HAL open science

Méthodes ascendantes pour l'ingénierie des connaissances

Nathalie Aussenac-Gilles

To cite this version:

Nathalie Aussenac-Gilles. Méthodes ascendantes pour l'ingénierie des connaissances. Informatique. Université Paul Sabatier - Toulouse III, 2005. tel-00089165

HAL Id: tel-00089165

https://theses.hal.science/tel-00089165

Submitted on 10 Aug 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITE TOULOUSE III - PAUL SABATIER

U.F.R. Mathématiques, Informatique et Gestion.

Synthèse des travaux

En vue de l'obtention de

l'Habilitation à Diriger des Recherches
de l'Université Paul Sabatier – Toulouse III
(spécialité informatique)

METHODES ASCENDANTES

POUR L'INGENIERIE DES CONNAISSANCES

Présentés par

Nathalie AUSSENAC-GILLES

Le 1e décembre 2005, devant le jury composé de

Catherine Garbay,	directeur de recherches au CNRS, IMAG (Grenoble)	rapporteur
Joost Breuker,	professeur à l'Université d'Amsterdam, SWL	rapporteur
Gilles Kassel,	professeur à l'Université d'Amiens, LaRIA (Amiens)	rapporteur
Claude Chrisment,	professeur à l'Université Paul Sabatier, IRIT (Toulouse)	président
Jean-Luc Soubie,	ingénieur de recherches (HDR) INRIA, IRIT (Toulouse),	directeur de recherche
Pierre Tchounikine,	professeur de l'Université du Mans, LIUM	examinateur

INSTITUT DE RECHERCHE EN INFORMATIQUE DE TOULOUSE
Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cedex – http://www.irit.fr/

RESUME

Les travaux présentés dans ce mémoire relèvent de l'ingénierie des connaissances et contribuent à mieux en cerner et définir le champ scientifique. Ce domaine vise la définition de solutions techniques, méthodologiques et organisationnelles pour identifier des connaissances, les modéliser et les restituer au sein d'applications informatiques. La modélisation est ici considérée comme une étape indépendante de l'opérationnalisation faite dans le système final.

L'approche retenue ici met l'accent sur les connaissances liées aux usages et pratiques des personnes concernées par le système à concevoir. L'objectif est de réaliser des aides à la modélisation, et précisément aux processus d'abstraction et de caractérisation des connaissances.

Les recherches développées s'appuient systématiquement sur une démarche expérimentale, grâce à des applications en entreprises et des études de cas, et sur des collaborations interdisciplinaires. Les contributions présentées comprennent des techniques et des logiciels de recueil et d'analyse de connaissances, des méthodes et des représentations des connaissances pour la modélisation conceptuelle ainsi que des plates-formes intégrant ces différents supports.

Ces méthodes et outils répondent successivement à trois problématiques différentes sur la modélisation, en phase avec les évolutions historiques du domaine. Le premier problème traité, assez large, est celui de la modélisation conceptuelle à partir de connaissances d'experts et d'activités humaines, faisant appel à des techniques et problématiques de psychologie et d'ergonomie. Une deuxième manière d'étudier la modélisation a consisté à s'intéresser à l'analyse des textes et aux approches linguistiques pour construire des modèles de domaines spécialisés, comme les bases de connaissances terminologiques et les ontologies. De l'ensemble des expériences menées pour ces évaluer ces propositions, il ressort que l'utilisation des modèles au sein des applications doit être prise en compte dès leur construction. Une troisième problématique porte donc sur l'étude de l'utilisation de modèles conceptuels, et plus particulièrement d'ontologies, dans des cadres applicatifs ciblés, pour ajuster les méthodes et logiciels requis pour leur construction. Les applications étudiées relèvent de la recherche d'information et de l'accès au contenu de documents. Cette problématique soulève des questions fondamentales sur la complémentarité des modèles et des documents en tant que vecteurs de connaissances.

Les perspectives de ce travail se situent selon deux axes liés aux documents. D'une part, si les modèles facilitent l'accès au contenu, comment définir les modalités de l'indexation sémantique de documents (à l'aide d'ontologies) ? D'autre part, les documents et les besoins étant sans cesse renouvelés, comment intégrer la question de la maintenance d'ontologies et de terminologies dans le processus de construction ? L'originalité de l'approche retenue est de traiter ces deux questions conjointement, et d'en chercher des solutions cohérentes s'appuyant sur les outils et méthodes de construction d'ontologies à partir de textes.

REMERCIEMENTS

Il m'est difficile de faire tenir en quelques lignes tous les remerciements que j'aimerais adresser à ceux et celles qui ont permis à mon travail d'exister et de progresser. C'est grâce à eux tous que ce mémoire a pu voir le jour.

Au-delà de la formalité d'usage, c'est avec un réel plaisir que je commencerai par remercier les membres de mon jury pour leurs encouragements, leurs conseils, le temps qu'ils ont consacré à l'évaluation de mon travail, à travers ce mémoire et ma soutenance.

- Joost Breuker était déjà rapporteur de ma thèse et c'est pour moi une grande chance qu'il ait accepté de prendre la mesure du travail accompli depuis. Je le remercie d'autant plus vivement de son évaluation que mon mémoire est en français.
- Je suis fière et heureuse que Catherine Garbay ait évalué ce mémoire pour en faire une lecture critique, approfondie et constructive. J'ai grandement apprécié tant la finesse et la justesse de son analyse que la pertinence et l'humanité de ses avis lors de nos échanges.
- La commission des thèses de l'UPS a eu la bonne idée de nommer Gilles Kassel comme rapporteur de mon travail. C'est donc à la fois un collègue et ami de longue date qui a fait une lecture minutieuse et encourageante de mon travail. Pour cela mais aussi pour toutes les voies qu'il a ouvertes dans notre domaine en montant une équipe et en défendant la problématique de l'ingénierie des connaissances auprès de nos tutelles, je lui adresse un grand merci.
- Je suis très reconnaissante à Pierre Tchounikine des discussions scientifiques animées qui m'ont aidée à mettre en évidence mes problématiques de recherche et à organiser leur présentation. Sans ses conseils patients et ses nombreux encouragements, je me serais encore plus souvent détournée du chemin de la rédaction.
- Mes remerciements vont aussi vers Claude Chrisment à plusieurs titres. Tout d'abord, il a été une des personnes qui m'ont encouragée régulièrement avant que je n'ose écrire cette habilitation. Ensuite, il a été un des vecteurs de l'orientation de mon travail vers l'utilisation des ontologies en recherche d'information, via des collaborations avec son équipe.
- Ce mémoire doit énormément à Jean-Luc Soubie, mon directeur de recherche pour ma thèse puis pour cette habilitation. Depuis 1986, sa confiance, ses fortes convictions scientifiques et son sens critique m'ont permis de progresser et de m'affirmer en tant que chercheur, même si ma problématique s'est progressivement éloignée de celle de la construction des SBCC.

Ce jury ne respecte pas les règles de parité politiquement correctes, et surtout il ne reflète pas la forte présence de collègues féminines en ingénierie des connaissances. Grâce à de nombreuses collaborations, j'ai rencontré quelques femmes hors du commun. Je leur exprime tout particulièrement ma gratitude car au-delà d'échanges scientifiques passionnants, nous avons établi des relations amicales, humaines, respectueuses et nourries d'encouragements réciproques.

- En suivant la chronologie de nos rencontres, je commencerai par Anne Condamines. Nous marchons côte à côte, ou presque, depuis 20 ans, et malgré nos personnalités si différentes, c'est toujours avec une grande confiance et beaucoup de complicité que nous continuons de nous lancer dans des projets communs. Elle m'a fait aimer la complexité de l'étude des langues et la rigueur nécessaire à leur manipulation informatique. Je la remercie pour sa patience face à mes retards si fréquents …

- Mon amitié pour Myriam Bras date également de notre thèse. En informaticienne enthousiasmée par l'étude du langage, elle a stimulé ma curiosité et donné l'envie de m'intéresser aux connaissances dans les textes. Je l'en remercie. Surtout, de randonnées en stages d'escalade, j'ai pu apprécier sa grande générosité, son énergie et sa finesse dans les relations humaines.
- J'ai connu Marie-Pierre Gleize à la même époque. Elle me montre le chemin à suivre avec quelques années d'avance, et depuis 2 ans, elle m'a soutenue avec une grande générosité et l'énergie qui la caractérise pour que je rédige cette habilitation. Un grand merci pour les matinées studieuses et les petits repas au calme de la rue du Bac.
- Enfin, dès ma thèse, j'ai aussi travaillé aux côtés de Corinne Chabaud. Grâce à elle, et aux étudiants qu'elle a impliqués dans nos projets, je suis entrée par la porte de l'activité dans le monde de l'ergonomie et de la psychologie cognitive. Au cours de ces années, elle m'a souvent soutenue avec complicité et je l'en remercie.
- Une de mes premières collaborations en dehors de l'IRIT m'a entraînée à Nice où j'ai fait la connaissance de Rose Dieng. Sa bonne humeur, sa grande énergie et sa maîtrise scientifique du domaine en font une collègue précieuse. Depuis 1990, nous nous retrouvons régulièrement sur les bancs des conférences et des groupes de travail, toujours avec la même amitié.
- J'ai une pensée très particulière pour Chantal Reynaud, rencontrée à la naissance du groupe de travail GRACQ, dans l'animation duquel nous nous sommes beaucoup investies ensemble. En mettant en commun MACAO et ASTREE, j'ai eu l'occasion d'apprécier son écoute, son exigence scientifique et ses qualités pédagogiques. Je lui témoigne ici toute mon amitié.
- MACAO II est le fruit du travail rigoureux et méthodique de Nada Matta. Les 5 années que Nada a passé dans notre équipe ont permis des avancées importantes, des expérimentations indispensables et une reconnaissance de nos travaux. Mes remerciements envers Nada vont bien au-delà de sa contribution scientifique, récompensée par son HDR il y a un an ! Grâce à elle, à ses sourires, ses services et sa générosité, nous avons compris ce qu'aimer la vie avec force voulait dire. Merci Nada pour ce témoignage précieux.
- Depuis 1997, mes collaborations avec Brigitte Biébow et Sylvie Szulman m'ont permis de découvrir le travail en équipe « éclaté » entre Toulouse et Villetaneuse. Outre une bonne connaissance de la gare du Nord, cette collaboration m'a apporté une grande richesse humaine et intellectuelle. Merci à elles deux pour leurs accueils chaleureux et pour nos discussions animées et stimulantes. En écrivant ce rapport, j'ai été étonnée de tout ce que nous avons produit ensemble ! L'efficacité de Brigitte à passer son HDR a été un exemple que j'ai eu du mal à suivre.
- Faute de place, je remercie plus rapidement, mais tout aussi chaleureusement, d'autres collègues du domaine : Françoise Carré, Daniéla Garcia, Isabelle Delouis, Josette Rebeyrolle, Sylvie Després, Régine Teulier, Nathalie Girard, Laure Vieu et je dois en oublier !

Avant de me focaliser sur le microcosme toulousain, je tiens à remercier quelques collègues des laboratoires avec qui j'ai travaillé. Mon travail doit énormément à Jean-Paul Krivine, depuis la fondation du GRACQ jusqu'à l'idée de m'intéresser à la modélisation à partir de textes, en passant par l'utilisation de langages réflexifs dans MACAO. L'amitié qui nous lie est précieuse.

C'est grâce à lui que j'ai commencé à travailler avec Didier Bourigault sur l'évaluation de Lexter. Mes remerciements vont aussi à Didier pour toutes les discussions animées que nous avons échangées, pour tous les services qu'il m'a rendus en analysant des corpus à l'aide de ses logiciels, pour sa passion pour la recherche et le traitement automatique des langues enfin.

Grâce au GRACQ, j'ai rencontré Philippe Laublet, complice des cours d'Avignon, des premières conférences IC et observateur attentif du domaine. J'ai grandement apprécié sa relecture attentive et critique de ce mémoire. C'est toujours via l'animation du GRACQ que j'ai travaillé avec Jean Charlet, et que nous partageons depuis 15 ans de fréquentes discussions scientifiques, des documents et des cours, et nous avons le souci commun d'animer la communauté scientifique française du domaine. Je le remercie vivement pour son soutien et ses encouragements réguliers à passer cette habilitation.

Le GRACQ s'est doté d'un bureau formé d'un ensemble de complices que je remercie ici pour tous les débats scientifiques (et religieux) qui m'ont permis de prendre du recul sur le domaine : Bruno Bachimont, dont l'HDR a été un modèle pour moi, Manuel Zacklad, Francky Trichet qui a repris la liste info-ic, … Plus tard, le groupe TIA m'a permis de rencontrer d'autres chercheurs en traitement automatique des langues et terminologie, comme Adeline Nazarenko et Valérie Delavigne, dont les travaux ont également beaucoup éclairé mes recherches. Plus récemment, les membres du groupe ASSTICCOT et J. M. Salaün qui m'a fait confiance pour animer cette action spécifique, m'ont également beaucoup apporté. Enfin, grâce au projet Arkeotek, j'ai découvert encore d'autres articulations entre langue, raisonnements et connaissances. Merci à Valentine Roux, Blanche de Saizieu et Philippe Blasco de m'avoir initiée avec passion à SCD et au monde de l'archéologie des techniques.

Depuis 1991, mon quotidien bénéficie de l'ambiance de complicité et d'amitié qui règne dans l'équipe CSC. Outre Jean-Luc et Corinne, merci à Jean Frontin pour son calme et son écoute, et pour les nombreux coups de main, de la mise au point de MACAO jusqu'à la gestion des serveurs Samba. Plus récemment, j'ai apprécié de travailler avec Bernard Rothenburger, de bénéficier de l'expérience et du recul de Jacques Virbel, de partager le dynamisme de Pascale Zaraté.

Une mention particulière aux doctorants que j'ai encadrés, dont les thèses ont contribué énormément aux résultats présentés dans ce mémoire, et avec qui ces relations de travail sont devenues des liens d'amitié. Après Nada Matta, c'est avec un réel plaisir que j'ai approfondi les questions de traitement automatique des langues et de modélisation avec Patrick Séguéla. Sa curiosité scientifique et son souci de l'évaluation ont permis à CAMELEON de voir le jour et de disposer d'une base de marqueurs considérable. Je regrette toujours de ne pas voir su lui permettre de trouver un poste pour continuer dans le monde de la recherche. Depuis 4 ans, merci à Mustapha Baziz pour son dynamisme, sa rigueur et la richesse des questions qu'il a étudiées sur l'utilisation des ontologies en recherche d'information et ontologies. Merci à Nathalie Hernandez de m'avoir fait comprendre de près les questions d'exploration de collections documentaires. Plus récemment, Axel Reymonet et Kévin Ottens ont la patience de « subir » une directrice de thèse un peu hyperactive. Je leur sais gré de leur patience et de leur exigence.

Pour réaliser les différents logiciels présentés ici, j'ai eu la chance d'encadrer des étudiants exceptionnels, des stagiaires ingénieurs CNAM, à qui je dois MACAO, SADE, GEDITERM et CONSULTERM, mais aussi le plaisir du travail en équipe. En contribuant toute une année à un projet, ils ont enrichi la vie de l'équipe : Marie-Hélène Rivière tout particulièrement, Monique Routaboul, Pascal Lépine, Dominique Fournier et Eric Lecorgne.

Enfin, la vie au laboratoire et les projets m'ont permis de rencontrer des collègues à qui je dois aussi de nombreuses discussions autant que des réunions de travail agréables : les membres de l'équipe SMAC, et surtout Pierre Glize, André Macchonin, qui connaissent de longue date mes retards légendaires ; les membres de l'équipe SIG, en particulier Josiane Mothe et Mohand Boughanem, qui m'ont patiemment initiée à la recherche d'information ; enfin les membres du laboratoire travaillant sur le traitement automatique du langage naturel, et particulièrement, Philippe Muller, Patrick Saint-Dizier et Farah Bénamara.

Au-delà des équipes, je remercie aussi les directeurs de mon laboratoire de rattachement, l'IRIT, Jean Vignolles, qui, avant de partir à la retraite, me disait déjà de commencer mon HDR, et Luis Fariñas del Cerro, pour la confiance qu'il fait aux chercheurs. Merci à Jean-Paul Denier de

m'avoir accueillie dans le laboratoire ARAMIIHS en stage postdoctoral de 1989 à 1991. Enfin, le laboratoire ne serait pas aussi agréable à vivre et notre travail efficace sans le personnel qui l'anime et que je remercie ici : Maryse Cailloux, Denise Roncier, Agathe Baritaud, Geneviève Cluzel, Alain Monnier, Chantal Morand et d'autres spécialistes des missions, des vacations et de la comptabilité, enfin Jean-Pierre Ceccato, Jean-Pierre Baritaud, l'ensemble de l'équipe informatique, et en particulier Max Delacroix.

J'ai une pensée toute particulière pour tous mes amis, car la rédaction de ce mémoire m'a bien éloignée d'eux, Véronique, Gisèle et Isabelle qui m'accueillent lors de mes séjours à Paris, mais aussi de fidèles amis qui m'ont soutenue : Dominique, Odile, Françoise, Fred, Fabienne, Dinh et Lucie, mes amis de l'ACI, avec une pensée spéciale pour Alain, Cathy et Hubert, mes amies de l'ACE, et surtout les mamans complices, Sylvie, Isabelle, Anne-Françoise, Anne-Pascale, Marie …

Je garde le plus précieux pour la fin. Ma famille m'a toujours encouragée, soutenue et aidée. Ma belle-famille a pris le relais avec générosité. Je n'aurai jamais entrepris ce métier et ce mémoire sans l'amour et la confiance de mes parents, mais aussi leur disponibilité et leur soutien matériel. Un grand merci à mes adorables filles, qui ont fait preuve d'une grande patience avec une maman pas toujours disponible, et qui sont soulagée que j'ai enfin fini … et un très grand merci à Bruno, entre autres pour les ballotins de chocolats, dans le rôle difficile du mari d'une chercheuse décalée, angoissée et laborieuse, hyperactive et speedée …

TABLE DES MATIERES

Résumé 1

Remerciements 3

Table des matières 7

Chapitre 1 - Présentation du mémoire 11

1. 1 - Contexte général 11
1. 2 - Historique de mes recherches 12
1.2.1 MACAO, une méthode pour l'acquisition de connaissances expertes 12
1.2.2 MACAO-II, modélisation de connaissances et opérationnalisation 13
1.2.3 Bases de connaissances terminologiques et analyse linguistique de textes 14
1.2.4 Analyse de textes pour modéliser ontologies et terminologies 15
1.2.5 Apport des modèles conceptuels à différents types d'applications 15
1. 3 - Ma contribution 16
1.3.1 Plan du mémoire 17
1.3.2 Bibliographie des chapitres et bibliographie générale 18

Chapitre 2 - L'ingénierie des connaissances : analyse du domaine 19

2. 1 - Présentation du domaine 20
2.1.1 L'acquisition des connaissances 20
2.1.2 Les applications à base de connaissances 21
2.1.3 Place vis-à-vis de l'informatique et de l'intelligence artificielle 25
2.1.4 L'IC, carrefour disciplinaire 26
2.1.5 La difficulté de l'évaluation 30
2. 2 - Analyse des travaux sur les modèles conceptuels 34
2.2.1 La notion de modèle conceptuel 34
2.2.2 De l'acquisition à l'ingénierie des connaissances 41
2.2.3 Les résultats relatifs à la modélisation 45
2. 3 - Analyse des travaux sur ontologies et textes 47
2.3.1 Textes et modèles de connaissances 47
2.3.2 Ingénierie des connaissances et documents 51
2. 4 - Bilan : la problématique de l'IC 53

Chapitre 3 - Problématiques de recherche 57

3. 1 - Principes généraux retenus 58
3.1.1 Approche constructiviste 58
3.1.2 Répondre aux besoins par des modèles spécifiques 59
3.1.3 S'appuyer sur les connaissances en usage 60
3.1.4 Prendre en compte la démarche d'ingénierie des connaissances dans sa globalité 62
3. 2 - Questions et orientations de recherche 63
3.2.1 Recueil et modélisation de connaissances expertes 63
3.2.2 Modèles terminologiques, ontologies et textes 64

Chapitre 4 - Méthodes et outils pour la modélisation de connaissances expertes 69
4. 1 - Choix d'une modélisation ascendante 69
4.1.1 Motivations....69
4.1.2 De MACAO à MACAO-II : historique....70
4. 2 - MACAO : Acquisition de connaissances expertes....71
4.2.1 Contexte....71
4.2.2 Une méthode de modélisation cognitive 75
4.2.3 Enseignements tirés de MACAO 79
4. 3 - MACAO-II : Le modèle conceptuel comme grille d'acquisition 80
4.3.1 Contexte....80
4.3.2 La méthode MACAO-II 83
4.3.3 Représentation des connaissances : du langage naturel au système opérationnel 88
4. 4 - Bilan sur la modélisation conceptuelle 93
4.4.1 Synthèse des résultats établis 93
4.4.2 Situation de MACAO-II par rapport à d'autres travaux....95
4.4.3 Bilan 95
4. 5 - Publications sur ces travaux....96

Chapitre 5 - Modèles pour les ressources terminologiques et ontologiques 99
5. 1 - Des experts aux textes, des modèles conceptuels aux ontologies....100
5.1.1 Nouvelles orientations thématiques 100
5.1.2 Textes, ontologies et bases terminologiques....102
5.1.3 Partis pris, choix retenus 107
5. 2 - Modèles de données pour les ressources terminologiques et ontologiques....108
5.2.1 Termes, notions et concepts 109
5.2.2 Un modèle pour les bases de connaissances terminologiques 111
5.2.3 Un modèle pour les ontologies régionales à composante terminologique 113
5. 3 - Logiciels pour la structuration et l'exploitation de terminologies....119
5.3.1 GEDITERM : gestion de bases de connaissances terminologiques 120
5.3.2 CONSULTERM : consultation de bases terminologiques 124
5.3.3 Évaluations expérimentales....125
5. 4 - Modèles de ressources terminologiques et ontologiques : bilan....129
5.4.1 Synthèse sur la notion de BCT : modèle neutre ou modèle lié à une application ?....129
5.4.2 Originalité des propositions relatives aux BCT 130
5.4.3 Des BCT aux ontologies....131
5. 5 - Publications sur ces travaux....131

Chapitre 6 - Des textes aux ontologies, des ontologies aux textes 133
6. 1 - Traitement Automatique des Langues et modélisation conceptuelle 134
6.1.1 Contexte : Traitement du langage naturel pour l'identification de connaissances 134
6.1.2 Aide au repérage et à la structuration de termes à partir de textes....136
6.1.3 Étude des relations sémantiques : CAMELEON 143
6. 2 - Méthodes et plates-formes pour construire des ressources terminologiques et ontologiques 154
6.2.1 Contexte....154
6.2.2 Modélisation d'ontologies à partir de textes : une méthode 157
6.2.3 Modélisation d'ontologies à partir de textes : plate-forme de modélisation TERMINAE 163
6.2.4 Expérimentations et validations 166
6.2.5 Impact de l'application ciblée sur le processus de construction de RTO 168
6. 3 - Modèles conceptuels comme accès au contenu de documents....169

6.3.1 Motivations 170
6.3.2 Apport de modèles construits à partir de textes à la consultation documentaire 171
6.3.3 Premières conclusions 173
6. 4 - Bilan 174
6.4.1 Traitement automatique des langues pour la construction de RTO 174
6.4.2 Propositions méthodologiques 176
6.4.3 Modèles conceptuels et consultation documentaire 179
6. 5 - Publications sur ces travaux 181
6.5.1 Publications sur les outils de TAL et les méthodes de construction de RTO 181
6.5.2 Publications sur modèles conceptuels et accès au contenu de documents 182

Chapitre 7 - Synthèse et perspectives : ontologies et documents 185
7. 1 - Nature de mes contributions 185
7.1.1 Cohérence et évolution de mes travaux 185
7.1.2 Résultats établis 186
7.1.3 Programme de recherche 189
7. 2 - Construction d'ontologies à partir de textes : valorisation des logiciels 190
7.2.1 De CAMELEON à l'exploration de corpus étiquetés 190
7.2.2 De TERMINAE à une plate-forme de modélisation à partir de textes 192
7.2.3 Situation par rapport aux perspectives du domaine 193
7. 3 - Ontologies, documents et recherche d'information 194
7.3.1 Ingénierie des connaissances et recherche d'information : convergences 195
7.3.2 RTO pour optimiser les résultats d'un moteur de recherche 195
7.3.3 Classification de documents pour la veille technologique 198
7.3.4 Ontologies pour l'annotation sémantique de documents structurés 201
7.3.5 Situation par rapport aux perspectives du domaine 205
7. 4 - Mise à jour et maintenance de modèles conceptuels 205
7.4.1 Observation des évolutions de l'expression de connaissances dans le temps 206
7.4.2 Maintenance des ressources terminologiques en lien avec des textes 207
7.4.3 Maintenance classique par enrichissement manuel 208
7.4.4 Agents adaptatifs pour construire et maintenir une ontologie dynamique 209
7.4.5 Situation par rapport aux perspectives du domaine 210
7. 5 - Conclusion 211
7. 6 - Publications sur ces travaux 212

Chapitre 8 - Bibliographie 215
8. 1 - Acquisition et Ingénierie des connaissances, modèles conceptuels 215
8. 2 - Ontologies, Web sémantique, terminologie et linguistique 218

CHAPITRE 1 - PRESENTATION DU MEMOIRE

1. 1 - Contexte général

Ce mémoire présente une contribution au domaine de l'ingénierie des connaissances, pris comme un champ de l'informatique qui s'intéresse à la mise au point de logiciels s'appuyant sur des connaissances pour assister un utilisateur dans sa tâche.

L'ingénierie des connaissances est un domaine dont les résultats significatifs sont parfois mal identifiés par les autres communautés scientifiques, malgré les analyses présentées dans divers ouvrages sur le domaine qui en rappellent régulièrement ses enjeux, fondements et contributions. Ce manque de lisibilité peut s'expliquer en partie par sa position, à la frontière entre l'intelligence artificielle et d'autres disciplines venant plus des sciences humaines. Il provient sans doute aussi de l'évolution de ses objets d'étude, qui donne une impression de surface où les effets de mode et la dispersion orientent plus les recherches que de vraies problématiques scientifiques. Si l'on observe le contexte actuel, par exemple, l'ingénierie des connaissances se trouve au cœur de demandes fortes au sein de la société et des entreprises en matière de gestion des connaissances, de veille technologique, de gestion documentaire et de recherche d'information. Ainsi, la notion d'ontologie est mentionnée dans une multitude de travaux, en particulier relativement à la perspective du Web Sémantique. Même lorsque le terme ontologie est utilisé de manière abusive pour renvoyer à des structures de données plus simples (listes de mots, thésaurus, …), cette tendance reflète bien le besoin essentiel et récurrent en modèles reflétant des connaissances structurées, pouvant donner du sens à des informations plus brutes ou moins accessibles. L'ingénierie des connaissances a donc une place fondamentale par le rôle qu'elle peut jouer en amont de tous ces besoins.

Le terme même d'ingénierie des connaissances me permet d'esquisser les enjeux et les méthodes de travail de ce champ disciplinaire, sur lesquels je reviendrai plus précisément par la suite. En tant qu'ingénierie, il s'agit de fournir des méthodes, des langages et des logiciels pour assister le cogniticien, ces propositions étant éventuellement reprises d'autres domaines de l'informatique ou d'autres disciplines. Cette ingénierie traite de connaissances dans la mesure où les systèmes informatiques à concevoir doivent assister un utilisateur dans des tâches mal définies, complexes et faisant appel à des savoir-faire. La difficulté de spécification du système a deux origines : d'une part, la résolution informatique ne s'appuie pas sur des algorithmes ou procédures connus ; d'autre part, pour que la tâche soit réalisée au mieux par l'utilisateur dans un ensemble de contextes, on ne cherche pas à en automatiser le plus possible le traitement, mais à en prendre en charge certaines parties. La question des connaissances en ingénierie des connaissances est donc délicate. Historiquement, elle a d'abord fait référence à une expertise humaine dont devait rendre compte la base de connaissances de systèmes experts. Le problème était d'une part de rendre explicite des savoir-faire, d'autre part de produire des formules logiques en rendant compte.

Ensuite, elle a évolué vers l'organisation structurée dans des modèles conceptuels de différents types de connaissance nécessaires à la résolution de problèmes. Les modèles ont permis de souligner qu'il s'agit de reconstruire de manière artificielle une représentation de connaissances choisies en fonction d'un objectif opérationnel. Aujourd'hui, le statut du modèle est toujours l'objet de débat, puisque certains travaux le considèrent comme un modèle formel décrivant une théorie du domaine, et d'autres comme un modèle à opérationnaliser rendant compte des connaissances sur un domaine. Ce choix n'est pas toujours formulé clairement. Le deuxième point de vue, retenu au sein de la communauté française d'ingénierie des connaissances, est celui que je défends. Les connaissances sont bien aux deux extrémités du processus d'ingénierie. Le statut des modèles que je retiens est donc celui de médiateurs, de représentations dont l'opérationnalisation va donner à l'utilisateur et au système les moyens de raisonner, d'agir et de produire des connaissances.

Je formule donc le problème de l'ingénierie des connaissances comme un problème de modélisation. Et au sein de cette problématique, c'est la partie amont, la mise en place de modèles adéquats à partir de traces de connaissances, qui me concerne particulièrement. De ce fait, parmi les différents rôles que jouent les modèles conceptuels dans la modélisation, je me focaliserai sur trois d'entre eux : celui de cible de la modélisation pour définir le système à construire, celui de structure permettant d'organiser et présenter ce qui est analysé et enfin celui de grille pour repérer les lacunes et orienter la suite du processus. Pour aborder ce problème, je développerai plusieurs propositions relevant de ce qui est appelé une « démarche ascendante », à savoir des méthodes, logiciels et langages définis en vue d'une meilleure localisation, explicitation et mise en forme de connaissances à partir de leurs usages afin de construire des modèles.

Ce mémoire rassemble donc mes diverses contributions, présentées suivant le type de connaissances abordées (expertises, activités humaines et documents techniques), et touchant à toutes les facettes de la modélisation : des méthodes, des techniques de recueil et d'analyse de connaissances ainsi que des logiciels d'aide à la modélisation et à la représentation des connaissances. Ces contributions comportent une dimension historique évidente, puisqu'elles couvrent des travaux répartis sur une quinzaine d'années. De ce fait, elles n'ont de sens que situées au cœur des réflexions plus fondamentales du domaine qui l'ont fait évoluer et qui ont conduit à des changements d'orientation successifs. Ainsi, par ce mémoire, j'espère donner une image riche et diversifiée de l'ingénierie des connaissances, qui, tout en reflétant mon point de vue, en souligne la continuité, les résultats significatifs et la contribution scientifique.

Avant de présenter mes contributions et le plan de ce mémoire, je dresse un historique rapide de mes travaux. Il donne une première image de leurs différentes facettes et de leurs résultats, qui ont conduit aux propositions énoncées par la suite.

1. 2 - Historique de mes recherches

1.2.1 MACAO, une méthode pour l'acquisition de connaissances expertes

Mes recherches dans le domaine de l'ingénierie des connaissances ont débuté avec ma thèse de 1986 à 1989, effectuée au sein de l'équipe SMI (Systèmes Multi-inférentiels) du laboratoire LSI, sous la direction de M. Borillo et encadrée par J.-L. Soubie. Ce sujet était motivé par les diverses expériences et difficultés rencontrées au sein de cette équipe pour construire les bases de connaissances d'un système multi-expert. L'objectif visé était de définir une méthode d'acquisition des connaissances expertes et un environnement de structuration de ces connaissances avant leur représentation sous forme de règles de production. Ce problème était peu abordé par les chercheurs français jusque-là. L'orientation choisie était d'emblée interdisciplinaire : d'une part, s'appuyer sur les travaux de la psychologie cognitive sur la nature des connaissances expertes pour mieux connaître les processus de résolution de problème, savoir les identifier et faire expliciter les connaissances mises en œuvre ; d'autre part, évaluer et adapter différentes techniques utilisées en

psychologie et surtout en ergonomie pour les proposer au sein d'une méthode dédiée à la construction de systèmes à base de connaissances. Cette thèse a produit une méthode, MACAO, proposant des repères et des supports à différentes techniques d'entretiens, une plate-forme de modélisation intégrant des outils de recueil de connaissances comme les grilles répertoires et une représentation de connaissances à l'aide de schémas. La méthode a été utilisée sur des cas d'école et dans le cadre du projet SAMIE[1].

Le domaine de l'acquisition des connaissances en était alors à ses débuts et une communauté scientifique commençait à s'organiser, en particulier grâce aux conférences KAW (Knowledge Acquisition Workshops) se tenant depuis 1986 en Amérique du Nord et depuis 1987 en Europe, à l'initiative de B. Gaines (chercheur à l'université de Calgary) et J. Boose (chercheur chez Boeing). En France, les premières journées scientifiques sur ce thème se sont tenues en 1988 et 1989 en lien avec le PRC-IA, et la première édition des Journées d'Acquisition des Connaissances (JAC) a eu lieu en 1990 à l'initiative de chercheurs travaillant en apprentissage automatique.

1.2.2 MACAO-II, modélisation de connaissances et opérationnalisation

Ce travail a été poursuivi dans le cadre d'une deuxième thèse, réalisée par Nada Matta de 1991 à 1995. L'objectif de cette thèse était de reprendre la méthode MACAO pour en dépasser certaines limites : prévoir une représentation des connaissances qui permette de mieux expliciter la méthode de résolution de problème présente dans le modèle conceptuel ; pouvoir, à partir de cette représentation, simuler la mise en œuvre de cette méthode ; mieux aider le cogniticien dans le passage du modèle conceptuel à la base de connaissances proprement dite. Ce travail s'est accompagné de l'étude des théories relatives aux différents modes de raisonnement (en psychologie et en IA). Il a conduit à la mise au point du langage MONA de modélisation des connaissances. La thèse de Nada Matta a également conduit à la mise en forme d'une nouvelle méthode, MACAO-II et au développement d'une plate-forme associée. MACAO-II permet de gérer des modèles de tâches de la bibliothèque de KADS et de les adapter pour construire le modèle conceptuel d'une expertise.

À partir de cette thèse, plusieurs développements ont été menés en collaboration avec d'autres laboratoires. Ainsi, un module d'opérationnalisation de modèles décrits en MONA a été développé, utilisant le langage ZOLA développé à l'IRIN par l'équipe de P. Tchounikine. Ce module permet de valider les modèles par simulation et surtout de construire des modèles adaptés à des systèmes coopératifs. Afin d'enrichir la représentation des connaissances du domaine avec MONA, une coopération avec F. Tort et C. Reynaud du LRI a permis de reprendre certaines propositions du système ASTREE, en particulier pour la formalisation des relations. Dans la continuité de ces collaborations, un travail conjoint avec ces deux équipes a visé une meilleure organisation des connaissances entre les niveaux domaine et résolution de problème ou tâche, ainsi qu'une articulation mieux contrôlée entre ces types de connaissance, à travers la structure de rôle.

Enfin, plusieurs évaluations expérimentales de cette nouvelle version de MACAO, dont le projet SADE (1993), ont été effectuées. Elles ont montré l'intérêt d'exploiter la complémentarité entre méthodes ascendantes (constructives) et descendantes (par réutilisation de modèles de résolution) pour la construction de modèle, et de mieux caractériser les résultats obtenus selon chacune d'elles. Enfin, dans le cadre d'une collaboration avec J. Breuker de l'Univ. d'Amsterdam, des outils ont été définis pour assurer une maintenance aisée et cohérente du modèle conceptuel et de la base de connaissances associée. À partir du module d'opérationnalisation en ZOLA, le langage MONA a été enrichi afin de mieux tracer le processus de modélisation et mesurer en quoi un modèle conceptuel facilite la maintenance du système.

[1] Projet mené avec la société MMS au sein du laboratoire ARAMIIHS au cours de mon stage post-doctoral

1.2.3 Bases de connaissances terminologiques et analyse linguistique de textes

C'est à partir de 1993 que mes travaux ont pris un tournant en se focalisant aussi sur les documents comme sources de connaissances et sur les outils d'analyse terminologique comme moyen de les exploiter. La motivation initiale, classique, visait un gain de temps pour le repérage de la terminologie et pour la structuration des concepts du domaine, travail correspondant à une des tâches du processus de modélisation. Ce glissement thématique a bénéficié d'une collaboration avec D. Bourigault, chercheur en Traitement Automatique des Langues à EDF à cette époque, et des linguistes du laboratoire ERSS de Toulouse Le Mirail. Une première expérience a consisté à dégager la terminologie du domaine à partir de documents techniques à l'aide du logiciel LEXTER au sein du projet SADE. LEXTER permet de trouver un ensemble très riche de termes du domaine, de les utiliser pour définir des concepts et de repérer des synonymies entre termes. L'avantage de son utilisation est aussi de disposer de liens entre le modèle conceptuel et les textes « source ».

Les résultats prometteurs obtenus m'ont amenée à explorer plus systématiquement la manière de conduire des analyses terminologiques en amont de l'acquisition des connaissances, en étroite collaboration avec des linguistes. Une convergence d'intérêt avec la linguistique de corpus et la structuration de terminologies a donné un caractère plus ambitieux à cette piste. La période entre 1993 et 1998 correspond donc à celle d'une évolution thématique qui s'est stabilisée avec la confirmation de l'intérêt de cette approche pour la construction d'ontologies et de l'importance (provisoirement exagérée sans doute) des ontologies dans les applications.

Cette période est également celle d'une transition dans le domaine de l'ingénierie des connaissances. Les projets européens (CommonKads et affiliés) ou américains (Protégé) étant parvenus à des propositions stables et convergentes en matière de méthodes et de modèles pour l'ingénierie des connaissances, plusieurs courants répondant à des besoins plus spécialisés ont vu le jour. Parmi ceux-ci, je citerai la conception de systèmes coopératifs (répartition dynamique des tâches entre système et opérateur), la prise en compte de l'organisation dans laquelle s'intègre le système, la définition de langages standardisés facilitant l'interopérabilité, ou encore l'exploitation de connaissances pour un meilleur accès au contenu du Web. Au croisement de ces trois dernières pistes, la notion d'ontologie, redéfinie par l'ingénierie des connaissances comme un modèle conceptuel consensuel d'un domaine, a pris une place croissante.

Un premier objet d'étude, les *bases de connaissances terminologiques* (BCT), a été le support de questions et de contributions convergentes venant de l'IC, de l'analyse linguistique de textes et de la linguistique de corpus. Ces bases contiennent des connaissances sur la terminologie d'un domaine, sous la forme d'un réseau conceptuel associé à des fiches terminologiques. Entre 1995 et 1998, mes travaux ont porté sur la définition et l'évaluation de ces structures de données et, pour cela, sur le développement d'un support logiciel pour les gérer. Élaboré en collaboration avec A. Condamines (ERSS), le modèle de données proposé pour les BCT associe textes, termes et un réseau conceptuel. La représentation des connaissances choisie se situe au niveau conceptuel puis elle est formelle (en logique de description). J'ai étudié la nature des changements qu'introduit la formalisation sur les données. Des outils ont été développés pour leur gestion (GEDITERM) et leur utilisation (CONSULTERM). Ces logiciels ont permis de conduire des recherches sur le passage de données lexicales à un modèle conceptuel, sur la traçabilité des choix de modélisation et le rôle des textes comme traces de ces choix.

Du point de vue méthodologique, j'ai essayé de rapprocher les méthodes et outils utilisés pour la modélisation conceptuelle pour construire une BCT. J'ai également cherché à évaluer en quoi une BCT pourrait être un produit intermédiaire utile pour la construction de modèles du domaine dédiés à des applications. Deux projets de valorisation menés avec A. Condamines (l'un avec la DDE de Midi-Pyrénées, l'autre, MOUGLIS, avec la DER d'EDF) ont permis d'utiliser et valider le modèle de données et les logiciels, ainsi que la méthodologie. Ils conduisent aussi à remettre en question l'intérêt de disposer d'une BCT pour construire un modèle du domaine. Or, l'hypothèse de travail des linguistes était double : une BCT pourrait permettre de rendre compte de

« toutes les connaissances » contenues dans un texte à partir des seules traces linguistiques ; et ainsi, une BCT pourrait servir de source de connaissances à modifier ensuite en fonction des objectifs de l'application. Cette hypothèse a été depuis complètement revue, d'une part à cause de l'observation pratique de l'impossibilité de rendre compte de manière neutre et exhaustive du contenu d'un texte, et, d'autre part, pour des motivations plus théoriques liées à la notion *d'interprétation*, bien présente dans le processus de construction de modèle.

