposer automatiquement des candidats potentiels au statut d'entité nommée. Le système surligne ces séquences à même le texte en utilisant un système de balisage de type hypertexte.

Nous situons d'abord le repérage des entités nommées parmi les nouveaux modes d'accès au texte puis nous présentons le système développé. Celui-ci est basé sur des ressources limitées *a priori* mais possède un module d'acquisition et d'enrichissement interactif de ses données par analyse de corpus. Un modèle de marquage au moyen de balises XML permet un accès assisté au texte. Enfin, une expérience portant sur un corpus issu du journal *Le Monde* donne lieu à une évaluation du processus d'acquisition et du repérage d'entités.

2 Analyse de textes pour l'aide à la décision

Le repérage des entités nommées permet une analyse rapide du texte facilitant la prise de décision. Cette tâche se situe dans la mouvance des nouveaux courants de recherche sur les moyens d'accès au texte.

2.1 L'extraction d'information comme nouveau moyen d'accès au texte

L'extraction d'information désigne l'activité qui consiste à remplir automatiquement une banque de données à partir de textes en langage naturel (Appelt *et al.* 1993) (Pazienza 1997). Il ne s'agit donc pas simplement de filtrage de documents, où le système renvoie un ensemble de textes pertinents par rapport à une

(1) Je remercie Adeline Nazarenko ainsi que les trois relecteurs anonymes de TIA pour leurs commentaires qui ont permis de notablement améliorer cet article. Ma réflexion a également été enrichie de conversations avec Benoît Habert (Limsi), Christian Jacquemin (Limsi) et Célestin Sedogbo (Thomson-CSF).

question. L'extraction met en œuvre une analyse fine du texte pour produire du factuel (remplir un formulaire prédéfini, en anglais *template*) et apporter des réponses précises aux questions des utilisateurs plutôt que du texte brut (Wilks 1997).

Les systèmes d'extraction d'information ont connu un fort développement depuis la fin des années 80 sous l'impulsion des conférences américaines MUC *(Messages Understanding Conferences)*. On assiste toutefois depuis quelques années à un infléchissement dans l'évolution de ce type de système. Une analyse strictement locale ne permet pas d'analyser correctement certains phénomènes linguistiques comme la résolution des anaphores. La portabilité des systèmes est également limitée. L'élaboration d'un système d'extraction passe par la définition de patrons identifiant une information donnée qui pourra ainsi être reconnue dans les textes. Il s'agit d'un travail long et fastidieux qu'il faut refaire pour toute application portant sur un domaine nouveau.

L'extraction d'information pose également question quant au statut à attribuer à l'information extraite. Le but traditionnel de l'extraction, que nous mentionnons ci-dessus, consiste à remplir une base de données à partir de textes en langage naturel. Mais les taux de réussite, qui varient de 50 à 90 % suivant la complexité de la tâche, ne permettent pas de considérer l'information extraite comme certifiée (Hirschman 1997). Diverses expériences qui ont pu être faites au contact de professionnels de l'information montrent un besoin de retour au texte, de vérification de l'information. Souvent, l'expert a besoin du texte original parce qu'il mène à partir du document une démarche d'interprétation particulière. La demande ne va donc pas tant vers des systèmes d'extraction que vers des systèmes de marquage d'éléments textuels pertinents aidant l'interprétation des textes.

On assiste ainsi, essentiellement depuis MUC-6 (1995), à un retour du corpus dans le cadre des systèmes d'extraction. Les systèmes doivent à la fois être centrés sur l'utilisateur (qui définit ce qu'il souhaite extraire) et sur le corpus (qui est sollicité de manière interactive par l'utilisateur). Ce rapport nouveau au texte a été constaté en amont pour la définition de patrons qui se base alors sur une analyse minutieuse du corpus et sur des procédures d'apprentissage automatique. Il est aussi nécessaire, en aval, de permettre les allers et venues entre l'information extraite et le corpus, pour fournir des systèmes d'aide à l'interprétation et à la décision. Par exemple, dans le domaine de la veille économique, mettre en évidence le nom de personnes et d'entreprises dans le texte permet à l'analyste de décider immédiatement si le texte est intéressant ou non pour son activité.

Les technologies d'extraction d'information s'insèrent donc naturellement dans le cadre des nouveaux moyens d'accès au texte. Ces systèmes permettent une navigation au moyen d'hyperliens et un accès à partir de bases de connaissances et d'index structurés (Assadi 1998). Le système d'extraction que nous présentons porte sur le repérage des entités nommées. Ce type d'analyse se révèle indispensable dans certains secteurs comme la veille économique et stratégique, où l'expert doit déterminer dans les dix secondes si le texte est intéressant ou non. Ces entités, en fournissant des indications sur les noms de personnes et d'entreprise en jeu, sur les dates et les lieux, permettent un jugement rapide sur la pertinence et l'importance d'un document.

2.2 Les systèmes d'extraction pour le français

Le projet européen *Écran* (LE-2110, 1996-98) a permis de développer des systèmes d'extraction complets pour plusieurs langues, notamment l'anglais et le français (Basili *et al.* 1998) (Poibeau 1998). C'est dans ce cadre qu'avait été écrit une première version d'un module de repérage des entités nommées du français, dont certains résultats ont été réutilisés lors du développement de notre système. Le système *Écran* souffre toutefois de plusieurs limitations importantes. En particulier, aucune analyse fine des mots inconnus n'est faite, ce qui provoque un nombre d'erreurs relativement important lors du repérage des entités nommées. De plus, aucun processus d'apprentissage n'avait été prévu dans le module développé pour *Écran* : il est dès lors nécessaire de procéder à un travail manuel important pour enrichir les dictionnaires de base lors du développement d'une application portant sur un nouveau domaine.

Écran est, à notre connaissance, le seul système d'extraction complet développé à ce jour pour le français et s'intéressant à des textes écrits[2]. La société Lexiquest a développé une grammaire des noms de personnes et de société. Ce système est en partie paramétrable (insertion de lexiques particuliers) mais il ne permet pas un développement incrémental des ressources. Plusieurs systèmes

(2) *Écran* permet d'extraire des textes non seulement des entités nommées mais aussi d'autres éléments comme les relations entre entités à partir d'une analyse minutieuse du verbe. Le projet *Exibum* (Kosseim & Lapalme, 98) vise à développer un système d'extraction mais ne comporte pas de module d'analyse des entités nommées.

d'analyse locale ont été développés à partir de lexique-grammaire. Maurel (1989) a ainsi développé un système de repérage des dates par automates et tables d'acceptabilité. Plus récemment, Belleil (1997) a présenté un système de repérage des toponymes français et Sénellart (1998) un système de repérage des noms de ministre à partir du journal *Le Monde*. Ces approches sont essentiellement basées sur des dictionnaires exhaustifs du sous-langage concerné. Sénellart propose une méthode interactive d'acquisition à partir d'automates, mais l'approche vise avant tout à constituer un dictionnaire exhaustif sur un domaine restreint.

Le projet *Xicop* (3) (eXtraction d'Information de COrpus de Parole), actuellement en cours de développement au Limsi/CNRS (Orsay), est un système de repérage d'entités nommées pour des corpus audio. S'il partage certains des principes adoptés ici (ensemble de règles permettant le repérage d'entités basé sur l'analyse d'« amorces » et de mots inconnus), en revanche il ne peut s'appuyer sur la distinction minuscule / majuscule et doit supporter des données bruitées. Les techniques que nous présentons sont davantage liées à l'analyse de documents écrits.

3 Un système de repérage d'entités nommées

L'architecture que nous adoptons est héritée des systèmes d'extraction d'information ayant participé à MUC. Ceux-ci procèdent par analyse locale progressivement étendue par un ensemble d'automates à état finis. On parle alors d'automates en cascade (Appelt *et al*. 1993).

(3) http://genesis.limsi.fr/~xicop.

3.1 Première étape : étiquetage lexical

Dans un premier temps, le système procède à la fois à un découpage du texte en unités minimales et à un étiquetage lexical :

- Les nombres sont analysés suivant des critères formels (un nombre est composé de chiffres et, éventuellement, d'un tiret, d'un point ou d'une virgule) ;
- Un dictionnaire de noms propres permet d'identifier et de typer un certain nombre de noms de personnes et de lieux. La reconnaissance est souvent partielle à ce stade (le système peut par exemple avoir reconnu le prénom mais pas le nom). Le dictionnaire est particulièrement développé au niveau des listes semi-fermées comme les prénoms, beaucoup moins pour les noms ;
- Un dictionnaire d'amorces permet de reconnaître certaines séquences importantes pour la suite comme *M.* (pour *Monsieur*) ou *SA* (pour *Société anonyme*) ;
- Un algorithme permet de traiter et de normaliser les sigles comme *I.B.M* ou *I.B.M.* (avec ou sans point à la fin). En revanche, le signe *IBM*, écrit sans point, est considéré comme un mot inconnu s'il n'a pas été stocké préalablement dans le dictionnaire de noms propres ;
- Les mots inconnus sont étiquetés en tant que tels au moyen d'un dictionnaire général de la langue. Ils reçoivent une étiquette différente suivant qu'ils sont en majuscules ou en minuscules, au début de phrase ou non ;
- Enfin, parmi les mots figurant dans le dictionnaire de la langue générale, ceux qui commencent par une majuscule sont étiquetés. Une distinction est faite suivant que le mot se trouve en début de phrase ou non.

Cette analyse est effectuée au moyen d'un arbre de décision. Les choix que doit faire le système s'appliquent dans l'ordre des points énumérés ci-dessus. Une fois cette étape franchie, le texte est enrichi de balises entourant les mots repérés. Nous donnons ci-dessous un texte étiqueté par l'analyseur. Chaque mot reconnu est entouré de l'étiquette qui convient.

<COMPANY> Fiat </COMPANY> possède <NUMBER> 90 </NUMBER> % de <PERSON> <COMPANY>Ferrari </COMPANY></PERSON>.

<COMPANY> Fiat </COMPANY> contrôle désormais <NUMBER> 90 </NUMBER> % du capital de <PERSON> <COMPANY> Ferrari </COMPANY></PERSON>, a annoncé le <NUMBER> 7 </NUMBER> <DATE type=MONTH> septembre </DATE> le groupe automobile <NATIONALITY> italien </NATIONALITY>. <COMPANY> Fiat </COMPANY> qui détenait déjà <NUMBER> 50 </NUMBER> % de la firme de <LOCATION> Modène </LOCATION>, précise qu'il a racheté «ces mois derniers» les <NUMBER> 40 </NUMBER> % qui appartenaient à <PERSON> Enzo </PERSON> <PERSON><COMPANY> Ferrari </COMPANY></PERSON>. L'opération s'est donc déroulée avant le décès du «<UNKNOWN> commandatore </UNKNOWN>», le <NUMBER> 14 </NUMBER> <DATE type=MONTH> août </DATE>. Les <NUMBER> 10 </NUMBER> % restants appartiennent au fils adoptif d' <PERSON> Enzo </PERSON> <PERSON><COMPANY> Ferrari </COMPANY></PERSON>, <TR_PERSON> M. </TR_PERSON> <PERSON>Piero </PERSON> <UFIRSTUNKNOWN> Lardi </UFIRSTUNKNOWN>. - (<UCASEUNKNOWN>AFP <UCASEUNKNOWN>.)

Dans l'exemple précédent apparaissent les principales étiquettes utilisées par le système, c'est-à-dire des noms de personnes (<PERSON>) et de société (<COMPANY>), des nombres (<NUMBER>), des dates (<DATE type=MONTH>), des amorces de noms de personnes (<TR_PERSON>) et des mots inconnus (<UNKNOWN>, <UFIRSTUNKNOWN>, <UCASEUNKNOWN> suivant que le mot est inconnu, qu'il est inconnu et que sa première lettre est en majuscule ou qu'il est inconnu et entièrement en majuscules). On remarque l'ambiguïté de *Ferrari*, qui peut être considéré soit comme un nom de personne, soit comme un nom de société. En conséquence, *Ferrari* est doublement étiqueté: <PERSON><COMPANY> Ferrari </PERSON></COMPANY>.

À la suite de cette étape, le fonctionnement du système ne repose plus que sur l'analyse des suites d'étiquettes, indépendamment des formes effectivement attestées dans le texte.

3.2 Deuxième étape: reconnaissance de séquences pertinentes

Une grammaire des entités nommées permet ensuite de repérer parmi les suites d'éléments repérés à l'étape précédente celles qui sont susceptibles de former une entité.

Cette grammaire est écrite sous forme de règles de réécriture avec comme partie droite une étiquette syntagmatique et comme partie gauche une expression régulière. Les règles sont compilées et s'appliquent en respectant certaines heuristiques:

- Les règles les plus longues (celles qui contiennent le plus d'éléments) s'appliquent les premières,
- Une règle ne peut plus s'appliquer à l'intérieur d'une séquence précédemment reconnue (autrement dit, une séquence reconnue forme par la suite un îlot inanalysable).
- Si deux règles de même longueur peuvent s'appliquer, le résultat est aléatoire (il dépend de l'ordre dans lequel les règles ont été compilées).

Ce principe, en évitant d'avoir à gérer des conflits, permet d'assurer la robustesse du système.

Nous donnons à la suite un exemple de règle. TR_PERSON désigne une amorce de nom de personne (« M.», «Mme»,…), PERSON? un nom ou un prénom optionnel et UFIRSTUNKNOWN+ un ou plusieurs mots inconnus dont l'initiale est en majuscule:

```
// M. Piero Lardi
// M. Lardi
TR_PERSON PERSON?
UFIRSTUNKNOWN+ ==>
PERSON
```

On voit ici que, même si le mot *Lardi* n'est pas connu du système, il est possible d'inférer, d'après cette règle, qu'il s'agit d'un nom de personne et compléter la base de noms propres. La séquence *M. Piero Lardi* forme par la suite un tout étiqueté <PERSON> qui ne peut plus être décomposé. L'évaluation tend à prouver que des règles simples déterminent avec un bon taux de réussite les éléments faisant partie de l'entité et son type (Sénellart 1998).

Il va de soi que le système fonctionne d'autant mieux qu'il a des listes de noms propres relativement complètes dès le début. En l'absence de telles listes, le système pourra faire des prédictions sur les mots inconnus apparaissant dans des règles spécifiques, comme la règle ci-dessus qui permet d'inférer que le mot inconnu représente un nom de personne. En revanche, la tâche sera plus longue pour les mots inconnus isolés, puisque ceux-ci se retrouvent dans des listes très hétérogènes (avec environ 20 % d'éléments pertinents seulement). L'avantage de notre système est d'être robuste et de permettre un fonctionnement en «mode dégradé», c'est-à-dire avec peu de connaissances préalables. Les performances en mode dégradé pourront toutefois être faibles et dépendent étroitement du domaine et des données disponibles (cf. la partie «évaluation», où le système sur un nouveau domaine ne couvre tout d'abord que 20 % des entités présentes).

Un système de préférence est mis en place pour résoudre les ambiguïtés comme pour le mot *Ferrari*. Les règles permettant de regrouper le plus grand nombre d'éléments s'appliquent d'abord. Cette procédure se fait en tenant compte des ambiguïtés et résout la plupart des cas. Par exemple dans le cas de <PERSON> Enzo </PERSON> <PERSON> <COMPANY> Ferrari </COMPANY></PERSON>, l'étiquette <COMPANY> ne sera pas retenue du fait qu'une règle permet de reconnaître *Enzo Ferrari* comme un nom de personne. Quand le contexte ne permet pas de désambiguïser une expression, un système de préférence permet de choisir l'étiquette la plus probable (*Ferrari* est classé comme étant préférentiellement un nom de société) ou, en l'absence d'élément répertorié permettant de décider, le système choisit une étiquette de façon aléatoire. On obtient alors le résultat suivant:

<COMPANY> Fiat </COMPANY> possède 90 % de <COMPANY> Ferrari </COMPANY>.

<COMPANY> Fiat </COMPANY> contrôle désormais 90 % du capital de <COMPANY> Ferrari </COMPANY>, a annoncé le <DATE> 7 septembre </DATE> le groupe automobile <NATIONALITY> italien </NATIONALITY>. <COMPANY> Fiat </COMPANY> qui détenait déjà

50 % de la firme de <LOCATION> Modène </LOCATION>, précise qu'il a racheté «ces mois derniers» les 40 % qui appartenaient à <PERSON> Enzo Ferrari </PERSON>. L'opération s'est donc déroulée avant le décès du « <UNKNOWN> commandatore </UNKNOWN>», le <DATE> 14 août </DATE>. Les 10% restants appartiennent au fils adoptif d'<PERSON> Enzo Ferrari </PERSON>, <PERSON> M. Piero Lardi </PERSON>. - (<UCASEUNKNOWN> AFP </UCASEUNKNOWN>.)

La sortie de cette étape d'analyse est un texte enrichi de nouvelles étiquettes indiquant les entités reconnues et leur type. Dans l'exemple ci-dessus, nous simplifions en effaçant, pour des raisons de lisibilité, une partie des étiquettes posées lors de l'étape précédente. En fait, chaque niveau de marquage rajoute son jeu d'étiquettes sur le texte sans supprimer les étiquettes du niveau précédent. L'utilisateur doit préciser les éléments qu'il souhaite voir apparaître dans l'interface(4).

3.3 Regroupement d'entités co-référentes

Une dernière passe permet de repérer de nouvelles entités, notamment parmi les mots inconnus et ceux qui n'ont pu être typés par une des règles de la grammaire. Ce niveau opère de façon simple en gardant simplement la mémoire des derniers éléments trouvés. On part du principe que les textes analysés ont une certaine cohérence et que celle-ci passe par la reprise des mêmes éléments au sein de périodes dans le texte. Nous entendons ici par période une suite de phrases ou de paragraphes marqués par la reprise d'éléments linguistiques (pronoms et autres anaphores, déictiques, etc.) ou thématiques communs.

<COMPANY id=1> Fiat </COMPANY> possède 90 % de <COMPANY id=2> Ferrari </COMPANY>.

<COMPANY id=1> Fiat </COMPANY> contrôle désormais 90 % du capital de <COMPANY id=2> Ferrari </COMPANY>, a annoncé le <DATE id=3> 7 septembre </DATE> le groupe automobile <NATIONALITY id=4> italien </NATIONALITY>. <COMPANY id=1> Fiat </COMPANY> qui détenait déjà 50% de la firme de <LOCATION id=5> Modène </LOCATION>, précise qu'il a racheté «ces mois derniers» les 40 % qui appartenaient à <PERSON id=6> Enzo Ferrari </PERSON>. L'opération s'est donc déroulée avant le décès du «<UNKNOWN> commandatore </UNKNOWN>», le <DATE id=7> 14 août </DATE>. Les 10 % restants appartiennent au fils adoptif d' <PERSON id=6> Enzo Ferrari </PERSON>, <PERSON id=8> M. Piero Lardi </PERSON>. - (<UCASEUNKNOWN> AFP </UCASEUNKNOWN>.)

Le texte que nous avons pris en exemple ne pose pas de problème particulier pour l'analyse des éléments co-référents. L'ambiguïté sur *Ferrari* en tant que nom de personne ou nom de société a été levée au niveau précédent. Ici, il suffit au système de s'assurer que les éléments co-référents sont de même type. Cette étape de l'analyse permet malgré tout de mettre en évidence, parmi les occurrences d'une chaîne de caractères ambiguë, celles qui réfèrent à un même objet. Ainsi, *Lardi*, qui était à l'origine un mot inconnu commençant par une majuscule (<UPPERUNKNOWN>) peut à l'issue du traitement être typé comme <PERSON> et être inséré automatiquement dans le dictionnaire.

Les pronoms et autres reprises anaphoriques ne sont pas pris en compte par le système à l'heure actuelle. Il ne permet pas non plus de relier des éléments comme *Fiat* et *la firme de Milan*: ceci serait possible mais nécessiterait une base de données gérant les synonymes. Enfin, les entités complexes comme le «fils adoptif d'Enzo Ferrari » ne sont pas analysées.

C'est également à ce niveau que certains mots inconnus peuvent être répertoriés et classés en fonction de leur contexte d'apparition. Un mot inconnu peut être étiqueté et typé correctement s'il peut être mis en rapport avec une entité déjà étiquetée. L'algorithme d'analyse est relativement simple: on co-indice un mot inconnu avec une entité connue si et seulement si elle possède une chaîne de caractère pertinente en commun. Les amorces (« trigger words ») permettant de reconnaître les entités ne sont pas considérées comme des chaînes de caractères pertinentes (<TR_PERSON> par exemple), à l'inverse d'éléments déjà étiquetés comme <PERSON> ou <DATE>.

Dans le cas présent, *AFP* a pu être repéré comme mot inconnu en majuscule. *Commandatore* est repéré comme mot inconnu sans majuscule mais aucun élément du contexte ne permet d'aller plus loin, c'est-à-dire de les typer plus précisément. Si les mots isolés sont ici pertinents (le «commandatore » est le surnom de *Enzo Ferrari*, *l'AFP* est une société), ce n'est évidemment pas toujours le cas.

(4) Le document XML, pour être visualisable dans un navigateur Web, doit au préalable être transformé en document HTML. Le système utilise une feuille de style associée qui permet de ne visualiser qu'une partie des éléments étiquetés.

3.4 Marquage hypertexte des séquences repérées

Les éléments repérés sont marqués au moyen de balises XML. XML est un langage permettant de définir une structure de document au moyen de balises. Ce type de marquage, où l'on peut définir ses propres balises, permet un repérage aisé des entités grâce à un langage de description normalisé. Par ailleurs, XML possèdera à terme ses propres « parseurs » et ne nécessitera pas, en principe, le développement de visualiseurs spécifiques.

Une DTD (Définition de type de document) est définie pour rendre directement compte du marquage vu précédemment. Une DTD permet de définir une grammaire décrivant une classe de documents. La DTD définie pour l'application contient deux blocs principaux d'instructions. Le premier consiste en un marquage minimal rendant compte de la forme du document, essentiellement pour préserver le découpage en paragraphes. Le deuxième bloc rend compte des étiquettes que nous avons vues précédemment pour le repérage des entités nommées proprement dites. La grammaire est stockée sous la forme d'un fichier de règles de réécriture classiques pour des raisons de lisibilité pour les personnes amenées à développer les ressources. Elle est équivalente et conforme à la DTD développée conjointement, au sens où la grammaire est conforme à la hiérarchie des balises définie dans la DTD. Pour l'instant, la cohérence entre la DTD et les règles est maintenue à la main, du fait de la stabilité et de la simplicité de la DTD. Il serait envisageable d'intégrer au système un parseur XML pour assurer cette cohérence de manière automatique.

En l'absence de visualiseurs de document XML au moment de l'implémentation, le système génère *in fine* un document au format HTML 4, avec feuille de style séparée. Le résultat est alors visible dans n'importe quel navigateur Web. Actuellement les entités apparaissent en surbrillance dans le texte original. Le système sera prochainement étendu pour pouvoir réagir de manière dynamique. Il sera alors possible de faire apparaître toutes les occurrences d'une entité donnée quelle que soit sa forme linguistique (entités co-référentes). Une fiche synthétique sera aussi accessible par simple clic sur l'entité afin d'avoir un système interactif, à base d'hyperliens.

4 Expérimentation et évaluation sur un corpus issu du journal *Le Monde*

Dans cette section, nous détaillons le protocole d'expérimentation et les résultats de l'évaluation qui a été menée.

4.1 Constitution du corpus

Un corpus d'évaluation a été constitué à partir de textes concernant les affaires internationales, tirés des archives du journal *Le Monde*. Les textes ont été sélectionnés en faisant une simple requête sur la base avec le mot clé *Affaires internationales*. Ces textes font usage d'un grand nombre de noms d'entités et constituent donc un excellent corpus d'évaluation. Celui-ci est distinct du corpus concernant l'actualité économique qui avait préalablement été utilisé lors du développement du système. Il s'agit, de plus, d'une source fréquemment utilisée en veille stratégique. Pour l'évaluation, nous avons isolé 25 textes d'environ 20 000 mots au total.

4.2 Évaluation du processus d'acquisition de noms d'entités

Nous avons évalué la couverture et la pertinence des mots inconnus relevés automatiquement par l'analyseur à partir d'un dictionnaire noyau de *Écran*. À l'instar des candidats termes en terminologie, on obtient des candidats entités, c'est-à-dire des groupes nominaux candidats au statut d'entité (de la même façon que le système Lexter permet d'obtenir des candidats termes (Bourigault *et al.* 1996). Ces candidats sont réparties en classes plus ou moins homogènes. Les éléments inconnus du dictionnaire qui apparaissent dans des séquences reconnues par les règles de repérage des entités nommées (étape 2, *cf.* point 3.2) forment une classe pertinente à plus de 95 %, comme en témoigne le tableau suivant :

Abdel	Baas
Abdullah	Baden-Baden
AFP	Bagdad
Ahmadi	Baker
Akaba	Balladur
Al	Bassorah
Amalric	Beaucé
Andréani	Bérégovoy
Andreotti	Bin
AP	Blum
Arafat	Bréhier
Aramco	Brent
Arens	Bush
Aziz	Bush-Baker

Mots inconnus ayant été repérés par au moins une règle de la grammaire et proposés comme candidats entités

Cette classe comporte aussi bien des mots correspondant à des noms de personnes qu'à des noms de lieux ou d'organisation. Il n'a pas été jugé utile de pousser plus loin le typage proposé, même si cela devrait être possible en fonction de la règle ayant permis de reconnaître un mot donné. La vérification du type proposé pour

un élément donné risque en effet de se révéler aussi coûteuse qu'un typage purement manuel.

Les classes de mots inconnus apparaissant en dehors de tout schéma identifié sont moins pertinentes. Elles permettent cependant de compléter la couverture du dictionnaire général actuel. Une évaluation manuelle révèle environ 20 % de candidats pertinents pour les entités nommées, 40 % de mots absents du dictionnaire général que l'on peut ainsi compléter et 40 % de segments divers qui ont échappé à l'analyse (notamment les séquences avec tiret comme *a-t-il*, *semble-t-il* ou *quasi-inconditionnel*, avec emploi incorrect du tiret devant l'adjectif).

Les règles de grammaires sont écrites de manière incrémentale, de manière à progressivement couvrir la majeure partie des éléments pertinents. Le développement de la grammaire est interactif: l'introduction de nouveaux éléments dans les dictionnaires permet de reconnaître de manière partielle de nouvelles entités. Leur pleine reconnaissance requiert de nouvelles règles qui à leur tour font apparaître de nouvelles entités... Par exemple, la règle

TR_PERSON PERSON? UFIRSTUNKNOWN+ ==> PERSON

ne permet de reconnaître que partiellement *M. Frederik de Klerk*. Il est alors nécessaire de créer une nouvelle règle de type

TR_PERSON PERSON? PREP UFIRSTUNKNOWN+ ==> PERSON.

Mais cette règle, à son tour, ne permet de reconnaître que partiellement un nom tel que *M. Jean de la Guerivière*. La règle précédente est alors modifiée comme suit:

TR_PERSON PERSON? PREP DET? UFIRSTUNKNOWN+ ==> PERSON.

On peut choisir de rendre aussi la préposition PREP optionnelle pour fusionner la règle avec la première mentionnée ci-dessus. En pratique, on évite d'écrire des règles avec trop d'éléments optionnels pour des raisons de lisibilité. Toute règle doit par ailleurs comporter un élément lexical non optionnel pour éviter de reconnaître une amorce ou une préposition de manière isolée.

Les performances du système sont de l'ordre de 20 % d'entités reconnues au début de l'expérience. Après 3 itérations (analyse du corpus, insertion des mots inconnus repérés par le système dans les différents dictionnaires, ajouts de règles, nouvelle analyse du corpus), un taux de reconnaissance d'environ 90 % est atteint grâce la méthode d'enrichissement incrémental proposée (5). Ces performances ont été atteintes en environ 3 heures de travail sur le corpus d'entraînement.

4.3 Évaluation du système

Une partie du corpus a été réservée pour l'évaluation du système, une fois celui-ci mis au point. Ce corpus a été étiqueté manuellement d'un côté et traité par le système de l'autre. On a enfin procédé à une comparaison des résultats du système avec ceux obtenus par l'analyste humain, en l'occurrence le concepteur du système (6).

Les règles écrites sous la forme d'expressions régulières lors de l'étape précédente sont compilées en un ensemble de règles simples. Dans le cadre de notre expérimentation, le système dispose d'une quarantaine de règles utilisant des métacaractères qui génèrent près de 200 règles simples. On garantit ainsi une analyse de complexité linéaire. Un corpus de 126 Ko a été traité en moins de 30 secondes sur un PC (K6 à 200 MHz), produisant comme résultat un fichier au format XML de 470 Ko.

Le taux de reconnaissance est d'environ 80 % des entités du texte correctement analysées. Les règles sont suffisamment contraintes pour avoir un bruit quasi nul. Les principales causes de reconnaissance partielle ou de silence sont les suivantes:

- Incomplétude de la grammaire ou du dictionnaire (l'expression *M. Valéry Giscard* a été reconnue alors que c'est l'expression *M. Valéry Giscard d'Estaing* qui figurait dans le texte);
- Transformations ayant échappé à l'analyse (*Charette Hervé* au lieu de *Charette Hervé de*, introduire une règle pour reconnaître ce type de transformation reviendrait à introduire énormément de bruit, puisqu'on reconnaîtrait beaucoup de noms de personne suivi de la préposition *de*);
- Orthographe approximative (*Langelier Jean Pierre* pour *Langelier Jean-Pierre*);
- Mot fortement ambigu (*Le Monde* qui dans le cas présent est un nom propre, mais qui est difficilement analysable hors contexte).

Ces résultats se situent légèrement en dessous des scores

(5) L'évaluation de la couverture du système sur le corpus d'entraînement a été faite pour donner un ordre d'idée. Le protocole d'évaluation que nous exposons dans la partie suivante (comparaison des résultats du système avec un étiquetage manuel) a été mis en œuvre, quant à elle, sur la partie du corpus qui n'avait pas servi au développement des ressources du système.

(6) En toute rigueur, un étiquetage par un utilisateur extérieur serait souhaitable.

obtenus par les systèmes anglo-saxons participant à MUC, qui oscillent entre 85 et 95 %. Mais le but de notre système est avant tout d'obtenir un score honorable après un temps d'adaptation limité (ici environ 3 heures), contrairement aux systèmes MUC qui nécessitent un travail manuel important pour développer les ressources les plus complètes possibles. Notons enfin que le processus d'acquisition fonctionne grâce à un système de règles. L'utilisateur peut ainsi facilement contrôler l'activité du système, contrairement aux résultats obtenus par des méthodes statistiques (Cucchiarelli, 98).

Le bon résultat que nous obtenons sur le corpus *Le Monde* devrait être comparé avec des résultats provenant d'autres corpus. Il est certain qu'il y a une corrélation assez forte entre le résultat et le genre textuel, le style ou l'auteur. Ainsi, le repérage des noms propres dans le journal *Le Monde* est facilité par le caractère quasi systématique du *M.* ou *M^me^* qui précède le nom. Il n'en irait pas de même dans un corpus issu du journal *Libération*, qui n'en fait pas un usage aussi systématique. Un travail d'adaptation en fonction du corpus est donc nécessaire.

5 Conclusion

Nous avons présenté un système d'extraction d'entités nommées fonctionnant à partir de ressources limitées. Un processus d'acquisition permet de proposer des éléments susceptibles d'entrer dans la formation de nouvelles entités et d'enrichir semi-automatiquement le dictionnaire. Ce processus d'acquisition permet aussi d'étendre progressivement la grammaire décrivant les entités. Le système développé est donc extrêmement portable d'un domaine à l'autre, voire d'une langue à l'autre. On a montré qu'on obtenait des résultats légèrement inférieurs à ceux des systèmes anglo-saxons ayant participé à MUC en un temps de développement limité. En injectant davantage de connaissances dans notre système, il est possible d'obtenir des résultats comparables.

Notre système se classe dans la famille des outils d'aide à l'accès à la documentation électronique. Il fait partie d'un ensemble plus important d'outils d'extraction d'information à partir de textes en cours de développement. Le but de ce système est de fournir à l'utilisateur les moyens de définir sa requête ainsi que des outils d'exploration interactive de corpus. Nous nous situons également dans la perspective d'une nouvelle ergonomie linguistique en laissant à l'utilisateur le soin de faire les choix qui lui incombent, par rapport à la tâche qu'il cherche à accomplir. Dans ce cadre, le repérage des entités nommées est une aide à la décision, qui doit permettre à l'analyste chargé d'effectuer une veille sur un domaine donné de déterminer rapidement si un document est pertinent ou non. Des applications sont en cours dans le domaine de la veille économique, où les mêmes noms de société et de personnes reviennent régulièrement et sont particulièrement représentatifs.

Poibeau, Thierry,
Thomson-CSF/LCR,
LIPN,
Université Paris-Nord,
Villetaneuse.

Bibliographie

Appelt (D.E.), Hobbs (J.), Bear (J.), Israel (D.), Kameyana (M.) & Tyson (M.), 1993, «FASTUS: a finite-state processor for information extraction from real-world text», dans *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, Chambéry, p. 1172-1178.

Appelt (D.E.), Bear (J.), Hobbs (J.), Israel (D.), Kameyana (M.), Kehler (A.), Martin (D.), Myers (K.) & Tyson (M.), 1995, *The FASTUS name recognition grammar*, Rapport Interne, SRI international.

Assadi (H.), 1998, *Constructions d'ontologies à partir de textes techniques*, Thèse de l'université Paris 6.

Basili (R.), Catizone (R.), Pazienza (M.T.), Stevenson (M.), Velardi (P.), Vindigni (M.) & Wilks (Y.), 1998, «An Empirical approach to lexical tuning», Acte du workshop «Adapting lexical and corpus resources to sublanguages and applications», dans *First International Conference on Resources and Evaluation*, Grenade.

Belleil (C.), 1997, *Reconnaissance, typage et traitement des coréférences des toponymes français et de leurs gentilés par dictionnaire électronique relationnel*, thèse de doctorat en informatique, Université de Nantes.

Bourigault (D.), Gonzalez-Mullier (I.) & Gros (C.), 1996, «LEXTER, a Natural Language Processing Tool for Terminology Extraction», dans *Proceedings Euralex'96*, Göteborg.

Cucchiarelli (A.), Luzi (D.) & Velardi (P.), 1998, «Using corpus evidence for automatic gazetteer extension», dans *First International Conference on Resources and Evaluation*, Grenade, p. 83-88.

Gaizauskas (R.), Wakao (T.), Humphreys (K.), Cunningham (H.) & Wilks (Y.), 1995, «University of Sheffield: description of the LaSIE system as used for MUC-6», dans *Proceedings of the sixth Message Understanding Conference*, Morgan Kaufmann Publishers, Los Altos, CA, p. 207-220.

Hirschman (L.), 1997, «Language Understanding Evaluations: A Case Study of MUC and ATIS», dans *Speech and Language Technology (SALT) Club Workshop on Evaluation in Speech and Language Technology* (Sheffield).

Kosseim (L.) & Lapalme (G.), 1998, «EXIBUM: un système expérimental d'extraction bilingue», dans *Actes Rifra'98* (Sfax), p. 129-140.

Mani (I.), McMillian (R.), Luperfoy (S.), Lusher (E.) & Laskowski (S.), 1996, « Identifying unknown proper names in newswire text », dans Pustejovsky (J.) & Boguraev (B.), dir., *Corpus processing for lexical acquisition*, MIT Press, Cambridge, MA.

Maurel (D.), 1989, *Reconnaissance des séquences de mots par automate, adverbes de date du français*, Thèse de Doctorat en Informatique, Université Paris 7.

MUC-6, 1995, *Proceedings of the sixth Message Understanding Conference*, Morgan Kaufmann Publisher.

Pazienza (M.T.), dir., 1997, *Information extraction (a multidisciplinary approach to an emerging information technology), International Summer School SCIE'97 (Frascati 14-18 juil. 1997)*, Springer Verlag (Lecture Notes in Computer Science).

Poibeau (T.), 1998, « Extraction d'information : adaptation lexicale et calcul dynamique du sens », dans *Actes Rifra'98* (Sfax), p. 141-153.

Sénellart (J.), 1998, « Locating noun phrases with finite state transducers », dans *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, Montréal, p. 1212-1217.

Wacholder (N.), Ravin (Y.) & Choi (M.), 1996, « Disambiguation of proper names in text », dans *Proceedings of the fifth Applied Natural Language Conference*, Washington, DC.

Wilks (Y.), 1997, « Information Extraction as a core language technology », dans Pazienza (M.T.), dir., *Information Extraction*, Springer Verlag.

Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés

Cet article présente une méthode pour l'extraction de relations sémantiques à partir de marqueurs lexico-syntaxiques. Notre démarche vise à assurer une articulation entre connaissances linguistiques «générales» et connaissances linguistiques «spécialisées», c'est-à-dire observables dans les textes techniques qui sont notre objet d'étude. Nous appuyant sur cette méthode, nous avons développé *Caméléon*, un environnement assistant complètement la gestion, l'adaptation et la réutilisation de bases de marqueurs suivant de nouveaux corpus, afin de générer des relations sémantiques candidates fiables entre les termes de ce corpus.

Termes-clés:
Recherche de relations conceptuelles; corpus spécialisés; construction de modèles; marqueurs de relations sémantiques.

1 Introduction

Nous proposons dans cet article une méthode et des outils pour enrichir des modèles du domaine dédiés à l'amélioration de la recherche d'information dans les systèmes de données techniques textuelles. Dans ces systèmes, l'introduction de nouveaux textes induit l'acquisition de nouvelles connaissances. Identifier plus rapidement les nouveaux concepts et surtout les nouvelles relations entre concepts pour permettre une exploitation la plus automatique possible de nouveaux textes constitue l'objectif global de nos travaux.

Pour cela, les outils d'aide à l'extraction de termes et les résultats sur les techniques d'extraction de relations sémantiques, issus de recherches en TALN (Traitement automatique du langage naturel), sont aujourd'hui très utiles. Ainsi, nous avons mis en œuvre une méthode d'extraction de relations entre concepts s'appuyant sur une technique de TALN de plus en plus utilisée: l'application de marqueurs linguistiques (généraux ou spécialisés) sur des corpus. Pour nous, un marqueur est défini comme un patron lexico-syntaxique qui désigne dans le discours une relation sémantique entre des termes, et ce, avec une certaine précision pour un corpus donné.

Nous nous intéressons plus particulièrement aux problèmes liés à la recherche, la gestion et la caractérisation de tels marqueurs sur de nouveaux corpus. Nous prônons une démarche qui vise à utiliser au maximum les connaissances linguistiques générales sur les marqueurs et garantit l'articulation entre ces connaissances «générales» et des connaissances linguistiques «spécialisées» qui sont des comportements spécifiques observés dans les corpus étudiés.

Notre méthode et nos outils se situent en amont de l'approche *Rex*(1) de gestion du retour d'expérience, mais ne sont pas spécifiques à cette méthode.

2 Extraire des relations sémantiques entre concepts

La plupart des travaux en TALN visant à extraire des relations sémantiques entre termes s'appuient sur l'hypothèse selon laquelle les classes de mots et les formules syntaxiques d'une grammaire sont en correspondance étroite avec les classes d'objets et leurs relations.

2.1 Étude statistique de la distribution de contextes

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. Par exemple, Smadja (1993) étudie les fréquences de cooccurrences de mots pour proposer des relations entre ces mots. Dans la continuité de ces travaux et en s'appuyant sur une analyse syntaxique du corpus, Grefenstette (1994) déduit

(1) La méthode *Rex* (Retour d'expérience), développée au CEA (Commissariat à l'énergie atomique), est actuellement diffusée par la société Euriware (Saint-Quentin-en-Yvelines) dans le cadre de son activité Knowledge Management.

de la distribution des contextes lexico-syntaxiques (Nom-adjectif, Nom-Nom, Nom-verbe) des mots et des classes de mots partageant les mêmes contextes. Plus récemment, d'autres travaux (Assadi (1998), Habert (1996), se sont intéressés à la composition des syntagmes nominaux d'un corpus et à la distribution de leurs unités pour mettre au jour des classes de termes partageant les mêmes *modifieurs* (adjectifs ou expansions). Toutefois, si ces méthodes statistiques robustes déduisent des classes de termes proches (souvent de mono-termes), elles ne préfigurent pas des relations entre les termes d'une classe qui peuvent tout à la fois être antonymes, synonymes, hyponymes ou simplement posséder un trait sémantique commun. Dans ces travaux, l'identification des relations entre termes est une tâche qui est assistée par l'analyse des contextes caractéristiques de la classe mais qui incombe totalement au cogniticien. Ces méthodes ne générant pas directement de relation sémantique étiquetée entre des termes, elles ne correspondent pas exactement à notre objectif.

2.2 Marqueurs linguistiques

Nous avons donc étudié les méthodes qui permettent d'identifier automatiquement des relations sémantiques entre termes à partir de patrons lexico-syntaxiques. Les principes de ce type de méthodes sont présentés par Hearst (1992).
M. Hearst, partant du constat qu'il existe de nombreuses façons par lesquelles une relation peut s'exprimer dans le langage, appelle marqueur un patron lexico-syntaxique qui apparaît fréquemment dans plusieurs genres de textes et indique, sous un certain point de vue et en dehors des phénomènes rhétoriques, presque toujours la même relation. Par exemple, le patron : NP0 tel que NP1, NP2, ...,(et|ou) NPn, (où NPi est un syntagme nominal) est associé à la relation d'hyponymie entre NP0 et NPi (i étant compris entre 1 et n). Afin de découvrir ces marqueurs, elle propose la méthodologie suivante :
• (1) Sélectionner une relation R ;
• (2) Établir une liste de termes entre lesquels on a identifié cette relation ;
• (3) Trouver dans le corpus des endroits où les termes reliés sont cooccurrents puis observer les régularités dans ces environnements ;
• (4) Une fois qu'un nouveau marqueur a été identifié, l'utiliser pour trouver d'autres couples en relation et revenir en (2).

Les résultats de cette méthode sont jugés encourageants pour la relation d'hyponymie mais Hearst souligne les difficultés qu'elle a eu à généraliser ce type d'approche à des relations plus difficilement caractérisables *a priori* comme la relation partie-de. Elle souligne néanmoins le fait qu'elle obtient de bons résultats pour l'identification de relations spécifiques par des marqueurs spécifiques, dédiés à son corpus. Cependant, avec cette méthode, il suffit d'une seule occurrence du patron lexico-syntaxique pour trouver une relation entre termes. À partir de cette méthode, plusieurs types de travaux ont proposé des méthodes pour établir des listes de marqueurs pour des relations binaires données, chacune considérant le rôle de la spécificité du texte comme plus ou moins déterminant dans la définition des marqueurs.

2.2.1 Généralisation de l'approche par marqueurs

Visant à généraliser l'approche d'extraction de relations à partir de marqueurs linguistiques, Jouis (1993) et Garcia (1998) ont construit d'importantes bases de marqueurs, respectivement pour les relations de type «statique» (identification, incompatibilités, mesures, comparaison, inclusion, appartenance, localisation, partie/tout, possession et attribution) et causales. Dans ces travaux, la démarche d'acquisition des marqueurs n'est pas exactement celle proposée par Hearst (1992). En effet, ici, une relation observable dans plusieurs domaines est choisie puis une étude des phrases où cette relation est identifiée est réalisée à partir de textes de genres et domaines divers, souvent en «langue générale», par opposition à langue spécialisée telle que défini dans Lerat (1995). Les différentes interactions entre les concepts désignés par la relation permettent d'identifier très précisément différents types de sous-catégories pour la relation choisie. La liste des marqueurs est alors ventilée sur les sous-catégories de la relation.

Afin de garantir la fiabilité des marqueurs, cette approche, adoptée dans *Seek* (Jouis 1993) et *Coatis* (Garcia 1998), a conduit à la création de listes de marqueurs lexico-syntaxiques très contraints, associés à des relations très particulières. L'identification des relations entre les termes du domaine se veut ici précise, même si elle peut s'avérer peu productive du fait de la complexité des marqueurs.

En validant de tels systèmes sur des corpus techniques, Jouis (1997) a montré que les listes de marqueurs ainsi établies étaient difficilement réutilisables car certains marqueurs ne désignent pas toujours la relation attendue dans un nouveau corpus et que des parties de marqueurs peuvent être polysémiques dans certains domaines.

Si l'adaptation de ces bases de marqueurs «génériques» à des corpus spécialisés engendre du bruit, le problème de la création de marqueurs dédiés n'est pas abordé dans ces systèmes où, semble-t-il, il n'est pas prévu l'acquisition de nouveaux marqueurs et de nouvelles relations selon le corpus, le type de discours ou l'objectif de modélisation.

2.2.2 Apprentissage de marqueurs à partir de textes

D'autres travaux considèrent que les marqueurs et les types de relations sont largement spécifiques aux domaines techniques, aux types de discours étudiés et aux objectifs liés à la reconnaissance de relations entre termes. Ainsi, Rousselot (1996) et Morin (1998) automatisent la méthode d'acquisition de marqueurs proposée par Hearst avec leurs logiciels *Reltex* et *Prométhée*. Ici, la relation étudiée n'est pas forcément générique ou fixée *a priori*. Elle est choisie par rapport à un objectif de modélisation. Les marqueurs trouvés par *Reltex* ou *Prométhée* sont issus du corpus et lui sont parfois très spécifiques.

Poussant encore plus loin la notion de spécificité, Riloff (1996) propose une méthode d'extraction de marqueurs pour les instances de classes sémantiques d'un domaine. Cette méthode s'appuie sur l'étude semi-automatique des contextes des termes de la classe, du point de vue lexico-syntaxique et de leur spécificité lexicale (les verbes spécifiques au domaine étant ici le point d'entrée privilégié pour l'identification de marqueurs). Dédiés à l'extraction d'information dans des domaines très spécialisés, les marqueurs trouvés ici ne sont pas réutilisables pour une autre classe de termes.

Enfin, pour certains linguistes terminologues, la projection de marqueurs permet d'assister le processus d'étude systématique des différentes acceptions des termes. Dans les travaux de Davidson (1998) ou Condamines (1998), construire une base de marqueurs revient à définir des filtres qui permettront de consulter des contextes d'occurrences «conceptuellement riches» autour d'un terme, de repérer les occurrences les plus intéressantes pour la définition des différentes acceptions dans le corpus. Ces travaux soulignent l'importance d'une phase d'adaptation des marqueurs «généraux» par rapport aux corpus et confirment la plus grande pertinence des marqueurs de relations lorsqu'ils sont dédiés au corpus.

Des différentes approches présentées, il ressort que la précision d'un marqueur, et donc la qualité des relations sémantiques qu'il propose, est un facteur difficilement caractérisable *a priori*, sans recours au corpus. Ceci est essentiellement dû au fait que les marqueurs linguistiques peuvent être ambigus s'ils sont peu contraints. Cependant, cette polysémie peut-être résolue dans certains corpus, donnant de fait au marqueur une meilleure validité à représenter la relation (*i.e.* «chez le X, Y» pour la relation d'hyponymie dans un corpus spécialisé en agro-alimentaire (Morin (1998)).

Afin de réduire cette polysémie potentielle, les marqueurs, ainsi que leurs relations associées, peuvent être restreints, c'est-à-dire spécialisés puis validés sur de gros corpus. Jouis (1997) et Davidson (1998) montrent néanmoins que si la restriction d'un marqueur entraîne automatiquement une diminution de ses occurrences, elle ne conduit pas nécessairement à un meilleur rappel et à une meilleure précision(2).

Considérant ces remarques, nous présentons dans la partie suivante une méthode et des outils permettant de gérer, réutiliser mais surtout adapter, préciser et enrichir une base de marqueurs à partir de nouveaux corpus.

(2) Les notions de rappel et de précision sont des mesures utilisées en recherche d'information. Le rappel = Nombre de bonnes réponses du système / Nombre total de bonnes réponses et la précision = Nombre de bonnes réponses du système / (Nombre de bonnes réponses du système + Nombre de mauvaises réponses du système)

3 Méthode de génération de relations sémantiques à partir de marqueurs

Si nous avons cherché à réutiliser au maximum les connaissances sur la langue générale qu'illustrent les marqueurs, c'est de façon empirique. Ainsi, c'est par rapport à un corpus que nous jugeons un marqueur précis quant à sa capacité à désigner une relation sémantique donnée. De fait, nous considérons qu'une méthode d'extraction de relations sémantiques à partir de marqueurs doit inclure une phase de validation des marqueurs par rapport au corpus. Nous concevons les bases de marqueurs comme des systèmes ouverts où certains marqueurs sont issus de connaissances linguistiques sur la langue générale (marqueurs dits génériques) et adaptables à la spécificité des corpus étudiés, et d'autres sont entièrement définis par rapport aux objets et discours propres au corpus.

Les marqueurs dits génériques sont structurés dans une base de marqueurs génériques. Nous évaluons leur précision pour chaque nouveau corpus (1). Puis, par une méthode proche de celle de Hearst (2), nous procédons à une phase d'acquisition de marqueurs spécifiques pour les relations de la base générique et d'autres relations jugées intéressantes pour un objectif applicatif donné. Les marqueurs validés (3), nous les projetons sur le corpus préalablement indexé par l'ensemble des termes du domaine (4) afin de proposer des relations sémantiques pour l'enrichissement des modèles du domaine (partie inférieure de la figure 1).