1.2.4 Analyse de textes pour modéliser ontologies et terminologies

En parallèle, dans la continuité des recherches sur le processus de modélisation conceptuelle, j'ai étudié l'intégration de l'analyse terminologique et de ses résultats dans ce processus. Au cours de différents projets, en collaboration avec A. Condamines et D. Bourigault de l'ERSS, j'ai expérimenté ou validé des logiciels d'aide à l'extraction d'éléments linguistiques porteurs de connaissances, et donc utiles pour la construction de modèles, comme les concordanciers (YAKWA), les logiciels d'extraction de termes et de réseau terminologique LEXTER puis SYNTEX, des logiciels d'analyse distributionnelle comme UPERY. Parce que les relations sémantiques sont un des moyens de repérer des concepts et de justifier leur définition, j'ai encadré la thèse de P. Séguéla (sous la direction de J.-L. Soubie) sur l'utilisation de patrons linguistiques pour le repérage de relations lexicales puis la mise en relation conceptuelle. Elle a débouché en 2000 sur la mise au point du logiciel CAMELEON, dont deux nouvelles versions ont été depuis développées afin de balayer des textes étiquetés grammaticalement.

Ces différents logiciels ont été évalués et j'ai spécifié leur intégration au sein d'une chaîne de traitements dans le cadre d'une méthodologie de construction de ressources terminologiques et ontologiques à partir de textes : TERMINAE. Initialement définie au LIPN par B. Biébow et S. Szulman pour représenter formellement des connaissances tirées de spécifications en langage naturel, cette méthode s'appuie sur un logiciel de modélisation dédié qui débouche sur une représentation en logique de description. Cette représentation a la structure d'une ontologie. Suite à l'expérience tirée de GEDITERM, j'ai collaboré avec le LIPN pour intégrer dans TERMINAE les éléments nécessaires à la gestion de BCT d'une part, et, d'autre part, des résultats des logiciels d'analyse de textes (en particulier des extracteurs de termes LEXTER puis SYNTEX). De ce fait, le modèle de données a évolué, intégrant désormais des fiches terminologiques et des éléments textuels (phrases ou paragraphes). De nouvelles interfaces de saisies et de nouveaux modules de développement des données ont été définis, à la spécification desquels j'ai contribué.

Plus fondamentalement, l'ensemble de ces travaux a permis une première réflexion sur l'apport d'éléments terminologiques et linguistiques pour améliorer la qualité et l'acceptabilité des modèles. En effet, on s'attend à ce que les systèmes qui les utilisent répondent mieux aux besoins des utilisateurs, car les éléments linguistiques contribuent à mieux en faire comprendre et accepter le contenu. Ensuite, la construction de modèles à partir de textes renouvelle la question du degré d'opérationnalisation des connaissances. Il paraît plus naturel de considérer qu'une partie des connaissances peut rester sous une forme peu opérationnelle, telle qu'elle se présente dans les textes, car ils sont accessibles à l'utilisateur et leur structure peut être maniée assez simplement par le système d'information. Seul un noyau de connaissances doit être formalisé, les connaissances sur lesquelles le système doit raisonner pour répondre aux besoins des utilisateurs. Jusque-là, l'objectif était de modéliser et d'opérationnaliser uniformément toutes les connaissances nécessaires au système. En renouvelant le questionnement de l'ingénierie des connaissances, cette approche élargit les types de réponse informatique envisagés face à des besoins d'utilisateurs.

1.2.5 Apport des modèles conceptuels à différents types d'applications

De ce fait, via des contrats de valorisation, j'ai également élargi l'éventail des applications possibles pour évaluer ces méthodes et outils, au-delà de l'aide à la résolution de problèmes :

gestion documentaire, mémoire d'entreprise, modélisation des utilisateurs, construction de systèmes coopératifs. Or chaque type d'application soulève des problèmes de recherche spécifiques, qui dépassent l'adaptation des logiciels. La majorité de ces expériences a concerné l'intérêt de modèles conceptuels ou terminologiques, donc de données sémantiques structurées, pour accéder à ou naviguer dans des éléments documentaires. La répartition de la « résolution de problème » au sein du couple système-utilisateur final est ici tout à fait inversée par rapport aux systèmes à base de connaissances. Le système exploite des connaissances du domaine pour orienter au mieux un utilisateur qui a l'initiative de la recherche, et surtout de l'interprétation du contenu documentaire en fonction du contexte dans lequel il réalise sa tâche. Au cours de différents projets, cette question a été déclinée sous plusieurs formes : (i) mesurer l'intérêt d'un modèle de tâches pour la consultation de guides de procédures (projet MOUGLIS), (ii) juger de l'apport d'un modèle conceptuel pour faciliter la sélection des termes et leur structuration dans un index (projet HYPERPLAN), (iii) structurer un index de site web selon une approche terminologique (IndexWeb), (iv) évaluer l'apport des ontologies pour la reformulation de requêtes (DEA de M. Baziz) ou encore (v) pour la consultation de documents structurés (projet ARKEOTEK). Ces différents questionnements exigent une réponse interdisciplinaire entre spécialistes des sciences de l'information ou de recherche d'information, du traitement automatique des langues et de l'ingénierie des connaissances, linguistes et ergonomes. Ma participation à des groupes de travail comme TIA du GRD I3 et l'action spécifique « corpus et terminologies » m'ont permis de mener ce type de réflexion avec des chercheurs de ces disciplines.

Mes recherches en cours reprennent ces différentes expériences pour en tirer des éléments méthodologiques sur les étapes et logiciels utiles à la construction de différents types de modèle terminologique ou ontologique pour chacune des classes d'application possibles. Il me semble important, pour proposer un cadre générique au sein duquel adapter des approches et des outils en fonction des applications, de s'appuyer sur des retours d'expérience. Plus que de définir un cadre méthodologique, aujourd'hui assez consensuel, la difficulté est de mettre en place pratiquement une chaîne de traitements du langage adaptés et d'aides à la modélisation.

Une autre problématique ressort de ces expériences : celle de la maintenance des modèles en cohérence avec le vocabulaire et les connaissances du domaine, les textes à indexer, consulter ou explorer. Le contexte d'usage des ontologies, et cela est encore plus criant dans le cas du web, est d'évidence en évolution permanente. Or les ontologies sont souvent considérées comme des représentations stables puisque consensuelles. Je voudrais prévoir un processus de maintenance d'ontologies dynamiques, afin qu'elles puissent être revues à la demande, en fonction des évolutions du contexte dans lequel elles sont utilisées. Une recherche est en cours sur ce thème, la solution proposée s'appuyant sur l'analyse de texte à l'aide d'agents adaptatifs.

1. 3 - Ma contribution

Ce mémoire rassemble mes contributions et réflexions tout en les situant dans la problématique globale de l'ingénierie des connaissances.

Dans mes recherches, le problème de l'ingénierie des connaissances est posé comme celui de l'identification de connaissances, de leur modélisation et de leur restitution au sein d'une application informatique. En réponse à ce problème, j'ai défini des méthodes, des techniques de recueil et d'analyse de connaissances ainsi que des logiciels d'aide à la modélisation et à la représentation des connaissances. Dans un premier temps, j'ai abordé la modélisation conceptuelle dans sa généralité, de manière à faire une proposition cohérente, en considérant l'expertise puis les activités humaines comme sources de connaissances privilégiées. La méthode et l'environnement de modélisation proposés s'appuient sur la prise en compte de résultats de psychologie cognitive et d'ergonomie. Dans un deuxième temps, je me suis focalisée sur l'étude des textes et des modèles de connaissances de domaines que sont les ontologies. J'ai alors approfondi les liens entre langage et connaissances pour définir une méthode et des outils faisant appel à la linguistique et au traitement

automatique des langues. Cette contribution élargit le champ initial de l'ingénierie des connaissances tant du point de vue des problématiques, des propositions pratiques que des types d'applications prises en compte.

Une des manières d'aborder l'ingénierie des connaissances est de prendre en compte les connaissances telles qu'elles sont exprimées et utilisées par les experts, les spécialistes, les utilisateurs ou dans les textes. C'est l'optique que j'ai retenue. Or l'identification, l'analyse et la modélisation de connaissances à partir de leur utilisation soulèvent des problèmes qui débordent des aspects informatiques et techniques. Ces tâches requièrent d'étudier les compétences, tâches et processus cognitifs des individus, ainsi que leur organisation sociale ou humaine. Ces études nécessitent de faire appel à d'autres disciplines. Je place au centre de mon approche des collaborations interdisciplinaires, avec des psychologues, des ergonomes puis des linguistes.

1.3.1 Plan du mémoire

Le chapitre 2 est une analyse du domaine de l'ingénierie des connaissances et de la manière dont sa problématique a évolué. À partir de la question initiale de l'acquisition des connaissances pour construire des systèmes experts, je montre comment la problématique du domaine s'est orientée vers celle de la modélisation conceptuelle puis, plus largement encore, vers celle de l'identification et de la mise en forme au sein d'application informatique de connaissances pour assister un opérateur dans des tâches cognitives. J'introduis la notion de modèle conceptuel pour faire un point sur les différentes contributions relatives à la modélisation de connaissances dans le domaine. Les écueils rencontrés par les systèmes tentant de prendre en charge la totalité de processus cognitifs comme la résolution de problème ou de la prise de décision ont orienté l'ingénierie des connaissances vers des applications où les modèles sont exploités pour produire des aides interactives. Les modèles des connaissances des domaines concernés ont pris alors une place croissante, qui se traduit par l'étude des ontologies. De plus, d'autres sources de connaissances, comme les textes techniques, ont été étudiées pour parvenir à des modèles stables et consensuels. Je retrace ces deux orientations.

Le chapitre 3 articule l'exposé de ce cadre général avec celui de mes travaux. J'en justifie l'objectif, qui est de réaliser des aides aux processus d'abstraction et de caractérisation des connaissances au cours de leur modélisation. Je présente d'abord la ligne directrice selon laquelle j'aborde la modélisation conceptuelle, un ensemble de choix qui en définissent la cohérence. Je dresse ensuite deux listes de questionnements et d'orientations scientifiques, correspondant à deux axes dans mes recherches : les premières questions portent sur la modélisation conceptuelle à partir de connaissances d'experts, les suivantes abordent les liens entre analyse de textes et modélisation. Le détail des travaux et contributions fera l'objet des chapitres 4, 5 et 6.

Dans le chapitre 4, je présente ma contribution au problème de l'acquisition et de la modélisation des connaissances expertes. Deux périodes ont jalonné ces recherches sur la modélisation ascendante. Le noyau initial de ma proposition, la méthode MACAO, vise la mise au point de systèmes experts, la priorité étant la modélisation cognitive de l'expertise. Cette méthode a évolué pour mieux prendre en compte le contexte de mise en œuvre de l'expertise et la tâche des utilisateurs du futur système. Un nouveau point de vue a été retenu ensuite et a donné lieu à la méthode MACAO-II. Le modèle conceptuel y est considéré comme un modèle du système, spécifiant les problèmes à traiter (et comment ils seront traités) sous la forme d'un modèle de raisonnement décrit à l'aide de tâches et de méthodes. Enfin, plusieurs valorisations de MACAO-II ont eu pour objectif, à travers des collaborations avec d'autres équipes, de diversifier les techniques de modélisation, et d'assurer le suivi jusqu'à la réalisation d'un modèle opérationnel.

Le chapitre 5 aborde la modélisation de connaissances issues d'analyses de textes. Je situe et présente les modèles construits selon ces analyses, appelés ensuite ressources terminologiques et ontologiques. Je souligne la position originale que je tiens sur le statut de ces modèles. Mes

contributions comportent deux modèles de données et des plates-formes de modélisation, l'une sur les bases de connaissances terminologiques (GEDITERM) et l'autre sur les ontologies (TERMINAE).

Au chapitre 6, je me focalise sur les apports du traitement automatique des langues et des approches linguistiques à la définition de logiciels pour faciliter la construction de ces modèles à partir de textes. Je présente le logiciel d'extraction de relations conceptuelles (CAMELEON) que j'ai défini et plusieurs bilans d'évaluation d'extracteurs de termes. J'aborde également la difficulté qu'il y a à définir l'utilisation combinée de ces différentes sortes de logiciels au sein d'une démarche méthodologique et d'une plate-forme de modélisation d'ontologies (TERMINAE). Je montre ensuite en quoi les modèles ainsi construits à partir de textes sont tout à fait adaptés à la définition de modes originaux de consultation de documents. À travers plusieurs expérimentations, je confirme l'impact du type d'application utilisant l'ontologie ou le modèle sur sa construction.

Le chapitre 7 récapitule mes contributions et présente les orientations amorcées pour les poursuivre. À partir des travaux en cours, mon projet de recherche s'oriente vers une meilleure prise en compte de l'utilisation et de l'évolution attendue des ontologies au moment de leur construction. D'une part, en me focalisant sur les applications documentaires et en recherche d'information, je voudrais mieux cibler le type de modèle adapté à ces applications et la manière de les utiliser. Je défends l'intérêt de construire ces modèles par analyse de textes et d'utiliser ces mêmes techniques pour faciliter annotation et indexation sémantiques. D'autre part, face à l'évolution constante des connaissances et des collections de documents d'un domaine, je vise la définition d'un cycle de maintenance des modèles en cohérence avec leur utilisation ainsi qu'avec le contenu des documents. Cette question me semble déterminante pour la généralisation de l'usage des ontologies. Je termine par les axes envisagés pour l'étudier et leur situation par rapport aux perspectives du domaine.

1.3.2 Bibliographie des chapitres et bibliographie générale

Les publications ou exposés sur mes travaux sont listés à la fin des chapitres 4, 5, 6 et 7, dans l'ordre chronologique et regroupés en un ou plusieurs thématiques suivant le contenu du chapitre. Dans le texte, les appels à ces références sont du type [Rapport-RIVIERE, 90] ou [EKAW, 97].

Les références à d'autres travaux sont rassemblées en fin de mémoire, en deux listes alphabétiques, l'une pour la modélisation de connaissances expertes, l'autre sur la modélisation de connaissances à partir de textes et les ontologies. Dans le texte, les appels à ces références sont du type (Linster, 1991) ou (Schreiber *et al.*, 1994).

CHAPITRE 2 - L'INGENIERIE DES CONNAISSANCES : ANALYSE DU DOMAINE

Ce chapitre s'organise comme suit : après l'introduction, je reviens à la problématique initiale du domaine, les méthodes et aides à la construction de systèmes experts, à partir de laquelle je trace son évolution (2.1). Cette évolution étant étroitement liée à la nature des applications visées, j'évoque un historique du domaine à travers celui des systèmes informatiques d'aide aux opérateurs dans des tâches cognitives. Ceci me permet de situer l'ingénierie des connaissances par rapport à d'autres disciplines, tout d'abord l'informatique et l'intelligence artificielle, puis toutes celles qui abordent avec elle sa problématique et qui font de l'ingénierie des connaissances un carrefour disciplinaire.

J'aborde ensuite les principaux débats et résultats relatifs à la notion de modèle conceptuel et au processus de construction de ces modèles (2.2). La notion de modèle a été renouvelée avec l'étude des textes comme sources de connaissances, puis suite à l'utilisation de modèles comme les ontologies pour la gestion des connaissances et la gestion documentaire. J'en évoque les nouveaux enjeux et résultats (2.3).

Depuis les premiers travaux sur l'acquisition des connaissances autour de 1980, le domaine de l'ingénierie des connaissances a pris sa place comme un champ de recherches scientifiques ancré dans l'informatique, proche de l'intelligence artificielle, mais aussi à la croisée de disciplines diverses concernées par la mise en œuvre, le repérage ou l'analyse de connaissances, comme la linguistique de corpus, la sémantique, l'ergonomie ou la sociologie. Sa problématique et ses résultats méritent d'être situés dans une dimension historique, marquée par l'évolution des systèmes qu'elle cherche à produire, qui sont des logiciels pouvant aider un utilisateur à mettre en œuvre des connaissances.

Les résultats des vingt années de recherches que compte l'ingénierie des connaissances ont conduit à des propositions de plus en plus fondées et mieux définies, au centre desquelles se trouvent les modèles conceptuels et le processus de mise au point de ces modèles. Ces notions-clés ont évolué en accord avec ce que représente un logiciel d'aide à l'utilisateur, mais aussi avec une analyse de ce qu'est un modèle de connaissances qui serve de base à la construction de tels systèmes.

Le recul actuel sur la part des tâches prises en charge dans l'aide donnée par un système informatique à son utilisateur a conduit à un renversement dans les types de connaissance auxquels s'intéresse la modélisation conceptuelle. Après avoir mis l'accent essentiellement sur les connaissances de résolution de problème, plusieurs courants se dessinent aujourd'hui, qui s'intéressent soit à la dimension collective des connaissances, soit aux interactions homme-système, soit encore à la contribution de modèles de connaissances d'un domaine pour accéder à

des sources existantes, comme les documents. Parmi les modèles étudiés, je me concentre sur les ontologies, qui décrivent les domaines de manière statique, sous forme de réseaux conceptuels.

Afin de situer mes travaux de recherche, je prends le temps de faire une présentation à la fois historique et critique de l'ingénierie des connaissances, qui ne se veut ni exhaustive ni neutre. Elle se focalise sur les facettes et questionnements du domaine en lien avec mes contributions.

2. 1 - Présentation du domaine

2.1.1 L'acquisition des connaissances

2.1.1.1 Motivations : les systèmes à base de connaissances

L'acquisition des connaissances se pose comme un des domaines de recherches associés au développement de logiciels à base de connaissances, les systèmes experts, depuis les années 1980. En effet, les systèmes experts ont été à cette époque une des facettes les plus visibles de l'IA auprès des industriels et du grand public. Or leur diffusion a autant souligné leurs atouts et intérêts que la difficulté de leur conception. Un des points forts des systèmes experts mis en avant par les chercheurs était la séparation entre une représentation déclarative de connaissances sous forme de règles de production d'une part, et un moteur d'inférence s'appuyant sur la logique pour raisonner à l'aide de ces règles d'autre part. Les règles de production ont souvent été présentées comme un formalisme lisible et facilitant la mise au point. Or les expériences des premiers systèmes tels Mycin se sont vite heurtées aux limites du langage de règle lors des phases de mise au point et de maintenance. Devant une grande quantité de règles, il est difficile de comprendre comment le système produit un résultat, quelles règles sont mises et jeu ou comment les corriger pour éliminer d'éventuelles erreurs. Dès 1976, Davis propose un module associé à Mycin dédié à la maintenance de la base de règles, qui évite d'aller lire les règles et favorise l'ajout ou la correction de connaissances par un dialogue avec l'expert. Ce module, TEIRESIAS (Davis, 1979), fournit des explications sur les déductions de Mycin et propose d'en corriger les erreurs en guidant le repérage des règles erronées. En cela, il est le premier système d'acquisition de connaissances.

À partir de cette première proposition, de nombreux travaux se sont engagés pour définir des méthodes et des outils assistant le cogniticien chargé de mettre au point une base de connaissances. Les difficultés abordées étaient tout d'abord de localiser l'expertise, de savoir comme « l'extraire » et ensuite sous quelle forme la représenter. Cette vision du problème laisse penser que les connaissances sont accessibles, gisent chez un ou plusieurs experts, et qu'il suffirait de trouver la bonne manière de les expliciter pour construire un système produisant les mêmes raisonnements. Cette vue, que l'on juge depuis un peu naïve, a eu le mérite de poser les bases de l'acquisition des connaissances comme champ disciplinaire : il s'agit *de s'intéresser aux connaissances pour elles-mêmes avant de considérer leur formalisation*, et de savoir en quoi le système à construire à partir de ces connaissances traite bien les problèmes qu'il doit résoudre.

Un autre facteur d'influence de cette orientation correspond à l'analyse en couches des systèmes informatiques par Newell (Newell, 1982). Celui-ci considère un système intelligent comme un agent rationnel, qui dispose de connaissances et sait effectuer des actions pour atteindre des buts. Il est rationnel dans la mesure où il choisit, avec ses connaissances, l'action suivante qui va le mener le plus directement au but. Newell considère comme une avancée significative la proposition de l'IA de décrire le fonctionnement d'un système au niveau symbolique, où sont explicitement représentés les connaissances et raisonnements. Il propose d'aller plus loin en différenciant connaissances et représentations, et en décrivant le comportement du système à l'aide de connaissances, indépendamment de leur formalisation, ce qui correspond au « niveau des connaissances ». Cette idée a été reprise largement en ingénierie des connaissances pour situer les modèles conceptuels à ce niveau des connaissances, comme des descriptions d'agents rationnels

indépendantes de la manière dont le système sera rendu opérationnel. Cette même proposition de Newell a également justifié l'organisation des modèles conceptuels en deux parties, les connaissances du domaine étant différenciées des actions et des buts.

2.1.1.2 Premières définitions

Je reprends ici une définition présentée en 1992 dans l'introduction d'un numéro spécial de revue sur l'acquisition des connaissances, qui était le premier effort de présentation de travaux de recherche français en la matière (Aussenac *et al.,* 1992). L'acquisition des connaissances y est définie comme le domaine « *chargé d'identifier et d'agencer les tâches requises pour élaborer un système à base de connaissances à partir de sources hétérogènes, et donc de fournir au système les connaissances qui seront à la base de ses compétences. Le cogniticien est la personne chargée d'orchestrer l'intervention des différents acteurs et la mise en œuvre des processus*. » Cette définition est complétée d'une analyse, qui souligne quatre points spécifiques de l'acquisition des connaissances, toujours d'actualité, au-delà de la problématique des systèmes experts :

- le caractère construit, incrémental et itératif du *processus de modélisation*, qui donne lieu à la définition d'un nouvel agencement de tâches pour chaque nouvelle application ;
- le rôle central d'une représentation des connaissances qui soit propre au processus d'acquisition, *le modèle conceptuel,* dont les caractéristiques sont de favoriser l'abstraction et la structuration progressive des connaissances, leur interprétation, leur évaluation et leur révision ;
- la nécessité *d'interactions* entre humains ainsi qu'entre humains et logiciel autour de ce modèle, et la place primordiale du cogniticien pour gérer la diversité tout en conservant comme objectif prioritaire la réponse aux besoins des utilisateurs ;
- la *complexité* du processus de l'IC, dans la mesure où les problèmes sont abordés en vraie grandeur, une des dimensions les plus fortes de cette complexité étant la maîtrise de la diversité : diversité des types de connaissance, des techniques et processus permettant d'en rendre compte, des besoins auxquels on peut chercher à répondre, des types de solution informatique envisageable pour y répondre, etc.

Ces aspects fondamentaux de l'ingénierie des connaissances ont été déclinés selon diverses perspectives au cours du temps, dans la mesure où le statut et la nature des modèles, les acteurs concernés ainsi que la manière de définir un système d'aide à l'utilisateur ont évolué et se sont largement diversifiés. Je reviendrai sur les périodes et les points de vue retenus en ingénierie des connaissances dans la partie 2.3 de ce chapitre.

2.1.2 Les applications à base de connaissances

L'évolution de l'ingénierie des connaissances (IC) est marquée par des périodes assez distinctes, qui se justifient par des influences de deux origines : des discussions et avancées au sein de la discipline d'une part, des interrogations plus larges au sein de l'IA sur la nature des applications visées d'autre part. Les débats au sein de l'IC portent sur la manière dont ces systèmes sont mis au point et dont les connaissances sur lesquelles ils s'appuient sont rendues opérationnelles. Parmi les questions soulevées, il s'agit de savoir si ces connaissances doivent (ou non) tout d'abord être mises à plat dans des modèles, si ces modèles se retrouvent ou non dans le système opérationnel, et plus important encore, de déterminer le statut de ces connaissances, ce qui fait leur validité et leur pertinence, etc. Or les choix et réponses à ces questions sont étroitement liés à la position de l'informatique et de l'IA sur ce qu'est un système intelligent mis à disposition d'un utilisateur et donc un système dont la mise au point relèverait de l'IC.

2.1.2.1 Éléments d'évolution des applications à base de connaissances

Au cours des trente dernières années, plusieurs caractéristiques de ces systèmes ont évolué : la nature des connaissances qu'ils contiennent, leur rôle et la part de la résolution de problème qu'ils prennent en charge, les interactions qu'ils permettent avec leurs utilisateurs, leurs capacités à s'adapter à différents types et contextes d'usage, ou encore leur place dans l'organisation. Finalement, c'est l'idée même de ce qu'est un « système intelligent » qui évolue : initialement localisée dans le seul système expert, « l'intelligence » se déplace vers le couple système-opérateur. La transition s'est opérée progressivement, les différents types d'applications et de points de vue continuant d'être développés et de faire l'objet de recherches. Cependant, on peut situer des périodes liées à l'introduction de chacun des nouveaux points de vue.

Si l'on se réfère aux applications, l'évolution correspond au passage des *systèmes experts*, mis en avant entre 1970 et 1985, aux *systèmes à base de connaissances* autour des années 1990, et depuis aux *systèmes à base de connaissances coopératifs* (SBCC) ou aux *systèmes d'aide interactifs* tels que les envisage l'approche « human centered design » : des supports intelligents aux activités humaines, éventuellement à base d'agents ou distribués, intégrant des raisonnements sur les connaissances à des tâches d'informatique plus classique. J'analyse, pour ces quatre types de système, leur situation par rapport aux caractéristiques que je viens d'énumérer. J'utiliserai plus tard cette typologie afin de définir les périodes et les approches jalonnant l'évolution de l'ingénierie des connaissances.

- *Les systèmes experts* : Les connaissances de référence sont ici celles d'un individu, l'expert. Techniques et spécialisées, elles correspondent à des savoir-faire de haut niveau, rarement verbalisés, qu'il s'agit souvent de pérenniser et de transmettre via le système informatique. D'ailleurs, ces types de connaissance ont été étudiés du point de vue de la psychologie cognitive, c'est-à-dire comme des représentations construites et utilisées par les individus experts à transposer fidèlement dans des représentations formelles. Le rôle du système est de résoudre automatiquement des problèmes que seul l'expert sait traiter, parfois d'expliquer les raisonnements qu'il a suivis, et, par effet de bord, de former les utilisateurs à cette expertise. La place du système et de son utilisateur au sein de l'organisation est rarement prise en compte pour la mise au point des premiers systèmes experts, pas plus que les capacités d'adaptation à l'utilisateur. Cependant, les besoins en maintenance ont donné lieu très tôt à des modules d'acquisition de connaissances, comme je l'ai développé en 2.1.1.
- *Les systèmes à base de connaissances* : Le changement majeur qui justifie ce nouveau type d'application est la prise de distance par rapport à l'expertise humaine. Les connaissances de référence peuvent être partagées par des spécialistes. Il s'agit de diffuser et de faire prendre en charge par le système des parties de procédures jusque-là mal identifiées ou peu disponibles. Cette évolution touche plusieurs facettes du système : son rôle est désormais d'aider l'utilisateur en interagissant éventuellement avec lui ; ses méthodes de résolution de problème privilégient l'efficacité et non la fidélité au raisonnement humain, et peuvent faire appel à des méthodes propres à l'IA. Affiner ces méthodes, les représenter et favoriser leur réutilisation sont alors les enjeux de la définition de ces systèmes. Le système devient alors un des éléments à intégrer dans l'environnement de travail de l'utilisateur, qui commence à être pris en compte dans les processus de conception, par des approches de type ergonomique.
- *Les systèmes à base de connaissances coopératifs* (SBCC) : Un pas plus avant vers la prise en compte de l'environnement et des utilisateurs est de définir des architectures et des modes de résolution favorisant une gestion coopérative de la réalisation des tâches. Les SBCC doivent s'adapter de manière dynamique aux divers contextes de leur utilisation, aux profils des utilisateurs et ceci même en cours de fonctionnement. Les connaissances sur l'environnement de travail, son organisation et les activités concernées par le système sont alors étudiées en complément des connaissances de résolution de problème pour définir la

manière dont le système et l'utilisateur vont coopérer, prendront en charge certaines tâches ou auront l'initiative de la poursuite du processus. L'introduction du système dans la situation de travail comporte donc une dimension sociologique.

- *Les systèmes interactifs d'aide à la réalisation de tâches* : Je désigne ainsi les applications actuelles concernant l'ingénierie des connaissances. Du côté des applications de l'IA, la réussite de la communication est considérée comme un enjeu aussi important que la qualité de la résolution de problèmes, les algorithmes ou les connaissances utilisées par les systèmes. Du côté des systèmes d'information (recherche d'information, applications du web, gestion des connaissances ou gestion documentaire), l'intégration de connaissances des domaines étudiés est de plus en plus fréquente, que ce soit visible ou non pour l'utilisateur. Les modules à base de connaissances constituent soit le cœur de l'application, soit une valeur ajoutée aux compétences de l'utilisateur dans la réalisation de sa tâche. Ils font appel à des savoirs techniques, consensuels et partagés, ou à des connaissances de sens commun. Le spectre des applications concernées est donc large et la nature des problèmes soulevés par leur conception justifie la diversité de travaux actuels en IC.

2.1.2.2 Élargissement de la notion d'application à base de connaissances

Paradoxalement, sur cette notion, la position classique de l'IA, parce qu'elle fait référence à l'objectif de parvenir à définir des systèmes dont les capacités de traitement puissent être qualifiées « d'intelligentes », reste stable : un système intelligent doit être capable de raisonner à partir de faits qui lui ont été fournis, qu'il a perçus ou qu'il a déduits d'une analyse du langage et de ses propres connaissances pour produire des résultats ou de nouvelles connaissances. La plupart des applications s'appuient sur une représentation logique des connaissances, qui constitue un modèle formel, et sur des capacités d'inférence logique. Ces systèmes se heurtent à des limites comme la difficulté de maintenance, la faible capacité à résoudre des problèmes nouveaux, à s'adapter aux évolutions du contexte et des utilisateurs, à apprendre de nouvelles connaissances, etc. Cependant, cette approche reste considérée comme le cœur de l'IA tant elle soulève de questionnements riches sur la complexité des raisonnements.

De ce fait, diverses remises en questions (comme celle de Clancey, 1993) ont débouché sur des approches alternatives qui renouvellent la notion de système d'IA. Ainsi ont vu le jour, entre autres, des recherches situant l'enjeu autour du couple système-utilisateur (*Human-centered design*), de l'interaction et de la coopération homme-machine (*systèmes coopératifs, problématique de l'interaction homme-système*). De nouvelles communautés ont travaillé sur la prise en charge de la réalisation de tâches ou de résolution de problèmes par des agents, sur l'émergence de nouvelles capacités au sein de communautés d'agents réactifs (*systèmes multi-agents, systèmes adaptatifs*). Plus récemment, la nature des applications intégrant des modules à base de connaissances ou faisant appel à des modèles s'est élargie à la gestion documentaire, à la recherche d'informations, aux applications du web sémantique ou encore à des aides à la réalisation de tâches intégrant des règles métier.

Ces alternatives font l'hypothèse que l'intelligence est dans l'interaction entre le système et son utilisateur. Le système ne raisonne plus à la place des individus, mais il doit donner à penser, s'intégrer dans les activités en servant de médiateur. Pour cela, il doit aussi pouvoir apprendre, évoluer ou intégrer de nouvelles données au lieu de reposer sur un ensemble figé de règles souvent vite dépassées.

L'ingénierie des connaissances constitue ainsi un domaine à la fois acteur de cette dynamique de diversification de l'IA, et devant s'y adapter, puisqu'il considère comme objets d'étude tout type de système faisant appel à des connaissances. Cette diversification a une double conséquence : la révision régulière des contours visibles de l'IC d'une part, la nécessité de collaborer étroitement avec les communautés de recherche considérées par ces applications ou approches nouvelles d'autre part.

2.1.2.3 Impact sur l'évolution de l'ingénierie des connaissances

Ainsi, l'ingénierie des connaissances relève de ce courant où, depuis 15 ans, la part de la technique n'a pas augmenté, au sens où l'humain tient une place de plus en plus importante. Très rapidement, en cohérence avec une analyse plus ergonomique et organisationnelle, l'IC a pris en compte le groupe d'acteurs concernés par la mise en place d'un système d'information, et pas seulement l'expert humain qui serait la référence en matière de « bonne résolution du problème ». L'implication des utilisateurs est reconnue comme indispensable pour que le système soit utilisé. Et surtout, la localisation de « l'intelligence » s'est déplacée. Alors que l'on ne mettait initialement l'accent que sur la base de connaissances et les capacités de raisonnement, donc sur la machine seule, il est clair maintenant que l'enjeu se situe tout autant dans les deux composantes du couple homme-machine et dans une troisième, l'interaction humain-machine. Finalement, et c'est une troisième facette de cette évolution, le problème lui-même traité par l'IC n'est plus tout à fait le même : *il traite moins de la formalisation du raisonnement et plus de la mise à disposition de connaissances (dans un système opérationnel et formel) comme un support à un raisonnement ou une activité.* Cette aide peut consister à réaliser une partie de la tâche, à guider l'utilisateur ou lui donner des connaissances à interpréter et rendre opératoire dans le contexte où il agit. En tout cas, on semble avoir renoncé à la logique comme seule solution au problème de la mise à disposition de connaissances opératoires auprès de l'utilisateur.

Cette analyse n'est pas propre à l'IC. Elle traduit un questionnement qui traverse tout un courant de l'informatique et place la réponse au besoin des utilisateurs au cœur de la problématique de conception des systèmes. Les domaines concernés correspondent aux recherches sur l'interaction homme-machine, les systèmes coopératifs ou encore à l'ingénierie des besoins. Bien sûr, ces recherches ne relèvent pas que de l'informatique et font appel à des disciplines comme l'ergonomie, la sociologie ou l'organisation des entreprises.

2.1.2.4 Systèmes à base de connaissances et ingénierie des connaissances aujourd'hui

Aujourd'hui, l'ingénierie des connaissances s'intéresse à la mise en place de systèmes informatiques s'intégrant dans des tâches humaines faisant appel à des connaissances spécialisées. Ces systèmes assistent leurs utilisateurs de manière individuelle ou collective, soit en effectuant des processus jusqu'ici non informatisés, soit en présentation des éléments d'information jusqu'ici non disponibles sous une forme utile à la réalisation de leur tâche ou à la prise de décisions. Les tâches à assister sont donc a priori complexes parce qu'on ne maîtrise pas complètement (de façon algorithmique) la manière de les traiter, parce que l'individu qui les réalise fait appel aussi à ses propres connaissances. L'introduction du système a pour but que le couple opérateur-système réalise conjointement la tâche mieux, plus rapidement ou plus efficacement que l'opérateur seul. Le système doit donc aussi maîtriser ces connaissances ou au moins la restitution, la mise à disposition ou l'opérationnalisation. Ces connaissances peuvent porter sur la capacité à résoudre le problème, sur des savoir-faire techniques ou opératoires, sur la description d'un domaine de manière encyclopédique ou finalisée.

Ce cadre large vise des applications qui vont de la gestion des connaissances (mémoire de projet par exemple) à l'aide au diagnostic, en passant par la recherche d'information ou la formation à distance. Les problématiques traitées couvrent également une grande diversité de questions, d'ordre méthodologiques, relatives aux formalismes et représentations des connaissances, ou encore des aspects technologiques dans la définition de logiciels d'aide et de plate-formes d'ingénierie. Cette diversité n'enlève rien à la lisibilité du domaine. Concernent l'ingénierie des connaissances toutes les questions relatives au repérage, à la structuration et mise en forme opératoire des éléments nécessaires pour assurer l'adéquation d'un système informatique d'aide à la réalisation d'une tâche aux besoins de ses utilisateurs.

L'ingénierie des connaissances est définie aujourd'hui comme le domaine qui « *s'intéresse à la conception de systèmes qui visent à traiter ou à aider des opérateurs à traiter des problèmes mal*

posés, faisant appel à des connaissances jusqu'ici non explicitées ou non modélisées. » (Charlet, 2002). Pour élaborer ces systèmes, l'IC propose de construire des modèles qui vont guider la définition d'une application informatique dont l'utilisation donne lieu à une interprétation en termes de connaissances par des utilisateurs. La discipline définit donc des méthodes, techniques, logiciels et formalismes pour organiser des connaissances au sein de modèles qui permettent de les restituer dans un environnement opérationnel. C'est tout le processus qui va de l'expression de besoins à la mise à disposition du système opérationnel qui intéresse l'IC.

2.1.3 Place vis-à-vis de l'informatique et de l'intelligence artificielle

Historiquement liée à l'intelligence artificielle (IA), l'ingénierie des connaissances entretient avec cette discipline des rapports continus : complémentarité des approches et questions, échange de résultats et de problématiques, mais aussi questionnement réciproque sur la place de chacune dans la mise au point de systèmes « intelligents ». Les recherches en acquisition des connaissances ont ainsi d'abord contribué à l'étude de la mise au point, de la validation, de la production d'explications pour les systèmes à base de connaissances et les systèmes coopératifs.

Plus largement, en s'intéressant à des problèmes réels, tels qu'ils se posent et non tels qu'ils se formalisent, en s'appuyant sur des expériences de terrain, en s'adressant aux acteurs des domaines visés, l'ingénierie des connaissances aborde, au-delà d'enjeux techniques, des problèmes scientifiques dans toute leur complexité. Son originalité est de s'intéresser aux difficultés liées à la réalité des applications. L'ingénierie des connaissances aborde des questions fondamentales sur la mise à disposition de connaissances dans des systèmes informatiques sous un angle différent des études plus théoriques de l'IA et en collaboration avec d'autres disciplines, en offrant des réponses originales. Par exemple, elle s'intéresse à la représentation des connaissances d'un point de vue pratique. Elle reprend pour cela des travaux formels pour les transposer dans des contextes opérationnels, comme les logiques de description pour la formalisation des ontologies. En cela, ses résultats complètent des approches théoriques, validées mathématiquement sur des exemples restreints, plus habituelles en IA.

Finalement, l'ingénierie des connaissances est un des champs issus de l'IA qui s'intéresse à l'ancrage dans un environnement humain des connaissances à partir desquelles sont construits des modèles : quelles représentations construire à partir d'une réalité donnée pour développer un système produisant le raisonnement attendu ou fournissant à l'utilisateur les éléments pertinents pour qu'il raisonne ? Cette question théorique interroge la nature des représentations, leur distance par rapport au réel, la façon dont elles le dénotent, mais aussi leur interprétation, formelle et humaine. Elle a évolué du mimétisme des processus cognitifs de l'expert (qui reste encore d'actualité pour des études à visée cognitive) vers la construction de systèmes dont les modes de raisonnement sont propres au système artificiel. Ce choix alimente le débat classique en IA entre anthropomorphisme et artificialisme.

Parce qu'elle s'intéresse à la mise au point de logiciels d'aide à l'utilisateur, la problématique de l'ingénierie des connaissances est très proche de celle du génie logiciel. En amont, elle traite de problèmes relevant de l'ingénierie des besoins. Recueillir et mettre en forme des besoins d'utilisateurs nécessite de recueillir des connaissances. De plus, la mise en forme de l'expression des besoins et des spécifications a beaucoup de points communs avec le processus de modélisation de l'IC. Les techniques et méthodes de travail de l'IC avec les experts et les utilisateurs peuvent apporter une réponse aux analystes, souvent démunis face aux utilisateurs.

Ensuite, ingénierie des connaissances et génie logiciel ont en commun la mise au point de modèles conceptuels. Ces modèles ont clairement le statut de spécification du système opérationnel en génie logiciel, alors qu'en ingénierie des connaissances, la distance entre le modèle et le système opérationnel varie d'une approche à l'autre. Le cogniticien définit son application en travaillant au niveau des modèles et non au niveau des programmes. Je considère que la priorité pour un modèle est d'être d'abord lisible par tous les acteurs concernés par le futur système.

Enfin, tout comme le génie logiciel vise la réutilisabilité de composants logiciels, l'acquisition des connaissances cherche à favoriser la réutilisation de composants de connaissances. Ce sont soit des modèles de domaine ou ontologies à partir desquels on peut développer plusieurs applications, soit des méthodes de résolution de problème applicables à différents domaines.

Les propositions de l'IC rencontrent également les recherches sur les interactions homme-machine. En effet, la modélisation apporte une solution aux problèmes de gestion de l'interaction entre opérateur et système intelligent. La formulation des capacités de résolution de problème du système au niveau conceptuel favorise une meilleure analyse de son comportement et une meilleure définition de ce que peut être la coopération avec ce système ainsi que sa mise à jour. La spécification de l'interaction homme-système au niveau conceptuel sous forme de modèle tâche-méthode est d'ailleurs au cœur de nombreuses recherches sur la définition de systèmes interactifs.