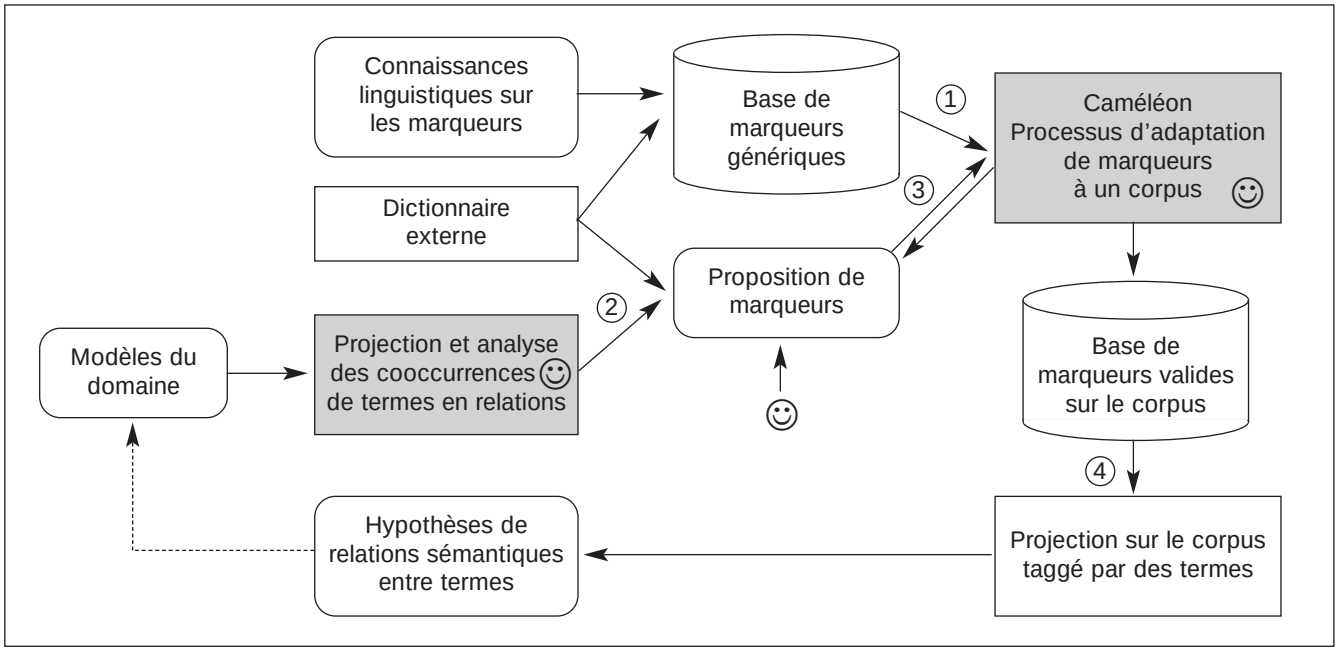


Figure 1.
Organisation de la méthode de génération de relations sémantiques par marqueurs.

3.1 Définition de la notion de *marqueur*

Notre système de gestion considère deux types de marqueurs : les marqueurs génériques et les marqueurs spécifiques.

Un marqueur générique est un patron lexico-syntaxique qui désigne une relation de façon plus ou moins stable à travers différents corpus et domaines techniques. Un marqueur spécifique est un patron lexico-syntaxique qui désigne une relation avec une certaine précision relativement à un corpus.

Au niveau informatique, un marqueur est caractérisé par un lien vers une relation et une forme A X B Y C, où {A,B,C} sont des patrons lexico-syntaxiques et {X,Y} des termes, futurs arguments de la relation désignée par le marqueur. Nous n'utilisons pas d'analyse syntaxique du corpus(3) et les catégories lexicales (lemmes et flexions) et syntaxiques sont déclarées à l'aide d'expressions régulières de types *Perl* (*Practical Extraction and Report Language*), directement dans le marqueur ou, pour les expressions plus complexes, dans un dictionnaire externe. À titre d'exemple, le marqueur exprimant l'action de décomposition d'un objet en fragments aux limites plus arbitraires défini par Jackiewicz (1996) se définit comme «X se VERBE_DECOMPOSITION (2*MOTS) en Y», où VERBE_DE_DECOMPOSITION = ((dé)?coup(e|ent|era|eront|ait|aient||erait|eraient)|partag(e|ent|era|eront|ait|aient||erait|eraient)|etc.) dans le dictionnaire externe et «(2*MOTS)» désigne le nombre de mots autorisés en insertion entre les deux patrons que sont «VERBE_DECOMPOSITION» et «en». Ce marqueur sera reconnu dans la phrase de l'exemple [1]:
[1] La grappe se décomposait radialement en n zones, les barreaux centraux (8) avaient un flux de N1, les barreaux périphériques avaient un flux de N2.

(3) L'utilisation d'une analyse syntaxique n'est pas indispensable dans les techniques de marqueurs, la plupart des marqueurs visant à retrouver des formes précises et non des fonctionnements syntaxiques.

3.2 La base de marqueurs génériques

La base de marqueurs génériques structure les marqueurs prévisibles et récurrents qui, pour une relation conceptuelle donnée, ont été jugés par des travaux de référence comme ceux de Borillo (1996), Jouis (1993) ou Jackiewicz (1996), ou par un cogniticien (lors de précédentes

applications) comme exprimant fréquemment et avec une certaine précision cette relation dans la langue dite générale.

Notre base de marqueurs génériques comprend deux types de relations lexicales: les relations d'hyponymie et de méronymie. Ces relations sont à la fois présentes dans beaucoup de domaines et structurelles, car, en tant que relations d'ordre, elles organisent les modèles en hiérarchies. Si la relation d'hyponymie est relativement bien définie en dehors de tout contexte (disons formellement), il n'en est pas de même pour la relation de méronymie que nous avons choisi de décomposer en un ensemble de sous-catégories d'après la taxinomie proposée par Winston (1987).

Dans la base générique, le lien entre le marqueur et la relation est un lien souple. Il peut être modifié lors de la phase d'adaptation de la base générique sur un corpus. Les marqueurs sont associés à un et un seul nœud de la hiérarchie de relation. Le marqueur présenté ci-dessus pouvant désigner plusieurs types de relations *a priori*, on choisira de l'associer à la relation de méronymie. Notre approche consiste à favoriser dans la base générique la déclaration de marqueurs les moins restreints possibles et se distingue donc des approches de Jouis (1993) et Garcia (1998) où la décomposition des relations conduit à une plus grande restriction des marqueurs.

À ce jour, la base générique de *Caméléon* possède 19 marqueurs d'hyponymie issus de Borillo (1996) et 69 marqueurs de type méronymie issus de Jackiewicz (1996) et Jouis (1993).

Nous proposons ici une méthode permettant de relativiser et restreindre les marqueurs de la base générique par rapport à un domaine de connaissance et une langue spécialisée.

3.3 Adaptation de la base de marqueurs génériques

Lors d'une nouvelle application, la base de marqueurs génériques est projetée sur le corpus. Afin de valider, spécifier ou restreindre nos marqueurs, nous avons défini une méthode de validation qui se décompose en 4 étapes:

• (1) Un sous-ensemble significatif d'occurrences du marqueur est présenté à l'utilisateur pour sa validation. Valider un marqueur signifie juger, en dehors des phénomènes rhétoriques, sa capacité à exprimer la relation sémantique entre deux termes qui lui est associée;

• (2) Suite à la validation du sous-ensemble d'occurrences, un taux de précision(4) est associé au marqueur;

• (3) Si le taux de précision est jugé valable, le marqueur est validé et associé à la base de marqueurs du corpus;

• (4) Si le taux de précision du marqueur est trop faible, il peut être restreint, c'est-à-dire rendu plus spécifique, par l'ajout de membres en {A,B,C}. Lorsque l'utilisateur contraint un marqueur, il est ramené en (1).

Caméléon assiste entièrement cette phase d'adaptation des marqueurs en permettant à un cogniticien, idéalement linguiste, d'observer les différents paramètres pendant le processus. *Caméléon* propose un mode «systématique», où tous les marqueurs sont validés par relation, ou un mode «ingénierie», où les marqueurs sont présentés par ordre croissant de productivité sur le corpus afin d'optimiser le temps de validation.

Une étape de l'évaluation du marqueur «X ADVERBE_DE_SPECIFICATION(5) Y» issu de Borillo (1996) est présentée sur la figure 3. La liste des 462 relations candidates trouvées sur ce corpus est présentée à l'utilisateur (partie droite de la fenêtre). L'utilisateur, pour un nombre paramétrable d'occurrences (ici 10), atteste ou nie le fait que le marqueur révèle la relation d'hyponymie entre les termes. Pour chaque relation à valider, nous proposons le contexte retrouvé par le marqueur et les termes repérés par *Caméléon* à gauche et à droite de ce marqueur.

Notons qu'au vu de la spécificité d'un corpus et des interactions entre les objets du domaine, un marqueur peut être associé à une relation plus précise ou une nouvelle relation, *Caméléon* autorisant l'utilisateur à

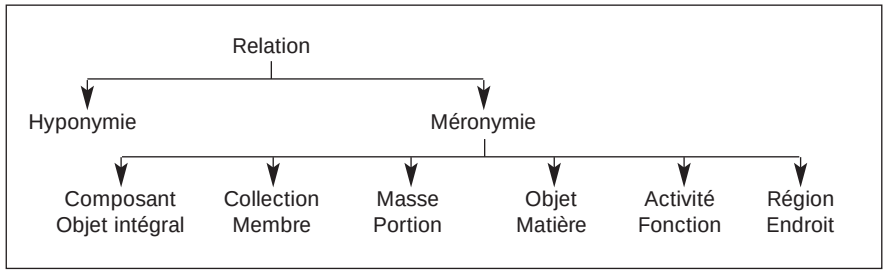


Figure 2.
Organisation des relations dans la base de marqueurs génériques.

(4) le taux de précision correspond au nombre d'occurrences déclarées valides sur le nombre total d'occurrences observées par le linguiste.

(5) où ADVERBE_DE_SPECIFICATION = (particulièrement|spécialement|en particulier|surtout|notamment|avant tout)

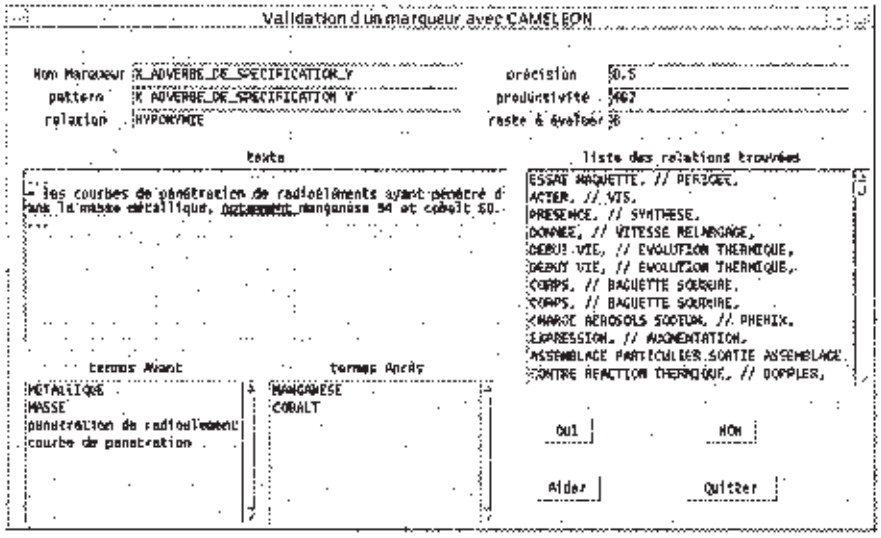


Figure 3.
Validation du marqueur «X ADVERBE_DE_SPECIFICATION Y» de la relation d'hyponymie avec *Caméléon*.

ajouter une relation propre au domaine dans la hiérarchie de départ présentée sur la figure 2.

À la fin de ce processus de validation, la base de marqueurs génériques est adaptée au corpus et est utilisée directement pour générer de nouvelles relations sémantiques entre termes.

3.4 Acquisition de marqueurs dédiés

Dans notre projet, nous disposons de modèles du domaine qui ont été construits selon le format et la démarche proposés par la méthode *Rex*. Ces modèles sont structurés à l'aide de trois types de relations conceptuelles: la relation «est-un», la relation «partie-de» et des relations spécifiques au domaine ou rôles (i.e. *le laboratoire X travaille sur le projet Y*). Nous appuyant sur la méthode d'acquisition de marqueurs proposée dans Hearst (1992) et exposée plus haut, nous générons automatiquement, pour un corpus et un modèle *Rex* associé, l'ensemble des cooccurrences de couples de termes qui désignent un couple de concepts dans le modèle. Cette projection est réalisée pour chacune des relations conceptuelles du modèle. Nous analysons ensuite systématiquement ces cooccurrences afin de trouver des candidats marqueurs qui seront validés avec *Caméléon*. La recherche de marqueurs se décompose en deux étapes, le repérage de marqueurs spécifiques désignant des relations de la base générique (hyponymie et méronymie) puis le repérage de marqueurs pour les rôles définis dans les modèles. Cette définition des nouveaux marqueurs ne nécessite aucune compétence informatique et est effectuée par un linguiste, directement dans l'environnement *Caméléon*. En effet, *Caméléon* met en place des structures indépendantes du domaine qui permettent de déclarer facilement des connaissances spécifiques au domaine.

Notons que, comme souligné plus haut, une phase de validation est toujours nécessaire avant l'introduction d'un nouveau marqueur. Une fois validé, ce nouveau marqueur est intégré à la base de marqueurs valides sur le corpus.

4 Évaluation de la méthode

Le corpus support de l'étude présentée dans la suite est composé de 2 millions de mots. Il contient 12 000 unités textuelles d'une page environ qui sont des interviews d'experts ou des extraits de documents de référence dans le domaine du nucléaire. Ces textes portent sur des travaux effectués ces dix dernières années. Ils ont une visée informative et traitent de sujets connexes; nous les définirons comme spécialisés (au sens de Condamines (1997)) et hétérogènes.

4.1 Projection de la base générique

Dans *Caméléon*, nous avons choisi de travailler à partir d'échantillons de 10 occurrences de marqueurs. Le processus d'affinage des marqueurs a été différent pour les relations d'hyponymie et de méronymie. Pour les premières, tous les marqueurs ont désigné la relation d'hyponymie avec une bonne précision (+ 75%). Aucun n'a été spécialisé. Si certains ce sont montrés très occurrents comme «Y ETRE ARTICLE_INDEFINI X» (1007 occurrences) ou «X ADVERBE_DE_SPECIFICATION Y» (462 occurrences), d'autres ne le sont que très faiblement ou pas du tout sur le corpus («Y ETRE LE X LE (plus|moins)», «X parmi (lequel|laquelle|lesquels|lesquelles) Y»). Ces marqueurs composés essentiellement de formes adverbiales, de structures d'énumération et de

construction autour du verbe être sont donc, d'après notre expérience, relativement fiables sur un nouveau corpus.

Les marqueurs de méronymie de la base générique sont, eux, construits sur des patrons qui portent essentiellement sur des syntagmes nominaux et des formes verbales (plus leurs formes dérivées). La validation de ces marqueurs nous a permis de relever que dans notre domaine technique, nombre des membres de ces marqueurs (groupe, assemblage, mélange) correspondent à des termes du domaine (groupe turbo-alternateur, assemblage combustible, mélange turbulent, etc.). Nous avons donc dû restreindre les marqueurs contenant ces formes verbales afin de ne pas générer trop de bruit. La plupart des marqueurs de méronymie ont été validés avec une assez faible précision (entre 50% et 75%). Ils ont été pour la plupart assez peu productifs (700 occurrences).

4.2 Acquisition de marqueurs à partir de modèles

Pour chaque couple de concepts en relation dans les modèles, *Caméléon* repère automatiquement les contextes de cooccurrences de termes les désignant. L'étude systématique de ces contextes nous a permis d'extraire de nouveaux candidats marqueurs.

4.2.1 Cas des marqueurs d'hyponymie

Pour cette relation, la projection de 60 couples de concepts a produit 180 contextes. L'étude de ces contextes nous a permis de retrouver les marqueurs les plus occurrents définis dans la base générique. 9 nouveaux candidats marqueurs nous ont semblé intéressant à tester dans *Caméléon*. Seulement 3 de ces marqueurs ont été validés; «Y ETRE ARTICLE X (dont|qui|pour)», «dans ARTICLE_DEFINI Y DE X» et «Y de certains X». Notons que les deux premiers sont des spécialisations de marqueurs déjà présents dans la base générique. Ils ont été validés avec une meilleure précision que le marqueur existant dont ils sont une restriction.

4.2.2 Cas de la relation de méronymie

300 contextes ont été analysés à partir de la projection de 180 couples de termes désignant des concepts reliés par cette relation. De très nombreux phénomènes de métonymie ont été observés dans ces contextes. Les candidats marqueurs de méronymie relevés sont étroitement liés à la spécificité des rapports entre les objets de notre domaine, comme l'implantation de certains types d'objets dans d'autres. Visant à généraliser ces candidats, nous avons été amené à chercher à partir d'un marqueur («X ETRE implantés? (2*MOTS) (dans|sur) Y») des équivalents avec d'autres formes grammaticales, que sont, par exemple, les formes verbales actives («IMPLANTER (dans|sur) X de Y»), les nominalisations alternatives au verbe (i.e. «implantation? DE Y (au centre de|sur|dans) X» et «implantations? (au centre de|dans|sur) X DE Y»), puis à étudier les contraintes qui pèsent sur les éléments du marqueur dans le but de le rendre le plus précis possible. Ce travail est cependant typiquement celui d'un linguiste. Nous nous sommes donc limité à généraliser nos marqueurs à partir des patrons de transformation proposés dans Jackiewicz (1996). 37 marqueurs ont ainsi été ajoutés à partir de 12 formes verbales et leurs transformations. Ils sont à la fois productifs et précis.

4.2.3 Cas des marqueurs de relations spécifiques aux modèles

Dans les modèles supports de l'expérience, les rôles ou relations spécifiques du domaine sont du type «X logiciels support de l'étude Y» ou «le projet X est étudié dans le laboratoire Y». Pour ces relations, l'étude des contextes des couples de termes nous a permis de trouver 6 marqueurs candidats comme: «calculs? DE Y (avec|à l'aide de|par) X» ou le marqueur typographique «X / Y» pour la relation structurante liant une structure (CEA) à une sous-structure (DAM) dans l'organigramme de l'entreprise dans CEA/DAM.

Notre méthode et nos outils se sont montrés très utiles pour la spécification de marqueurs de relations mal définies *a priori*, comme les relations de méronymie. En effet, si des travaux linguistiques nous permettent de connaître précisément les acceptions possibles de certaines formes, il n'est pas facile de préjuger quel sens sera utilisé dans un corpus donné, puis comment spécifier cette forme pour en faire un marqueur précis. Dans *Caméléon*, nous avons envisagé la phase de définition d'un nouveau marqueur comme une phase de spécification descendante: nous déclarons un marqueur peu restreint puis le spécialisons de façon empirique, à partir de preuves émergeant du corpus.

Nous aurions pu assister le processus d'extraction de candidats marqueurs à l'aide d'outils de mise en évidence de similarités dans les contextes de cooccurrences de concepts (Rousselot (1996) Morin (1998)). Cependant, le faible nombre d'occurrences (souvent une seule) de candidats marqueurs dans les contextes des termes aurait rendu difficile l'utilisation de telles méthodes statistiques sur ce corpus avec ces modèles.

La validation systématique de la base de marqueurs génériques a demandé 5 heures de travail à l'auteur. L'étude des contextes de cooccurrences est un processus relativement long qui a demandé un jour de travail à l'auteur. Soulignons que si des connaissances linguistiques théoriques permettent d'accélérer le processus de mise au point des marqueurs, nous pensons qu'elles ne se substituent pas à la validation des candidats sur corpus.

5 Extraction de relations et construction de modèles du domaine

Afin de générer des relations candidates entre termes à partir des marqueurs valides, nous avons tout d'abord indexé notre corpus par les candidats termes extraits par *Nomino*(6) et *Ana* (Enguehard 1993). *Caméléon* a généré automatiquement environ deux mille relations candidates entre termes. La souplesse de notre méthode visant à ne pas restreindre *a priori* les marqueurs, mais à les spécialiser à partir de données attestées, nous permet de trouver un nombre important de relations. Ces relations constituent un point de départ privilégié pour la construction et l'enrichissement de modèles. L'intégration de nouvelles relations dans un modèle du domaine n'est cependant pas un processus direct.

En effet, certaines relations candidates ne représentent pas la relation exprimée dans le texte. Des phénomènes linguistiques, comme l'anaphore, nous font trouver automatiquement des termes en position X ou Y qui ne sont pas ceux qui sont réellement en relation. De plus, si la validation d'une relation est un processus complexe, son intégration dans un modèle l'est tout autant. Intégrer une nouvelle relation demande d'interpréter les modèles existants et de juger de la possibilité et l'intérêt de l'ajout d'une nouvelle relation. Certaines relations entre termes sont trop «éloignées»; ainsi, pour la relation d'hyponymie, l'hypéronyme est parfois un terme trop général par rapport à l'hyponyme et la relation ne peut être intégrée sous cette forme dans les modèles. Ce que l'on choisira de modéliser dépend des connaissances déjà exprimées dans nos modèles et d'un point de vue, celui de l'utilisation des modèles dans la future application. Ainsi, même si une relation est valide, on peut choisir de ne pas la modéliser.

Une méthode et des outils pour assister l'expert dans ce processus de modélisation à partir de relation candidates sont proposés dans Séguéla (1999).

6 Conclusions et perspectives

Notre méthode assistée par *Caméléon* permet de gérer, enrichir et spécialiser une base de marqueurs linguistiques relativement à de nouveaux corpus spécialisés. Notre outil ouvert et souple permet d'optimiser l'utilisation de connaissances linguistiques générales sur des corpus spécialisés. Il vise, à la fois, à assister le processus de spécialisation de ces connaissances générales et à assister l'acquisition de connaissances spécifiques aux corpus. Thésaurisant les différentes actions de l'utilisateur, *Caméléon* est un outil très adapté à l'étude des différents phénomènes linguistiques liés aux marqueurs.

Nous nous sommes attaché à associer aux marqueurs une précision afin d'obtenir des relations candidates fiables en amont du processus de modélisation. Nous pensons que cette démarche garantit une certaine qualité vis-à-vis de la méthode de projection de marqueurs pour l'extraction de relations sémantiques en vue de la construction de modèles.

Cependant, lors de la modélisation et afin d'éviter l'absence de relations candidates lors de l'étude de certains concepts, nous pensons que ce type de méthodes fines et précises doit être complémentaire d'une méthode d'extraction de relations moins précise, comme l'étude de la distribution des termes dans les textes ou la décomposition de syntagmes complexes.

Patrick Séguéla,
Laboratoire d'aide à l'exploitation des réacteurs,
Direction des réacteurs nucléaires,
St-Paul-Lez-Durance
et
Centre compétence objet,
Euriware,
Saint-Quentin-en-Yvelynes.

(6) http://www.ling.uqam.ca/nomino

Bibliographie

Assadi (H.), 1998: *Construction d'ontologies à partir de textes techniques, Application aux systèmes documentaires.* Thèse de doctorat, Université de Paris 6.

Borillo (A.), 1996: «Exploration automatisée de textes de spécialité: repérage et identification de la relation lexicale d'hypéronymie», dans *Linx*, n° 34/35, p. 113-124.

Condamines (A.), 1997: «Langue spécialisée ou discours spécialisé?», dans *Mélanges offerts à Kocourek*, Université de Dalhousie, Alfa, p. 171-185.

Condamines (A.), Rébeyrolle (J.), 1998: «CTKB: A Corpus-based Approach for Terminological Knowledge Base», dans *Proceedings of the first workshop on computational terminology (Computerm'98), Workshop of Coling'98, Montréal, august 15 1998.*

Davidson (L.), Kavanagh (J.), Mackintosh (K.), Meyer (I.), Skuce (D.) 1998 : « Semi-automatic extraction of knowledge-rich contexts from corpora : examples and issues », dans *Proceedings of the first workshop on computational terminology (Computerm'98), Workshop of Coling'98. Montréal, august 15, 1998*.

Enguehard (C.), 1992 : *Ana, Apprentissage Naturel Automatique d'un réseau sémantique,* Thèse de doctorat, Université de technologie de Compiègne.

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

Grefenstette (G.), 1994 : *Explorations in automatic thesaurus discovery*, Boston, Kluwer Academic Publishers.

Habert (B.), Nazarenko (A.), 1996 : « La syntaxe comme marche-pied de l'acquisition des connaissances : Bilan critique d'une expérience », dans *Actes des septièmes Journées Acquisition des Connaissances, Sète, 8-10 mai, 1996*, p. 137-148.

Hearst (M.) 1992 : « Automatic Acquisition of Hyponyms from Large Text Corpora », dans *Proceedings of the international conference on computational linguistics (Coling 92), Nantes July 25-28, 1992*, p. 539-545.

Jackiewicz (A.), 1996 : « L'expression lexicale de la relation d'ingrédience (partie-tout) », dans *Faits de Langues*, n°7, Paris, Orphys, p. 53-62.

Jouis (C.), 1993 : *Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes. Réalisation d'un prototype : le système Seek*, Thèse de doctorat, École des hautes études en sciences sociales de Paris.

Jouis (C.), Biskri (I.), Desclés (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 », dans *Actes de la première journée scientifique et technique du réseau francophone de l'ingénierie de langue de l'Aupelf-Uref, Avignon, 15-16 avril*, p. 427-432.

Lerat (P.), 1995 : *Les langues spécialisées*, Paris, Puf.

Morin (E.) 1998 : « Prométhée : un outil d'aide à l'acquisition de relations sémantiques entre termes », dans *Actes de la cinquième conférence du traitement automatique du langage naturel (TALN-98), Paris, 10-12 juin 1998*, p. 172-181.

Riloff (E.), 1996 : « Automatically Generating Extraction Patterns from Untagged Text », dans *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96), Portland, 4-8 August*, p. 1044-1049.

Rousselot (F.), Frath (P.), Oueslati (R.), 1996 : « Extracting Concepts and Relations from Corpora », dans *Proceedings of ECAI Workshop on Corpus-Oriented semantic analysis, Budapest, August 12, 1996.*

Séguéla (P.), 1999 : « Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine », dans *Actes de la conférence Ingénierie des Connaissances*, Palaiseau, (à paraître).

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

Winston (M.), Chaffin (R.), Herrmann (D.), 1987 : « A Taxonomy of Part-Whole Relations », dans *Cognitive Science* n°11, p. 417-441.

Détection de liens de synonymie : complémentarité des ressources générales et spécialisées

Dans le cadre d'une aide à la structuration de terminologies, l'utilisation de données sémantiques générales nous a conduits à proposer des règles d'inférence de liens de synonymie entre des candidats termes complexes. Au regard de l'évaluation faite par un expert du domaine en contexte applicatif, il s'avère que nous obtenons des résultats intéressants. Ces premiers résultats montrent que, contrairement à une opinion assez largement admise, l'utilisation de ressources générales comme un dictionnaire de langue pour le traitement de documents techniques se justifie. Nous cherchons ici à caractériser plus précisément l'apport de ces ressources. Nous avons confronté les résultats obtenus aux liens inférés à partir de ressources lexicales plus spécialisées. Il s'avère que peu de liens inférés sont communs d'une ressource à l'autre. Ceci souligne la complémentarité des différentes sources et l'intérêt spécifique des informations de la langue générale pour la structuration de terminologies.

Termes-clés : structuration de terminologie ; variation sémantique ; synonymie ; ressources lexicales ; langue spécialisée *vs.* langue générale.

1 Introduction

Le travail présenté ici s'inscrit dans un projet de développement d'outils d'aide à la structuration et à la mise à jour de terminologies. Il résulte d'une collaboration entre le *LIPN* et la Direction des études et recherche d'électricité de France (DER-EDF). La terminologie constituée est ensuite utilisée dans un système de consultation de documents techniques (*SCDT*). Notre objectif est de fournir des liens sémantiques, de synonymie notamment, entre termes extraits d'un corpus technique. Ces liens doivent faciliter la navigation dans les documents techniques. Nous présentons et confrontons les résultats de l'utilisation de plusieurs types de ressources sémantiques pour inférer des liens de synonymie : un dictionnaire de langue, des classes de synonymes construites manuellement et le thesaurus EDF.

Ce travail s'inscrit dans le débat concernant le statut des langues de spécialité et l'apport des ressources lexicales comme un dictionnaire de langue pour le traitement des documents techniques. On constate un fort contraste dans la description des unités lexicales figurant dans les dictionnaires d'usage et leurs emplois dans des langues de spécialité : les mots des documents techniques ne sont pas toujours répertoriés dans les dictionnaires et quand ils le sont, c'est souvent avec des sens plus variés et différents. L'hypothèse couramment admise est donc que ce type de ressources n'est pas exploitable pour le traitement de documents spécialisés pour lesquels on s'attache à construire des ressources spécifiques.

Les premiers résultats que nous avons obtenus soulignent cependant l'apport d'un dictionnaire de langue pour l'aide à la structuration de terminologie. Dans cette expérience préliminaire (Hamon *et al.* 1998), le dictionnaire *Le Robert* a été utilisé pour inférer des liens de synonymie entre termes. La validation des résultats par un expert du domaine concerné montre que 37% des liens inférés expriment effectivement une relation de synonymie (*action de protection* / *action de sauvegarde*) et plus largement que la moitié des liens est utile pour la structuration de terminologie (*rapport de sûreté* / *analyse de sûreté*, que l'expert analyse comme un lien de méronymie).

Sur la base de ces premiers résultats, nous cherchons ici à caractériser plus précisément l'apport de ce type de ressources lexicales en les confrontant à des données plus spécialisées dans la perspective du développement d'outils d'aide à la navigation dans les documents techniques. Dans ce qui suit, nous opposons donc des ressources « générales » comme un dictionnaire d'usage à des ressources spécialisées. En pratique ces ressources diffèrent surtout par l'usage plus ou moins spécialisé qui en est fait.

1.1 Utilisation de liens sémantiques dans un SCDT

De nombreuses applications dans les domaines de spécialité nécessitent l'utilisation de terminologies: indexation contrôlée, aide à la rédaction, consultation de documents, etc. Nous nous intéressons ici à ce dernier type d'applications. La taille croissante de la documentation technique conduit en effet les entreprises à développer des outils de navigation dans leurs documents.

Le système de consultation de documentation technique développé par EDF (Gros *et al.* 1996), (Gros *et al.* 1997) fournit un accès hypertexte au document suivant différents modes:

• Une table des matières;
• Un accès plein texte par mots-clés, étendu par la consultation d'une terminologie;
• Un index du domaine intégrant des liens de synonymie et d'hyponymie entre une entrée et une sous-entrée;
• Un index de l'activité modélisant la tâche de l'utilisateur.

Le processus d'aide à la construction et à la structuration de terminologies doit faciliter l'intégration de nouveaux documents dans un système de consultation de documents techniques en fournissant des liens sémantiques entre les termes extraits du document. Les liens de synonymie sont introduits dans le système pour enrichir l'index et la terminologie proposés aux utilisateurs.

1.2 Structuration d'une terminologie

Le processus de constitution d'une terminologie se divise en deux grandes phases (Dagan *et al.* 1994). Dans un premier temps, les termes candidats sont extraits d'un corpus pertinent pour le domaine étudié. L'ajout de liens sémantiques permet ensuite d'obtenir un réseau terminologique candidat, plus complexe.

Le système d'EDF repose sur cette démarche. L'extraction des candidats termes est assurée par le logiciel d'extraction de terminologie *Lexter* (Bourigault 1994). Les candidats termes sont organisés en un réseau syntaxique. Les liens sémantiques sont ajoutés dans le réseau syntaxique sous la forme de classes conceptuelles (Assadi 1997) ou de liens de causalité (Garcia 1998). Notre travail porte sur cette deuxième étape: il vise à enrichir le réseau initial de liens de synonymie entre candidats termes.

Notre méthode repose sur l'utilisation de ressources lexicales telles qu'un dictionnaire de langue. Les bases de connaissances lexicales en langue de spécialité étant rarement disponibles, nous avons évalué la pertinence et l'utilité des informations sémantiques générales dans les documents techniques (Hamon *et al.* 1998). À partir du lien de synonymie (*commande / ordre*) un lien de synonymie est inféré entre les candidats termes *commande manuelle* et *ordre manuel*.

Les résultats obtenus montrent la pertinence des informations sémantiques générales et apportent une réponse expérimentale au débat sur le rôle de ces connaissances dans le traitement des textes de spécialité. Cependant, bien que les résultats soient intéressants du point de vue de l'expert, il est nécessaire de les combiner à des données spécialisées. Nous avons donc cherché à caractériser la contribution respective des différents types de données lexicales. Nous avons confronté le dictionnaire de langue avec deux ressources spécialisées disponibles au sein de la DER-EDF et *a priori* pertinentes pour l'un de nos corpus d'expérimentation.

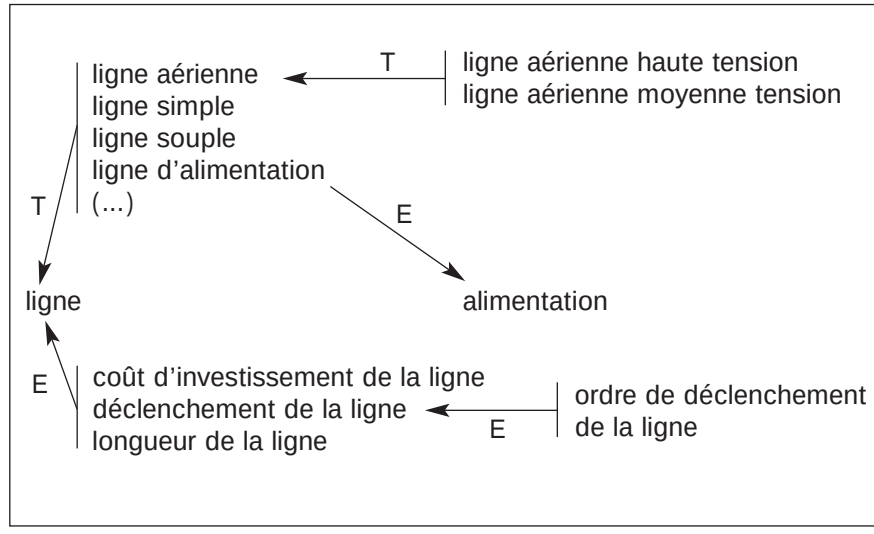


Figure 1:
Fragment du réseau syntaxique Lexter (T = tête, E = expansion).

1.3 Présentation du corpus de travail

Le corpus des dossiers de système élémentaire (*DSE*) comporte environ 160 000 mots et décrit en partie le fonctionnement des centrales nucléaires.

Le corpus est analysé par *Lexter* (Bourigault 1994) qui en extrait 17 675 candidats termes (2 865 noms, 1 306 adjectifs et 13 504 groupes nominaux), structurés en réseau syntaxique (*cf.* figure 1). Chaque candidat terme complexe (par ex. *ligne d'alimentation*) est décomposé en une tête (*ligne*) et une expansion (*alimentation*).

2 Inférence de liens de synonymie

2.1 Principe général

Notre définition de la synonymie est proche de celle proposée dans *WordNet* (Miller *et al.* 1993). Alors que la synonymie peut être vue comme une relation graduée, nous la considérons comme une relation d'équivalence contextuelle. Ainsi, à l'instar de la synonymie cognitive de (Cruse 1986), nous définissons une relation de synonymie cognitive contextuelle entre deux termes X et Y dans un contexte C si les deux termes sont syntaxiquement identiques et substituables – *salve veritate* – dans le contexte C.

L'inférence d'un lien de synonymie entre termes complexes repose sur l'hypothèse que la compositionnalité des termes complexes préserve la synonymie. Cette hypothèse est évidemment simplificatrice : ce n'est pas parce que *Le Robert* donne *arrêt* et *interruption* comme synonymes dans certaines acceptions que le terme complexe *arrêt du réacteur* doit s'entendre au sens de *interruption du réacteur*. En pratique, l'inférence d'un lien de synonymie suppose que deux termes complexes comportant des éléments synonymes et construits selon le même schéma syntaxique soient attestés en corpus. Nous considérons que deux termes sont synonymes si leurs composants sont identiques ou synonymes. Dans l'exemple ci-dessus, *interruption du réacteur* n'étant pas attesté dans le corpus, aucun lien de synonymie n'est inféré. En revanche dès lors que *arrêt de l'appoint* (ellipse pour *arrêt de l'appoint en acide borique*) et *interruption de l'appoint* sont tous deux des termes attestés, nous faisons l'hypothèse qu'ils sont synonymes.

Cette démarche se rapproche en fait de celle de Basili *et al.* (1997) dans le sens où elle exploite des données de la langue générale pour des corpus spécialisés lorsque ces données sont corroborées en corpus par l'existence de constructions parallèles (dans notre cas) ou par la similarité des contextes d'apparition (pour R. Basili et ses collègues).

2.2 Détection des candidats termes synonymes

La méthode générale d'inférence des liens de synonymie est présentée dans Hamon *et al.* (1998). Elle se décompose en deux étapes.

La première est une étape de filtrage qui réduit la taille des données utilisées lors de l'application des règles d'inférence. Un lien entre deux termes est conservé si ceux-ci sont tous les deux présents dans le document étudié. Par exemple, le lien (*portion / tronçon*) est retenu si ces lemmes figurent sous une forme quelconque dans le corpus.

La deuxième étape est le processus inférentiel à proprement parler. Nous avons conçu trois règles pour inférer des liens de synonymie entre candidats termes complexes. Un lien de synonymie est ajouté entre deux candidats termes du réseau syntaxique si l'une des trois conditions suivantes est vérifiée :

- Règle 1 : les têtes sont identiques et les expansions sont synonymes (*action de protection / action de sauvegarde*) ;
- Règle 2 : les têtes sont synonymes et les expansions sont identiques (*capacité faible / puissance faible*) ;
- Règle 3 : les têtes sont synonymes et les expansions sont synonymes (*classement d'équipement / classification de matériel*).

Nous contraignons les composants des termes à posséder la même catégorie syntaxique. Par ailleurs, nous avons choisi de ne pas tenir compte des prépositions et des formes fléchies des termes. Ce parti pris réduit le coût du calcul des liens sémantiques et a peu d'incidence sur les résultats.

Les liens initiaux sont d'abord utilisés pour amorcer la détection des candidats termes complexes. Puis, tant que de nouveaux liens sont trouvés, nous réitérons le processus en prenant en compte les liens précédemment détectés.

La méthode a été testée sur des corpus de taille différente : *Menelas* (85 000 mots), *DSE* (160 000 mots), *Crater* (750 000 mots). Un algorithme efficace permet l'application à des corpus techniques importants. Une interface de validation est en cours de réalisation.

2.3 Protocole de validation

Un expert du domaine a validé les résultats. Les liens inférés ont été acceptés ou rejetés en tenant compte du contexte d'application de ces résultats : un système de consultation de documents techniques. De plus, le statut terminologique des candidats termes liés a été pris en compte contrairement à la première expérience (Hamon *et al.* 1998). Ainsi, lors de la validation des résultats, les liens dont l'un des

candidats termes liés ne pouvait avoir le statut de terme, ont été rejetés.

Nous avons choisi de présenter les candidats termes liés sous la forme fléchie de leur première occurrence rencontrée dans le corpus. Lors de la validation, l'expert a pu accéder aux groupes nominaux maximaux ainsi qu'aux phrases dans lesquels se trouvent les candidats termes liés.

Toutefois, bien que notre objectif soit la détection de liens de synonymie entre des candidats termes complexes, nous avons constaté que les liens inférés peuvent être typés différemment (Hamon *et al.* 1998). Nous laissons donc à l'expert la possibilité de modifier le type du lien. En effet, bien qu'il ne s'agisse pas de liens de synonymie, il est intéressant, dans le cadre d'une aide à la structuration de terminologie, de conserver tous les liens sémantiquement pertinents. L'évaluation des résultats tient compte de cette caractéristique.

Afin d'assurer une certaine cohérence, les liens inférés ont été structurés et présentés suivant deux modes de regroupement :

• Structuration par famille : les liens sont regroupés en fonction du lien initial utilisé par les règles d'inférence. Par exemple, les liens (*débit maximum / volume limite*, *débit nécessaire / volume requis*, *débit d'acide borique / volume d'acide borique*, *débit d'eau / volume d'eau*, *débit total / volume total*) sont présentés ensemble puisqu'ils sont inférés à partir du même lien initial (*débit / volume*) ;

• Structuration par classe : nous avons regroupé les liens qui constituent les chemins de longueur *n* entre deux termes pour les proposer ensemble à la validation. Ainsi, par exemple, les liens *gamme logarithmique / échelle logarithmique*, *échelle logarithmique / mesure logarithmique* sont regroupés dans la même classe.

L'expert a essentiellement utilisé cette deuxième présentation des résultats pour valider les liens, les candidats termes liés apparaissant dans le graphe avec leur voisinage. Ce type de présentation propose une vue globale sur un ensemble de liens et favorise la cohérence de la validation.

3 Utilisation de différentes ressources lexicales

Cette partie présente les résultats de l'application de l'inférence de liens de synonymie sur le corpus de travail à partir de trois sources lexicales différentes. La confrontation de ces différents ensembles de résultats permet de mieux évaluer l'apport respectif de chaque source lexicale.

3.1 Exploitation d'un dictionnaire de langue

Nous avons utilisé pour cette étude les informations sémantiques du dictionnaire *Le Robert* fournis par l'Inalf. Même s'il comporte des indications de synonymie, ce n'est pas à proprement parler un dictionnaire de synonymes. Cependant, ce dictionnaire, largement disponible, est reconnu comme un standard. Il permet d'effectuer des expériences dans des conditions réelles. Des listes de liens de synonymie ont été extraites. Une entrée peut comporter différentes listes de synonymes correspondant à chacun de ses sens mais les synonymes eux-mêmes ne sont pas désambiguïsés. Les indications de sens n'étant ni explicites ni homogènes, nous les avons négligées (Ploux *et al.* 1998). De plus, les liens de synonymie extraits du dictionnaire pouvant être très contextuels ou exprimer, par exemple, des relations d'analogie, l'application de la propriété de transitivité provoque de nombreuses erreurs : nous n'exploitons pas cette propriété de la synonymie.

Le dictionnaire couvre 40% des candidats termes extraits du corpus des DSE. À partir des données du dictionnaire, nous avons cherché à inférer des liens de synonymie sur ce corpus. L'étape de filtrage conserve 3 129 liens exprimant principalement des relations de synonymie entre mots simples. Ces liens permettent ensuite d'inférer 590 liens sur les candidats termes complexes. L'expert a jugé que 199 liens inférés (33,7 %) sont pertinents : *air de l'enceinte / atmosphère de l'enceinte*, *changement de gamme / modification d'échelle*, *fiche de correction / fiche de modification*. Parmi ces liens, 101 liens ont été retenus comme exprimant des relations de synonymie (*débit nul de refroidissement / débit nul de réfrigération*, *fluide actif / liquide radioactif*, *prescription de sûreté / règle de sûreté*) et 84 comme des liens de type Voir-Aussi (*liaison d'alimentation / ligne d'alimentation*). Les autres liens sont typés comme des relations d'hyperonymie ou de méronymie (*phénomènes naturels / phénomènes physiques*). Une partie des erreurs est due au fait que des candidats termes liés ne peuvent avoir le statut de termes. Nous estimons que le taux de précision est proche de celui de la première expérience (37 % de liens de synonymie valides, 50 % de liens sémantiquement valides) si le statut terminologique de candidats termes n'est pas pris en compte.

3.2 Amorçage à l'aide de classes de synonymes construites manuellement

Des classes de synonymes ont été constituées manuellement par un expert pour un corpus du même domaine que celui de notre corpus de travail. Nous avons à notre disposition un ensemble de 500 classes constituées de 1 335 termes. Considérées comme des classes d'équivalence, ces classes fournissent

3 456 liens de synonymie. Ces liens reposent sur des relations morpho-syntaxiques (*appoint en acide borique / appoint en bore*) ou des relations sémantiques (*eau de refroidissement / fluide réfrigérant*). Il s'agit de liens entre des termes complexes mais aussi entre termes simples et termes complexes (*appareil de mesure / capteur*). De telles ressources spécialisées sont précieuses mais rares. Elles sont coûteuses à construire et demandent à être mises à jour régulièrement.

Lors de l'étape de filtrage 281 liens sont retenus. Les règles permettent d'inférer 167 liens sémantiques entre des termes complexes. Lors de la validation, 143 liens sont jugés pertinents (85,6 %): *concentration en bore du circuit primaire / teneur en bore du circuit primaire*, *Procédure contrôle / procédure d'essais*. Parmi ces liens, 106 liens sont des relations de synonymie (*circuit d'alimentation / réseau d'alimentation*, *bilan de fuite global du circuit primaire / bilan de fuite global du primaire*) et 23 liens de type Voir-Aussi (*état normal d'exploitation / conditions normales d'exploitation*)

Nous avons également appliqué les règles d'inférence en utilisant les liens extraits des classes de synonymes à la suite de liens du dictionnaire. On constate que 41 liens peuvent être inférés par l'une ou l'autre des deux ressources indifféremment. L'utilisation conjointe de ces deux ressources lexicales permet d'inférer 40 nouveaux liens supplémentaires. Il s'agit de liens obtenus par la règle 3. Si un lien (ex. *tronçon du circuit de réfrigération intermédiaire / portion du circuit RRI*) ne peut être inféré qu'à partir de deux liens initiaux issus de deux sources différentes (*tronçon / portion*, fourni par le dictionnaire et *circuit de réfrigération intermédiaire / circuit RRI*, fourni par les classes de synonymes), seule l'utilisation conjointe de ces deux ressources permet de l'inférer. L'expert a validé 9 des 40 liens supplémentaires, 3 liens exprimant des relations de synonymie (*liaison d'alimentation / ligne de distribution*) et 2 des relations Voir-Aussi (*vidange du réservoir / purge des bâches*). Le faible nombre de liens pertinents est dû à l'application de la règle 3. En effet, comme nous l'avons déjà constaté (Hamon *et al.* 1998), cette règle ne permet pas de proposer beaucoup de liens pertinents. Cette règle est cependant précieuse puisqu'elle permet d'inférer des liens difficiles à trouver manuellement par les terminologues.

3.3 Inférence de liens à partir d'un thesaurus

Le thesaurus EDF (EDFDOC) contient 20 000 termes simples ou complexes organisés entre 330 champs sémantiques (Circuit électrique, Technologie des câbles, Organisation administrative), eux-mêmes regroupés en 45 points de vue (classes très générales). Trois types de liens sont proposés: les liens associatifs (Voir-Aussi: *source autonome / alimentation de secours*), les liens hiérarchiques (Hyponymie: *sécurité des personnes / protection de l'opérateur*) et les liens de synonymie (Est employé pour: *panier filtrant / tamis*). Nous avons utilisé tous les liens sémantiques du thesaurus, soit 25 000 liens.

Bien que la plupart des liens du thesaurus expriment des relations sémantiques ou morpho-syntaxiques entre termes complexes, nous avons cherché à saturer le réseau terminologique candidat de la même manière que pour les classes de synonymes. Ainsi, 389 liens sont conservés lors du filtrage et 55 liens sont inférés. Ce faible résultat tient à la complexité des termes présents dans le thesaurus. Les liens inférés portant sur des termes plus complexes, peu de liens sont détectés. Lors de la validation, 36 liens sur 55 (65,4 %) sont retenus par l'expert (*capteurs de pression / mesures de pression*, *signalisations lumineuses locales / voyants locaux*, *arrêt de l'appoint / interruption de l'appoint*). Les liens de synonymie ne représentent qu'une faible partie des liens validés (4 / 36) alors que 15 liens sont typés comme des liens Voir-Aussi (*indicateurs locaux / mesure locale*, *capteurs de niveau / régulation de niveau*).

4 Caractérisation de l'apport spécifique de chaque source lexicale

La table 1 présente les résultats obtenus à partir des trois sources. La proportion de liens inférés par rapport au nombre de liens retenus lors de l'étape de filtrage varie suivant le type de source utilisé. Cette proportion est de 3/5 pour les classes de synonymes et seulement de 1/5 pour le dictionnaire dont la productivité est donc faible. Néanmoins, du fait de la taille de ce dernier, les liens inférés à partir du dictionnaire représentent 78% du total des liens inférés tandis que seulement 22% des liens sont inférés à partir des classes de synonymes. Les informations sémantiques extraites du dictionnaire permettent d'inférer beaucoup de liens à un faible coût alors que la constitution des classes de synonymes est très coûteuse.

Les résultats présentés ci-dessus font apparaître des différences numériques significatives dans l'apport des différentes sources lexicales utilisées. Les sections qui suivent comparent les résultats obtenus par inférence avec les données présentes dans les classes de synonymes et le thesaurus qui contiennent eux-mêmes des termes complexes. Nous cherchons ainsi à mieux caractériser l'apport spécifique

	Filtrage des liens		Inférence des liens sur le corpus	Validation
	Nombre de liens conservés	Nombre de termes	Nombre de liens inférés	Nombre de liens
Dictionnaire	3 129	1 299	590	199 (33,7 %)
Classes de synonymes	281	344	167	143 (85,6 %)
Thesaurus EDF	389	478	55	36 (65,4 %)
Dictionnaire puis classes	3 376	1 547	756	315 (41,6 %)

Tableau 1 :
Résultats de la méthode d'inférence

de chaque source lexicale et leur complémentarité.

4.1 Complémentarité des classes de synonymes construites manuellement et du dictionnaire

La confrontation des liens inférés à partir du dictionnaire de langue avec les liens fournis par les classes de synonymes ou inférés à partir de ceux-ci est riche en enseignements. Le tableau 2 synthétise ces résultats.

On remarque tout d'abord que très peu des liens établis manuellement par l'expert (19 liens) peuvent être trouvés par inférence. Dans la mesure où les classes de synonymes contiennent beaucoup de liens entre termes complexes, on pourrait s'attendre à ce que les règles d'inférence, appliquées sur des classes de synonymes, retrouvent des liens déjà présents dans ces mêmes classes. Ainsi le lien de synonymie (*condition de fonctionnement / régime de fonctionnement*) qui est donné dans les classes est également inféré à partir du lien (*condition / régime*) lui aussi établi par l'expert. Le nombre de ces liens est cependant faible (19 liens) par rapport au nombre de nouveaux liens inférés à partir des classes de synonymes (148 liens). Il s'avère que l'expert, lors de la construction des classes, n'a pas eu le souci d'en faire la clôture inférentielle : il privilégie la cohérence intrinsèque des classes. Ceci souligne la complémentarité des deux démarches, humaine et algorithmique, pour la détection des liens de synonymie. L'expert a validé 17 des 19 liens inférés également présents dans les classes sémantiques : *conditions de fonctionnement / régimes de fonctionnement*, *classe sismique / classification sismique*, *limitation de durée / limitation de temps*. Le typage de ces liens se répartit équitablement entre la relation de synonymie et la relation Voir-Aussi.

	Nombre de liens inférés également construits par l'expert	Nombre de liens inférés non construits par l'expert
Classes de synonymes	19	148
Dictionnaire	18	572
Dictionnaire puis classes	32	724

Tableau 2 :
Proportion de liens construits par l'expert parmi les liens inférés

On observe par ailleurs que peu de liens inférés à partir du dictionnaire (18 sur un total de 590) sont donnés par l'expert dans les classes de synonymes. Il y a donc beaucoup de liens qui sont inférés à partir du dictionnaire que l'expert valide effectivement comme liens de synonymie quand on les lui soumet mais qu'il n'a pas pensé à inclure dans ses classes de synonymes. Ceci s'explique par le fait que l'expert travaille de manière privilégiée sur la langue technique. De son propre aveu, rechercher les liens de langue générale est un surcroît de travail. Ceci souligne l'intérêt spécifique des ressources générales. Parmi les liens proposés par l'expert dans les classes sémantiques, tous les liens sont validés : *contrôles périodiques / inspections périodiques*, *baisse de pression / réduction de pression*. Cependant, peu de liens expriment des relations de synonymie (3 liens). Il s'agit essentiellement de liens de type Voir-Aussi (15 liens). La proportion de ces liens inférés à partir du dictionnaire et présents dans les classes est également faible au regard du nombre total de liens dans les classes.

Le recouvrement entre les deux sources lexicales est donc faible : les données spécialisées fournies par l'expert et les données issues du dictionnaire de langue apparaissent largement complémentaires.

4.2 Le thesaurus, un apport plus marginal

De la même manière, le recouvrement entre les liens inférés à partir du dictionnaire ou des classes et les liens donnés par le thesaurus est

	Nombre de liens inférés déjà présents dans le thesaurus	Nombre de liens inférés non présents dans le thesaurus
Dictionnaire	2	588
Classes de synonymes	1	166
Dictionnaire puis classes	2	754

Tableau 3:
Proportion de liens du thesaurus parmi les liens inférés

très faible. Ne figurent dans le thesaurus que deux liens inférés à partir des liens extraits du dictionnaire et un seul lien inféré à partir des liens construits par l'expert (voir le tableau 3): *circuit de réfrigération / circuit de refroidissement* et *appareil de mesure / dispositif de mesure*. Le premier lien est validé comme un lien de synonymie alors que le second est retenu comme un lien de type Voir-Aussi. Dans le cas du thesaurus, ce faible recouvrement indique en fait un intérêt très marginal pour la détection de liens de synonymie entre termes.

Nous pouvons avancer deux types d'explications qui demandent à être confirmées par un expert du domaine: la couverture et la normalisation du thesaurus. Le thesaurus semble ne couvrir que très partiellement le domaine du corpus. Nous ne retrouvons que 28 liens du thesaurus parmi les classes de synonymes, qui ont été construites à partir d'un corpus du domaine. La couverture est d'autant plus faible que les termes présents dans le thesaurus sont en grande partie complexes. De surcroît, la faible productivité du thesaurus pour la détection de liens de synonymie montre que les liens initiaux ne reflètent souvent pas des liens terminologiques du domaine. Les termes liés dans le thesaurus n'entrent pas dans les constructions parallèles recherchées par les règles d'inférence.

Les liens du thesaurus ont été construits afin d'organiser conceptuellement les termes en fonction de différents domaines d'activité. Il est possible que les objectifs de normalisation sous-jacents dans le thesaurus rendent difficile son utilisation pour le traitement de corpus. Si les candidats termes extraits automatiquement sont en fait des variantes des formes normalisées du thesaurus, il faut mettre en œuvre des techniques plus lourdes et plus complexes telles que la génération de variantes (Jacquemin 1997) pour inférer des liens de synonymie à partir du thesaurus.

À maints égards, le thesaurus semble avoir un statut intermédiaire entre le dictionnaire de langue et la source spécialisée que sont les classes de synonymes. Malheureusement, dans la visée qui est la nôtre, il combine les faiblesses de l'un et de l'autre. Le thesaurus, après l'étape du filtrage, ne fournit qu'un nombre réduit de liens initiaux (comme les classes de synonymes) mais, à la différence des liens initiaux des classes, ceux du thesaurus ne sont pas directement exploitables pour le corpus et ont une faible productivité (ce qui rapproche le thesaurus et le dictionnaire).

Cette confrontation des résultats obtenus en utilisant des sources lexicales de natures différentes confirme l'intérêt des dictionnaires de langue pour le traitement de corpus spécialisées. Quand elles existent, les ressources spécialisées construites manuellement sont souvent incomplètes du point de vue de la synonymie. Un dictionnaire de langue, à la différence d'une source lexicale comme le thesaurus EDF, fournit un apport numériquement conséquent et qualitativement complémentaire.

5 Le rôle des ressources lexicales dans l'acquisition de connaissances spécialisées

Ces dernières années ont vu se développer les travaux portant sur le traitement automatique des textes techniques en vue de l'acquisition de connaissances spécialisées (lexicales, terminologiques ou ontologiques). Les méthodes reposant sur les seules données du corpus ayant montré leurs limites, ces travaux reposent souvent sur une approche mixte combinant des sources de connaissances préexistantes et des corpus spécialisés. Ils diffèrent cependant par le type des sources utilisées.

Il s'agit le plus souvent de sources spécialisées. Naulleau *et al.* (1996) utilisent certaines informations sémantiques du thesaurus EDF pour construire des classes de termes dans une perspective de filtrage de documents. Morin (1998) exploite des liens d'hyponymie d'un thesaurus d'agronomie pour amorcer l'acquisition de nouveaux liens à partir de corpus. Habert *et al.* (1998) confrontent à un corpus portant sur la médecine coronarienne une nomenclature médicale assez générale pour l'étendre et l'adapter à ce domaine particulier. Maynard *et al.* (1998) désambiguïsent les termes d'un corpus médical en s'appuyant à

la fois sur leurs distributions en corpus et sur un thesaurus médical.

Les données de la langue générale sont plus rarement utilisées, l'idée qu'elles sont peu pertinentes pour les textes techniques étant assez largement admise. Basili *et al. (*1997) montrent cependant comment la confrontation de *WordNet* et d'un corpus technique permet de construire un *WordNet* spécialisé, adapté au domaine considéré. Nous avons tenté de montrer que la contribution d'un dictionnaire de langue à la structuration d'une terminologie spécialisée est réelle (Hamon *et al.* 1998). Au regard de ces expériences mais aussi de Habert *et al.* (1998), il ressort que les données lexicales générales sont utiles dès lors qu'elles sont contrôlées par des attestations en corpus.

En pratique, c'est souvent faute de ressources spécialisées adéquates que l'on en vient à exploiter des sources lexicales générales. Peu de travaux ont exploré les deux pans de l'alternative et cherché à caractériser la contribution respective des sources lexicales spécialisées et des sources lexicales décrivant la langue générale.

C'est ce que nous avons tenté ici dans le cadre d'une expérience particulière. Des données lexicales générales et spécialisées sont combinées et projetées sur un corpus pour élaborer de nouvelles connaissances spécialisées. Les résultats montrent la complémentarité de ces sources pour la structuration de la terminologie du corpus étudié.

6 Conclusion

L'utilisation de ressources lexicales de différents types dans une méthode d'inférence de liens sémantiques permet de détecter un nombre important de liens sémantiques entre des candidats termes extraits d'un corpus technique. L'étude et la confrontation des résultats suivant le type de ressources a montré la complémentarité de données de la langue générale, comme *Le Robert*, et de données très spécialisées construites manuellement par un expert du domaine.

De plus, le faible recouvrement entre les liens inférés à l'aide de ces deux ressources et les liens présents dans la source de statut intermédiaire, le thesaurus EDF, justifie l'utilisation d'informations de la langue générale pour la structuration d'une terminologie du domaine étudié. Les résultats de cette expérience laissent penser que dans le contexte d'un système d'accès à l'information, il est probablement préférable de constituer manuellement de petites ressources très spécialisées et adaptées au corpus plutôt que d'exploiter un thesaurus lui-même coûteux à maintenir. L'utilisation d'un thesaurus spécialisé n'apporte pas une aide significative à la structuration d'une terminologie alors que des informations issues d'un dictionnaire de langue largement disponible, combinées à des petites ressources construites à partir du corpus, permet d'obtenir de meilleurs résultats et une relativement bonne couverture du corpus étudié.

Ce travail appelle un double prolongement. En montrant la nécessité de combiner différentes ressources pour la structuration de terminologies, cette étude fait apparaître deux aspects cruciaux des outils d'aide à la construction de terminologie. La réutilisation des données exploitées dans cette expérience a demandé la mise au point de quelques algorithmes. L'intégration de connaissances hétérogènes pour l'extraction d'information étant un problème en soi, il est nécessaire de proposer des outils et algorithmes permettant d'assurer la cohérence de ces informations lorsqu'elles sont utilisées conjointement. De plus, afin d'aider le terminologue au moment de la validation mais aussi lors de l'évaluation, il est important de structurer les résultats en fonction de mesures de pertinence. Cette classification des liens inférés doit tenir compte des validations et permettre également d'éliminer rapidement un certain nombre de liens erronés.

Remerciements

Ce travail est le fruit d'une collaboration avec la DER-EDF. H. Boccon-Gibod, Y. Abbas, M.-L. Picard (DER-EDF) et D. Bourigault (CNRS) ont mis leurs données et outils à notre disposition. Ce travail a bénéficié des discussions que nous avons avec eux ainsi qu'avec C. Jacquemin (Limsi) et B. Habert (UMR 8503).

Thierry Hamon,
Laboratoire d'informatique de Paris-Nord,
Université Paris-Nord,
Villetaneuse,
France.

Daniela Garcia,
Direction des études et recherches d'électricité de France,
Clamart,
France.

Adeline Nazarenko,
Laboratoire d'informatique de Paris-Nord,
Université Paris-Nord,
Villetaneuse,
France.

Bibliographie

Assadi (H.), 1997 : « Knowledge acquisition from texts : Using an automatic clustering method based on noun-modifier relationship », dans *Proceedings of the 35th Annual Meeting of the ACL - Student Session, Madrid, Spain.*

Basili (R.), Pazienza (T.) et Velardi, (P.), 1997: «Acquisition of selectional patterns in sublanguages», dans *Machine Translation*, n°8, p. 175-201.

Bourigault (D.), 1994: *Lexter, un logiciel d'extraction de terminologie. Application à l'extraction des connaissances à partir de textes.* Thèse de mathématiques, informatique appliquée aux sciences de l'homme, EHESS, Paris.

Cruse (D. A.), 1986: *Lexical semantics*, Cambridge University Press.

Dagan (I.) et Church (K.), 1994: *Termight*: «Identifying and translating technical terminology», dans *Proceedings of ANLP'94, Stuttgart, Germany*, p. 34-40.

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

Gros (C.) Assadi (H.), Aussenac-Gilles (N.) et Courcelle (A.): «Task models for technical documentation accessing», dans *Proceedings of the 9th European Workshop on Knowledge Acquisition (EKAW'96), Nottingham*.

Gros (C.) et Assadi (H.), 1997: «Intégration de connaissances dans un système de consultation de documentation technique», dans *Actes de ISKO'97*, Presses universitaire du Septentrion.

Habert (B.), Nazarenko (A.), Zweigenbaum (P.) et Bouaud (J.), 1998: «Extending an Existing Specialized Semantic Lexicon», dans *Proceedings of LREC'98, Granada*, p. 663-668.

Hamon (T.), Nazarenko (A.) et Gros (C.), 1998: «A step towards the detection of semantic variants of terms in technical documents», dans *Proceedings of Coling-ACL'98, Montreal, août 1998*, p. 498-504.

Jacquemin (C.), 1997: *Variation terminologique: Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus*, Mémoire d'habilitation à diriger des recherches en informatique fondamentale, Université de Nantes, Nantes.

Maynard (D.) et Ananiadou (S.), 1998: «Acquiring Contextual Information for Term Disambiguation», dans *Proceedings of the First Workshop on Computational Terminology, Montreal, August*, p. 86-90.

Miller (G. A.), Beckwith (R.), Fellbaum (C.), Gross (D.) et Miller (K.), 1993: *Introduction to WordNet: An on-line lexical database*, Technical Report CSL Report 43, Cognitive Science Laboratory, Princeton.

Morin (E.), 1998: «Prométhée, un outil d'aide à l'acquisition de relations sémantiques entre termes», dans *Actes de la Conférence TALN 1998, Paris*.

Naulleau (E.), Monteil (M.-G.) et Habert (H.), 1996: «Recycling an existing thesaurus to characterize and process terms», dans *Proceeding of Euralex'96, Göteborg*.

Ploux (S.) et Victorri (B.), 1998: «Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes», dans *Revue Tal*, vol. 39 n°1, p. 161-182.

Construction d'un lexique bilingue des droits de l'homme à partir de l'analyse automatique d'un corpus aligné

Cet article présente une expérience de construction d'un lexique français/anglais des droits de l'homme à partir de l'analyse automatique d'un corpus bilingue constitué d'arrêts rendus par la Cour européenne des droits de l'homme de Strasbourg. Ce corpus a été aligné automatiquement au niveau des phrases à l'aide d'heuristiques simples exploitant la structure logique des arrêts, identique dans les deux langues. Le logiciel *Lexter* a extrait des candidats termes de la partie française du corpus. Les juristes terminologues ont construit le lexique en repérant dans les phrases anglaises les équivalents des candidats termes français jugés pertinents.

Termes-clés :
extraction de terminologie, alignement terminologique, lexique bilingue, droits de l'homme

1 Introduction

Les systèmes de mémoire de traduction rencontrent un succès grandissant. Il est admis que leur utilité n'est avérée que dans les situations où la quantité de documents à traduire sur un même domaine est très importante. La présence de tels systèmes ne condamne donc pas le recours aux terminologies bilingues. Mémoire de traduction et terminologies multilingues sont des outils complémentaires dans l'usage que le traducteur peut en faire. Ces outils peuvent aussi être complémentaires dans leur mode d'élaboration réciproque. Ainsi, une terminologie bilingue peut être exploitée pour aligner un corpus parallèle, de même que, réciproquement, un corpus aligné peut être exploité pour (re)construire une terminologie bilingue. Dans cet article, nous exposons comment nous élaborons un lexique bilingue des droits de l'homme, à partir de l'analyse (semi-automatique) d'un corpus bilingue français/anglais, aligné au niveau des phrases, constitué d'un ensemble d'arrêts rendus par la Cour européenne des droits de l'homme de Strasbourg. Nous présentons dans la section 2 la problématique générale de l'alignement terminologique dans les recherches en traitement automatique des langues. Dans la section 3, nous posons le cadre générale du projet «lexique multilingue des droits de l'homme», auquel collaborent juristes, terminologues et linguistes informaticiens. Nous décrivons dans la section 4 les traitements informatiques effectués, concernant l'alignement du corpus et l'extraction terminologique automatique sur la partie française du corpus, et nous présentons une analyse quantitative des premiers résultats obtenus.

2 Alignement de terminologie

2.1 Alignement de phrases, de mots, de termes

Dans le domaine de la recherche en traitement automatique des langues (TAL), le thème de l'alignement multilingue suscite un grand nombre de travaux depuis plusieurs années. On s'est intéressé d'abord à l'alignement de phrases dans un corpus parallèle, c'est-à-dire un corpus bilingue dont l'une des parties est une traduction de l'autre. L'objectif est de construire des couples de phrases, extraites de chacune des parties, qui soient les traductions l'une de l'autre. On élabore ainsi un corpus aligné. Différents algorithmes et différentes techniques de type statistique ou linguistique, s'appuyant sur le niveau lexical ou sur celui des caractères, sont mis en œuvre (Brown *et al.* 1995, Church 1993, Gale et Church 1993). À partir d'un corpus aligné au niveau des phrases, on peut chercher à aligner des mots, et donc à construire des lexiques bilingues (Dagan *et al.* 1993). Les opérations d'alignement de phrases et d'alignement de mots sont souvent

interdépendantes, puisqu'on peut s'appuyer sur des couples de mots pour identifier des associations de phrases, et réciproquement, sur des couples de phrases pour identifier des associations de mots.

Depuis quelques années, les efforts portent sur l'alignement de termes. L'objectif est, à partir d'un corpus aligné au niveau des phrases, de construire non seulement des couples de mots simples, mais aussi des couples de séquences de mots, qu'elles soient désignées sous le nom de (candidats) termes, de collocations ou de syntagmes nominaux (Smadja et McKeown 1994, Kupiec 1993). Pour l'alignement de termes, parmi tous les choix méthodologiques à opérer pour élaborer un système d'alignement, l'un concerne l'extraction monolingue: dans certains travaux, par exemple Hull (1998), l'extraction de candidats termes est effectuée indépendamment sur chacun des deux corpus, et les candidats termes automatiquement extraits sont alors appariés à l'aide d'algorithmes basés sur des principes analogues à ceux adoptés pour l'alignement de mots; dans d'autres travaux, par exemple Gaussier (1998), l'extraction automatique est effectuée uniquement sur l'un des deux corpus, et les algorithmes mis en œuvre réalisent de façon conjointe les tâches d'extraction des termes dans l'autre corpus et d'appariement avec les termes extraits automatiquement du premier corpus.

2.2 Une expérience d'alignement de terminologie sans appariement statistique

Dans l'expérience que nous décrivons ci-dessous, notre démarche a été la suivante: à partir d'un corpus bilingue français/anglais, aligné au niveau des phrases, nous avons d'abord effectué une extraction automatique de candidats termes sur la partie française du corpus; ensuite, ce sont les juristes terminologues eux-mêmes qui sont allé chercher les équivalents anglais dans les contextes. Cette démarche est rendue possible, et efficace, grâce à une interface de validation spécialement conçue pour cette tâche, dans laquelle le juriste terminologue accède directement à l'affichage des couples de phrases, pour lesquels le candidat terme en cours d'analyse est présent dans la phrase française (*cf.* figure 1).

Sur le plan technique et algorithmique, notre démarche est donc «pauvre», puisqu'aucun appariement statistique n'est réalisé. Cependant, rien ne prouve qu'elle ne supporte pas la concurrence avec des techniques plus sophistiquées, qui seraient en mesure de proposer des appariements pour les candidats termes les plus fréquents. Se pose ici le problème de l'évaluation en ingénierie linguistique. Toutes les techniques d'alignement de terminologie sont présentées par leurs auteurs comme étant susceptibles d'apporter une aide à un utilisateur humain chargé de construire une terminologie bilingue. Les modes d'évaluation systématiquement évoqués par ces auteurs font appel aux notions de taux de précision (le plus souvent) et de taux de rappel (parfois). Pour mesurer le taux de précision, on évalue la proportion de couples corrects dans une liste de couples extraits automatiquement. Nous estimons que ce paramètre, certainement utile à un moment donné de l'élaboration du système, n'est pas le paramètre à prendre en compte de façon prioritaire pour évaluer les systèmes d'extraction multilingue. Le problème est identique à celui de l'extraction monolingue (*cf.* Bourigault et Habert 1998): puisque l'objectif n'est pas une extraction automatique, mais une aide efficace à l'utilisateur, ce sont les gains en temps et les gains en qualité, apportés par l'utilisation d'outils d'extraction et d'alignement, qu'il convient de mesurer. Une évaluation sérieuse des techniques d'extraction et d'alignement passe donc de façon incontournable par une phase d'observation de leur utilisation dans des expériences en grandeur réelle de constitution de terminologie. C'est ainsi que l'on peut espérer pouvoir mesurer leur apport effectif et donc leur intérêt, et identifier leurs lacunes et donc déterminer les directions de recherche futures en extraction automatique de terminologie.

Dans la suite de cet article, nous relatons une expérience au cours de laquelle des juristes terminologues ont construit «à la main» un lexique bilingue des droits de l'homme, à partir des résultats fournis par un outil informatique d'aide au dépouillement terminologique. Notre propos dans cet article est en particulier de fournir des indications quantitatives concernant le nombre de couples construits et le temps passé par les juristes terminologues. Ces données, avec le lexique lui-même, pourraient servir de base pour appréhender quel pourrait être l'intérêt de fournir aux juristes terminologues les résultats d'un appariement statistique de candidats termes.

3 Le projet «lexique des droits de l'homme»

Le travail a été effectué dans le cadre d'un projet financé par le Ministère français de la Recherche et de l'Enseignement supérieur(1), et s'est déroulé sous la forme d'une collaboration entre linguistes

(1) Nous remercions Anne Guyon, de la Direction de l'information scientifique et technique et des bibliothèques, pour son soutien constant et amical.

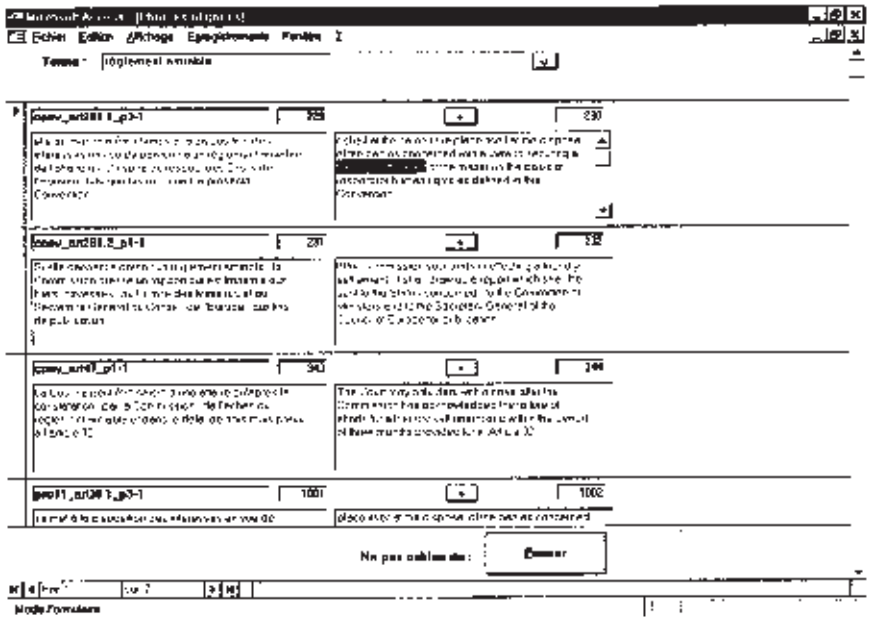


Figure 1 :
Illustration de la démarche. Pour chaque candidat terme (français) extrait par *Lexter*, l'interface HTL affiche les phrases du corpus français dans lesquelles il a été détecté avec en regard les phrases du corpus anglais qui en sont les traductions.

informaticiens et juristes terminologues. Ce lexique a été constitué à partir de l'analyse d'un corpus textuel bilingue composé de la Convention de sauvegarde des droits de l'homme et des libertés fondamentales, et de sa douzaine de protocoles, et de 36 arrêts rendus par la Cour européenne des droits de l'homme (CDEH) de Strasbourg en 1995.

La Cour européenne des droits de l'homme règle les litiges relatifs à l'interprétation et à l'application de la Convention de sauvegarde des droits de l'homme et des libertés fondamentales signée le 4 novembre 1950 et entrée en vigueur le 3 septembre 1953. Le texte de cette convention a été révisé depuis cette date par des protocoles (une dizaine) qui font partie intégrante de celle-ci. La compétence de la Cour s'exerce à l'égard des États qui l'ont reconnue de plein droit ou ont donné leur agrément à la saisine de celle-ci dans une affaire déterminée. À ce jour, une quarantaine d'états ont accepté la juridiction obligatoire de la Cour. Il existe deux versions officielles des textes précités : l'une en français, l'autre en anglais. Du 20 avril 1959 – date de son entrée en fonction – à 1997, la Cour a rendu plus de huit cents arrêts (dont environ 600 au cours des sept dernières années). Ces arrêts sont rédigés – comme la Convention et des protocoles – en anglais et en français. Les deux langues font également foi et il est impossible de distinguer une langue source et une langue cible. La Convention, les protocoles et les arrêts constituent ainsi un corpus juridique bilingue délimité. À une phrase dans une langue correspond, en effet, presque toujours exactement une phrase dans l'autre langue.

Une lecture attentive de ces textes fait apparaître certaines disparités de terminologie : ainsi pour un terme ou une expression français correspond parfois – dans les différents textes – plusieurs termes ou expressions anglais... et réciproquement. Les noms des institutions (*e.g.* Cour suprême autrichienne), en particulier, sont traduits de façon très variable d'un arrêt à l'autre... ce qui préoccupe vivement les traducteurs de la Cour. C'est pour mieux mettre en lumière la spécificité et la richesse du vocabulaire employé par la Cour dans les deux langues et, à terme et en accord avec la volonté de celle-ci de « normaliser » – dans une certaine mesure – la terminologie dans le domaine que nous avons entrepris ce travail. Celui-ci correspond en outre à un besoin : il n'existe aucun lexique fiable et récent en la matière et la CEDH elle-même ainsi que d'autres institutions (le Conseil constitutionnel en France, notamment) réclament un tel document.

Dès le départ du projet, et devant l'ampleur de la tâche, les juristes terminologues ont souhaité utiliser un outil informatique d'aide au dépouillement terminologique. C'est ainsi que s'est mise en place une collaboration entre les juristes terminologues du Centre de terminologie et de néologie (CTN) du Laboratoire de linguistique informatique de Villetaneuse, et les linguistes informaticiens de l'Équipe de recherche en syntaxe et sémantique de Toulouse.

4 Les traitements informatiques

4.1 Préparation des corpus : balisage et alignement

Le corpus de travail nous a été fourni sous forme électronique par la Cour européenne des droits de

l'homme de Strasbourg. Rappelons qu'il s'agit, pour chacune des deux langues, de la Convention accompagnée de ses protocoles et de 36 arrêts rendus par la Cour pendant l'année 1995. Il s'agissait de fichier Ascii pauvres, c'est-à-dire sans balisage logique des sections et encore moins de phrases. Dans cet état, plutôt que d'un corpus aligné, nous disposions d'un corpus «à aligner». En effet, la structure des arrêts est extrêmement bien marquée, et de façon équivalente pour les deux langues, sur le plan physique: parties, sections, alinéas, paragraphes, etc (*cf.* figure 2). De ce fait, pour n'importe quel lecteur humain, maîtrisant le français et l'anglais, il est aisé, à la lecture conjointe des deux corpus, d'associer à chaque phrase de l'un des corpus sa traduction dans l'autre. Mais pour permettre les traitements informatiques ultérieurs, il convenait de transformer cette structuration physique (visuelle), en une structuration logique manipulable par l'ordinateur: la tâche a consisté à identifier les phrases de chacun des deux corpus, et à associer le même identifiant aux couples de phrases équivalentes.

Dans le cas présent, étant donné la forte structuration du corpus, l'alignement a été effectué sur des bases uniquement formelles, il n'a pas été nécessaire de recourir aux techniques statistiques d'alignement de phrases, telles que celles évoquées dans la section 1. Les tâches de balisage et de segmentation ont été réalisées par une chaîne de programmes, qui ont été réalisés au fur et à mesure en s'assurant que l'application des règles s'effectuaient de façon identique sur chacun des deux corpus. La chaîne se décompose en 7 étapes:

1) Repérage du numéro de l'arrêt. Ce numéro (450 dans l'exemple ci-après) apparaît toujours dans le même contexte: *« TITLE: Affaire ALLENET de RIBEMONT c. France, CASE: 3/1994/450/529»*.

2) Repérage des grandes parties. Chaque arrêt se décompose en un certain nombre de grandes parties, chacune étant marquée par un titre normalisé: «*PROCÉDURE ET FAITS (*ang. *PROCEDURE AND FACTS)»,* dans laquelle est décrite la procédure qui a été suivie dans le pays d'origine avant que la Cour ne soit saisie, «*EN FAIT (ang. AS TO THE FACTS»*, qui présente les faits, *«EN DROIT (ang. AS TO THE LAW)»*, où sont détaillés les éléments du droit pertinents pour le cas, *«PAR CES MOTIFS, LA COUR (ang. FOR THESE REASONS, THE COURT)»*, qui expose la décision des juges de la Cour, *« OPINION DISSIDENTE (ang. DISSENTING OPINION)»*, dans le cas où certains des juges ne se sont pas rangés à l'avis de la majorité.

3) Repérage des numéros de section. Chaque arrêt est découpé en sections numérotées. C'est essentiellement ce découpage rigoureux en section, très largement répandu dans les textes de droit, qui rend possible l'approche formelle adoptée pour le découpage et l'alignement des corpus.

4) Repérage des paragraphes. Au sein de chaque section, le texte peut être organisé en paragraphes (portion entre deux retours chariots). Dans une grande majorité des cas, le découpage en paragraphes est identique dans les deux corpus. Dans les cas contraires, nous nous sommes autorisé l'insertion de marques de paragraphe pour rétablir un parallèle exact.

5) Repérage des citations. Dans le texte d'un arrêt peuvent apparaître des citations, soit des propos rapportés d'un des antagonistes du cas traité, soit des extraits de textes de loi des pays concernés. Pour distinguer ces passages du discours de la Cour lui-même, il a été jugé indispensable de repérer ces situations, heureusement marquées de façon régulière par un décalage du texte sur la droite et par des guillemets.

6) Élimination d'éléments divers non textuels. Les arrêts fourmillent d'élements textuels, peu intéressants sur le plan terminologique, que nous avons jugés bons d'éliminer pour simplifier la tâche, et de l'extracteur terminologique, et des juristes. Il s'agit en particulier des références à des cas déjà jugés, donnés sous la forme de leur titre ou de leur numéro (ex. «*Minelli c. Suisse, n° 266-A, p. 13* »); des références à des articles de loi (ex.: «*article 6 par. 2 (art. 6-2)»*); des dates; et enfin des noms d'individus.

7) Segmentation en phrases. Le programme de découpage en phrases s'appuie de façon très classique sur le repérage des ponctuations fortes, qui, hélas, diffèrent légèrement entre le français et l'anglais. Une évaluation de la qualité de l'alignement au niveau du paragraphe était fournie par le comptage du nombre de phrases pour chaque paragraphe. En cas de distorsion, nous nous sommes autorisé, là aussi, à insérer un certain nombre de modifications mineures dans les corpus (ajout, élimination ou changement de signes de ponctuations, ajout de retours chariots) de façon à ce que pour chaque paragraphe les nombres de phrases en français et en anglais soient les mêmes. Un exemple du résultat obtenu est présenté sur la figure 3.

Chaque corpus, qui compte environ 300 000 mots, a été segmenté en 12 131 phrases. La mise au point de ces programmes, qui ont été écrit en *Flex* sous Linux et la vérification du balisage ont pris une vingtaine d'heures. C'est très peu en regard du temps passé ensuite par les juristes à analyser le corpus pour construire le lexique. Cette phase de balisage et d'alignement était nécessaire pour la mise en place de la suite des opérations. À l'issue de ces

(...)
EN DROIT
I. SUR LA VIOLATION ALLÉGUÉE DE L'ARTICLE 6 PAR. 2 DE LA CONVENTION
31. M. Allenet de Ribemont dénonce les propos tenus lors de la conférence de presse du 29 décembre 1976 par le ministre de l'Intérieur et les hauts fonctionnaires de police qui l'accompagnaient. Il invoque l'article 6 par. 2 (art. 6-2) de la Convention, ainsi libellé:
«Toute personne accusée d'une infraction est présumée innocente jusqu'à ce que sa culpabilité ait été légalement établie.»
32. Le Gouvernement conteste en substance l'applicabilité de l'article 6 par. 2 (art. 6-2), en se fondant sur l'arrêt Minelli c. Suisse du 25 mars 1983 (série A n° 62). D'après lui, une atteinte à la présomption d'innocence ne peut provenir que d'une autorité judiciaire et ne se révéler qu'à l'issue de la procédure en cas de condamnation si la motivation du juge permet de supposer que celui-ci considérait a priori l'intéressé comme coupable.

(...)
AS TO THE LAW
I. ALLEGED VIOLATION OF ARTICLE 6 PARA. 2 OF THE CONVENTION
31. Mr Allenet de Ribemont complained of the remarks made by the Minister of the Interior and the senior police officers accompanying him at the press conference of 29 December 1976. He relied on Article 6 para. 2 (art. 6-2) of the Convention, which provides:
«Everyone charged with a criminal offence shall be presumed innocent until proved guilty according to law.»
32. The Government contested, in substance, the applicability of Article 6 para. 2 (art. 6-2), relying on the Minelli v. Switzerland judgment of 25 March 1983 (Series A no. 62). They maintained that the presumption of innocence could be infringed only by a judicial authority, and could be shown to have been infringed only where, at the conclusion of proceedings ending in a conviction, the court's reasoning suggested that it regarded the defendant as guilty in advance.

Figure 2:
Extraits des corpus français et anglais avant balisage et alignement

#A450_DR_1-p2-1
SUR LA VIOLATION ALLÉGUÉE DE L'ARTICLE 6 PAR. 2 DE LA CONVENTION
#A450_DR_a31_1-p1-1
<elim nompr>M._Allenet_de_Ribemont</elim> dénonce les propos tenus lors de la conférence de presse <elim date>du 29 décembre 1976</elim> par le ministre de l'Intérieur et les hauts fonctionnaires de police qui l'accompagnaient.
#A450_DR_a31_1-p1-2
Il invoque l'<elim article>article 6 par. 2</elim> <elim art>(art. 6-2)</elim> de la Convention, ainsi libellé:
#A450_DR_a31_1_CIT1-p1-1
Toute personne accusée d'une infraction est présumée innocente jusqu'à ce que sa culpabilité ait été légalement établie.
#A450_DR_a32_1-p1-1
Le Gouvernement conteste en substance l'applicabilité de l'<elim article>article 6 par. 2</elim> <elim art>(art. 6-2)</elim>, en se fondant sur l'arrêt <elim cas> Minelli c. Suisse</elim> <elim date>du 25 mars 1983</elim> (série A <elim numero>n° 62)</elim>.
#A450_DR_a32_1-p1-2
D'après lui , une atteinte à la présomption d'innocence ne peut provenir que d'une autorité judiciaire et ne se révéler qu'à l'issue de la procédure en cas de condamnation si la motivation du juge permet de supposer que celui-ci considérait a priori l'intéressé comme coupable.

#A450_DR_1-p2-1
ALLEGED VIOLATION OF ARTICLE 6 PARA. 2 (art. 6-2) OF THE CONVENTION
#A450_DR_a31_1-p1-1
<elim nompr>Mr Allenet de Ribemont</elim> complained of the remarks made by the Minister of the Interior and the senior police officers accompanying him at the press conference of <elim date>29 December 1976</elim>.
#A450_DR_a31_1-p1-2
He relied on <elim article>Article 6 para. 2 </elim><elim art>(art. 6-2)</elim> of the Convention, which provides:
#A450_DR_a31_1_CIT1-p1-1
Everyone charged with a criminal offence shall be presumed innocent until proved guilty according to law .
#A450_DR_a32_1-p1-1
The Government contested, in substance, the applicability of <elim article>Article 6 para. 2. 2</elim> <elim art>(art. 6-2) </elim>, relying on the l'arrêt <elim cas>Minelli v. Switzerland</elim> judgment of <elim date>25 March 1983</elim>. (Series A <elim numero>no. 62</elim>).
#A450_DR_a32_1-p1-2
They maintained that the presumption of innocence could be infringed only by a judicial authority, and could be shown to have been infringed only where, at the conclusion of proceedings ending in a conviction, the court's reasoning suggested that it regarded the defendant as guilty in advance.

Figure 3:
Extraits des corpus français et anglais après balisage et alignement. À l'issue de la phase de balisage et d'alignement, les deux corpus (300 000 mots chacun) sont découpés en 12 131 phrases, alignées.

sept étapes, le découpage ainsi effectué est exact jusqu'au niveau du paragraphe. En ce qui concerne le découpage en phrases au sein des paragraphes, une évaluation au moment de la recherche des équivalents dans l'interface de validation conduit à une estimation du taux d'alignement correct d'environ 90%. Les erreurs d'alignement ne sont pas préjudiciables au moment de la construction du lexique, puisque que dans tous les cas la phrase équivalente précède ou suit la phrase alignée, et l'utilisateur y accède donc rapidement.

4.2 Extraction terminologique sur le corpus français

Après le balisage et l'alignement, la seconde phase du travail informatique a consisté à faire traiter le corpus français par l'extracteur de terminologie *Lexter* (Bourigault 1993, Bourigault *et al.* 1996). *Lexter* reçoit en entrée un corpus de texte, en français, portant sur un domaine quelconque, effectue une analyse morpho-syntaxique de ce corpus et extrait une liste de candidats termes, c'est-à-dire de mots ou de séquences de mots susceptibles d'être retenus comme termes du domaine. Ces candidats termes sont soumis au terminologue par l'intermédiaire d'une interface hypertextuelle de validation, dite « Hypertexte terminologique *Lexter* » (HTL). Pour chacun des candidats termes, le terminologue accède, entre autres, à l'ensemble des phrases du corpus dans lesquelles le logiciel a détecté le candidat.

Lexter est habituellement utilisé dans des contextes monolingues. Dans ce projet, il a pu être utilisé pour élaborer une terminologie bilingue parce que nous disposions d'un corpus aligné. La démarche a donc été la suivante. *Lexter* a traité le sous-corpus français et a extrait des candidats termes français qui ont été validés par les juristes. L'interface de validation HTL a été légèrement modifiée de façon à ce qu'en regard les phrases du sous-corpus français dans lesquelles a été détecté un terme soient affichées les phrases du sous-corpus anglais qui en sont les traductions. Grâce à cela, les juristes retrouvent très facilement dans ces phrases, dans la grande majorité des cas, les équivalents anglais des candidats termes français (*cf.* figure 1).

4.3 Premiers résultats quantitatifs

L'expérience n'a pas encore atteint son terme. Nous estimons avoir accomplir les trois quarts du travail[2] (avant la soumission de nos résultats aux traducteurs experts de la Cour). Le lexique comporte actuellement un peu plus de 4 000 couples de termes français/anglais. Nous ne présentons ici que quelques indications de type quantitatif. Une analyse des résultats d'un point de vue théorie de la traduction et théorie du droit est exposée dans Humbley *et al.* (1999).

Conformément à ce que nous avons annoncé dans la section 2.2, c'est aux paramètres de la qualité du lexique et du temps d'élaboration qu'il convient de s'intéresser. La qualité est assurée par le fait que tous les couples de termes sont construits « manuellement » par les juristes terminologues par observation des attestations en corpus. Le lexique est donc une image fidèle, une photographie, de la façon dont les traducteurs de la Cour travaillent. Bien entendu, cela ne signifie pas que toutes les traductions sont bonnes, et l'un des résultats du travail pourrait être une tentative de normalisation, qui devrait être entreprise par les traducteurs eux-mêmes sur la base des résultats que nous leur fournissons.

Sur le plan du temps d'élaboration, les termes peu fréquents dans le corpus sont très rapidement analysés, puisque le juriste terminologue peut observer d'un coup d'œil l'ensemble des couples de phrases, pour choisir le ou les équivalents. Bien entendu, pour les termes très fréquents l'analyse est plus longue. Nous estimons à une vingtaine de jours pleins la durée d'élaboration du lexique dans son état actuel. C'est sur cette base que devrait être mesuré l'intérêt de l'introduction de techniques statistiques d'appariement de termes, dont les résultats ne sont précis que pour les termes extrêmement fréquents.

Dans le tableau 1, nous présentons les résultats chiffrés concernant les termes complexes (syntagmes nominaux). Environ 60% des candidats termes fournis par le logiciel ont été retenus par les juristes terminologues. Ce chiffre se situe dans la fourchette des taux habituellement observés en extraction automatique de terminologie. Précisons d'une part que l'élimination des candidats termes jugés non pertinents est une opération très simple et très rapide, et d'autre part que parmi les 40% éliminés, seuls quelques pour-cents doivent être considérés à proprement dit comme du bruit, consécutif à des erreurs d'analyse du logiciel, soit au moment de l'étiquetage, soit au moment du repérage des syntagmes nominaux. Une bonne partie des syntagmes non retenus apparaissent dans des parties du corpus décrivant les faits jugés (parties «EN FAIT», *cf.* section 4.1), et ne présentent pas de pertinence d'un point de vue du droit.

Il convient de constater que les hapax – les candidats termes extraits

(2) Nous remercions Delphine Bailly, stagiaire en traductrion juridique, pour sa contribution efficace au projet.

	Extraits	Vus	Non retenus	Retenus	analysés
fréquence = 1	12 193	6 375 *43%*	2 720 *57%*	3 655	1 183
fréquence > 1	4 283	3 185 *33%*	1 058 *66%*	2 127	2 127
TOTAL	16 476	9 560 *40%*	3 778 *60%*	5 483	3 310

Tableau 1.
Nombre de candidats termes extraits par *Lexter*, vus, retenus et analysés par les juristes terminologues.

une seule fois dans le corpus – présentent un intérêt certain. 57% des hapax ont été retenus (3 655/6 375), 36% des termes retenus sont des hapax (1183/3310). Ceci laisse entrevoir d'éventuelles limites des outils d'appariement statistique, qui se basent sur la récurrence des associations.

Le travail de constitution du lexique a très vite fait apparaître que les cas de traductions multiples étaient très fréquents, et ce dans les deux sens. C'est ainsi que sur les 2 127 termes français dont le nombre d'occurrence est supérieur à 2, 553 (soit 26%) ont au moins deux équivalents différents! 17% (168/981) des termes français qui n'apparaissent que deux fois dans le corpus se voient associés à deux équivalents anglais différents. Les chiffres sont du même ordre de grandeur de l'anglais vers le français. Une analyse des cas de traduction multiple est présentée dans Humbley *et al.* (1999). Un exemple pour conclure: le terme français *détention provisoire* est traduit par: *detention on remand, pre-trail detention, detention pending trial, imprisonment in default, remand custody, remand detention*.

5 Perspectives

En ce qui concerne le projet de lexique, notre objectif est d'achever la validation et l'analyse des candidats termes, de façon à atteindre une couverture maximale du corpus. Ce lexique sera ensuite soumis aux traducteurs de la Cour. Par ailleurs, le lexique, dans son état actuel, a servi de base pour l'intégration de deux nouvelles langues: le polonais et le roumain. Pour ces deux langues, seule la convention et certains protocoles ont été analysés. L'élaboration de lexiques multilingues incluant d'autres langues que les langues officielles de la Cour est indispensable pour les pays adhérents du Conseil de l'Europe qui souhaiteraient faire traduire les arrêts de la Cour dans leur(s) langue(s) nationale(s). Nous souhaitons étendre le lexique aux verbes et syntagmes verbaux. Le travail a été entièrement réalisé, à la main, sur la Convention et ses protocoles. Un travail à grande échelle, sur l'ensemble du corpus, est envisageable dans un avenir proche, dès que le logiciel *Lexter* aura été étendu à l'extraction des syntagmes verbaux. Sur le plan des recherches en extraction terminologique bilingue, nous avons entamé une collaboration avec les chercheurs du Centre de recherche de Xerox à Grenoble pour une confrontation et une évaluation comparative de nos approches.

Didier Bourigault,
Équipe de recherche en syntaxe et sémantique,
CNRS et Université Toulouse 2.

Christine Chodkiewicz,
Centre de terminologie et néologie,
Laboratoire de linguistique informatique,
Université Paris XIII.

John Humbley,
Centre de terminologie et néologie,
Laboratoire de linguistique informatique,
Université Paris XIII.

Bibliographie

Bourigault (D.), 1993: «Analyse syntaxique locale pour le repérage de termes complexes dans un texte», dans la *Revue Tal*, volume 34, n°2.

Bourigault (D.), Gonzalez-Mulliez (I.) et Gros (C.), 1996: «Lexter, a Natural Language Tool for Terminology Extraction», dans *Actes du 7e congrès international Euralex*, Göteborg, Suède.

Bourigault (D.) et Habert (B.), 1998: «Evaluation of Terminology Extractors: Principles and Experiments», dans Rubio (A.), Gallardo (N.), Castro (R.) et Tejada (A.), Éditeurs, dans *Actes de la première conférence internationale sur les ressources linguistiques et l'évaluation*, volume I, p. 299-305, Grenade, Espagne.

Brown (P.), Lai (J.) et Mercer (R.), 1991: «Aligning sentences in parallel corpora», dans *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics (ACL'91).*

Church (K. W.), 1993: «Char_align: A program for aligning parallel texts at the character level», dans *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics (ACL'93)*, Columbus.

Dagan (I.), Church (K. W.) et Gale (W. A.), 1993: «Robust bilingual word alignment for machine aided translation»,

dans *Proceedings of the workshop on Very Large Corpora* (*VLC'93*), Columbus.

Gale (W. A.) et Church (K. W.), 1993, «A program for aligning sentences in bilingual corpora», dans *Computational Linguistics*, 19(1).

Gaussier (E.), 1998: «Flow network models for word alignment and terminology extraction from bilingual corpora», dans *Actes de la 17e conférence internationale de linguistique informatique (COLING-ACL'99)*, volume I, pp. 444-450, Montréal, Canada.

Hull (D.) 1998: «A pratical approach to terminology alignment», dans Bourigault (D.), Jacquemin (C.) et L'Homme (M.-C.) éditeurs, dans *Proceedings of the first workshop on Computational Terminology (COMPUTERM'98)*, Montréal.

Humbley (J.), Chodkiewicz (C.) et Bourigault(D.), 1999: «Using *Lexter* to establish a glossary of Human Rights», (à paraître) dans *Actes de la conférence Terminology and Knowledge Engineering (TKE'99)*, Vienne.

Kupiec (J.), 1993: «An algorithm for finding noun phrase correspondences in bilingual corpora», dans *Proceedings of the 31st Annual Conference of the Association for Computational Lingustics (ACL'93)*, Columbus.

Smadja (F.) et McKeown (K.), 1994: «Translating collocations for use in bilingual lexicons», dans *Proceedings of the ARPA Human Language Technology Workshop*, Plainsboro, New Jersay.

Enrichissement terminologique en anglais fondé sur des dictionnaires généraux et spécialisés

Nous présentons une méthode d'extraction (enrichissement) terminologique en anglais fondée sur l'utilisation de ressources linguistiques et terminologiques déjà existantes. Nous utilisons les «termes simples significatifs» de chaque domaine, i.e. les substantifs, adjectifs, adverbes et participes apparaissant parmi les termes déjà répertoriés, pour rechercher des nouvelles séquences qui contiennent certains de ces termes simples significatifs, complétés éventuellement par des mots grammaticaux ou néologismes. Cette méthode, indépendante de la taille du corpus traité, permettra de construire un logiciel d'aide à la traduction.

Termes-clés: traitement automatique du langage naturel; extraction terminologique; acquisition de terminologie; dictionnaires électroniques; traduction assistée par ordinateur.

Introduction

Dans cet article nous présentons un prototype d'outil d'extraction automatique de termes, attaché à une grande base de données terminologique multilingue, *LexPro*, contenant plusieurs millions de termes pour une trentaine de domaines techniques. Cette ressource terminologique très riche est pour nous le point de départ pour la recherche de termes complexes d'un texte, qui sont soit déjà recensés dans la base, soit construits à partir du même matériau lexical que les termes déjà connus. D'autre part, les lexiques très complets de la langue générale, notamment ceux de l'anglais et du français, élaborés au LADL(1), nous permettent de proposer parmi les nouveaux candidats-termes ceux qui contiennent éventuellement des néologismes(2), des noms propres etc. Les premiers résultats prometteurs obtenus dans le domaine de l'informatique, pour lequel nous disposons d'un lexique de 85 000 termes anglais, simples et composés, nous permettent d'envisager l'adaptation de notre extracteur à d'autres langues et domaines de *Lexpro*, ainsi que son utilisation en tant qu'outil d'aide à la traduction.

Au cours de l'extraction, nous nous limitons à la recherche de termes complexes, i.e. contenant au moins deux mots simples (pour la définition du mot simple voir note 5). L'extraction se fait sur un texte étiqueté (partiellement ambigu) par les dictionnaires spécialisés et généraux, et elle est fondée sur la recherche de patrons syntaxiques dans le texte. Nous verrons que déjà des patrons assez simples peuvent donner des bons résultats (très bon rappel, précision relativement bonne) si les dictionnaires utilisés sont suffisamment riches.

Avant que l'extracteur puisse intervenir, une phase importante est celle de la préparation des ressources de *Lexpro*, i.e. le nettoyage des dictionnaires, le classement des termes par catégories grammaticales et par la structure syntaxique, le codage des mots inconnus et la flexion automatique des termes simples et composés. Ce travail, seulement partiellement automatisable, a beaucoup d'importance pour la qualité du résultat final. Ceci peut être vu comme l'inconvénient principal de notre méthode. Remarquons néanmoins qu'une fois la phase préparatoire accomplie, nous pouvons utiliser les mêmes dictionnaires comme une base très fiable non seulement pour l'extraction terminologique, mais aussi pour d'autres tâches, telles que l'automatisation de la traduction, l'indexation documentaire etc.