En matière d'applications informatiques, l'IC est une des disciplines pouvant contribuer aux questions de gestion des connaissances dans les entreprises. À côté de la gestion ou de la sociologie d'entreprise, concernées par les aspects organisationnels et humains du problème, l'informatique concernée par les aspects techniques de mise en réseau, de sauvegarde ou d'exploitation de données (fouille de données, gestion de réseaux, …), il s'agit d'identifier les connaissances en jeu, leur circulation et la manière de les mettre à disposition. Enfin, les évolutions récentes vers la prise en compte de connaissances, de la sémantique dans la gestion des connaissances documentaires font se rencontrer ingénierie des connaissances et recherche d'information ou gestion documentaire.

2.1.4 L'IC, carrefour disciplinaire

Même si les solutions techniques envisagées relèvent de l'informatique, l'IC est, par essence, un carrefour disciplinaire. En effet, tout d'abord, les questions des connaissances, de leur modélisation ou leur opérationnalisation, dépassent le champ de l'informatique. Elles nécessitent de prendre en compte les analyses d'autres disciplines, comme la psychologie cognitive, l'ergonomie, les sciences des organisations, la linguistique ou le traitement automatique des langues (TAL). Ensuite, en tant qu'ingénierie s'intéressant à des situations réelles, l'IC doit faire référence à des concepts, points de vue et résultats d'autres disciplines pour aborder les facettes de cette réalité. Je développe ces deux aspects fondamentaux des recherches en IC dans les deux parties suivantes avant de discuter de la nature de ces échanges disciplinaires dans une 3e partie.

2.1.4.1 Place des connaissances

Depuis qu'elle se démarque d'une visée de simulation cognitive et met l'accent sur l'efficience des systèmes à construire, l'IC ne cherche plus à élaborer une nouvelle définition des connaissances. Elle accorde le statut de connaissances, parfois de manière abusive, à des objets d'étude très divers qui en sont les traces et manifestations ou des représentations informatiques. En entrée du processus d'ingénierie, ce sont d'une part des besoins en connaissances qui sont exprimés en lien avec la réalisation d'activités, individuelles ou collectives, et d'autre part des traces de connaissances qui sont observables (activités, textes, expressions d'experts). En sortie de ce processus, ce sont des matérialisations informatiques (Bachimont parle *d'inscriptions*) qui doivent soit permettre au système d'effectuer des traitements adaptés, soit les rendre accessibles à l'utilisateur en lui permettant de se les approprier et de les mettre en œuvre à son tour.

Les difficultés propres à l'ingénierie des connaissances se situent donc à la fois dans la spécification de réponses pertinentes à des besoins et dans l'identification puis l'explicitation sous une forme observable de connaissances requises pour élaborer ces spécifications. Or ces réponses sont des systèmes opérationnels manipulant des représentations informatiques qui ne sont plus systématiquement des modèles logiques de connaissances au sens de l'IA. Les modèles conceptuels de l'IC ne prétendent plus rendre compte de connaissances ni au sens cognitif ni au sens logique. Ils

sont seulement des représentations intermédiaires définies pour guider l'explicitation de connaissances, d'une part, et la mise au point de représentations informatiques, d'autre part.

L'évolution historique des systèmes informatiques dits « à base de connaissances » et de la nature des modèles conceptuels correspond à l'identification progressive des différentes dimensions des connaissances à prendre en compte lors de la réalisation de ces systèmes. Cet élargissement de point de vue porte par exemple sur les détenteurs des connaissances ; sur leur caractère situé et lié à une activité ou au contraire encyclopédique ou livresque ; sur leur caractère implicite ou explicite, accessible ou non. Pratiquement, l'IC fait référence aujourd'hui à quatre types de source de connaissance, parfois de manière complémentaire au sein d'un même projet, et s'appuie pour étudier chacun d'eux sur des concepts, techniques et méthodes venant de disciplines s'intéressant à ces types de connaissance :

- Pour étudier les connaissances individuelles, expertes, spécialisées ou même didactiques, elle fait appel à des résultats de la psychologie, cognitive en particulier, comme les techniques expérimentales d'entretiens et de validation cognitive de modèles ;
- L'analyse des pratiques, des activités et des usages individuels renvoie à des approches ergonomiques ; l'IC les reprend pour spécifier les contenus de modèles conceptuels, l'interaction et la coopération homme-système, ou la validation en usage de ces modèles.
- La prise en compte des organisations et des collectifs fait référence à la sociologie et aux sciences des organisations ; elle conduit l'IC à proposer des supports aux échanges et aux utilisations collectives de connaissances, et à penser des solutions au niveau organisationnel ;
- Enfin, l'étude des connaissances à travers leur expression dans le langage et les textes s'appuie sur des travaux en linguistique de corpus, TAL, terminologie ou sciences de l'information.

À titre d'exemple de ce type d'emprunt à d'autres disciplines, plusieurs typologies de connaissances sont référencées dans de nombreux travaux d'IC. Parmi elles, la typologie en niveaux de Rasmussen, venant de la psychologie, situe les savoir-faire et les heuristiques par rapport aux autres connaissances requises pour la résolution de problèmes par un individu (Rasmussen, 1985). La distinction entre connaissances tacites (implicites) et connaissances explicites, classique en psychologie cognitive (Leplat, 1986), est transposée au niveau des connaissances de l'entreprise par Nonaka et Takeuchi et croisée à un autre axe individuel / collectif (Nonaka et Takeuchi, 1997). Dans ce dernier cas, les connaissances sont caractérisées par leur potentiel d'action et par les effets de leur mise en œuvre. Ces caractérisations permettent de mieux comprendre les processus de diffusion, circulation et transmission de connaissances, et soulignent leur caractère dynamique et évolutif. Elles contribuent à l'étude de la gestion des connaissances.

Plus précisément, ces contributions permettent aussi de mesurer les enjeux de la réalisation d'un système pour un groupe d'utilisateurs et une situation donnée. En soulignant la diversité des types de connaissance humaine mis en jeu, elles appellent à envisager une solution non pas seulement de son seul point de vue technique, mais aussi sous forme de nouvelles organisations. Elles invitent à diversifier les solutions techniques au sein du système final ainsi que les types d'aide ou d'interaction fournies par ces systèmes. En amont, et cela concerne l'ingénierie des connaissances, ces différents points de vue sur les connaissances ont été pris en compte alternativement pour définir des méthodes d'analyse et de modélisation.

2.1.4.2 Quand un domaine de recherche est une ingénierie …

Je n'entrerai pas dans le débat hasardeux de savoir si l'ingénierie des connaissances est ou non une science. J'ai parlé jusqu'ici de domaine scientifique, car il me semble primordial de l'identifier comme un domaine ayant des objectifs, des concepts et des méthodes propres. En revanche, je souhaite développer en quoi sa problématique relève bien d'une d'ingénierie. Tout

d'abord, je m'appuie sur l'analyse de B. Bachimont situant une ingénierie par rapport à une technique ou une science. Ensuite, je souligne que le domaine est (et devrait être plus encore) alimenté à la fois par les retours d'expérience de praticiens et par des recherches plus théoriques.

Science, technique et ingénierie

Dans son habilitation, B. Bachimont situe la notion d'ingénierie par rapport à celles de *science, technique et ingénierie* dans différentes analyses en philosophie pour placer l'ingénierie des connaissances en tant que discipline (Bachimont, 2004).

Il rapporte d'abord ce qu'est une ingénierie chez les philosophes « modernes ». « L'ingénierie s'entend comme l'application à un problème concret de solutions trouvées à partir d'une démarche scientifique ». L'ingénieur sera d'autant plus efficace qu'il maîtrisera les connaissances scientifiques. En effet, la science permet de modéliser de manière exacte les processus pour les reproduire techniquement et en faire des procédés techniques autonomes.

Aujourd'hui, pour les philosophes contemporains, le croisement étroit entre technique et nature fait que les objets techniques peuvent être observés et étudiés comme des objets naturels, et inversement les objets naturels peuvent être modifiés par une intervention technique. Cette fusion est possible parce qu'à chaque niveau de la nature, on dispose d'un moyen technique de l'observer. Plus que cela, une autre particularité des techno-sciences est de modifier la nature des événements à venir, de les influencer ou de les réorienter. Cette influence s'illustre par exemple par la difficulté à anticiper l'activité d'un opérateur alors qu'on conçoit un système informatique s'intégrant dans cette activité.

Pour B. Bachimont, la technique aussi permet la construction de cadres d'analyse scientifique, y compris pratiques. En cela, elle déplace les questions qu'elle est supposé résoudre, car elle modifie les termes qui permettent de la poser. Elle génère encore plus de technique, pour mieux arraisonner ce devenir qui n'arrête pas de lui échapper. L'IC, et plus encore toute l'ergonomie de conception, n'échappe pas à ce phénomène, et même elle l'illustre complètement.

B. Bachimont considère que l'IC a pour objet l'inscription numérique des connaissances, que c'est une ingénierie des supports techniques des connaissances. L'IC propose des solutions technologiques pour définir ces supports, élaborer des dispositifs techniques et les alimenter, ainsi qu'une critique portant sur leur mobilisation et sur leur interprétation comme connaissance par les utilisateurs. La dimension pluridisciplinaire vient alors de ce que l'élaboration des dispositifs techniques fait appel aux sciences physiques et à l'informatique alors que la critique des connaissances inscrites dans ces supports relève des sciences humaines.

IC, science de l'ingénieur : place et capitalisation des savoir-faire

Parce qu'elle est une ingénierie, l'IC s'apparente aux sciences de l'ingénieur, qui sont souvent classées comme « sciences empiriques », au sens où elles s'appuient surtout sur l'expérience pour élaborer de nouvelles connaissances. Les savoir sont à la fois des savoirs techniques, relevant de la maîtrise de langages, de formalismes, de méthodes et de principes, et aussi des savoir-faire, qui s'appuient sur une maîtrise de la mise en œuvre de ces techniques, du choix de ces langages ou formalismes dans des contextes spécifiques. Le savoir-faire se manifeste dans la capacité à adapter des résultats plus théoriques, des propositions techniques et des solutions technologiques à de nouveaux cadres applicatifs selon leurs caractéristiques.

De plus, les problèmes étudiés par l'ingénierie des connaissances sont ouverts, c'est-à-dire que les éléments à prendre en compte ne sont pas connus à l'avance, au début d'un projet, mais peuvent se révéler au fur et à mesure de son avancement. En particulier, l'expression des besoins des utilisateurs s'affine en cours de projet, la nature des connaissances à structurer se dévoile au fur et à mesure de l'observation de leur mise en œuvre, de leur expression par les experts ou de ce qui peut en être perçu par des écrits. La manière dont le système opérationnel va les intégrer et les utiliser pour les rendre opératoires mérite d'être ajustée en fonction de ces deux paramètres, tâches

et besoins des utilisateurs et nature des connaissances. L'IC ne peut donc que proposer des cadres relativement généraux, des familles de techniques de recueil, structuration ou formalisation, des principes et niveaux d'analyse des problèmes, et non des solutions opératoires précises, des formules directement applicables ou des outils prédéfinis. L'expertise, les compétences et la formation du cogniticien contribuent tout autant au succès des projets. Les retours d'expérience et leur mise en forme sont autant de résultats à capitaliser.

Ainsi, la capitalisation des connaissances en IC passe à la fois par la mise en forme de résultats s'appuyant parfois sur des travaux théoriques, et le plus souvent sur des propositions attestées par leur mise en œuvre expérimentale. De ce fait, la difficulté de cette capitalisation tient entre autres à la lourdeur, à la complexité et à la durée des expériences à mener avant de pouvoir en tirer de nouvelles méthodes et de nouvelles approches validées.

Cependant, les contributions dans le domaine ne se réduisent pas à des travaux « académiques ». Par exemple, l'engouement pour les systèmes experts a donné lieu à quantité de projets et de développements en entreprise. Les ingénieurs chargés de ce travail se sont rapidement heurtés à la difficulté de l'acquisition des connaissances. Leurs retours d'expérience ont été rapportés dans des ouvrages proposant des techniques d'entretien organisées au sein de méthodes (Berry, 1988). La recherche universitaire a peu repris ces propositions, qui considèrent l'acquisition comme une extraction de connaissances. Plus tard, le transfert des résultats de recherche comme CommonKADS vers les entreprises a permis de valider la méthode. Je considère comme fondamental que la recherche en IC valide et corrige ses propositions via des utilisations de ses résultats en vraie grandeur, qu'elle se nourrisse d'échanges entre praticiens et chercheurs.

2.1.4.3 Bilan sur les échanges disciplinaires en IC

Afin d'approfondir les enjeux de recherches impliquant plusieurs disciplines autour de la construction d'artefacts informatiques, C. Garbay a proposé une analyse de différentes formes d'échanges et de collaborations entre disciplines (Garbay, 2003). Elle souligne la place charnière des Sciences et Technologies de l'Information et de la Communication (STIC), dont relève l'IC, comme développant des systèmes qui donnent lieu à des bouleversements forts dans le statut des systèmes techniques et de leurs usagers. De nouvelles pratiques émergent qui constituent de nouveaux objets d'étude à définir et à aborder par l'analyse croisée de plusieurs disciplines. Ainsi, elles en remettent en question les frontières, et renouvellent le travail interdisciplinaire. Cette analyse reflète tout à fait les mouvements et la dynamique qui touchent l'ingénierie des connaissances et des disciplines proches (comme l'ergonomie, la psychologie cognitive, la sociologie ou la gestion) concernées par la mise au point de nouveaux systèmes à base de connaissances. Les échanges entre l'ingénierie des connaissances et ces disciplines peuvent suivre deux parcours.

Certaines recherches correspondent à la transposition d'une théorie ou d'un résultat venant d'une discipline vers une autre. Ainsi, la définition des primitives de modélisation en IC s'inspire de représentations proposées en IA (comme les frames ou les graphes conceptuels), l'analyse de l'activité a été empruntée à l'ergonomie, ou encore de techniques d'analyse de textes issues de la linguistique de corpus.

Après s'être nourrie de résultats venant d'autres disciplines, l'évolution du domaine de l'IC passe par des travaux interdisciplinaires. Pour une classe d'applications donnée et pour traiter un des problèmes particuliers de l'IC, ces recherches supposent la mise au point d'une collaboration avec une ou plusieurs disciplines afin de définir des méthodes, outils et supports techniques adaptés à de nouvelles situations. Ces recherches concernent différentes disciplines amenées à définir un questionnement et éventuellement un cadre théorique commun. La définition de solutions aux problèmes abordés par l'IC suppose en effet de s'intéresser aux relations entre humains et systèmes techniques, du point de vue de l'appropriation des systèmes par les usagers, des apprentissages et nouveaux savoirs que cet usage développe, mais aussi de leur construction à partir de savoirs et de besoins identifiés. Le caractère cyclique de ce processus, non seulement à l'échelle individuelle

mais aussi au niveau collectif, souligne la forte dépendance entre l'étude des situations d'interaction du point de vue des personnes, collectifs et organisation impliqués d'une part et du point de vue technique d'autre part. Ainsi, pour définir des méthodes et outils spécifiant la tâche d'un système à base de connaissances, l'IC ne peut se contenter des seules questions de représentation, de langage ou d'architecture informatique. Définir les moyens d'analyse des tâches de futurs utilisateurs au sein d'une organisation nécessite des analyses menées avec des ergonomes ou des spécialistes en gestion des entreprises par exemple. La mise au point de ressources terminologiques et d'ontologies à partir de textes va bien au-delà d'un échange d'outils avec le TAL et de techniques avec la linguistique. Il s'agit, entre autres, de revoir conjointement le statut des concepts et des termes au regard des nouveaux usages prévus par ces ressources, de poser un débat sur le caractère normatif, générique, universel ou non de ces représentations.

L'ingénierie des connaissances engage donc une dynamique qui la pousse à aller à la rencontre d'autres disciplines. Cette « curiosité scientifique » fait de l'IC un carrefour disciplinaire innovant, en renvoie une image positive. Mais le prix à payer en est le manque de lisibilité des contours réels de l'IC, et même de ses contributions propres. Comme le souligne C. Garbay, une réflexion approfondie s'impose entre informatique et sciences humaines qui passe par une mise en réseau des chercheurs, des individus concernés par ces questions, et par des motivations intellectuelles et scientifiques suffisamment fortes et stimulantes pour inviter à ces collaborations. La manière dont l'IC s'est interrogée au cours des quinze dernières années a ainsi permis des échanges forts au niveau national et international d'abord avec des ergonomes et sociologues, puis avec des linguistes et terminologues. Toutefois, une particularité française est d'afficher cette interdisciplinarité avec plus de conviction.

2.1.5 La difficulté de l'évaluation

2.1.5.1 Un manque d'évaluation ?

La question de l'évaluation en ingénierie des connaissances est une question complexe, régulièrement abordée car elle conditionne la crédibilité du caractère scientifique des approches proposées. Elle est pourtant un des points faibles de l'IC parce que les chercheurs ont du mal à établir des critères de validité de la plupart des résultats qu'ils proposent. L'absence de métrique, de standards d'évaluation donne l'impression d'une mauvaise maîtrise de la portée et de la qualité de ces résultats : en quoi un modèle, une méthode ou une technique peut-il s'appliquer dans un autre contexte d'application, à un autre type de tâche ou de domaine ? en quoi un résultat constitue-t-il une réelle avancée dans le domaine ?

S'interroger sur l'évaluation revient à préciser les objets, méthodes et critères d'évaluation :

- La nature des résultats à valider : faut-il évaluer les méthodes, techniques et outils proposés par l'IC par eux-mêmes ? ou les systèmes résultants de la mise en œuvre de ces moyens méthodologiques ? ou encore l'apport de ces systèmes auprès de leurs utilisateurs, la qualité de la réponse au besoin ?
- La pertinence d'un résultat : s'agit-il de la réponse à un besoin, l'adéquation à une théorie de la connaissance, l'utilisation effective d'un logiciel par un cogniticien ou un gain réel dans l'activité assistée par le logiciel développé ?
- Les méthodes d'évaluation : Que peut-on vraiment mesurer à partir d'applications en vraie grandeur ? comment mener des expérimentations, des études de cas représentatives ? des évaluations de résultats intermédiaires, de techniques particulières ou d'outils supposent de bien délimiter leur portée. En quoi ces évaluations partielles garantissent-elles la validité de l'ensemble ?

Ces questions se croisent. Il paraît indispensable de parvenir au final à évaluer des résultats globaux dans des situations réelles d'usage, c'est-à-dire la qualité du système développé à l'issue

du processus d'ingénierie des connaissances. Or ce type d'évaluation globale se heurte à plusieurs difficultés : le coût, la durée, la multiplicité des paramètres non maîtrisés, le grand nombre de concepts ou d'hypothèses à évaluer. Un compromis consiste à s'appuyer sur des validations plus ciblées, soit en se focalisant sur une partie de l'approche proposée (langages, modèles, techniques, logiciels d'extraction, etc.) soit en s'appuyant sur des expériences mieux maîtrisées comme les études de cas ou les simulations.

2.1.5.2 Quelques cadres d'évaluation

Afin de réduire la complexité du problème et de se donner des éléments de comparaison des résultats produits en IC, la communauté scientifique s'est organisée à partir de 1992. Elle a proposé des études de cas à utiliser lors de campagnes d'évaluation.

- Les premières propositions, les projets Sisyphus I (1992-1994), ont porté sur la modélisation de connaissances de résolution de problème. Un énoncé simple de problème d'affectation (de bureaux aux membres d'un laboratoire) était proposé pour construire un modèle permettant de résoudre ce problème. Aucune méthode de résolution n'était suggérée a priori. La proposition s'est affinée dans le projet Sisyphus II sous forme d'une grille de questions pour permettre de comparer la méthode suivie par chaque participant pour obtenir le modèle, les logiciels utilisés, les choix de modélisation, la représentation des connaissances ou encore la manière d'opérationnaliser le modèle.
- Le deuxième projet Sisyphus III (dit V.T.) (autour de 1995) proposait un problème plus complexe de conception d'un ascenseur par assemblage de pièces à choisir en fonction de contraintes matérielles. Outre le type de tâche, le changement venait de la richesse des informations fournies pour guider la modélisation et l'enrichissement de la grille de comparaison des travaux. Cette expérience a permis d'aborder un nouveau type de problème, la conception.
- Un dernier projet Sisyphus IV (autour de 1996-1997), dans le domaine de la géologie, a porté sur de la modélisation de connaissances à partir de textes. Ce projet était beaucoup plus ambitieux et sa complexité plus proche d'une étude de cas réelle. Le modèle à construire devait répondre à un besoin d'utilisation donné aux participants, la source de connaissance était un ensemble de documents tels qu'ils avaient été étudiés dans un projet.

Ces projets ont favorisé la connaissance réciproque des travaux de recherche. Ils ont permis de mieux comprendre chacune des approches, méthodes et solutions des participants, de dégager des éléments de confrontation, de présenter bien plus précisément que dans des articles classiques les langages de modélisation. Ils ont eu des impacts très bénéfiques au niveau des échanges scientifiques, des réflexions méthodologiques et de l'utilisation des logiciels. Cependant, ils n'ont pas vraiment permis de (et même se sont refusés à) juger de la qualité des modèles obtenus : comment décider qu'un modèle était meilleur, plus valide ou plus performant, qu'un autre ? ils n'apportaient pas davantage de réponse qualitative à la comparaison des méthodes.

La construction d'ontologies a donné un nouvel élan à ce type de projet, car il semble plus facile d'établir des métriques pour juger de leur pertinence, les comparer et confronter des approches pour leur construction. Des protocoles ont été définis pour mieux mesurer les avancées, comparer des travaux et qualifier ou quantifier des résultats, enfin d'en assurer une visibilité à l'extérieur de la communauté concernée.

- Plusieurs conférences EON (Evaluation of Ontology based Tools) ont porté sur l'évaluation d'outils utilisés pour la construction d'ontologies. Une étude de cas très simple proposait une page décrivant les éléments d'un domaine et une application supposée

utiliser l'ontologie. Lors de la première expérience[2], la comparaison a porté sur les choix de modélisation et le langage de représentation des connaissances. Lors du workshop suivant[3], l'évaluation a porté sur l'interopérabilité des modèles et des outils, la capacité des éditeurs à lire des modèles construits par d'autres équipes, et celle des modèles à être réutilisés ou adaptés. Enfin, l'édition de 2004[4] s'intéressait aux méthodes d'alignement d'ontologies. Ces workshops ont dynamisé et fait progresser la communauté de recherche, en favorisant une meilleure connaissance des capacités des méthodes et des logiciels. Ils ont rendu plus visibles ces résultats. Cependant, ils ne permettent pas de dire si une méthode ou une représentation des connaissances est plus adaptée ou plus efficace qu'une autre. La difficulté est propre au domaine, où il n'existe pas un « bon » modèle pour une application et une description de connaissances données.

- Depuis 5 ans, le développement de systèmes pour la construction d'ontologies à partir de textes rapproche l'IC de deux communautés plus familières de campagnes d'évaluations qualitatives et quantitatives : d'une part, la communauté du traitement automatique des langues, qui fournit les outils d'analyse des textes, et d'autre part celle de la recherche d'information qui utilise ces ontologies pour indexer ou classer des documents, améliorer les capacités des moteurs de recherche, etc. Le succès récent de l'utilisation de l'apprentissage et de l'extraction d'information n'a fait que renforcer ce phénomène et rend possible de mesurer plus précisément certains traitements sur les textes et leurs apports à l'enrichissement des ontologies. Actuellement, le réseau d'excellence OntoKnowledge, en lien avec une série de workshops OLP, a mis en place le projet PASCAL qui propose plusieurs tâches liées à l'analyse automatique de textes et à l'enrichissement d'ontologies à partir de textes. Les tâches consistent à acquérir des types de connaissance à partir de textes (termes, concepts, relations, etc.) et à les utiliser pour enrichir ou construire complètement une ontologie dans un domaine particulier.

Le risque est de tomber dans des travers technologiques analogues à ceux auxquels est confrontée la recherche d'information avec les grandes campagnes d'évaluation (TREC, CLE, etc.). Les tâches proposées ne sont pas forcément très représentatives de la problématique effective de construction d'ontologie. Par exemple, la référence pour la validation d'un modèle reste le contenu du texte alors que l'on sait que l'application visée oblige à s'éloigner du contenu du texte. Ces campagnes ne s'intéressent pas vraiment à l'utilisabilité des modèles, à leur pertinence en usage, car mesurer la satisfaction des utilisateurs supposerait des protocoles complexes et lourds, et à des résultats qualitatifs difficilement comparables.

2.1.5.3 Validation versus évaluation des modèles

En ce qui concerne les modèles conceptuels, et plus encore pour les ontologies et terminologies (RTO), la communauté donne un sens particulier à *validation* et *évaluation* [RIA, 04]. Par *validation* d'un modèle, on entend le moment où l'analyste présente le modèle à l'expert, et lui demande de valider ou d'invalider les choix de modélisation effectués. L'enjeu est de s'assurer avec les experts que la conceptualisation représentée n'est pas en contradiction avec leurs connaissances. Une fois le modèle construit, s'engage un processus d'*évaluation* selon les procédures de base du génie logiciel. Il s'agit de vérifier si le modèle satisfait le cahier des charges et répond aux attentes spécifiées au début du projet. Définir un cadre pour ce genre d'évaluation présente deux types de difficulté. Tout d'abord, l'ontologie n'est qu'un élément de l'application cible, qui est le dispositif à évaluer. Il faut donc concevoir des expériences et des bancs d'essais qui permettent de cibler l'évaluation sur la seule ressource. Ensuite, chaque cas étant particulier, les

[2] http://km.aifb.uni-karlsruhe.de/eon2002
[3] http://km.aifb.uni-karlsruhe.de/ws/eon2003
[4] http://km.aifb.uni-karlsruhe.de/ws/eon2004

procédures d'évaluation doivent être adaptées aux types d'application, à la manière dont elles utilisent le modèle, etc.

2.1.5.4 Évaluation des outils

À côté de l'évaluation des modèles, se pose le défi de l'évaluation des outils de construction des modèles. La source des difficultés est double : d'abord il s'agit d'outils d'aide, ensuite chaque outil est rarement utilisé seul. Dans le cas d'un outil automatique, du type « boîte noire », ses performances peuvent être mesurées en comparant ses résultats à des résultats attendus (« gold standard »). Dans le cas des outils d'aide, les résultats fournis sont interprétés par l'analyste. Or les conséquences de cette interprétation sont variables : une modification, un enrichissement du modèle à un ou plusieurs endroits, voire l'absence d'action immédiate, sans que cela signifie nécessairement que les résultats en question soient faux ni même non pertinents. De plus, cette interprétation s'appuie normalement sur une confirmation par retour aux sources de connaissances. Il n'y a pas systématiquement de lien direct entre un résultat de l'outil et une portion de modèle. Si on rajoute à cela, qu'une portion de modèle n'a pas de sens prise isolément, et que la ressource elle-même ne peut être évaluée qu'en contexte, on saisit l'ampleur de la tâche. Il y a un parcours interprétatif considérable entre les résultats de l'outil et la ressource construite. De ce fait, la comparaison des résultats de l'outil et d'une ressource de référence ne peut apporter que des évaluations limitées, même si elle fournit des indications intéressantes pour faire évoluer l'outil (Nazarenko *et al.,* 2001).

L'idéal serait par exemple de comparer deux modèles, l'un construit avec tel outil, et l'autre sans, en termes de temps de réalisation et de qualité. Quand on connaît le temps de développement d'un modèle, on imagine la lourdeur et la difficulté de mise en œuvre d'une telle méthodologie. Le problème reste ouvert. Pour mesurer, ne serait-ce que d'un point de vue qualitatif, l'intérêt des outils, considérons pour le moment qu'il est primordial de les tester dans des contextes nombreux et variés et aussi réels que possible pour faire avancer la recherche.

2.1.5.5 Valeur des expériences en IC

L'accumulation d'expériences bien ciblées, décrites et mesurées semble donc une des seules pistes envisageables pour évaluer les propositions de l'IC. Or B. Bachimont a invité à la prudence dans l'utilisation du terme « expérience », affirmant de manière provocatrice qu' « il n'y a pas d'expérience en IC » (Bachimont, 2004). Il rappelle alors qu'expérience correspond ici à la validation expérimentale d'une théorie, d'hypothèses concernant des lois établies pour décrire une ensemble de phénomènes observés. Or justement, n'étant pas une science mais une ingénierie, l'IC ne cherche pas à décrire ou prédire, à travers les modèles qu'elle élabore, des hypothèses scientifiques sur la connaissance. Ces modèles contribuent à définir le comportement du système informatique visé, et leur validation, empirique, se mesure à leur capacité à rendre le système final pertinent.

Pour lui, les modèles conceptuels définissent des normes qui doivent s'ajuster aux usages et pratiques. Et par l'accumulation d'utilisations et d'adaptations de ce type, se définit un cadre d'applicabilité de chaque modèle ou type de modèle pour des usages futurs. L'IC doit être capable de critiquer chaque nouvelle situation et de dégager des principes d'adaptation aux nouveaux usages à partir de ceux déjà réalisés. Une application, un projet ne valide donc pas une technique ou une méthode, mais augmente la possibilité de les critiquer et de préciser comment les adapter ou les réutiliser dans de nouveaux contextes.

De plus, l'IC se nourrit des concepts d'autres disciplines pour en produire des spécifiques. Pour B. Bachimont, cela impliquerait que les modes de validation correspondent aux méthodes des disciplines d'origine de ces concepts. Ceci expliquerait une partie du malaise des praticiens et chercheurs de l'IC. Soit ils valident séparément et selon des approches d'autres disciplines des

contributions ponctuelles au processus de modélisation. Soit ils mettent en place des expériences portant sur la globalité de leurs propositions, dont il est délicat d'identifier ce qu'elles valident.

Face à cette vue qui me semble réductrice, puisqu'elle donne l'impression d'une faible créativité et d'un manque de repères pour évaluer les propositions faites, j'oppose près de quinze années de recherche dans le domaine. Ces travaux ont produit des résultats propres à l'IC, et pas seulement empruntés, essentiellement sur les représentations et les typologies de modèles conceptuels, leurs composantes ou encore leur mode de construction. Le souci constant de la validation a pris des formes originales pour chacune de ces contributions prises séparément, qui passe souvent par le développement de prototypes, de modèles ou d'applications illustrant la mise en œuvre de ces propositions. Globalement, la particularité du domaine est de rappeler la nécessité d'utiliser les concepts proposés dans des situations réelles, pour contribuer à des projets opérationnels avec leurs contraintes. Ces utilisations ne sont peut être pas des expériences au sens de la physique : il faudrait construire l'application une première fois selon une approche « classique » puis une deuxième fois selon la méthode ou les outils originaux proposés, et mesurer le gain apporté. Toutefois, elles offrent un retour suffisant pour juger de l'intérêt des propositions mises en place, pour estimer leurs apports effectifs ou leurs limites. La confrontation de plusieurs expériences pose la base d'une réflexion sur le statut des modèles, sur ce qui fait leur pertinence, ainsi que sur le rôle et les compétences des personnes concernées, en particulier le cogniticien chargé de mener à bien le processus. Je rejoins ici B. Bachimont : les expériences ne peuvent pas tenir lieu de validation absolue, mais elles orientent des choix et infléchissent des directions.

2. 2 - Analyse des travaux sur les modèles conceptuels

Autour des années 1990, il a été proposé de construire un système à base de connaissances en commençant par la description des connaissances du système indépendamment de leur implémentation. Cette représentation, abstraite et finalisée, devait tenir compte des multiples facettes et types de connaissance utiles pour que le système réponde aux besoins identifiés. Le support pour rendre compte de cette représentation a été appelé modèle conceptuel. Depuis, cette notion est devenue centrale en ingénierie des connaissances, et elle a largement évolué pour recouvrir des réalités différentes suivant les besoins auxquels elle devait répondre, suivant l'évolution des approches et des recherches dans le domaine. Je présente ici cette notion et les différents types de réalité qu'elle recouvre : statut de ces modèles, nature de leur contenu, utilisation qui en est faite, composantes qui les constitue. (2.2.1). Le processus de construction de ces modèles a été considéré entre 1990 et 2000 comme le cadre essentiel permettant de situer les recherches en ingénierie des connaissances. Partant des questions pratiques que soulève sa mise au point, j'en présente les étapes. Je récapitule les divers points de vue qui se sont succédés sur la manière de conduire ce processus et qui ont marqué des jalons dans les recherches du domaine (2.2.2). Je présente ensuite les principales contributions (méthodes, logiciels d'analyse ou de modélisation de connaissances, langages de représentation) relatives à la modélisation (2.2.3).

2.2.1 La notion de modèle conceptuel

2.2.1.1 Des représentations en amont de la formalisation

La notion de modèle conceptuel a été utilisée en ingénierie des connaissances un peu avant 1990. Elle a succédé à des notions plus intuitives, qui répondaient à une vue du problème comme relevant de l'extraction de connaissances.

La première notion est celle de « mediating representation » (que l'on pourrait traduire par « représentation support » ou « intermédiaire ») utilisée vers 1989 pour désigner une forme intermédiaire matérialisant les connaissances recueillies auprès d'un expert (Gaines, 1980). Cette représentation devait favoriser les échanges, le dialogue avec les experts et l'expression de

nouvelles connaissances. Elle précédait leur formalisation sous forme de règles dans une base de connaissances. Un exemple de « mediating representation » est la visualisation des classements effectués à l'aide de techniques, le plus souvent issues de la psychologie, comme les grilles répertoires (Boose, 1988) (Gaines & Shaw, 1989) ou le tri de cartes (Shadbolt, 1988).

Deux autres notions sont celles de « modèle de tâche » telle qu'elle est proposée par Chandrasekaran (Chandrasekaran, 1986) ou de « méthode de résolution de problème » utilisée par Clancey pour caractériser Mycin (1986) et que l'on retrouve dans les systèmes développés par l'équipe de Mc Dermott au MIT : MORE, MOLE ou SALT présentés dans (Marcus, 1988). Ces *modèles* sont des caractérisations de haut niveau, en terme « de connaissances », des capacités d'inférence du système construit. Ici, leur intérêt est de mettre à plat la manière dont le système va résoudre les problèmes de manière efficace.

Pour ces deux courants, l'opérationnalisation se fait ensuite sous forme de règles. Les modèles ont servi à caractériser la nature de ces règles, à localiser leur intervention au cours des inférences associées à la résolution d'un problème, ou encore à vérifier que toutes les règles requises ont bien été construites.

Dans la méthode KOD (1988), Vogel propose de mettre en place plusieurs modèles à partir d'entretiens d'experts. Le premier, linguistique, reflète un découpage proche de l'expression des connaissances en langage naturel. Le second, le modèle cognitif du système, est à rapprocher de la notion de modèle conceptuel. Il se détache de l'aspect terminologique et linguistique pour organiser les connaissances à un niveau plus abstrait. Enfin, le troisième modèle est une opérationnalisation du précédent, pour laquelle Vogel est le premier à proposer d'adopter le paradigme objet.

Une hypothèse implicite forte est que ces modèles reflètent la manière dont l'expert résout le problème. Ainsi, à partir de cette période et jusqu'aux remises en questions de KADS débouchant sur CommonKADS, une confusion existe sur la nature des méthodes de résolution de problème. Par exemple, dans la bibliothèque de méthodes réutilisables proposées par KADS à partir de 1993, cohabitent des algorithmes de parcours de graphes propres à l'IA, comme hill-climbing ou A*, des méthodes renvoyant à des caractérisations vagues de résolution de problèmes (comme « la classification heuristique » de Clancey) ou encore des méthodes plus proches de la manière des experts de traiter certaines tâches (comme « model-based diagnosis »).

2.2.1.2 Définitions

À partir de 1991, il est clairement posé qu'acquérir des connaissances, c'est *construire* des *modèles conceptuels* (Krivine et David, 1991). J'insiste autant sur la définition donnée alors [RIA, 92][5] de ces modèles que sur les restrictions associées :

Un modèle est une abstraction qui permet de réduire la complexité en se focalisant sur certains aspects, en fonction de certains buts. MAIS un modèle devrait permettre plus : manipuler les objets et interpréter les résultats de la manipulation.

L'ensemble met en avant la place du modèle comme *outil* pour l'acquisition et la structuration des connaissances. Pour cela, le modèle doit tenir deux rôles presque orthogonaux. Le modèle doit pouvoir être interprété par un individu. Il doit aussi pouvoir être utilisé dans un système formel pour produire des résultats et être évalué. Toutes les facettes et toute la complexité du statut du modèle conceptuel sont ainsi posées. M. Linster évoque cette tension (Linster, 1992) à partir de son expérience de définition de la méthode et du langage OMOS, en identifiant deux facettes *: modèle pour abstraire des connaissances*, support au dialogue entre humains versus *modèle pour permettre au système de raisonner*, opérationnel et calculable. Le modèle a donc une

[5] Cette référence, et toutes celles de la forme [.., XX] citées dans ce chapitre, sont indiquées en fin de chapitre 4.

double sémantique : interprétative à la lecture par des humains et formelle, pour une interprétation informatique.

Cette dichotomie reste tout à fait d'actualité bien que la nature des applications visée aujourd'hui se soit diversifiée. En effet, les interprétations formelles envisagées dans ces définitions peuvent correspondre à la résolution de problème, sens le plus communément entendu en 1992, mais aussi à d'autres types de traitement (classification, navigation, etc.). Les modèles peuvent donc être utilisés plus largement pour ces autres types d'applications. Ensuite, le modèle doit toujours être interprété par des humains : après la phase de construction, le modèle peut être désormais rendu lisible directement dans des applications d'aide à la navigation dans des documents ou dans la mise en œuvre de tâches.

Des définitions plus orientées vers l'ingénierie et la gestion des connaissances ont été proposées par la suite. Je retiendrai la définition proposée dans le projet CommonKADS, qui tient lieu désormais de référence en la matière. "*The results of knowledge analysis are documented in the "knowledge model". It contains a specification of the information and knowledge structures and functions involved in a knowledge-intensive task.*" Les auteurs soulignent son double rôle pour la gestion du processus de modélisation et pour faciliter le développement du système. Ils insistent également sur la parenté de cette démarche avec celle du génie logiciel, l'ingénierie des connaissances s'intéressant à des logiciels particuliers, faisant appel essentiellement à des connaissances et soulevant donc des difficultés spécifiques. Le modèle de connaissances n'est alors qu'une des aides proposées parmi d'autres, au sein d'une démarche de gestion de projet qui définit un cycle de vie comparable à celui des logiciels classiques. Sont également explicités au niveau conceptuel (en s'affranchissant dans un premier temps des contraintes liées à la programmation) le contrôle sur les tâches à réaliser, la communication homme-système ainsi que le contexte organisationnel au sein duquel le couple système-utilisateur va réaliser la tâche.

Finalement, ces définitions convergent aujourd'hui, et tendent à fixer la nature des connaissances que couvre le modèle. Ainsi, je retiens la définition suivante de B. Bachimont :

Un modèle conceptuel d'IC exprime les connaissances d'un domaine relatives à une tâche dans un langage de modélisation ... Une fois rendu opérationnel dans un système*, il constitue un outil pour agir sur le monde, qui doit être pertinent en usage* (Bachimont, 2004).

L'usage auquel il est fait référence ici est celui du modèle au sein du système opérationnel qui va assister l'utilisateur final. Or un autre usage à prendre en compte est celui du modèle pendant la phase de sa mise au point, comme support à la structuration. Il tient leu de « brouillon » pour tester, imaginer l'impact de différentes organisations des connaissances. Il aide aussi à repérer des connaissances manquantes à recueillir, à rechercher ou à faire expliciter, ou comme une métrique qui mesure la couverture de ce qui a été modélisé par rapport au rôle du système final. *Le modèle conceptuel est donc tout d'abord un instrument du processus d'ingénierie des connaissances lui-même avant d'en être un résultat, c'est-à-dire une composante du système à base de connaissances en usage auprès d'utilisateurs*.

2.2.1.3 Des modèles cognitifs aux modèles de systèmes

La nature du contenu d'un modèle conceptuel qualifie ce à quoi renvoient les connaissances que le modèle matérialise : connaissances expertes, système informatique, théorie du domaine étudié, ... Il s'agit des éléments de référence qui donnent sa validité, sa valeur de vérité au modèle.