Dans la section suivante nous argumentons le choix de notre méthode qui peut être classée comme fortement fondée sur des dictionnaires, et indépendante de la taille des corpus traités. La section 2 explique l'utilité d'un extracteur terminologique en tant qu'outil d'aide à la traduction. La section 3 décrit les formats des dictionnaires utilisés. Dans la section 4 nous montrons les

(1) Laboratoire d'automatique documentaire et linguistique, Université Paris 7

(2) Dans ce contexte un néologisme sera pour nous un mot non reconnu ni par un dictionnaire spécialisé, ni par un dictionnaire général.

phases du travail de l'extracteur, et la section 5 présente les résultats obtenus en anglais dans le domaine de l'informatique, traité avec un dictionnaire spécialisé de 85 000 entrées. Dans la section 6 nous analysons les aspects novateurs de notre approche. Finalement, dans la section 7 nous montrons les perspectives de notre logiciel : les possibilités d'affinement des patrons de recherche, l'adaptation au français et à d'autres langues, la prise en compte des formats non Ascii des textes, etc.

1 Pourquoi cette approche ?

De nombreuses sociétés, centres scientifiques et corps administratifs effectuent des travaux visant le recensement et l'unification des terminologies propres à leurs domaines d'activités. Ainsi des dictionnaires spécialisés et des listes terminologiques de tailles souvent très importantes sont accessibles en vente, ou bien fonctionnent comme outils internes, dont la maintenance et l'enrichissement sont parfois confiés à une équipe de terminologues.

D'autre part, des scientifiques en ingénierie linguistique, comme Daille (1994), Bourigault (1994), proposent des outils d'extraction automatique de termes, qui dans la plupart des cas admettent le corpus traité comme le seul point de départ. Ceci est idéal pour le traitement de nouveaux domaines techniques ou pour des utilisateurs ne possédant pas de ressources terminologiques. Néanmoins, ceux qui ont déjà fait un effort de constitution de bases terminologiques ou bien ceux qui utilisent des dictionnaires de domaines bien définis, n'ont pas la possibilité, avec ces logiciels, de réutiliser leurs ressources déjà disponibles.

Un des travaux importants fondés sur un lexique spécialisé existant et permettant l'*enrichissement* terminologique plutôt que l'acquisition initiale, est celui de Jacquemin (1997). Ses résultats étant de très bonne qualité du point de vue de la pertinence des candidats-termes proposés, il est nécessaire d'utiliser un corpus de taille importante (l'auteur travaille sur un texte de 1,6 million de mots) afin d'obtenir un rendement satisfaisant (3300 nouveaux termes à partir d'un lexique de 70 000 entrées).

Pourtant, un très grand corpus du domaine traité n'est pas toujours disponible. En particulier dans le cadre d'aide à la traduction technique, dans lequel nous nous plaçons, les traducteurs ont rarement affaire à des documents qui dépassent 1 mégaoctet de texte. Ils disposent par contre presque toujours d'un ou plusieurs dictionnaires techniques, et éventuellement de lexiques personnels ou fournis par le client. Il nous a donc paru intéressant de proposer un extracteur terminologique qui permette de réutiliser des listes de termes disponibles, et dont la qualité de résultats ne dépende pas de la taille du corpus traité.

2 Extraction terminologique au service d'un traducteur technique

Dans le travail d'un traducteur technique un rôle important est attribué à la constitution d'un glossaire du document à traduire. Le traducteur, pas toujours expert du domaine traité, lit le texte en langue source et répertorie tous les termes simples et complexes inconnus ou difficiles à traduire, accompagnés d'exemples de leurs occurrences dans le texte. Cette liste est ensuite envoyée au client qui fournit ses propres traductions ou valide celles proposées par le traducteur. Gouadec (1997) propose une méthodologie très précise de création d'un tel glossaire, appelé chez lui un concordancier, et explique son rôle en tant que garant de l'homogénéité terminologique, ainsi que sa valeur contractuelle entre le traducteur et son client.

La constitution et la validation du glossaire du texte devraient en principe être effectuées avant le début de la phase de traduction. Ceci peut entraîner des délais importants, surtout pour des documents volumineux. C'est ici qu'un programme d'extraction automatique de candidats-termes pourra intervenir. Il analysera le texte et sortira instantanément une liste de candidats-termes, classés selon leurs fréquences d'apparitions, parmi lesquels le traducteur choisira ceux qu'il voudra inclure dans son glossaire.

Dans ce cadre, l'extracteur doit viser le maximum de rappel possible, car, pour la constitution du glossaire, le traducteur ne travaillera plus sur le texte entier, mais sur la liste de candidats. Il n'aura donc aucune possibilité de «rattraper» les termes qui ont échappé à l'extracteur. En même temps, la liste des candidats ne peut pas être excessivement longue (précision relativement bonne), en particulier le temps de sa consultation ne peut pas dépasser celui de la création du lexique «à la main». Remarquons néanmoins la difficulté de définir les notions de rappel et de précision dans notre contexte, liée à la question de ce qui doit être considéré comme un bon terme. Pour un terminologue qui dépouille de grandes quantités de textes, un bon terme est celui avec un statut établi constaté dans différentes sources et chez plusieurs auteurs. En revanche, pour un traducteur, un terme qu'il faut retenir est celui qui pose un problème de traduction dans le texte traité. Un candidat retenu par le traducteur pour le glossaire d'un texte

donné peut ne plus l'être pour un autre texte, un autre client ou un autre contrat de traduction.

3 Dictionnaires électroniques généraux et spécialisés

Les points de départ pour la recherche de termes seront pour nous deux grandes ressources linguistiques et terminologiques: les dictionnaires électroniques généraux(3) élaborés selon la méthodologie du LADL, et *Lexpro*, une grande base de données terminologiques multilingue(4) et multidomaine. À l'état actuel, notre outil de recherche de termes n'utilise que la partie anglaise de ces ressources, mais nous envisageons d'étudier son adaptation au français et ensuite à d'autres langues.

3.1 Dictionnaires électroniques du LADL

Le dictionnaire électronique des mots simples(5) *Delas*(6) de l'anglais, contient les formes de base (l'infinitif pour les verbes, le singulier pour les noms etc.) des mots simples avec leurs codes flexionnels. Le code flexionnel décrit la façon d'obtenir toutes les formes fléchies d'un mot simple à partir de sa forme de base. Par exemple l'entrée *loaf, N6* indique le code *N6*, équivalent à l'ensemble de terminaisons (*<E>:s ,1ves:p*) , qui signifient que le singulier est égal à la forme de base (il faut ajouter <E> i.e. une séquence vide à la forme de base) et que le pluriel *loaves* s'obtient en enlevant une lettre de la fin et en rajoutant la terminaison *ves*. Le *Delas* anglais fournit une bonne couverture de la langue générale, car il contient à présent plus de 90 000 entrées. À partir du *Delas*, on obtient automatiquement le dictionnaire des formes fléchies des mots simples, le *Delaf*, contenant plus de 170 000 entrées. À chaque entrée est attribuée une étiquette indiquant sa catégorie (nom, adverbe, adjectif etc.) et éventuellement ses traits morphologiques (nombre, genre, personne etc.). Par exemple, le mot *permits* est décrit par deux lignes suivantes:

[1] permits,permit.N:p
[2] permits,permit.V:P3s

C'est soit le pluriel du nom *permit*, soit la troisième personne du singulier indicatif présent du verbe *to permit*.

Les mots composés(7) sont recensés dans un autre dictionnaire, le *Delac*, qui, pour chaque entrée, donne sa catégorie, ses traits flexionnels et la façon dont elle se fléchit, par exemple:

[3] point(point.N1:s) of view,N:s/+N

Cette entrée du *Delac* indique que *point of view* est un nom composé au singulier (N:s) et pouvant se mettre au pluriel (*/+N* signifie qu'il y a une flexion en nombre) par la mise au pluriel de ses *constituants caractéristiques*, i.e. les composants simples pour lesquels le code flexionnel est indiqué, ici *point* (pour les détails de la flexion automatique des composés, consulter Chrobot (1998)). Ces informations permettent de générer automatiquement le *DELACF*, le dictionnaire des mots composés fléchis, comme le montrent les exemples suivants:

[4] point of view,.N:s
[5] points of view,point of view.N:p

Le *Delac* anglais, qui est au cours de réalisation, comprend à présent près de 60 000 entrées, dont environ 50 000 noms composés, 4 000 adverbes composés (e.a. *all of a sudden*), 3 500 adjectifs composés (e.a. *left-handed*), 300 prépositions composées (e.g. *in front of*) et 100 conjonctions composées (e.a. *as well as*).

3.2 *Lexpro*

La base terminologique *Lexpro*, construite à partir de 120 dictionnaires spécialisés traditionnels, mis sur un support informatique, contient actuellement près de 5 millions de termes en 11 langues, dont environ 2 millions en anglais. L'exhaustivité et la qualité des données varient beaucoup d'un dictionnaire à l'autre et d'un domaine à l'autre. De nombreuses entrées nécessitent la correction orthographique, la mise en évidence de certaines abréviations, la séparation

(3) Il s'agit de bases de données lexicales pour la morphologie flexionnelle de la langue générale. À présent, les dictionnaires généraux du français, de l'anglais, l'allemand, l'italien et l'espagnol sont accessibles.

(4) Anglais, français, russe, allemand, espagnol, portugais, italien, néerlandais, danois, suédois, arabe.

(5) Un mot simple est pour nous une séquence contiguë de lettres de l'alphabet (contenu pour chaque langue dans un fichier à part), délimitée par deux séparateurs: blancs, apostrophes, tirés, points, ou autres caractères de ponctuation. Cette définition est purement orthographique, car e.a. en anglais air et airplane sont des mots simples, tandis que air-bed et air force sont des composés.

(6) Pour la description détaillée du format des dictionnaires LADL consulter Courtois et Silberztein (1990).

(7) Un mot composé est, en bref, une séquence contigüe de deux ou plus de mots simples, dont les propriétés sémantiques et/ou syntaxiques ne peuvent pas être déduites de celles de ses constituants.

des variantes, ou l'unification du format (effacement des déterminants initiaux, remise au singulier etc.). Très peu d'auteurs de dictionnaires indiquent les propriétés grammaticales de leurs termes, telles que catégorie, genre, nombre, existence du pluriel etc., indispensables pour notre extracteur. C'est pourquoi l'exploitation de notre base demande une phase préparatoire, seulement partiellement automatisable. Elle permet d'obtenir une très haute qualité des données utilisées, qui est la source principale de l'efficacité de notre approche (voir section 6). Cette phase doit comprendre entre autres:

• L'analyse lexicale des termes du *LexPro* à l'aide du dictionnaire *Delaf* général décrit ci-dessus, afin de retrouver tous les mots simples inconnus que l'on doit ensuite coder, i.e. fournir pour chacun un code flexionnel comme cela a lieu dans le dictionnaire *Delas* (ce codage est manuel).

• La correction des termes mal orthographiés (semi-automatique).

• Pour chaque terme, le marquage de sa catégorie (semi-automatique).

• Pour les termes complexes, le marquage de ses *constituants caractéristiques*, ou *têtes* (semi-automatique) – voir section 3.1.

Tout le contenu de la base *LexPro* n'est pas pertinent du point de vue de la recherche de termes. Nous n'avons pas besoin de certains champs, comme définitions, précisions, commentaires, sources des données etc. Nous allons donc extraire de la base ce que nous appelons les «formes écrites»: les termes principaux, leurs synonymes, leurs abréviations, leurs antonymes etc., i.e. les « vraies » unités terminologiques telles qu'elles peuvent être trouvées dans des textes. Ensuite nous convertirons les données ainsi obtenues en deux dictionnaires ressemblant au *Delas* et le *Delac* du LADL: l'un pour les termes simples et l'autre pour les termes composés.

Nous avons accompli le prétraitement des données décrit ci-dessus pour deux dictionnaires du domaine de l'informatique, celui de De Sollier (1999) et de Hildebert (1998), et nous avons obtenu un *Delas* spécialisé de 27 000 entrées et un *Delac* spécialisé de 57 000 entrées, dont voici des exemples:

[6] interrupt,N1+Spec
[7] interrupt,V7+Spec
[8] arithmetic overflow indicator(indicator.N1:s),N+Spec+Comp:s/+N

La flexion automatique de ces deux dictionnaires a engendré le *Delaf* et le *DELACF* informatiques de 74 000 et 108 000 entrées respectivement. Le trait supplémentaire *+Spec* renseigné pour chacun des termes simples et composés permettra, en cours de la recherche de nouveaux termes, de faire la distinction entre les étiquettes provenant des Delaf/DELACF généraux et celles des Delaf/DELACF spécialisés. En revanche, le trait *+Comp* est celui qui différenciera les termes spécialisés composés et simples.

4 Phases de l'extraction

Nous ramenons le problème de l'extraction de termes à celui de la recherche de patrons syntaxiques dans un texte. Le texte est d'abord soumis à l'analyse lexicale qui attribue à chaque unité atomique une ou plusieurs étiquettes syntaxiques provenant des dictionnaires utilisés. Ensuite, dans le corpus ainsi étiqueté, nous cherchons toutes les séquences qui correspondent au patron syntaxique donné. Ces séquences seront les candidats que l'utilisateur va pouvoir valider, i.e. décider s'ils sont ou non des termes.

Le schéma de fonctionnement de l'extraction est montré sur la figure 1, où les éléments ovales représentent les différentes phases de l'algorithme, tandis que les éléments rectangulaires correspondent aux entrées et sorties de ces phases (les dictionnaires *Delaf* et *DELACF* entourés des rectangles dessinés en gras sont consultés en mode prioritaire – voir section 4.1). Nous pouvons voir que le nombre de données en entrée est le plus important dans l'étape de la recherche des mots simples et composés du texte. En effet, les résultats de cette étape sont décisifs pour l'efficacité de la méthode.

Tous les 4 algorithmes utilisés – l'indexation, l'analyse lexicale des mots simples, l'analyse lexicale des mots composés, et la recherche de patrons – ont été récupérés du système Intex développé au LADL. Pour la représentation des dictionnaires, aussi bien que pour le dépouillement des corpus, ils emploient un modèle à états finis. Chaque dictionnaire utilisé pour l'étiquetage est converti en un automate à états finis, ce qui permet sa consultation en temps linéaire en fonction de la longueur du mot recherché. Les patrons syntaxiques de l'extraction sont, eux aussi, des automates finis. Pour les détails de l'implémentation, consulter Silberztein (1997).

4.1 Étiquetage du texte

Deux phases préliminaires du traitement sont celles de l'identification des items(8) du texte, et de la constitution de l'index qui, pour chaque item, donne la liste de

(8) Un item de texte, tel qu'il est défini dans *Intex*, correspond à une suite contigüe soit de lettres, soit de séparateurs (caractères non alphabétiques).

toutes ses occurrences. L'indexation nécessite un temps supplémentaire de traitement, mais elle accélère, surtout pour des corpus volumineux, les autres étapes de l'extraction. L'analyse lexicale s'occupe ensuite de la reconnaissance des mots simples et composés du texte selon 5 dictionnaires *Delaf/DELACF* et deux niveaux de *priorité*.

Les mots simples sont recherchés dans 3 dictionnaires: le *Delaf* général et le *Delaf* spécialisé, décrits dans la section 3, ainsi qu'un dictionnaire des mots grammaticaux, qui a le même format que le *Delaf*, et qui contient près de 500 prépositions (*about, after, through…*) déterminants (*the, no, one…*), conjonctions (*though, and, or…*), adverbes (*above, almost, yet…*), pronoms (*who, another, few…*) et certains verbes (*are, can, have, may, will…*). À ce petit lexique est attribuée une priorité supérieure aux deux autres dictionnaires *Delaf*. Cela signifie que chaque mot simple du texte est d'abord recherché parmi les mots grammaticaux, et seulement s'il n'y figure pas, sa recherche est poursuivie dans les *Delaf* général et spécialisé. Cette mesure a été introduite pour éviter le superflu des séquences incorrectes extraites plus tard par le patron syntaxique: il n'est pas rare qu'un auteur décide de fournir dans son dictionnaire spécialisé les traductions de certains mots grammaticaux, même si elles sont, en principe, universelles. Ceci fait apparaître dans notre *Delaf* spécialisé des entrées non significatives du domaine. Le dictionnaire prioritaire nous garantira que ces entrées ne participeront pas à la recherche de patrons, fondés essentiellement sur l'étiquette *+Spec* comme nous le verrons dans la section suivante.

Il arrive qu'un mot, qui dans la plupart des cas a une fonction grammaticale, puisse avoir dans un langage spécialisé une signification spécifique. Par exemple, en informatique, *or* désigne une opération logique, donc il est, en principe, incorrect d'interdire à l'analyseur lexical l'accès à l'étiquette *or,.N+Spec:s*. Ceci ne permettrait pas la reconnaissance de certains bons candidats-termes informatiques, comme *or gate*. Nous espérons néanmoins que le silence ainsi introduit est minimal, au moins pour les domaines bien couverts par les dictionnaires *Lexpro*, car les termes «curieux » comme *or gate* figurent déjà dans le *DELACF* spécialisé, et ils participeront éventuellement à la recherche des candidats plus larges comme *exclusive-or gate*.

Parallèlement à la reconnaissance des mots simples, l'analyseur lexical consulte les deux dictionnaires des mots composés: *DELACF* général et *DELACF* spécialisé, décrits dans la section 3. Les deux *DELACF* ont priorité sur tous les autres dictionnaires. Ceci veut dire que si une séquence contigüe d'items du texte est reconnue en tant qu'entrée du *DELACF* spécialisé ou général, elle est dans la suite traitée *en bloc*, i.e. on ne cherche plus à étiqueter ses sous-séquences par les autres

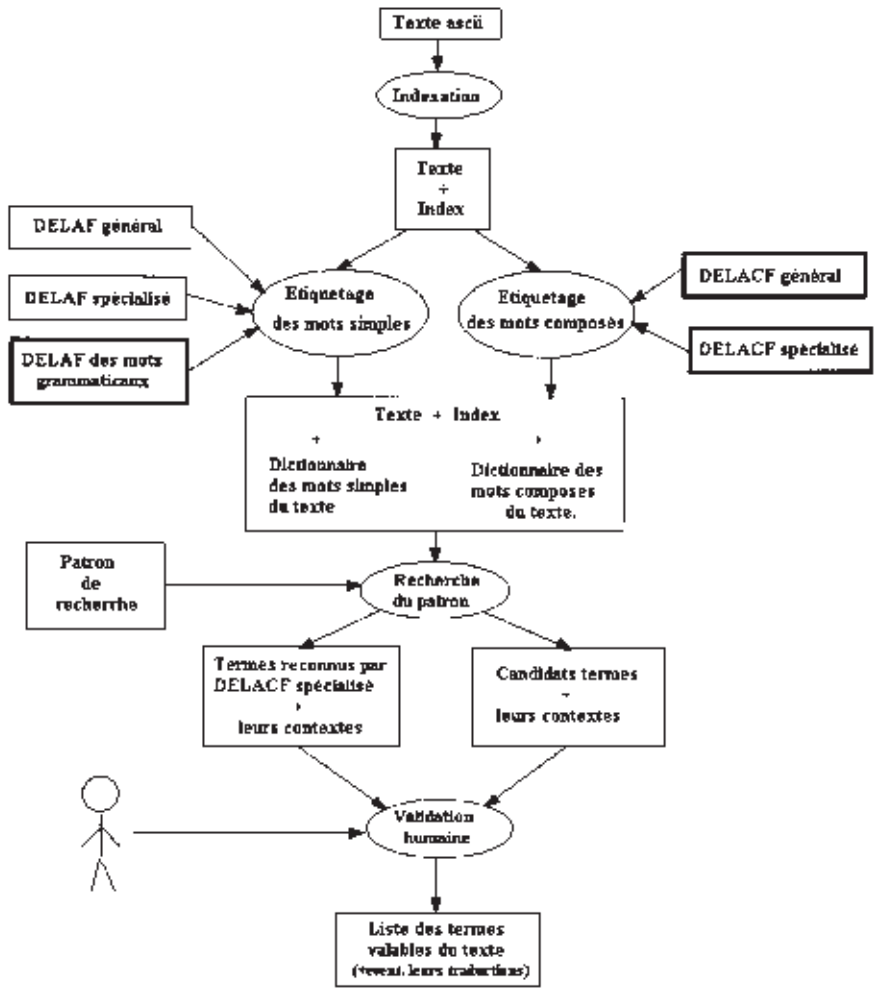


Figure 1.
Schéma de fonctionnement de l'extraction.

dictionnaires, et dans la recherche de patrons elle sera équivalente à un mot simple.

L'analyse lexicale produit deux dictionnaires associés au texte de départ – le dictionnaire des mots simples et celui des mots composés reconnus dans le texte – dont le format est identique à celui des *Delaf* et *DELACF* généraux et spécialisés, décrit dans la section 3. Les unités d'un ou plusieurs items reçoivent une ou plusieurs étiquettes grammaticales, pour lesquelles nous n'effectuons aucune désambiguïsation, mis à part l'utilisation des dictionnaires prioritaires. Ainsi un mot se retrouve souvent avec 2, 4 ou 6 étiquettes, dont certaines identiques au trait *Spec* près, par exemple le dictionnaire des mots simples du texte peut contenir des entrées suivantes:
[9] registers, register.N+Spec:p
[10] registers, register.N:p
[11] registers, register.V+Spec:P3s
[12] registers, register.V:P3s

L'ambiguïté entre les étiquettes [9] et [10], ainsi que [11] et [12] ne pose pas de problèmes pour la reconnaissance du patron que nous avons choisi. En revanche, l'attribution par le *Delaf* spécialisé de catégories différentes pour le même mot (ambiguïté entre [9] et [11]), peut être à l'origine d'un certain nombre de candidats termes incorrects.

Les entrées du dictionnaire des mots composés du texte peuvent aussi être ambigües, si elles figurent à la fois dans le *DELACF* général et dans le *DELACF* spécialisé, ou bien si un terme complexe a réellement plusieurs emplois avec des catégories différentes. Là aussi seul ce dernier type d'ambiguïté peut influencer les résultats de la recherche du patron syntaxique.

4.2 Recherche de patrons

Nos patrons de recherche se présenteront sous forme d'automates à états finis, dont l'alphabet sera celui des étiquettes grammaticales décrites dans la section 3 et attribuées aux unités du texte dans la phase de l'analyse lexicale. Nous allons toujours chercher à extraire les séquences contigües et maximales décrites par les patrons, sans nous préoccuper de l'existence de sous-termes ou insertions éventuelles à l'intérieur de ces séquences. Entre autres, nous n'avons pas pris en compte la possibilité de rechercher des conjonctions, comme *application name and location, flash and SRAM cards*, et d'autres variantes terminologiques qui font objet de l'étude détaillée par Jacquemin (1997).

Jusqu'à présent, nous avons mis au point un seul patron, qui est représenté par le graphe sur la figure 2. Un graphe est équivalent à un automate, ce que l'on peut voir si pour chaque nœud du graphe: a) on ajoute un état avant ce nœud; b) on transforme le nœud en une transition, qui sera étiquetée par le symbole de l'intérieur de ce nœud; c) le nœud le plus à gauche devient l'état initial; d) le nœud entouré d'un cercle devient l'état final. La direction des transitions est toujours celle du côté droit d'un nœud vers le côté gauche d'un autre nœud.

Analysons quelques exemples de séquences extraites par les différents chemins du graphe. La branche centrale, contenant l'étiquette *<N+Spec:s>*, permet de trouver toutes les suites de noms spécialisés. En effet, comme le montrent nos dictionnaires informatiques, les termes complexes du schéma $N_1N_2\ldots N_k$ (avec k>=2), e.g. *access frequency, program reference table, data transmission control unit*, sont de loin les plus nombreux. Nous admettons l'insertion éventuelle de la marque «'s» ou «' » du génitif, ainsi que du séparateur «/», entre deux noms, pour ne pas manquer les candidats du type *Windows NT User's Guide, server's hostname, matrix' mode, I/O activity*. La contrainte sur le nombre dans l'étiquette *<N+Spec:s>* a été introduite pour éviter le bruit trop important provenant des ambiguïtés entre les verbes à la troisième personne du singulier et les noms au pluriel, comme dans les contextes suivants (les séquences extraites à tort sont soulignées): *the analyzer displays information on the following…, an array supports write caching if it has…, the hit ratio shows cache efficiency and…* Cette contrainte risque d'introduire un certain silence, car il est en principe possible qu'un terme du type $N_1N_2\ldots N_k$ contienne un nom au pluriel sur une des positions 1… k-1, comme ceci a lieu dans des termes déjà connus, e.g. *active contents type, advanced communications system, american national standards institute.*

Les deux chemins supérieurs du graphe, contenant les étiquettes *<ADV+Spec>*, *<A+Spec>* et *<V+Spec:K>*, assurent la prise en compte des adjectifs, participes passés et adverbes spécialisés à l'intérieur des séquences du type *AN, ANN, NAN, AdvVN etc.*, comme *long return, parallel access array, storage system's physical disks, locally attached arrays.*

La partie du graphe, utilisant les étiquettes *<MOT>* (n'importe quel mot) et *<NB>* (nombre), permet d'extraire les mots liés par le trait d'union ou le soulignement, qui marquent souvent le caractère figé des séquences concernées, à condition qu'il n'y ait pas d'espace autour de ces séparateurs (ceci est exprimé par le signe #). Cette branche du patron correspondra à des candidats comme *operating system-related restrictions, DAE-to-DAE interconnection, dual-initiator/dual-bus configuration, mia_output_disable, 512-byte data block.*

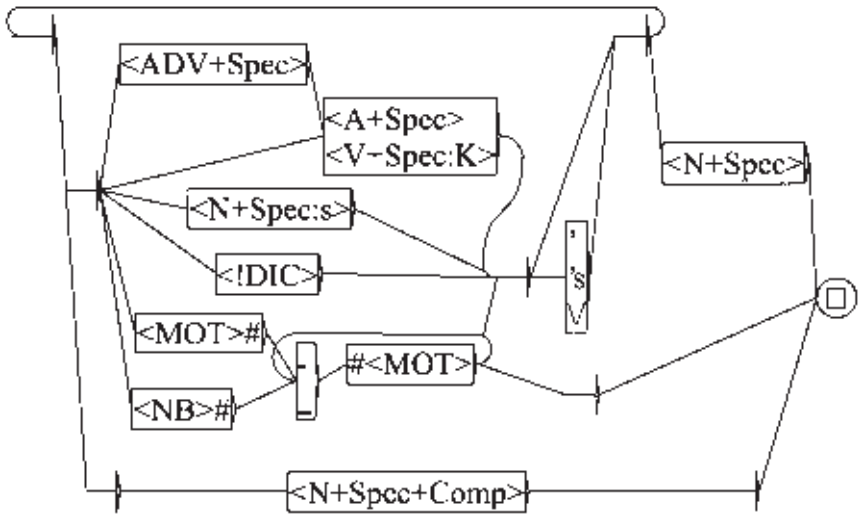


Figure 2.
Patron de recherche de nouveaux termes.

Le nœud du graphe étiqueté par le symbole *<!DIC>* est celui qui donne la possibilité de prendre en compte les néologismes, i.e. les mots (communs et propres) non reconnus ni par le *Delaf* général ni par le *Delaf* spécialisé. Parmi les exemples de ce type trouvés dans nos corpus se trouvent (les néologismes sont soulignés): *OpenManage Data Administrator, midplane connectors, nonmirrored write, powerup initialization sequence*.

Finalement, le chemin inférieur du graphe permettra, grâce aux traits *+Spec+Comp*, d'extraire les noms composés terminologiques reconnus déjà par notre *DELACF* spécialisé au cours de l'analyse lexicale, qui n'ont pas encore été inclus dans des fréquences plus longues extraites par les autres parties du patron. Ces composés, étant des termes établis du domaine, ont une grande chance d'être retenus par l'utilisateur pour le texte donné.

Remarquons que toutes les étiquettes du graphe contenant le trait *+Spec*, à part celle avec en plus le trait *+Comp*, peuvent correspondre non seulement à des mots simples spécialisés, mais aussi à des composés, ce qui permet la reconnaissance des surcompositions obtenues par ajouts de nouveaux modifieurs ou têtes à des termes déjà connus, comme le montrent les candidats suivants (les composés existants déjà dans le *DELACF* spécialisé sont soulignés): *ac power distribution, disk-based application, user free memory*.

La mise au point du graphe ci-dessus a été faite d'une façon expérimentale, par des allers-retours constants entre le patron de recherche et un corpus informatique de 700 kilooctets fourni par un traducteur technique. Nous avons essayé de trouver un juste milieu entre le rappel et la précision introduits, dont nous essayons d'estimer les proportions dans la section 4.3.

4.3 Validation

La figure 3 présente l'interface de la phase de validation. Cette validation est effectuée par le traducteur qui utilise le logiciel pour créer son glossaire de traduction.

Après l'ouverture du texte, a lieu la phase d'extraction décrite dans la section précédente. Ensuite, les

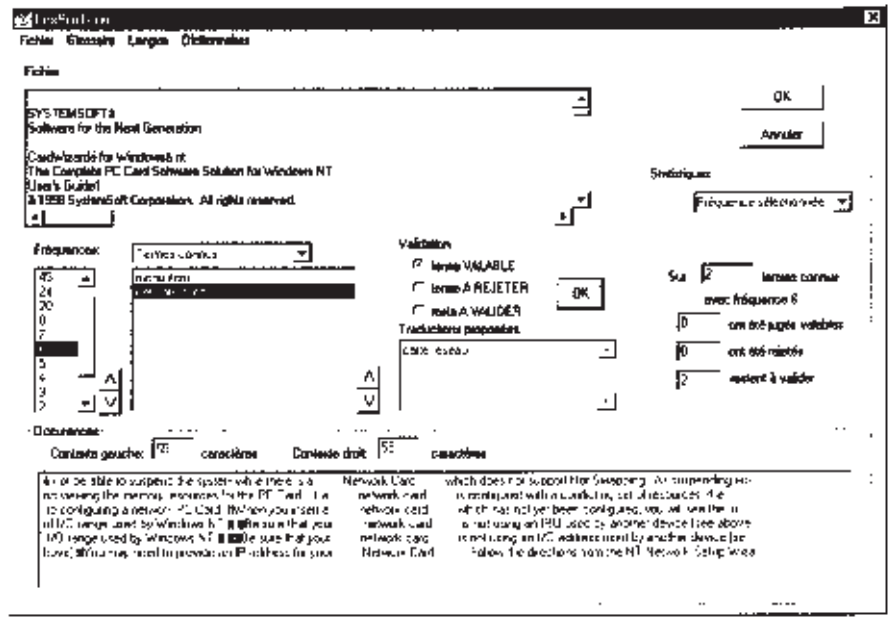


Figure 3. Interface de validation.

séquences extraites sont regroupées par variantes orthographiques : deux séquences sont considérées comme variantes si elles sont égales à l'emploi des minuscules et des majuscules près. Pour chaque ensemble de variantes du même candidat, celle qui emploie le moins de majuscules est choisie comme la forme représentative. Toutes les formes représentatives sont triées selon les fréquences de leurs variantes dans le texte et divisées en deux listes : « Termes connus » et « Nouveaux termes », selon qu'elles figurent ou non dans le *DELACF* spécialisé (et sont donc des termes connus). La sélection d'un candidat de chaque liste entraîne l'affichage de toutes ses occurrences avec leurs contextes gauche et droit de longueurs réglables. L'utilisateur peut consulter la liste de candidats soit dans l'ordre alphabétique (option « Toutes » sur la liste « Fréquences »), soit selon leurs fréquences. Dans le deuxième cas, seuls les candidats ayant la fréquence sélectionnée s'affichent.

Le rôle principal de l'utilisateur est de valider les candidats-termes en activant l'un des trois boutons du milieu de l'écran (on choisit le bouton intitulé « reste À VALIDER », si l'on n'est pas sûr du statut d'un candidat), et éventuellement de proposer une ou plusieurs traductions pour chaque terme jugé valable. Les statistiques à droite de l'écran indiquent le nombre de candidats déjà retenus ou rejetés et de ceux qui restent à valider. On peut interrompre la validation à tout moment. Typiquement, on ne va examiner que les fréquences les plus élevées, mais cette stratégie n'est pas toujours la bonne : de nombreux candidats corrects n'apparaissent qu'une seule fois, même dans des corpus volumineux.

Les résultats finaux de la validation peuvent être exportés dans un fichier lisible par un logiciel permettant la consultation et la création de bases terminologiques (comme Access etc.) afin de mettre en forme le glossaire de traduction du texte (voir section 2).

5 Premiers résultats

Pour estimer l'efficacité de notre logiciel, nous nous servons des deux critères habituels, ceux de précision et de rappel. La précision est définie comme la proportion de bons termes parmi tous les candidats-termes proposés par l'extracteur. Le rappel signifie la proportion de bons termes proposés par l'extracteur parmi tous les termes existant dans le texte traité.

Nous avons déjà mentionné à la section 2 qu'il était difficile de décider si une séquence est ou non un terme dans le contexte de création du glossaire d'un texte à traduire. Néanmoins, nous avons fait un test pouvant nous donner des premières indications quant à la qualité de notre outil. Il a été réalisé sur un petit corpus de 52 kilooctets (8500 mots) de texte anglais sur le domaine informatique, fourni par un traducteur technique.

Nous avons d'abord effectué un prétraitement du corpus, qui consistait à marquer manuellement toutes les occurrences de termes. Ce choix a dû être parfois arbitraire, car, à part les termes informatiques connus, comme *AC power*, *hard disk*, *power management*, nous avons sélectionné certaines séquences sans statut terminologique établi, mais devant être, à notre avis, traitées comme unités de sens au cours de la traduction, par exemple : *cleanup feature*, *easy-to-read displays*, *non-network function*, *active termination device*, *CardWizard for Windows NT Notify Options screen*, *Card View Display Options*, *notification message timeout*. Nous prenions en compte toujours la séquence maximale, sans rechercher ses sous-termes éventuels. Le glossaire du texte ainsi obtenu, contenant 839 occurrences(9), a été comparé aux listes de séquences extraites du même texte par notre extracteur. Parmi les 839 termes, 240 ont été reconnus par le *DELACF* spécialisé, et 450 ont été extraits par le patron syntaxique, ce qui donne un rappel égal à 82%.

Cette valeur, pas encore assez élevée du point de vue de notre application, est due en grande partie aux limites introduites dans le patron de recherche. Les termes non reconnus sont entre autres ceux qui : contiennent des prépositions (46% de cas), comme *PC Card support for Windows NT* ; contiennent des noms au pluriel sur des positions non terminales (15%), comme *options menu* ; sont contenus dans des séquences plus longues (20%), comme *PC Card information screen displays* (le bon terme est souligné, *displays* a été extrait à tort). Ce dernier exemple montre le problème très important en anglais d'ambiguïtés entre les noms et les verbes, renforcé encore dans le domaine de l'informatique par le phénomène fréquent de conversion(10) de noms en nouveaux verbes ou de verbes en nouveaux noms. Par exemple, le mot *network*, fonctionnant dans la langue générale en tant que nom, gagne un nouveau sens verbal dans la langue spécialisée : *to network*, avec toutes les formes fléchies associées : *networks*, *networked*, *networking*. Le phénomène inverse est encore plus courant : de nombreux verbes deviennent des noms désignant l'action de ces verbes. Ainsi l'on obtient *an interrupt*, *a merge*, *a reset*,

(9) Ce nombre se réduit à 417, si l'on compte une seule fois les différentes occurrences et versions orthographiques du même terme.

(10) Pour une discussion sur la conversion en anglais consulter e.g. Bauer (1983).

an assert, qui peuvent aussi se mettre au pluriel, ce qui peut être ambigu par rapport à la troisième personne du singulier des mêmes verbes.

Pour évaluer le taux de précision de l'extraction, il faut comparer le nombre de candidats-termes corrects avec celui de tous les candidats proposés. Puisque l'utilisateur ne consultera chaque candidat qu'une seule fois, nous faisons ce calcul, contrairement à celui du rappel, sur les listes sans doublons. Le nombre de toutes les séquences uniques extraites par le patron est égal à 644 (100 séquences proviennent du *DELACF* spécialisé). Dans cet ensemble, 339 candidats (dont 87 entrées du *DELACF*) sont pertinents, donc le taux de précision est égal à 53%. L'utilisateur retiendra donc à peu près 1 candidat sur 2, ce qui nous semble raisonnable pour le travail de constitution du glossaire de texte à traduire.

6 Aspects novateurs

L'originalité de notre méthode n'est pas dans les algorithmes employés, car:

• La recherche de patrons dans un texte étiqueté est une technique souvent appliquée dans la tâche d'extraction (par exemple chez Daille 1994 et Auger *et al.* 1996 en français, ou chez Justeson et Katz 1995 en anglais).

• La méthodologie de construction et d'utilisation des dictionnaires électroniques est celle employée au LADL (voir Courtois et Silberztein 1990).

• Les principaux programmes informatiques ont été repris du système *Intex*.

• L'analyse lexicale du texte n'effectue qu'un minimum de désambiguïsation des mots.

Le point fort principal et l'originalité de notre approche est dans le fait de fournir à ces algorithmes des données de très haute qualité et complétude. Nous avons récupéré les résultats des travaux d'experts en lexicographie, terminologie et traduction. Leurs dictionnaires, généraux et spécialisés, étant l'effet de l' «extraction» humaine, sont de très bonne qualité du point de vue de la pertinence des mots et séquences qu'ils contiennent. De plus, nous nous sommes penchés sur la préparation de ces ressources, nécessaire pour le traitement automatique: la correction orthographique, le marquage des catégories et des traits flexionnels, la génération des formes fléchies, etc. Ainsi, nous disposons d'un noyau lexical très fiable que nous pouvons ensuite enrichir par une méthode automatique, standard du point de vue algorithmique, mais originale et efficace grâce à la qualité des ressources.

Les autres aspects novateurs de notre méthode sont à voir dans les points suivants:

1. Application de l'extraction dans le domaine de traduction assistée par ordinateur, qui présente des caractéristiques et exigences particulières, telles que:

• La nécessité de traiter des textes de tailles très variées, et rarement aussi importantes que les corpus auxquels sont traditionnellement appliqués les outils d'extraction. Cette contrainte exclut l'application efficace de toute méthode d'extraction qui comprend des calculs statistiques, telles que Daille (1994), Justeson et Katz (1995), Nakagawa (1998) et autres, dont un panorama est présenté chez Jacquemin (1997), p. 24-29 et chez Daille (1994).

• L'importance pour la qualité de traduction d'un très bon rappel des termes extraits (la même condition est prise en compte par Ladouceur et Cochrane (1996), mais leur article ne précise malheureusement pas les algorithmes employés).

• La spécificité de la notion du terme valable (il ne doit pas obligatoirement avoir un statut terminologique établi – voir section 2).

2. La variété des ressources utilisées et de leurs rôles dans le processus d'extraction:

• Le dictionnaire des mots composés terminologiques (le *DELACF* spécialisé) sépare les séquences qui ont un statut terminologique déjà reconnu.

• Les dictionnaires des mots simples et composés terminologiques (le *Delaf* et le *DELACF* spécialisés) fournissent les étiquettes qui sont à la base du patron de recherche (trait *+Spec*).

• Les dictionnaires des mots composés généraux et terminologiques (le *DELACF* général et le *DELACF* spécialisé) permettent de traiter «en bloc» certaines séquences figées (i.e. nous ne cherchons pas de nouveaux termes à l'intérieur des mots composés connus).

• La complétude du dictionnaire des mots simples généraux (le *Delaf* général) permet de considérer les mots simples non reconnus comme néologismes du domaine traité et de les prendre en compte dans le patron de recherche.

3. L'utilisation d'un analyseur lexical qui tienne compte des mots composés. Cet aspect est absent ou très limité dans les étiqueteurs employés par les extracteurs existants.

4. L'hypothèse que le matériau lexical à l'intérieur d'un domaine est relativement stable par rapport à la croissance très importante de la terminologie. Ainsi, nous admettons que la création d'un nouveau terme se fait le plus souvent par une combinaison grammaticalement correcte de termes simples et composés déjà existants. Cette hypothèse est reflétée dans le patron de recherche utilisé et confirmée par les résultats des tests (elle apparaît aussi chez Nakagawa (1998), mais les

mots simples caractéristiques du domaine y sont recherchés non pas dans un dictionnaire mais dans le corpus par une méthode statistique).

7 Perspectives

La méthode d'extraction présentée ci-dessus n'est que le début de notre travail. Il nous reste à mettre en forme tous les dictionnaires *Lexpro*, comme nous le décrivons dans la section 3.2. Beaucoup de ces dictionnaires sont peu volumineux, et donc le nombre de termes simples, sur lesquels est fondée une grande partie du patron de recherche, peut s'avérer trop bas. Dans ce cas, nous pouvons entreprendre, avant de commencer l'extraction, une mesure supplémentaire d'auto-enrichissement de dictionnaires spécialisés par récupération des « termes simples significatifs » de chaque domaine, i.e. des substantifs, adjectifs, adverbes et participes apparaissant parmi les termes complexes déjà répertoriés. Ces nouvelles entrées, soumise ensuite à la flexion, peuvent compléter les *Delaf* existants.

Il sera aussi nécessaire d'ajouter, au cours de l'extraction, la lemmatisation des candidats-termes, afin de ne plus proposer la même séquence au singulier et au pluriel (e.g. *disk module* et *disk modules*) comme deux candidats indépendants. Remarquons que cette lemmatisation est à faire sur les termes complexes entiers, i.e. seuls leurs constituants caractéristiques doivent être lemmatisés, et non pas tous leurs composants, comme ceci est fait par Daille (1994).

Un problème important qui reste à résoudre est celui des ambiguïtés des mots simples et composés apparues suite à l'étiquetage du texte. Le rattachement à notre système d'un des étiqueteurs disponibles, par exemple de celui de Brill (1994), peut s'avérer difficile à cause de la spécificité des dictionnaires que nous utilisons, entre autres ceux des mots composés. Néanmoins, nous envisageons de tester cette possibilité pour augmenter la précision de notre logiciel.

Nous souhaitons aussi élaborer de nouveaux patrons de recherche. Premièrement, des nouvelles structures syntaxiques, entre autres celles contenant des prépositions, comme *arrangement of slots*, doivent faire l'objet d'une étude détaillée. Deuxièmement, nous voudrions élaborer des méthodes fondées sur l'observation que les listes de termes connus contiennent de nombreuses séries, e.a. *access control group*, *access control list*, *access control machine*, *access control profile*, *access control register*, etc. Il est probable qu'une séquence contenant le même affixe qu'une des séries recensées soit un bon candidat-terme.

Actuellement, notre logiciel ne traite que le texte pur. La prise en compte des formats enrichis, e.g. de l'emploi d'une police spéciale pour certaines parties du document, ainsi que le traitement des données textuelles inclues dans des tableaux, illustrations etc., sera un affinement important du point de vue d'un traducteur technique.

Finalement, l'adaptation du logiciel à d'autres langues de Lexpro est à envisager. Nous sommes consciente que ceci représente un travail considérable, car les patrons de recherche pour une langue seront rarement utilisables dans une autre langue. Nous espérons néanmoins que, le point fort de la méthode étant la taille et la qualité de nos dictionnaires généraux et terminologiques, nous allons pouvoir fournir des résultats intéressants pour les utilisateurs multilingues.

Conclusion

Nous voyons l'un des avantages importants de notre méthode d'extraction de termes dans le fait que ses résultats(11) ne dépendent pas de la taille des documents sur lesquels elle est effectuée. En effet, si l'on soumet à l'extraction seulement une partie d'un corpus, les candidats termes proposés seront exactement les mêmes que ceux qui dans la même partie ont été trouvés lors de l'extraction sur le corpus entier. Seules les fréquences des candidats (donc leur ordre de présentation) pourront varier, mais ceci n'affecte pas le contenu de la liste(12).

Si l'on dispose d'un dictionnaire couvrant précisément le domaine traité, et si la terminologie disponible dans ce dictionnaire est assez complète et de bonne qualité, les résultats de l'extraction seront riches. Nous espérons qu'avec la taille de *Lexpro* déjà très importante et toujours croissante, notre extracteur terminologique fera ses preuves.

Agata Chrobot,
Laboratoire d'automatique documentaire et linguistique,
Université Paris 7 et
LCI (Langage communication informatique),
Jouy-en-Josas,
France.

(11) Nous comprenons ici par résultat d'extraction la liste des candidats retenus par le logiciel avant toute intervention humaine.

(12)Nous avons mentionné que, dans le cadre de la traduction, il est important de ne pas négliger les termes d'une fréquence basse d'occurrences. Si néanmoins l'utilisateur choisit de ne valider que les candidats de fréquences élevées, son résultat final, i.e. la liste validée, dépendra évidemment de la taille du corpus.

Bibliographie

Auger (P.), Drouin (P.), Auger (A.), 1996 : « Filtact : un automate d'extraction des termes complexes », dans Grarson (M.), éd., dans *Terminologies nouvelles, Banques de terminologie, Actes de la table ronde, Québec, 18 et 19 janvier 1996*, n°15, Bruxelles, p. 48-51.

Bauer (L.), 1983 : *English Word-Formation*, Cambridge University Press.

Bourigault (D.), 1994 : *Lexter, un logiciel d'extraction de terminologie. Application à l'extraction des connaissances à partir de textes.* Thèse de doctorat en mathématiques, informatique appliquée aux sciences de l'homme, École des hautes études en sciences sociales, Paris.

Brill (E.), 1994 : Supervised part of speach tagger, http://www.cs.jhu.edu/~brill.

Chrobot (A.), 1999 : « Flexion automatique des mots composés », dans Lamiroy (B.), Klein (J.), Peirret (J.-M.), éd., dans *Cahiers de l'Institut de linguistique de Louvain. Actes du XVI^e colloque européen sur les lexiques et la grammaire comparés des langues romanes, Louvain-la-Neuve, septembre 1997*, Louvain-la-Neuve.

Courtois (B.), Silberztein (M.) éd., 1990 : *Dictionnaires électroniques, Langue française 87*, Larousse, Paris.

De Sollier (F.), 1998 : *Dictionnaire encyclopédique de l'informatique*, Paris, La Maison du dictionnaire.

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

Gouadec (D.), 1997 : *Terminologie et phraséologie pour traduire – le concordancier du traducteur*, Paris, La Maison du dictionnaire.

Hildebert, 1999 : *Dictionnaire des sciences de l'informatique*, Paris, La Maison du dictionnaire.

Jacquemin (Ch.), 1997 : *Variation terminologique : reconnaissance et acquisition automatique de termes et de leurs variantes en corpus*, habilitation à diriger des recherches en informatique, Irin, Université de Nantes.

Justeson (J.), Katz (S.), 1995 : « Technical terminology : some linguistic properties and an algorythm for identification in text », dans *Natural Language Engineering, 1(1)*, p. 9-27.

Ladouceur (J.), Cochrane (G.), 1996 : « Termplus, système d'extraction terminologique », dans *Terminologies nouvelles, Banques de terminologie, Actes de la table ronde, Québec, 18 et 19 janvier 1996, n°15*, Bruxelles, p. 52-56.

Nakagawa (H.), Mori (T.), 1998 : « Nested Collocation and Compound Noun For Term Extraction », dans *Preceedings of COPUTERM, the First Workshop on Computational Terminology, August 15*, 1988, University of Montreal.

Silberztein (M.) , 1997 : *INTEX 3.4. Reference Manual*, LADL, Université Paris VII, Paris.

Le modèle de représentation et de gestion hypertexte des concepts d'un domaine dans le système *CoDB-Web*

Pour élaborer des bases de concepts (BdC), nous adoptons une approche s'appuyant sur les concepts et techniques de la terminologie. Une BdC comporte trois niveaux d'abstraction: informations terminologiques, caractère définitoires et représentation. Sa mise en place est basée sur l'utilisation corrélée d'un SGBD relationnel et d'un système hypertexte. Pour son implémentation, nous avons développé le système *CoDB-Web* qui assure la gestion et l'exploitation hypertextuelle des concepts via le système hypermédia distribué *World Wide Web*. Une instanciation de BdC, *MedTrad*, a été réalisée pour capitaliser la connaissance africaine sur les plantes médicinales.

Termes clés:
base de concepts; terminologie; information terminologique; caractère définitoire; hyperdocument; médecine traditionnelle.

1 Introduction

Le projet «base de connaissance généralisée » sur lequel travaille notre équipe Ingénierie des bases de connaissances généralisées (ICOG) est né du constat de la convergence de trois domaines: bases de données (BD), bases de connaissances et hypertextes. Dans le cadre de ce projet, Nathalie Forest (1994) a proposé un modèle d'hyperdocument expert. La construction d'un tel hyperdocument expert nécessite une méthode et un outil d'aide à la structuration et à la représentation des concepts du domaine que l'on veut modéliser par une application hypertexte. Des modèles de structuration de concepts ont été proposés dans d'autres disciplines telles que: la terminotique avec D. Gouaderec (1992, 67) pour la description des concepts à l'aide de fiches terminologiques, l'ontologie avec C. Garcia (1996, 95) dans le sens restreint d'ensemble des types atomiques de concepts (« le vocabulaire de base »), les BD avec Michel Bonjour (1994, 263) pour l'intégration de schémas de bases de données, etc. Notre objectif est donc la conception d'un modèle de BdC (*cf.* M. Brou 1997) générique et réutilisable afin de fournir à toutes ces disciplines dont la structuration des concepts est un élément important, une méthode et un outil d'aide à la représentation et à la gestion des concepts d'un domaine donné. *«Nous définissons une BdC comme étant une base de données de concepts qui utilise des facilités hypertextuelles pour la gestion et l'exploitation des concepts d'un domaine».*

Dans cet article, nous proposons un modèle conceptuel pour la représentation et la gestion des concepts d'un domaine. Dans la partie 2, nous situons la notion de BdC par rapport aux travaux existants. Puis, dans la partie 3, nous décrivons le modèle de la BdC, en insistant plus particulièrement sur ses trois niveaux d'abstraction, sur son modèle hypertexte et sur son exploitation. La partie 4 aborde l'implémentation de la BdC en présentant le système *CoDB-Web* qui est le noyau d'intégration entre les différentes composantes de la BdC. À la partie 5, nous présentons une instanciation de la BdC à travers l'application *Medtrad* qui est une BdC sur la médecine traditionnelle africaine. La conclusion fait le bilan du travail que nous avons réalisé.

2 Critiques de l'existant

Le but poursuivi par la définition de la notion de BdC est à rapprocher avec celui des bases de connaissances terminologiques (BCT), à savoir l'utilisation de la terminologie dans les systèmes à base de connaissances (SBC), notamment dans l'acquisition de connaissance. Tout comme la BCT, la BdC est différente d'une base de données terminologiques (BDT). En effet, une BDT contient uniquement des termes associés à quelques informations linguistiques, spécifiques à un domaine donné. Comme le fait remarquer Nathalie

Aussenac (1995), ces BDT ne conviennent pas pour les SBC à cause de la pauvreté de la description conceptuelle, limitée à la définition des concepts donnée en langage naturel.

Une BCT intègre à la fois une base terminologique conceptuellement et sémantiquement structurée et une base de connaissance. Elles contiennent des informations conceptuelles que peuvent utiliser les cogniticiens pour la construction du modèle conceptuel d'un SBC. De nombreux outils ont été développés dans le cadre des BCT. Parmi ces outils on peut citer:

• *Lexter*, Bourigault (1994): pour l'analyse d'un corpus de texte afin de dériver une liste de termes candidats;
• *Terminae*, Brigitte Biebow (1997): basé sur l'outil d'analyse Lexter, il permet de construire des fiches terminologiques, de garder des traces des choix de modélisation et de gérer une ontologie;
• *Macao* (Aussenac 1988): la modélisation conceptuelle du domaine dans Macao est à rapprocher d'un réseau de concept de BDC; de plus, la méthode préconise une analye de textes en amont pour y repérer des termes utilisés ensuite pour définir les concepts du domaine;
• *Code* (D. Skuce 1993): est un outil complet de gestion de la connaissances. En plus de l'analyse des textes, il permet de créer un lexique de termes, de stocker, rechercher et extraire de la connaissance conceptuelle dans ces textes.

L'outil que nous avons développé pour construire une BdC et gérer efficacement les termes et les concepts d'un domaine s'appelle *CoDB-Web*. Il s'appuie à la fois sur les bases de données (qui assurent la gestion et la recherche d'informations) et les hypertextes (qui offrent une interface ergonomique d'accès aux concepts par association d'idées). Couplé à un outil d'analyse de texte, on peut le comparer à *Terminae*; dans la gestion et la manipulation de concepts, on peut le comparer à *Macao* et à *Code*. Cependant, la BdC est caractérisée par trois niveaux d'abstraction (terminologique, définitoire et de représentation) dans la représentation des connaissances. Elle permet de construire et de consulter des réseaux conceptuels modélisant un domaine en relation avec des textes où sont décrites les informations de ces réseaux. Elle cherche à différencier la représentation des connaissances terminologiques, conceptuelles et leur implémentation; de plus, elle facilite le retour au texte d'où sont extraites ces informations.

3 Le modèle conceptuel de la BdC

La terminologie est la science qui étudie les termes d'un domaine en précisant les relations qui existent entre ces différents termes. Comme le fait remarquer M. Van Campenhoudt (1999), elle est fondée sur un modèle tripartite dont les mots clés sont: le concept, l'objet et le terme. La figure 2.1 montre les relations qui existent entre ces trois entités à l'intérieur d'un domaine.

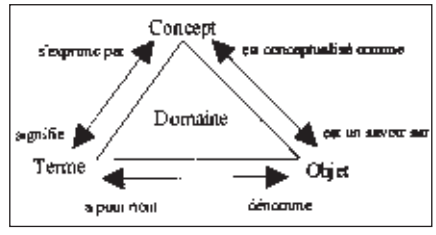


Figure 2.1
Relations entre concept, objet et terme.

• Domaine: «un domaine est un ensemble de concepts reliés les uns aux autres par des relations sémantiques».
• Concept: «un concept est une idée abstraite donc générale pour définir un ensemble d'objets ayant des caractères ou propriétés communs». Par exemple le concept *plante* est une notion générale qui convient pour désigner toutes les plantes. Un concept peut être élémentaire ou complexe.
• Objet: «un objet est un élément de réalité qui peut être perçu ou conçu». Il peut être matériel (exemple: plante) ou immatériel (exemple: *maladie*). Un objet est une instanciation d'un concept, c'est la représentation concrète d'un concept. Par analogie au modèle objet, il peut être comparé à un objet d'une classe, c'est-à-dire une instance de classe.
• Terme: «un terme désigne au moyen d'une unité linguistique un concept défini dans une langue de spécialité». Il peut être constitué de un ou plusieurs mots. Exemple: *plante* et *plante ligneuse* sont deux termes qui désignent deux concepts différents.
• Caractère: «un caractère est une propriété d'un concept; il peut être assimilé à un attribut d'une relation du modèle relationnel ou à un attribut d'un objet du modèle objet». Par exemple le concept *plante* peut avoir les caractères suivants: famille, durée de vie, feuille.

Une BdC est donc constituée de trois niveaux d'abstraction qui sont: terminologique, caractère, et représentation. Le « niveau information terminologique » contient des informations de nature textuelle ou graphique permettant de prendre connaissance de la sémantique des concepts. Ces informations sont très peu adaptées pour effectuer des opérations de recherche pertinente sur les concepts, d'où le deuxième «niveau caractère». Quant au «niveau représentation», il

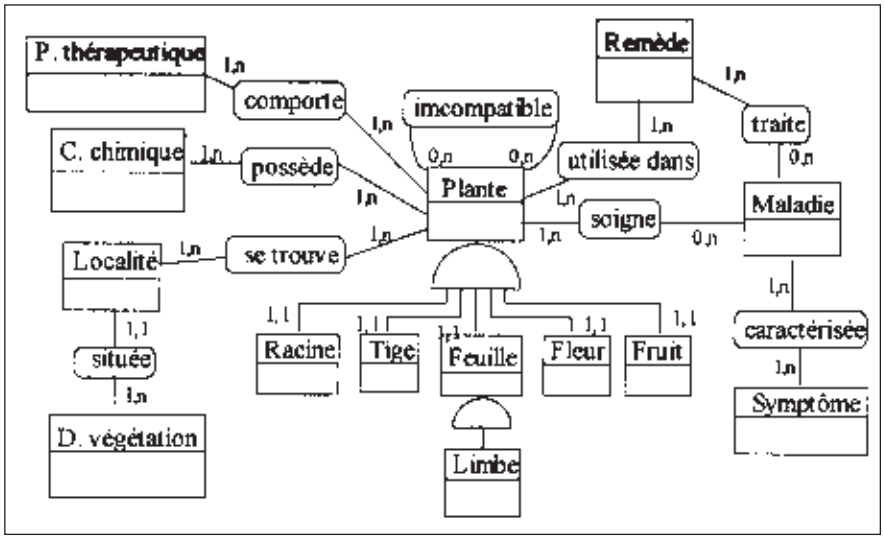


Figure 3.2
Schéma conceptuel du domaine.

indique comment sont représentées et gérées les informations de la BdC.

Pour étayer nos propos, nous utiliserons les concepts du domaine de la phytothérapie, c'est-à-dire le traitement des maladies par les plantes (Figure 2.2).

La figure 3.2 illustre un schéma entité/association partiel simplifié des concepts manipulés dans ce domaine. Une plante est composée d'une racine, d'une tige, d'un type de feuille, d'un type de fleur et d'un type de fruit. Elle possède une composition chimique, elle peut soigner plusieurs maladies, elle peut être utilisée dans plusieurs remèdes. Un remède peut traiter plusieurs maladies.

3.1 Le niveau de l'information terminologique

Le niveau terminologique permet une représentation structurée mais non formelle des connaissances; la structure de concept, à ce niveau, contient des connaissances linguistiques relatives aux termes qui le désignent, et des connaissances conceptuelles qui le situent dans une hiérarchie. On utilise la notion de fiche terminologique pour décrire les concepts à l'aide d'une structure commune. Chaque champ de cette structure est appelé information terminologique (IT), il peut contenir des informations textuelles ou graphiques permettant de prendre connaissance de la sémantique des concepts. Les différents champs de cette structure sont:

• Terme: un terme est la dénomination du concept, par exemple *plante*;

• Langue: dans un domaine, une langue est choisie et est utilisée par les experts du domaine afin d'avoir la même interprétation des concepts qu'ils utilisent; par exemple le *latin* est la langue véhiculaire en botanique;

• Synonyme: désignation du même concept (dans une même langue) par des termes différents; par exemple, le concept *Acacia senegal* (une plante) a pour synonyme *Acacia verek* en latin;

• Équivalent: désignation du même concept dans une autre langue; par exemple le concept *Acacia senegal* a pour équivalents *gommier blanc* en français et *ngobup uki* en serer (une langue du Sénégal);

• Homonyme: à un concept peuvent être associés un ou plusieurs homonymes c'est-à-dire des concepts ayant des termes identiques mais des valeurs de référenciation différentes. Cette IT permet d'éviter les erreurs d'interprétation dues à la polysémie de certains termes. Par exemple, en botanique, le concept *oignon* désigne une plante potagère à bulbe comestible, il désigne également le bulbe de certaines plantes telles que le lis et la tulipe;

• Concept générique: à un concept peut être associé un concept plus général appelé concept générique. Cette IT permet de traduire la notion d'héritage entre concepts. Le concept *plante ligneuse* a pour concept

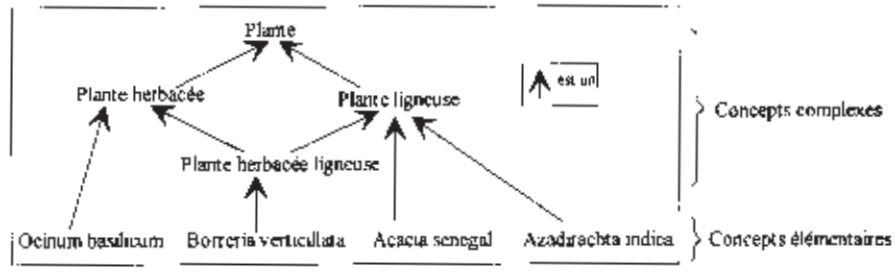


Figure 3.3
Structuration hiérarchique du concept *plante*.

générique le concept *plante* (Fig. 2.3);
• Sous-concept: de même qu'à un concept peut être associé un concept générique, à un concept générique va être associé ses concepts spécifiques ou sous-concepts. Ces deux notions concept génériques et sous-concepts permettent de modéliser le domaine en hiérarchie de concepts (Fig. 3.3);
• Concept lié: cette IT permet de traduire toutes les relations sémantiques qui existent entre les concepts autres que la synonymie, l'homonymie et la hiérarchie. Elle peut permettre de décrire une relation d'association (par exemple *soigne, traite, est utilisée*, Fig. 3.2), d'antonymie, etc. entre concepts;
• Définition: la définition fournit la signification du terme qui désigne le concept; elle est donnée sous forme textuelle et peut contenir des notions faisant référence à d'autres informations de la BdC. Par exemple, dans la définition du concept *Acacia senegal* ci-dessous, les mots soulignés font référence à d'autres concepts du domaine (ex. *composées bipennées* fait référence au concept *feuille*).
[1] « L'Acacia senegal est un petit arbre de 4 à 7 m de haut, à épines courtes et courbes, longues de 4 à 7 mm, réunies par trois à la base du pétiole. Les feuilles sont composées bipennées...».
• Contexte: quant au contexte, il indique les conditions d'utilisation d'un concept. Il est de la même nature que l'information terminologique définition;
• Image: un concept peut être illustré par une image; cette image peut permettre d'appréhender d'autres concepts de la BdC. Il est donc intéressant d'avoir une image réactive afin d'accéder aisément à la sémantique de tous ces concepts (*cf.* annexe Fig. 6.2).

Le niveau information terminologique contient des informations de nature textuelle ou graphique permettant de prendre connaissance de la sémantique des concepts. Ces informations sont très peu adaptées pour effectuer des opérations de recherche pertinente sur les concepts, d'où le deuxième niveau d'abstraction de la BdC que nous détaillons à présent.

3.2 Le niveau du caractère définitoire

À un concept est associé un ensemble de caractères qui sont les propriétés de leurs instances. Ce niveau permet de décrire l'ensemble de ces caractères spécifiques communs à un groupe de concepts. On peut assimiler un caractère ou caractère définitoire (CD) à la notion d'attribut d'une relation en BD. D'après D. Gouaderec (1992: 67), le terme définitoire est utilisé en terminologie pour indiquer qu'un terme contient des éléments de définition; ceci est le cas des caractères d'un concept parce qu'ils sont contenus dans leur définition. Considérons les CD de *plante* regroupés dans la table suivante:

plante	Famille	durée de vie	racine	tige	feuille	fleur	fruit
Acacia senegal	Mimosaceae	Pérenne	pivotante	aérienne	composée bipennée	Complète	sec déhiscent
Azadirachta indica	Méliaceae	Pérenne	pivotante	aérienne	composée paripennée	Complète	charnu

La première ligne contient les CD de *plante*, les autres contiennent une instanciation de ces CD pour deux instances du concept *plante*: *Acacia senegal* et *Azadirachta indica*. Les tuples de certaines tables relationnelles peuvent être également des CD des concepts; par exemple la table composition chimique suivante contient des tuples qui sont des CD des instances du concept *plante*.

Composition chimique

partie utilisée	élément chimique	quantité
Tronc (gomme arabique)	eau	10 à 15 %
Tronc (écorce) Fruit	tanin vitamine	28 % trace

Les CD peuvent être élémentaires (famille durée de vie) ou complexes c'est-à-dire composés aussi d'autres CD (racine, composition chimique). Ils sont utilisés dans les opérations de recherche d'information dans la BdC.

Une BdC doit pouvoir répondre aux interrogations suivantes: sur la sémantique d'un CD (Qu'est ce qu'une famille?), sur la valeur d'un CD (Qu'est ce que *Mimosaceae*?), sur les concepts (Quelles sont les plantes de la famille des *Mimosaceae*?).
Un modèle de requête simple, basé uniquement sur les informations terminologiques et les caractères définitoires a été proposé par Marcellin Brou (1997) pour faciliter la recherche d'informations dans la BdC.

3.3 Le niveau de la représentation

Le troisième niveau indique le moyen utilisé pour la représentation et la gestion efficace des informations qui sont manipulées dans la BdC.

On peut aborder la représentation des informations de la BdC selon trois approches: hypertexte, base de données, couplage BD/hypertexte.

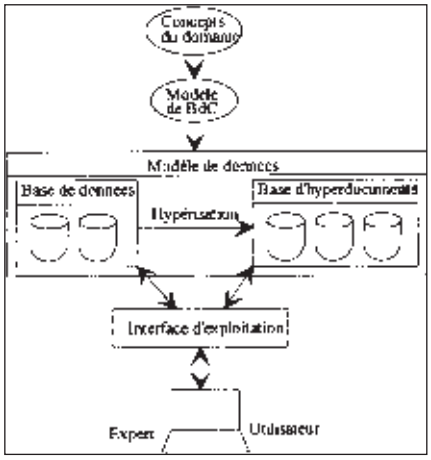


Figure 3.4
Architecture générale d'une BdC.

La solution que nous avons retenue est fondée sur la troisième approche c'est-à-dire une BD relationnelle intégrant des facilités hypertextes qui s'inspire des solutions propriétaires O2 Web (1996), Oracle (1996) et Internet Information Server (1996). Cette approche propose non seulement une interface d'accès aux informations d'une BD, mais surtout un modèle de représentation et de gestion des informations de type hypertexte stockées dans une BD.

En effet, dans la BD nous stockons les informations de type hypertexte non pas sous forme d'un hypertexte déjà créé, mais plutôt leurs deux composantes: contenu informatif et structure logique. L'hypertexte lui-même est généré par un processus d'hypérisation (fig. 3.4). D'après Mark Frisse (1988, 247) l'hypérisation est le travail qui permet de transcrire un document textuel linéaire en un réseau de nœuds hypertextes.

Cette approche facilite la maintenance de la base d'hyperdocuments (BHD), notamment dans les opérations de vérification de sa cohérence (*cf.* § 3.5). Elle permet également de créer un point de vue (vue partielle de la BdC modélisée par la liste des informations terminologiques et des caractères définitoires) qui permet de dériver une partie de la BHD contenant les informations qu'un utilisateur maîtrise le mieux et qu'il souhaite manipuler. Cette notion de point de vue permet de limiter la surcharge cognitive de l'utilisateur, qui est une source de désorientation, c'est-à-dire la perte de vue du but de sa recherche.

3.4 Le modèle conceptuel de données de la BdC

Le schéma de la figure 3.5 présente dans le formalisme Entité/Association, le modèle conceptuel de données de la BdC qui est constitué de deux parties: la partie de droite modélise le contenu informatif de la BdC et la partie de gauche modélise sa structure logique.

• L'aspect informatif: les IT de type attribut (*terme* et *langue*) et de type hypertexte (*définition*, *contexte*, *image*) sont modélisées par l'entité concept. Les IT de type relation conceptuelle (*synonyme*, *homonyme*, *concept générique*, *sous-concept*, *concept lié*) par l'association «Relation_conceptuelle». Les CD sont modélisés par les entités Cd_simple et Cd_type_relation. L'entité description permet de modéliser certaines informations du domaine qui sont référencées par les IT de type hypertexte ou par les CD.

• L'aspect logique: un concept peut être stocké dans au moins une unité informative (UI) et une UI peut être composée d'autres UI; l'association «est_composée » permet de modéliser la structure hiérarchique de la BHD. Une UI peut contenir des liens de référence; ces liens de référence sont modélisés par l'association lien. Un lien possède une ancre source et une ancre destination qui sont des UI. Un

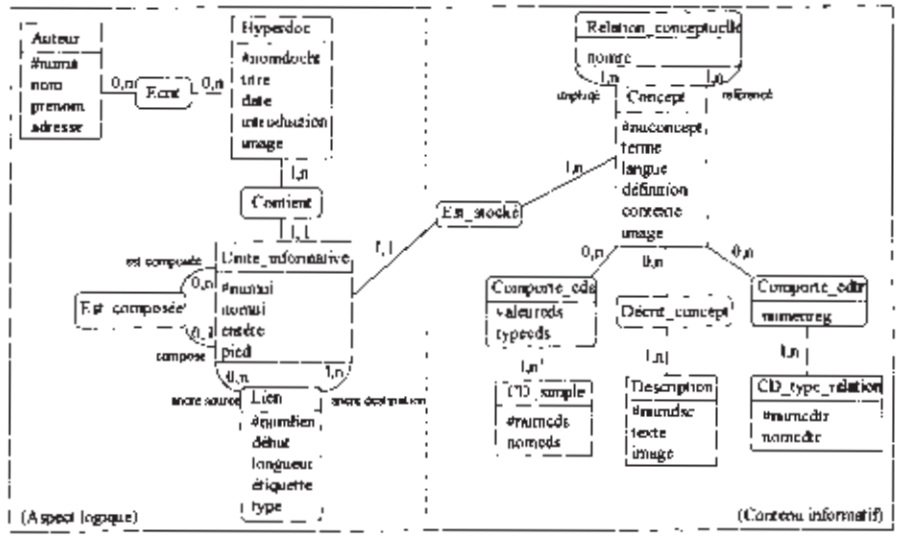


Figure 3.5
Schéma conceptuel de données de la BdC.

hyperdocument contient au moins une UI, et il peut être l'œuvre de plusieurs auteurs.

3.5 Le modèle hypertexte de la BdC

Le modèle hypertexte de la BdC s'appuie sur le modèle d'hyperdocument expert défini par Nathalie Forest (1994) et Renaud Deschamps (1995). Dans ce modèle, les hyperdocuments sont décomposés en unités informatives reliées les unes aux autres par des liens hiérarchiques ou de référence.

• Les unités informatives (UI): une UI est une extension des nœuds hypertextes classiques par l'introduction d'un ensemble de connaissances supplémentaires destinées à intégrer un aspect dynamique à leur contenu informatif; on peut la comparer aux documents actifs de Jocelyne Nanard (1993). Dans le cadre de la BdC, nous avons défini six types d'UI: UI du domaine, UI concept complexe, UI Concept élémentaire, UI informations terminologiques, UI caractères définitoires et UI description (*cf*. fig. 4.1).

• Les liens: on distingue deux sortes de liens: les liens hiérarchiques et les liens de référence.

Les liens hiérarchiques sont déduits des arbres conceptuels du domaine. Ils ont pour ancre les IT concept générique et sous-concept. Pratiquement, ils sont modélisés par la relation Unité_informative dont nous présentons une instanciation comme suit:

Unité informative

numui	nomui	entête	numui2	nomdocht	numconcept
1	C1it	IT Plante		Medtrad	1
2	C5it	IT Acacia senegal	1	Medtrad	5
3	C5d	Définition Acacia senegal	2	Medtrad	5
5	C15i	Image Feuille		Medtrad	15
6	C5d1	Description Inflorescence	4	Medtrad	5
7	C5d2	Description épi cylindrique	4	Medtrad	5
8	C15d7	Description sommet	5	Medtrad	12

Où *numui2* est le numéro d'une UI dont l'UI de numéro *numui* et de nom *nomui* est une composante; entête est le titre de l'UI; nomdocht est le nom de la page d'accueil; numconcept est le numéro du concept qui est décrit par l'UI.

Les UI C5it et C5d sont des composantes de l'UI C1it, les UI C5d1 et C5d2 sont des composantes de l'UI C5d, elle-même composante de l'UI C5it. Cette structuration hiérarchique est réalisée grâce à l'attribut numui2.

Les liens de référence, dans la sont typés. On en distingue 5 types. Ces liens sont modélisés par la relation Lien, dont nous présentons une instanciation:

Pour chaque type de lien, les attributs *début* et *longueur* ont une interprétation particulière:

1) Lien dont l'ancre est située dans un hypertexte, par exemple le lien n° 1: l'ancre source est située dans l'UI C5d, cette ancre commence au 39e caractère du contenu informatif de l'IT définition et a une longueur de 15 caractères.

2) Lien dont l'ancre est un CD ou la valeur d'un CD (lien n° 3 et 4). L'attribut *début* indique le rang du CD qui est l'ancre du lien. L'attribut *longueur* permet de déterminer si c'est le CD lui-même qui est l'ancre du lien (longueur = 0) ou si c'est sa valeur (longueur = 1).

Lien

numlien	début	longueur	étiquette	type	numui	numui2
1	39	15	Caractère définitoire	définition	5	3
2	368	17		définition	5	6
3	6	0		cds	3	6
4	6	1		cds	3	7
5	360	170		image	4	8

3) Lien dont l'ancre est située dans une zone réactive circulaire dont le centre a pour abscisse *début* et pour ordonnée *longueur* (lien n° 5).

• Maintenance de la BHD: après les opérations de mise à jour de la BHD, elle doit rester cohérente au niveau de sa structure logique et au niveau de ses liens de référence.

Dans les hypertextes classiques, il n'y a pas de séparation entre leur contenu informatif et leur structure logique (cas des documents HTML), ceci rend leur maintenance difficile. Considérons l'hyperdocument HTML ci-dessous. Ce document contient deux liens, le premier fait référence à la partie caractères définitoires simples du document *C5cd.htm*. L'ancre de ce lien est *4 à 7 m de haut*, son activation permet d'accéder à la partie *caractère définitoire simple* du document *C5cd.htm*. Supposons que l'on supprime cette partie, pour que la BHD reste cohérente, il faut supprimer manuellement tous les liens des hyperdocuments qui font référence à cette partie.

<H1> Définition Acacia senegal </H1>
L'Acacia senegal est un petit arbre de <A HREF="C5cd.htm #Caractères définitoires simples"> 4 à 7 m de haut </A>, à épines courtes et courbes, longues de 4 à 7 mm, réunies par trois à la base du pétiole. Les feuilles sont <A HREF= "C15it.htm"> composées bipennées </A>...

Dans la BdC, la maintenance est aisée. En effet, les opérations de mise à jour sont faites automatiquement de façon relationnelle dans les tables et le processus d'hypérisation la répercute sur la BHD.

Considérons l'UI U_{ij} contenant des liens qui pointent dans le vide (qui n'ont pas d'UI destination). Dans la BdC, la vérification de cohérence de la BHD au niveau des liens de référence est effectuée en trois étapes:

1) la première étape consiste à rechercher les liens théoriques de cette UI dans la table lien grâce à une requête de type SQL comme suit:

```
SELECT numui, numui2, étiquette,
début, longueur INTO
Lien_théorique
FROM lien
WHERE numui = j
```

Lien_théorique

numui	numui2	étiquette	début	longueur
4	1	Caractère définitoire simple	40	15
4	2		167	19

2) Ensuite, grâce à un programme, nous recherchons les liens effectifs qui existent dans le code hypertexte de l'UI Uij.

Lien_effectif

UIS	UID	étiquette	ancre
C5d	C5cd.htm	Caractère définitoire simple	4 à 7 m de haut
C5d	C15it.htm		Composées bipennées

3) Enfin, par la comparaison entre les éléments de ces deux tables (numui, UIS), (numui2, UID), (début, longueur, ancre) nous déterminons les causes de cette incohérence.

4 Implémentation de la BdC

4.1 Le système *CoDB-Web*

Le schéma de la figure 3.1 présente le système *CoDB-Web* (Concept DataBase Web) qui est le système de gestion et de représentation hypertexte de la BdC fondé sur le *World Wide Web*. C'est une interface particulière entre le SGBD qui gère la BdC et le système hypertexte qui gère la BHD. Il est constitué de quatre modules interconnectés.

Le cadre de développement de la BdC est celui d'une architecture client-serveur qui respecte les protocoles de l'Internet (adressage HTTP, langage HTML…), moyen idéal de diffusion de la connaissance stockée dans la BdC.

Le système hypertexte est donc le système hypermédia W3 qui offre une interface graphique d'accès aux ressources d'Internet. Les documents hypertextes qu'il manipule sont décrit à l'aide du langage HTML.

L'avantage de W3 est que son formalisme est bien défini et normalisé, il est mondial et disponible sur toutes les plates-formes (environnement *Windows* ou *Unix*). On peut noter cependant l'absence d'outils de conception et de maintenance des hyperdocuments, d'où notre proposition de la notion de point de vue et du système *CoDB-Web* pour gérer efficacement la BdC.

Quant au système de gestion de base de données, nous avons choisi le SGBDR *Access* qui permet de développer rapidement des applications BD dans l'environnement graphique *Windows*.

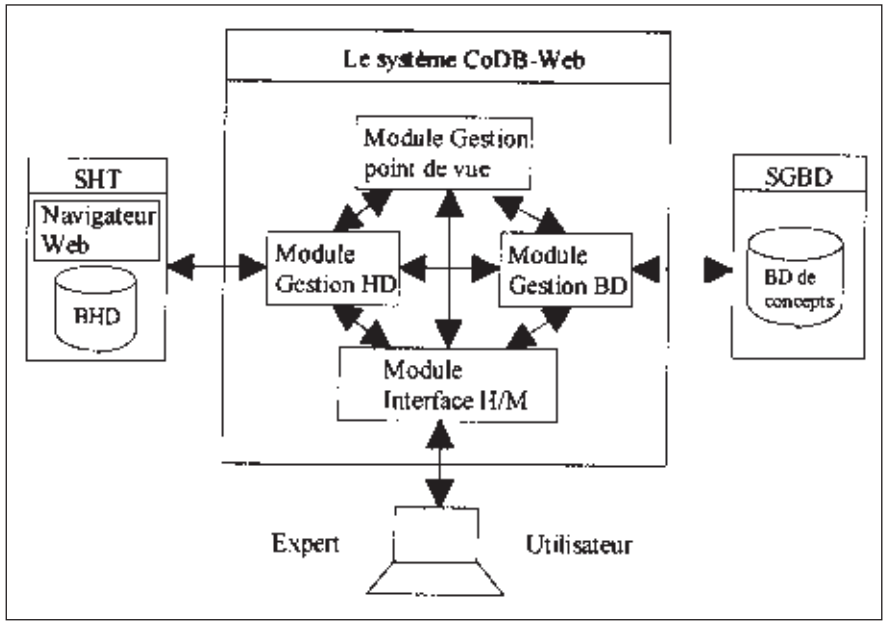


Figure 3.1
Schéma de fonctionnement général de la BdC.

Entre le système hypertexte, le SGBD et le système *CoDB-Web*, il fallait trouver un langage fédérateur. Nous avons choisi de ce fait comme langage d'implémentation de la BdC le langage *JavaScript Intrabuilder* qui est une version enrichie de *JavaScript*, lui-même une implémentation simple de *Java*. *JavaScript* Intrabuilder permet d'intégrer un aspect dynamique aux pages Web. Sa principale qualité est son interface d'accès aux BD localisées sur les serveurs. Il faut signaler cependant sa lenteur due au fait qu'il est interprété.

4.2 Méthodologie d'instanciation de la BdC

Pour la construction d'une BdC, nous avons proposé une méthodologie comportant six étapes:

1) Délimitation du domaine: elle consiste à cerner l'ensemble des connaissances du domaine à étudier afin d'isoler les concepts du domaine;

2) Collecte d'informations: il s'agit de rechercher les documents de référence du domaine étudié tels que les dictionnaires, les règlements, les livres, etc. Ces documents permettront d'isoler les concepts clés du domaine et fourniront des renseignements utiles à la détermination des valeurs des informations terminologiques et des caractères définitoires;

3) Modélisation du domaine sous forme d'arbres conceptuels: il s'agit d'organiser les concepts du domaine de façon arborescente afin de faire apparaître les relations hiérarchiques qui existent entre les concepts. Ces relations permettront de trouver les concepts génériques.

4) Modélisation de l'application sous forme d'un schéma conceptuel: pour la mise en place de ce schéma, nous préconisons l'utilisation du modèle entité/association pour la simplicité de son formalisme; dans ce formalisme, nous représentons chaque concept par une entité, et la relation sémantique entre deux concepts est traduite par une association. Le schéma conceptuel aide à la détermination des caractères définitoires des concepts (attributs des entités) et les concepts liés (entités liées par certaines associations);

5) Établissement des fiches terminologiques: une fiche représente un concept à l'aide d'un ensemble de champs textuels ou graphiques contenant les valeurs propres au concept. Il s'agit de déterminer dans un premier temps le type des informations du domaine qui vont permettre d'alimenter les champs de ces fiches, et dans un deuxième temps d'instancier ces fiches pour chaque concept. Il existe deux sortes de fiches: des fiches « informations terminologiques » et des fiches «caractères définitoires». Ces fiches facilitent la saisie des données dans les tables relationnelles de la BdC.

6) Utilisation de la BdC: cette étape comporte 4 parties: la génération des tables relationnelles de la BdC réalisée par un programme *Access Basic*, l'instanciation de ces tables grâce à des fiches de saisie *Intrabuilder*, le lancement d'un programme *JavaScript Intrabuilder* pour générer la BHD. La dernière étape est l'utilisation proprement dite de la BdC selon deux modes d'accès: la navigation ou la recherche. L'utilisateur peut utiliser une version limitée de la BdC grâce à la version *Access* ou la version navigation qui est plus complète. Nous présentons en annexe quelques images écrans de la version navigation relatives à l'application *MedTrad*.

5 L'application *MedTrad*

MedTrad est une BdC sur la médecine traditionnelle africaine à base de plantes médicinales. La médecine traditionnelle occupe une place très importante dans les pays en

voie de développement et plus particulièrement en Afrique, mais malheureusement cette connaissance tend à disparaître d'une part à cause du modernisme, mais surtout à cause de la transmission de ce savoir qui est mal assurée.

La problématique de cette médecine traditionnelle a été abordée par Claude Frasson (1992) dans le cadre du système *Seiboga*. Ce système est basé sur une exploitation graphique nécessitant parfois une certaine compétence de l'utilisateur.

L'application *Medtrad* a aussi pour objectifs, d'une part, de capitaliser cette connaissance ancestrale africaine sur la phytothérapie qui présente des aspects rationnels et des résultats convaincants éprouvés par de nombreux chercheurs africains tels que Guy Maynart (1990), Sijelmassi (1993) et Assi (1996). D'autre part, elle doit apporter une aide aux tradipraticiens dans leur processus de diagnostic et de proposition d'une thérapeutique rigoureuse.

La figure 5.1 schématise la structuration de la base d'hyperdocuments de *MedTrad*. La page d'accueil (*UI domaine*) est composée d'une introduction qui décrit l'application et d'un sommaire constitué par la liste des concepts clés du domaine; ces concepts sont des ancres de liens permettant d'accéder à leurs unités informatives qui sont des *UI concept complexe*. Chaque concept est représenté par deux fiches terminologiques (UI informations terminologiques et UI caractères définitoires) comme celle du concept *Acacia senegal*.

L'application *MedTrad* permet l'apprentissage de la phytothérapie en se basant sur une interface hypertextuelle bien adaptée à une mémorisation à long terme et à une compréhension profonde (assimilation).

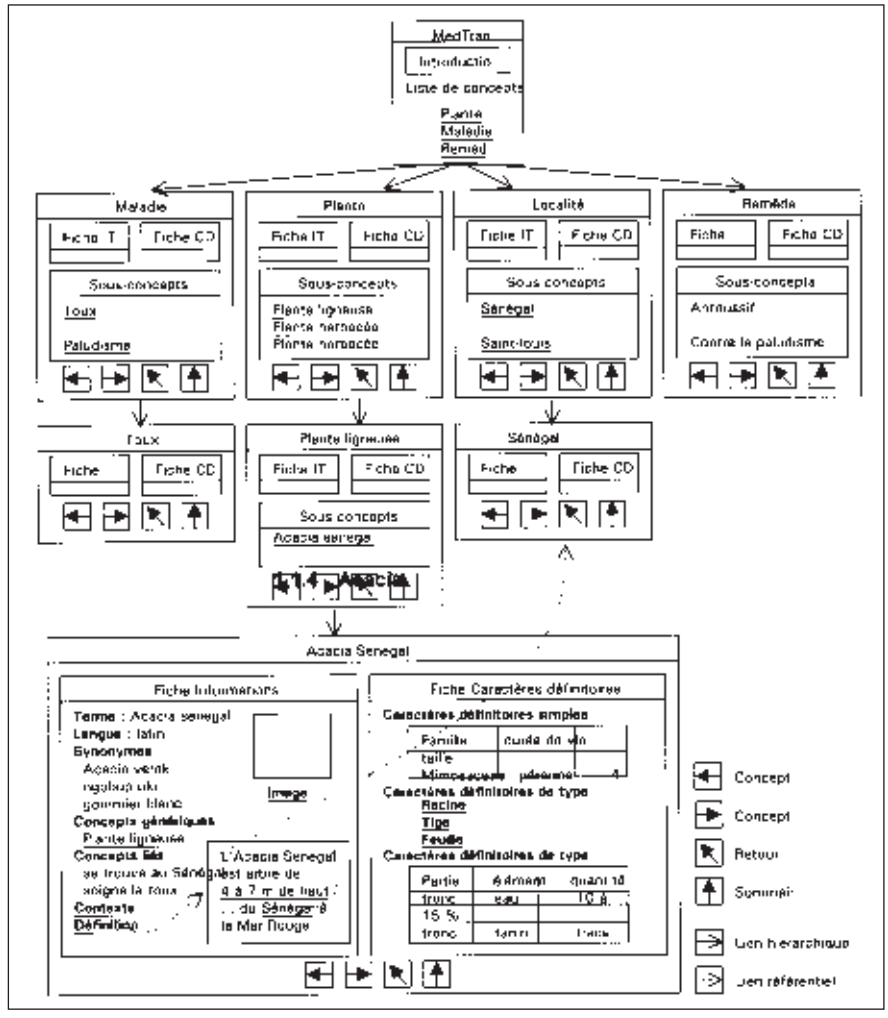


Figure 5.1
Base d'hyperdocuments de *MedTrad*.

6 Conclusion

L'objectif de cet article était de proposer un modèle de représentation et de gestion des concepts d'un domaine afin de construire des ontologies facilement accessibles et réutilisables. L'outil d'exploitation de ce modèle appelé base de concepts (BdC) est *CoDB-Web*, il a été développé en utilisant les bases de données, des techniques hypertextes et des résultats sur la représentation des connaissances. Nous avons présenté une utilisation possible de la BdC dans le cadre de l'acquisition de connaissance à travers l'application *MedTrad* dont le but est de capitaliser et de diffuser la connaissance ancestrale africaine sur les plantes médicinales. Il valide le premier point de notre objectif, à savoir la construction d'une ontologie facilement accessible. Quant au second point lié à la réutilisation d'ontologie, nous n'avons pas suffisamment d'expérience pour l'affirmer. Outre le développement

d'application, on peut citer d'autres cas d'utilisation de la BdC tels que : l'explication de la connaissance statique d'un SBC par le couplage de *CoDB-Web* avec un système expert, la création d'interface ergonomique d'accès à une base de données, l'aide au traducteur qui cherche à comprendre un texte dans un domaine dont il n'est pas le spécialiste…

L'instanciation de la BdC a travers *MedTrad* a soulevé les problèmes suivants :

• La détermination des termes et des concepts : elle est faite manuellement par l'expert du domaine, peut-on l'automatiser ?

• La classification des concepts : peut-on la faire automatiquement ;

• Le calcul de la position des ancres dans le contenu textuel des informations terminologiques de type hypertexte ;

• Le mode d'hypérisation : dynamique ou statique ?

Dans l'avenir, nous essaierons d'apporter des éléments de réponse à ces problèmes.

A. Hocine,
Département informatique
Université de Pau et des pays de l'Adour
France.

Konan Marcellin Brou,
Département mathématique-informatique
Institut national polytechnique Yamoussoukro,
Côte d'Ivoire.

Bibliographie

Aussenac (N.), Bourigault (D.), Condamines (A.) : « How Can Knowledge Acquisition Benefit from Terminology ? », dans *Proceedings of KAW95*, Banff (Can.), fév. 1995, http://www.irit.fr/ACTIVITES/EQ_SMI/PUBLI/banff95.html.

Assy (L.-A.), 1996 : *Plantes utilisées dans la médecine traditionnelle en Afrique de l'Ouest*, Édition Roche Basse, Suisse.

Biebow (B.), Szulman (S.), 1997 : « Avancée sur le concept de base de connaissance terminologique », 6e journée nationale du PRC-GDR intelligence artificielle.

Bonjour (M.), Falquet (G.), Léonard (M.), 1994 : « Base de concepts et intégration de bases de données », dans *actes du congrès INFORSID'94*, Aix-en-Provence, 17-20 mai 1994, p. 263-280.

Borland, 1996 : *Borland Intrabuilder pour Windows 95 & Windows NT*, manuel d'introduction, 1996.

Bourigault (D.), Lepine (P.), 1994, « Une méthode d'utilisation de Lexter en acquisition des connaissances », dans 5e Journées d'acquisition des connaissances, Strasbourg.

Brou (K.-M.), 1997 : *Base de concepts : Contribution à la représentation et à l'exploitation hypertexte de concepts – le système CoDB-Web*, thèse de doctorat de l'Université de Pau, novembre 1997.

Deschamps (R.), 1995 : *Bases de connaissances généralisées : une approche fondée sur un modèle hypertexte expert*, thèse de doctorat de l'Université Paul Sabatier de Toulouse, janvier 1995.

Forest (N.), 1994 : *Bases de connaissances généralisées : le modèle de gestion d'informations de types hypertexte* », thèse de doctorat de l'Université de Pau, janvier 94.

Frasson (C.), Houtsa (F.), Obenson (P.), 1992 : « Interface visuelle pour l'aide au diagnostic médical en médecine traditionnelle » dans *ICO*, vol. 4, n° 1 et 2, 1992, p. 17-25.

Frisse (M.), 1988 : « *From text to hypertext* », Byte, octobre 1988, p. 247-254.

Gouadec (D.), 1992 : « Terminologie et terminotique, outils, modèles et méthodes », actes de la première Université d'automne en terminologie, Rennes 2 - 21 au 26 sept 1992, Édition la Maison du dictionnaire, p. 67-120.

Garcia (C.), 1996 : « Construction coopérative d'ontologies dans un cadre de multi-expertise : ébauche d'un outil », dans JAC 96, Journée acquisition apprentissage, Sète 8-10 mai 1996, p. 95-107.

Internet Information Server, 1996 : *Documentation technique Windows NT Server 4.0*, 1996.

Maynart (G.), Lô (M.), Fortin (D.), 1990 : *Plantes médicinales du sahel*, Série études recherches, 1990.

Nanard (J.), Nanard (M.), Massotte (A.), Djemaa (A.), Joubert (A.), Betaille (H.), Chauché (J.), 1993 : « Integrating knowledge-based hypertext and database for task-oriented access documents », dans *actes DEXA'93*, p. 721-732.

O2 Technology, 1996 : *O2 web user Manual*, release 4.6, January 1996.

Oracle Corporation, 1996, *Guide des solutions micro et Internet*, 1996.

Sijelmassi (A.), 1996 : *Les plantes médicinales du Maroc*, Édition Le Fennec, 3e édition 1993.

Skuce (D.) : *CODE4 : A Unified Sytem for Managing Conceptual Knowledge*, document Web, http://www.csi.uottawa.ca/~ingrid/cogniterm.html.

Van Campenhoudt (M.) : *Abrégé de terminologie multilingue*, document Web, http://www.refer.fr/termisti/centre.htm#bref.

Annexe

La figure 6.1 présente la page d'accueil du système *MedTrad*; elle comporte une introduction qui pose la problématique de la médecine traditionnelle africaine et un sommaire constitué par les deux modes d'accès à *MedTrad* (la navigation et la recherche). La figure 6.2 présente les fiches terminologiques du concept *acacia senegal*. Cette fiche comporte deux parties: la partie gauche décrit les informations terminologiques, et celle de droite les caractères définitoires. La figure 6.3 présente l'image du concept *feuille* (image réactive); elle permet aussi de découvrir les concepts *limbe, nervure, pétiole et renflement.*

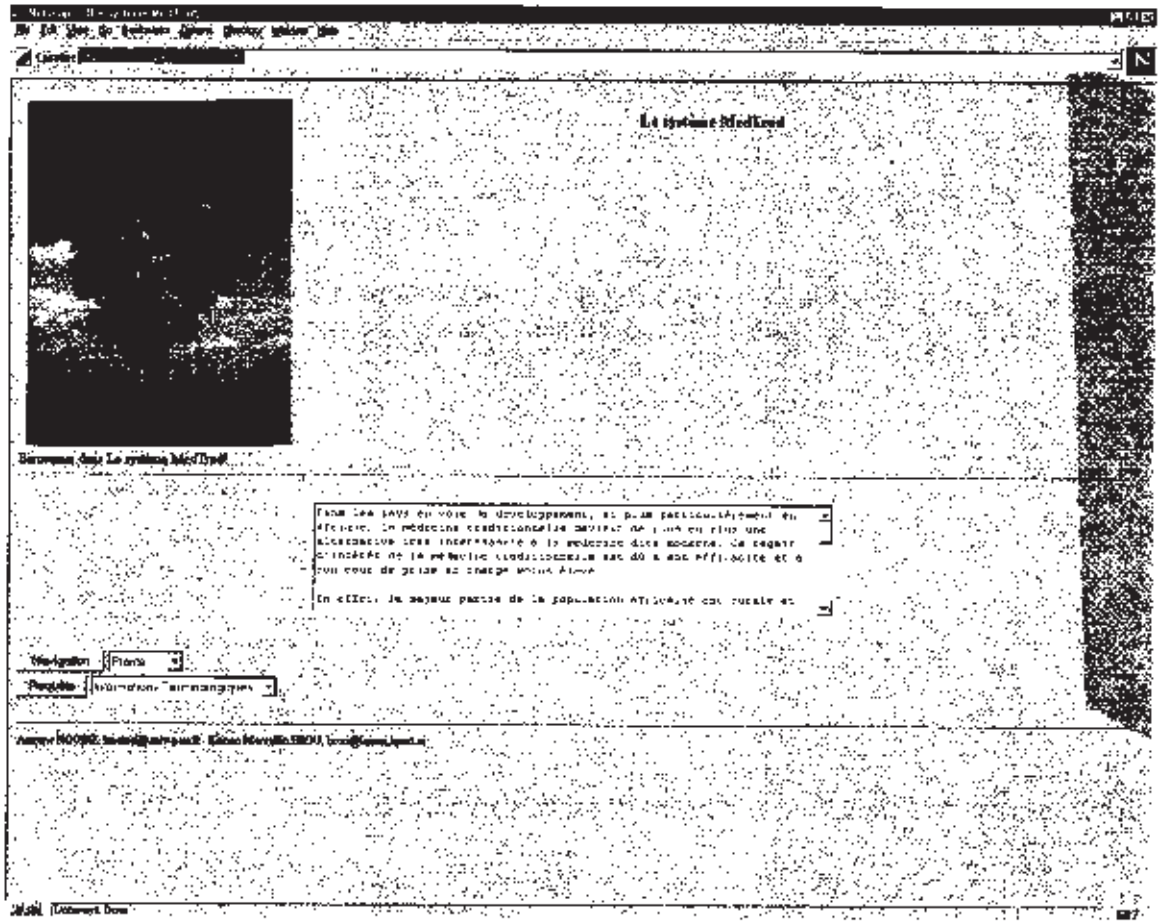


Figure 6.1
La page d'accueil du système *MedTrad*.

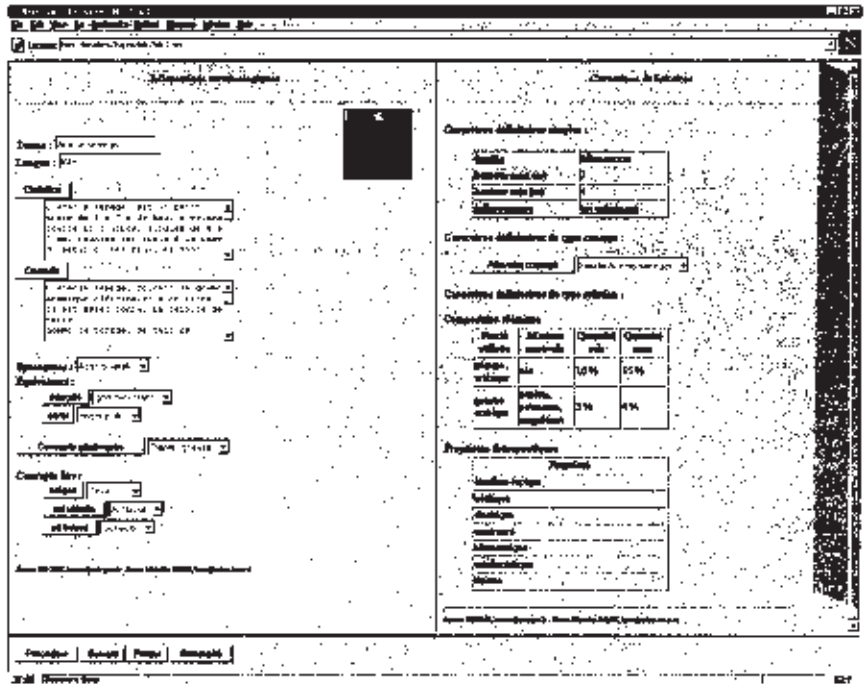


Figure 6.2
Fiches terminologiques du concept *Acacia senegal*.

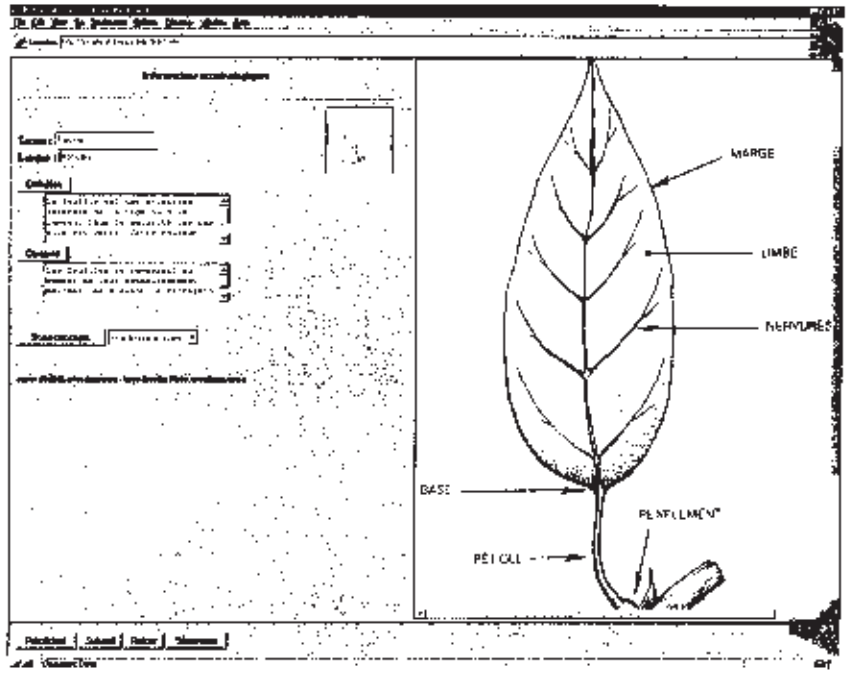


Figure 6.3
Activation d'un lien hypertexte image.