Les premiers modèles sont le fruit d'une démarche visant à restituer les connaissances et modes de raisonnement des experts. Ils correspondent à une *modélisation cognitiviste* qui fait l'hypothèse que l'intelligence du système pour traiter une classe de problèmes donnée sera d'autant plus grande que ce modèle sera proche des connaissances et processus mis en œuvre par des individus experts. Cette vision est retenue dans les méthodes KOD (Vogel, 1988), dans des systèmes d'enrichissement de base de règles comme TEIRESIAS (Davis, 1979), ou encore à la base de techniques comme les grilles répertoires (système ETS de J. Bradshaw et J. Boose, 1988). Pour

que ces modèles aient le statut de modèles cognitifs, les techniques de recueil et d'organisation des connaissances doivent faire appel à l'étude des phénomènes cognitifs humains. Elles sont donc empruntées des techniques, des méthodes, des hypothèses de fonctionnement et de représentations des connaissances en mémoire proposés en psychologie cognitive et en psychologie du travail. Cette vue cognitiviste a correspondu à l'hypothèse de développement des systèmes experts, dont la validité est jugée par comparaison tant aux performances qu'aux modes de raisonnement humain. Elle continue de prévaloir dans des contextes particuliers de simulation cognitive en vue de modéliser et agencer des situations de travail en ergonomie par exemple, ou encore en vue de transmettre des pratiques ou des modes particuliers de résolution dans des systèmes de formation.

À l'opposé, le modèle conceptuel peut être vu comme une pure construction artificielle. On parle dans la littérature *d'approche constructiviste*. Ainsi, construire des modèles à partir de méthodes génériques de résolution de problèmes relève d'une approche constructiviste. Des exemples de méthodes génériques sont les *Generic Tasks* de Chadrasekaran (1989) ou les *Role Limiting Methods* de Mc Dermott (1988). Dans les deux cas, la manière dont le système va traiter un problème s'appuie sur des modèles de résolution standardisés, adaptés à une classe de problèmes. Par exemple, la méthode « cover and differentiate » au sein du système MOLE guide la mise au point de système d'aide au diagnostic. Pour être adapté à une application particulière, chacun des systèmes utilisant *une Role Limiting Method* (MORE, MOLE, SALT …) suit le déroulement de la méthode pour interroger un expert, qui fournit les connaissances heuristiques propres au domaine concerné et requises par la méthode choisie. L'intérêt de la réutilisation de cadres génériques a été souligné encore plus fortement avec les méthodes KADS puis CommonKADS.

Or, l'approche constructiviste ne clôt pas le débat sur la *distance à prendre avec les modes de raisonnement des individus*. L'éventail des réponses possibles est large, et le choix dépend du type d'application. Ce débat rejaillit sur la manière de construire ces modèles, et la *part de la réutilisation* de modèles génériques. Le changement de point de vue entre les méthodes KADS et CommonKADS à ce sujet reflète les limites de propositions méthodologiques s'appuyant sur des modèles génériques relativement figés, et pas toujours bien identifiés. Ainsi, KADS suggérait de décrire l'organisation des tâches réalisées par le système à concevoir ainsi que la méthode de résolution de problème en reprenant des parties entières de modèles génériques. Ceci sous-entend que le modèle au sein du système peut être arbitrairement défini du moment qu'il est efficace. Or les applications de KADS ont souligné l'impact fondamental du modèle choisi et de son adaptation au contexte de l'application pour parvenir à un système utilisé. La méthode CommonKADS s'est donc enrichie d'indications précises, sous forme de questions et de critères, pour choisir puis adapter des modèles prédéfinis. Le modèle conceptuel est bien construit non pas pour rendre compte des processus de raisonnement d'experts, mais pour s'intégrer efficacement dans (et donc être adapté à) la pratique d'utilisateurs effectuant une tâche.

Actuellement, les modèles construits tirent leur légitimité de l'adéquation du système, ou même du couple système-utilisateur, à la tâche prévue. Ce sont les fonctionnalités du futur système, les problèmes à résoudre ou les tâches à traiter avec ce système qui déterminent le contenu du modèle. Ainsi, les modes de résolution de problème implémentés peuvent être relativement éloignés des pratiques des experts, ou encore le domaine peut être décrit de manière optimale par rapport à cette résolution, et non comme se le représentent des individus.

De ce fait, la distance entre les méthodes adoptées par l'utilisateur et celles qui sont implémentées dans le système n'est jamais grande pour deux raisons. Premièrement, le modèle sert de repère, de grille pour interroger les pratiques ou les savoirs dont il sera le support, et à ce titre, le cogniticien se doit d'argumenter et justifier les choix qui conduisent à faire traiter certaines tâches selon des méthodes différentes de celles des individus. Deuxièmement, en phase d'utilisation, le système doit interagir avec ses utilisateurs, et donc être acceptable. Certains traitements peuvent être effectués de manière transparente à l'utilisateur, et donc éloignée des pratiques. En revanche, le niveau de résolution perceptible par l'utilisateur et le faisant intervenir doit être acceptable et adapté. Finalement, même si seul le comportement du modèle opérationnel, ou même du système

utilisant le modèle, sert à juger sa validité, cette évaluation suppose également l'acceptabilité du système par ses utilisateurs, et ne permet pas de plaquer des méthodes de résolution choisies uniquement pour des raisons de performance qui ne seraient pas compatibles avec les usages.

2.2.1.4 Des modèles formels à des modèles pragmatiques

On peut se demander alors quelle est la part d'arbitraire dans la manière de fixer le contenu de tels modèles, et ce qui va finalement leur donner une validité. Là encore, des courants différents privilégient soit une validation opératoire ou formelle (*vue formaliste*), soit la fidélité au système de connaissances tel que la langue l'exprime (*vue normalisatrice*), soit encore des aspects pragmatiques liés à l'acceptabilité par les utilisateurs (*vue pragmatique*). Ces divergences de vue se retrouvent actuellement dans les travaux sur les ontologies mais aussi sur les modèles de tâches.

Ainsi, les ontologies peuvent, pour certains chercheurs, tenir lieu de théorie formelle d'un domaine ou des connaissances en général. Il s'agit par exemple des travaux sur les ontologies formelles inspirés de la philosophie comme ceux de (Guarino et Welty, 2002). Cette option est aussi implicitement retenue pour argumenter la définition d'ontologies réutilisables et partagées dans de nombreuses recherches, en particulier dans la perspective du web sémantique (Gómez-Pérez *et al.,* 2004).

L'autre *point de vue* est de tradition *linguistique*. Il pose le modèle conceptuel comme rendant compte d'un système linguistique supposé stable, partagé et pouvant être explicité au travers de définitions et de relations sémantiques. Cette hypothèse est à l'origine de structures de données lexicales ou terminologiques (comme WordNet ou MicroCosmos) qui comportent un noyau conceptuel unique pour rendre compte du sens de mots dans une ou plusieurs langues, ce noyau étant bâti sur une étude fine de la langue générale.

Enfin, le dernier type de modèles conceptuels, évoqué dans la partie précédente, décrit des connaissances telles qu'elles sont utiles à la *réalisation d'une tâche* et validées par la pratique. Ces modèles peuvent s'appuyer sur les pratiques des individus, leurs savoirs explicités dans des formations ou sur des documents contenant des traces de ces pratiques. Mais ils peuvent aussi être enrichis de nouvelles connaissances, adaptés en fonction des contextes, plus ou moins fidèles à ces pratiques en fonction du rôle du système final auprès des utilisateurs.

2.2.1.5 Rôles d'un modèle conceptuel

Le type de contenu d'un modèle lui confère un statut particulier, qui induit un mode de construction et de validation propre. Prendre l'un de ces partis cache un ensemble d'implications et de présupposés qui, d'une part, sont parfois discutables, et d'autre part, expliquent de nombreuses hypothèses sur la construction, l'utilisation ou la réutilisation des modèles. Par exemple, des modèles d'un domaine reflétant des connaissances théoriques de spécialistes de ce domaine, dans un objectif normatif de réutilisation et de représentation formelle du monde, justifient des approches privilégiant la réutilisation de connaissances génériques. Je récapitule ici les différents rôles que peut jouer un modèle conceptuel :

1) *Langage partagé par les acteurs de la modélisation,* en particulier le cogniticien et l'expert. Le modèle sert alors de support à la négociation pour spécifier comment le système va fonctionner, sur la base de quelles connaissances. Les enjeux de la définition d'un langage de modélisation touchent alors à la lisibilité, la présentation explicite et claire du contenu. L'interprétation étant faite par des individus, l'utilisation d'une terminologie proche de celle des acteurs du domaine est fondamentale. Ce sont eux qui donnent sa validité au modèle à ce niveau, quitte à ce qu'elle soit revue par l'opérationnalisation. Un autre point important selon ce rôle est la capacité de modifier, retoucher, faire évoluer facilement ce modèle, autant que celle de le commenter pour mieux en faciliter la lecture et en justifier la structuration auprès d'autres personnes. La définition du langage CML au sein du projet CommonKADS cherche à privilégier ce rôle.

2) *Cadre d'expression des connaissances.* Le modèle rassemble en général de manière inédite des connaissances identifiées comme pertinentes pour une pratique, un usage ou la réalisation d'une tâche. Il est le *support qui sert à les expliciter* et en assure une vérification syntaxique. Il doit donc assurer une expression précise, riche et non ambiguë des connaissances, en s'appuyant sur un langage structuré disposant d'une syntaxe bien définie. Ainsi dans ASTREE (Tort, 1996), logiciel de modélisation de connaissances du domaine, un langage inspiré des modèles Entité/Association permet de vérifier, par la syntaxe, des propriétés comme l'unicité de définition des concepts ou la présence de relations pour les différencier.

3) *Trait d'union entre connaissances mises en oeuvre et connaissances calculables*. Le modèle ayant vocation de déboucher sur un système opérationnel, il doit être finalement interprétable par le système informatique, *« compréhensible par l'artefact »*. Suivant les approches, le passage à un modèle opérationnel peut suivre un continuum où le modèle formel reprend une formalisation logique de chacun des éléments du modèle conceptuel, ou bien correspondre à un changement qualitatif, une traduction des éléments conceptuels en nouvelles primitives opérationnelles. Dans tous les cas, les enjeux portent sur la non-ambiguïté sémantique du langage et sa calculabilité (capacité à produire des raisonnements - classification, déduction, etc. - ou à produire le comportement prévu pour l'application finale en un temps fini). Par exemple, la classification de concepts est souvent un des objectifs qui justifient l'utilisation d'une logique de descriptions. Ou encore, les opérations possibles sur les graphes justifient une formalisation en graphes conceptuels dans des applications de recherche d'information et d'annotation sémantique par exemple (comme avec le système CORESE (Corby *et al.,* 2002). L'opérationnalisation est aussi un moyen de valider la sémantique du modèle.

4) *Langage de méta-niveau permettant de s'adapter à différentes applications*. Ce point de vue concerne plus le langage de modélisation lui-même que les modèles produits à l'aide de ce langage. Il correspond à la volonté d'adapter ou de spécialiser des primitives conceptuelles de haut niveau. L'objectif est de les agencer dans de nouveaux modèles de manière à répondre à des besoins particuliers en matière de gestion du raisonnement ou des tâches réalisées par le système, ou bien des interactions entre le système et l'utilisateur. Dans cet esprit, les langages ZOLA (Isténès, 1996) et DSTM (Trichet *et al.,* 1997) permettent de redéfinir les structures de tâche et méthode ainsi que la manière de les exploiter pour développer des systèmes dont le contrôle de la résolution de problème possède des propriétés particulières.

En dissociant différentes fonctions d'un modèle, on peut mieux caractériser la nature des structures et des langages permettant de les réaliser. L'intégration de l'ensemble des points de vue est ensuite un problème de compromis et de choix entre ces caractéristiques, parfois contradictoires, au regard des objectifs d'une classe d'applications. La complexité de la mise au point d'un modèle conceptuel découle donc de plusieurs difficultés : identifier les connaissances qu'il doit représenter, mais aussi les rôles prioritaires que l'on veut lui faire jouer, connaissant les contraintes qu'imposent ces rôles.

2.2.1.6 Différents types de connaissance présents dans un modèle conceptuel

Une des motivations historiques à la définition de modèles conceptuels pour des systèmes à base de connaissances est bien de caractériser la nature des connaissances à utiliser et de les organiser en fonction de leur utilisation dans la résolution de problèmes. La notion de moteur générique pour les systèmes à base de règles supposait déjà que l'on puisse isoler d'une part, une représentation déclarative des connaissances sous forme de règles de production, et d'autre part, la manière de les exploiter, l'algorithme de parcours des règles étant implémenté dans le moteur d'inférence. Or on retrouvait des types de règle qui intervenaient à des étapes différentes du raisonnement, ces règles jouant des rôles particuliers. Le modèle conceptuel essaie de rendre explicites ces rôles et de les organiser. Il renferme donc plusieurs types de connaissance dont la caractérisation est un des objectifs de la modélisation.

Composantes d'un modèle conceptuel

Aussi trouve-t-on dans presque toutes les approches au moins deux parties dans un modèle conceptuel : la caractérisation à un haut niveau de la manière de résoudre un problème, plus ou moins indépendamment du domaine traité, constitue le *modèle du raisonnement* ; les éléments propres au domaine considéré (concepts, relations ou règles heuristiques) forment le *modèle du domaine.* Ce découpage guide l'analyse des connaissances. Il facilite aussi la mise en correspondance entre des méthodes de résolution établies par l'IA et des problèmes réels. Établir cette passerelle est bien un des objectifs de l'ingénierie des connaissances.

À partir de là, la littérature des années 90 à 95 comporte une grande diversité de propositions quant à la manière de décrire la résolution de problème. Ainsi, L. Steels sépare les composantes tâche et méthode, alors que la méthode KADS suggère trois niveaux : stratégie, tâche (buts, décomposition en sous-tâches et contrôle sur leur activation et utilisation des opérateurs de la méthode de résolution) et inférence (méthode de résolution de problème sous forme de diagramme de données). Le niveau stratégie disparaîtra avec CommonKADS. Le contrôle n'y est plus considéré comme une couche de haut niveau, mais comme un modèle à part entière, lui-même décrit en trois couches, dont l'objet est de raisonner sur le raisonnement lié au domaine. Dans AIDE (Gréboval et Kassel, 1992) et OMOS (Linster, 1992), le niveau du contrôle regroupe la définition des tâches et de la méthode.

Le niveau *contrôle* est parfois isolé pour mieux être décrit et paramétré en fonction du contexte et des utilisateurs. L'objectif est que le système puisse modifier la manière de résoudre un problème ou d'interagir avec l'utilisateur en fonction du contexte. On parle alors de système réflexif, capable d'observer et de corriger son propre contrôle. À ce niveau du modèle conceptuel, les paramètres d'adaptation ou de modification du contrôle, les contextes agissant sur ces paramètres et leur action sur le raisonnement sont décrits explicitement. Cela n'a de sens que si cette dynamique est reproduite au niveau opérationnel, et donc si l'implémentation restitue la réflexivité du contrôle à l'exécution, comme le permettent les langages REFLECT (Reinders *et al.,* 1991) Lisa (Delouis & Krivine, 1995), *L*-AIDE (Gréboval *et al*., 1996), Mapcar (Tchounikine, 1994) puis ZOLA (Isténès, 1997), DSTM (Trichet, 1998) et def-* (Kassel *et al*., 2000). Ces langages assurent l'opérationnalisation du modèle en respectant sa structure au niveau conceptuel.

Un consensus plus net fait du *modèle du domaine* un réseau sémantique, où sont définies les classes conceptuelles sur lesquelles porte le raisonnement, leurs propriétés ainsi que des connaissances heuristiques (règles, contraintes, faits ou axiomes) sur ces classes. Seule K. Causse propose d'isoler un *niveau heuristique* à la charnière entre domaine et raisonnement pour représenter les inférences (Causse, 1993). Cependant, un débat existe pour savoir si ce modèle comporte ou non les instances de ces classes, ou bien si celles-ci ne sont définies que dans le système opérationnel. On parle alors de *modèle du cas traité* comme dans AIDE (Gréboval *et al*., 1996) pour désigner une résolution particulière, et donc un modèle instancié. Dans OMOS, le modèle opérationnel, appelé *modèle de la Tâche*, correspond à une application de la méthode de résolution de problème au domaine décrivant un problème particulier (Linster, 1996). La recherche d'invariants entre applications d'un même domaine a donné lieu à la notion d'ontologie dès 1990 (Gruber, 1993). L'ontologie est supposée contenir les connaissances propres à un domaine, indépendamment des raisonnements effectués dans ce domaine, organisées de manière à être réutilisables. J'y reviendrai par la suite.

Liens entre les composantes d'un modèle conceptuel

La définition des composantes du modèle va de pair avec celle des *interactions entre composants*, des passerelles qui permettent de passer de l'un à l'autre, ou de les utiliser de manière complémentaire pour réaliser une tâche. Ces liens peuvent être *explicites et directs* : les buts et opérations du raisonnement sont alors définis à partir d'éléments du domaine. Dans ce cas, la gestion des composantes du modèle ne peut se faire indépendamment. Leurs contenus, étroitement influencés l'un par l'autre, et ont peu de chance d'être réutilisables dans d'autres contextes.

Les *interactions* peuvent être *indirectes et limitées* : chacun des niveaux est alors défini avec ses propres éléments et les liens entre niveaux sont établis séparément. Dans cet esprit, afin de privilégier réutilisabilité et généricité, la méthode KADS s'appuie sur une hypothèse d'interaction la plus réduite possible entre les niveaux. Selon cette hypothèse, construire un modèle conceptuel, c'est représenter séparément les tâches réalisées, les méthodes de résolution utilisées et les concepts du domaine mis en jeu. Puis des liens précis sont posés entre ces niveaux. Or pratiquement, des influences réciproques entre niveaux existent : toute description structurée d'un domaine ne convient pas pour toute méthode de résolution, ou encore la manière de définir les tâches à réaliser dépend de la méthode retenue. Ainsi, si un rôle comme `hypothèse` est en entrée de l'opérateur `classement`, les connaissances du domaine associées doivent pouvoir être classées et un critère de classement explicité.

Les interactions peuvent donc être *indirectes mais fortes et explicitées*. La notion de poignée conceptuelle sert par exemple à associer des classes d'objets (concepts) à des types (des rôles) au sein de modèles construits à l'aide du langage ZOLA (Isténès, 1995). La méthode CommonKADS insiste sur les engagements ontologiques que déterminent les rôles au sein d'une méthode de résolution (Schreiber *et al.*, 2000). De ce point de vue, les différents composants d'un modèle doivent être mis au point conjointement, les choix retenus à un niveau ayant une incidence sur les autres. Ainsi, la caractérisation de la méthode de résolution de problème définit les sous-buts ou sous-tâches à atteindre, et inversement.

2.2.2 De l'acquisition à l'ingénierie des connaissances

L'historique de l'ingénierie des connaissances peut être retracé à partir des changements de points de vue sur le processus de modélisation qui ont jalonné les recherches du domaine. Avant de présenter ce parcours historique et les différentes approches de la modélisation encore d'actualité, je m'appuie sur les questions pratiques que soulève la mise au point d'un modèle.

2.2.2.1 Questions pratiques d'un projet d'IC

Un projet d'ingénierie des connaissances porte, dans un premier temps, sur des analyses de besoin, des analyses de tâche et d'activité, qui débouchent sur un travail d'inventaire de ressources d'une part, et de spécification de logiciel d'autre part. À titre d'illustration, le tableau 1 présente des exemples de questions abordées pour mener ces premières étapes.

- A quel besoin veut-on répondre ?
  - quelle est la tâche à assister ? Qui en maîtrise la réalisation actuellement ?
  - qui sont les utilisateurs ? avec qui collaborent-ils ? au sein de quelle organisation ?
  - comment leur tâche doit-elle évoluer ?
- Comment répondre à ce besoin ?
  - Quel système informatique définir ?
  - quelle part de la tâche va-t-il prendre en charge ? comment l'utilisateur réaliserait-il sa tâche à l'aide de ce système ?
  - comment le système va-t-il traiter la part qui lui revient ? selon quelle méthode ?
  - quelle interaction vise-t-on entre le système et l'utilisateur ?
- Comment accéder aux connaissances qui permettront au système de réaliser la tâche définie ?
  - Quels types de connaissances viennent-ils répondre à ces besoins ?
  - Quelqu'un détient-il les connaissances correspondant au domaine de l'application et à la méthode choisie ? La méthode relève-t-elle d'une solution algorithmique connue ?
  - Peut-on les trouver dans des documents, formations, thesaurus ou bases de données ?
  - Ces connaissances sont-elles déjà diffusées, explicitées, structurées ou bien sont-elles implicites et locales ?
  - Qui, du système ou de l'utilisateur, va prendre des décisions sur le déroulement de la tâche ?

Tableau 1 : Exemple de questions relatives à un projet d'ingénierie des connaissances.

Les réponses apportées à ces questions ont évolué avec la notion de système à base de connaissances et en fonction des bilans tirés des expériences de développement de l'ingénierie des connaissances. Suivant les réponses, le système à développer devra comporter ou non une opérationnalisation de la tâche à réaliser, codée à l'aide de connaissances ou d'algorithmes connus. Il comportera ou non un modèle explicite des connaissances du domaine permettant de guider ou d'orienter l'utilisateur.

- Comment déterminer puis recueillir, rassembler de manière explicite et représenter les connaissances ainsi spécifiées ?
  - quelles techniques et logiciels utiliser pour le recueil d'expertise ou de connaissances spécialisées, pour l'analyse de textes ou le dépouillement de données ;
  - choix ou définition de représentations de connaissances adaptées pour modéliser la tâche à réaliser et le domaine concerné ;
  - adopter une approche orientant l'utilisation des logiciels de recueil, le dépouillement de leurs résultats pour organiser un modèle et l'articulation entre les différentes composantes du modèle
  - valider ce modèle
- Comment les rendre opérationnelles conformément au rôle retenu pour le système ? comment s'assurer qu'elles vont permettre au système de réaliser la tâche attendue ?
  - quel formalisme utiliser ? quelle architecture retenir pour le système final ?
  - jusqu'où conserver le modèle ou la structure du modèle dans le système final ?

Tableau 2 : Exemples de questions relatives à la modélisation des connaissances.

Dans un deuxième temps, se posent des questions relatives au processus à mettre en place pour modéliser ces connaissances et spécifier le système à concevoir, questions plus au cœur de l'ingénierie des connaissances (tableau 2). Il s'agit de définir ou d'utiliser des techniques d'identification et de recueil de connaissances, d'explicitation, de choix, de structuration et de modélisation. On s'intéresse ensuite à l'opérationnalisation des modèles, à leur validation opératoire ou leur capacité d'interprétation par le système opérationnel.

Ces deux groupes de questions ne sont pas indépendants : les modalités de l'opérationnalisation font partie des choix méthodologiques, et peuvent avoir des conséquences sur la représentation des connaissances. Ils doivent également être traités en ayant à l'esprit les problèmes amont d'étude de besoins. Cependant, suivant que la priorité sera donnée à l'un ou l'autre, les recherches de l'IC s'orientent vers des propositions de natures différentes. Dans le premier cas, l'accent est mis sur les aspects méthodologiques, sur le support au travail du cogniticien. Les recherches peuvent par exemple porter sur la manière de mettre en forme une méthode et la rendre accessible à des cogniticiens, sur la manière de garder des traces des choix de modélisation dans le cadre d'une méthode donnée ou encore sur la nécessité ou non d'établir une correspondance structurelle entre un modèle et son implémentation.

2.2.2.2 Étapes du processus de modélisation

J'ai déjà évoqué le cheminement historique qui a conduit à introduire des modèles conceptuels en amont de la définition de systèmes à base de connaissances. Ces modèles caractérisent les différents types de connaissance utiles à la résolution de problèmes. L'objectif de construire un modèle conceptuel guide l'identification (le recueil) et la représentation des connaissances. Le processus de modélisation comporte alors deux types d'activités (étapes 1 et 2 de la figure 2.3) : rassembler des indications, entretiens, observations sur la manière dont les experts procèdent puis les organiser, les structurer dans un modèle.

Avec la notion de « système expert de 2e génération » puis, plus tard, de système à base de connaissances coopératif, on admet que le système peut raisonner selon sa propre méthode. Le modèle n'est plus alors systématiquement le reflet exact de la méthode mise en oeuvre par l'expert. De plus, c'est la méthode choisie qui va déterminer les connaissances requises : elle oriente

l'acquisition. On parle *d'acquisition guidée par les modèles*. Ce processus correspond à l'ajout de l'étape 3 sur la figure 2.3 : pour parvenir au modèle conceptuel visé, le modèle en cours de construction sert de cadre, fournit des repères pour orienter la recherche de connaissances complémentaires. Une première version du modèle (incomplet) permet de revenir sur les étapes 1 et 2 en indiquant de manière plus précise la nature des connaissances à rechercher auprès des sources de connaissances. Ce schéma est une abstraction de certaines propriétés des connaissances utiles au système visé.

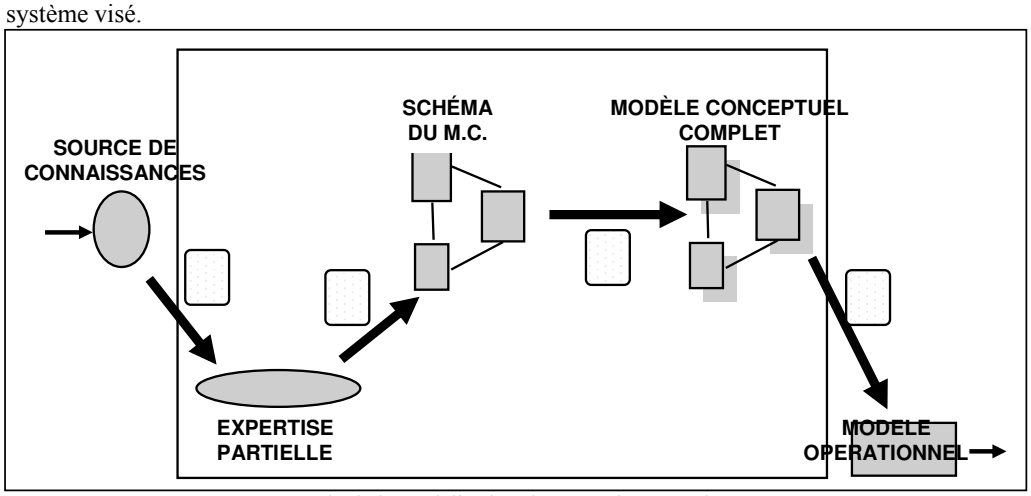

Figure 2.3 : Cycle de la modélisation des connaissances d'après [RIA, 92].

Enfin, parce que l'opérationnalisation (étape 4 sur la figure 2.3) va bien au-delà d'une simple traduction du modèle conceptuel, cette dernière étape fait partie du processus d'ingénierie des connaissances. L'observation du comportement et des erreurs éventuelles du modèle opérationnel peut nécessiter de préciser des connaissances, de corriger ou de compléter le modèle.

Cette analyse, proposée dès 1992 [RIA, 92], a été par la suite revue et affinée. En particulier, la notion de schéma de modèle conceptuel rend compte non pas d'un modèle partiel mais d'un modèle caractérisant les connaissances à un niveau plus abstrait. Il guide l'acquisition et la structuration de connaissances plus précises. Par exemple, dans la méthode KADS, la structure d'inférence oriente le recueil de concepts et d'heuristiques du domaine jouant les rôles définis par cette structure.

2.2.2.3 Aspects ascendants versus descendants, réutilisation

Cette analyse présente l'intérêt de fournir un cadre encore pertinent pour rendre compte des tâches impliquées dans le processus de modélisation d'une part, et pour situer et comparer les propositions de l'état de l'art au cours des années.

Une première dimension d'analyse est celle de la formalisation, qui correspond à l'axe horizontal sur la figure 2.3. La modélisation conceptuelle progresse de manière continue selon cet axe, jusqu'à l'obtention d'un modèle opérationnel. Suivant les approches, le modèle obtenu se situe à la fin de l'étape 3 (cas de MACAO) ou à l'étape 4 (cf. CommonKADS). Dans le cas où le même langage serait utilisé pour décrire le modèle et pour son opérationnalisation, comme dans les *Role Limiting Methods* (Marcus, 1988) ou COMMET (Steels, 1990), les étapes 3 et 4 sont imbriquées.

Une deuxième dimension caractérise les tâches d'abstraction versus celles de spécialisation. Elle correspond à l'axe vertical sur la figure 2.3, où le schéma du modèle conceptuel et le modèle lui-même se situent à un niveau plus abstrait que le système opérationnel puisque, par nécessité, le modèle opérationnel manipule aussi des instances.

L'étape 2 est finalement le cœur du processus : elle inclut le choix ou l'identification de caractéristiques pertinentes de la résolution de problème et des concepts du domaine. La manière

de mener cette étape a fait l'objet de nombreuses recherches. Elle peut être considérée comme une *tâche ascendante* où, partant des données recueillies, on cherche à dégager une représentation conceptuelle caractérisant les raisonnements que le système mettra en œuvre et les éléments du domaine associés. Elle peut aussi faire appel à la *réutilisation* de structures génériques (en général des méthodes de résolution de problème), propres à des types de problème particuliers, suivie d'une *tâche descendante* d'adaptation, d'affinement de ces structures et d'inventaire des connaissances du domaine associées, remplissant les rôles définis par ces structures. Ces modèles guident alors le recueil et l'analyse des données auprès d'experts ou dans des textes.

2.2.2.4 Évolution des points de vue sur la modélisation

Ce cadre unifiant les approches de modélisation constitue un moyen de mieux identifier les résultats produits, les points variables, les changements de points de vue et d'approches au cours de l'histoire de l'ingénierie des connaissances.

J'ai déjà évoqué deux des paramètres qui ont évolué : le *type d'application visée*, la manière de rendre opérationnelles les connaissances dans un système informatique ; et le point de *référence du modèle* (modèle cognitif versus modèle du système). En lien avec le type d'application, un autre paramètre est le *degré d'opérationnalisation* et de formalisation du modèle : l'histoire de l'IC tend à refléter une prise de distance avec l'IA : l'objectif visé n'est pas en priorité de faire un système intelligent, mais bien d'augmenter l'efficience de l'utilisateur aidé par le système. De ce fait, alors que les premiers modèles prenaient en charge la totalité de la résolution de problème, les modèles suivants ont facilité la répartition des tâches entre le système et l'utilisateur. Finalement, dans les applications actuelles, le modèle ne reflète qu'une partie des connaissances (connaissances du domaine comme les ontologies, manipulées par un moteur de recherche ou autre application, ou la partie interaction système utilisateur, etc.).

Dans ce contexte, les connaissances stabilisées, consensuelles et partagées prennent le pas au sein des modèles sur les savoir-faire et l'expertise individuelle. Il s'agit d'un saut qualitatif considérable, qui traduit une sorte de renoncement à la vocation première des systèmes experts et qui répond à des besoins un peu différents. Ce changement reflète une vision plus réaliste et plus générale de l'aide à l'utilisateur.

Au fil des expériences, la *nature des sources de connaissances interrogées* s'avère très diverse (expert humain, collectif de spécialistes ou de futurs utilisateurs, textes techniques, documents liés à un domaine ou à l'activité …). De ce fait, la place de l'expert est souvent réduite au profit de celle de l'analyste, chargé entre autres d'observer et analyser les activités humaines, rassembler et analyser des textes, etc. L'expert n'intervient que pour des validations fondamentales.

Plus radicalement encore, le statut du modèle conceptuel est en train de prendre un nouveau tournant qui reflète un point de vue différent sur les connaissances (ce point de vue n'est pas nouveau en sciences cognitives ou en philosophie, c'est son appropriation en IC qui est récente). En effet, si le modèle conceptuel est un modèle de connaissance, ce n'est pas parce qu'il rend compte de connaissances humaines sous forme d'un objet manipulable et palpable, sous la forme de représentations. C'est parce qu'il permettra, à travers un système informatique l'utilisant, d'en observer des traces, des manifestations ou le fruit de leur utilisation, ou encore plus simplement de donner accès à des sources d'information que l'utilisateur interprète pour construire des connaissances.

Le processus d'ingénierie des connaissances, après avoir été considéré comme une extraction, une acquisition puis une modélisation de connaissances, doit maintenant prévoir un support à la co-construction de connaissances pendant l'usage du système visé. L'IC doit donc outiller la gestion de modèles permettant aux systèmes informatiques de restituer des traces, des inscriptions de connaissances auprès de leurs utilisateurs.

2.2.3 Les résultats relatifs à la modélisation

Des travaux spécialisés offrent un support à certaines tâches du processus de modélisation. Ils proposent soit des langages, soit des techniques de recueil, soit des éditeurs de modèles. Au contraire, d'autres recherches plus ambitieuses intègrent plusieurs outils au sein de plates-formes ou s'intéressent aux aspects méthodologiques.

La production de résultats dans le domaine passe par une visée de réutilisation, qui cherche à *généraliser les questions et les résultats propres à chaque expérience,* et à dégager des invariants de la diversité et de la complexité que j'ai signalées. Cette démarche suppose de dépasser l'étude de cas d'école, de propriétés théoriques ou formelles de représentations, ou de la validation isolée d'outils d'aide (Tchounikine, 2002). Elle requiert de se fixer des contextes d'évaluation en vraie grandeur, tout en se donnant les moyens de mesurer la contribution et l'intérêt de chacun des éléments des démarches proposées. Loin de s'imposer facilement, ce type d'analyse se développe dans la mesure où l'on peut s'appuyer maintenant sur un plus large échantillon de recherches, d'approches et de retours d'expérience.

Ces analyses conditionnent les avancées dans le domaine et l'établissement de résultats qui puissent avoir un écho, une diffusion au-delà du cercle de l'ingénierie des connaissances. Les rapprochements attendus portent, par exemple, sur les types de connaissance permettant de répondre à des types de besoin ou d'application. Autre exemple, des contributions plus précises devraient déboucher dans les années à venir sur la manière d'intégrer différents types de connaissance et réutiliser au mieux les ressources existantes. Les propositions actuelles en matière de langages, de modèles, de méthodes ou de techniques ne peuvent s'affranchir de l'identification des contextes (types de connaissance, d'application ou d'utilisateur) dans lesquels ils sont pertinents. Ainsi, les efforts de standardisation autour des langages de représentation d'ontologies pour le web sémantique sont tout à fait révélateurs de cette volonté. Ils concernent différentes communautés (représentation des connaissances ou IA en général, réseaux et télécommunication et gestion documentaire entre autres). La part de l'IC dans cet effort est de restituer son savoir faire sur l'évaluation des formalismes. Elle rappelle que ces formalismes doivent prendre en compte les contextes d'usage particuliers, permettre des raisonnements adaptés aux systèmes les utilisant.

2.2.3.1 Techniques et outils

Pour faire des propositions pratiques en matière d'accès aux connaissances à travers les personnes ou les documents qui sont supposés en fournir des indications, l'IC a forgé des solutions qui lui sont propres. Elle s'est aussi largement inspirée de disciplines proches en fonction de la source de connaissances considérée : ces disciplines ont couvert successivement la psychologie cognitive puis l'ergonomie puis la terminologie et la linguistique de corpus. Des panoramas de ces techniques sont présentés dans [thèse-AUSSENAC, 89], (Schreiber, 1996) ou (Dieng et al., 2000).

Par *techniques,* je fais référence à des « modes opératoires » préconisant des modes de choix ou de création de situations de production ou d'utilisation de connaissances, puis la manière de repérer-recueillir-extraire ou analyser ces données et enfin des propositions pour interpréter, dépouiller, structurer les fruits de cette analyse. La psychologie cognitive a fourni de nombreuses indications sur les différentes techniques d'entretien, leur analyse, les avantages et limites de chacune. L'IC a repris à l'ergonomie des techniques d'analyse de la tâche et de l'activité.

Dans certains cas, des *outils* ont été définis en IC pour faciliter la mise en place de certaines techniques ou le dépouillement des données ainsi recueillies. Ces logiciels assistent les modes opératoires définis ou en créent de nouveaux. Par exemple, certains logiciels sont des supports aux techniques de classification, de tri de cartes (comme 3DKAT), ou encore aux grilles répertoires comme ETS et AQUINAS (Boose, 1987), (Boose, 1989) … Certains logiciels facilitent la simulation de l'activité ou l'extraction de connaissances à partir des données recueillies. Par exemple, des outils permettent d'annoter des textes, des retranscriptions d'entretiens, d'identifier des connaissances particulières, comme le repérage de taxèmes ou d'actèmes avec la K-Station

associée à la méthode KOD. D'autres outils simplifient le stockage de résultats (bases de données plus ou moins structurées, bases de cas, d'entretiens retranscrits …) et l'analyse de résultats (tris, recoupements, analyses statistiques, etc.)

2.2.3.2 Langages pour la modélisation

Avec le modèle conceptuel, on espère se situer au-dessus de l'opérationnalisation informatique et ainsi de ne pas contraindre les représentations par des critères de performance ou de calculabilité. L'IC ne reprend donc pas exactement les formalismes de l'IA (car ils privilégient la capacité à raisonner et produire des inférences), mais cherche à faire des propositions nouvelles en s'appuyant sur d'autres travaux cognitivistes. On peut citer par exemple les structures de représentations mentales proposées par la psychologie cognitive, modèles de tâches de l'ergonomie, représentation proposées en gestion des connaissances (graphes de connaissances) ou pour l'analyse du langage naturel (graphes conceptuels). L'IC a en effet besoin de proposer des langages de modélisation qui favorisent l'expressivité, la caractérisation abstraite des connaissances tout en préparant leur formalisation.

Des langages comme CML dans CommonKADS (Schreiber *et al.*, 2000) ou le langage dérivé d'Entité-Association dans ASTREE (Tort, 1996) facilitent la mise au point du modèle par le cogniticien, et le rendent facilement lisible par un humain. D'autres privilégient la formalisation, la précision et la capacité de raisonnement. Dans cette dernière catégorie, on trouve encore deux types de langage. D'une part, ceux qui rendent calculable le modèle conceptuel en restant au plus près des structures initiales, que l'on appelle langages de prototypage ou de formalisation ou d'opérationnalisation parfois ; par exemple, ForKADS (Wetter, 1992), OMOS (Linster, 1992) ou Def_* (Kassel *et al.,* 2000). D'autre part, ceux qui rendent le modèle opérationnel dans le système final. On parle ici aussi de langage d'opérationnalisation ou d'implémentation comme LISA (Delouis, 1994) ou KARL (Fensel, 1994).

Ces points de vue divergent sur l'intérêt qu'il y a ou non à préserver, dans l'application finale, l'organisation structurelle du modèle ; sur l'importance accordée aux performances et la calculabilité versus la lisibilité ou l'expressivité ; enfin sur la complexité et le coût que peut générer la nécessité de passer successivement par trois représentations successives : un modèle structurant sans formaliser, un modèle formalisant pour valider, puis un codage permettant au système final de fonctionner de manière optimale. Les choix dépendent de la taille et de l'ambition du projet, des besoins en maintenance mais aussi du type d'application.

Pour gérer la tension qu'impose de répondre aux différents rôles que l'on veut faire jouer à un modèle conceptuel (représenter, raisonner, échanger des connaissances …), la plupart des travaux proposent de s'appuyer sur plusieurs représentations successives et sur différents formats des mêmes primitives. Par exemple, un format graphique ou un langage d'expression simple de type frame convient bien pour les phases initiales d'organisation des données recueillies. Ensuite, une représentation structurée et une sémantique plus précise permettent de guider la modélisation. Citons par exemple les langages tâches-méthodes pour représenter le raisonnement ou bien RDFs ou OWL pour le domaine. Enfin, la vérification syntaxique ou sémantique s'appuie souvent sur une représentation logique.

2.2.3.3 Méthodes, éditeurs de modèles et plates-formes

Certains concepteurs de langages prévoient que le cogniticien écrive directement un modèle à l'aide du langage (cas des langages OMOS, LISA ou Def-*). D'autres proposent un éditeur spécifique, qui guide l'interaction avec l'utilisateur, vérifie la syntaxe des structures définies et des liens entre structures. Par exemple, le langage CML de CommonKADS dispose d'un éditeur. De même, les différents langages de représentation d'ontologies sont accessibles via des éditeurs, comme OILEd pour OIL.

Choisir et utiliser des techniques de recueil et d'analyse de données pour alimenter et structurer un modèle selon un langage donné requiert un véritable savoir faire. Le cogniticien chargé de ce travail doit pour cela adopter un point de vue sur les connaissances, sur l'application finale et sur ce qu'est un modèle conceptuel pertinent dans ce contexte. Pour guider l'ensemble du processus, des propositions méthodologiques ont donc été définies dans des projets d'envergure, surtout au niveau européen, comme KADS, VITAL, COMMET puis CommonKADS, ou encore KOD. Pour assurer cohérence et continuité du processus défini par chacune de ces méthode, la plate-forme informatique associée regroupe à la fois des outils de modélisation et de recueil.

2. 3 - Analyse des travaux sur ontologies et textes

L'analyse des textes en ingénierie des connaissances a toujours été présente en ingénierie des connaissances, mais la manière de l'aborder a radicalement changé après 1990 avec la référence aux bases terminologiques et au traitement automatique des langues (Bourigault, 1994). Elle a pris un véritable essor avec le déploiement des ontologies. En effet, les textes fournissent des éléments stables, consensuels et partagés d'un domaine, comme des descriptions d'objets et de concepts tels qu'on peut en avoir besoin pour former un modèle du domaine ou une ontologie. L'analyse de textes a ciblé la construction de ressources proches que sont les ontologies, les thesaurus, les index, les lexiques ou les bases de connaissances terminologiques. C'est dans ce sens que je l'entendrai dans tout ce mémoire.

Cette problématique se justifie en effet par deux finalités. D'un côté, il s'agit d'exploiter une source de connaissances complémentaire ou se substituant à l'expertise humaine - les textes - avec l'espoir que leur analyse plus ou moins automatique pourrait accélérer la construction des modèles conceptuels. De l'autre, plusieurs types d'applications requièrent des représentations des contenus des textes, sous la forme de ressources rendant compte d'éléments lexicaux, terminologiques ou sémantiques. Leurs concepteurs viennent chercher dans l'ingénierie des connaissances des propositions de structures de données pour ces ressources et de méthodes pour les construire.

Je présente ici ces deux nouvelles facettes de la modélisation en lien avec des textes, l'une justifiée par la construction de modèles, l'autre par la gestion de documents.

2.3.1 Textes et modèles de connaissances

2.3.1.1 L'acquisition de connaissances à partir de textes

L'acquisition des connaissances à partir de textes est un thème présent dans les recherches depuis le début de l'acquisition des connaissances. Ainsi, un des premiers écrits des précurseurs de la modélisation conceptuelle, B. Wielinga et J. Breuker (Wielinga & Breuker, 1984), porte sur l'exploitation de retranscriptions d'entretiens d'experts pour la modélisation conceptuelle. Loin de considérations linguistiques, leur proposition est justement de ne pas rester au plus près du fil du texte, mais de caractériser la nature du problème traité, de la tâche effectuée et de la méthode choisie pour la traiter. Cependant, à cette même époque, d'autres auteurs proposent de s'intéresser au matériau linguistique que présentent les textes en tant que traces de connaissances. Dans sa thèse, D. Bourigault repère deux courants successifs (Bourigault, 1994a) :

- Dans un premier temps, au cours des années quatre-vingt, l'acquisition de connaissances à partir de texte s'appuyait sur des analyses manuelles, le cogniticien cherchant à effectuer un « transfert » des textes vers les modèles.
- Ensuite, à partir de 1990, des travaux ont fait appel au Traitement Automatique des Langues (TAL), s'appuyant sur une vraie réflexion linguistique et terminologique, voire ontologique, sur le passage d'une analyse de surface des textes à des modèles et représentations.

Chacune de ces périodes a donné lieu à une reformulation de la problématique :

Lors de la première période, il s'agit de définir comment repérer des connaissances dans des textes et dans des entretiens d'experts. Plus rarement, la question posée est de disposer d'une théorie générale de la langue pour repérer des éléments de modèle à partir de phrases. Ainsi, la méthode KOD (Vogel, 1988) propose une correspondance entre des primitives linguistiques (ou plutôt sémiotiques, comme les schèmes et les sèmes) pour repérer des concepts et éléments d'actions ensuite représentés sous forme d'objets d'un modèle. Ce repérage manuel dans des entretiens non dirigés et retranscrits suppose une lecture exhaustive et un dépouillement minutieux. La méthode Cognosys (Woodward, 1989) propose un découpage manuel de paragraphes, ensuite analysés phrase à phrase pour repérer des règles heuristiques. Dans les deux cas, il manque une prise en compte de la nature des entretiens, une vraie réflexion sur la constitution des corpus ou sur le recul à prendre entre l'information trouvée dans les textes et le modèle pertinent pour le système à concevoir.

Les questions se posent différemment dès lors que l'on envisage d'outiller l'analyse de textes : Comment des logiciels d'analyse, des extracteurs et des techniques linguistiques peuvent-ils aider à repérer des connaissances à partir de textes ? Comment les termes en usage sont-ils révélateurs de structures conceptuelles, lesquelles ? Sont-ils des indices pour construire des représentations informatiques ? Comment les représenter, en rendre compte pour qu'elles soient pertinentes pour une application donnée ? Les premières réponses à ces questions ont été apportées par les recherches sur les ontologies génériques (Sowa, 2000), sur les bases de connaissances terminologiques (Skuce et Meyer, 1992) ou encore sur l'extraction terminologique (Reimer, 1990) (David et Planque, 1996).

2.3.1.2 Des modèles conceptuels aux ontologies

Une autre tendance a fortement marqué l'IC à partir de 1992 : le développement des premiers travaux sur les ontologies (Gruber, 1991) (van Eijst, 1995). Depuis cette période, les ontologies sont progressivement devenues le cœur de la majorité des nouvelles recherches en IC, occultant les problématiques de modélisation du raisonnement ou de la résolution de problème, jusque-là considérées comme le cœur de la problématique de l'IC. Ce glissement radical revient à renouveler la répartition des connaissances au sein du système. La réalisation des tâches et leur enchaînement, nécessitant pourtant une modélisation parfois complexe, n'est plus au cœur des études. En effet, elle est supposée prise en charge par le système de manière plus ou moins algorithmique, ou laissée à l'utilisateur. Au contraire, les connaissances propres au domaine et manipulées par les raisonnements ou directement par le système sont devenues centrales et l'objet de toutes les attentions. Il s'agit de ne plus organiser un réseau conceptuel de manière intuitive ou empirique en fonction des besoins de la tâche, et de trouver des critères explicites de bonne structuration de ces connaissances. La notion d'ontologie est venue répondre à ces besoins. Leur étude correspond à un changement plus profond de point de vue sur ce qu'est et peut faire un système « à base de connaissances ».

Je ne reviendrai pas ici sur l'historique de la notion d'ontologie en philosophie, sur les traditions philosophiques définissant le concept ni sur les typologies d'ontologies et courants présents en IC. Le lecteur peut consulter les habilitations de J. Charlet (Charlet, 2003) et B. Bachimont (Bachimont, 2004) qui abordent ces points de manière très complète.

Les recherches sur les ontologies en intelligence artificielle font suite à des initiatives croisées sur la réutilisation de connaissances du domaine, leur meilleure indépendance par rapport au raisonnement (van Eijst, 1995) et surtout sur la définition de représentations des connaissances facilitant un meilleur échange de bases de connaissances, en particulier le Knowledge Sharing Effort aux USA (Neches *et al*., 1991). Les motivations initiales au développement des ontologies sont essentiellement la réutilisation des modèles de connaissances du domaine d'une application à l'autre, l'interopérabilité des systèmes les utilisant, leur maintenance ainsi qu'une plus grande validité, faisant consensus entre spécialistes du domaine (Gruber, 1991) (Valente *et a*l., 1996). À

ces éléments, le développement massif d'applications pour le web a soudain ajouté de nouveaux enjeux, à la fois techniques et économiques, ayant des conséquences sur la forme mais aussi le fond des modèles attendus. Les propositions d'architecture ou d'applications pour le futur web dit « web sémantique » font systématiquement appel aux ontologies : elles doivent fournir des représentations partagées utilisables par des agents logiciels, des bases de méta-données pour annoter ou indexer des documents ou encore assurer la mise à disposition de tous de bases de connaissances consensuelles. Ainsi, les ontologies couvrent des réalités différentes suivant qu'elles sont destinées à être des connaissances partagées entre agents logiciels, des supports pour des systèmes interagissant avec l'utilisateur ou encore des ressources de méta-données pour indexer ou annoter des documents.

2.3.1.3 Les ontologies en ingénierie des connaissances

L'évolution de la définition des ontologies reflète les débats dont elle fait l'objet en IC, et le fait que différentes études ont été nécessaires pour stabiliser ce qu'elles peuvent être. Ainsi, les premières définitions mettent l'accent sur l'aspect représentation formelle des connaissances (Gruber 1991), ainsi que sur les aspects « vocabulaire et définitions des concepts d'un domaine » (Ushold & Gruninger, 1996). Je présente ici une définition actuelle, proposée dans (Studer *et al.,* 1998), car elle fait consensus dans le domaine et intègre de manière pertinente plusieurs clarifications relatives à la définition assez fondatrice pour l'IC donnée par (Gruber, 1993) :

An ontology is a formal, explicit specification of a shared conceptualisation. Conceptualisation refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. Explicit means that the type of concepts used, and the constraints on their use are explicitly defined. Formal refers to the fact that the ontology should be machine-readable. Shared reflects the notion that an ontology captures consensual knowledge, that is, it is not private of some individual, but accepted by a group.

Ainsi, l'ontologie répond à des exigences complémentaires et symétriques : (i) en tant que spécification, elle définit une sémantique formelle pour l'information permettant son exploitation par un ordinateur ; (ii) en tant que reflet d'un point de vue – partiel – sur un domaine, que l'on cherche le plus consensuel possible, elle fournit une sémantique qui doit permettre de lier la forme exploitable par la machine avec sa signification pour les humains.

Une autre définition, proposée par Charlet (Charlet, 2003), a servi de repère à mes travaux à partir 1998, et peut compléter la première :

Une ontologie est une spécification normalisée représentant les classes des objets reconnus comme existant dans un domaine. Construire une ontologie, c'est aussi décider d'une manière d'être et d'exister des objets de ce domaine.

Cette définition, inspirée des travaux de B. Bachimont (Bachimont, 2004), met l'accent sur l'importance du point de vue retenu pour construire l'ontologie sur son contenu. Elle souligne également la notion de normalisation, qui fait allusion aux critères « ontologiques » de structuration qui guident l'identification et la description de concepts selon le point de vue retenu. Enfin, elle sous-entend que cette spécification est formelle, qu'elle définit des concepts qui vont servir de « vocabulaire » à une théorie logique représentant formellement l'ontologie. L'école constructiviste fait l'hypothèse forte que cette théorie peut refléter toutes les connaissances que l'on cherche à représenter, toute la conceptualisation visée, et cela de manière universelle. Ce courant a fortement influencé les approches de l'ingénierie des connaissances, qui ont mis en avant le caractère générique et réutilisable de ces modèles.

Plus nuancés, les travaux de (Guarino et Giaretta, 1995) parlent au sujet de cette théorie d'« engagement ontologique » pour souligner qu'elle ne peut rendre compte que partiellement de la conceptualisation visée. B. Bachimont accentue cette restriction (Bachimont, 2004), entre autres parce qu'il s'appuie sur l'usage de la langue pour dégager des concepts, et aussi parce qu'il propose des principes différentiels appliqués à des unités linguistiques comme principes ontologiques. Pour

lui, c'est la position d'un concept dans le réseau conceptuel qui va déterminer sa signification. Son libellé pourra être utilisé comme une primitive logique s'il a été défini en respectant des principes différentiels. Et c'est l'ensemble des prescriptions interprétatives données par ces principes qui définissent l'engagement sémantique à la base de l'ontologie. Or cet engagement ne peut être universel. Il n'est valable que localement, « régionalement », dans le cadre du domaine et de la tâche qui ont permis de définir les textes choisis pour disposer d'une image des usages réels des unités linguistiques.

À la suite des travaux de B. Bachimont, les membres du groupe TIA ont défendu que les ontologies construites par l'IC et utiles pour des systèmes d'information à base de connaissances sont des ontologies régionales. Je reviendrai sur la portée de cette affirmation plus tard dans ce chapitre.

2.3.1.4 Textes, terminologies et modèles de connaissances : une convergence pluri-disciplinaire

Les premiers travaux sur les ontologies menés en intelligence artificielle et en IC couvrent les aspects formels de la représentation des connaissances et s'intéressent beaucoup moins à la définition des contenus de ces modèles. On s'attend à ce que la puissance et la qualité des langages de représentation, assurant une unicité des définitions et une certaine interopérabilité, viennent résoudre les ambiguïtés des notions sous-jacentes et réduire la difficulté à bien définir des concepts. La formalisation est supposée garantir la validité des interprétations possibles.

Or à la même époque, c'est dans d'autres domaines, ceux de l'ontologie formelle (Smith, 1998), de la terminologie (Skuce et Meyer, 1992), de la sémantique formelle et des réseaux sémantiques (Sowa, 1991) que les questions du sens et de sa représentation, en lien avec la langue qui permet de l'exprimer, sont posées. Ces interrogations conduisent chacun des domaines à renouveler ses approches ou à pousser plus loin ses questionnements.

Un exemple de renouvellement concerne la représentation des connaissances. Les limites dues au manque de précision des réseaux sémantiques, comme la polysémie des étiquettes de relations ou le manque de repères pour structurer la définition des concepts ont donné lieu à des réflexions théoriques sur la représentation des relations (Woods, 1975) et des concepts (Brachman, 1977), leur interprétation formelle et humaine. Pour approfondir le lien historique fondamental entre ontologies et réseaux sémantiques, on peut consulter l'habilitation de B. Biébow (Biébow, 2004).

Un autre renouvellement est celui qui a touché la terminologie en tant que discipline. Je détaille cette évolution, également décrite dans (Condamines, 2003) et (Biébow, 2004), parce que le regard posé sur les textes par les informaticiens construisant des ontologies reprend celui de la terminologie. La terminologie vise à étudier, inventorier et décrire sous forme de fiches les termes d'un domaine, avec une visée normalisatrice. Ses fondements, posés par Wüster dans les années trente dans la lignée du courant positiviste, affirmaient que la connaissance scientifique était basée sur le raisonnement logique, et proposaient le terme comme unité minimale de cette connaissance. Ce postulat entraîne deux affirmations liées au statut du terme : l'unicité de son interprétation (de son sens) au sein d'un domaine et de la manière de le fixer de façon « définitive » ; la possibilité claire de délimiter des domaines complémentaires et disjoints, dont la réunion couvrirait la diversité des champs scientifiques. Or la plupart de ces principes ont été mis en défaut par la pratique des terminologues : même au sein d'un domaine, on trouve des termes polysémiques ; leur sens ne fait pas toujours consensus et cette normalisation relève du choix du terminologue et, de plus, elle peut être remise en question régulièrement ; les domaines n'ont pas de frontières très claires et définitives dans l'absolu ; enfin, l'usage d'une terminologie donnée influence fortement la manière d'en déterminer le contenu (Slodzian, 1995). Finalement, les rares objets d'étude tangibles et bien identifiés sont les textes où les termes sont utilisés. Le domaine a donc évolué vers une *terminologie textuelle* (Bourigault et Slodzian, 1999) Les limites d'un domaine sont alors délimitées par le corpus des textes étudiés et l'usage prévu du produit terminologique. Le terme et

sa description sont alors le fruit d'une analyse de l'ensemble de ses usages en corpus, guidée par la pertinence par rapport au corpus et à l'application. En fait, la terminologie, en tant que base de données, n'a pas de raison d'être en soi : elle est déterminée par ses usages, et l'on parlera de *ressource* plus que de *produit* terminologique.

La question du sens est également au cœur de certaines approches du traitement automatique des langues. Ce domaine de l'informatique comporte des facettes très variées, parmi lesquelles l'analyse du langage écrit rejoint la problématique de la recherche de « connaissances » dans des textes. Après 1990, les systèmes ou algorithmes proposés sont désormais compatibles avec les besoins de l'IC car la manière d'aborder l'analyse du langage a, elle aussi, évolué. En effet, différentes expériences ont montré l'écueil de développer des approches exhaustives, visant des analyses complètes, à tous les niveaux de description du langage, et des interprétations sémantiques du langage écrit. La notion de traitements de surface, partiels et robustes a permis de mettre au point des logiciels moins ambitieux, n'abordant qu'une partie des phénomènes linguistiques, combinant des aspects linguistiques et statistiques, pour produire des vues sur les textes analysés (Daille, 1994). Ces logiciels, parvenus à une certaine maturité, s'avèrent performants et adaptés à la problématique de la modélisation conceptuelle dans la mesure où certains ont le souci de faciliter l'exploration des données ainsi tirées des textes (concordanciers, KWIC, extracteurs de termes, …). L'idée d'interpréter des textes pour la seule visée de compréhension du langage étant remise en question (Reimer, 1990) (Hahn *et al.,* 2002), les chercheurs développant ces logiciels ont cherché de nouveaux terrains d'expérimentations et d'applications possibles de ces traitements.

La convergence de ces différentes disciplines avec l'IC a été possible car, à leur tour, leurs recherches se sont tournées vers l'intelligence artificielle afin d'y trouver (ou de lui proposer) des formalismes logiques pour raisonner (cas des réseaux sémantiques), des structures de données pour gérer de gros volumes de données (cas de la terminologie), ou des terrains d'application. Ainsi, à la frontière entre informatique et terminologie, la notion de base de connaissances terminologiques a permis de proposer une structure de données suffisamment riche et souple pour rendre compte des liens complexes entre les termes d'un domaine, les textes où ils sont utilisés et des concepts informatiques.

2.3.2 Ingénierie des connaissances et documents

Même si l'intérêt de l'ingénierie des connaissances pour les documents, essentiellement les documents électroniques, est surtout visible au sujet du Web (sémantique) et des ontologies, il va bien au-delà.

2.3.2.1 Textes et documents en ingénierie des connaissances

L'IC s'intéresse aux documents comme porteurs de sens et révélateurs de connaissances depuis les premières études sur l'acquisition des connaissances pour les systèmes experts (années 90). Il s'agissait déjà de repérer des connaissances heuristiques, de rendre compte de raisonnements explicités plus ou moins dans des documents existants ou élaborés pour l'étude (retranscriptions d'entretiens). Dans cette perspective, les documents ne sont pris en compte que pour leur contenu. Ils sont considérés comme une source de connaissance complémentaire ou alternative aux experts et spécialistes du domaine. Le document est l'objet d'étude dans un premier temps, mais ensuite, il est abandonné au profit du modèle qu'il a permis de construire. Ce modèle sera intégré dans une application (éventuellement pour revenir au document).

Dans la perspective de la gestion des connaissances, c'est le document en tant que tel qui est central, il est un élément support de connaissance à part entière. Il peut s'agir alors de textes mais aussi d'images, de vidéo, etc. La gestion des documents produits et utilisés au sein de l'activité individuelle et collective étudiée, mais aussi, en tant que telle, la gestion de fonds documentaires (images, sons, vidéos) intéresse alors l'ingénierie des connaissances. Ces applications font appel

aux technologies relevant de la gestion documentaire et permettant le partage, la diffusion, l'archivage, l'indexation, la structuration ou la classification de documents ou de flux de documents. Ces technologies sont propres à la nature des supports, et les problématiques diffèrent un peu selon que l'on traite des images, du son, des vidéos ou des textes. La difficulté est d'appliquer ces solutions technologiques aux bons documents de manière à répondre au mieux aux besoins des utilisateurs, et qu'ils y trouvent les supports (entre autres les connaissances) utiles à la réalisation de leurs tâches.

Parce que de plus en plus de projets d'IC intègrent la gestion de documents sous des formes très variées, les chercheurs du domaine ne peuvent s'affranchir d'une réflexion approfondie sur la notion de document, et particulièrement de document numérique. Ainsi, plusieurs chercheurs contribuent aux travaux du réseau thématique pluridisciplinaire sur le document (RTP-DOC) et à ses productions (Pédauque, 2003) (Pédauque, 2005).

Dans la suite de ce mémoire, je me focaliserai essentiellement sur le premier point de vue. Dans ce cas, le document (on parle de "texte") est exploré, manuellement ou à l'aide de logiciels de traitement automatique des langues, pour y repérer des éléments utiles à la construction d'un modèle conceptuel. Le document n'est perçu que par son contenu, au détriment de la sémantique que peut porter sa mise en forme, son statut dans l'organisation qui l'utilise ou son historique par exemple. Il fait l'objet d'une analyse souvent parcellaire, "microscopique" (études au niveau de la phrase ou du paragraphe) et morcelée en unités de lecture qui minimisent l'intérêt de l'unité et de la structuration de l'ensemble. Les nouveaux outils de TAL au service de ce type d'étude essaient justement d'aller au-delà d'une analyse locale, de faciliter les recoupements à travers l'ensemble du document, de retrouver par exemple tous les contextes d'usage de syntagme pour en appréhender un sens global. Néanmoins, le travail d'analyse consiste, à partir de l'observation de fragments de texte, à les interpréter, les sélectionner ou les rejeter, à y identifier des connaissances, à les représenter et les formaliser. Ce jugement est tout autre que l'assignation directe d'une valeur de vérité à une expression en langue pour la traduire en une formule logique. Il consiste à évaluer une pertinence et à faire des choix de structurations des connaissances à partir de plusieurs analyses. Il s'agit de pondérer la confiance que l'on fait au fragment de texte vis-à-vis du document, au document par rapport au modèle à construire, et enfin de mesurer l'intérêt des connaissances identifiées par rapport à ce modèle. Des experts du domaine sont sollicités pour valider ou corriger ces choix.

Ainsi, la perspective du modèle à construire, de la mise en forme du raisonnement en lien avec une expertise, prime sur la fidélité au texte ou sur le souci d'en rendre compte avec précision.

2.3.2.2 Gestion des connaissances et construction d'ontologies

Avant même le vaste programme du Web sémantique, les recherches sur les ontologies et sur la gestion des connaissances sont venues bousculer de plusieurs manières ce rapport de l'IC aux documents. D'une certaine manière, il en est devenu plus précis et plus complet. Les évolutions mentionnées ci-dessous le sont à titre d'exemple et ne se veulent pas exhaustives.

- Ainsi, dès que l'on aborde la gestion des connaissances, le statut des différents documents dans l'organisation, leur rôle, leur circulation et leur histoire sont analysés finement pour déterminer ceux qui peuvent favoriser les échanges de connaissances, pour définir les organisations ou les logiciels facilitant leur conception et leur diffusion, etc. La matérialité des documents est prise en compte, ainsi que leurs auteurs, objectifs, lecteurs ciblés, les différents modes d'accès ou de lecture à prévoir, ce qui replace le document à la fois dans sa dimension matérielle et dans sa dimension sociale.

- Un autre impact de travaux en gestion des connaissances concerne la manière de rendre compte des modèles de raisonnement et de connaissances. L'échec des systèmes de résolution de problèmes indépendants (systèmes experts) a obligé d'inventer de nouvelles manières de rendre compte des savoir-faire au service de la réalisation d'une tâche. La proposition globale est de fournir un système d'aide à l'opérateur (et non qui se substitue à lui), avec des propositions variées

sur la nature de cette aide : réalisation automatisée de certaines tâches, d'autres étant laissées à l'opérateur, guidage, présentation d'objectifs ou de méthode de résolution, systèmes hypertextuels de navigation dans des connaissances rédigées et structurées. Le document, numérisé, structuré et organisé pour différents modes de lecture en lien avec la tâche, est intégré dans l'application non plus comme une source de connaissances mais comme un moyen de se les approprier après qu'elles aient été modélisées, structurées et rédigées.

- D'autres exemples de renouvellement du lien IC et documents peuvent être pris dans les travaux sur les ontologies. Les ontologies ont été définies comme des représentations d'un domaine, accentuant la dissociation (parfois provisoire) entre le raisonnement heuristique et la description des concepts manipulés par ces heuristiques. Les ontologies se focalisent donc sur l'essence d'un domaine (comme la médecine, ou un champ de la médecine par exemple), sur son vocabulaire et, au-delà, sur le sens dont il est porteur. Le rapprochement a été fait rapidement avec les thésaurus ou les terminologies pour mieux marquer les différences et les apports des uns et des autres.

Cette confrontation dépasse la simple comparaison de structures de données pour revenir aux modes de construction et aux usages qui sont faits de ces représentations. Les méthodes de l'IC se sont alors enrichies d'échanges avec les terminologues, lexicographes ou linguistes de corpus, ou spécialistes du TAL. Le regard sur les textes s'est également affiné par la mise en oeuvre d'approches syntaxiques et sémantiques.

2.3.2.3 Une dernière étape : le web sémantique

Enfin, l'état d'avancement actuel des propositions du W3C au sujet du Web Sémantique accorde une place privilégiée aux ontologies. Ceci augmente ainsi artificiellement les attentes à leur égard, concernant le potentiel de leur utilisation par différents types d'applications comme les services web, les agents logiciels ou la recherche d'information. Malgré ce risque, l'enjeu est passionnant parce qu'il élargit encore et positivement le spectre des applications concernées par les connaissances en y ajoutant les applications classiques en recherche d'information. Les ontologies sont alors vues comme une structuration plus riche que les thesaurus ou les lexiques utilisés jusqu'ici car elles introduisent d'une part une dimension sémantique (le réseau conceptuel) et formelle (gérable par les applications informatiques) et d'autre part, dans certains cas, une dimension lexicale qui améliore les accès aux documents.

Si l'on voit les ontologies sous l'angle de leur utilisation pour le Web Sémantique, elles sont effectivement avant tout un réservoir à méta-données pour mieux caractériser le contenu des ressources du Web. L'IC aborde donc depuis peu seulement, et avec un bagage tout autre que celui des Sciences de l'Information ou de la recherche d'information, le problème des méta-données et des méta-langages. En cela, elle a effectivement intérêt à se rapprocher de ces disciplines, qui, par leurs acquis, lui rappellent que les questions qu'elle traite ont déjà été abordées dans un contexte différent mais comparable.

2. 4 - Bilan : la problématique de l'IC

Cette analyse du domaine au cours des vingt dernières années reflète notre point de vue sur ce qu'est l'ingénierie des connaissances. Cette vision comporte certes une dimension historique et renvoie à des positions différentes au cours du temps. Cependant, elle donne une cohérence à mes propositions qui sont présentées dans la suite de ce mémoire. Je tente de l'expliciter ici, sous forme d'une liste d'affirmations qui ont été ou seront argumentées par ailleurs.

L'IC est fondamentalement **une ingénierie à l'intersection d'autres disciplines scientifiques**, dont elle se nourrit pour créer ses propres modèles et méthodes. Je l'illustrerai par mes travaux à différentes périodes, qui ont fait appel à la psychologie cognitive (méthode et plate-forme MACAO), à l'ergonomie (MacaoII), puis à la terminologie, à la linguistique et au TAL (CAMELEON et TERMINAE) ou plus récemment, les sciences de l'information. En tant qu'ingénierie,

l'IC définit des concepts qui permettent d'outiller et de critiquer des solutions techniques et non de décrire des phénomènes. Cependant, ces solutions ont la particularité de donner lieu à une interprétation en termes de connaissances par des utilisateurs. Elle se définit donc comme la discipline qui se focalise sur la manière d'organiser ces connaissances au sein de modèles qui permettent de les restituer dans un environnement opérationnel.

L'IC **se nourrit de la capitalisation de retours d'expériences sur des problèmes réels**, et sa manière de poser la conception de systèmes à base de connaissances ne peut se satisfaire ni d'études de cas limitées ni de preuves formelles ou théoriques. Les expériences de mise à l'épreuve de propositions méthodologiques, de représentations ou d'outils d'analyse n'ont aucune des caractéristiques des expériences scientifiques, au sens où elles ne correspondent pas à une mise à l'épreuve d'hypothèses de modèles de connaissances sur le monde. Je renvoie ici à l'analyse faite par B. Bachimont (Bachimont, 2004), qui qualifie ces expériences d'expériences humaines dont la visée est de critiquer des propositions, des moyens d'instrumenter l'interprétation de sources de connaissances par des utilisateurs. J'ai multiplié ce genre d'expériences au cours de mes travaux, et je les mettrai en regard de mes contributions au fil des chapitres pour montrer les leçons et le regard critique que j'ai pu en tirer.

L'IC a vécu des changements d'orientation successifs, qui en font **un champ de recherche en devenir**, sans doute pas tout à fait parvenu à maturité, comme en témoignent les fluctuations terminologiques d'une part (cf. les analyses d'articles scientifiques du domaine exposées à IC 2000 et IC 2002). Un autre signe de ce caractère mouvant et en recherche se reflète à travers les changements réguliers des thèmes de recherche : après un engouement pour les méthodes de résolution de problème et l'expertise humaine, l'actualité accorde une place sans doute exagérée aux ontologies, même si ce type de structure de données pose le problème de la représentation des connaissances sous un angle nouveau qui mérite d'être exploré. L'IC est en constante évolution de l'intérieur (nouvelles analyses, nouvelles perspectives, manières originales de poser les problèmes ou nouveaux concepts théoriques) et de l'extérieur (les types d'applications ciblés ont changé au fil des années, avec des contextes d'usage qui se renouvellent, de nouvelles contributions d'autres disciplines viennent apporter des méthodes et des concepts nouveaux).

Ces évolutions élargissent progressivement le champ de l'IC : **les nouveaux cadres théoriques proposés englobent les travaux précédents** au sein de perspectives renouvelées. Même si certains changements de points de vue correspondent à des ruptures (comme le passage d'une vue cognitiviste à une vue constructiviste), les résultats historiques du domaine se complètent au fil du temps et peuvent être repris sous un nouvel angle. Mon parcours est tout à fait cohérent avec l'analyse historique de l'ensemble du domaine que j'ai présentée dans la première partie. Ainsi, les techniques proposées pour prendre en compte l'expertise humaine viennent compléter et valider l'observation de l'activité, ou encore les résultats des analyses des textes. Au-delà de leur complémentarité en tant que sources de connaissances, l'association d'éléments liés à l'activité et d'autres à l'expression des connaissances dans la langue, au sein d'un modèle co-construit par les acteurs du domaine, est un moyen d'accroître son acceptabilité et la possibilité de se l'approprier une fois qu'il sera opérationnel. Les différentes propositions que j'ai pu faire ont encore du sens aujourd'hui si l'on retient la définition de l'IC ci-dessous :

L'IC a pour objectif *de définir les moyens de construire des modèles et des systèmes (artefacts informatiques) qui mettent à disposition d'utilisateurs, sous la forme la mieux adaptée à la réalisation d'une tâche, des connaissances pertinentes qu'ils sauront interpréter et utiliser comme instrument au travail intellectuel au sein d'une organisation ou d'un collectif.* Il s'agit d'une activité constructive où l'enjeu est de modéliser les connaissances à un niveau d'abstraction adapté, qui fasse sens pour les acteurs impliqués (cogniticiens, experts métiers, utilisateurs, etc.), qui leur permette de s'approprier le comportement du système, les connaissances qu'il manipule ou présente, et d'interagir avec lui. Ceci peut supposer aussi de s'intéresser au fonctionnement cognitif de l'utilisateur ou du groupe d'utilisateurs dans leur relation au futur système. **L'évaluation de travaux en IC** est donc par nature empirique. Les démarches proposées sont valides si elles sont reproductibles, instrumentées techniquement (méthodes et outils), fondées sur des approches

rigoureuses et si elles conduisent à des applications utilisées. La validité des modèles construits passe donc par l'acceptabilité (au sens ergonomique) des applications qui les utilisent.

Les avancées de l'IC passent par la volonté de **mieux synthétiser les résultats** produits pour en identifier points forts et points faibles, en particulier par type d'application, et par **la capacité à les adapter à de nouvelles demandes** (complexité des nouveaux cadres applicatifs, multiplicité et complémentarité des nouveaux supports de connaissances). Le problème récurrent d'une ingénierie est en effet de gérer au mieux l'adéquation méthode-besoin dans un contexte technologique et scientifique mouvant et novateur.

CHAPITRE 3 - PROBLEMATIQUES DE RECHERCHE

Ce chapitre synthétique tient lieu d'articulation entre notre première analyse du domaine et l'exposé de mes travaux. Après l'introduction, je présente, dans une première partie, la ligne directrice de ces travaux, un ensemble de choix qui en définissent la cohérence. Je dresse ensuite dans une deuxième partie deux listes de questionnements et de choix associés, correspondant à deux périodes dans mes recherches : les premières questions portent sur la modélisation conceptuelle à partir de connaissances d'experts, les suivantes abordent les liens entre analyse de textes et modélisation. Le détail des travaux et contributions fera l'objet des chapitres 4, 5 et 6.

Le chapitre 2 a exposé un point de vue particulier sur la recherche en ingénierie des connaissances et l'évolution de sa problématique. Ce point de vue a mis en avant la place centrale des modèles conceptuels, tout d'abord définis comme des spécifications du comportement attendu de l'application finale. Or le domaine a évolué vers une problématique de restitution des connaissances plus que d'opérationnalisation. Une illustration particulière concerne les applications en lien avec des documents, où les modèles définis sont des ontologies souvent construites par analyse de textes.

En effet, un des enjeux de la modélisation conceptuelle est de parvenir à mettre en forme des représentations issues de et donnant accès à des connaissances ou permettant de les reconstruire. La complémentarité entre sources de connaissances existantes, modèles conceptuels et utilisateurs doit se retrouver dans l'application finale et être favorisée par les interactions prévues entre l'utilisateur et le système. Le modèle peut alors servir à fournir des connaissances sans pour autant « raisonner ».

Dans ce chapitre 3, je me focalise sur les questions à l'origine de mes recherches avant d'en présenter les résultats dans les chapitres suivants.

Le fil conducteur de mes travaux est de fournir des aides pour assurer la construction de modèles conceptuels, et ce selon des approches ascendantes. J'entends par là une analyse précise de ce que sont les besoins en connaissances et les traces liées à leur utilisation, y compris dans la langue écrite. Mes questions et les choix retenus pour y répondre ont en commun la volonté de se démarquer d'une démarche cognitiviste, mais de s'appuyer sur l'analyse des connaissances en usage pour construire des modèles spécifiques et adaptés à chaque situation. Une autre caractéristique est de s'intéresser au processus de modélisation dans sa globalité. De ce fait, les résultats produits comprennent des langages de modélisation, des techniques et des logiciels de recueil, d'analyse et de modélisation ainsi que des propositions méthodologiques.

En revanche, cette recherche a été déclinée sous forme de questions différentes selon les catégories d'applications visées. Ainsi, mes premières contributions portent sur la modélisation conceptuelle dans son ensemble, à partir de l'étude d'activités expertes. Je les ai recentrées ensuite

sur la modélisation du domaine et les ontologies, et leur construction à partir de textes. Cette évolution correspond aussi à une ouverture dans la mesure où les applications ciblées couvrent un spectre bien plus large que les systèmes à base de connaissances, et où les questions soulevées concernent l'articulation langue-connaissances.

Dans les deux cas, ces questions situent le cœur de la modélisation dans la prise en compte de trois contraintes parfois contradictoires : les besoins des utilisateurs de l'application ciblée, les connaissances telles qu'elles sont perçues dans les sources étudiées (activités ou entretiens de spécialistes du domaine, textes) et enfin les normes de « bonne structuration » des modèles. Ces questions précisent la problématique d'ensemble : comment guider le cogniticien chargé de construire un modèle en tenant compte de ces trois contraintes ? quels outils lui fournir pour accélérer son travail, augmenter la qualité des modèles ? comment agencer, combiner de tels outils au niveau matériel (dans une plate-forme) et méthodologique ? quels principes organisateurs retenir pour organiser le modèle ? sous quelle forme le représenter ? Comment produire des résultats plus contrôlés, dont on maîtrise la portée applicative et l'apport au domaine.

3. 1 - Principes généraux retenus

Pour aborder la modélisation des connaissances, j'ai retenu une approche constructiviste, où les priorités sont à la fois la construction de modèles assurant une bonne réponse aux besoins et l'ancrage de ces modèles dans les connaissances en usage. Dans l'esprit des travaux sur la modélisation conceptuelle, je découple modélisation et opérationnalisation, en référence à l'hypothèse du « knowledge level » de Newell. Ces modèles doivent servir de support au dialogue entre acteurs de la modélisation avant d'être le résultat de ce processus. Afin de mieux intégrer les usages des connaissances, je propose de développer ces modèles selon une démarche ascendante, qui débouche sur des représentations spécifiques et utiles plus que réutilisables. Enfin, mes recherches et propositions prennent en compte le processus de modélisation dans son ensemble, depuis l'identification de sources de connaissances jusqu'à la mise à disposition du modèle dans une application. Pour intégrer cette complexité et les diverses dimensions de la modélisation, j'ai visé la définition et la construction de plates-formes qui intègrent des techniques et des logiciels de recueil et d'analyse, des représentations des connaissances et un cadre méthodologique. J'argumente les raisons de ces choix dans la suite de ce paragraphe.

D'autres principes croisent les précédents ou en découlent. Je les justifie au fil de l'exposé des premiers :

- se référer à des travaux d'autres disciplines (psychologie cognitive, ergonomie, linguistique et terminologie, traitement automatique des langues) ;
- exploiter différentes sources de connaissances : expertise individuelle, activités liées aux tâches mais aussi textes du domaine ;
- multiplier les expériences en « vraie grandeur » dans le cadre de diverses applications dans différents domaines, si possible en collaboration avec des entreprises.

3.1.1 Approche constructiviste

Je situe mon approche parmi les démarches dites « constructivistes ». Par cela, j'entends que le modèle ne se veut pas un reflet d'une réalité mais un artefact, une construction originale utile comme support pour rendre accessibles des connaissances dans une situation donnée.

Préciser cette position pourrait sembler trivial et ne plus être d'actualité, car les chercheurs en ingénierie des connaissances affichent désormais à l'unanimité ce même choix. Les modèles conceptuels dont je parle, que je construis, ne cherchent pas à être fidèles à des modèles de connaissances individuelles ou collectives. Ils sont des constructions artificielles, des modèles des

futurs systèmes à concevoir ou des modèles utilisés par ces systèmes. Ils contiennent différents types de représentation, qui vont de représentations formelles interprétables par un système informatique à des fichiers d'image ou de texte qui seront interprétés par l'utilisateur. Or ce choix me semble important à plusieurs titres.

Tout d'abord, cette vue constructiviste a mis du temps à s'imposer (l'acquisition des connaissances n'a pris ce tournant que dans les années 90). Mes premiers travaux avaient une orientation cognitiviste. Il s'agissait de bien connaître la nature des connaissances expertes pour les représenter au mieux, au plus près de ce qu'elles sont, avant de les opérationnaliser.

Ensuite, dans le cadre des sciences cognitives, de nombreuses recherches continuent d'être réalisées avec une finalité d'étude et de simulation des processus cognitifs humains, individuels ou collectifs. Ces travaux utilisent parfois les mêmes concepts (ceux de modèle conceptuel, d'ontologie ou encore de base de connaissances) avec une perspective tout à fait différente : le système visé sert à appuyer et évaluer des théories sur les raisonnements, les prises de décision ou d'autres types de résolution de problèmes dans différentes situations, et non à assister un opérateur, un « utilisateur ».

Enfin, ce choix peut sembler paradoxal alors que j'insiste sur mon souci de répondre aux besoins spécifiques des utilisateurs par une analyse ascendante des activités, des entretiens d'experts ou des textes du domaine. Le modèle est construit à partir des usages pour mieux être utilisé, mais il n'est en aucun cas un reflet des processus cognitifs ou des modèles mentaux qui conduisent à ces usages ou encore des comportements collectifs qui en rendent compte.

Soulignons que le mot *constructiviste* prend un autre sens en linguistique et terminologie. Le point de vue constructiviste suppose que le sens des mots est construit en usage, et peut donc évoluer ou être reconstruit dans de nouveaux contextes. Il s'oppose à un point de vue normatif et structuraliste, qui fait l'hypothèse que le sens des mots leur est propre et intrinsèque, et peut donc être analysé une fois pour toutes. La construction d'ontologies universelles et d'ontologies formelles au sens de la philosophie répond à l'hypothèse structuraliste en offrant une représentation des connaissances qui rend compte du sens des mots et de leurs relations sémantiques.