Ontologie et terminologie : le modèle *OK*

Les ontologies, comprises ici comme des vocabulaires communs sur des termes et leurs significations, constituent un axe de recherche primordial pour les sociétés de l'information. Dans ce cadre, elles doivent répondre à un double critère de consensus et de cohérence, condition *sine qua non* d'un modèle computationnel réellement exploitable. Cet article décrit un modèle d'ontologies par différentiation spécifique, héritier direct des arbres de Porphyre, et le langage de description associé ainsi que sa mise en œuvre à travers un environnement logiciel.

Termes-clés : terminologie ; ontologie ; différence spécifique ; langage de description ; représentation.

1 Problématique

Cette introduction a pour but, non pas tant de montrer l'intérêt des ontologies pour la représentation des signifiés, que de présenter les conditions qui ont présidé à une telle approche. Si l'on considère que les usages qu'une communauté fait de ses signes en déterminent le sens, il est important d'en expliciter les cadres d'utilisation. Ces cadres justifient les choix qui seront effectués. Ainsi, la notion d'inter-opérabilité, pris au sens de communication d'agents logiciels hétérogènes, qui n'a de sens que pour les sociétés de l'information, plaide ici pour une modélisation formelle des significations et de leurs relations.

1.1 Cadre général

Notre problématique se situe dans le cadre des sociétés de l'information, où un ensemble d'acteurs, qu'ils représentent des personnes, des organisations, des logiciels ou des machines, sont amenés à communiquer par l'intermédiaire de moyens informatiques.

Le paradigme *agent* permet la communication d'acteurs hétérogènes par leur encapsulation sous une forme normalisée. Les systèmes multi-agents distribués sur des réseaux informatiques communiquant par envois de messages, partageant et échangeant des connaissances et non plus seulement des données, constituent un modèle privilégié des sociétés de l'information. L'ingénierie simultanée basée sur une approche multi-agents (agent-based concurrent engineering), où chaque agent coopère et collabore à la conception et la fabrication d'un produit, qu'il soit logiciel ou manufacturé, est une bonne illustration de cette approche (Roche 1998). Citons à titre d'exemples les projets Shade (*SHAred Dependency Engineering*) (McGuire 1993) et PACT (Palo Alto Collaborative Testbed) (Cutkosky 1993) pour l'ingénierie simultanée ; ainsi que les projets Tove (*Toronto Virtual Enterprise*) (Fox 1992) et *Enterprise Project* (Stader 1996) pour la modélisation et l'intégration d'entreprise.

1.2 Les problèmes posés : communication, partage et échange de connaissances

Une société de l'information n'existe que s'il y a communication, partage et échange de connaissances entre ses acteurs. Elle peut faire sienne le leitmotiv de l'ingénierie collaborante : *«Product development is a knowledge and communication intensive process»* (Gruber 1992).

Cela nécessite l'emploi d'un langage de communication compréhensible pour tous les agents. L'utilisation d'un même langage de communication, dénommé *ACL* pour *Agent Communication Language*, permet de résoudre le problème syntaxique. Ces langages, tels que KQML et ses dérivés (Knowledge Query Manipulation Language (Labrou 1997)) sont basés sur les

actes de langages au sens des projets Darpa.

Le problème sémantique est repoussé à la charge des agents. Seul a été considéré le problème de la sémantique des termes échangés lors des communications : une communication ne pourra être prise en compte que si l'agent receveur en « comprend » les termes. C'est pourquoi la structure des phrases d'un ACL permet de préciser l'ontologie utilisée par l'émetteur d'un message.

1.3 Objectifs et limitations

Les premières réalisations d'architecture multi-agents ont mis en évidence l'importance du problème de la sémantique des termes échangés : il ne peut y avoir communication, et *a fortiori* collaboration et coopération, sans compréhension des mots qui composent le message. Les ontologies, comprises ici comme des vocabulaires communs sur des termes et leurs significations (Gruber 1992), constituent donc un axe de recherche primordial. Notons qu'en toute rigueur une ontologie n'est pas une terminologie, mais une représentation et une structuration particulières de connaissances conceptuelles. Cet abus de langage, en usage dans cette problématique et que l'on retrouvera dans cet article, peut s'expliquer par une préséance des connaissances conceptuelles et une acceptation de l'arbitraire du signe dans le domaine technique.

Notre travail porte sur la représentation de la signification des termes dénotant des connaissances conceptuelles. Nous nous limitons aux domaines techniques où le lexique est propre à un groupe social défini par une activité spécifique et où l'élaboration du sens d'un mot s'appuie sur l'élaboration d'une idée. S'écartant des problèmes que posent la langue naturelle, en espérant de pas tomber dans ceux d'une novlangue, nous avons résolument adopté une approche aristotélicienne de la définition de la signification des termes qui nous permet de garantir une cohérence forte de nos ontologies, condition *sine qua non* d'un modèle computationnel réellement exploitable. Le modèle *OK* obtenu est un héritier direct des arbres de Porphyre (Porphyre, Eco 1988, Rastier 1987).

Ce travail a abouti à une première version d'un environnement logiciel, la *OK Station*, pour *Ontological Knowledge Station*, dédié à l'acquisition, la représentation et l'exploitation de connaissances ontologiques.

2 Le Modèle *OK*

S'il existe d'ores et déjà plusieurs ontologies tant générales que spécialisées, certaines parfois considérables en nombre de termes définis : *Cyc*, *Mikrokosmos* (Mahesh), *Generalized Upper Model*, *Sowa's ontology, Tove* (Fox 1992), *Enterprise Ontology* (Stader 1996)… il est cependant difficile de trouver des informations sur leur réutilisabilité et leur compatibilité. Alors que les ontologies ont une visée normative, il est surprenant d'en constater les divergences conceptuelles.

Ainsi, dans le cadre des ontologies d'entreprise, comment concilier les définitions de Tove et celles d'*Enterprise Ontology* ? Prenons pour exemple le concept d'activité. *Enterprise Ontology* définit une activité comme étant décomposable en sous activités, réalisée par un exécutant et nécessitant des ressources. Elle hérite de la classe 'Activity-Or-Spec' définie en Ontolingua par la fonction :

```
(Define-Class Activity-Or-Spec (?X)
«The union of Activity and Activity-Spec»
:Iff-Def (And (Eo-Entity ?X) (Or
(Activity ?X) (Activity-Spec ?X)))
:Axiom-Def (Partition Activity-Or-
Spec (Setof Activity Activity-Spec)));
```

alors que pour Tove une activité est l'opération élémentaire de changement d'état. Elle correspond à un graphe (activity cluster) liant un état initial, dans lequel doit se trouver le système pour que l'activité soit applicable, à un état final. Les activités peuvent être structurées pour définir des activités plus complexes. Ainsi, un plan d'action sera défini par l'instruction :

```
(define-class plan_action (?a):def
 (forall (?alpha ?f ?s)
    (=> (holds (agent_constraint
          ?alpha (fluent_goal ?f)) ?s)
          (forall (?ap ?s1 ?s2)
       (=> (and (subaction ?ap ?a)
            (leq ?s1 ?s2) (Do ?ap ?s1 ?s2
            (intended ?s2))
            (holds ?f ?s2)))))
```

De même, que peut-on attendre des définitions de Mikrokosmos, quand la présentation de ce projet débute par ces avertissements : « *In this ontology, you should not expect to find : any kind of guarantees, warrantees, or liability for correctness or precision, formally clean or theoretically «pure» concepts, complete consistency ; guaranteed absence of contraditions ; etc* ».

Les principes épistémologiques (au sens de la théorie de la connaissance) de certaines ontologies, de par leur imprécision, peuvent expliquer ces problèmes. Si le calcul des prédicats offre un formalisme rigoureux pour la définition du sens d'un terme, il ne permet pas néanmoins de différencier entre concept et ensemble si l'on considère que le premier porte sur l'essence des objets qu'il subsume et le second sur leur état (un ensemble regroupe les objets, éventuellement de nature différente, vérifiant une propriété logique portant sur les valeurs de leurs attributs définissant leur état). Tove illustre ce type de problème. De

même, la confusion entre substance et qualité ou quantité, dans le cas de *Mikrokosmos*, peut être source de problèmes : dans une telle ontologie le mercure serait-il à la fois un métal et un liquide ou un métal dont l'état, sous certaines conditions, est liquide ? Le problème de la langue naturelle, miroir incontournable mais déformant de la réalité, devra être pris en compte lors de l'acquisition des connaissances ontologiques.

2.1 Principes d'*OK*

Afin d'éviter de tels écueils, le modèle *OK* repose sur des principes épistémologiques forts, voire contraignants, qui, s'ils ne permettent pas d'aborder tous les problèmes de la sémantique lexicale, permettent d'obtenir des définitions consensuelles et cohérentes souvent suffisantes dans le cadre de domaines techniques.

Notre objectif est la définition de la signification de termes dénotant des connaissances conceptuelles (concepts et ensembles), c'est-à-dire portant sur une pluralité de choses, et ce à partir de quoi elles sont définies (différences et attributs). L'ensemble de ces termes est structuré en quatre vocabulaires : celui des concepts, des ensembles, des différences et des attributs. L'association d'une signification à chacun de ces termes constitue une ontologie *OK*.

Le modèle *OK* s'appuie sur le classique triangle sémiotique :
[« signifiant » – <signifié> – *référent*]
où le sens d'un « mot » est le < concept > qu'il dénote. Savoir ce que signifie un mot se ramène à connaître l'idée dont il est le signe. Cette approche relève donc davantage d'une théorie des idées que de la linguistique proprement dite. Pour la suite de l'article, la notation « m » désignera le terme et <m> sa signification.

• Une distribution du sens

L'approche adoptée par *OK* n'impose pas une ontologie commune, acceptée et partagée par tous les agents. Le propre de l'ingénierie simultanée est d'avoir pris conscience de cette distribution du sens en autant d'ontologies locales aux agents que nécessaire. Néanmoins l'existence d'une ontologie minimale partagée par tous, l'ontologie invariante, est indispensable à la communication entre agents, tout comme l'est la compatibilité des ontologies locales avec l'ontologie commune.

• Une définition de la définition

Le modèle *OK* repose sur une définition précise des signifiés conceptuels (concepts) :
le signifié d'un terme désignant un concept est défini à partir d'un autre signifié conceptuel (concept) en précisant sa différence spécifique.

De cette définition découlent les propriétés du modèle *OK*. Ce modèle repose sur une sémantique référentielle puisque le sens d'un terme est l'idée dont il est le signe, et différentielle dans la mesure où ce sens est défini à partir d'un signifié existant par différence spécifique.

OK postule une classification précise des connaissances mises en jeu pour la définition des signifiés, distinguant d'une part les concepts des ensembles, et d'autre part les différences des attributs. Les notions de concept et de différence sont centrales dans *OK*, c'est pourquoi nous nous limiterons à la présentation de ces deux seules notions.

• Les différences

La différence est l'unité sémique élémentaire à partir de laquelle se construit par différence la sémantique des concepts. Elle traduit une qualité essentielle des objets subsumés par le concept qui ne peut être valuée (si *mortel* est une différence, *âge* est un attribut). Nous retrouvons ici l'approche des arbres de Porphyre (Porphyre).

De par la définition de la construction du signifié d'un concept, la différence est une unité divisive et constitutive : divisive au sens où elle sépare les concepts en ceux qui la possèdent et ceux qui ne la possèdent pas ; constitutive au sens où elle participe à leur l'essence. Une différence ajoutée à un concept <C> permet donc d'engendrer deux nouveaux concepts <C1> et <C2>, le premier la possédant, le deuxième ne pouvant la posséder. Une telle différence est dite différence spécifique pour le concept <C1>. Une différence divise l'ensemble des <concepts> subsumés en deux sous-ensembles de <concepts>, ceux qui la possèdent et ceux qui ne peuvent la posséder. Il en découle que l'ensemble des signifiés conceptuels se structure sous la forme d'un arbre binaire (arbre de Porphyre). La figure n°1 en est un exemple.

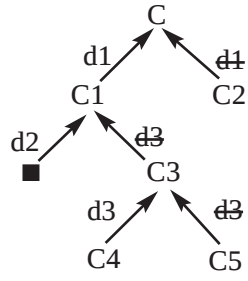


Figure 1

Les différences sont définies par paires de différences opposées (par exemple d1 et ~~d1~~ de la figure n°1) correspondant à des antonymes complémentaires. En tant qu'unité sémique élémentaire, aucun signifié n'est associé à une différence qui ne prend de sens que dans la mesure où elle participe à la définition des <concepts>. C'est pourquoi la détermination de ces sèmes élémentaires doit être consensuelle. Bien que consensuelle, la détermination de ces différences reste arbitraire et subjective. Elle est

fonction d'une application donnée et d'une communauté particulière.

• Les concepts

Le concept s'intéresse à l'essence et à la structure des choses de même nature, concepts ou objets, indépendamment de l'état des objets qu'il subsume. Il se distingue de l'ensemble qui permet de regrouper des objets pouvant être de nature différente mais satisfaisant des propriétés communes pouvant porter sur les valeurs de leurs attributs (état de l'objet). Ainsi, si *Homme* est un concept, *Adolescent* est un ensemble défini en intention par l'expression logique: Homme(x) (âge(x)<18.

Par définition les signifiés conceptuels sont structurés sous la forme d'un arbre binaire et héritent des différences spécifiques des <concepts> qui le subsument.

• Les catégories

Il existe des concepts qui par essence n'entretiennent aucune relation quant à leur définition (*êtres animés, usinages, temps*…). Leurs signifiés sont alors structurés en autant d'arbres binaires différents correspondant à autant de catégories *OK*. Les concepts racines de ces catégories ne sont pas définis par différentiation spécifique, ce sont des concepts primitifs dont le choix détermine une modélisation particulière du monde.

• Propriétés

La définition d'un concept par différenciation spécifique à partir d'un signifié existant basée sur l'utilisation de couples de différences correspondant à des antonymes complémentaires offre de multiples avantages qui ont justifié ce choix:

• Cette approche de la définition, pour rigide qu'elle puisse paraître, a l'avantage d'être admise par tous;

• Elle se concentre sur l'essence des choses sans la confondre avec la notion d'état soumis aux changements;

• Elle évacue le problème de la hiérarchie multiple par définition

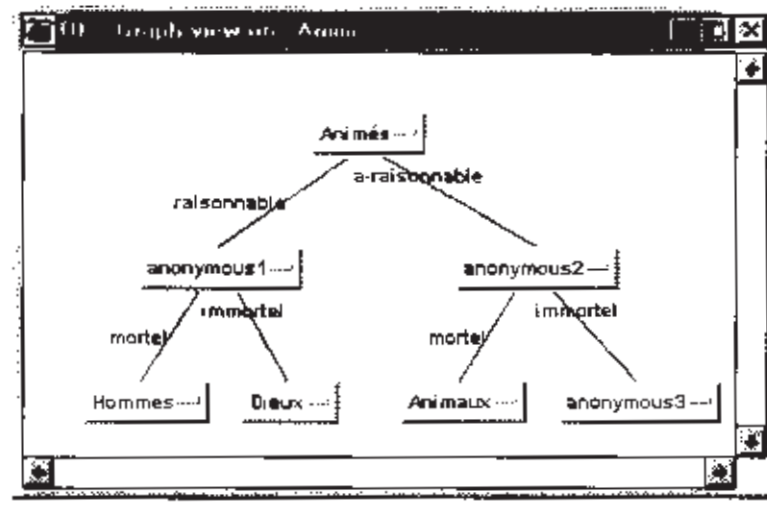


Figure 2

même des concepts. De plus, les différences étant définies par paires correspondant à des antonymes complémentaires, un concept ne peut hériter de deux concepts différents, nécessairement opposés par une paire d'antonymes (sachant qu'un concept *OK* ne peut appartenir, par définition, qu'à une seule catégorie);

• Elle assure aux modèle *OK* des propriétés logiques intéressantes, garantes de la cohérence des ontologies. Ces propriétés sont exploitées tant pour la définition ou l'acquisition de nouveaux concepts que pour le classement;

• Elle permet de définir des équivalences entre arbres du fait que seules importent les différences et ce quel que soit leur ordre,

• Elle réduit le problème des définitions consensuelles à la seule détermination des différences. Ce consensus est relativement facile à obtenir dans la mesure où les différences sont des unités de sèmes élémentaires. Il l'est d'autant plus que le domaine est technique et que les différences désignent des faits concrets: «enlèvement de matière», «rotation de la pièce», «traitement thermique »…

2.2 le langage *Lok*

La gestion des «signifiants » et des <signifiés> se fait à l'aide d'un langage dédié, le langage *Lok* pour *Language for Ontological Knowledge*. Ce langage comporte plus de 150 instructions, avec une syntaxe «à la Lisp» (notation fonctionnelle préfixée parenthésée). Ces instructions se répartissent en deux ensembles.

• Les instructions de définition d'ontologies

Un premier ensemble d'instructions *Lok* permet la définition et la modification des associations «signifiant » – <signifié>. Prenons pour exemple l'arbre des concepts de la figure n°2.

Les instructions *Lok* suivantes permettent de définir de nouvelles paires de différences antonymes. Lorsqu'une seule différence est définie, la différence opposée est automatiquement créée à partir de cette différence en la préfixant du «a-» privatif.

```
(defineDifference raisonnable')
  returns ('raisonnable  a-raisonnable')
(defineDifference mortel immortel')
                  returns ('mortel immortel')
```

La définition d'un nouveau signifié conceptuel s'effectue à partir d'un signifié existant en précisant la différence spécifique.

```
(defineConceptFrom Animés
  (leftConcept ?
    (specificDifference raisonnable')))
     returns ('anonymous1 anonymous2')
(defineConceptFrom Animés
raisonnable
  (leftConcept Hommes
    (specificDifference mortel'))
  (rightConcept Dieux
    (specificDifference immortel')))
              returns ('Homme Dieux')
```

Il est à noter que *Lok* permet de définir des signifiés sans signifiant («?») qui seront dénotés par une désignation, c'est-à-dire par un terme dénotant un concept et une suite de termes dénotant des différences: *Animés raisonnable* désignant le concept anonyme *anonymous1* de la figure n°2.

Toutes les instructions de ce premier ensemble modifient l'ontologie. C'est pourquoi une ontologie *OK* sera définie par une suite d'instructions *Lok* appartenant à ce premier ensemble.

• Les instructions de manipulation d'ontologies

Le deuxième ensemble d'instructions est utilisé pour l'exploitation des ontologies: interrogation et recherche selon différents critères, tant sur les termes que sur leurs significations.

L'évaluation de l'instruction suivante retourne la définition d'un concept sous la forme d'une chaîne de caractères:

```
(prettyDefinitionOf Homme')
returns   concept 'Homme is Animé
                  raisonnable with
                  specific difference
                  mortel category
                  Animé
```

Ces instructions ne modifient pas l'ontologie et ne participent donc pas à sa définition.

2.3 Représentation computationnelle

Une ontologie *OK* est représentée de façon déclarative par un fichier d'instructions *Lok*. Le résultat de la compilation d'un tel fichier est une représentation computationnelle de l'ontologie qui permet de gérer cinq vocabulaires de termes, celui des concepts, des différences, des attributs, des relations et des ensembles, ainsi que leurs significations (représentation et organisation des concepts).

2.4 *Lok*: langage pivot

Un des buts des ontologies est de permettre le partage et l'échange de connaissances (termes et significations). C'est pourquoi il est nécessaire de pouvoir fournir des ontologies écrites dans un formalisme plus commun que *Lok*. En fait, chaque acteur peut utiliser son propre formalisme de représentation de connaissances: réseau sémantique, graphe conceptuel, schéma, calcul des prédicats, base de données relationnelles, etc. dont la diversité illustre l'ampleur du problème. Plutôt que de traduire une représentation *Lok* dans chacun de ces formalismes, ce qui impliquerait l'écriture d'autant de traducteurs, nous avons choisi la solution des formats d'interchange qui permet de réécrire une ontologie *OK* dans un formalisme accepté par tous (le terme d'interchange que l'on préférera à celui d'échange, tout peut être échangé, insiste d'avantage sur le caractère commun du formalisme). C'est pourquoi les ontologies *Lok* peuvent être traduites en *Kif* (*Knowledge Interchange Format*) ou en graphes conceptuels. Ces fichiers sont plus conséquents en taille et moins lisibles du fait que ces formalismes sont plus généraux que *Lok* et qu'il est dès lors nécessaire de traduire également la sémantique du modèle *OK*.

2.5 Architecture de la *OK Station*

La *OK Station* est un environnement logiciel qui permet à

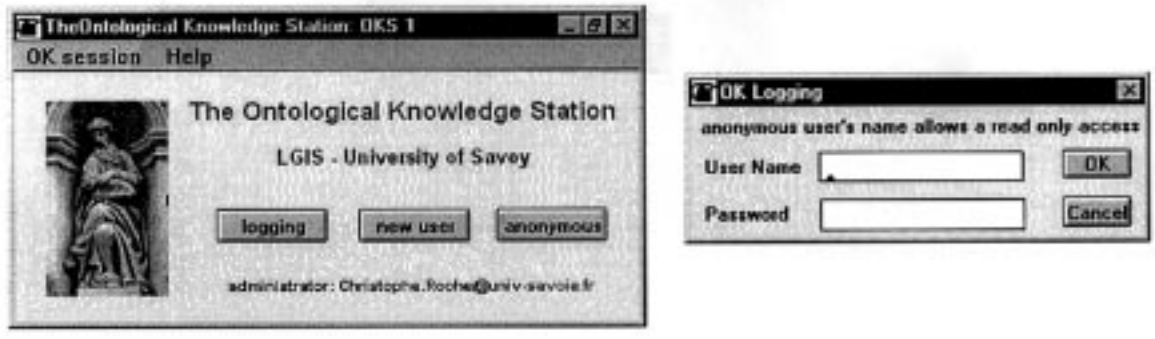


Figure 3

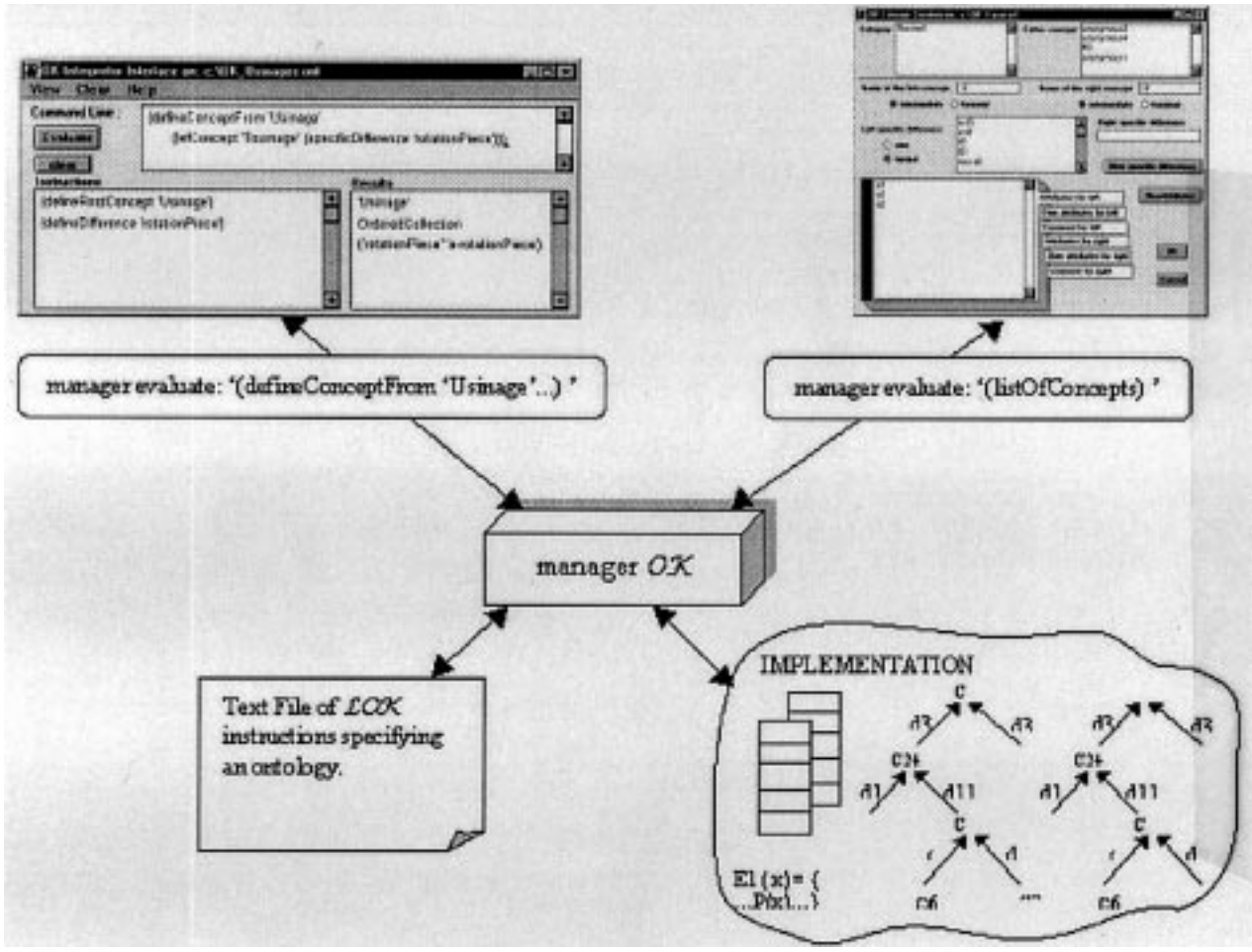


Figure 4

un ensemble d'utilisateurs d'acquérir, de définir, d'exploiter et de gérer des ontologies.

L'architecture de la *OK Station* repose sur trois modules : un module de gestion des utilisateurs, un module de gestion des ontologies comprenant un ensemble d'outils pour la définition et l'exploitation des termes qui les composent ; et un module d'acquisition décrit dans le troisième paragraphe.

• Les utilisateurs

La *OK Station* permet à un ensemble d'utilisateurs de gérer (définir et exploiter) leurs ontologies en offrant les sécurités classiques d'un tel système : gestion des utilisateurs, mot de passe, droits d'accès, gestion des ressources critiques. La figure n°3 présente la bannière d'accueil de la *OK Station* et la fenêtre de connexion des utilisateurs.

• Les « managers »

L'utilisation d'une ontologie définie par un fichier d'instructions *Lok* se fait par l'intermédiaire de sa représentation computationnelle et de son manager associé. La représentation informatique n'étant pas accessible, le manager *OK* est le passage obligé pour l'exploitation des termes et de leurs significations. Ainsi, pour une ontologie donnée, la création d'un nouveau concept, la recherche de la définition d'un terme, sont autant de requêtes écrites en *LOK* que le manager associé à l'ontologie devra interpréter. La figure n°4 décrit les rapports entre ces différents composants.

• Le « launcher »

La gestion des ontologies d'un utilisateur s'effectue à l'aide d'un *launcher OK* qui lui est associé lors de sa connexion à la *OK Station*. Ce *launcher* lui permet d'une part de gérer les ontologies en tant que fichier : suppression, création, édition, traduction en format *Kif* (*Knowledge Interchange Format*) ou en graphes

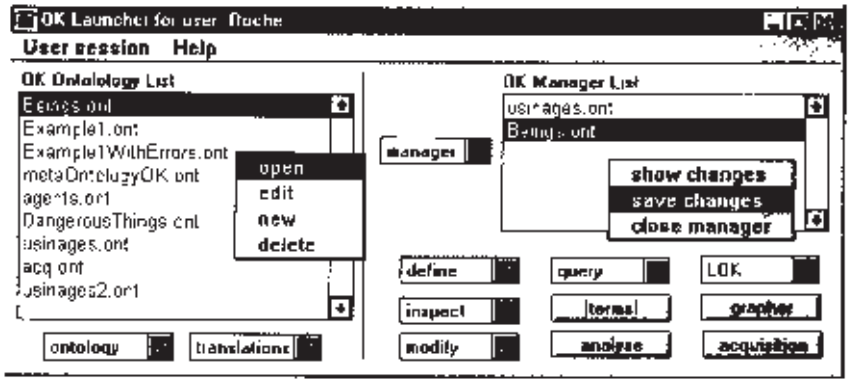


Figure 5

conceptuels; et d'autre part de gérer leurs représentations computationnelles par l'intermédiaire des managers associés aux ontologies lors de leur ouverture: gestion des termes et de leur signification: définition, modification et recherche. La figure n°5 présente l'ensemble de ces fonctionnalités.

• Les outils

Enfin, la *OK Station* offre un ensemble d'outils pour la manipulation des termes et de leur signification: éditeur graphique interactif, interpréteur d'instructions *Lok* , fouineurs de définition, de suppression, d'édition, de recherche… Tous ces outils, interfacés graphiquement, sont accessibles pour un utilisateur donné à partir de son *launcher* et portent sur l'ontologie gérée par le manager sélectionné.

• Applications

La *OK Station* est utilisée dans le cadre d'un projet Eurêka pour la définition des termes utilisés au sein de l'équipe de développement (ontologies agents) et pour la définition d'ontologies d'entreprise (usinages, modélisation de l'entreprise).

3 Acquisition

Le chapitre 3 présente des travaux menés en parallèle avec ceux décrits dans les deux chapitres précédents. Il présente en particulier comment certaines techniques d'acquisition peuvent être particularisées pour le modèle *OK*.

Le travail de modélisation de la connaissance peut être couplé avec l'acquisition de connaissances. L'acquisition de connaissance consiste à extraire et à analyser les connaissances d'un domaine et à les décrire d'une manière à la fois compréhensible et rigoureuse. Ce n'est pas une simple activité de collecte, mais un réel travail de modélisation. Ainsi, avec notre approche ontologique, nous voyons apparaître la nécessité d'adapter les méthodes d'acquisition choisies à notre modèle de représentation: Les concepts fondamentaux de la modélisation doivent transparaître à travers le guide offert par les différentes méthodes d'acquisition.

Le processus d'acquisition de connaissance est un processus incrémental. (Marty 1991). Chaque itération du processus permet d'acquérir de nouvelles connaissances et de raffiner ou d'enrichir la version courante de l'ontologie. L'intégration de plusieurs méthodes d'acquisition permettra de mieux tirer parti des avantages de chaque méthode. Le processus d'acquisition étant incrémental, on peut donc fort bien considérer que chaque cycle est en fait l'application d'une technique d'acquisition (la plus adéquate à cette étape) afin de raffiner l'ontologie.

Chaque technique est caractérisée par une stratégie de recueil particulière, un mode de représentation des connaissances spécifique et des moyens d'analyse associés pour valider la modélisation obtenue. L'utilisation de l'éventail des techniques proposées permet ainsi de gagner en productivité, en qualité et en précision:

• En productivité car nous incluons des techniques d'extraction qui permettent de structurer et d'analyser les connaissances directement au cours d'une interview structurée; la phase de transcription en texte libre est ainsi supprimée. La productivité est aussi améliorée car les principes de base de ces techniques, issus de la psychologie cognitive, facilitent l'expression des experts.

• En qualité puisque les techniques reposent sur des stratégies systématiques (moins de risque d'oubli, de digression...) et aident à détecter les ambiguïtés, les contradictions ou les manques.

• En précision grâce à la synergie entre techniques: l'ingénieur de la connaissance peut appliquer à chaque étape la technique de modélisation la plus appropriée. La connaissance ainsi acquise pourra être ultérieurement utilisée par d'autres outils d'acquisition.

Chacune de ces activités contribue à l'extraction, à la précision ou à la réorganisation de la connaissance nécessaire pour résoudre un problème dans un domaine donné. On peut donc imaginer autour de notre noyau de modélisation ontologique, un atelier de génie cognitif composé de différents outils supportant des méthodes variées et complémentaires d'acquisition de connaissance. Cette plate-forme devra être ouverte afin de pouvoir intégrer de nouvelles techniques si besoin est. Nous avons déjà identifié au moins les besoins suivants:

• Détection et corrélation des éléments du domaine

Une des activités du cogniticien consiste à analyser les données brutes du domaine (documentations, transcriptions d'interviews) en identifiant et en corrélant les concepts du domaine, les attributs, les relations. Cette activité structure principalement la connaissance textuelle. Elle produit un vocabulaire des termes les plus importants du domaine ainsi que des liens sur leurs parties explicatives. Elle permet également d'annoter les différentes informations et de les classer les unes par rapport aux autres, procurant ainsi une première hiérarchie taxinomique des informations du domaine.

• Hiérarchie taxinomique des concepts: création et raffinement

Les concepts significatifs identifiés dans le domaine doivent être regroupés dans une hiérarchie entre concepts. Les attributs significatifs à prendre en compte pour ces concepts doivent être fournis. Des règles de classification peuvent émerger de cette activité.

Bien souvent, cette classification s'effectuera lors d'une séance d'interview supplémentaire avec l'expert du domaine. Une méthode adéquate pour cette activité est la méthode du tri par cartes (Gamack 1987) qui aide souvent à identifier des groupes de concepts liés entre eux.

La méthode de grille répertoire (Bannister 1977) permet au cogniticien de raffiner la hiérarchie des concepts en construisant une grille contenant d'une part des concepts du domaine et d'autre part des attributs de ces concepts qui ont deux extrêmes.

Nous avons particularisé ces méthodes classiques afin que le cogniticien se rapproche le plus possible de la formalisation en phase d'acquisition et qu'il prenne en compte les éléments majeurs de notre modèle *OK*, comme par exemple raisonner par attribut différentiateur.

En ce qui concerne la détection et la corrélation des éléments du domaine, la première itération peut consister en une interview libre. Comme notre système est ouvert, on pourrait imaginer ici l'intégration de techniques d'extraction statistiques telles que celles utilisées dans *Lexter* (Bourigault 1995). Actuellement, cette première élicitation du vocabulaire du domaine se fait en collaboration avec un expert, qui énonce les principaux termes du domaine et donne une définition pour chacun de ces termes. La définition donnée par l'expert fait apparaître de nouveaux termes, qui sont rajoutés à la liste, et dont l'expert doit également donner une définition. On s'arrête lorsque tous les termes apparaissant dans les définitions ont été eux-mêmes définis. On estime alors que l'on a obtenu une liste de termes suffisamment complète pour passer à l'étape suivante. Bien entendu, on peut revenir à cette étape dans une itération ultérieure, si une étape d'acquisition suivante fait apparaître de nouveaux termes.

Ensuite, on réalise un premier tri par cartes sur ces termes du domaine. L'objet de ce tri par cartes est de catégoriser le domaine (au sens des catégories *OK*). Ici, c'est l'expert qui détermine ces différentes catégories par le choix des tas. Ce tri permet également de simplifier les étapes suivantes de l'acquisition, en fournissant de plus petits groupes de termes.

L'étape suivante de notre processus incrémental consiste à élaborer la hiérarchie taxinomique des concepts. Nous nous rapprochons alors de notre modèle de représentation des connaissances, et les techniques d'acquisition vont devenir directives pour respecter ce modèle. Ainsi, on utilise de nouveau un tri par cartes, dédié à la terminologie *OK*, dont les tas correspondent aux types de termes. On obtient ainsi un tas de concepts, un tas de différences spécifiques, un tas d'attributs, etc.

Il reste alors à déterminer l'organisation des concepts entre eux, à l'aide des différences spécifiques. On utilise ici la technique des grilles répertoires, particularisée pour *OK*. Dans notre cas, au lieu d'utiliser des attributs pour différencier les concepts, on utilise les différences spécifiques. Celles-ci n'étant par définition pas valuées, on ne remplit pas les cases de la grille avec des notes, comme on le fait habituellement pour les attributs dans la technique de grille répertoire. On remplit plutôt les cases de la grille en déterminant, pour chaque différence et chaque concept, si le concept possède ou non cette différence, s'il peut la posséder, ou si celle-ci n'a pas de sens pour le concept. Lorsque la grille est remplie, on l'analyse et on la complète avec l'aide de l'expert.

Cette étape d'analyse de la grille, habituellement utilisée pour rapprocher les concepts ou les attributs du domaine, sera dans notre cas utilisée pour déterminer si la grille est complète et cohérente. Par exemple, si deux concepts ont les mêmes valeurs pour toutes les différences, cela signifie soit qu'ils sont synonymes, soit qu'il manque un élément dans leur définition. S'il s'agit du deuxième cas, l'expert les différencie en rajoutant à la grille une nouvelle différence spécifique. De même, le système détecte les différences spécifiques susceptibles d'être synonymes ou opposées. Lorsque la grille est complète, on construit la hiérarchie des concepts d'après leurs différences.

Notre processus incrémental peut alors entrer dans une étape de raffinement. Dans certains cas, la génération de hiérarchies de concepts fournit une forêt et non un arbre. Dans ce cas, l'expert peut valider les résultats obtenus, auquel cas il s'agissait de plusieurs catégories différentes. Il peut au contraire

compléter la grille de manière à relier ces hiérarchies de concepts, en ajoutant par exemple une différence spécifique qui permettrait de les opposer.

Conclusion

Une société de l'information n'existe que s'il y a communication, partage et échange de connaissances entre ses acteurs. Or il ne peut y avoir communication, et *a fortiori* collaboration et coopération, sans compréhension des mots qui composent le message. C'est pourquoi les ontologies, comprises ici comme des vocabulaires communs sur des termes et leurs significations, constituent un axe de recherche primordial. Dans ce cadre particulier d'applications, où l'on est confronté à des problèmes qui lui sont propres tels que la notion d'inter-opérabilité, pris au sens de communication d'agents logiciels hétérogènes, les ontologies doivent répondre à un double critère de consensus et de cohérence, condition *sine qua non* d'un modèle computationnel réellement exploitable. Pour cela, elles doivent reposer sur des fondements épistémologiques «clairs» qui ne peuvent puiser à une seule discipline mais tenir compte des enseignements de la linguistique, des sciences cognitives, de la sémantique, de la théorie de la connaissance et des modèles computationnels.

Une approche atomiste de la signification basée sur la notion de différence, unité sémique élémentaire à partir de laquelle se construit la sémantique des termes désignant des concepts, directement issue des arbres de Porphyre, permet de répondre de façon satisfaisante à ces exigences en ce qui concerne les domaines techniques. Le modèle *OK*, pour *Ontological Knowledge*, présenté dans cet article, en est une illustration. La définition d'un concept par différenciation spécifique basée sur l'utilisation de couples de différences correspondant à des antonymes complémentaires offre de multiples avantages qui ont justifié ce choix. Ainsi, l'accent est mis sur l'essence des choses sans la confondre avec la notion d'état soumis aux changements; le problème de la hiérarchie multiple est évacuée par définition même; le modèle offre des propriétés logiques intéressantes et garantes de la cohérence des ontologies; il est possible de définir des «équivalences» entre arbres des signifiés du fait que seules importent les différences et ce quel que soit leur ordre; et enfin le problème des définitions consensuelles est réduit à la seule détermination des différences.

Une ontologie trouve sa justification dans la pleine exploitation des relations entre les signifiés. Pour cela nous avons défini un langage déclaratif de représentation ontologique, le langage *LOK*, et un modèle computationnel associé. Le résultat en est un environnement logiciel, la *OK Station*, qui permet à un ensemble d'utilisateurs d'acquérir, de définir, d'exploiter et de gérer des connaissances ontologiques.

De nombreux problèmes restent en suspens tels que la distribution du sens à travers des ontologies locales, la réutilisabilité d'ontologies existantes et l'évolution des significations. Ici aussi, nous pensons qu'une approche analytique, même si elle est réductrice par certains aspects, peut apporter des solutions intéressantes dans le cadre des sociétés de l'information.

Christophe Roche,
Jean-Charles Marty,
Stéphanie Lacroix,
LGIS – Université de Savoie.

Bibliographie

Bourigault (D.), Lépine (P.), 1994: «Extraction et structuration automatiques de terminologie pour l'aide à l'acquisition des connaissances à partir de textes», dans *Actes du 9e Congrès AFCET RFIA'94, janvier 1994*.

Bachimont (B.), 1996: *Herméneutique matérielle et artéfacture: des machines qui pensent aux machines qui donnent à penser*, Doctorat de l'École Polytechnique, 1996

La banque des mots, 1995: *Terminologie et intelligence artificielle*, numéro spécial 7-1995, revue de terminologie française.

Cutkosky, Engelmore, Fikes, Gruber, Genesereth, Mark, Tenenbaum et Weber, 1993: «PACT: An experiment in integrating concurrent engineering systems», dans *IEEE Computer*, vol. 26, n°1, January 1993.

Eco (U.), 1988: *Sémiotique et philosophie du langage*, PUF, Paris, 1988

Fox (M.S), 1992: «The TOVE Project: Towards a Common Sense Model of the Enterprise», dans *Enterprise Integration, C. Petrie (Ed.), 1992*, Cambridge MA: MIT Press.

Gamack (J.G), 1987: «Different Techniques and Different Aspects of Declarative Knowledge», dans Kidd (A.L.) editor, *Knowledge Acquisition for Expert Systems: A Practical handbook*, Plenus Press, New York, 1987.

Fransella, D. Bannister, 1977: *A Manual for Repertory Grid Technique*, Academic Press, 1977

Gruber (T. R.), J.M. Tenenbaum and J.C. Weber, 1992: «Towards a knowledge Medium for Collaborative Product Development», dans *Proceedings of the Second International Conference on Artificial Intelligence in Design, (Pittsburgh, Ill., USA, Jun. 22-25, 1992)*, Kluwer Academic Publishers.

Gruber (T.R.): «Toward Principles for the Design of Ontologies Used for Knowledge Sharing», dans *Formal Ontology in Conceptual Analysis and Knowledge Representation*, edited by Nicola Guarino and Roberto Poli, Academic Publishers.

McGuire, Kuokka, Weber, Tenenbaum, Gruber et Olsen, 1993 : « SHADE : Technology for Knowledge-Based Collaborative Engineering », dans *Concurrent Engineering: Research & Applications*, Volume 1, n° 3, September 1993

Labrou (Y.), Finin (T.), 1997 : *A proposal for a new KQML Specification*, Internal Report TR CS-97-03, Computer Science and Electrical Engineering Department (CSEE), University of Maryland Baltimore County (UMBC).

Mahesh (K.) : *Mikrokosmos*, http://crl.nmsu.edu/Research/Projects/mikro/htmls/ontology-html/onto.index

Marty (J.C), Ramparany (F.), Doize (M.S), Jullien (C.), 1991 : « ACKnowlegde : An Intelligent Workbench for the Knowledge Engineer », dans *Actes The World Congress on Expert Systems, December 1991*.

Porphyre : *Isagoge*, traduction J. Tricot, Vrin 1984

Rey (A.), 1992 : *La Terminologie, noms et notions*, Paris, PUF.

Rastier (F.), 1987 : *Sémantique interprétative*, Paris, PUF.

Rastier (F.), 1991 : *Sémantique et recherches cognitives*, Paris, PUF.

Roche (C.), Dumond (Y.), 1998 : « From Concurrent Engineering to Collaborative Engineering : An Agent-Oriented Approach », dans *ECEC'98, 5th European Concurrent Engineering Conference (SCS : Society for Computer Simulation), Erlangen-Nurember, Germany, April 26-29 1998*

Roche (C.), 1999 : « Ontology : A Key Point for Concurrent Engineering », dans *Actes CAPE'99 : the 15th International Conference on Computer-Aided Production Engineering. Durham, UK, April 19-21 1999*

Stader (J.), 1996 : « Results of the Enterprise Project », dans *Actes Expert Systems 96, the 16th Annual Conference of the British Computer Society Specialist Group on Expert Systems ; Cambridge, UK ; December 1996*

Géditerm : un logiciel pour gérer des bases de connaissances terminologiques

Les bases de connaissances terminologiques organisent des connaissances terminologiques en différenciant les aspects linguistique et conceptuel. Cet article fait le point des outillages disponibles pour gérer des BCT et délimite les besoins des terminologues en la matière ; il présente *Géditerm*, logiciel de saisie et gestion de BCT que nous avons développé, ses spécifications et ses fonctionnalités ; il se termine par les questions théoriques qu'a fait évoluer cette réalisation, en particulier en contribuant au choix du modèle de données et du niveau de représentation des connaissances.

Termes clés :
Bases de connaissances terminologiques ;
représentation des connaissances ;
modélisation conceptuelle.

Introduction

Voilà près de dix ans que la notion de base de connaissances terminologiques (BCT) a vu le jour, en particulier autour des travaux d'I. Meyer (1992) ou d'A. Condamines (1993). Les BCT visent à répondre au besoin de rendre compte de connaissances terminologiques, en différenciant les niveaux linguistique et conceptuel dans une même structure de données. Le concept de BCT, aujourd'hui parvenu à une certaine maturité, se stabilise dans des travaux de recherche au confluent de la terminologie et de l'intelligence artificelle. Il reste cependant peu précis, de telle sorte que les réalisations concrètes, peu nombreuses, sont toutes assez différentes. Ainsi, les BCT sont encore loin d'être répandues et diffusées, que ce soit pour des applications industrielles ou auprès des terminologues. On peut trouver plusieurs raisons à cela :

1) Leur mise au point est longue et les méthodes applicables encore en cours de définition. Pour dépasser les techniques classiques d'exploration manuelle de corpus ou de définition de termes sans référence à leur usage, les terminologues exploitent désormais des corpus selon une approche linguistique, manuelle ou automatisée à l'aide de logiciels de traitement automatique de langage naturel (TALN). Or l'utilisation judicieuse de ces outils et leur intégration requièrent un travail méthodologique rigoureux associé à des développements informatiques complémentaires.

2) Leur intérêt, bien qu'intuitivement fort, présente plusieurs facettes dont certaines n'ont pas été complètement démontrées. Ainsi, les BCT constituent un pas en avant indéniable pour la pratique terminologique : on passe de listes de termes, chacun ayant un statut référentiel, à des termes reliés entre eux sémantiquement via un réseau conceptuel, établi en fonction de leur usage en corpus. Dans une BCT, le terme est donc rattaché à une sémantique différentielle, plus pertinente dans de nombreux cas, comme le soulignent F. Rastier (1995), M. Slodzian (1995) ou B. Bachimont (1995). En revanche, beaucoup reste à faire pour démontrer l'apport de l'utilisation des BCT au développement d'applications. Il faudrait préciser comment les exploiter au mieux et mesurer en quoi elles augmentent la qualité des modèles produits et l'efficacité de leur construction. Pour le moment, la contribution des BCT a été surtout étudiée pour construire des ontologies par D. Skuce (1994), des terminologies ou des index par C. Gros (1998).

3) Les outils informatiques adaptés à leur construction sont encore insuffisants. Ils peuvent être de deux sortes. Une première classe de logiciels(1) aide le terminologue-linguiste à analyser les corpus et à automatiser ses techniques d'exploration. Ce sont des logiciels de TALN, les extracteurs de listes de mots, des outils d'aide à la mise au point ou à l'application de

(1) Un panorama d'outils de ce type a été dressé par A. Condamines et J. Rebeyrolles en 1997.

marqueurs, des concordanciers ou des outils statistiques, etc. Une deuxième famille regroupe les environnements de saisie et de gestion des données d'une BCT. Ces logiciels, peu nombreux, correspondent souvent à une interprétation très informatique de la notion de BCT. Plutôt que de répondre au plus près aux besoins des linguistes-terminologues, ils privilégient la mise au point d'un réseau conceptuel et se focalisent d'avantage sur la formalisation des connaissances. Ils servent en effet souvent à construire des bases de connaissances ou des ontologies.

Nous avons développé un logiciel de saisie de BCT, dans le cadre plus large d'un projet(2) qui vise l'évaluation de la pertinence des BCT pour différents types d'applications. En contribuant au choix du modèle de données puis du niveau de représentation des connaissances, notre intervention a fait progresser la définition de ce concept, tout en mettant en évidence à la fois ses limites et sa pertinence. L'article s'articule donc autour de trois parties : la première fait le point de logiciels disponibles pour gérer des BCT et délimite les besoins des terminologues auxquels nous avons cherché à répondre ; la deuxième présente le logiciel développé, *Géditerm*, les spécifications retenues et ses fonctionnalités ; la dernière met l'accent sur les questions théoriques qu'a fait évoluer cette « concrétisation informatique » du concept de BCT.

(2) Ce projet a été financé par le Gis « Sciences de la Cognition », la région Midi-Pyrénées et la DER d'EDF-GDF de 1996 à 1998.

1 Quelques systèmes de gestion informatique de BCT

Pour présenter les différents logiciels ou langages qui existent aujourd'hui pour construire des BCT, nous nous plaçons dans la perspective du linguiste-terminologue qui souhaite enregistrer des données alors qu'il analyse un corpus, identifiant des termes ou des connaissances sur un domaine. Nous mettons à part les bases de données terminologiques, versions informatiques de listes sur papier, présentant les mêmes faiblesses : les termes sont définis indépendamment les uns des autres, sans référence précise à leur usage.

Nous avons donc évalué en quoi les structures de données produites par différents systèmes étaient des BCT, quelles étaient leurs fonctionnalités et en quoi elles facilitaient la tâche de leurs utilisateurs. Nous avons retenu les critères énoncés par F. Lemaire et F. Rechenmann (1995) : la structure d'une BCT doit permettre d'éviter la confusion entre relations linguistiques (grammaticales) et relations sémantiques (conceptuelles) ; de différencier la représentation des connaissances terminologiques et conceptuelles. L'offre actuelle est très disparate, suivant que les outils prennent en compte ou non l'utilisation qui sera faite de la BCT, et qu'ils s'appuient ou non sur une représentation formelle des connaissances. Le fait d'utiliser une représentation formelle permet, lors de la construction d'une BCT, de réduire les ambiguïtés en obligeant à formuler des critères de définition et de différenciation, de classer au fur et à mesure les concepts définis, de vérifier leur cohérence, etc. La contribution de l'intelligence artificielle est ici significative : la plupart des formalismes utilisés sont inspirés des réseaux sémantiques, des logiques de description et des graphes conceptuels. Par contre, la formalisation impose des contraintes. Nous reviendrons sur la question de savoir si une BCT doit être formelle ou non lorsque nous présenterons nos propres choix, dans la partie 2.