3.1.2 Répondre aux besoins par des modèles spécifiques

J'ai choisi d'aborder la modélisation systématiquement de manière *ascendante* : mes propositions s'appuient sur une étude des traces de connaissances que produisent les acteurs du domaine, et non sur une idée a priori de ce que doit être le modèle visé. Les modèles ainsi obtenus sont *spécifiques* et répondent en priorité à des besoins et à des utilisations particulières. Leur mise au point fait intervenir les acteurs concernés et éventuellement l'adaptation de modèles déjà existants ou prédéfinis. Je fais le choix de *partir des usages pour revenir vers les utilisateurs*. Le contenu des modèles dépend de l'étude des besoins, il est lié à l'analyse des activités des spécialistes ou des usages terminologiques que révèlent leurs écrits.

Cette approche se démarque de méthodes qui privilégient les capacités de résolution de problème et orientent la modélisation des connaissances en fonction de la méthode de résolution que le système opérationnel va mettre en œuvre sur les connaissances. De ce fait, mes choix complètent sans les rejeter des approches par réutilisation de composants génériques. En revanche, ils sont orthogonaux à celles qui visent la construction de modèles génériques et universels (comme les ontologies génériques). J'ai donc peu abordé le problème de la réutilisation, bien que ce soit un axe important dans le domaine et une solution pour réduire les coûts de construction de systèmes.

Ce besoin est particulièrement fort dans des domaines ou pour des applications qui cherchent à transmettre des pratiques, des connaissances ou des savoir-faire précis auprès d'utilisateurs qui ainsi se forment, apprennent ou suivent des procédures précises. Dans ce cas, le modèle sert autant à organiser, à mettre en forme un support de connaissances qu'à définir un système opérationnel. Le modèle doit pouvoir être consulté et faire sens auprès de ceux qui détiennent des connaissances

puis auprès de ceux qui vont l'utiliser. C'est le cas par exemple des logiciels éducatifs, où des élèves et des enseignants doivent accéder à des connaissances et les assimiler par le biais du modèle (Tchounikine, 1998). Les critères de pertinence d'un modèle sont alors à la fois d'ordre cognitif et pédagogique. On ne cherche pas à optimiser la manière de résoudre le problème, on cherche à transmettre des connaissances en respectant ce que l'enseignant veut que l'élève perçoive du domaine. Et pour cela, l'enseignant doit pouvoir prendre part à la construction, comprendre, manipuler et critiquer le modèle et ce qui en sera fait dans le système final. Le même type de choix est fait dans des applications à la médecine qui visent à transmettre des guides de bonne pratique (Charlet, 2002). Le modèle, par exemple une ontologie de la spécialité concernée, est alors un support à l'exploration du guide par l'utilisateur. Il doit à la fois respecter la logique des pratiques conseillées par les spécialistes, et donner à l'utilisateur médecin la possibilité de comprendre, intégrer ou adapter les pratiques conseillées.

Étudier la définition de modèles spécifiques n'est pas le seul objectif de l'IC, mais c'est une visée intéressante car également présente, dans une moindre mesure, dans des approches basées sur la réutilisation ou visant une grande généricité. La réutilisation suppose une adaptation en fonction des particularités des connaissances en jeu (on revient donc à une démarche d'étude – minimale – de cette situation). La généricité d'un modèle passe par l'abstraction de situations plus spécifiques ou la caractérisation dans l'absolu de connaissances particulières en fonction de classes ou de modèles prédéfinis, ce qui suppose encore une étude de situations spécifiques.

3.1.3 S'appuyer sur les connaissances en usage

Pour réaliser des modèles de connaissances, je privilégie l'étude des connaissances telles qu'elles sont utilisées. Les usages sont pris en compte à deux niveaux. D'une part, les situations dans lesquelles sont mises en œuvre des connaissances ainsi que les documents associés constituent des supports et des traces des connaissances qui sont analysés pour construire les modèles. D'autre part, les usages renvoient aux pratiques des futurs utilisateurs, aux besoins formulés relativement à l'application à construire. Il s'agit là d'une deuxième contrainte prise en compte pour définir le modèle conceptuel. Enfin, les connaissances effectivement utilisées sont parfois propres aux acteurs du domaine, et différentes de celles que pourraient fournir des manuels didactiques, des experts ou une théorie du domaine. Par exemple, les travaux sur corpus dans le domaine du droit conduisent à des terminologies et à des ontologies prenant en compte la jurisprudence et des éléments de sens commun (Lame, 2002). Elles se démarquent d'ontologies génériques composées par les théoriciens du domaine (Breuker *et a*l., 2004).

Le modèle construit en prenant en compte ces deux types d'usage des connaissances ne prétend pas pour autant refléter les représentations mentales ou les processus cognitifs des experts ou des utilisateurs du domaine. Il s'agit d'un modèle « artificiel », selon le point de vue constructiviste mis en avant plus haut. L'analyste qui construit le modèle prend donc une distance nécessaire par rapport aux traces de connaissances qu'il étudie. C'est en fonction de sa propre interprétation qu'il peut donner du sens à des éléments qui ont souvent trait à la forme, à la régularité de certaines actions, de gestes, à l'utilisation de termes ou de tournures syntaxiques. C'est aussi lui qui décide comment élaborer ou figer ce sens dans un modèle. Enfin, le processus de modélisation doit aussi prévoir comment restituer ce sens auprès des utilisateurs et lui donner la capacité d'interpréter correctement le fonctionnement et les informations présentées par le système. Notre approche de la modélisation consiste donc à définir les moyens d'accéder aux connaissances en usage, de les analyser et de construire des modèles à partir d'une interprétation raisonnée en fonction de l'application ciblée.

La prise en compte des usages répond à des motivations à la fois pragmatiques et théoriques.

D'un point de vue pratique, le critère de validation des travaux de l'IC est bien l'efficience des systèmes. Le modèle à construire est celui d'un système, qui doit *servir à ses utilisateurs* à produire, à travers une interface, des interactions, des textes ou des propositions d'action qui font

sens pour lui. L'utilisateur doit pouvoir adhérer aux propositions ou réalisations du système, raisonner sur ces éléments, utiliser ses connaissances et le système pour en produire de nouvelles utiles à sa tâche dans son contexte de travail.

D'un point de vue plus théorique, en m'intéressant aux connaissances en usage, j'ai intégré les préoccupations de l'ergonomie d'une part et de la linguistique de corpus d'autre part.

Pour mieux connaître et décrire les situations initiales des utilisateurs et des spécialistes, et ainsi anticiper les solutions informatiques à mettre en place, j'ai retenu les méthodes et résultats de l'ergonomie. Au-delà d'un emprunt de techniques, j'ai identifié un objectif commun : analyser des situations de travail, de résolution de problème ou de prise de décision, pour construire des applications informatiques qui permettront de les assister. En collaborant avec des chercheurs de ce domaine, j'ai défini des techniques et éléments méthodologiques d'étude des connaissances en usage. En effet, l'ergonomie accorde une place importante à l'étude de l'activité comme préalable à la mise en place de systèmes d'aide à l'opérateur, de nouveaux services ou à l'amélioration de l'organisation des situations de travail. L'analyse de l'activité dépasse une simple observation qui serait neutre et exhaustive. Elle s'appuie sur des principes et des théories qui vont orienter ou éclairer les observations afin de favoriser des actions améliorant la capacité à réaliser cette tâche (Spérandio, 1996). Elle dresse un diagnostic et débouche éventuellement sur des recommandations ou anticipe les conséquences des choix de conception.

De même, une collaboration avec des linguistiques travaillant sur corpus m'a semblé indispensable pour étudier la construction de modèles à partir de textes. En effet, la particularité de la linguistique de corpus, par contraste avec les pratiques classiques en linguistique (basées sur l'introspection), est justement d'accorder une place essentielle à l'usage de la langue dans les écrits ou les oraux retranscrits (Condamines, 2003). Cette approche renouvelle la sémantique : ce sont les usages en corpus qui contribuent alors à la caractérisation sémantique, en soulignent la variabilité et la difficulté de la fixer a priori. Le travail sur corpus oblige à ne plus ignorer l'influence sur la sémantique d'éléments pragmatiques comme la mise en forme, le genre textuel, le contexte de lecture ou encore d'éléments sociologiques dont certains courants fondateurs de la linguistique ont cherché à s'affranchir. Il suppose de définir des outils informatiques de validation d'hypothèses linguistiques pour explorer les corpus sous forme numérique. Ces outils sont par exemple l'étude de cooccurrences de termes ou encore des patrons de fouille basés sur la morphologie, la syntaxe ou le lexique. Je les ai détournés de leur objectif linguistique par la modélisation conceptuelle, et j'ai évalué leur aide pour retrouver dans les textes des informations particulières, à travers les contextes d'utilisation des termes.

Au-delà des outils, j'ai voulu mener une réflexion commune avec la linguistique de corpus et la terminologie sur le statut de la langue comme révélateur de connaissances, à travers l'articulation entre termes et concepts par exemple. En effet, la plupart des travaux sur les ontologies ont un caractère normatif. Elles se situent implicitement dans une lignée constructiviste et normalisatrice de la sémantique qui, même si elle a permis des avancées, a montré ses limites. Le point de vue constructiviste ignore les divergences de points de vue sur un domaine, ne rend pas compte des phénomènes de glissement de sens au cours du temps ou de variations entre la langue générale et celle utilisée dans des domaines techniques. Or la linguistique de corpus et la terminologie textuelle vont à l'encontre de cette vue. En s'appuyant sur les usages linguistiques attestés par les pratiques et les textes techniques, elles remettent en question l'existence a priori de concepts associés à des termes. Elles soulignent le rôle essentiel de l'interprétant dans l'analyse de manifestations linguistiques et dans la sélection de termes qui permettent de définir les concepts a posteriori, en fonction des objectifs visés. J'ai trouvé là un écho particulier à mon expérience de l'analyse de texte et mes hypothèses sur le rôle d'interprétation et de construction de l'analyste.

3.1.4 Prendre en compte la démarche d'ingénierie des connaissances dans sa globalité

En tant qu'ingénierie, l'ingénierie des connaissances doit apporter des solutions pratiques et adaptées aux situations réelles d'utilisation des systèmes qu'elle permet de construire. De ce point de vue, ces solutions n'ont d'intérêt que si elles prennent en compte la complexité et la globalité des contextes d'utilisation des applications. De plus, même si elles sont étudiées séparément, elles doivent s'intégrer de manière cohérente pour couvrir l'ensemble du processus de modélisation, depuis l'identification de sources de connaissances jusqu'à la mise à disposition du modèle dans une application. La complexité des situations réelles et la diversité de la modélisation peuvent être prises en compte grâce aux choix suivants :

- diversifier les techniques et outils proposés pour pourvoir aborder des types différents de situations, d'applications, d'expertises ou de sources de connaissances ;
- proposer plusieurs types d'aides complémentaires pour les différentes tâches du cogniticien au cours de la modélisation : méthodes, langages de modélisation des connaissances, techniques et outils de recueil et d'analyse ;
- tenir compte d'emblée de la complémentarité des aides proposées et les intégrer de manière cohérente au sein d'ateliers ;
- couvrir le cycle de modélisation, depuis l'étude des besoins jusqu'à la conception d'une application ; toutefois, ce sont les phases initiales de la modélisation que j'ai le plus étudiées ;
- évaluer ces propositions via plusieurs projets et expériences à partir de besoins réels dans des situations de travail.

Ces choix sont motivés aussi par la volonté de situer différentes aides à la modélisation les unes par rapport aux autres dans l'ensemble du processus, d'en mesurer l'intérêt et l'impact. De ce fait, ma contribution trouve son originalité dans la *cohérence* de l'ensemble plus que dans chacune des aides (méthode, outils ou représentation de connaissances) prise séparément. Une autre particularité est l'attention portée à *la place et la coordination des acteurs* concernés, c'est-à-dire d'un côté les personnes qui détiennent les connaissances (experts, spécialistes ou auteurs des textes) et d'un autre les utilisateurs et leurs interlocuteurs au sein de l'organisation. Ce sont eux qui révèlent les connaissances en usage ou qui contribuent à la définition des spécifications. Donner toute leur place à ces acteurs confirme la nécessité de poser le problème non pas comme un défi technologique relevant seulement de l'informatique ou de la logique, mais bien comme un problème à traiter en collaboration avec d'autres disciplines (psychologie cognitive, ergonomie, terminologie, linguistique et TAL).

La complexité provient entre autres de la difficulté à accéder aux connaissances, de la *diversité de leur nature*, de la variété des sources et de leur complémentarité. Pour cela, mes recherches ont traité successivement de connaissances expertes ou spécialisées et de leur accès selon des approches de type ergonomique, puis de connaissances présentes dans des textes techniques et de leur accès par l'analyse du langage naturel. La complémentarité entre ces deux sources de connaissances va bien au-delà de la validation ou de la spécification de l'un par l'autre. Chacun produit des connaissances de nature et de validité différentes, et conduit à des modèles adaptés à des usages particuliers.

3. 2 - Questions et orientations de recherche

3.2.1 Recueil et modélisation de connaissances expertes

Comme je l'ai retracé au chapitre 2, l'ingénierie des connaissances s'est débord fixé comme objectif la définition de méthodes et d'outils pour définir et représenter les connaissances nécessaires à un système à base de connaissances. Ce premier objectif a été reformulé comme celui du recueil et de la modélisation de connaissances d'experts. Mes recherches ont commencé à cette période et se sont naturellement intégrées dans cette problématique. Elles ont d'abord porté sur des techniques et des logiciels pour recueillir, organiser et représenter des connaissances d'experts et des savoir-faire.

J'ai abandonné le point de vue de l'« acquisition des connaissances » par la suite, mon objectif n'étant plus la réalisation d'un système expert reproduisant exactement les raisonnements associés à ce savoir-faire. Cependant, dans la plupart des projets d'ingénierie des connaissances, il s'agit de rendre accessibles des connaissances jusque-là ni explicitées ni structurées et peu diffusées, qu'elles soient partagées par un collectif ou relevant du savoir-faire d'experts. Pour cela, les mêmes techniques et logiciels peuvent servir, appliqués à d'autres types de connaissance ou utilisés pour rechercher des types particuliers de connaissances. Elles permettent d'interroger des personnes et des situations de travail relativement à une activité particulière en vue de faire réaliser cette tâche par les mêmes personnes ou par de nouvelles, dans un contexte modifié, puisque leur activité sera assistée par l'utilisation d'un support informatique.

Ces objectifs m'ont conduite à considérer un ensemble de questions de recherche propres à la problématique du recueil et de la modélisation de connaissances d'experts, et à aborder ces questions en cohérence avec les principes généraux retenus. Je résume ci-dessous ces questions et l'orientation retenue pour mes recherches.

1) *Comment identifier et recueillir des connaissances expertes ?* Pour répondre à cette question, j'ai choisi de proposer une approche *ascendante* qui parte des traces de connaissances telles qu'elles sont recueillies auprès d'experts pour en dégager des parties de modèles. Ce choix s'oppose aux approches descendantes (top down) qui suggèrent de guider le recueil et la modélisation en s'appuyant sur des cadres interprétatifs fixés et génériques (comme les méthodes de résolution de problème de KADS ou de Mc Dermott). Mon choix m'a conduite à inventorier, définir ou adapter des techniques dites « ascendantes » (entretiens, simulations, études de cas, exercices de tri, de comparaisons, ou de mise en situation, etc.) pour évaluer leurs apports, limites et complémentarités au sein d'une approche de modélisation. J'ai poursuivi en approfondissant la maîtrise de ces techniques. En effet, elles sont indispensables lors des phases d'expression des besoins et de validation, pour gérer les entretiens avec des acteurs du domaine. Pour progresser dans leur mise en œuvre, j'ai voulu souligner l'impact du type de projet, de la nature du système visé et de l'état d'achèvement du modèle sur leur apport à la modélisation. J'ai mené ces études en *collaboration* avec des spécialistes de la psychologie cognitive et de l'ergonomie. Ces disciplines contribuent à une étude originale des connaissances et savoir-faire liés à la réalisation de tâches.

2) *Quel type de modèle utiliser pour faciliter explicitation, interprétation et structuration des connaissances ?* Parmi les différents rôles d'un modèle conceptuel, je privilégie celui de *médiateur* guidant l'explicitation et la structuration de connaissances, plutôt que l'intérêt du modèle pour préparer l'opérationnalisation. Ce choix met en avant le fait que le modèle matérialise le résultat de la modélisation, afin d'en critiquer le résultat et de le faire évoluer. Ainsi, sa structure et son contenu « guident » la recherche puis l'organisation de nouvelles connaissances. Ce modèle doit pouvoir être compris, interprété et validé par les acteurs de la modélisation. J'ai cherché à définir des primitives de représentation lisibles, qui favorisent les commentaires et la justification des choix de modélisation, qui

distinguent et articulent les différents types de connaissance et qui facilitent le travail du cogniticien (construction incrémentale du modèle, simplicité de mise à jour, …). De ce point de vue, une représentation des connaissances est jugée non pour son pouvoir d'expression et de raisonnement mais en tant qu'aide à la structuration, pour sa facilité d'utilisation et de visualisation, graphique par exemple. Une autre contrainte retenue est que le modèle représenté selon un formalisme puisse être opérationnalisé en respectant la même logique de modélisation.

3) *Comment construire des modèles s'inspirant d'activités expertes ?* Ayant choisi d'appuyer la modélisation sur une étude des usages, j'ai voulu construire des modèles tenant compte des activités des experts et des futurs utilisateurs, soit pour s'en inspirer, soit pour s'en démarquer. Le modèle définit, en lien avec cette activité, la manière dont le système va traiter la tâche qui lui revient. Pour assurer son adéquation aux usages, j'ai retenu l'approche ergonomique, basée sur les analyses de l'activité et de la tâche, et les représentations rendant compte de leurs résultats. Ces analyses constituent des techniques de recueil pertinentes et fournissent des représentations utiles pour amorcer un modèle de la résolution de problème non opérationnel. Pour cela, les résultats et approches de l'ergonomie doivent être adaptés à la modélisation conceptuelle. Une autre technique ergonomique retenue est l'analyse de l'activité des experts à travers des simulations et des études de cas, dont on peut conserver des représentations à titre d'exemples de résolution de problèmes.

4) *Comment un modèle peut-il favoriser la coopération homme-machine ?* Plusieurs recherches ont montré l'intérêt de langages associant de manière dynamique une méthode à chaque tâche en fonction du contexte d'utilisation, de manière à mieux coopérer avec l'utilisateur. Or ces langages se situent au niveau opérationnel. J'ai choisi de m'inspirer de ces travaux pour proposer des structures équivalentes pour la modélisation conceptuelle. Situées en amont des langages opérationnels, ces structures de modélisation de tâche et de méthodes doivent privilégier la facilité de mise au point et la documentation du modèle, avant d'être affinées pour définir des primitives d'un langage opérationnel associé. Elles doivent permettre de spécifier les différentes méthodes à proposer pour la réalisation d'une tâche ainsi que les conditions déterminant la mise en œuvre de chacune d'elles.

5) *Comment qualifier la nature des modèles ainsi construits ?* Dans le cadre de la définition d'une approche ascendante, une des difficultés est de situer la place des études de cas et des modèles de tâche par rapport au modèle du système à construire. J'ai choisi une première approche méthodologique qui prévoit des étapes pour définir le modèle du système à partir des études de cas ou du modèle de la tâche. Une autre direction expérimentée est de confronter les modèles ainsi construits aux modèles obtenus selon d'autres principes de construction, comme la réutilisation de composants génériques par exemple. L'idée est toujours d'affiner les recommandations méthodologiques et de fournir un logiciel d'aide adapté.

6) *Quelles propositions pour organiser le travail du cogniticien ? Quelle est leur portée ?* Je me suis fixé de produire des méthodes intégrant de manière cohérente différentes propositions : des techniques et outils de recueil ou d'analyse alimentant les modèles d'une part et des modèles et représentations de connaissances servant à les décrire ou à les opérationnaliser d'autre part. Cela suppose non seulement de définir des guides et des recommandations, mais aussi de les rendre accessibles et opératoires sur des supports adaptés (livres, guides méthodologiques ou encore ateliers logiciels).

3.2.2 Modèles terminologiques, ontologies et textes

Mon intérêt pour les textes se justifie d'abord par la nécessité de trouver une alternative aux entretiens avec les experts, trop coûteux. L'analyse des textes est devenue un des axes essentiels de

mon travail. La problématique soulevée est d'autant plus intéressante que je l'ai abordée dans le cadre de collaborations avec la linguistique et le traitement automatique des langues. Ce type d'approche s'avère particulièrement pertinent pour construire des ontologies, ce qui m'a conduite naturellement à me focaliser sur ce type de modèle et des modèles proches, comme les terminologies. En parallèle, la forte expansion du volume des documents numériques disponibles ont renforcé les enjeux de ces travaux. Ces recherches se trouvent aujourd'hui très sollicitées par les nouvelles approches de la gestion documentaire et de la recherche d'information.

Les textes, et en particulier les textes techniques, représentent un potentiel de connaissances qui peut permettre d'accélérer la modélisation d'un domaine tout en sollicitant moins les spécialistes du domaine. Pour mettre les connaissances contenues dans les textes à disposition des utilisateurs, deux types d'approches sont possibles. Les premières considèrent le texte comme un objet important à conserver et à présenter à l'utilisateur, ce qui est souvent le cas en gestion des connaissances. Pour la deuxième, le texte est analysé pour produire une autre représentation (un modèle conceptuel en IC, une base de données en data mining par exemple). Je me suis placée dans le deuxième courant, puisque mon objectif premier était d'élaborer un modèle conceptuel. Cependant, j'ai vite constaté que les modèles produits sont de meilleurs modèles pour revenir aux textes, pour les indexer ou les annoter par exemple.

En accord avec une approche d'ingénierie, je pose le problème de manière pragmatique, sans avoir l'ambition de proposer une théorie formelle de la sémantique par exemple. Construire un modèle, ce n'est pas représenter toutes les connaissances pouvant être tirées d'un texte ni permettre différents types de raisonnement sur ces connaissances. C'est identifier, à travers l'usage de la langue, des concepts et leurs descriptions, qui sont retenus en fonction du type de modèle à construire et de l'utilisation qui en est prévue. Ainsi, à travers cette problématique, je ne perds pas de vue la place et le rôle du modèle conceptuel dans un dispositif opérationnel auprès d'un utilisateur. Avec les ontologies et des applications documentaires, la part d'initiative laissée à l'utilisateur devient plus grande, il doit reconstruire les connaissances à mettre en oeuvre.

Je décline ici une série de questions que je soulève à travers cette problématique, et indique comment j'ai choisi de les aborder :

1) *Comment exploiter les textes en tant que sources de connaissances ?* Par cette question, je fais référence à la nécessité de choisir un point de vue sur les textes, de se situer par rapport aux différentes manières de les traiter informatiquement ou de les interpréter. D'emblée, j'ai choisi de m'appuyer sur l'expérience de la terminologie, de la linguistique de corpus et sur les logiciels de traitement automatique des langues. Le cadre du groupe pluridisciplinaire TIA a favorisé ce type d'échange. Je me suis donc posé cette question a posteriori, de manière à clarifier le point de vue sur les textes qui est impliqué par notre approche et à justifier son adéquation.

2) *Quelles structures de données pour rendre compte des connaissances tirées des textes ?* J'ai reformulé cette question successivement sous deux formes. Dans un premier temps, j'ai donné priorité à l'analyse des textes : étant donné une analyse de texte, quelle structure de donnée permettrait d'en rendre compte le plus exhaustivement possible ? En collaboration avec des linguistes et terminologues, j'ai étudié et revu la notion de Base de Connaissances Terminologiques dans cet objectif. Dans un deuxième temps, j'ai fixé la structure de données ciblée, celle d'ontologie. La question devient alors : comment faire enrichir la structure des ontologies pour intégrer d'autres données disponibles à partir de l'analyse des textes et pertinentes dès que l'on veut utiliser l'ontologie pour explorer, indexer ou annoter des textes. Dans les deux cas, une réponse possible est de mettre au point un modèle rendant compte de données terminologiques, de leur sémantique (souvent à travers un réseau sémantique) et éventuellement de leur usage (textes).

3) *Comment repérer et exploiter des indices linguistiques de connaissances ?* Autrement dit, comment trouver dans les textes les éléments de connaissance pertinents pour renseigner les structures de données de modélisation ? Plusieurs niveaux de connaissances peuvent

ainsi être repérés dans les textes : classes sémantiques définissant des concepts, relations conceptuelles et propriétés, ou encore instances de concepts. Il s'agit donc d'identifier ou de développer des logiciels permettant d'identifier ces types de connaissance, puis d'étudier comment combiner l'utilisation de plusieurs types de logiciel. Une dernière question se situe au niveau de l'exploitation des résultats : selon quels critères présenter les résultats ? mettre en avant des éléments « importants » du domaine ? Pour l'ensemble, j'ai choisi de cibler des logiciels de traitement automatique des langues (TAL) et des approches linguistiques. Je considère aussi que la réponse à ces questions passe par une collaboration avec des chercheurs en traitement automatique des langues pour adapter ou développer des logiciels d'analyse et des méthodes à la construction de modèles.

4) *Selon quelle méthode définir une ontologie à partir de textes ?* Cette interrogation générale se décompose en questions plus précises : comment sélectionner des textes et former un corpus pertinent ? quels logiciels de TAL utiliser ? comment en coordonner l'utilisation conjointe à celle d'outils de modélisation ? quel statut accorder aux données extraites par rapport à un modèle ? La nature du modèle et son objectif d'utilisation ont-ils une influence sur la manière d'interpréter ces données ? sur la manière de construire le modèle en général ? Selon quels principes organiser les connaissances dans un modèle ? Pour répondre à ces questions, j'ai mené des études expérimentales en collaboration avec les concepteurs de logiciels de TAL. L'objectif de ces études est d'abord de mieux maîtriser le rôle et la contribution de chaque logiciel, son adéquation aux caractéristiques des textes, la nature de ses résultats et sa contribution possible à la construction du modèle. Un deuxième objectif est de définir une méthode. J'ai choisi de collaborer avec des linguistes pour bénéficier de leur éclairage sur l'analyse de textes.

5) *Comment tirer profit du parallèle évident entre l'extraction d'éléments de modèles à partir de textes et l'annotation des textes à l'aide d'éléments de modèles* ? Construire un modèle à partir de textes convient particulièrement lorsque ce modèle facilite l'accès « par le contenu » à des documents, en recherche d'information par exemple. Cette question surgit au cœur des recherches sur le web sémantique. En quoi la dimension terminologique que j'ajoute aux modèles obtenus à partir de textes favorise-t-elle leur utilisation pour l'indexation sémantique et l'annotation de textes à l'aide de méta-données ou de mots-clés ? Comment les techniques, logiciels et méthodes utiles à la construction des modèles peuvent-ils aussi servir pour annoter ou indexer de nouveaux textes à partir de ces modèles ? J'ai retenu une démarche expérimentale pour répondre à ces questions et de me focaliser sur des applications de recherche d'information avec des tentatives de réponses variées : en évaluant l'apport de logiciels de TAL, d'ontologies génériques ou spécifiques, à des étapes particulières de la recherche d'information.

6) *Validation : pour quelles classes d'applications ces modèles sont-ils plus pertinents ? comment adapter les approches, méthodes et outils à ces applications* ? Valider une approche de modélisation à partir de textes comporte plusieurs facettes : la validation des outils, des techniques et des modèles obtenus, mais aussi l'intérêt de l'approche (coût versus qualité de la réponse apportée à un besoin particulier). J'ai choisi de me situer dans une démarche d'ingénierie : la validation de la méthode découle de la pertinence de l'application finale, et donc de l'usage fait du modèle conceptuel. Je considère que l'évaluation des outils pris indépendamment les uns des autres n'est qu'une étape intermédiaire, réductrice et pas toujours révélatrice. Mon intuition est que les performances des logiciels sont importantes, mais marginales par rapport aux facilités de navigation et de sélection de leurs résultats. Je n'ai donc pas visé des évaluations qui se réfèrent à des tests de performance, mais des utilisations pour des projets réels. Face à la diversité des types de modèle, je cherche à capitaliser les retours d'expérience pour établir progressivement, le type de modèle et l'approche de construction associée qui convient pour chaque classe d'application.

7) *Quelle est la généricité, la possibilité de réutilisation des modèles obtenus* ? Construites pour répondre à des besoins spécifiques dans des domaines ciblés, les ressources ontologiques et terminologiques sont plus ou moins proches des usages des termes dans les textes. Dans le cas des ontologies, j'ai choisi de faire référence à des principes de structuration ontologique, et de pousser l'analyste à expliciter un point de vue pour définir et organiser des concepts. Ce choix est motivé par la volonté de s'écarter des textes et d'anticiper la prise en compte de nouveaux usages. On peut se demander jusqu'où les modèles obtenus peuvent être effectivement réutilisés et comment.

CHAPITRE 4 - METHODES ET OUTILS POUR LA MODELISATION DE CONNAISSANCES EXPERTES

Ce chapitre s'organise en quatre parties. Je rappelle d'abord mes motivations et je justifie le choix d'une approche ascendante (4.1). Je développe ensuite les travaux et résultats relatifs à l'acquisition de connaissances d'experts et la méthode MACAO (4.2). A partir de là, je montre comment la méthode MACAO-II et la représentation des connaissances MONA permettent d'accompagner un processus de modélisation et d'acquisition guidée par le modèle (4.3). Je dresse enfin un bilan de ces travaux pour mettre en avant les principes repris dans mes recherches (4.4).

J'aborde ici ma contribution au problème de l'acquisition et de la modélisation des connaissances expertes, qui recouvre des techniques, des logiciels et une méthode ascendante, partant de l'analyse des connaissances des experts tels qu'ils les utilisent. Deux périodes ont jalonné ces recherches. Une proposition initiale, dont la méthode MACAO constitue le noyau, vise la mise au point de systèmes experts, la priorité étant la modélisation cognitive de l'expertise. Cette méthode a évolué pour mieux prendre en compte le contexte de mise en œuvre de l'expertise et la tâche des utilisateurs du futur système. Un nouveau point de vue a été retenu ensuite et a donné lieu à la méthode MACAO-II. Le modèle conceptuel y est considéré comme un modèle du système, spécifiant les problèmes à traiter (et comment ils seront traités) sous la forme d'un modèle de raisonnement décrit à l'aide de tâches et de méthodes. La représentation des connaissances a été revue pour proposer un langage conceptuel adapté à la modélisation de systèmes coopératifs. Enfin, plusieurs valorisations de MACAO-II ont eu pour objectif, à travers des collaborations avec d'autres équipes, de diversifier les techniques de modélisation, et d'assurer le suivi jusqu'à la réalisation d'un modèle opérationnel.

4. 1 - Choix d'une modélisation ascendante

4.1.1 Motivations

Ces travaux couvrent la période qui va de 1986 à 1995, et correspondent aux deux premières parties de l'historique présenté au chapitre 1. Ils sont le fruit de mon travail de thèse au LSI, de sa valorisation et de son enrichissement au sein du laboratoire mixte ARAMIIHS jusqu'en 1991, puis de la contribution de Nada Matta dont j'ai encadré la thèse de 1992 à 1995. J'ai choisi de développer ces résultats, pourtant anciens, pour plusieurs raisons :

a) En matière de recueil de connaissances auprès d'un expert ou d'une communauté de spécialistes dont les savoir-faire n'ont pas été écrits ou mis en forme, **les techniques à la base des**

approches cognitivistes restent toujours d'actualité, moyennant une adaptation du statut accordé aux données obtenues et de leur interprétation en termes de connaissances. En particulier, dans des projets de gestion des connaissances ou de définition d'aides à la décision, le point de départ est souvent le recueil des paroles d'experts. Or les diverses techniques adaptées à l'ingénierie des connaissances et leurs caractéristiques sont finalement mal connues, y compris dans le cercle du génie logiciel et de l'IA, ce qui conduit parfois à des approches naïves des entretiens.

b) La manière d'en exploiter les fruits (à savoir les entretiens enregistrés, etc.) pose toujours problème si l'on ne se place pas dans une perspective de modélisation. La large diffusion de KADS puis de CommonKADS met plus l'accent sur la réutilisation comme méthode privilégiée pour construire un modèle. Or cette réutilisation suppose aussi des analyses des méthodes de résolution utilisées par les spécialistes du domaine. Un des résultats intéressants de l'ingénierie des connaissances est donc de fournir des **recommandations pour conduire la partie ascendante de la modélisation**.

c) Plusieurs classes d'applications nécessitent une démarche ascendante : les systèmes d'aide à la décision, les systèmes tuteurs (EIAH) ou encore la mise au point de l'interaction ou de la coopération entre un système et ses utilisateurs. En effet, leur raisonnement et les tâches qu'ils réalisent doivent **s'appuyer sur un modèle conceptuel proche de la manière dont l'utilisateur réalise sa tâche**.

4.1.2 De MACAO à MACAO-II : historique

Différentes périodes ont jalonné ces recherches sur la modélisation ascendante. Les résultats établis ainsi que les nouveaux points de vue introduits à chaque période sont présentés ici. Ces évolutions sont en phase avec l'historique du domaine dressé au chapitre 2.

Le noyau initial de cette proposition, la **première version de la méthode MACAO** [Thèse-AUSSENAC, 89] vise la mise au point de systèmes experts, la priorité étant la modélisation cognitive de l'expertise. Les techniques de recueil de connaissances proposées autant que le langage de modélisation font référence aux résultats établis en psychologie cognitive sur les connaissances expertes. Le logiciel MACAO a été une des toutes premières plates-formes de modélisation conceptuelle selon une démarche ascendante, qui ne soit pas basée sur un modèle de résolution de problème pré-établi. En effet, ce n'est que vers 1990 qu'a été disponible la K-Station (vendue par Ilog), plate-forme associée à la méthode KOD, pourtant diffusée avant MACAO. La méthode a été utilisée sur des cas d'école (projet Sisyphus 1) [EKAW, 91b] et dans le cadre du projet SAMIE avec la société MMS au sein du laboratoire ARAMIIHS au cours de mon séjour postdoctoral dans ce laboratoire [SAMIE, 90].

La **deuxième version de MACAO**, définie vers 1992, cherche à mieux prendre en compte le contexte de mise en œuvre de l'expertise et la tâche des utilisateurs du futur système. Le modèle conceptuel est alors issu d'un processus d'abstraction de résolution de problèmes par l'expert, mais il est mis en rapport avec le modèle de la tâche de l'utilisateur. Les techniques de recueil proposées intègrent des analyses ergonomiques de l'activité. Des évolutions du logiciel et de la représentation des connaissances, en particulier une représentation arborescente des schémas, permettent au cogniticien, au cours de la modélisation, de s'appuyer sur les raisonnements experts pour produire un modèle qui s'en éloigne afin de mieux s'adapter à la tâche et au contexte de travail de l'utilisateur. Le point fort de cette nouvelle version est certainement l'intégration de démarches d'analyse ergonomique ainsi que l'articulation entre représentation du raisonnement et connaissances du domaine.

Ces deux groupes de travaux sont présentés dans la partie 4.2 de ce chapitre.

Ensuite, la **méthode MACAO-II** (finalisée vers 1995) pousse plus loin la volonté de définir le modèle conceptuel comme un modèle du système, de spécifier les problèmes qu'il doit traiter et comment ils seront traités sous forme d'un modèle de raisonnement décrit à l'aide de tâches et de

méthodes [thèse-MATTA, 95]. Ce langage, MONA, permet de définir progressivement des composantes du raisonnement puis de les formaliser en conservant la structure conceptuelle. MACAO-II est certainement influencée par les résultats établis dans la méthode KADS (Schreiber *et al.*, 1994), qui a pris une place prépondérante à partir de cette période, et par les représentations du raisonnement sous forme de tâches et méthodes, comme dans les langages OMOS (Linster, 1991), LISA (Delouis, 1993) ou la méthode COMMET (Steels, 1992). Cependant, la méthode conserve la particularité de privilégier une analyse ascendante des manières de procéder des spécialistes du domaine, le modèle du système étant obtenu après analyse de modèles de cas produits par un expert. Elle offre donc une alternative intéressante aux approches privilégiant la réutilisation de composants génériques, comme cela a été évalué dans le cadre du projet SADE mené en partenariat avec EDF.

Enfin, les dernières **valorisations de MACAO-II** ont eu pour objectif, à travers des collaborations avec d'autres équipes, de diversifier les techniques de modélisation d'une part, et d'assurer le suivi jusqu'à la réalisation d'un modèle opérationnel d'autre part. Ainsi, le système ASTREE (Tort, 1995) a été défini à partir du langage MONA et permet de guider la mise au point d'un modèle de raisonnement en tirant profit des caractéristiques de structuration des concepts du domaine ainsi que d'opérations de raisonnement élémentaires. Pour l'opérationnalisation, le langage ZOLA (Isténès, 1996) a été utilisé pour définir des structures opérationnelles conformes à celles de MONA (Beaubeau, 1997). La correspondance structurelle permet ainsi une validation opératoire du modèle conceptuel en phase de mise au point. Afin de faciliter la modélisation du domaine, souvent fastidieuse, une tentative d'utilisation d'un logiciel d'extraction de terminologies, LEXTER, a été réalisée au sein du projet SADE. Les résultats fructueux obtenus ont posé les bases des contributions possibles des analyses terminologiques à MACAO-II et plus largement en ingénierie des connaissances.

Ces deux derniers groupes de recherches sont présentés conjointement dans la partie 4.3.

4. 2 - MACAO : Acquisition de connaissances expertes

4.2.1 Contexte

4.2.1.1 Premiers travaux en acquisition de connaissances

Comme je l'ai précisé dans mon analyse de l'ingénierie des connaissances (chapitre 2), les premiers travaux dans le domaine datent de 1980 et les premiers résultats notables de 1985. L'objectif était alors de définir comment procéder pour recueillir les connaissances d'un système expert, selon une approche cognitiviste, c'est-à-dire en restant au plus près de la manière de raisonner de l'expert. Mon travail, amorcé en 1986, s'est inspiré d'approches venant de l'intelligence artificielle et de plusieurs travaux convergents en acquisition de connaissances. Je rappelle ici les plus influents sur la communauté scientifique et sur mes travaux.

Newell et le Knowledge Level : Une référence commune à toutes ces recherches est celle d'A. Newell qui a donné une base théorique aux travaux sur les modèles conceptuels en proposant l'analyse d'un système à base de connaissances au « knowledge level » (niveau des connaissances). Ce niveau se situe au-dessus du niveau formel dans une description en couches de plus en plus abstraites du fonctionnement des systèmes informatiques. Le système, perçu comme un agent rationnel par un observateur, y est spécifié en termes de buts et des connaissances utilisées par les atteindre. Cette analyse a été largement reprise en ingénierie des connaissances. Tout d'abord, elle a inspiré les bases d'une représentation de haut niveau des buts pour organiser les heuristiques mises en œuvre par un système expert. Elle a aussi promu la notion de modèle conceptuel situé au « knowledge level », comme représentation à part entière en amont d'un modèle formel et du système opérationnel. Enfin, elle souligne les limites de la représentation des connaissances,

Newell insistant sur le fait que les « connaissances ne peuvent être vues autrement que comme le résultat de processus d'interprétation s'appliquant à des expressions symboliques ».

Generic Tasks de Chandrasekaran (Ohio State University) : ces « tâches génériques » sont des primitives de haut niveau, englobant des règles de production, utilisées pour décrire les buts et raisonnements produits par un SBC. Ces structures organisent les règles en fonction des buts au moment de la conception du système. Définies à un niveau générique, elles sont propres à un type de problème et peuvent être utilisées dans différents domaines pour les mêmes tâches. Plusieurs travaux, en particulier Components of Expertise (Steels, 1992) et KADS, ont développé ces deux idées : rendre explicites des buts pour spécifier le SE et représenter le raisonnement sous forme d'arbre de tâches (choix aussi retenu dans MACAO-II) ; définir des blocs génériques réutilisables pour de nouveaux systèmes.

Role Limiting Methods (équipe de J. Mc Dermott au MIT) : Plusieurs logiciels ont été définis selon ce principe (Marcus, 1988) : ils exploitent une représentation explicite de haut niveau de la résolution de problème pour guider un dialogue avec l'expert et l'enrichissement de la base de règles. Les règles jouent des « rôles » prédéfinis dans la résolution de problème, d'où le nom de l'approche. Ces systèmes ont permis d'établir plusieurs résultats repris ensuite : la méthode de résolution de problème peut être explicitée ; elle peut guider l'organisation des connaissances du domaine et les règles (principe utilisé dans MACAO-II) ; ces méthodes de résolution de problème sont adaptées à des types de problème, et ne conviennent pas à tous ; enfin, ces méthodes sont des méthodes d'IA, qui caractérisent l'algorithme de parcours des règles plus que la manière de raisonner de l'expert (repris dans MACAO-II pour définir des méthodes du raisonnement expert).

TEIRESIAS (Davis, 1979) : Ce module de transfert d'expertise a été associé au premier système expert, MYCIN, pour en corriger les règles. Le système dialogue avec l'expert à partir des erreurs signalées pour guider le repérage des règles ayant conduit à cette erreur. TEIRESIAS valide l'hypothèse d'une acquisition des connaissances interactive, s'appuyant sur le système d'inférence et s'adressant directement à l'expert du domaine.

KOD (Vogel, 1988) : C. Vogel propose une méthode qui définit des modalités d'entretien avec les experts, puis d'analyse linguistique de leurs paroles, pour construire un *modèle cognitif* qui sera opérationnalisé en système expert. Vogel souligne la nécessité de comprendre les mécanismes cognitifs experts et cherche à ce que le raisonnement produit par ce système colle au plus près de celui de l'expert. Ces mêmes principes sont adoptés dans MACAO. Le *modèle cognitif* est vu comme une représentation intermédiaire entre le langage naturel, moyen d'expression des connaissances, et le système opérationnel. Enfin, Vogel a été un des premiers à proposer que l'opérationnalisation suive le paradigme objet et non uniquement à base de règles, ce qui a été retenu pour MACAO-II.

ETS (Boose, 1986) puis AQUINAS (Boose, 1988) : Ce système s'appuie sur des travaux en psychologie, la « personal construct theory » de Kelly, pour proposer une technique de classification de concepts et de valeurs : les *grilles répertoires*. ETS propose une interface graphique d'expression des connaissances, puis de classification pour établir des corrélations entre valeurs et concepts, mais aussi et surtout des corrélations entre concepts eux-mêmes. Ces corrélations débouchent sur la définition de nouvelles règles de production. Un des mérites de ce système est de susciter un questionnement sur des connaissances à acquérir à partir de la proximité inattendue de concepts qui semblent éloignés a priori. J'ai retenu de ces travaux la notion de « mediating representation », c'est-à-dire l'intérêt, pour construire un modèle, de s'appuyer sur des représentations intermédiaires, si possible sous une forme graphique simple, favorisant l'expression et l'interprétation des connaissances par l'expert. J'ai également choisi la technique des *grilles répertoires* comme technique classificatoire de recueil de connaissances du domaine dans MACAO.

KADS : Les recherches préalables à la première version de KADS ont rapidement été identifiées comme une proposition prometteuse car intégrant les différentes contributions et réflexions précédentes. Les premiers écrits de Breuker et Wielinga (Wielinga, 1984 et 1986) posent

le problème de l'acquisition des connaissances dans les mêmes termes que dans MACAO : un problème d'analyse de données verbales, pour lesquelles le modèle des connaissances expertes fournit une grille d'interprétation. Les points forts de leur proposition découlent de l'organisation hiérarchique retenue pour ce modèle d'interprétation. Chaque niveau permet de s'intéresser successivement à une des facettes des connaissances ; ces niveaux sont faiblement reliés, ce qui assure une cohérence à l'ensemble ; chaque niveau fait référence à des primitives de représentation reconnues en IA (tâches, méthodes et domaine sous forme de réseau conceptuel), ce qui garantit la capacité d'opérationnaliser ces modèles dans un système formel ; enfin, les niveaux associés au raisonnement sont supposés suffisamment génériques pour être réutilisables d'un domaine à l'autre.

4.2.1.2 Apports des sciences cognitives

La modélisation d'expertise ne relève pas d'un seul problème de formalisation informatique, mais bien d'accès aux connaissances, de choix du contenu qui sera représenté formellement. L'ingénierie des connaissances ne peut donc être traitée du seul point de vue de l'informatique. Les points de vue de la psychologie cognitive, de la sociologie et de l'ergonomie sont autant d'atouts pour comprendre ce qu'est un expert, son rôle social, ses manières de raisonner, la nature de ses connaissances et les moyens possibles pour les lui faire expliciter. En effet, les sciences cognitives ont toujours étudié de près le projet de l'intelligence artificielle. Avec le développement des systèmes experts, de nombreuses études ont été menées en **psychologie cognitive** dans le but de mieux caractériser la nature des connaissances expertes (Leplat, 1987). Elles précisent leurs spécificités par rapport à celles de novices (Falzon, 1988) (Kolodner, 1983) ou les représentations cognitives qu'elles mettent en jeu, organisées en niveaux de connaissances (Rasmussen, 1985) ou en niveaux de contrôle (Hoc, 1987). Des études ont aussi cherché à s'assurer de la pertinence des règles de production pour leur représentation (Grumbach, 1987). De ce point de vue, la technologie des systèmes experts a même été envisagée par les psychologues comme un support pour la validation de leurs hypothèses de représentation ou d'acquisition de connaissances (Matthieu, 1984). Enfin, les premiers travaux sur les modèles des utilisateurs (Richard, 1986) ont souligné les différences fondamentales entre les logiques des utilisateurs et celles des experts consultés, chacun ayant un niveau d'expertise différent ou une tâche différente à réaliser (par ex. diagnostic versus conception).

Les hypothèses sur l'organisation en mémoire des connaissances et savoir-faire, et de leur mobilisation pour la résolution de problème ont été débattues et depuis remises en question. Mais elles ont servi de point de repère fort pour évaluer la pertinence de l'hypothèse de la simulation cognitive, et proposer une « ingénierie cognitive ». En effet, à ces modèles sont associés des modes d'analyse particuliers et des techniques de recueil, dites « d'extraction de connaissances », ou d'analyse des protocoles associés (Caverni, 1986). La transposition de ces techniques dans le contexte de la conception de systèmes experts a été envisagée très tôt, par exemple dans l'ouvrage de (Gallouïn, 1988). Ma thèse fait un état de l'art assez complet sur ces sujets.

En revanche, la diffusion des travaux **d'ergonomie** est plus récente et suppose des collaborations approfondies pour intégrer des spécificités des systèmes informatiques et de leurs contextes d'usage. À partir du moment où l'on veut mieux prendre en compte les utilisateurs, la conception d'un système informatique se pose en termes de modification d'une tâche suite à l'introduction d'un nouvel outil de travail. Or cette problématique relève de l'ergonomie cognitive (Spérandio, 1996). Cette discipline propose des méthodes d'analyse de l'activité et de la tâche, ainsi que des supports pour rendre compte de ces études. Ces analyses débouchent sur des diagnostics de situations de réalisation de tâches, des projections permettant d'envisager la tâche future dans un environnement modifié, et sur d'éventuelles propositions d'ajustement. Ces réflexions et travaux ouvrent la problématique de la conception des systèmes experts sur un horizon plus large, celui de l'assistance à l'opérateur à l'aide de connaissances. Cette orientation m'a amenée à intégrer la démarche ergonomique dans la méthode MACAO vers 1992, et de manière encore plus forte dans MACAO-II.

Mon appréhension de ces travaux n'a été possible que grâce au contexte pluridisciplinaire qu'offrait le projet national PIRTTEM[6], dont un des groupes toulousains a étudié de 1985 à 1990 l'impact des systèmes experts dans les environnements de travail. Rassemblant psychologues cognitivistes, sociologues, ergonomes et informaticiens, ce groupe a débouché sur des études relatives à l'utilisabilité et à la validation des systèmes (Chabaud *et al.*, 1989), à leur mise au point (objet de ma thèse) et à leur adéquation pour différents types de tâche (De Terssac *et al*., 1988). Il m'a permis de collaborer avec l'équipe de psychologie cognitive de J.M. Cellier de l'Université Toulouse 2, et en particulier avec B. Michez, en thèse de psychologie du travail. C'est également au sein de PIRTTEM que des contacts ont été établis avec des ergonomes, particulièrement C. Chabaud avec qui j'ai largement collaboré par la suite.

4.2.1.3 La communauté de recherche en acquisition des connaissances

Une communauté de chercheurs organisés

Sur le plan international, la communauté scientifique s'est organisée autour de 1985. Au plan européen, les partenaires du premier projet Esprit sur ce thème, KADS, se sont mobilisés dès 1985 et ont assuré une visibilité dans les conférences d'IA dès 1984. En Amérique du Nord, à la même période (1986), B. Gaines (Univ. de Calgary) et J. Boose (Boeing) ont pris l'initiative d'organiser les conférences KAW (Knowledge Acquisition Workshops). Ils ont transposé l'idée en Europe, ce qui a donné naissance au premier EKAW en 1987. En France, les premières journées scientifiques sur ce thème se sont tenues en 1988 et 1989 en lien avec le PRC-IA, et la première édition des Journées d'Acquisition des Connaissances a eu lieu en 1990 à l'initiative de J.-G. Ganascia et J. Sallantin.

Dynamisme de la communauté française

Afin de fédérer les efforts de recherche nationaux d'une part, de permettre des avancées et de mieux connaître des travaux internationaux, J.-P. Krivine, P. Laublet et moi-même avons constitué le GRACQ en 1991, groupe de travail rattaché au PRC-IA et à l'AFIA puis au GDR-I3. Les activités de ce groupe de 1991 à 1995 ont été d'un dynamisme révélateur des espoirs mis dans ces recherches, de la diversité des questionnements ainsi que des approches, des moyens réels dont disposaient les laboratoires grâce à des contrats et, il faut le reconnaître, d'un effet de mode ! Entre 1991 et 1994, le groupe s'est réuni régulièrement tous les 2 mois, rassemblant entre 60 et 80 personnes au cours de réunions plénières mais aussi au sein de 6 groupes de travail. Pratiquement, cet effort a débouché sur deux dossiers dressant un panorama des travaux du domaine en France, parus dans le bulletin de l'AFIA l'un en 1991 et l'autre en 1998. La fréquence des réunions et les objectifs du groupe ont ensuite pris un tournant pour devenir aujourd'hui une structure d'animation.

Difficulté de diffusion et reconnaissance des travaux français

Caricaturalement, les recherches françaises sur ce domaine n'ont pas vraiment réussi à passer le cap européen et donc international. Malgré leur qualité, elles restent peu connues car peu publiées en anglais et peu intégrées à d'autres travaux ou projets. Cela est particulièrement vrai pour MACAO. En 1989, les propositions faites étaient originales et reconnues. En 1995, MACAO-II a été le support de collaborations nationales. Des travaux de valorisation, menés avec EDF, le LRI et l'IRIN ou au sein des expériences Sisyphus de comparaison de travaux en IC, ont été publiés et appréciés. Cependant, ces recherches n'ayant pas été intégrées au sein d'un logiciel maintenu et largement diffusé, elles ne font plus référence qu'au sein de la communauté française.

[6] Pôle Interdisciplinaire de Recherche du CNRS sur les Technologies, le Travail, l'Emploi et le Mode de vie.

4.2.2 Une méthode de modélisation cognitive

La méthode MACAO organise l'acquisition des connaissances expertes et s'appuie sur un environnement de structuration de ces connaissances avant leur représentation sous forme de règles de production. Différentes techniques d'entretiens (parfois sous forme de logiciels), des principes d'analyse ainsi qu'une représentation des connaissances sont proposés au cogniticien au sein de la plate-forme de modélisation MACAO associée à la méthode.

4.2.2.1 Le modèle cognitif

Rôles du modèle conceptuel : Dans MACAO, le modèle conceptuel est défini comme une représentation « médiatrice », un cadre sémantique partagé, qui doit faciliter le dialogue entre les acteurs concernés [EKAW, 89] et [KAW, 89]. En plus de « langage partagé », le modèle joue deux rôles en priorité : celui de représentation lisible permettant à la fois d'orienter l'acquisition de nouvelles connaissances et de rendre compte de celles déjà recueillies et structurées ; celui de représentation structurée à la base de l'opérationnalisation du système final.

Modèle conceptuel et modèle cognitif : Dans la toute première version de MACAO, le modèle conceptuel se voulait un reflet du modèle cognitif de l'expert, ou du moins d'un sous-ensemble correspondant à la classe de problèmes à résoudre [CREIS, 88]. Il s'agissait d'expliciter les connaissances expertes jusque-là « implicites » [RIS, 92]. Dès 1992, ce modèle a été envisagé comme la représentation que se fait le cogniticien du modèle du système. La priorité était alors de représenter les tâches que le système va réaliser, en restant proche de la méthode des experts.

Nature des connaissances décrites dans le modèle conceptuel : Dans la plupart des approches, le modèle conceptuel est formé de classes abstraites qui renvoient à des objets du domaine dont on caractérise les propriétés ou à des opérations, à des tâches ou des procédures qui caractérisent les raisonnements. Dans MACAO, le modèle conceptuel comporte aussi des exemples, appelés *modèles de cas*. Les structures de modélisation servent à la fois à rendre compte du modèle final et à modéliser des études de cas puis des méthodes de résolution associées à des classes de problèmes (*modèles de types de problème*). Ce choix est lié à la méthode, ascendante, qui repose sur des abstractions faites progressivement à partir de ces études de cas.

Première représentation des connaissances : Les choix précédents ont eu une incidence sur le choix d'une représentation des connaissances dans le modèle. Plutôt que de partir de formalismes classiques en IA, comme les réseaux sémantiques ou les règles de production, de nouvelles structures ont été définies afin de rendre compte d'un modèle cognitif d'expert. Un premier langage de représentation, proche des réseaux sémantiques, a été proposé en 1988 [EKAW, 88]. La structure de base, *l'unité fonctionnelle*, permettait de rendre compte soit de connaissances du domaine, soit d'heuristiques. Les *unités fonctionnelles* étaient organisées au sein d'un réseau à l'aide de liens, qualifiés de *statiques* ou *dynamiques* suivant qu'ils traduisent des relations indépendantes du raisonnement ou générées par celui-ci. Cette mise à plat des connaissances ne favorisant pas la structuration, elle n'avait pas atteint l'objectif visé de guider l'acquisition.

Représentation des connaissances à l'aide de schémas [KMET, 91] : un travail de DEA [DEA, 91] mené en 1991 a débouché sur une proposition à base de *schémas*, proches de frames au sens de la psychologie cognitive. Ces structures de plus haut niveau regroupent les connaissances associées aux actions ou inférences mises en œuvre dans une classe de situations (exemple en figure 4.2) pour atteindre un but particulier. Elles s'instancient en schémas empiriques, qui correspondent aux connaissances mobilisées dans un contexte spécifique. Les schémas contiennent un ensemble de procédures (les unités fonctionnelles) et d'actions portant sur les concepts du domaine qui sont présents dans ce contexte.

Schema-name: OFFICE-ASSIGNMENT
Name-comment: assignment problem of persons into offices respecting constraints
Parameters: Input parameters: person-list, office-list
Constraints: assignment rules, compatibility rules, priority lists
Output parameters: assigned person-list
Context-comment: no member of the team has been assigned an office yet
The number of places in rooms is greater or equal to the number of persons
(This second condition is not checked by the model).
Context: for all i such as person(persi) and not(*is-assigned*(persi))
Goal-comment: Assign an office to all the persons respecting constraints such as priority in considering persons and offices, centrality considerations and assignment preferences.
Goal: for all i such as person(persi), is-assigned(persi)
Strategy: ()
Strategy-comment:
Process: WHILE determinant *type-list-not-empty*
select-person-type
group-compatible-persons
aff-off-group

Figure 4.2 : Le schéma OFFICE-ASSIGNMENT : Le contexte indique les conditions dans lesquelles s'applique ce schéma. Sa mise en oeuvre suppose que les schémas SELECT-PERSON-TYPE et GROUP-COMPATIBLE-PERSONS soient toujours activables, alors que AFF-OFF-GROUP peut échouer si le nombre de bureaux vérifiant les contraintes de taille et de situation est insuffisant.

Finalement, la structure de *schéma* représente un pas de raisonnement associé à la réalisation d'un but dans un contexte. Les schémas font appel à d'autres schémas et leur enchaînement forme le *modèle du raisonnement*, qui peut être visualisé sous forme d'arbre, comme sur la figure 4.3. Les concepts du domaine sont organisés en réseau conceptuel, le *modèle du domaine*, également affiché sous forme graphique. Des modèles sont construits à trois niveaux d'abstraction : problèmes (cas), classes de problèmes et expertise globale. À chaque niveau, le modèle comporte deux volets : les concepts et les relations forment un réseau, le modèle du domaine.

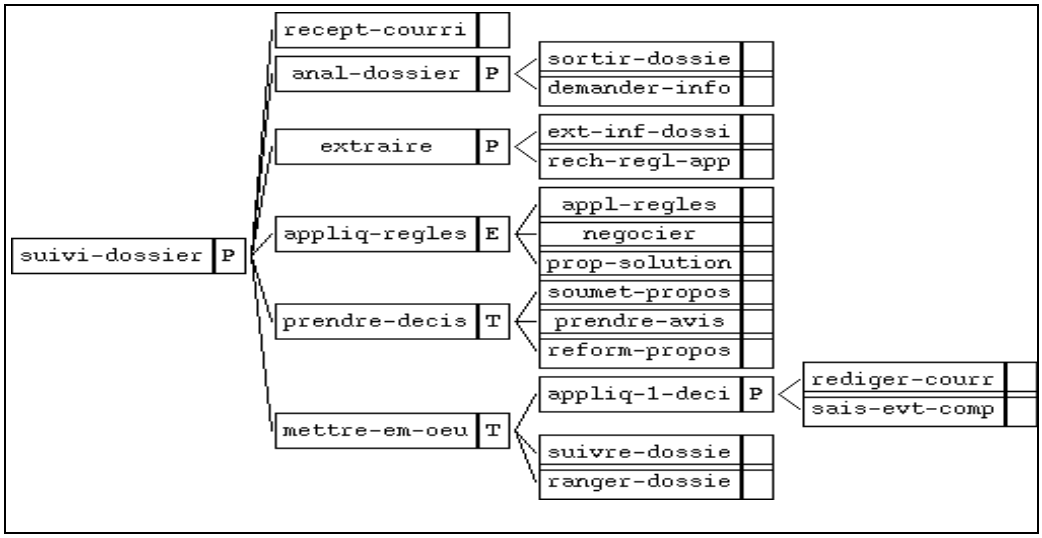

Figure 4.3 : Exemple de MC du raisonnement construit avec MACAO (projet SADE). E=ET, P=PUIS, O=OU, T=TANT QUE, C=CAS.

4.2.2.2 Les outils et la méthode

Principes : Plusieurs techniques de recueil et d'outils d'aide sont proposés afin de rendre compte des différents types de connaissance experte. La méthode guide dans leur choix, chaque technique étant documentée par un exposé de ses caractéristiques. Le cogniticien choisit celles qui

conviennent pour le problème étudié. Ces techniques conduisent à se focaliser sur l'activité de l'expert en priorité [COGNITIVA, 90].

Techniques proposées : la méthode suggère une progression dans leur utilisation [EKAW, 91b] : commencer par des entretiens informels liés à l'expression des besoins, auprès des experts et des utilisateurs, et par des observations et des analyses de l'activité ; poursuivre par des entretiens centrés sur la sélection de cas, en utilisant la technique des grilles répertoires pour dégager des classes de problèmes ; poursuivre par des études de cas (simulations avec verbalisations simultanées, verbalisations explicatives consécutives puis analyse de protocoles) ; terminer par des entretiens centrés liés à la validation du modèle au fur et à mesure de sa construction.

Grilles répertoires : le logiciel guide la mise en oeuvre de cette technique. Il interroge l'utilisateur (l'expert en l'occurrence) sur des exemples de problèmes, les caractéristiques qui les différencient, puis propose de les regrouper en *catégories de problèmes* en fonction des réponses fournies. Cette phase de classification prépare le processus d'abstraction.

Une modélisation ascendante : la méthode s'intègre dans un processus classique de génie logiciel. Sa mise en œuvre suppose d'avoir mené au préalable une analyse des besoins et de la demande au cours d'une étude d'opportunité (IEA, 91). Elle consiste en quatre étapes principales, chaque étape comportant des activités de recueil, de conceptualisation, de structuration et de validation. A partir de la représentation de la manière dont l'expert traite des exemples, on dégage une démarche générale via le niveau intermédiaire des catégories de problèmes :

(1) *Identification de l'expertise* : Tout d'abord orientée vers la spécification du rôle et des fonctionnalités du système à construire, cette étape a été enrichie par une analyse ergonomique du travail et de l'organisation des experts et des futurs utilisateurs ; puis elle a été complétée par une analyse de documents et l'élaboration d'un lexique.

(2) *Étude du raisonnement sur des exemples* : Il s'agit de recueillir un large éventail de cas rencontrés par l'expert puis de les lui faire classer en types de problème selon la technique des grilles répertoires ; ensuite, de définir des conditions de simulation ; puis de faire simuler par l'expert la résolution de cas représentatifs. Ces simulations font l'objet d'explications par l'expert puis d'une analyse par le cogniticien qui conduit à la représentation de modèles de cas (pb1.1, pb1.2 etc. sur la figure 4.4). La part des commentaires dans les structures de schéma est importante et permet de leur associer des parties entières de retranscriptions des verbalisations.

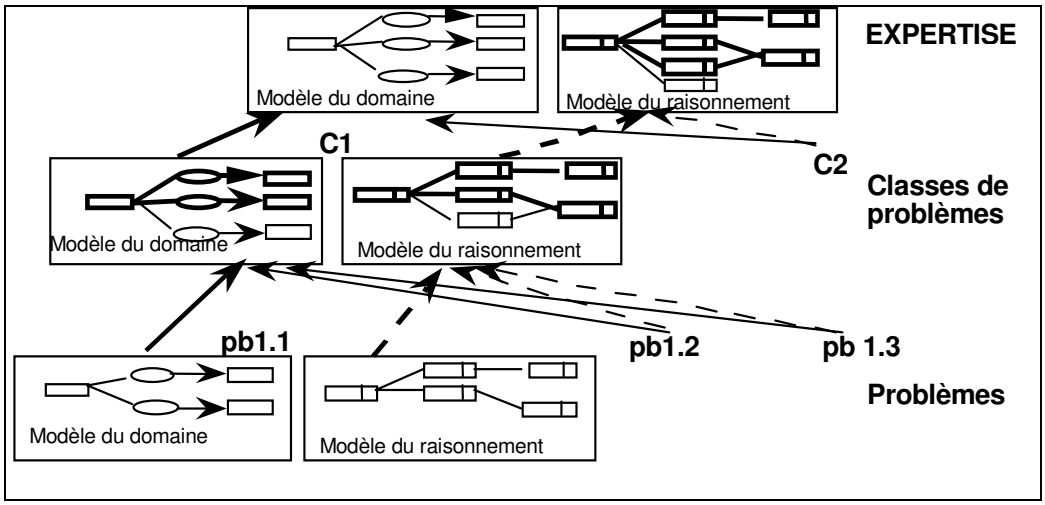

Figure 4.4 : Le processus de modélisation ascendante avec MACAO.

(3) *Modélisation de l'expertise* : La démarche ascendante consiste à dégager les points communs et les conditions de variation dans les différents raisonnements mis en œuvre. Tout

d'abord, des modèles sont construits pour chaque classe ou catégorie de problème identifiée (C1 et C2 sur la figure 4.4). Finalement, le modèle conceptuel du système doit rendre compte des différentes facettes de l'expertise sous une forme unifiée, que ce soit pour la partie raisonnement que pour la représentation du domaine.

(4) *Validation* : La validation se fait en vérifiant que le modèle permet de représenter les exemples traités initialement ; l'expert valide également le modèle.

La plate-forme MACAO est un logiciel comportant des éditeurs permettant de gérer les différents cas traités (énoncés et retranscriptions des protocoles recueillis lors de leur résolution simulée), les différents modèles conceptuels (des cas, des types de problème et de l'expertise) avec pour chacun la partie résolution de problème et la partie raisonnement. Ces éditeurs permettent de définir des schémas et de les organiser dans un modèle du raisonnement, ou des concepts et des relations pour former un modèle du domaine. Un effort a été fait dès les premières versions de MACAO pour favoriser une représentation graphique du réseau conceptuel (modèle du domaine) et des arbres de schémas pour la représentation du raisonnement, l'éditeur de graphe permettant de créer ou modifier les modèles de manière interactive [JAC, 90] [EKAW, 91b]. La plate-forme intègre également un module de classification selon la technique des grilles répertoires. La première version a été développée en Le_Lisp en 1990 dans le cadre d'un projet CNAM [Rapport-RIVIERE, 90], et enrichie lors de deux autres stages [Rapport-ROUTABOUL, 91].

4.2.2.3 Expérimentation et évaluation

Étude de cas Sisyphus I et II

Une première mise en œuvre de la méthode dans sa globalité correspond à l'étude de cas Sisyphus I, projet de comparaison de systèmes de modélisation mené dans le cadre de la conférence EKAW. L'énoncé d'un problème d'affectation simple était fourni comme point de départ. L'objectif était de produire un modèle de connaissances et un système capable d'effectuer cette tâche. L'énoncé ne fournissait aucun exemple de résolution de problème ni de solution que pourrait adopter un expert. Utiliser MACAO sur cet exemple a permis de mieux qualifier la nature du modèle obtenu et le caractériser par rapport aux modèles obtenus selon d'autres approches [EKAW, 91b].

Un travail approfondi de comparaison avec le système 3D-KAT a été mené lors d'une collaboration avec R. Dieng car 3D-KAT privilégiait aussi une démarche ascendante et une visualisation graphique [EKAW, 91a]. 3D-KAT se focalise sur l'expression de paramètres de décision et des relations de cause à effet entre ces paramètres. Il convient bien pour des tâches de conception pour lesquelles les contraintes entre paramètres, appelées topoi, sont traduites par des relations entre les nœuds d'un graphe (connaissances de type « plus …, plus … » ou « moins …, plus … » etc.). Ce logiciel implémente une technique particulière pour exprimer des connaissances causales, mais ne permet pas d'organiser un modèle conceptuel au sens où il n'explicite pas à un niveau abstrait l'ensemble des connaissances requises pour la résolution de problèmes.

L'objectif de Sisyphus II était de comparer sur les mêmes bases différentes approches et outils de modélisation de connaissances. À partir du problème de Sisyphus I, une grille de questions plus précises portait sur la manière dont le modèle était construit, les caractéristiques de l'approche de modélisation, les propriétés du modèle obtenu ainsi que son opérationnalisation. En particulier, la robustesse du modèle face à des changements d'énoncé ou face à des situations conflictuelles devait être évaluée. Les limites rencontrées pour adapter le modèle à de nouveaux cas ont conduit à envisager une autre représentation des connaissances. Elles ont été à l'origine de l'évolution de la représentation et de la méthode, qui ont conduit à MACAO-II (IJHCS, 94).

Projet SAMIE

La méthode et le logiciel MACAO ont ensuite été utilisés dans le cadre d'un projet en entreprise, pour évaluer la faisabilité d'un système d'aide au diagnostic en ligne pour des techniciens en informatique. Ce projet SAMIE a permis d'approfondir toute la partie initiale relative à l'analyse de l'activité et à l'étude de différentes sources de connaissances (ici, des spécialistes ayant différents degrés d'expertise et des bases d'incidents) [SELF, 90] et [SAMIE, 90]. L'analyse de l'activité a permis de décrire l'activité du futur utilisateur du système, des spécialistes du domaine mais aussi de leurs partenaires. Elle a conduit à un modèle de la circulation de l'information et de la gestion des tâches au sein de l'équipe impliquée par l'introduction du système à base de connaissances. De nouveaux éléments méthodologiques ont ensuite été intégrés dans la méthode. L'approche par catégorie de problèmes a permis de mener une étude de faisabilité sur un sous-ensemble du domaine à couvrir, puis de projeter le résultat qui aurait pu être obtenu sur le domaine entier, le coût et la durée de développement d'un tel système.

4.2.3 Enseignements tirés de MACAO

4.2.3.1 Comparaison aux travaux de cette période

Finalement, la méthode MACAO reprend les principes de plusieurs systèmes développés autour de 1990, et anticipe des travaux qui seront développés par la suite et inspirés par KADS. La méthode est une des toutes premières disponible dans la communauté française, avec KOD, et elle est la première à proposer un environnement de modélisation, la K-station n'ayant vu le jour que vers 1992. Parmi les travaux francophones de la même époque, outre 3D-KAT, dédié aux problèmes de conception, quelques systèmes d'acquisition ont été consacrés à des types de problème particulier : diagnostic médical pour ACTE (Charlet, 1990), DIVA (David et Krivine, 1990). Les systèmes ADELE (Reynaud, 1989) et ACTE présentent l'originalité d'exploiter la caractérisation des types de connaissance dans les systèmes experts dits « de deuxième génération ». Les « connaissances profondes », c'est-à-dire des relations entre éléments d'un modèle (par exemple des relations structurelles entre composants à diagnostiquer), guident le recueil d'heuristiques (dites « connaissances de surface ») utilisées par les experts.

Comme les méthodes KRITON et BABYLON (Linster, 1988) (Linster, 1989), la méthode MACAO souligne la nécessité de combiner plusieurs techniques d'analyse de l'expertise en amont, dont les grilles répertoires et l'analyse de protocoles. La représentation des connaissances de ces méthodes propose aussi d'utiliser des concepts et des relations, mais, en revanche, elle ne donne pas un réel statut au modèle conceptuel.

KOD (Vogel, 1988) ayant été une source d'inspiration importante, les deux méthodes partagent une approche cognitiviste et une démarche ascendante pour définir un modèle conceptuel de l'expert ou modèle cognitif. Cependant, KOD fait une référence bien plus forte à la linguistique pour systématiser le repérage d'éléments de modèle dans les retranscriptions d'entretien. Sur les techniques de recueil de connaissances, elle préconise une méthode d'entretien très « analytique » : l'expression spontanée non guidée, qui peut sembler peu efficace.

Enfin, une des méthodes de référence à cette période est Protégé (Musen, 1986), méthode associée à un logiciel. Les modèles conceptuels construits avec Protégé s'appuient sur un formalisme de tâche pour une méthode de résolution particulière : le raffinement de plan. Comme dans les Role Limiting Methods, la méthode de résolution est exploitée pour guider le recueil : des écrans de saisie sont générés pour faciliter l'organisation des éléments du domaine requis pour résoudre les problèmes visés. Protégé favorise une acquisition interactive et guide donc beaucoup plus précisément que MACAO dans le recueil de connaissances. Bien sûr, dans sa version de 1986, son champ d'application était plus réduit. Protégé a ensuite évolué de manière à s'adapter à toute méthode de résolution définie par l'utilisateur. Son point fort est justement d'être très ouvert, et

d'autoriser la définition par l'analyste du modèle de raisonnement ou de niveau « méta » qui va ensuite guider le recueil de connaissances du domaine.

4.2.3.2 Limites identifiées

La mise en œuvre de MACAO dans le projet SAMIE ainsi que sa confrontation à d'autres approches dans le cadre de Sisyphus ont souligné des limites fortes à l'approche MACAO. Ces limites dessinent les axes des futurs développements de la méthode réalisés entre 1992 et 1997 :

- l'incapacité à simuler le modèle pour le valider et à produire rapidement un prototype ; cette limite est inhérente à la nature même du modèle conceptuel, qui n'est pas formel ;
- les limites du formalisme des schémas pour rendre compte d'un modèle de tâche ;
- le manque d'indications pour assurer le passage du modèle conceptuel à la base de connaissances proprement dite ;
- l'insuffisance en matière d'outils ou d'éléments méthodologiques quant à l'utilisation de ressources documentaires comme la base de données des incidents, exploitée à la main ;
- la lourdeur et le coût des entretiens et du travail avec des experts ;
- la nature peu synthétique, très descriptive du modèle conceptuel obtenu à l'issue du processus ;
- la rigidité du processus qui anticipe mal les problèmes de maintenance ou la nécessité d'un éventuel apprentissage de connaissances du domaine.

4. 3 - MACAO-II : Le modèle conceptuel comme grille d'acquisition

4.3.1 Contexte

4.3.1.1 Modélisation et ingénierie des connaissances

Les recherches en acquisition et ingénierie des connaissances ont connu autour de 1992 un essor considérable. Entre 1990 et 1994, cinq projets européens au moins ont été lancés afin de pousser plus loin les propositions du projet KADS. De nombreux livres, de numéros spéciaux de revues sont parus, tant des recueils d'articles que des monographies basées sur la pratique de cogniticiens. Tous affirment que l'orientation « modélisation des connaissances » soulève de nouveaux enjeux liés à l'imbrication étroite et maintenant bien identifiée entre le rôle du système visé, le modèle à construire pour le définir et la manière de recueillir les connaissances à organiser dans ce modèle. Les questions de recherche abordées définissent le domaine de l'ingénierie des connaissances ou *knowledge engineering*. Tout d'abord, l'accent est mis sur le processus de construction de modèles, à partir des pratiques d'experts mais aussi de choix arbitraires, et de la nature désormais constructiviste de l'approche. Puis les analyses mettent mieux en avant une deuxième facette de la modélisation, le rôle du modèle comme grille d'acquisition. Ce rôle est bien illustré par exemple par les systèmes OMOS (Linster, 1993) ou Protégé-II (Musen, 1993).

Une caractéristique de cette période est le très fort impact des résultats du projet KADS. Certaines propositions issues de KADS ont très vite fait office de référence car elles ont concrétisé la convergence d'idées venant de plusieurs travaux et d'influences jusque-là moins bien explicitées. Par exemple, la nécessité de faire référence au génie logiciel était primordiale pour la diffusion des travaux du domaine, et KADS a permis de mieux l'affirmer. C'est aussi suite à ces travaux (et d'autres) qu'il est devenu systématique de considérer le modèle construit comme celui du système

et non de l'expert, de décrire le modèle du raisonnement en termes de tâches et méthodes, ou encore de privilégier la réutilisation lorsque cela était possible. L'influence de KADS se retrouve dans les systèmes et approches proposés à partir de 1995. Par exemple, dès 1993, Protégé a évolué vers Protégé-II et a intégré des bibliothèques de modèles de résolution de problème.

Un autre courant d'influence est lié à la remise en question des SBC, évoquée en 2.2. Ces analyses débouchent sur une focalisation non plus sur le contenu des modèles mais sur la manière dont ils sont utilisés en interaction avec l'utilisateur au sein de systèmes dits coopératifs. La visée du processus d'ingénierie est alors plus ambitieuse : il s'agit d'améliorer les performances du couple système-opérateur et non d'optimiser les capacités de résolution du système seul. La vue anthropomorphique, selon laquelle le système développe des modes de résolution de problème identiques à ceux de l'expert, est abandonnée au profit d'une vue coopérative (Soubie, 1996).

4.3.1.2 Émergence de nouveaux thèmes de recherche

Les thèmes de recherche d'actualité entre 1992 et 1998 s'intègrent dans la problématique de la modélisation présentée au 2.2 et voient l'émergence de la notion d'ontologie.

Sur les modèles conceptuels, de nombreuses recherches ont porté sur les langages *d'opérationnalisation des modèles* ou sur l'utilisation de la logique pour leur validation formelle. En effet, la rupture étant trop marquée entre modèles conceptuels et règles de production, il a semblé nécessaire de définir des langages permettant de retrouver la structure du modèle dans le système opérationnel. Des langages réflexifs ont été définis pour assurer plus de flexibilités et favoriser le développement des systèmes coopératifs.

D'autres travaux ont revu la notion de *méthode de résolution de problème et de composant réutilisable* : les éléments de la bibliothèque de KADS étant jugés parfois mal adaptés, on a cherché à réutiliser des éléments plus simples (les « mécanismes » de Protégé-II, les « pas d'inférence » dans Astrée ou CommonKADS). L'objectif était de pouvoir adapter plus facilement les modèles et de mieux qualifier la notion de type de problème (ainsi, CommonKADS suggère désormais qu'un projet peut relever de plusieurs types de problème dont les enchaînements sont contraints).

La volonté de favoriser *l'interopérabilité* entre systèmes a conduit à définir des langages d'échange (comme KIF aux USA) puis au-delà du format, à étudier des modèles de données ou de connaissances qui puissent être échangées, les *ontologies*. Ce nouveau regard sur les connaissances d'un domaine a convergé avec une réflexion sur une meilleure articulation entre les différents niveaux d'un modèle. Ces niveaux sont moins indépendants que les hypothèses initiales (d'interaction limitée) le supposaient. Des notions comme celle *d'engagement ontologique* d'une méthode sur le niveau domaine ont été définies pour rendre compte de l'impact des niveaux les uns sur les autres.

Enfin, c'est au cours de cette période que ces principes de modélisation ont été appliqués en dehors du champ strict des systèmes à base de connaissances, comme pour l'utilisation des modèles de tâches pour définir l'interaction homme-système.

4.3.1.3 Élargir le spectre des sources de connaissances : apport des sciences humaines

Prise en compte des aspects sociaux, les connaissances dans l'organisation

Dans la lignée des recherches faisant intervenir ergonomie et psychologie cognitive sur la définition de systèmes d'aide à l'opérateur, un courant de recherche s'est constitué, plaçant l'utilisateur au centre du processus de conception du système (Clancey, 1992). Très proches de l'ingénierie des connaissances et faisant aussi appel à la modélisation conceptuelle, ces travaux ont la particularité de mettre autant l'accent sur l'interaction homme-système que sur les capacités de traitement du système (la résolution de problème). Une autre originalité est la prise en compte de la

situation de travail au sein de laquelle le système est utilisé, bien au-delà de la sphère de l'individu, pour rejoindre celle de l'organisation au sein de laquelle il travaille. Des analyses menées par des sociologues, psychologues ou ergonomes montrent l'enjeu qu'il y a à sous-estimer l'impact des choix de conception dans les différentes sphères de l'activité de l'utilisateur. Pratiquement, ces travaux ont conduit à quelques évolutions méthodologiques en ingénierie des connaissances et surtout à la définition de modèles complémentaires au modèle conceptuel : le modèle de coopération et le modèle d'interaction homme-machine (Schreiber *et al.,* 1994) (Zacklad, 2000). Dans notre équipe, une architecture de système coopératif a été proposée pour définir les principes fondateurs de ces systèmes (Hadj Kacem *et al.,* 1993a). Elle fait appel à plusieurs types de modèle conceptuel à côté de celui du système : un modèle de l'utilisateur, un modèle de l'environnement et un modèle de la coopération homme-système gérant le contrôle de l'exécution des tâches et la communication homme-machine.

Une autre illustration de cette analyse est l'étude menée dans l'équipe sur la validation des SBC du point de vue non seulement de la qualité intrinsèque des résultats fournis, mais aussi de leur pertinence pour l'utilisateur et pour garantir l'utilisabilité du système final. Ergonomes et informaticiens ont étudié l'attitude d'utilisateurs vis-à-vis de différentes applications (Chabaud *et al.,* 1989). La fréquente mauvaise utilisation des systèmes ou même leur rejet ont souligné la nécessité d'une validation à chaque étape du cycle de vie du système et même au sein du processus de modélisation (Soubie, 1996). Les analyses en amont et les validations précoces des principes d'interaction et de résolution du système doivent permettre d'éviter de mal répondre aux besoins des utilisateurs [IEA, 91].

Les propositions les plus originales ont d'abord été situées à la marge de l'ingénierie des connaissances (IC) car les systèmes répondant aux besoins ne font appel que de manière partielle à l'IA. Aujourd'hui, la plupart des applications sont dans ce cas : il s'agit souvent de systèmes d'information, de gestion de connaissances ou d'aide à la réalisation de tâches qui comportent quelques modules ou un noyau faisant appel à des connaissances. Cette convergence justifie un regain d'intérêt de l'IC pour les approches coopératives et sociologiques.

Exploitation des textes comme source de connaissances

Comme le montre D. Bourigault dans sa thèse (Bourigault, 1994), l'acquisition de connaissances à partir de textes a été abordée tout d'abord d'un point de vue très finalisé, laissant de côté toute considération linguistique ou terminologique. Les avancées en matière de représentation terminologique ainsi que le recul des terminologues sur leurs approches et pratiques ont permis de souligner la convergence avec l'ingénierie des connaissances (Skuce et Meyer, 1992). Ce tournant a donné lieu vers 1992 à de nouvelles initiatives. Il correspond de plus à l'atteinte d'une certaine maturité des recherches sur le traitement automatique du langage naturel, et donc à la disponibilité de logiciels adaptés à l'analyse de textes par des non-linguistes, comme les extracteurs de termes ou les étiqueteurs. En effet, un certain nombre de travaux ont remis en question des approches très générales ou très formelles, comme des analyses syntaxiques complètes, une véritable compréhension du langage à l'aide de grammaires ou encore une représentation formelle permettant de raisonner sur différents aspects des connaissances exprimées à l'aide du langage. C'est après les années 1980 que des approches plus pragmatiques ou plus focalisées ont montré leur meilleure efficacité et robustesse face à des données réelles. Ces approches ne prétendent pas traiter toutes les facettes de l'analyse du langage mais des problèmes précis et ainsi apporter une aide à certains types d'analyses ou d'exploitations de textes en langage naturel. Parmi celles-ci, on peut citer l'analyse contextuelle (Desclès & Minel, 1995) ou encore les analyses syntaxiques partielles des étiqueteurs.