Nous distinguons ici les réalisations qui visent à développer une représentation structurée en vue d'utilisations potentielles non définies de celles qui visent à développer des BCT en sachant à quoi elles vont servir (ces BCT formeront une partie d'application). Les premières s'intéressent plus au processus de construction des BCT et à la composante terminologique alors que les secondes privilégient la base de connaissances de la BCT et son utilisation.

1.1 Des logiciels pour enregistrer des BCT

Les environnements qui guident l'organisation des connaissances et la construction de la BCT, sans avoir *a priori* une idée sur l'utilisation qui en sera faite, ne reposent pas sur un profil précis de l'utilisateur, à la fois linguiste et cogniticien. Le modèle des données et l'interface dissocient termes et concepts, tout en facilitant leurs associations et le retour au corpus de référence. On peut donc bien parler de BCT au sens donné par Meyer (1992), Condamines (1993) ou Rechenmann (1995). Ces travaux ont également en commun le souci de mener une étude « théorique » de ce que doit être une BCT.

HTL, l'interface hypertextuelle de consultation des résultats de *Lexter* (conçu par D. Bourigault, 1994) sert avant tout à visualiser et valider les candidats termes produits à partir d'un corpus. HTL permet aussi d'organiser les termes en familles de synonymes, désignées par un des termes, le terme-vedette, qui est

proche d'une étiquette de concept. Les vedettes peuvent d'ailleurs être associées par des relations conceptuelles étiquetées. De plus, le corpus est intégré dans HTL : il donne accès à toutes les occurrences d'un terme et permet de justifier l'organisation conceptuelle en fonction de l'usage des termes en corpus. En ce sens, les données contenues dans HTL se rapprochent d'une BCT.

La méthode d'analyse conceptuelle interactive (ACI) d'H. Assadi (1998) s'appuie sur une documentation technique et sur des outils d'analyse des termes contenus dans ce corpus pour élaborer une ontologie régionale documentée. Ce résultat est assimilable à une BCT, dans la mesure où les concepts sont organisés en réseau puis reliés à des expressions linguistiques et au corpus d'où ils sont extraits. De plus, la définition des concepts est locale au corpus. Dans le prototype d'atelier associé à la méthode ACI, le réseau conceptuel est représenté par des *frames*, gérés formellement au moyen de règles. Ce langage permet de construire progressivement l'ontologie et de garantir, par induction, la bonne organisation de la taxonomie des concepts et l'héritage de leurs propriétés.

Pour construire une BCT formelle, N. Capponi a envisagé une démarche en deux temps (1995, p. 11) : construire un modèle terminologique du domaine par analyse linguistique du corpus ; puis formaliser ce modèle pour en faire une BCT exploitable par un résolveur de problème. La formalisation utilise *Classic*, une logique de description, pour réduire une partie du modèle terminologique (des listes de candidats-termes produites par *Lexter*) puis le structurer en définissant des concepts et des relations. Cette BCT n'intègre pas le corpus et ne distingue pas explicitement terme et concept, puisque les informations linguistiques sont associées aux concepts, qui sont les classes du formalisme. De ce fait et à cause des contraintes liées au langage choisi, cette BCT convient mal pour enregistrer les résultats d'une analyse linguistique plus fine.

1.2 Des outils pour utiliser des connaissances terminologiques

D'autres travaux ont recours à la définition d'une BCT parce que cette structure de données semble adaptée à certains traitements sur les connaissances contenues dans les textes : les classer, les vérifier, les comparer, les spécialiser ou les généraliser. Les exemples retenus ici illustrent des applications variées mais étroitement liées au texte : recherche documentaire, vérification de spécifications décrites dans le texte. Le réseau conceptuel est alors directement formalisé dans un langage choisi en fonction des opérations à faire sur les connaissances dans l'application envisagée. Mais, de ce fait, ces outils privilégient les utilisations du réseau conceptuel, et non les besoins liés à la construction de la BCT. Nous ne mentionnons pas certains langages et environnements de construction d'ontologies, qui gèrent les termes désignant les concepts, car ces travaux s'éloignent de notre objectif.

Terminae, logiciel développé par B. Biébow et S. Szulman (1997), sert avant tout à construire, à partir de textes, des ontologies pour une application donnée, la vérification de spécifications. Ce logiciel permet d'étudier les occurrences des termes en corpus, de les organiser et de les formaliser dans une BCT. À partir d'une lecture du texte, les concepts sont organisés dans une hiérarchie différentielle, ce qui permet de les classer et de valider leurs descriptions. Les termes associés et les liens vers le texte sont conservés. Le formalisme utilisé, une logique de description, permet ensuite de raisonner sur ces connaissances ou de les traduire vers d'autres formalismes.

CG-KAT, logiciel développé par P. Martin (1995), permet de décrire formellement des informations pour faciliter la recherche de connaissances dans les documents d'où elles sont extraites, de présenter le document sous forme hypertextuelle et de l'indexer. Ces informations sont représentées à différents niveaux de granularité (concept, phrase, paragraphe, etc.), selon différents points de vue. À la fois proches du niveau linguistique et formalisées à l'aide de graphes conceptuels (GC), elles constituent une structure apparentée à une BCT mais inexploitable dans un autre contexte. Ces graphes sont construits par analyse automatique des textes (contenu et forme), puis enrichis en exploitant le réseau terminologique *Wordnet*.

Plus général, le logiciel *Hytropes*(3) est une interface de gestion de bases écrites en *Tropes*, un langage de *frames* proposé par J. Euzenat (1996). Il autorise la formulation de requêtes pour retrouver les concepts d'un domaine (les classes d'objets *Tropes*) selon des critères sémantiques (leurs champs ou attributs). *Hytropes* facilite la visualisation et la structuration des concepts, de leurs sous-classes et des différents points de vue associés. Il permet donc de gérer des ontologies à rapprocher de BCT dans la mesure où des textes sont associés aux concepts (et inversement), de manière à illustrer des définitions, justifier la représentation des connaissances et faciliter la navigation au sein des connaissances.

(3) Logiciel visible sur l'Internet à l'adresse http://ksi.cpsc.ucalgary.ca/KAW/KAW96/euzenat/euzenat-demo.html

Dans la lignée de Code4 (1994), le logiciel *DocKMan* (4), développé par D. Skuce (1998), permet de modéliser les connaissances présentes dans des textes afin de les rechercher par requêtes. *DocKMan* propose un niveau d'abstraction intermédiaire pour représenter ces connaissances avant d'en construire une ontologie formelle. Son originalité est de gérer des assertions, unités de connaissances tirées du texte, plus riches que le terme ou le concept. Chaque assertion traduit de manière structurée un terme et toutes les informations linguistiques relatives au texte associé (sujet/prédicat/objets). Explorer les connaissances présentes dans le texte se fait ensuite par des requêtes sur la base des assertions, qui renvoient les différentes utilisations des termes. Une base de ce type est amorcée par une analyse syntaxique automatique et sommaire du texte, puis enrichie par un linguiste qui organise et décrit précisément les assertions, au plus près du texte.

2 Le logiciel *Géditerm* : présentation

2.1 Prise en compte des besoins des terminologues

Notre objectif a été de fournir aux terminologues *Géditerm*, un outil de saisie de connaissances pour les organiser en une BCT. Le terminologue peut extraire ces connaissances d'un corpus en utilisant ses compétences et techniques de linguiste, sa connaissance de la langue de sens commun et ses outils de traitement automatique de la langue.

(4) Logiciel accessible sur l'Internet à l'adresse http://dkm.site.uottawa.ca/beta/dockman.html

Méthodologie de référence

Notre logiciel a été spécifié dans le contexte d'un projet de recherche, au cours duquel les linguistes-terminologues étaient en train de mettre au point leur démarche. Celle-ci n'était donc pas figée en début de spécification, et il a été difficile d'imaginer des fonctionnalités adaptées. Finalement, le travail de définition des spécifications a participé à la réflexion méthodologique, en obligeant à préciser les méthodes de travail et la façon dont les données sont renseignées. Les linguistes savaient quels outils elles utilisaient, comment elles procédaient de manière manuelle, mais n'avaient jamais mis à plat leur démarche, ni caractérisé leurs productions lorsqu'elles utilisent des logiciels. Le fruit de cette réflexion a été publié par A. Condamines et J. Rebeyrolle (1997). Avant d'exposer les fonctionnalités de *Géditerm*, nous présentons la démarche suivie pour recueillir les besoins, les besoins retenus et les choix relatifs à la représentation des connaissances.

Démarche suivie

Au centre de notre démarche de recueil des besoins, nous avons conduit une série d'entretiens auprès des linguistes puis des réunions de travail centrées sur le modèle de données. Ces réunions ont eu pour support une formulation écrite des spécifications, affinée et corrigée tout au long des échanges. Très rapidement, une maquette a été développée pour suggérer des fonctionnalités, vérifier leur adéquation aux méthodes de travail et aux besoins des linguistes en phase de construction d'une BCT. À partir de là, des modèles UML ont été construits pour organiser les spécifications, assurer une bonne documentation du projet et minimiser les erreurs lors des évolutions.

Une autre référence pour nous aider à spécifier les fonctionnalités a été l'analyse comparée et l'évaluation d'un logiciel proche, l'interface HTL présentée ci-dessus, que les linguistes ont utilisé pour valider la liste des candidats-termes identifiés par *Lexter*. HTL autorise une présentation des termes et la saisie de certaines de leurs propriétés en listes. Les linguistes n'ont pas utilisé ce type de présentation. Mais les termes eux-mêmes sont aussi accessibles sous forme de fiche rassemblant toutes les informations qui les concernent. Lorsqu'une vedette est associée à un terme, on peut consulter aisément l'un puis l'autre. Les différentes informations rattachées à une vedette (termes associés et relations conceptuelles) sont présentées sur un même écran. Une évaluation positive de ces dernières fonctions nous a conduit à favoriser une saisie par fiche ou carte plus que par liste.

Le modèle de données

Les besoins correspondent d'abord aux exigences théoriques (classiques) relatives à la séparation des notions clés de terme et de concept, ainsi que leur relation avec le corpus, et touchent donc le modèle de données. Ainsi, le système doit distinguer les mots en contexte des structures de terme et de concept, tout en rendant compte de la complexité de leur couplage. Les termes étant enregistrés sous une forme «standard» proche du lemme, il faut conserver leurs formes brutes pour les retrouver dans le corpus et accéder à leurs différentes occurrences. Une originalité de *Géditerm* est de conserver le corpus comme partie intégrante de la BCT, puisque, de fait, il définit le domaine couvert par cette BCT. Les informations présentes dans le corpus, l'usage des termes, donnent leur sens

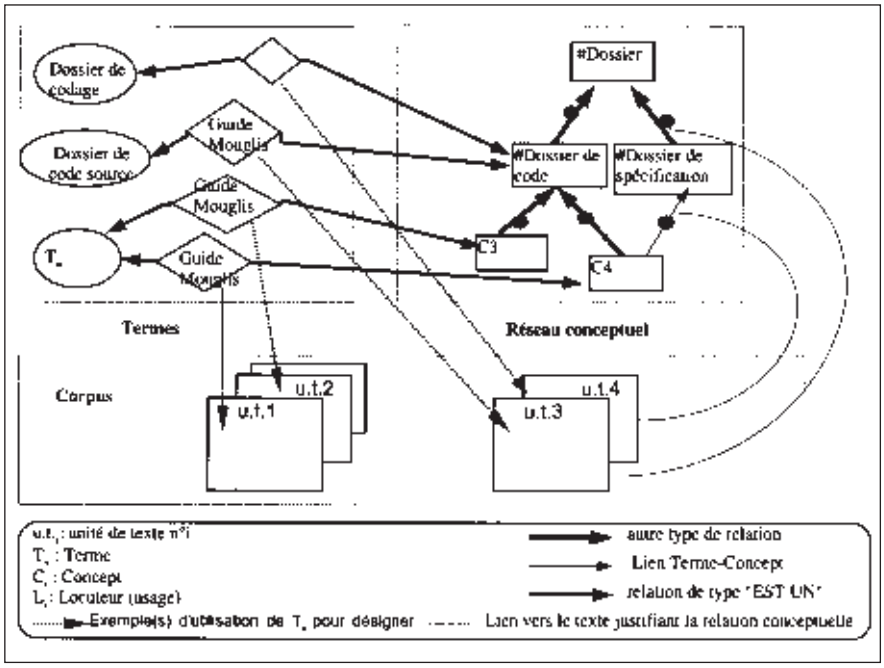


Figure 1 :
Organisation des données dans la BCT. Le réseau conceptuel, à droite, est relié aux données terminologiques, à gauche, par des liens qui renvoient à des parties du texte justifiant le fait que ce terme désigne ce concept. Les liens vers les textes partant des relations conceptuelles renvoient à des occurrences illustrant la relation.

aux concepts associés, dont les définitions ont donc une validité locale au corpus et non universelle. L'organisation des concepts est «stabilisée mais provisoire »: c'est celle révélée par le texte.

Les données terminologiques et conceptuelles ainsi que les relations entre ces données et le corpus sont enregistrées dans la BCT, selon un modèle de données qui a fait l'objet d'une étude approfondie. Une première version de ce modèle, sa justification pratique et théorique, ont été présentées par N. Aussenac et P. Séguéla (1997). Le modèle stabilisé correspondant à l'organisation actuelle des données dans le logiciel fait l'objet de la figure 1 (Simon 98). Pour organiser la base de données de la BCT, et en vue d'anticiper la prise en compte de concepts matérialisés par des sons ou des images, le modèle conceptuel de la base comprend une entité signifiant (à ne pas prendre au sens de strict de Sausssure) dont les entités terme, image et son sont des spécialisations. Dans la suite, nous assimilerons terme et signifiant.

Besoins relatifs aux fonctionnalités

Une autre partie des besoins exprimés concerne les fonctionnalités de saisie, mise à jour et consultation des données de la base. On suppose que le logiciel est utilisé soit pour saisir à la main des termes et des concepts un à un, décrits au fur et à mesure, soit qu'il récupère des données déjà validées, produites par un logiciel. L'élimination systématique de candidats qui ne seraient pas des termes parmi les listes produites par des logiciels d'extractions de candidats termes comme *Lexter* ou *Nomino* doit se faire en amont de l'utilisation de *Géditerm*. De ce fait, on privilégie un affichage et une saisie par fiches et non sur des listes (comme c'est le cas dans HTL par exemple).

Une autre des caractéristiques du travail des linguistes est de progresser dans la description et dans la compréhension du domaine en s'intéressant successivement à des concepts ou à des types de relations conceptuelles à partir de leur traces dans le corpus. Ces concepts sont repérés par des groupes de termes ayant le même comportement en contexte et définissant des classes conceptuelles plus ou moins larges. Pratiquement, les linguistes ont donc besoin de rechercher des termes reliés, des synonymes ou antonymes, de vérifier qu'un mot dans différents contextes correspond au même terme et ensuite au même concept, etc. Ils souhaitent aussi sauvegarder toutes les relations associées à un concept dans le corpus.

Enfin, le logiciel sera d'autant plus utile qu'il sera simple à utiliser et qu'il permettra à ses utilisateurs de vérifier et récapituler rapidement ce qui a été saisi. Une visualisation du réseau conceptuel sous forme de graphes, déjà évaluée par les linguistes dans le cadre de précédentes études, semble être un moyen efficace de montrer et valider le réseau conceptuel. La sélection de sous-ensembles des données en fonction de critères sémantiques avant leur affichage est indispensable pour éviter la visualisation d'énormes graphes.

Parmi les besoins implicites, il a fallu clarifier le type d'aide, de guide que le logiciel doit apporter, que ce soit à travers la représentation des connaissances, l'enchaînement ou la nature des fonctionnalités. Avec les linguistes, nous avons choisi de ne pas définir le logiciel comme un processus actif qui impose pas à pas de suivre une méthode particulière ou d'utiliser des fonctionnalités dans un certain

ordre. Le logiciel doit proposer un ensemble de fonctions d'édition, appelées à la demande selon les informations que l'utilisateur veut enregistrer ou modifier. Par contre, il doit pouvoir consigner des éléments méthodologiques, comme les marqueurs lexico-sémantiques ayant servi à identifier les relations conceptuelles, comme les extraits du corpus justifiant les relations conceptuelles (exemples) ou justifiant les relations terme-concept (occurrences pour lesquelles le terme désigne le concept).

2.2 Principes retenus pour la représentation des connaissances

Un des principes importants retenus pour la représentation des connaissances a été de préférer une représentation non formelle, qui rende compte du réseau conceptuel sans permettre de raisonner ou d'inférer sur ces connaissances. Pour décider de ce choix, nous avons tenu compte de la pratique des linguistes que nous avons confrontée aux logiciels présentés dans la partie 1. Les rapports de S. Simon (1998) et P. Séguéla (1996) rendent compte de nos évaluations de différentes représentations formelles des connaissances. Reprenant ici leurs conclusions, nous expliquons pourquoi nous considérons qu'utiliser un langage formel est trop contraignant pour le linguiste qui construit une BCT.

Tout d'abord, la formalisation contraint le linguiste à répondre à des questions dont la réponse n'est pas forcément dans le texte ou est ambigüe. En effet, la description formelle des connaissances oblige à poser des critères de définition des concepts. Créer un nouveau concept dans la hiérarchie EST-UN doit se justifier par la présence d'un attribut ou d'une relation (rôle) qu'il ne partage ni avec ses frères ni avec son père. Pour toute notion repérée à partir d'un terme, il faut d'abord décider comment la représenter (sous forme de concept, de relation ou d'attribut). Ensuite, si un concept est identifié, il faut trouver des indicateurs pour le placer systématiquement au bon endroit dans la hiérarchie EST-UN, puis le décrire en établissant des relations ou attributs qui le différencient de ses frères ou de son père; savoir si ces relations et attributs lui sont propres ou sont hérités, etc. Or ces connaissances ne sont pas toujours dans les textes. Elles doivent être demandées aux experts. De plus, il est difficile de garantir une «bonne définition» des concepts en l'absence de finalité précise, c'est-à-dire de critère de décision pour trancher sur la définition à retenir. Or le linguiste ne se préoccupe pas encore à ce niveau de la finalité de la BCT.

Ensuite, la formalisation impose d'avoir une vision globale des connaissances à représenter, pour distinguer d'abord les concepts et relations dits primitifs de ceux qui seront définis à partir des premiers. De plus, il faut définir les concepts dans un ordre lié à l'organisation conceptuelle des données (placer les concepts les plus généraux puis les spécialiser) et non dans l'ordre où les données sont trouvées dans le texte. Au contraire, le linguiste dépouille le corpus progressivement et souhaite les représenter au fur et à mesure, alors qu'il ne possède pas tous les éléments nécessaires. Le linguiste a donc besoin d'une structure souple, peu contraignante, qui joue le rôle d'un outil d'annotation de résultats.

La représentation formelle des connaissances serait donc utile pour aider le linguiste à procéder de manière systématique, à ne rien oublier. Mais elle exige d'anticiper l'utilisation qui sera faite des données, d'avoir recours plus souvent aux experts du domaine pour des validations et pour compléter le corpus. De ce fait, elle conduit à enregistrer des connaissances plus éloignées du corpus, parfois sans justification linguistique, ce qui n'est pas notre objectif premier. Envisageable dans un deuxième temps, nous l'avons prévue dans l'environnement d'exploitation de BCT que nous sommes en train de développer.

2.3 Fonctionnalités

Paramètres du logiciel

De manière à pouvoir utiliser le logiciel dans différents projets, un certain nombre de données de l'application sont paramétrables: les différents statuts de validité donnés aux termes ou aux concepts, la langue du corpus, les différentes catégories grammaticales, genres et nombres possibles des termes dans cette langue, le nombre de phrases que l'on veut voir simultanément lorsqu'on consulte le corpus.

Stocker, enregistrer

Les données enregistrées correspondent aux éléments du modèle de données: *signifiants* (termes), *concepts*, *types de relations conceptuelles* et *relations conceptuelles*, *liens* entre signifiants et concepts, *textes*, ainsi que quelques éléments méthodologiques. L'enregistrement du corpus requiert un traitement préalable pour le découper en unités. Nous avons repris sur le format utilisé dans HTL, un découpage en phrases ayant chacune un code identifiant. Nous avons envisagé assez tard d'intégrer des éléments méthodologiques, alors que cela semble un atout important pour permettre d'enrichir la méthodologie des terminologues et, plus immédiatement, pour assurer une maintenance correcte de la BCT. Il s'agit essentiellement des marqueurs lexico-sémantiques utilisés pour

repérer des traces de relations conceptuelles dans le corpus. Ils serviront à retrouver ces relations sur de nouveaux textes si jamais le corpus évolue par exemple. De plus, dans chaque structure de donnée, le linguiste dispose de zones de commentaires dans lesquelles il peut consigner la justification des choix de modélisation et des occurrences.

Pour accélérer l'inventaire des termes et des unités de texte à consigner dans une BCT, il est préférable d'utiliser en amont un outil d'extraction de candidats termes comme *Lexter* ou *Nomino. Géditerm* offre une fonction permettant d'intégrer dans la BCT le corpus, une liste de termes candidats (préalablement validés) et même des hypothèses de concepts tirés de HTL. Cette fonction déclenche un transfert des données systématique puis une rapide validation interactive par le linguiste pour corriger leur organisation.

Consulter ou modifier : des listes aux cartes

Accès aux listes

Les composants de base sont : les signifiants, les concepts, les types de relation et les textes. Les liens signifiants-concepts et les relations conceptuelles ne sont jamais visibles indépendamment des entités qu'ils mettent en relation. Ces informations sont accessibles à partir de listes de noms, complètes ou partielles lorsqu'elles sont filtrées selon des critères précisés par l'utilisateur. Les listes de concepts ou de signifiants affichent par défaut l'intégralité des noms des structures correspondantes de la BCT. Depuis une liste, on peut créer un nouveau composant de ce type ou bien demander à consulter un composant existant.

Accès par jeux de cartes

Choisir un nom dans une liste permet de visualiser le composant associé (un signifiant, un concept, un lien terme-concept ou une relation conceptuelle), sous forme d'un écran de saisie appelé « fiche » ou « carte » dans la suite. La carte présente les informations spécifiques à cette donnée et ses composants reliés : pour un signifiant, la liste des concepts qu'il désigne ; pour un concept, la liste des signifiants qui le désignent et la liste des concepts reliés par des relations conceptuelles. Sélectionner un de ces composants permet d'accéder à une nouvelle carte le décrivant. Les cartes se superposent (3 au maximum) à l'écran et des onglets permettent de passer rapidement des unes aux autres. L'utilisateur dispose alors des informations relatives à plusieurs composants ayant une parenté sémantique : soit un triplet concept-relation-concept, soit un triplet terme-lien-concept. L'accès au texte se fait alors depuis les liens ou les relations.

Sur la figure 2, la carte présente les informations relatives au concept *#acteur externe*. Les autres cartes présentes contiennent les informations relatives au concept *#humain*, à un signifiant, au lien entre le signifiant et l'un des concepts *#acteur externe* ou *#humain* ainsi que la liste des concepts. En cliquant sur l'un des onglets, la carte correspondante apparaît.

Aider à la définition et à la structuration des données

Guider l'organisation des données

Pour guider l'organisation des concepts, les relations conceptuelles sont typées, et l'ensemble des types de relations est répertorié, organisé en hiérarchie et accessible via des listes ou sous forme d'arbre (Figure 3). L'organisation hiérarchique est supposée refléter le caractère plus ou moins général des relations et s'appuie sur les liens de spécialisation tels que les interprète le linguiste. Tout nouveau type de relation doit être ajouté à la hiérarchie et défini par le type des concepts qu'il relie avant d'être utilisé.

Faciliter les choix

Géditerm simplifie la saisie d'informations en présentant des listes lorsqu'il s'agit de choisir parmi un ensemble fini de données. En favorisant de plus un repérage rapide

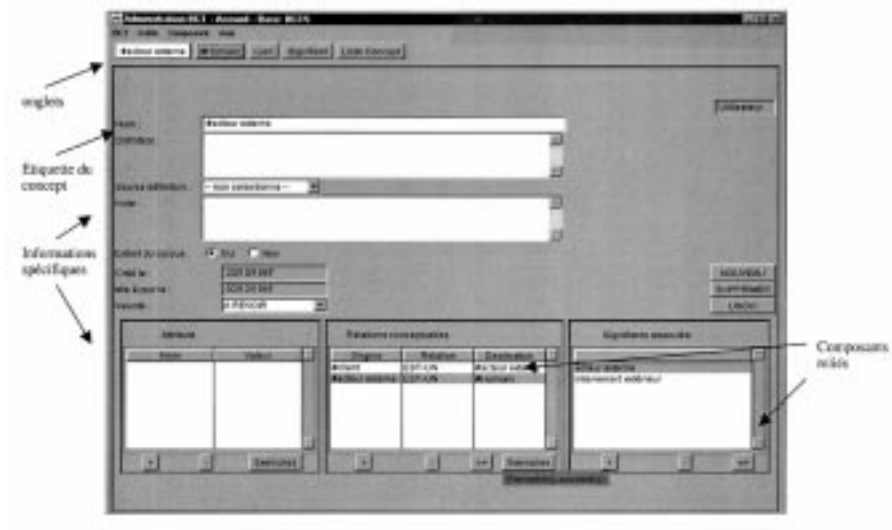


Figure 2 :
Exemple de carte de concept active : le concept *#acteur interne*.

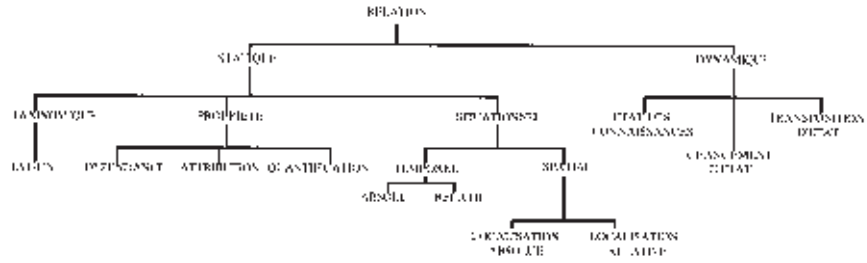


Figure 3:
La hiérarchie des types de relations conceptuelles

dans les listes (recherche des premiers caractères d'un mot) et en autorisant le paramétrage des options à choisir dans ces listes, l'interface essaie d'alléger les tâches les plus fastidieuses. Par exemple, le paramétrage permet à l'utilisateur de compléter la hiérarchie des relations avec des types de relations propres au corpus. Ensuite, lorsqu'il pose une relation entre deux concepts, il choisit son type parmi les types de cette hiérarchie.

Aider à sélectionner des données

Pour rechercher des données, on peut appliquer des filtres, ensembles de critères sémantiques, sur les listes. Ces critères portent sur les attributs des données et sur les relations entre données. Par exemple (Figure 4), les critères de sélection sur une liste de signifiants peuvent être un syntagme nominal; une variante de forme (ellipses, abréviations et formes les plus utilisées d'un terme); un locuteur; un concept (les signifiants retenus auront au moins un lien avec ce concept); un degré de validité.

Visualisation graphique

De même, on peut filtrer les données de la BCT avant de créer une vue qui sera visualisée graphiquement. Ainsi, on peut sélectionner les concepts ou les termes reliés à un concept précis ou fixer un type de relation particulier. La figure 5 présente un sous-ensemble de la BCT Mouglis: ce sont tous les concepts reliés par une relation autre que EST-UN au concept *#cycle de développement produit*.

Visualisation du corpus

La consultation du texte de référence de la BCT se fait soit depuis un composant du modèle (lien terme-concept ou relation conceptuelle), soit depuis le menu principal. Les extraits du corpus sont affichés par groupes ou paragraphes (Figure 6) alors que le texte ne provenant pas du corpus est affiché de manière isolée. Un paragraphe comprend une unité textuelle et les *n* unités textuelles qui la précèdent et la suivent (*n* peut être modifié). On accède à la demande aux paragraphes précédant et suivant le paragraphe courant.

Vérifier, valider

Les spécifications ont établi l'ensemble des informations à vérifier, mais surtout les principes à suivre et le moment de ces vérifications. Nous avions le choix entre des vérifications systématiques dès qu'une structure (terme ou concept) est modifiée, des vérifications à la demande sur ces structures ou des vérifications massives sur des listes (portant sur les structures complètes ou par critère). Nos premiers choix étaient de retenir une validation morcelée et systématique après la mise à jour de données. Pour cela, toute donnée possède un degré de validité, qui peut prendre plusieurs valeurs dont une seule est valide. Lorsqu'on valide une donnée, le processus de vérification est déclenché à la fermeture de la fiche, et les informations manquantes ou incorrectes sont signalées. Voici quelques exemples de vérifications prévues:

• Les relations conceptuelles par rapport à leur sémantique: afin de garantir que les étiquettes des types de relations conceptuelles soient interprétées de la même façon au cours de la construction de la BCT, puis à son utilisation, les types de relations comportent une «signature», c'est-à-dire que l'on indique les classes sémantiques (concepts de haut niveau) qu'ils relient. Toute relation spécifique entre deux concepts doit donc relier des concepts fils (indirectement dans la hiérarchie des concepts) de ces classes.

• La place des concepts dans la hiérarchie EST-UN. Tout concept doit, en fin de construction de la BCT, être situé dans la hiérarchie

Figure 4:
Fenêtre de définition d'un filtre pour la liste des signifiants

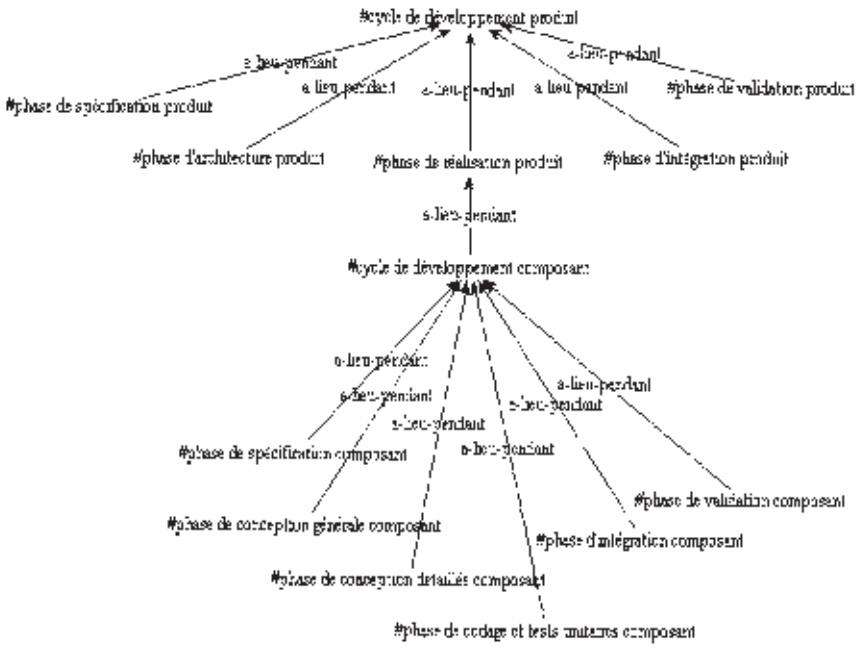


Figure 5 :
Extraits du réseau conceptuel de la BCT Mouglis.

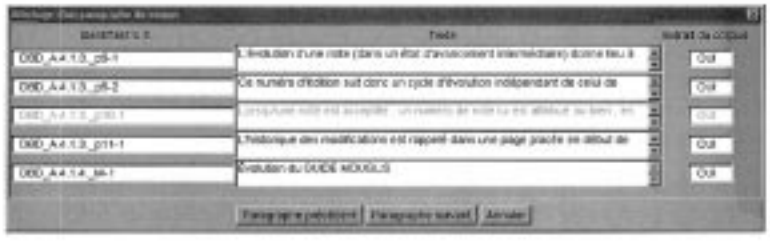


Figure 6 :
Fenêtre de consultation du corpus : un paragraphe d'unités textuelles avec n=2

EST-UN. Afin d'éviter un travail massif en fin de processus, cette vérification, initialement prévue à la création de chaque concept, n'est lancée qu'à la demande.

• La description des structures (termes ou concepts). Les spécifications ont conduit à imposer qu'un certain nombre d'informations minimales soient connues pour toute structure de données, comme son nom, mais aussi un terme relié pour un concept, sa langue pour un terme, etc. On considère qu'une BCT n'est terminée que si ces informations minimales ont été renseignées pour tous les termes et concepts qu'elle contient. Cette vérification massive n'est lancée qu'à la demande.

Ce choix a été remis en question par la pratique des linguistes, qui, à l'utilisation de *Géditerm*, n'ont jamais déclaré les données valides afin d'éviter les vérifications. En effet, celles-ci sont trop contraignantes et demandent des connaissances parfois non disponibles. Ces principes de validations correspondent en fait à des critères de « bonne modélisation », adaptés du principe de différenciation en vue de préparer une formalisation. Or, comme nous l'avons dit plus haut, l'application de ces règles directives requiert des connaissances pas toujours présentes dans le corpus.

2.4 Scénario d'utilisation

Pour illustrer l'utilisation des fonctions disponibles dans *Géditerm*, nous décrivons sur un exemple comment enrichir la BCT. L'utilisation de *Géditerm* suppose d'avoir sélectionné le corpus et d'en disposer sur support informatique, découpé en phrases avec des identifiants. Si on a analysé le corpus à l'aide de *Nomino* ou *Lexter*, il est préférable de filtrer ces listes de candidats-termes avant de les importer dans la BCT.

Les points de départs de toute recherche ou saisie d'information sont les listes d'étiquettes(5) de concepts et de signifiants (termes). Que ce soit directement ou à partir d'un élément relié, elles permettent de créer de nouvelles données ou de retrouver un composant. Supposons que les linguistes en soient à la phase d'identification des concepts et des relations conceptuelles. Ils appliquent des marqueurs sur le corpus, comme par exemple le marqueur suivant (TC = Terme candidat) :
Dét_déf TC1 Vêtre présent dét_indéf (partie, élément, composant...) de TC2

(5) Sur les cartes de saisie des signifiants (termes) et des concepts, le nom renvoie à l'étiquette et non à la catégorie grammaticale du terme (qui peut être un groupe nominal ou un verbe).

Ce marqueur permet de trouver le contexte suivant : *La documentation de réalisation est une composante essentielle du produit logiciel*, dans lequel « documentation de réalisation » et « produit logiciel » sont des termes qui vont être associés à des concepts. Ce contexte indique une relation de « composition » entre ces deux concepts. Enfin, d'autres occurrences de « document de réalisation » permettent d'établir que ce terme est synonyme de « dossier de réalisation ». Les termes « dossier de réalisation » (figure 7), « documentation de réalisation » et « produit logiciel » sont déjà présents dans la BCT car identifiés par *Lexter*, ainsi que le concept *#dossier de réalisation*. Le linguiste doit créer un nouveau concept à partir du terme « produit logiciel », et relier le terme « documentation de réalisation » au concept *#dossier de réalisation*.

Ensuite, il doit poser la relation conceptuelle « est-un-composant-de » entre ces deux concepts. Supposons que ce type de relation n'ait pas encore été défini. Il doit tout d'abord le créer en modifiant la hiérarchie et situer cette relation par rapport aux relations existantes. Il doit également définir cette relation (figure 8) en indiquant les types de concepts qu'elle relie. Enfin, depuis la carte d'un des concepts concernés (figure 2), il utilise le bouton « + » associé à la liste des relations conceptuelles pour définir la relation.

Figure 8 :
Interface de définition d'un type de relation

Ainsi de suite, le linguiste travaille par association sémantique, en allant d'un concept à un autre ou à un terme en fonction des données extraites du corpus. Il peut vouloir accéder à un concept ou un terme inconnu, mais caractérisé par des propriétés connues. Pour cela, à partir d'une liste, il utilise les filtres pour la réduire en fonction des propriétés recherchées, puis sélectionner par le nom le composant recherché ou une entité reliée. Enfin, pour visualiser le réseau déjà saisi, en partie ou en totalité, selon des critères particuliers, il utilise la visualisation graphique. Il faut d'abord créer une vue (sélectionner des concepts ou des termes), lancer le logiciel *Graphlet* puis visualiser la vue sous forme de graphe (figure 5). De là, on peut corriger la présentation à l'écran à l'aide d'algorithmes ou à la main, mais on ne peut pas encore modifier les données.

Figure 7 :
La carte du terme dossier de réalisation donne accès au concept *#dossier de réalisation*

3 Discussion : impact de la réalisation d'un logiciel sur la réflexion théorique

3.1 Évolutions du modèle de données

Bien que les grandes lignes du modèle des données reprennent les propositions énoncées par A. Condamines et P. Amsili (1993), la réalisation concrète d'un logiciel a obligé les linguistes à préciser comment décrire termes et concepts,

quelles informations associer au lien terme-concept et surtout comment gérer les notions d'usage et de point de vue. La notion d'usage n'existe plus en tant que telle, mais ce sont les liens entre termes et concepts qui, renvoyant à du texte, illustrent les usages. De même, la notion de point de vue n'est pas gérée explicitement. Par contre, un lien terme-concept est valide pour un ou plusieurs locuteurs. Ces locuteurs peuvent avoir des points de vue sur un concept. Dans ce cas, le terme qu'ils utilisent est polysémique : il prend un sens différent dans différents contextes. Pour les linguistes, il est important de consigner dans ce cas le point de vue dominant pour ce locuteur, c'est-à-dire le sens qu'il donne le plus couramment à ce terme. On doit donc indiquer sur les liens entre termes et concepts quels locuteurs utilisent ce terme pour désigner ce concept et, pour chacun d'eux, s'il s'agit d'un point de vue dominant.

Une dernière évolution du modèle des données a été de renforcer les liens vers le texte par rapport à ce qui était initialement prévu. Ainsi, des exemples peuvent être associés aux relations conceptuelles et aux attributs des concepts. Ils correspondent à des contextes dans lesquels on trouve des marqueurs de cette relation. Ces exemples justifient les relations conceptuelles et aident à les interpréter.

3.2 Différenciation de deux types de BCT

La construction d'une BCT avec *Géditerm* nous a conduite à vouloir renforcer le caractère neutre de l'analyse linguistique. Il nous semble important de marquer la rupture qui existe entre le point de vue du linguiste, que nous avons cherché à respecter, et celui du cogniticien qui modélise les connaissances. Le linguiste a le souci de ne pas biaiser l'interprétation du texte en construisant une BCT. Pourtant, on sait bien que, du moment qu'il traduit son interprétation dans une structure de données, il fait des choix et construit une abstraction qui n'est pas complètement neutre par rapport au texte. De plus, pour rendre le modèle exploitable par une autre personne, l'organisation des connaissances doit être commentée. Nous appelons tout de même ce résultat une «BCT-Corpus». Nos réflexions sur la représentation des connaissances et sur les fonctions de validation nous ont confirmé qu'une BCT-C doit être informelle et incomplètement structurée.

Au contraire, si, dans un deuxième temps, on construit une BCT en vue d'une application précise, on parlera de «BCT-Applicative». Les informations y sont organisées en fonction des besoins de l'application, peuvent différer du contenu du texte, et méritent d'être formalisées pour être mieux exploitées. Les systèmes comme *CG-KAT, Terminae* ou *Hytropes* sont, pour nous, des outils pour construire directement des BCT-A.

3.3 Quelle validité pour les données d'une BCT ?

Dans la BCT-C, c'est le corpus qui sert de justification à l'organisation des connaissances. Un autre élément de validation, plus implicite et moins facile à cerner, est la connaissance de la langue que possède le terminologue. Pour le moment, seule une vérification structurelle est possible. Or, si on veut pouvoir exploiter les connaissances, il faut imposer rapidement des critères de définition et vérifier qu'ils sont respectés. Il faut également faire valider les connaissances par des experts compétents par rapport à l'application envisagée. Il est clair que l'outillage formel est particulièrement adapté dans ce cas, et que la traduction formelle des données du réseau conceptuel est un moyen de s'assurer de la validité des relations entre concepts par rapport à leur définition, du respect des critères de différenciation et de la bonne utilisation de la hiérarchie EST-UN pour organiser les concepts.

3.4 Vers un environnement de consultation de BCT

Le logiciel *Géditerm* est adapté à la construction d'une BCT, mais il en offre une vue trop morcelée pour pouvoir en exploiter le contenu efficacement. Le fait d'adapter le contenu d'une BCT à des contraintes spécifiques relatives à une application requiert des outils et des fonctions bien différentes, comme la possibilité de réorganiser les concepts dans la hiérarchie, de les différencier plus systématiquement, de les représenter formellement. Cette remarque confirme le bien fondé de la séparation BCT-C / BCT-A.

Pour cela, nous avons commencé à prototyper un environnement de consultation de BCT, destiné à préparer des BCT-Applicatives à partir de BCT-Corpus, comme construire un index ou formaliser les connaissances pour construire une BC. Notre hypothèse est que les personnes qui consultent une BCT sont celles qui développent une application à partir des connaissances de la BCT, pas celles qui utiliseront ces connaissances dans leur travail. Ce prototype propose des fonctions d'aide à la sélection des données dans des BCT existantes, pour les modifier et les réorganiser, en fonction d'un besoin spécifique. Il permet ainsi de construire de nouvelles BCT en fonction d'objectifs applicatifs particuliers. Pour cela, nous avons prévu de faire expliciter puis appliquer des critères de définition précis. Ce logiciel doit produire des

données d'un format facilement exportable vers des applications. Pour le moment, ce sont des tables de bases de données. On prévoit également des sorties papier sous forme de rapport de formats divers.

4 Conclusion et perspectives

Tout comme les résultats méthodologiques, les logiciels font partie de l'outillage nécessaire à l'évaluation expérimentale des BCT. Le développement de *Géditerm* représente un résultat pertinent à double titre. Tout d'abord, la réflexion approfondie menée pour sa mise au point a permis d'aborder des questions fondamentales sur la représentation des connaissances (quelles structures utiliser pour représenter des connaissances avant de les rendre interprétables par la machine? Sous quelle forme présenter ces connaissances?) et de préciser le modèle de données dans la base.

Ensuite, même si le concept de BCT s'avère très pertinent, la quantité de données requises pour leur constitution puis l'exploitation des données qu'elles contiennent ne peuvent se faire que sur support informatique. Il est donc indispensable, pour confirmer l'utilisabilité de ce concept, de fournir aussi l'environnement qui le rende exploitable concrètement, dans l'esprit d'une plate forme terminologique (moins orientée vers la modélisation que la plate-forme d'ACI proposée par H. Assadi (1998). Sinon, il restera une belle construction intellectuelle sans diffusion possible en contexte industriel. Cet environnement doit prendre en compte la démarche de construction, et donc intégrer un éditeur comme *Géditerm*. Il doit comprendre aussi des logiciels automatisant les traitements sur corpus qui réduisent le coût de la démarche tout en maintenant le degré de validité des données. Un éditeur hypertextuel du document y est indispensable pour passer aisément du texte (d'un terme pris dans le texte) aux termes ou aux concepts, pour retrouver le texte sous sa présentation d'origine ou pour mettre en valeur les termes qu'il contient.

4.1 Perspectives techniques

Concernant le logiciel, il nous faut encore développer de nouvelles fonctionnalités, les unes liées aux mises à jour du modèle depuis la représentation graphique, les autres pour pouvoir imprimer des données sous forme de rapport. Nous devons également procéder à une évaluation systématique de *Géditerm* selon une démarche ergonomique. En particulier, il semble souhaitable de rendre plus simple et commode l'accès au corpus. Une autre piste est de rendre l'outil encore plus paramétrable, pour que le terminologue puisse ajouter ou retirer des propriétés aux termes et aux concepts en fonction des informations dont il a besoin pour les décrire. Initialement, nous avions fait l'hypothèse que ces informations soient indépendantes de l'utilisation prévue des données de la BCT. Or il est clair que le modèle actuel ne peut anticiper tout type de besoin. Par exemple, il serait insuffisant s'il fallait utiliser une BCT pour l'aide à la traduction.

4.2 Perspectives théoriques

Une première perspective concerne la représentation des connaissances. Le modèle de données actuel est limité pour rendre compte de relations complexes, faisant intervenir plus de deux concepts, ou pour rendre compte de liens possibles entre relations. Par exemple, dans une relation de découpage en partie, on voudrait pouvoir préciser quels concepts sont complémentaires et forment ensemble l'objet entier. Autre exemple: rendre compte de schémas de type agent/verbe/objet/moyen. Dans cette perspective, deux pistes sont à étudier. L'une, dans l'esprit de *DocKMan* de D. Skuce (1998), serait de conserver, dans la partie terminologique, des schémas syntactico-sémantique ou lexico-syntaxiques au lieu de syntagmes nominaux seuls; l'autre serait de construire, dans le réseau conceptuel, leur équivalent sous forme de *frames* sophistiqués rassemblant des sous-ensembles de réseaux. La difficulté est alors de ne pas trop figer ces représentations de plus grande taille pour pouvoir revenir à leurs composants, termes et concepts.

Une autre perspective est de poursuivre l'évaluation de l'intérêt des BCT pour construire des modèles d'un domaine, des modèles de produits ou des bases de connaissances. *Géditerm* est en effet le point de départ indispensable pour ce type d'évaluation, et il doit être utilisé dans des contextes différents pour bien évaluer toutes ces fonctionnalités et permettre de favoriser l'utilisation des BCT.

Remerciements

Ce travail est le fruit d'une collaboration étroite avec une équipe de l'ERSS de Toulouse (Anne Condamines et Josette Rebeyrolle), qui est à l'origine du modèle de données retenu et qui a contribué aux spécifications de *Géditerm*. De plus, je remercie vivement Dominique Fournier qui a développé ce logiciel.

Nathalie Aussenac-Gilles,
Irit – UMR 5505 du CNRS,
Université Paul Sabatier,
Toulouse,
France.

Bibliographie

Assadi (H.), (1998): *Construction d'ontologies régionales à partir de textes techniques: application aux systèmes documentaires*, Thèse de doctorat de l'Université Paris VI en Informatique. 286 p.

Bachimont (B.), (1995) «Ontologie régionale et terminologie: quelques remarques méthodologiques et critiques», dans *Banque des mots*, n° spécial Terminologie et intelligence artificielle, 7, p. 65-84.

Biébow (B.), Szulman (S.), (1997): «Méthodologie de création d'un noyau de base de connaissances en logique terminologique à partir de textes», dans *Actes des 2e rencontres Terminologie et intelligence artificielle. TIA'97*, Toulouse, avril 1997, p. 69-84.

Bourigault (D.), (1994): *Lexter, un logiciel d'extraction de terminologie: application à l'acquisition de connaissances à partir de textes*, Thèse de doctorat en informatique linguistique de l'EHESS, Paris.

Capponi (N.), (1995): *Modélisation d'une base de connaissances terminologiques.* Mémoire de DEA de l'Univ. de Nancy I. CRIN/LORIA, Nancy, 45 p.

Condamines (A.), Amsili (P.), (1993): «Terminologie entre langage et connaissances: un exemple de base de connaissances terminologiques», dans *Terminology and Knowlegde Engineering*. Francfort, 316-323.

Condamines (A.), Rebeyrolle (J.), (1997): «Construction d'une base de connaissances terminologique à partir de textes: expérimentation et définition d'une méthode», dans *Actes des Journées d'Ingénierie des Connaissances IC'97, mai 1997*, (INRIA), Roscoff (F), 191-206.

Condamines (A.), Rebeyrolle (J.), (1998): «Description d'une BCTC: base de connaissances terminologiques modélisée à partir d'un Corpus», dans *Workshop COGNITERM'98, International Conference on Computational Linguistics (COLING), August 1998*, Montreal, Canada.

Euzenat (J.), (1996): «Hytropes: a www front-end to an object knowledge management system», dans *Knowledge Acquisition Workshop, KAW'96*, Fiche démonstration, Banff, Canada.

Fournier (D.), (1998): *Étude et conception d'un logiciel de gestion de base de connaissances terminologiques*, Mémoire d'ingénieur CNAM, Toulouse.

Gros (C.), Assadi (H.), (1997): «Intégration de connaissances dans un système de consultation de documentation technique», dans *Acte des 1res journées du Chapitre Français de l'ISKO*. Lille, 16-17 oct. 1997, J. Maniez et W. Mustafa el Hadi (Eds), Édition du conseil scientifique de l'université Charles de Gaulle Lille 3, Collection «Travaux et Recherche».

Lemaire (F.), Rechenmann (F.), (1995): «Intégration de connaissances terminologiques dans les grandes bases d'objets – Exemple en biologie moléculaire», dans *Banque des mots. n° spécial Terminologie et Intelligence Artificielle*, 7, p. 103-112.

Martin (P.), (1995): «Knowledge Acquisition using Documents, Conceptual Graphs and a Semantically Structured Dictionary», dans *Proc. of the 8th Knowledge Acquisition Workshop, KAW'95*, Banff, Canada.

Meyer (I.), Skuce (D.), Bowker (L.), Eck (K.), (1992): «Toward a new generation of terminological Ressources: An Experiment in Building a Terminological Knowledge Base», dans *Proc. of the 13th International Conference on Computational Linguistics*, Nantes, p. 956-960.

Rastier (F.), (1995): «Le terme, entre ontologie et linguistique», dans *Banque des mots, n° spécial Terminologie et intelligence artificielle*, 7, p. 35-64.

Rousselot (F.), Frath (P.), Oueslati (R.), (1996): «Extracting concepts and relations from corpora», dans *12th European Conference on Artificial Intelligence, ECAI'96; Workshop on Corpus-oriented semantic analysis*, Ed. By W. Wahlster, Pub. by John Wiley & Sons Ltd.

Séguéla (P.), Aussenac-Gilles (N.), (1997): «Un modèle de base de connaissances terminologiques», dans *Actes des 2e rencontres Terminologie et Intelligence Artificielle*. TIA'97, avril 1997. Toulouse, p. 47-68

Simon (S.), (1998): *Représentation formelle de connaissances issues d'une base de connaissances terminologiques*, Mémoire de DEA, RCFR de l'Univ. P. Sabatier, Toulouse.

Skuce D., Lethbridge (T.C.), (1994): «CODE4: A multifunctional knowledge management system», dans *Proc. of the 8th Knowledge Acquisition Workshop, KAW'94*, Banff, Canada.

Skuce (D.), (1998): «Intelligent Knowledge Management: Integration documents, Knowledge bases, Databases and Linguistic Knowledge», dans *Proc. of the 10th Knowledge Acquisition and management Workshop, KAW'98*, April 1998, Banff, Canada.

Slodzian (M.), (1995): «Comment revisiter la doctrine terminologique aujourd'hui?», dans *Banque des mots, n° spécial Terminologie et intelligence artificielle*, 7, p. 11-18.

Exemple de pratique terminographique en entreprise

La constitution d'une terminologie de référence destinée à l'entreprise passe par plusieurs étapes: l'extraction des termes à partir du corpus, la validation au cours de laquelle les experts sont guidés par le terminologue et l'organisation des termes validés en domaines et sous-domaines. La dernière partie de cet article présente la difficulté de suivre certains principes de la théorie générale de la terminologie afin de produire une terminologie destinée à l'entreprise.

Termes-clés: corpus; domaine; extraction terminologique; pratique terminographique; théorie de la terminologie; validation.

1 Introduction

Suite à un projet de la DE (Direction de l'équipement d'EDF), une terminologie de référence a été amorcée. Cette dernière doit être composée de termes extraits de la documentation EDF et organisés en domaines et sous-domaines selon un cahier des charges établi par la DER (Direction études et recherches d'EDF).

La constitution d'une terminologie de référence entre dans le cadre du projet RMI (Référentiel méthodologique d'ingénierie) mené par la DE. Cette dernière se charge de la conception, la réalisation et l'exploitation des centrales nucléaires et mène le projet d'harmoniser et de mettre en cohérence l'ensemble des documents liés aux métiers de l'ingénierie: des doctrines, des dictionnaires, des thesaurus, des modèles de documents, etc. La DE souhaite faciliter, par le biais de la terminologie de référence, la rédaction de documents (contrôle terminologique, contrôle de la normalisation et de la cohérence et aide à la traduction) dans un premier temps et à terme, fournir des moyens avancés pour la consultation des documents techniques.

2 Description du corpus de référence

Le corpus que la DE a fourni et à partir duquel les termes sont extraits est constitué d'un ensemble de documents techniques, les DSE (Dossiers de systèmes élémentaires) destinés aux constructeurs de centrales nucléaires. Ces dossiers de spécifications sont multi-auteurs. Des spécialistes de chaque métier définissent les systèmes élémentaires du palier N4 (dernière génération de centrale nucléaire) et ce sous divers angles tels que la conception, la conduite, le fonctionnement et la sûreté. Ces thématiques sont reflétées de manière très explicite dans la structure même des DSE. En effet, ils sont tout d'abord organisés en ensembles désignés par des lettres de l'alphabet. Ces lettres peuvent faire référence à une zone de la centrale nucléaire comme la lettre [1] R pour le réacteur ou bien des traitements tels que [2] J pour l'incendie. Chaque groupe est divisé en plusieurs systèmes élémentaires. Ainsi, l'ensemble R compte 17 sous-ensembles tels que le [3] RCP qui concerne le circuit élémentaire ou le [4] RHY qui indique l'hydrogénation et les besoins nucléaires. Les sous-ensembles faisant référence aux différents systèmes élémentaires présentent une certaine homogénéité. Ils se composent de rubriques qui elles-mêmes se divisent en sous-rubriques. Cette structure est commune à une majorité de DSE. À telle enseigne que [5] le document D0 contient le sommaire général. Les sous-rubriques abordent notamment l'historique de ces documents, le fonctionnement du système élémentaire, son rôle, sa base de conception et la description des matériels.

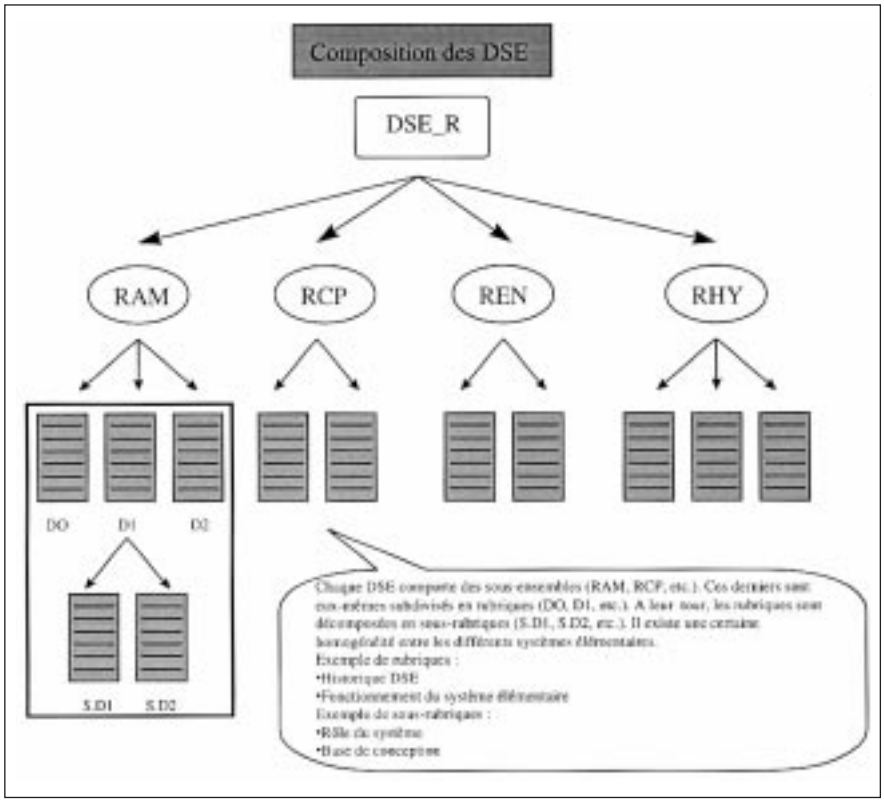


Figure 1 :
composition des DSE

Suite à une réunion avec les experts de la DE chargés de la validation au cours de laquelle une première terminologie composée de termes extraits des DSE a été présentée, le projet a quelque peu changé d'orientation. Compte tenu de la nature très technique des documents constituant le corpus de départ, les futurs valideurs ont jugé nécessaire de construire parallèlement une seconde terminologie illustrant le savoir-faire de l'entreprise à partir d'un nouveau corpus, le *Mir* (*Manuel d'ingénierie de référence*). Ce changement de cap dénote l'évolution constante du projet et le manque de stabilisation du corpus. Bien que le traitement du *Mir* ainsi que sa structuration ne soient pas abordés dans cet article en raison des priorités données pour l'instant dans le projet RMI à la terminologie technique basée sur les DSE, il est intéressant de noter la différence de contenu de ces deux corpus. Cette nouvelle source textuelle ne sera pas traitée sur le plan de la structure mais du contenu uniquement du fait que le corpus ne contient que des extraits du *Mir* et non la totalité.

Le *Mir* a été créé afin de répondre aux besoins de la DE essentiellement en matière de normalisation de documents d'ingénierie, de formalisation des méthodes d'ingénierie et de mise en cohérence entre projets, unités et types de documents propres à la DE. Il permet d'établir la structure documentaire de la DE et d'acquérir des connaissances au sein de cette direction.

Le *Mir* traite à la fois du savoir-faire tel que la méthode de fabrication de vannes et des objets techniques, à savoir les vannes en elles-mêmes. Il comporte tout d'abord les codes de classement des plans. Ils se présentent sous la forme d'une liste. Celle-ci est composée d'un repère (une valeur numérique) et de sa référence (du texte). Les codes désignent par exemple les terrains, les installations chantiers et les différents éléments entrant dans la construction des bâtiments. D'autre part, le Mir réunit des spécifications dont celle de la typologie des documents d'ingénierie. L'objectif de cette spécification est de répertorier et d'identifier les différents types de documents de contenu technique réalisés à la DE ou par des tiers.

3 Mise en place d'une méthodologie de travail

La constitution d'une terminologie à partir du corpus des DSE implique les trois étapes suivantes :
• Extraction des candidats termes par l'outil terminologique *Lexter* (Bourigault 1994) ;
• Validation par des experts de la DE ;
• Organisation des termes validés en domaines et sous-domaines.

3.1 Extraction des candidats termes : une méthodologie incrémentale

Cette méthodologie consiste à soumettre les DSE à un outil d'extraction terminologique *Lexter* afin d'effectuer l'extraction de candidats termes. Les listes obtenues sont soumises aux experts de la DE afin d'être validées et organisées en domaines et sous-domaines. Enfin, les termes sont intégrés dans la terminologie de référence. Compte

tenu de la taille du corpus des DSE (près de 18 000 pages de documents), il est préférable que le traitement s'amorce à partir d'un sous-corpus puis que la méthodologie adoptée soit systématisée afin d'être appliquée au reste des documents.

Le traitement itératif des données

Il ressort du traitement d'un sous-corpus test une méthodologie permettant une extraction et une validation plus efficaces.

Les ensembles DSE sont traités un à un. Le traitement *Lexter* d'un DSE N produit une liste de termes candidats que l'on compare à la liste de termes validés du DSE précédent (N-1). Cette phase de mise en parallèle représente une pré-validation destinée à faciliter le travail des experts de la DE. Elle consiste à attribuer la validité des termes du DSE N-1 existant dans le DSE N. Ce premier élagage permet d'obtenir une liste plus réduite qui fera l'objet d'une validation définitive toujours par les experts. Dès lors, les termes obtenus sont introduits dans la terminologie de référence.

Initialisation du traitement : comparaison Paluel – N4

Le traitement itératif est amorcé avec les DSE_R du palier *N4*. Le sous-corpus de départ est constitué de sept sous-ensembles du DSE_R (RAM, RAZ, REA, REN, RPR, RRI, et RRM). Ce choix a été motivé par la phase de comparaison. Une comparaison est faite avec les 7 sous-ensembles équivalents d'un palier antérieur, le *Paluel*. Ils permettent la pré-validation des candidats termes obtenus après l'extraction terminologique. Les DSE *Paluel* sont organisés de la même manière que les DSE_R du *N4* en ensembles de lettres et en sous-ensembles de systèmes élémentaires. Ils ont fait l'objet d'une validation dans le cadre d'un projet EDF. Ce travail de sélection explique le choix des documents *N4* faisant référence aux mêmes systèmes élémentaires. Les termes candidats ont été extraits à l'aide de *Lexter* puis la liste a été validée puis structurée en domaines par Henry Boccon-Gibod, ingénieur expert.

Les termes potentiels issus de la comparaison *Paluel – N4* sont soumis aux experts. Une fois validés, ils représentent la première version de la terminologie à laquelle seront ajoutées, au fur et à mesure des traitements, de nouvelles listes de termes provenant d'autres DSE. Le traitement *Lexter* des sept sous-ensembles du DSE *Paluel* avait permis d'obtenir 11 500 termes candidats. Après la validation de l'expert Henry Boccon-Gibod, 4 000 termes avaient été validés «oui» et 7 500 validés « non». Les sept sous-ensembles *N4* ont produit, après extraction terminologique, 24 500 termes potentiels.

La comparaison *Paluel – N4* permet d'obtenir les résultats suivants :

- Les termes communs aux deux paliers sont au nombre 3 500 dont 1 500 validés «oui »;
- Les termes candidats *Paluel* absents du N4 s'élèvent à 8 000 dont 2 500 validés «oui »;
- Les termes candidats *N4* absents de *Paluel* atteignent le nombre de 20 000.

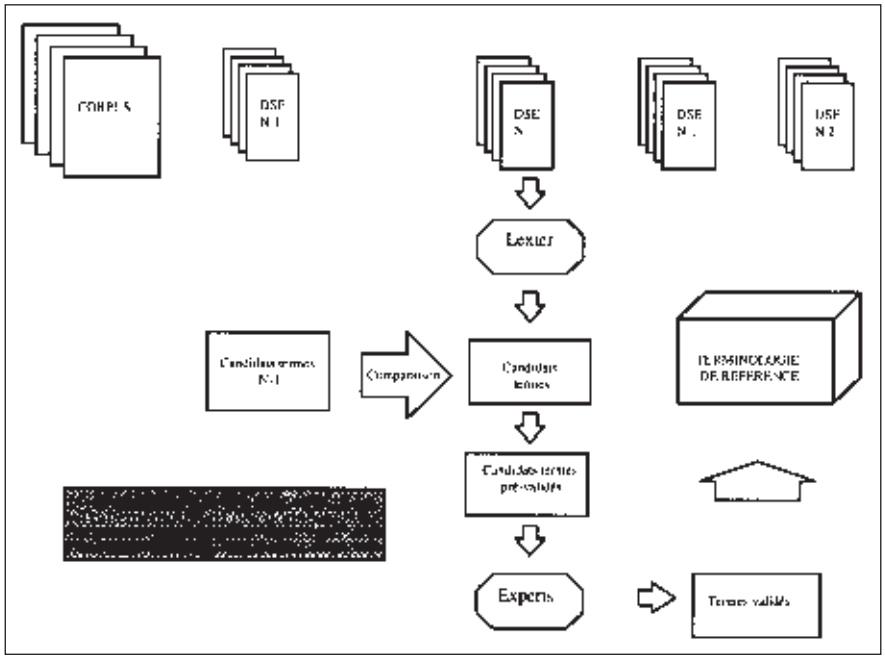


Figure 2 :
Traitement itératif des données

3.2 Travail de validation des experts : importance des consignes

La présentation aux experts de la DE d'un ensemble de termes extraits des DSE_R met en évidence la nécessité de produire des consignes de validation. Une interface de validation simple permet d'afficher les termes potentiels et de leur attribuer

quatre validités possibles : « non-validé », « oui », « non » et « à voir ». Les syntagmes nominaux sont au nombre de 3 300 et sont passés en revue par trois experts. Ces derniers valident près de 500 termes en deux heures. 534 termes sont vus par au moins un expert. Les résultats du travail de validation sont les suivants :

- 111 sont validés oui/oui/oui ;
- 70 sont validés non/non/non ;
- 343 sont litigieux.

Au cours de la phase de validation, les experts mettent l'accent sur l'absence de consignes de validation précises. En effet, les seules recommandations transmises aux valideurs sont de sélectionner les termes candidats dans la perspective d'une intégration dans une application d'aide à la rédaction et à la compréhension de documents techniques et de ne pas en éliminer trop, certains, bien que très généraux, pouvant constituer des noeuds dans la terminologie de référence.

Les experts sont très vite confrontés aux « joies » de la validation. La première question est de déterminer s'ils doivent retenir ou rejeter les candidats termes par rapport à leurs spécialités (aspect « électricité », aspect « électronique », etc.) ou bien par rapport à l'ingénierie DE de manière générale. Les termes de gestion tels que [6] *repère signalétique* peuvent être communs à plusieurs métiers. Par contre [7] *disjoncteur de départ de tranche* figurera dans la terminologie d'un mécanicien mais pas dans celle d'un électricien. Ce dernier ne sélectionnera pas le terme [8] *pression relative de calcul* propre au métier de mécanicien.

Un autre cas de figure se pose au cours de la sélection des termes : la présence de deux termes faisant référence au même objet mais relatifs à deux paliers distincts. Ainsi, [9] *salle de conduite de tranche* est propre à *Paluel* et existe dans le *N4* sous le nom de [10] *salle de contrôle-commande*. Faut-il garder une trace de ce changement de dénomination ou supprimer les anciens usages ?

Le besoin de retourner aux documents afin de vérifier le contexte dans lequel le terme est utilisé se fait ressentir de manière considérable. Ainsi, [11] *activité élevée*, un terme imprécis pris hors contexte, ne peut être retenu qu'à la condition qu'il désigne un seuil bien déterminé dans un domaine particulier. Il en est de même pour [12] *haute température*. Le seul moyen de faciliter la validation et de sélectionner la bonne occurrence du terme est de sauvegarder le lien existant entre texte et terme.

Les incertitudes des experts s'estompent après discussion et sélection des applications de la terminologie de référence : l'aide à la rédaction de documents techniques. Ils appréhendent de manière plus concrète la manière dont la terminologie permet un gain en productivité et en précision sur le plan de la rédaction sans compter la cohérence et ce après avoir assisté à la démonstration de deux outils d'aide à la production de documents techniques.

En effet, la terminologie doit être représentée graphiquement à l'aide de l'outil *WorldTrek* Picard *et al.* (1999), conçu à la Direction études et recherches d'EDF. Des propriétés telles que celles souhaitées par la DE à savoir la définition, la catégorie grammaticale, l'équivalent en langue étrangère et le contexte peuvent accompagner le graphique. La DE souhaite du reste que cette organisation s'inspire des métiers de la Direction de l'équipement. Ainsi, le travail de réflexion porte dorénavant sur les domaines et sous-domaines. Bien que les valideurs s'inspirent directement de leurs spécialités au cours du travail de sélection, il n'est pas garanti que ce découpage soit en adéquation avec l'organisation proposée dans les paragraphes suivants.

La terminologie doit également être intégrée dans l'application *TermChecker* (Boccon-Gibod 1999). Cet outil facilite la vérification de terminologie lors de la rédaction d'un document à l'aide d'un logiciel de traitement de texte classique. Il fonctionne à la manière d'un correcteur orthographique et propose des termes auxquels des statuts sont affectés. Ces derniers sont choisis en fonction des besoins des utilisateurs. À ce sujet, les remarques des experts sont très précieuses. Il ne s'agit pas uniquement de présenter les termes ayant changé de dénomination avec l'évolution des techniques mais de faire une mémoire de rédaction permettant de faire la transition entre les documents propres aux différents paliers des centrales nucléaires. Les experts expriment également la volonté d'attribuer des statuts conseillant l'usage d'un terme ou au contraire l'interdisant. Cette préoccupation part d'une volonté de normalisation et de stabilisation de la langue technique de la DE à la base du projet RMI.

Il ressort du suivi de la validation des termes potentiels par les experts la mise en évidence d'un autre aspect du métier de terminologue : celui de médiateur. Il détermine, avec les futurs utilisateurs qui s'avèrent être également les valideurs, les applications et le rôle de la terminologie. De plus, il ne se contente pas d'établir des règles afin de guider les experts dans leur lourde tâche. Il a pour rôle d'observer puis d'analyser le retour d'expérience de la validation et de formaliser les consignes afin de faciliter le travail de sélection et de le mettre en cohérence notamment lorsque les experts sont plusieurs. Il prend part de manière significative au travail de normalisation réalisé par le ou les valideurs.

3.3 Organisation des termes DSE en domaines et sous-domaines

La dernière étape de traitement, l'organisation des termes validés extraits des DSE en domaines et sous-domaines suscite beaucoup d'intérêt auprès des experts. La première tentative s'inspire de la structure existante du thesaurus EDF.

Le thesaurus EDF

Le thesaurus EDF, réalisé par le centre de documentation de la DER tout d'abord manuellement puis de manière automatique, a été créé exclusivement à des fins d'indexation. Il réunit des mots provenant de documents techniques de natures variées, caractérisés par un profil, auxquels est attribué le statut de descripteur. Ces mots-clés sont essentiels pour la recherche d'information, le classement et la diffusion de documents. Ces termes ont fait l'objet d'une collecte par des documentalistes puis d'une validation par des experts. Le thesaurus est mis à jour régulièrement par souci de coller au corpus dont les descripteurs sont extraits.

Le thesaurus est structuré en 45 points de vue dont 15 font l'objet d'un regroupement académique. Ils varient de la biologie à la thermique en passant par le droit, l'économie, les sciences de la terre, etc. Ils se divisent en domaines tels que l'électricité, le magnétisme et l'optique pour le point de vue sciences physiques ou assurance, famille et protection sociale pour le point de vue environnement social.

Inadaptation de la structure du thesaurus EDF et proposition d'une nouvelle organisation

Appliquer le découpage du thesaurus organisé en points de vue et en domaines s'avère très rapidement une erreur et ce sur plusieurs points. Une seconde répartition des termes est proposée aux experts : calquer les domaines et sous-domaines de la future terminologie de référence sur la structure physique des DSE notamment la division en rubriques (les documents) et sous-rubriques (les sous-documents) présentée dans la partie consacrée à la description du corpus de référence.

La complexité des DSE porte tant sur la taille que sur le contenu. La très grande masse de documents (près de 18 000 pages) motivent indirectement l'organisation. Un ensemble de 7 dossiers de systèmes élémentaires a fait l'objet d'une extraction terminologique. L'outil informatique utilisé, *Lexter*, a permis d'obtenir 24 500 termes potentiels. Respecter la structure des DSE peut permettre un traitement plus rapide à condition que cela soit en adéquation avec ce qui est attendu de la terminologie de référence.

Sur le plan du contenu, la hiérarchisation des DSE, avant tout en sous-ensembles, à savoir les systèmes élémentaires, puis en rubriques et sous-rubriques, a pour avantage de constituer la base de travail des ingénieurs. Elle reflète des pratiques, des habitudes, une certaine vision des métiers de l'ingénierie. Cette structuration est donc directement rattachée à la réalité, à l'expérience des futurs utilisateurs de la terminologie de référence. S'inspirer du thesaurus EDF ne permet pas d'adapter la structuration à cette nécessité. Cet ouvrage concerne tout d'abord EDF en tant qu'entreprise puis la production d'énergie sous tous ses aspects. Le point de vue «sciences humaines » semble répondre difficilement à des besoins au sein de la Direction de l'équipement, plus particulièrement en ce qui concerne les domaines tels que «les arts et lettres » et «la philosophie». Il reste très général du fait que le regroupement est majoritairement académique et ne peut convenir à un public spécialisé. Il est important de rappeler que l'objectif «aide à la rédaction » de la terminologie de référence est très différent de celui du thesaurus conçu pour la recherche documentaire (Monteil 1995). En effet, celui-ci a pour applications la recherche documentaire et l'indexation. Ce découpage du thesaurus n'est pas adapté à la nature très technique et précise des métiers de la DE. Ces caractéristiques sont d'autant plus importantes lors de la création de documents techniques.

Le regroupement par rubriques et sous-rubriques est d'autant plus facile en raison de la sauvegarde au cours de l'extraction et de la validation du lien entre le terme et la séquence de texte dans laquelle il apparaît. Ainsi, les balises indiquant l'origine textuelle du terme reflètent les domaines et sous-domaines de la terminologie. La principale préoccupation est de connaître le taux de recouvrement des termes par rapport à la structure des DSE afin de confirmer la concordance entre les métiers de la DE et la structure des DSE. Le cas échéant, devrait-on donner la priorité à la disposition des documents en rubriques et sous-rubriques ou bien se fier davantage à la classification métiers des experts de la DE ?

Il existe un autre point de divergence à propos du contenu : la nature des liens unissant les termes d'un même domaine et les domaines d'un même point de vue au sein du thesaurus. Se reporter au thesaurus EDF favorise la constitution de groupes de termes de type «générique-spécifique » majoritaires dans ce document. Ainsi, [13] les *échangeurs de chaleur* sont rassemblés par type à savoir [14] *les échangeurs à contact*, [15] les *échangeurs par mélange* et [16] les *échangeurs par surface* et non en fonction de leur rôle au sein d'un système donné

notamment le *transfert de chaleur* du [17] *cœur du réacteur* vers [18] le *circuit secondaire*. Par contre, le fait que les DSE reflètent davantage un métier implique une diversité de liens au sein d'un même domaine. Dans un DSE, la rubrique fonctionnement du système élémentaire réunit des termes décrivant le déroulement d'un processus. Le terme [19] *groupe turbo-alternateur* appartient à la même sous-rubrique «cycle de la vapeur » que les termes suivants: [20] *turbine*, [21] *sécheur-surchauffeur*, [22] *condenseur*, [23] *pompe d'extraction* et [24] *pompe alimentaire*. Ils sont unis par des liens fonctionnels. Par contre, la sous-rubrique description du matériel regroupe des termes illustrant la composition d'un système élémentaire. Dans cet exemple précis, les liens sont plutôt de type «tout-partie » et le même terme groupe *turbo-alternateur* figure avec [25] *chaudière nucléaire* et [26] *condenseur* dans le domaine «centrale nucléaire».

Il ressort de cette analyse une nette différence entre le contenu du thesaurus EDF et celui du corpus DSE. L'élément déterminant est la personnalisation du travail linguistique qu'il s'agisse d'un thesaurus ou d'une terminologie de référence. La prise en compte de la réalité à savoir le contexte d'utilisation, le contenu des données à traiter et le public utilisateur est essentielle.

4 La terminologie de référence EDF face à la doctrine

Dans le cas particulier de la constitution d'une terminologie de référence destinée à la Direction de l'équipement, certains principes de la théorie générale de la terminologie ne sont pas adaptés à la pratique terminographique.

La présentation aux experts de la DE d'une première terminologie réunissant exclusivement des termes DSE a changé le projet d'orientation. En effet, un nouveau corpus a été soumis afin d'illustrer davantage l'aspect méthodologique des métiers d'ingénierie. Le corpus est étroitement lié à cette volonté de construire un terminologie multi-facettes. Déterminer les caractéristiques des documents composant le corpus permet de mieux cerner la future composition de la terminologie sur le plan des termes et de l'organisation en domaines et sous-domaines.

L'impossibilité de revenir au texte source pose problème. Au cours de leur travail de validation, les experts souhaitent à maintes reprises se replonger dans le contexte d'utilisation d'un futur terme. Cela souligne l'importance de sauvegarder un lien entre la terminologie et le corpus, du moins le terme et la séquence de texte dont il a été extrait. Bien que la théorie générale de la terminologie préconise le détachement des termes des textes, ce besoin se fait ressentir très concrètement au cours de la validation.

L'attribution de statuts aux termes extraits des DSE tels que «accepté», «interdit » ou «à éviter » afin de guider le rédacteur pose le problème de la polysémie et de la synonymie en terminologie (Felber 1984). Selon la doctrine, les changements de sens donnent naissance à des termes très ambigus. Il est donc préférable de créer des termes nouveaux pour représenter des notions nouvelles. Or, en ce qui concerne la synonymie, elle sème la confusion et donne l'impression qu'il existe plusieurs notions. Cependant, il est difficile de faire abstraction de ces phénomènes. Ces manifestations linguistiques reflètent les conditions de création et d'utilisation d'un terme. Yves Gambier (1991: 11) évoque les phénomènes de polysémie et de synonymie en ces termes: «Les transferts (métaphoriques) de termes répondent aux besoins pressants de dénomination, au croisement des disciplines, au rapport motivé et non pure convention entre le signe et la notion». Il est important de donner au terminographe la possibilité d'avoir recours à toutes les manifestations du langage, à partir desquelles il fait des choix selon leur pertinence par rapport au travail terminologique à réaliser. Il reste que réduire la polysémie lors de la description d'un domaine est nécessaire afin de garantir la stabilité des significations. En même temps, le terme prend sens dans son contexte d'utilisation. Tenir compte des nuances de synonymie est donc important. Cela permet de faire état de l'évolution d'un terme au sein d'une même discipline. Ainsi, dans le cas particulier des centrales, cela permettra de garder une trace de l'évolution de technologie lors du passage d'un palier à un autre.

La prise en considération de l'interdisciplinarité au cours du travail terminologique se pose concrètement dans le cas EDF. En terminologie classique, la création de classes universelles ne permet pas de refléter ce contact entre les différentes disciplines. Or, le terme résulte de l'innovation. Selon François Gaudin (1993: 83), «l'innovation » qui constitue l'essentiel de la production terminologique «naît de réseaux transversaux et la circulation langagière, l'échange et la contamination de concepts entre les disciplines sont des moteurs puissants d'innovation». Cette innovation résulte des contacts entre différentes communautés scientifiques. «Ce qui motive dans ce cas les interactions, et les recatégorisations, qu'elles induisent, c'est l'existence d'un but commun, d'un objectif poursuivi par un travail». Ce phénomène de circulation des connaissances se

manifeste dans les rubriques et sous-rubriques composant les DSE. De ce fait, le terme découle directement de la pratique d'une ou plusieurs communautés. Il ne s'agit pas de partage d'universaux comme le laisse présumer la théorie générale de la terminologie à travers la constitution de grandes classes canoniques dans le but de décrire une discipline mais d'une pratique entre des acteurs prenant part à la même expérience professionnelle. Ce cheminement de l'innovation au terme permet de constater que ce dernier provient directement de l'interdisciplinarité entre métiers.

Enfin, chaque discipline peut être abordée sous des angles différents en fonction des objectifs d'une application. Comme le souligne François Rastier (1994: 62), «Les modes de structuration varient selon les domaines». Les domaines et sous-domaines d'une terminologie sont délimités en fonction du métier, de l'application et de l'utilisateur. Au lieu d'en réduire le nombre afin d'obtenir des classes de termes très générales, il faut au contraire garder à l'esprit les spécificités de chaque pratique afin de les décrire au mieux sur le plan notionnel. Toutefois, il est important d'opérer une sélection pertinente. En effet, Rastier (1994: 77) précise que «les différents axes sémantiques seront choisis selon les conditions de la description: pour opposer métro et autobus, on peut choisir la catégorie ferré *vs* routier dans un texte technique, mais aussi lent *vs* rapide si l'on décrit les raisons du choix des usagers... Bien entendu, ces divers axes ne s'excluent pas, mais une description pertinente doit rejeter les catégories inutiles». Il est nécessaire d'avoir une vue d'ensemble de la discipline dont on dresse la terminologie afin de ne pas éliminer les éléments pertinents.

5 Conclusion

Le travail sur corpus, les différents contacts avec les experts de la DE ainsi que les tentatives d'organiser les termes potentiels extraits puis validés en domaines et sous-domaines ont eu pour effet de mettre l'accent sur certains éléments importants.

L'impact du corpus sur la nature de la terminologie à produire ne s'est jamais autant confirmé. L'analyse de sa structure et de son contenu par rapport aux résultats à produire permet de reproduire cet aspect multi-domaines présent en entreprise. De plus, dans le cas particulier d'EDF, l'étude du corpus a une lourde influence sur les modèles d'organisation en domaines et sous-domaines proposés pour la terminologie de référence.

La contextualisation des termes constituant la terminologie de référence est nécessaire tout d'abord aux valideurs qui ont besoin de sélectionner un terme en fonction de son sens dans la phrase mais également aux utilisateurs. La création de liens termes-textes dans une perspective d'aide à la rédaction a pour avantage de souligner les nuances entre termes proches sur le plan sémantique.

L'inadaptation de la structure du thesaurus EDF afin d'organiser les termes en domaines et sous-domaines met l'accent sur l'importance de la pertinence en terminologie. Il faut avoir à l'esprit qu'une terminologie efficace est «éphémère » du fait qu'elle s'attache à un besoin particulier, à un moment particulier, à un public particulier.

Enfin, le travail effectué sur le projet d'une terminologie de référence renforce la vision du terminologue en tant que médiateur. L'assistance, l'aide à la décision ainsi que l'orientation des experts au cours de la validation met en relief la nécessité de coopérer avec le terminologue afin que le travail terminologique réalisé soit le plus pertinent possible.

Remerciements

Nos remerciements à Henry Boccon-Gibod, Didier Bourigault et Monique Slodzian pour leurs conseils éclairés et leurs précieuses relectures.

Yasmina Abbas,
Centre de recherches en ingénierie multilingue,
Institut national des langues et civilisations orientales,
Paris,
France.

Marie-Luce Picard,
Groupe SOAD (Statistiques, optimisation et aide à la décision) du Service IMA (Informatique et mathématiques appliquées) de la DER (Direction des études et recherches d'EDF),
Clamart,
France.

Bibliographie

Boccon-Gibod (H.), 1999: «Enjeux de la maîtrise de la terminologie pour la production et la consultation de documents – Application à la documentation technique des installations de production d'électricité», dans *DocuWorld 1999*, Paris.

Bourigaut (D.), 1994: *LEXTER, Un Logiciel d'Extraction de TERminologies-Application à l'acquisition des connaissances à partir de textes*, Thèse de l'École des hautes études en sciences sociales, Paris.

EDF-DER, SID (Département systèmes d'information et de documentation), Mise à jour 1997, *Thesaurus EDF*, français-anglais.

Felber (H.), 1984: *Manuel de terminologie*, Infoterm (Centre international d'information pour la terminologie), Unesco, Paris.

Gambier (Y.), 1991 : « Terminologie et sociolinguistique », dans *Cahiers de linguistique sociale*, n° 18, p. 43-44.

Gaudin (F.), 1993 : « Socioterminologie, Des problèmes sémantiques aux pratiques institutionnelles », dans *Publications de l'Université de Rouen*, n°182, p. 83.

Monteil (M.-G.), 1995 : « Indexation automatique et manuelle, comparaison et perspectives », *IDT'95*, Paris.

Picard (M.) et Boudailler (E.), 1999 : « Worldtrek for authoring and comprehension », *IUI'99*, Redondo Beach, Californie, USA.

Rastier (F.), Cavazza (M.) et Abeille (A.), 1994 : *Sémantique pour l'analyse*, Masson, Paris.

Slodzian (M.), 1995 : « Comment revisiter la doctrine terminologique aujourd'hui ? », dans *La banque des mots*, n° spécial 7, p. 11-18.

Publications

Discours professionnels en français

Il peut paraître pervers de rassembler en Scandinavie et de publier en Allemagne un recueil d'études sur les langues de spécialité du français. Certes, le rédacteur de l'ouvrage et initiateur du projet, Yves Gambier, est professeur de français dans un institut de traduction en Finlande et l'éditeur, Peter Lang, est plus ouvert aux études de linguistique appliquée que la plupart des maisons d'édition françaises, mais la vraie raison de cette anomalie géographique se trouve ailleurs. D'une part, le regard que porte l'étranger sur la situation en France est garant d'objectivité et, d'autre part, ce sont les étrangers qui sont en passe de prendre le relais des Français, qui délaissent, à quelques brillantes exceptions près, ce domaine de recherche.

Cette crise des études des langues de spécialité fait l'objet d'une analyse perspicace signée du rédacteur lui-même, ici l'auteur du plus long article du recueil. Il commence par esquisser la situation des LSP à l'échelle internationale, où l'anglais domine, comme le sigle en usage même dans les écrits francophones le laisse entendre. Historiquement, c'est la nécessité d'enseigner rapidement la langue des études universitaires (donc l'anglais, le français, naguère le russe) aux étudiants allophones qui a motivé les premières études des langues de spécialité. C'est aussi l'aspect des langues de spécialité qui accuse le plus grand retard en France aujourd'hui. Traditionnellement, c'était le français, langue étrangère (Fle), qui suscitait le plus grand nombre d'études sur les langues de spécialité du français, tradition qui s'essouffle, sauf dans le nord de l'Europe. Gambier examine ensuite la langue de spécialité par rapport à des domaines connexes, qui représentent en général une application apparentée: la terminologie, la rédaction technique, la traduction. La méthode fait appel à des critères souvent flous: le domaine, la situation de communication, l'interaction, la vulgarisation, la sociolinguistique (langue de spécialité comme instrument de légitimation professionnelle), ces dernières étant pourtant susceptibles d'évolution, donc porteuses de renouveau. Ce tour d'horizon est une raison suffisante pour justifier l'utilité de ce recueil. Mais les autres articles, la démonstration pour ainsi dire, illustrent l'enrichissement possible des langues de spécialité, lorsque l'analyse se fonde sur une approche linguistique plus complète.

Savoir ce que constituent exactement les langues de spécialité tracasse les linguistes depuis longtemps. C'est «la question des questions», comme le dit Lothar Hoffmann, cité en exergue par Finn Frandsen, qui, pour sa part, tente une définition qui se libère du carcan de l'opposition simpliste entre langue générale et langue de spécialité. Pour ce chercheur, il vaut mieux examiner les entités selon des perspectives différentes et complémentaires: celle de l'épistémologie d'une part et de l'ontologie de l'autre. L'une définit les formes d'usage d'une langue, l'autre les données elles-mêmes. La nouvelle définition des langues de spécialité qu'il propose est donc asymétrique et si elle résout en le déplaçant le problème de la langue générale, elle comporte l'inconvénient de devoir constituer un nouveau métalangage, que l'auteur n'essaie pas d'improviser.

Après tant de critiques dirigées contre l'approche qui privilégie le niveau lexical, il n'est pas étonnant de constater que la majorité des articles accusent une orientation résolument textuelle. Privilégier le textuel, et parfois l'intertextuel, n'exclut pas une multiplicité de regards, qui se révèlent tout à fait complémentaires.

Une des lectures possibles des textes de spécialité est la sémantique interprétative de François Rastier. W. Johanssen analyse une brochure présentant une entreprise à la lumière des isotopies sémantiques dégagées dans le texte, effet de la récurrence syntagmatique d'un même sème. Cette analyse fait ressortir non seulement la cohésion du texte, dont le point de départ est différent de

celui de Hassan et Halliday, par exemple, mais aussi un aperçu de la culture d'entreprise.

André Avias pour sa part fait appel à une approche prototypique inspirée de J. M. Adam pour analyser un article du *Figaro* sur le budget. Il détermine d'abord, non sans difficulté, le découpage structurel de son article en séquences textuelles, puis il choisit des prototypes. Bien entendu, cette démarche ne se limite pas aux textes spécialisés, mais elle apporte une contribution au traçage de l'argumentation dans le discours.

Le caractère argumentatif, voire polémique d'un autre genre de texte, le «Mot du P.D.G.», sorte de préface du rapport d'activité annuel d'une entreprise, qui fournit le thème de l'article de K. Flottum. L'analyse de ce genre textuel est d'ailleurs une spécialité scandinave, et l'auteur a l'avantage de pouvoir puiser dans un fonds de recherche établi. Elle fonde son analyse sur les études qu'a réalisées Ducrot sur la négation syntaxique, sur la qualité de polyphonie, surtout sur la lecture qu'en fait H. Nölke. Une analyse des différents types de négation fait ressortir l'usage dominant de l'un d'entre eux, la négation polémique, déni des «idées fausses qu'on peut avoir sur l 'entreprise». Non seulement cette analyse – surtout syntaxique – rappelons-le, permet de résoudre des questions posées dans les études antérieures et de déterminer qu'il s'agit d'un type de texte argumentatif et contre-argumentatif. L'auteur apporte la preuve par ailleurs que ce discours n'est pas impersonnel, comme une lecture superficielle pourrait le laisser entendre, car les destinataires sont bien présents, mais implicites.

Ce sont les sciences cognitives qui viennent enrichir l'analyse d'un texte juridique français par L. Lundquist, dans le cas présent un jugement de la Cour de cassation. Les outils que l'auteur emploie sont empruntés à l'intelligence artificielle dans le cas du cadre (*frame*), et aux études psychologiques, dans celui de l'espace mental. Celui-là sert à situer l'analyse de l'ensemble, celui-ci structure les relations inférieures. L'auteur expose ensuite le cadre du jugement, permettant au néophyte non francophone de situer les relations structurelles, tâche malaisée surtout en comparaison de la transparence structurelle du discours juridique danois déjà familier. L'idée d'espace mental permet à son tour de bien repérer les éléments et les relations au fil du discours. Cet article est un bon exemple de perspicacité obtenue par un regard extérieur porté sur le discours juridique français, regard qui est loin d'être naïf.
A. Askelund s'inspire de la grammaire des cas pour son analyse des textes juridiques, mais c'est la traduction qui l'intéresse. Elle explique comment l'emprise du texte de départ est différente selon qu'il s'agit de la langue maternelle ou de la langue étrangère, et on apprécie son cadre d'analyse original et explicite de la «métamorphose» du texte traduit.

Les deux derniers articles du recueil sont à orientation lexicale. Celui de G. Dyrberg et J. Tournay renoue avec la très riche tradition danoise des dictionnaires de spécialité dont le contenu et la présentation sont constamment modifiés en fonction des besoins perçus des usagers. Le dictionnaire envisagé dans cet article est juridique d'après son contenu et asymétrique selon sa présentation. Destiné aux usagers danois, il envisage un savant mélange d'exemples et d'explications, tantôt en danois, tantôt en français. Ce sont les informations encyclopédiques qu'il convient d'inclure qui constituent le sujet de l'article, mais on passe en revue la formation de la définition, les stratégies d'exemplification de l'équivalence partielle, particulièrement bien illustrées d'ailleurs, la typologie des exemples, et la place dans le dictionnaire de ce qu'elles appellent, d'après A. Kjaer, les «formules de routine». Ces dernières auraient mérité un article de plus – en quoi consistent-elles en fait? Faut-il les présenter sous un élément de la suite, ou de façon conceptuelle, quel traitement informatique proposer pour résoudre les problèmes sans issue sur le papier.

Le dernier article du recueil, signé de P. Lederlin, traite des constructions de type N de N dans les textes économiques. L'auteur cherche des règles de production simples et maniables à proposer à ses étudiants norvégiens. Pour son cadre théorique, d'une part, il se limite aux études relativement anciennes (M. Wilmet 1986); d'autre part, il sous-estime l'explication de la lexie complexe, pourtant évoquée. Selon cette approche, l'étudiant apprend *appel d'offres*, *assiette de l'impôt* comme des unités lexicales (ou terminologiques) et non comme des locutions.

Le recueil comporte une bibliographie de 71 études portant sur les langues de spécialité françaises publiées récemment en Scandinavie.

Après avoir pris connaissance de plusieurs nouvelles approches présentées ici, on ne peut partager le pessimisme de Gambier. Certes, la plupart des auteurs peinent à définir les langues de spécialité. Toutefois, en les décrivant, ils contribuent à fournir une définition par extension, plutôt que par compréhension, comme diraient les lexicographes, mais qui est en même temps plus intuitive, plus pratique, et plus consensuelle.

Une lecture de John Humbley,
Centre de terminologie et de néologie,
Laboratoire de linguistique informatique,
Université Paris 13.

Gambier (Yves), éd.,1998: *Discours professionnels en français*, Francfort, Peter Lang GmbH, 233 p.

Terms in context

Dans son récent ouvrage, Jennifer Pearson entend avant tout éveiller l'intérêt des communautés scientifiques spécialisées dans la terminologie et dans la linguistique de corpus, ainsi que des enseignants de langues de spécialité sur le potentiel des corpus de textes spécialisés exploitables en matière de terminographie.

Si les outils actuels d'extraction terminologique sont principalement axés sur la recherche de termes dans les corpus de textes, l'auteur montre ici qu'il est tout à fait possible d'accéder également aux données définitoires accompagnant les termes en déployant une stratégie de recherche semi-automatique fondée principalement sur les structures grammaticales environnant ces derniers.

La méthodologie suivie par l'auteur s'articule en plusieurs étapes : après avoir sélectionné des corpus de textes adéquats sur la base du critère des cadres de communication, l'auteur procède à une identification manuelle conventionnelle des structures lexico-syntaxiques des termes pour chacun des corpus en vue de créer une liste finie de structures de termes qui servira à extraire automatiquement les candidats-termes du corpus. Elle établit ensuite des restrictions supplémentaires visant, d'une part, à écarter les mots qui ne sont pas des termes (par le recours au critère du statut générique, que l'on peut informatiser en spécifiant qu'un terme ne peut être précédé d'aucun article à l'exception de l'article indéfini *a*) et, d'autre part, à filtrer les termes accompagnés de contextes définitoires (un jeu de signes linguistiques a été conçu à cet effet, comprenant entre autres des expressions telles que *called*, *known as*, *e.g.*). Finalement, dans le but de repérer les structures définitoires, elle emploie une batterie de marqueurs linguistiques et métalinguistiques particuliers à divers types de définitions, classées entre autres selon leur degré de précision et selon qu'elles se répartissent sur une ou plusieurs phrases, qu'elles incluent un terme générique ou non.

Préalablement à ce minutieux travail d'extraction de termes et de données définitoires, l'auteur veille à délimiter son terrain d'investigation et à bien en définir les éléments-clés. Ainsi a-t-elle exploré les théories appliquées au terme, aux langues de spécialité et aux sous-langages, à la compilation de corpus de textes et à leur typologie, à la classification des textes, à la formulation et à la catégorisation des définitions spécifiques aux dictionnaires et aux textes. Les théories répondant le mieux aux besoins de la recherche ont été confrontées aux données disponibles dans les corpus de textes et ont donc pu être soit infirmées, soit corroborées et approfondies, ce qui ouvre la voie à de nouvelles pistes de recherche.

Une lecture de Sylviane Descotte,
Centre de recherche Termisti,
Institut supérieur de traducteurs et interprètes,
Bruxelles.

Pearson (J.), 1998 : *Terms in Context*, Amsterdam, John Benjamins Publishing Company.

Je soussigné souhaite recevoir gratuitement la revue *Terminologies nouvelles.*

Nom : _____

Entreprise, organisme : _____

Fonction : _____

Adresse : _____

_____

Ce bulletin d'abonnement est à adresser au module dont vous relevez (adresse au dos de la revue)

Descriptif bibliographique :
Enguehard (Chantal)
et Condamines (Anne), éd. :
Terminologie et intelligence artificielle, actes des 3es rencontres « Terminologie et intelligence artificielle » (Nantes, 10 et 11 mai 1999),
dans *Terminologies nouvelles*,
n° 19, décembre 1998 - juin 1999,
Bruxelles, Agence de la francophonie
et Communauté française de Belgique,
ISSN : 1015-5716.

ISSN : 1015-5716

Édit. resp. :
M. Garsou, 44 Boulevard Léopold II,
1080 Bruxelles, Belgique.

Numéros déjà parus

Consultables à partir du n° 14 à l'adresse www.rint.org

1, mai 1989 :
Le Rint : objectifs et perspectives

2, décembre 1989 :
La formation en terminologie

3, juin 1990 :
Harmonisation des méthodes en terminologie (actes des séminaires de Talence et de Hull)

4, décembre 1990 :
Numéro général

5, juin 1991 :
Terminologie et informatique

6, décembre 1991 :
Terminologie et développement I (actes du séminaire de Rabat)

7, juin 1992 :
Numéro général

8, décembre 1992 :
Terminologie et environnement

9, juin 1993 :
Terminologie et développement II (actes du séminaire de Cotonou)

10, décembre 1993 :
Phraséologie (actes du séminaire de Hull)

11, juin 1994 :
Numéro général

12, décembre 1994 :
Implantation des termes officiels (actes du séminaire de Rouen)

13, juin 1995 :
Terminologie et entreprise

14, décembre 1995 :
Numéro général

15, décembre 1996 :
Banques de terminologie (actes de la table ronde de Québec)

16, juin 1997 :
Enquêtes terminologiques

17, décembre 1997 :
Terminologie et formation

18, juin 1998 :
Terminotique et documentation

19, décembre 1998 - juin 1999 :
Terminologie et intelligence artificielle (actes du colloque de Nantes)

À paraître

20, décembre 1999 :
De nouveaux outils pour la néologie

Adresses des organismes membres du Rint

Afrique centrale et de l'Est

Coordination: Centre de linguistique théorique et appliquée
BP 4956
Kinshasa/Gombe
Zaïre.

Afrique de l'Ouest

Coordination: Centre de linguistique appliquée
Université Cheikh Anta Diop
Dakar — Fann
Sénégal.

Canada

Terminologie et Normalisation
Bureau de la traduction
Travaux publics et
Services gouvernementaux
Portage II, 3e étage
165, rue Hôtel-de-Ville
Hull (Québec)
K1A 0S5
tél.: 1 (819) 994-5934

Communauté française de Belgique

Ministère de la
Communauté française
Service de la langue française
44, Bd Léopold II
B-1080 Bruxelles
tél.: 32 (2) 413 32 74

France

Délégation générale à la langue française
1, rue de la Manutention
F-75116 Paris
tél.: 33 (1) 40 69 12 00

Haïti

Faculté de linguistique
Université d'État d'Haïti
38, Rue Dufort
(Quartier Bois-Verna)
Port-au-Prince
tél.: (509) 45 12 33

Madagascar

Centre des langues de l'Académie malgache
BP 6217
Antananarivo 101.

Maroc

Institut d'études et de recherches pour l'arabisation
BP 6216
Rabat — Instituts
tél.: 212 (7) 77 30 12

Québec

Office de la langue française
200, chemin Sainte-Foy,
Québec (Québec)
G1R 5S4
tél.: 1 (418) 643-4144

République centrafricaine

Conseil national d'aménagement linguistique
BP 888
Bangui.

Suisse

Chancellerie fédérale suisse
Services linguistiques centraux
Section de terminologie
Gurtengasse 2-4, 4e étage
CH 3003 Berne
tél.: 41 (31) 324 11 49

Tunisie

Innorpi
10bis, rue Ibn el Jazzar
1012 Tunis — Belvédère
tél.: 216 (1) 785 922

Modules associés

Union latine
Bureau de Paris
131, rue du Bac
F-75007 Paris
tél.: 33 (1) 45 49 60 60

Coédité par:
L'Agence de la francophonie
et la Communauté française de Belgique
(Service de la langue française
du ministère de la Communauté française
et Commissariat général
aux relations internationales)

Secrétariat du Rint:
Office de la langue française
200, chemin Sainte-Foy,
Québec (Québec)
G1R 5S4 Canada

Le Rint sur Internet:
http://www.rint.org