4.3.1.4 Objectifs de MACAO-II

Dans ce contexte, poursuivre les travaux de recherche dans la lignée de MACAO se justifiait parce que MACAO était une des rares méthodes mettant en avant une démarche ascendante. Il

semblait important d'évaluer la complémentarité de cette approche par rapport à des modélisations basées sur l'adaptation de méthodes et modèles génériques, et enfin parce que le support logiciel disponible pouvait être le support de l'évaluation de ces recherches.

Les objectifs de MACAO-II répondent aux limites de MACAO en élargissant les moyens proposés pour la modélisation. La modélisation y est étudiée jusqu'à l'opérationnalisation des modèles, pour savoir en quoi un modèle opérationnel peut être évalué à travers la simulation de son fonctionnement. La nature et représentation des méthodes de résolution de problème ont été revus à la lumière de résultats établis en psychologie cognitive et en intelligence artificielle. En particulier, il s'agit de juger en quoi la notion de rôle peut servir d'articulation entre les objets ou concepts d'un domaine et les éléments sur lesquels porte le raisonnement.

À ces objectifs initiaux, deux thèmes de recherche, issus des expérimentations, se sont rajoutés. Le premier est lié à la nécessité d'intégrer de nouvelles connaissances ou de réviser celles déjà modélisées au cours du cycle de vie du logiciel. Cela suppose de considérer le processus d'acquisition comme un processus itératif. Une partie des recherches sur MACAO-II a donc porté sur la maintenance du modèle conceptuel et de la base de connaissances associée. Le deuxième thème vise à diversifier les sources de connaissances par le biais de l'exploitation plus systématique de documents, et surtout par l'approfondissement des apports de l'ergonomie à l'étude de l'activité.

4.3.2 La méthode MACAO-II

De nouvelles propositions, enrichissements et évaluations méthodologiques constituent la première facette des résultats établis avec MACAO-II, la deuxième, développée dans la partie suivante, portant sur la représentation des connaissances. Ainsi, la méthode MACAO-II fait des propositions sur la manière d'intégrer l'analyse de l'activité (§ 4.3.2.1). Elle précise comment le modèle peut guider l'acquisition avec la notion de schéma de modèle conceptuel (§ 4.3.2.2). La méthode a été mise en forme dans un document (§ 4.3.2.3). Elle permet de gérer de manière cohérente la construction du modèle et sa maintenance (§ 4.3.2.4). Enfin, elle s'ouvre à un nouveau type de sources de connaissances : la terminologie et les documents du domaine (§ 4.3.2.5).

4.3.2.1 Une approche ergonomique de l'activité

Les expériences menées avec MACAO, comme le projet SAMIE, ont permis de collaborer intensivement avec des ergonomes, de mettre en pratique un certain nombre de recommandations et pratiques liées à l'analyse de la tâche des spécialistes et des futurs utilisateurs et, par là même, de confirmer le caractère indispensable de ces études. En effet, le système à construire va s'intégrer dans l'activité de l'utilisateur et, à ce titre, le contexte d'usage doit être pris en compte dès la modélisation, soit pour faire un modèle spécifique (CommonKADS prévoit un modèle de l'organisation (Wielinga *et al.,* 1992)) soit pour moduler le modèle conceptuel par les attentes de l'utilisateur et spécifier les interactions qu'il aura avec le système. La méthode MACAO-II recommande donc de mener très tôt **l'analyse de la tâche** des spécialistes et des utilisateurs à partir d'observations et de simulations, ainsi qu'une modélisation de l'organisation dans laquelle cette tâche se déroule. Elle donne également des indications en matière d'analyse de besoins.

Comme MACAO, la méthode MACAO-II insiste sur le fait que le **modèle** conceptuel produit soit **compréhensible par les différents intervenants** du projet. Pour cela, la représentation des connaissances proposée, MONA, cherche à faciliter la documentation précise du modèle ainsi que le suivi des différentes présentations des connaissances, leur trace, au cours de leur mise en forme, depuis leur formulation par l'expert jusqu'à leur opérationnalisation. La partie conceptuelle des structures de représentation des connaissances favorise l'interprétation humaine et sert de support au dialogue entre expert, cogniticien et programmeur. La partie opérationnelle établit un lien entre le modèle et le code correspondant dans le système final, servant de référence commune au programmeur et au mainteneur.

Des principes d'ergonomie ont également été appliqués à la mise au point du **guide méthodologique MACAO-II** et du logiciel de modélisation lui-même. Le logiciel a fait l'objet d'une évaluation ergonomique qui a débouché sur l'ajout d'aides en ligne. Le guide présente les quatre grandes étapes de la méthode et leur réalisation sous forme d'une arborescence de tâches (IRIT-94-17-R, 94). La mise en forme de la méthode avait également bénéficié des efforts de présentation de ces caractéristiques dans le cadre de Sisyphus 2. Dans le guide, chaque tâche est située dans son contexte de manière chronologique et par rapport à l'avancement de la modélisation. Toutes les tâches sont décrites de manière structurée selon un plan identique, incluant une phase de validation, et illustrées d'un exemple. Lorsqu'elles font appel au logiciel associé, les fonctionnalités requises et leur mode d'utilisation sont présentés.

4.3.2.2 Évaluations expérimentales liées à MACAO-II

Ariane 4

Le premier retour d'expérience de MACAO-II sur des données réelles correspond à une application de diagnostic dans le domaine spatial, au sein du laboratoire ARAMIIHS. Un système d'aide au diagnostic a été développé en 1992 pour faciliter l'intégration et les tests de la case à équipements du lanceur Ariane 4. MACAO-II a été utilisé pour modéliser les connaissances de l'expert, expérience rapportée dans (Rapport-SOLER, 92]. La représentation des connaissances s'est faite à l'aide de schémas pour le raisonnement et d'un réseau conceptuel pour le domaine. La démarche adoptée suit le processus ascendant de MACAO-II. Un prototype du système a été développé puis validé opérationnellement. Ce projet a ensuite servi de terrain d'étude pour un projet sur l'apport des modèles conceptuels à la maintenance des systèmes à base de connaissances.

SADE

Les contrats liés à l'attribution de prêts immobiliers aux personnels des entreprises E.D.F. et G.D.F. et leur suivi sont gérés sous forme de dossiers. Si un contrat n'est plus honoré, le dossier est fermé. Il a été décidé de développer un système expert, SADE, pour capitaliser le savoir-faire des personnes chargées des dossiers fermés, leur permettre une meilleure analyse des dossiers en amortissement de prêts et mettre en oeuvre des procédures de recouvrement de créances mieux adaptées [Rapport-LEPINE, 93]. Le traitement des dossiers ne nécessite pas seulement l'application de règles juridiques écrites (clauses du contrat de prêt et réglementation). Il s'appuie également sur l'expérience des experts. Il est aussi influencé par plusieurs aspects (social, politique et économique) du recouvrement des créances. L'expertise a été modélisée selon une démarche essentiellement basée sur MACAO-II mais reprenant aussi des résultats de KADS. L'analyse comparée des deux approches est développée dans [JAC, 94a] et reprise dans [Livre-AC, 96b]. Pour les besoins de ce projet, MACAO, jusque-là disponible sur station de travail SUN, a été porté sur PC.

Dans le projet SADE, la documentation juridique autant que les dossiers comportent une partie importante de la connaissance à modéliser et contiennent tout le vocabulaire du domaine. Ce projet était donc un terrain favorable pour expérimenter l'utilisation de logiciels de terminologie pour accélérer la modélisation du domaine et en définir la terminologie, tâche qui fait partie de la première étape de MACAO. Le projet ayant lieu dans le cadre d'une collaboration avec la DER d'EDF, un logiciel d'extraction de terminologie, LEXTER, basé sur des principes novateurs et tout à fait adapté au repérage des concepts d'un domaine, a aussi été utilisé. Le corpus analysé a été choisi en fonction de l'expertise à représenter. LEXTER s'est avéré un support très efficace pour réduire le temps consacré à l'inventaire, la structuration et à la définition des concepts du domaine nécessaires à la modélisation de l'expertise [JAC, 94b] repris dans [Livre-AC, 96b].

4.3.2.3 Expliciter une méthode de résolution de problème

La démarche préconisée par MACAO-II vise à produire des modèles conceptuels plus synthétiques que ceux qui sont obtenus avec MACAO. L'accent est mis sur la **caractérisation** de la résolution de problème qui se traduit au final par la représentation explicite d'un modèle de tâches décrivant la ou les méthode(s) de résolution de problème adoptée(s). Pour parvenir à ce niveau de description plus abstrait, deux axes sont proposés, traduisant une influence des approches par réutilisation : d'une part, profiter des modèles existants quand leur réutilisation est possible ; et d'autre part, dégager les aspects génériques des modèles d'expertise via le schéma du modèle conceptuel. Le premier axe a consisté essentiellement en une confrontation de MACAO-II et CommonKADS, développée dans la partie suivante.

Clarifier la nature, le rôle et la manière d'obtenir le **schéma de modèle conceptuel** a été un moyen de préciser à la fois comment le modèle conceptuel peut guider le processus d'acquisition et en quoi consiste une caractérisation plus abstraite de la résolution de problème. Cette notion, introduite dans le cadre général de la modélisation conceptuelle dans [RIA, 92] a été redéfinie au sein de MACAO-II [JAC, 94a]. Une étude plus complète sur sa construction et son utilisation dans MACAO-II est récapitulée dans la revue IJHCS [IJHCS, 94]. Elle s'appuie entre autres sur l'expérience ARIANE 4.

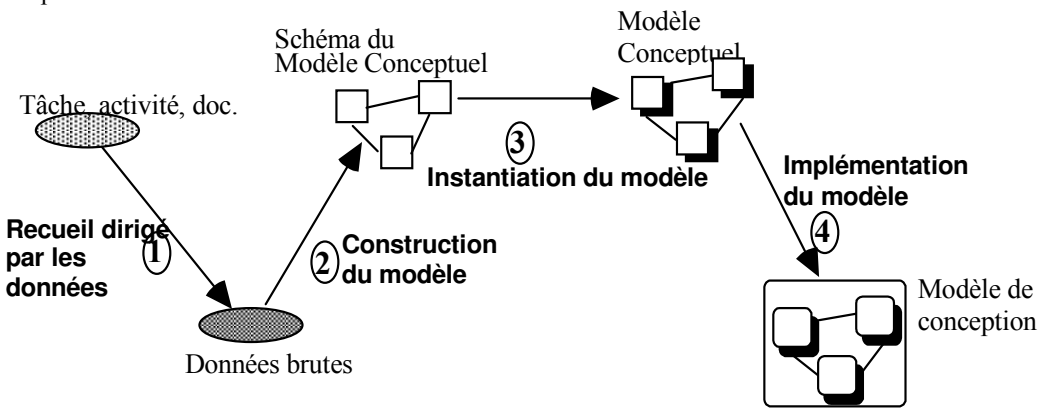

Figure 4.3.2 : Place du schéma du modèle conceptuel dans le processus de modélisation

Le Schéma du modèle conceptuel de MACAO-II est composé d'un Schéma du Modèle du Raisonnement et un Schéma du Modèle du Domaine (fig. 4.3.2). Le Schéma du Modèle du Raisonnement est une représentation abstraite du raisonnement, non détaillée et plus ou moins indépendante du domaine de l'application, à rapprocher de la notion de méthode de résolution de problèmes utilisée dans KADS. Exprimé dans MACAO à l'aide de la structure de *schémas*, il est montré sous forme d'arbre. Le Schéma du Modèle du Domaine correspond à la partie abstraite du modèle du domaine, au réseau des concepts et relations sans leurs instances.

Cette notion a été reprise et clairement définie dans MACAO-II. Le modèle du raisonnement y est représenté sous forme d'un arbre de tâches et méthodes. Le schéma du modèle de raisonnement correspond à une caractérisation de la méthode adoptée pour réaliser la tâche assignée au système. Par rapport au modèle complet du raisonnement, le schéma ne contient pas toutes les heuristiques ou les connaissances utilisées dans ce raisonnement. Construit de manière incrémentale, il guide le recueil de nouvelles connaissances dans la mesure où il permet d'identifier les lacunes et de rechercher les éléments qui vont donner une cohérence et du sens au modèle.

4.3.2.4 Combiner approche ascendante et réutilisation

La complémentarité entre approche ascendante (MACAO-II) et réutilisation (KADS puis CommonKADS) a été analysée grâce aux trois projets Sisyphus 2, ARIANE 4 et SADE [EKAW, 94]. La démarche MACAO-II commence par la modélisation de résolutions de problèmes

spécifiques. À partir de ces modèles, le cogniticien construit un modèle par catégorie de problème. Puis il doit parvenir à un modèle rendant compte de la démarche générale de résolution de problème. Ce processus d'abstraction en trois niveaux est représenté sur la figure 4.3.2.4. Selon MACAO-II, cette démarche ascendante peut être combinée à l'adaptation de modèles génériques proposés par KADS. Le processus d'abstraction impose de prendre le temps d'une analyse fine de l'activité experte, puis d'une étude précise des types de problème qu'il traite et de la manière dont il les résout. Ces tâches fournissent des éléments pour ensuite choisir et adapter avec plus de pertinence une méthode « générique » de résolution. La place accordée à ces recueils préalables centrés sur l'activité humaine et non sur des modèles a priori est une des originalités de MACAO-II. La méthode se démarque des démarches descendantes qui caractérisent rapidement la tâche et ne recherchent des éléments d'expertise qu'en fonction des questions qui se posent au cours de l'adaptation d'une méthode générique.

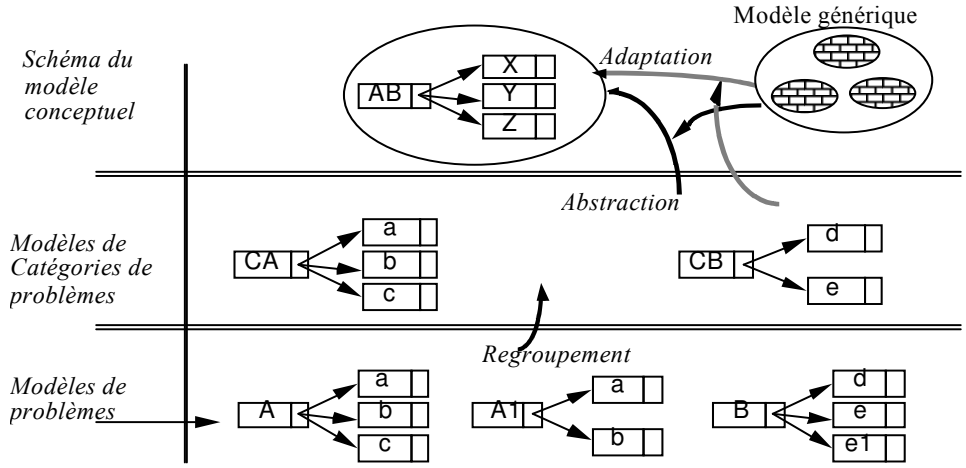

Figure 4.3.2.4 : Démarche ascendante et réutilisation dans MACAO-II. Les flèches sombres correspondent à un processus d'abstraction alors que les flèches grisées matérialisent une adaptation pour la réutilisation.

4.3.2.5 Maintenance des modèles conceptuels et SBC

Le modèle de diagnostic pour les tests d'intégration de la case à équipement d'ARIANE 4 a été repris en 1993 en intégrant des résultats de KADS en vue d'étudier le problème de sa maintenance en lien avec celle du système opérationnel[7]. Un spécialiste des modèles de diagnostic de KADS a identifié au plus vite le type de diagnostic effectué dans cette application, puis sélectionné un modèle de méthode de résolution de problème parmi les modèles disponibles dans la bibliothèque de CommonKADS [Rapport-BENJAMINS, 93]. Le modèle ainsi construit a été confronté à celui obtenu avec MACAO et présenté aux experts du domaine. Les deux modèles présentent des caractéristiques symétriques et complémentaires. Le modèle MACAO est très détaillé, concret, lié au domaine étudié et proche de la manière de raisonner des experts. Certaines étapes de raisonnement pourraient être regroupées ou généralisées. Le code est donc relativement morcelé. Le modèle obtenu avec KADS est très synthétique, plus abstrait mais les experts du domaine y reconnaissent difficilement leur approche. Il garantit une interprétation plus sûre par le cogniticien et permet d'organiser le code du système opérationnel. Clairement, ce dernier modèle favorise mieux la maintenance que le premier.

[7] Ce travail a été réalisé en collaboration avec l'équipe fondatrice de KADS à l'Univ. d'Amsterdam, dans le cadre d'un projet financé par un contrat avec la Région Midi-Pyrénées (MP-930, 94) et le programme « Sciences de la Cognition » (MEN-92-727, 94)

Une étude bibliographique du processus de maintenance, jusque-là encore peu exploré, a été réalisée. L'objectif était surtout d'identifier les conditions dans lesquelles un modèle conceptuel facilite la maintenance du logiciel qu'il permet de réaliser, puis de préciser les propriétés qu'il doit vérifier. Il s'agissait de définir des outils, méthodologiques et logiciels, appropriés pour assurer un gain et non un coût supplémentaire au cours de la maintenance. Ce projet a comporté également une analyse psychologique de l'activité de mise à jour d'un modèle conceptuel. Cette étude a débouché sur un rapport de DEA [DEA-SILVIN, 94]. Surtout, elle a eu des conséquences sur la méthode MACAO-II autant que sur la plupart des choix faits dans MONA. Aux deux niveaux, les principes de représentation des connaissances et l'ajout de commentaires favorisent la traçabilité des connaissances ainsi que la caractérisation du raisonnement à un niveau plus abstrait.

Le modèle ARIANE 4 a finalement été repris avec le nouveau formalisme, MONA, de MACAO-II pour valider les capacités de représentation de la résolution sous forme de tâches.

4.3.2.6 Documents comme sources de connaissances

Le nouveau regard porté à partir de 1992 sur l'exploration de textes, considérés comme des sources de connaissances, à côté des coûteux entretiens avec les experts, a ouvert de nouvelles perspectives pour la mise au point d'applications à base de connaissances. Jusque-là, l'utilisation des textes comme sources de connaissances se faisait à l'aide de trois types de logiciel (Bourigault, 1994) : des *outils de transfert*, visant la production automatique de structures de connaissances à partir de textes, des *outils hypertextuels* de navigation dans les documents découpés en fragments, et enfin des *outils de dépouillement* permettant d'associer les textes aux modèles, à titre de documentation. Or aucun de ces choix n'est véritablement satisfaisant, et c'est la combinaison des trois types de fonctionnalité qui peut apporter une aide efficace à la modélisation à partir de textes. De plus, la focalisation sur la modélisation des connaissances du domaine concerné, et avec elle sur la maîtrise du vocabulaire associé, permet d'envisager l'utilisation d'outils linguistiques et d'approches terminologiques (Skuce et Meyer, 1991). Ces logiciels facilitent une recherche focalisée des connaissances auxquelles la référence aux textes assure une plus grande validité.

Dans cet esprit, j'ai engagé deux types d'expériences autour de MACAO-II. L'une, autour de l'identification du vocabulaire et des concepts du domaine, a consisté à évaluer l'apport du logiciel d'extraction de terminologie LEXTER. Pour l'autre, j'ai utilisé le logiciel COATIS qui permet de repérer des connaissances causales à partir de textes afin de représenter des relations conceptuelles dans le modèle du domaine. L'ensemble de ces expériences a été possible grâce à plusieurs projets de collaboration avec la DER d'EDF (Clamart).

Ces deux initiatives ont fortement orienté la suite de mes recherches. Les perspectives ainsi ouvertes ont débouché sur les thèmes que j'ai développés à partir de 1995, et que je présente dans les chapitres suivants (5 et 6).

Extraction de terminologie et modélisation du domaine : LEXTER

LEXTER est un logiciel d'aide à l'acquisition et à l'interprétation de terminologie à partir de textes techniques. Ce logiciel réalise une analyse de corpus pour en dégager un ensemble de candidats termes, c'est-à-dire des groupes nominaux susceptibles de désigner des concepts du domaine (on se reportera à (Bourigault, 1994a ou 1994b) pour une présentation détaillée de son fonctionnement). Ces termes sont organisés au sein d'un réseau terminologique reflétant la manière dont ils entrent dans la composition les uns des autres. Une interface permet de naviguer dans ce réseau et de consulter la liste des termes produits selon différents critères numériques (fréquence ou productivité) ou syntaxiques (termes comportant un terme donné en tête ou comme complément). L'intérêt de ce logiciel est donc d'automatiser l'extraction des termes avec des résultats de qualité, mais aussi de faciliter l'exploitation des termes extraits grâce à l'interface de navigation.

Initialement conçu pour contribuer à la constitution et à la mise à jour de thesaurus, LEXTER a été utilisé à des fins d'acquisition de connaissances pour la première fois au sein du projet SADE,

en conjonction avec la méthode MACAO. Le système Sade a pour vocation de fournir des connaissances pour guider la gestion de dossiers d'accession à la propriété en amortissement de prêts. Cette expérience, récapitulée dans (Bourigault et Lépine, 1996), a montré comment LEXTER pouvait répondre à deux besoins liés à l'analyse de textes en acquisition des connaissances : (1) aide à la constitution de la terminologie du domaine ; (2) accès permanent et simple à la documentation du domaine. Elle a aussi permis de mettre en évidence différentes modalités d'utilisation des résultats de l'extraction terminologique. Lors de phases d'analyse ascendante des données identifiées, ce sont les fonctionnalités de dépouillement de LEXTER qui sont utiles : les critères numériques permettent de lister les termes selon différents points de vue et d'opérer une sélection, de regrouper les termes synonymes ou encore des relations de généricité entre termes. Ces critères facilitent aussi le repérage de termes importants et pertinents pour le domaine. Lors des phases d'acquisition descendante, guidées par le modèle construit, LEXTER est utilisé comme outil de fouille. À partir de termes pertinents, le réseau terminologique permet de repérer des termes plus précis ou portant sur les mêmes éléments, ainsi que les termes co-occurrents (voisins en corpus), qui peuvent avoir des relations sémantiques avec les termes étudiés.

Ce projet a eu le mérite de poser les bases de démarches depuis développées au niveau national (au sein du groupe TIA en particulier) et international (Mikeev & Finch, 1995) (Ahmad & Holmes-Higgin, 1995) ou (Maedche & Staab, 2000). Ces différents modes d'utilisation ont été par la suite revus ou affinés, et peuvent nécessiter d'avoir recours à d'autres types d'outils ou d'analyses. En particulier, la phase de validation systématique de la liste des termes ne semble plus utile dans le cas de la construction de modèles. Pour ce qui est de la fouille de texte, des logiciels comme les concordanciers ou encore des systèmes de fouille automatique de textes sont tout à fait complémentaires.

Extraction de relations de causalité et modélisation du domaine : COATIS

COATIS est un outil d'aide au repérage d'expressions d'actions reliées par des relations causales dans les textes techniques en français (Garcia, 1996). Les expressions d'actions ainsi trouvées constituent alors un index structuré du texte analysé. Pour cela, COATIS exploite le réseau de candidats à être des termes du domaine que produit l'extracteur de terminologie LEXTER à partir du texte. Mis au point par Daniela Garcia durant sa thèse, COATIS s'appuie d'une part sur un modèle en construction des relations causales telles qu'elles sont exprimées en français, et d'autre part sur la stratégie de l'exploration contextuelle. Deux expériences menées avec COATIS ont permis de montrer que ce système peut aider à l'acquisition des connaissances causales par exploration de textes, dans le cadre de la réalisation d'un modèle conceptuel. Les résultats de COATIS facilitent d'abord la compréhension du domaine à partir d'une lecture focalisée des textes. De plus, COATIS fournit des éléments pour construire le modèle : il permet de relever des concepts saillants, d'identifier des actions associées à ces concepts, de trouver ou de vérifier des règles du domaine et, surtout, d'établir un réseau des causalités entre ces concepts [JAC, 96] et [Livre-IC, 00]. L'aide apportée par COATIS diffère selon la vocation du modèle construit. Dans le projet HYPERPLAN au cours duquel COATIS a été utilisé, le système visé aide à la consultation de documents textuels. Pour juger de l'intérêt dans le cadre de l'aide à la résolution de problème, ces résultats ont été confrontés à différents travaux utilisant des modèles causaux pour l'aide au diagnostic technique (Charlet *et al*., 1996).

4.3.3 Représentation des connaissances : du langage naturel au système opérationnel

Dans MACAO, le modèle conceptuel est considéré avant tout comme un support pour comprendre et interpréter les informations provenant de l'expert et, seulement dans un deuxième temps, comme un moyen de formalisation. De ce fait, la représentation des connaissances cherche à favoriser des mises à jour faciles ou des représentations provisoires, utiles pour clarifier les idées du cogniticien. Elle sert de base à des visualisations simples et faciles à comprendre par un expert.

Cette représentation des connaissances offre ainsi d'abord un cadre minimal pour assurer la compréhension de l'expertise, puis elle permet de nouveaux raffinements avant d'être formalisée.

4.3.3.1 Des primitives paramétrables : MONA

Afin de faciliter le passage d'un modèle conceptuel à un modèle opérationnel, la représentation des connaissances dans le modèle conceptuel a été adaptée d'un langage opérationnel. Ce nouveau langage, MONA, un des résultats de la thèse de N. Matta, est le formalisme de représentation des connaissances défini pour MACAO-II [thèse-MATTA, 95] [rapport IRIT/94-02-R]. Il a été codé comme une surcouche de Le_Lisp (produit Ilog) et intégré au sein de la plate-forme MACAO-II.

Caractéristiques communes

L'originalité des structures MONA se situe dans leur découpage en trois parties, chacune étant relative à un degré de formalisation pertinent au cours de l'acquisition. La description de la structure conceptuelle comprend deux vues complémentaires : graphique et langage naturel. Une structure MONA est formée des trois parties suivantes :

- Dans la *description en langage naturel*, le cogniticien décrit les connaissances représentées, argumente et exprime ses choix de modélisation. Découpée selon un plan prédéfini que le cogniticien est libre de modifier, cette description est complétée par des liens vers des extraits de documents de l'expertise, les *références*.

- Une *expression formelle* rend explicites les relations entre structures pour rendre plus précise l'organisation des connaissances au sein du modèle. Les différents éditeurs graphiques du logiciel associé à MACAO-II facilitent la définition de ces relations car ils rendent transparente la syntaxe formelle.

- Une *expression opérationnelle* établit le lien entre la structure MONA et la partie correspondante dans le modèle opérationnel en LISA (Delouis et Krivine, 1995). Le passage des structures MONA vers le code LISA se déroule en deux temps : une traduction automatique génère un cadre opérationnel puis un codage manuel vient le compléter. Ceci permet de respecter le plus possible l'organisation des connaissances du modèle conceptuel dans le modèle opérationnel.

Structures du langage MONA

De manière très classique, les connaissances du domaine sont représentées, au niveau terminologique, par des *termes*, et, au niveau conceptuel, à l'aide de *concepts* et de *relations* étiquetées entre concepts. Les concepts sont génériques. Leurs instances peuvent également être représentées. Les termes sont associés aux concepts qu'ils désignent. Tout nouveau type de relation doit être situé dans une hiérarchie de relations prédéfinies. Cette classification des relations a pour objectif de mieux en comprendre leur signification avant que leur sémantique ne soit définie par le typage des concepts reliés (co-domaine - valeur).

La représentation du raisonnement a évolué par rapport à MACAO. À partir d'une collaboration avec I. Delouis, le langage LISA (Delouis et Krivine, 1995) a été retenu comme point de départ. LISA est un langage de modélisation opérationnel, réflexif, et permettant d'adapter dynamiquement la résolution de problème à l'utilisateur. Le contrôle gérant l'exécution tient compte du contexte pour choisir, pour chaque tâche, la méthode qui va la réaliser. Le langage MONA, correspondant à l'équivalent, au niveau conceptuel, des primitives de LISA, a alors été défini. MONA propose donc, comme LISA, des structures de *tâche* et de *méthode* pour représenter le raisonnement. Comme montré sur la figure 4.3.3.1, tâches et méthodes sont d'abord décrites par des commentaires puis à l'aide de concepts, de rôles et de relations avec d'autres tâches, et enfin opérationnalisées en LISA.

À la différence de LISA, MONA propose aussi une structure de *rôle* pour décrire les entrées et sorties des méthodes. Notion reprise à de nombreux travaux comme OMOS (Linster, 1993), CML (Schreiber *et al.,* 1994) ou VITAL (Leroux *et al.,* 1993), un rôle est utilisé pour nommer de manière abstraite les objets du raisonnement. Cette structure a été retenue dans MONA afin de parvenir à une meilleure lisibilité des modèles et à une meilleure caractérisation de la résolution de problème dans le schéma du modèle conceptuel. Ainsi, on espère pouvoir plus facilement réutiliser des modèles empruntés à d'autres projets et aboutir à des modèles plus réutilisables.

Tout comme dans MACAO, le modèle du domaine forme un réseau proche d'un réseau sémantique : le *Graphe du domaine, générique ou instancié* suivant que sont représentés des *concepts* ou leurs instances. Ces graphes peuvent être vus, entièrement ou partiellement, à l'aide d'éditeurs de graphe, qui en facilitent la construction progressive.

Tâche
- nom : Affectation-bureau
- définition : cette tâche concerne ....
- but : personnes-placées
- contraintes/contexte : les bureaux doivent contenir toutes les personnes
- contraintes/but : toutes les personnes doivent être placées
- critère de satisfaction : on ne cherche pas une solution optimale.

Méthode
- nom : Affecter
- définition : c'est une méthode ...
- paramètres : personnes, bureaux, contraintes.
- résultats : personnes-placées
- contraintes/paramètres : les bureaux doivent contenir toutes les personnes
- contraintes/résultats :
- contexte favorable : l'affectation se fait en respectant des contraintes
- traitement : TANT QUE il-existe-des-critères : choix-critère, affectation-personnes/critère

Rôle
- nom : composant
- définition : un composant à placer
- connaissances du domaine : personnes.

Figure 4.3.3.1 : Exemples de structures de représentation du raisonnement en MONA.

4.3.3.2 Simuler le comportement du modèle conceptuel

À l'issue du processus de modélisation, la version formelle du modèle des connaissances en LISA est appelée *modèle formel* ou modèle opérationnel. L'opérationnalisation permet entre autres de valider le comportement du modèle et de le vérifier syntaxiquement. La correspondance entre structures des modèles conceptuel et opérationnel a pour but de propager facilement au niveau conceptuel les modifications requises à partir de problèmes constatés au niveau opérationnel. Parvenir jusqu'au modèle opérationnel présente plusieurs avantages : on en attend une meilleure précision du modèle et un passage plus facile à la base de connaissances finale.

La méthode MACAO-II prévoit plusieurs étapes dans le passage du modèle conceptuel au modèle opérationnel : au début, des structures LISA incomplètes correspondant aux structures MONA du modèle sont générées automatiquement ; ensuite, le cogniticien doit interpréter le contenu des structures MONA pour les compléter. La plate-forme MACAO-II intègre le support de cette transition. Enfin, cette possibilité a ouvert des perspectives de prototypage et de construction incrémentale du modèle.

4.3.3.3 Une autre forme d'opérationnalisation : prototypage avec ZOLA

En LISA, il n'y a pas de rôle : les tâches et les méthodes manipulent directement les objets du domaine (concepts et relations). Les concepteurs de LISA justifient ce choix par le fait qu'au niveau opérationnel, le modèle n'a plus lieu d'être générique ou abstrait : il doit être le plus adapté possible au domaine pour être performant.

Une nouvelle expérimentation de l'opérationnalisation des modèles MONA a donc été menée à l'aide d'un autre langage réflexif, Zola (Isténès et Tchounikine, 1996), qui permet de coder des primitives opérationnelles adaptées à des structures conceptuelles. En développant ce langage opérationnel, appelé ZTM, dans un premier temps, les définitions des structures MONA ont été précisées, puis la notion de rôle a été redéfinie [IRIT-96-23-R]. Ainsi, le passage des rôles à des concepts opérationnels a souligné l'importance de conserver une trace du contexte dans lequel les connaissances du domaine peuvent jouer ce rôle. Le contexte associé à un concept est précisé à l'aide d'un nom de tâche ou de méthode et du nom du champ (entrée-sortie-ressource) qui indique comment les connaissances sont utilisées dans la structure. De plus, afin de bien gérer l'association entre rôles et concepts du domaine tout au long d'une résolution, des liens fixent les concepts pouvant jouer chaque rôle dans les différents contextes de leur utilisation, comme illustré sur la figure 4.3.3.3 pour le rôle « ensemble de composants ».

Rôle : ens-de composants
Valeurs Possibles:
Val 1 Lien vers un concept: ens-de-personnes
Contexte : entrée, tâche Affecter-emplacement-composant
Val 2 Lien vers un rôle: ens-de-composants, entrée, tâche Affecter-emplacement-composant
Contexte: paramètre, méthode affectation-selon-critères
Val ...

ens-de-composants désigne les entrées de la tâche *Affecter-emplacement-composant* (valeur 1) et les paramètres de la méthode *Affectation-selon-critères* (value 2). Pour la valeur 1, *ens-de-composants* peut être assimilé à un rôle statique : il indique les concepts du domaine pouvant être utilisés en entrée de la tâche *Affecter-emplacement-composant*. Pour la valeur 2, *ens-de-composants* peut être assimilé à un rôle dynamique ; il n'est pas relié à des connaissances du domaine, mais fait référence à un autre rôle. Il signifie que les paramètres de la méthode *Affectation-selon-critères* correspondent aux entrées de la tâche *Affecter-emplacement-composant*.

Figure 4.3.3.3 : La notion de rôle dans MACAO-II.

Le parti retenu dans MONA est de représenter les rôles par des structures indiquant en extension les connaissances du domaine pouvant les jouer mais ne les caractérisant pas. La limite évidente de ce choix est qu'on doit prévoir tous les concepts susceptibles de jouer le rôle, ce qui rend la description du rôle dépendante du domaine, ce qui limite l'intérêt de les définir.

4.3.3.4 Complémentarité avec ASTREE

Mis au point par F. Tort et C. Reynaud au LRI, ASTREE est un logiciel d'aide à la modélisation du raisonnement à partir d'une analyse des caractéristiques structurelles des concepts, définis au sein d'une ontologie du domaine (Tort, 1996). Le modèle du raisonnement est décrit sous forme de tâches et méthodes adaptées de LISA, comme pour MONA. Les entrées et sorties d'une méthode (appelées paramètres et résultats) sont les mêmes que les entrées et sorties (contexte et but respectivement) de la tâche qu'elle réalise, et font dans les deux cas référence aux termes de l'ontologie du domaine. Les méthodes sont identifiées par interprétation des connaissances du domaine et notamment de la sémantique du langage Entité-relation (E/R) utilisé pour les représenter. L'originalité d'ASTREE est aussi de permettre d'exprimer des contraintes sur les concepts ou sur leurs valeurs.

Plusieurs convergences fortes justifient d'essayer d'intégrer ASTREE et MACAO : un des points faibles de MACAO-II est la représentation du domaine ; le modèle du raisonnement dans ASTREE utilise les mêmes structures que celui de MACAO ; enfin, ASTREE est un outil supplémentaire pour aider à organiser systématiquement le modèle du raisonnement en adéquation avec celui du domaine (l'ontologie). Pratiquement, une expérience a été conduite en collaboration avec le LRI au sein du projet Hyperplan (chapitre 6). Le langage de type entité-association d'ASTREE a été utilisé pour représenter le modèle du domaine. Les règles de normalisation et de structuration associées se sont avérées des aides pertinentes pour définir un modèle plus compact et vraiment en adéquation avec les raisonnements effectués. Elles influencent également le niveau auquel sont définis les propriétés des concepts et leur nature. Au final, ce modèle comporte moins de concepts qu'un premier modèle fait avec MACAO-II, il est plus explicite et mieux structuré. Cette expérience a confirmé l'intégration possible entre MACAO-II et ASTREE, et l'importance de règles de structuration comme celles de normalisation en base de données [CMKB, 97].

4.3.3.5 La notion de rôle

Ces deux collaborations ont souligné le besoin d'approfondir la notion de *rôle* avec l'équipe de l'IRIN, qui a défini une surcouche de ZOLA, le langage DSTM, et l'équipe du LRI développant ASTREE. Cette analyse conjointe a eu aussi pour but de mieux faire converger les différents travaux liés à MACAO, en particulier ASTREE, en s'appuyant sur la philosophie de DSTM. À partir d'une étude de l'état de l'art, il est ressorti que la notion de rôle recouvrait des réalités très différentes qu'une étude conjointe avec le LRI et l'IRIN a permis de sérier [JICAA, 97] [EKAW, 97]. A minima, le *rôle* est juste un label qui, parce qu'il est plus abstrait, indépendant du domaine et relatif au raisonnement, permet au lecteur du modèle de changer de point de vue sur les connaissances du domaine. Les *rôles* sont parfois exprimés sous forme de contraintes portant sur les concepts du domaine et leurs propriétés, sans pour autant être étiquetés. Enfin, dernier cas de figure, les *rôles* peuvent être définis explicitement, soit en tant que tels comme dans VITAL (Leroux *et al*., 1994) soit au sein des méthodes.

Bien que n'utilisant pas la notion de *rôle*, ASTREE permet de caractériser les entrées et sorties des tâches et méthodes au même niveau d'abstraction, et ceci grâce à des contraintes et à un degré élevé de structuration. Il est non seulement possible de définir la nature et la structure des entrées et sorties mais aussi leur syntaxe et leur statut. Ce choix correspond au 2e cas ci-dessus. Dans ASTREE, l'absence de structure *rôle* est justifiée par le fait que l'objectif n'est pas la description de méthodes génériques. Néanmoins, son ajout a été envisagé pour mieux articuler raisonnement et modèle du domaine.

DSTM est un noyau d'opérationnalisation de modèles conceptuels qui propose une architecture en 3 niveaux de manière à définir des systèmes réflexifs raisonnant sur leur contrôle. Les primitives de modélisation du raisonnement (tâche et méthode) se trouvent au niveau le plus bas. Ensuite, des notions abstraites sont définies au niveau au-dessus, comme la notion de « méthode déclenchable » ou de « méthode-favorable ». Ces notions sont manipulées par des actions de haut niveau (comme « identifier les méthodes déclenchables ») définies au 3e niveau de l'architecture. La notion de rôle peut donc intervenir à deux niveaux : au niveau du raisonnement, les rôles caractérisent les manières dont les connaissances du domaine sont manipulées dans le raisonnement, et au niveau abstrait, les notions abstraites définissent les rôles que jouent des méthodes au sein des actions de haut niveau.

Enfin, dans MONA, la première manière de représenter les rôles en ZOLA correspond non à une caractérisation mais à un lien fixe entre domaine et raisonnement. Ce choix, bien que non satisfaisant si on cherche à rendre indépendant du domaine le modèle du raisonnement, est pratique du point de vue opérationnel. Il souligne l'importance de définir les rôles comme des structures à part entière, ainsi que l'atout de la souplesse de ZOLA pour les définir facilement avec les opérateurs qui les manipulent dans le système opérationnel. Les propositions d'Astrée fournissent

alors des outils pour une caractérisation syntaxique qui réponde à l'exigence d'une définition explicite et indépendante du domaine. Cependant, ces perspectives n'ont pas été poursuivies.

4. 4 - Bilan sur la modélisation conceptuelle

4.4.1 Synthèse des résultats établis

Les méthodes MACAO puis MACAO-II constituent mes réponses aux questions posées au chapitre 3 sur la modélisation conceptuelle.

Les premières recherches sur MACAO répondent en partie aux questions de la représentation des connaissances au sein de modèles, de la nature de ces modèles et des aspects méthodologiques (questions 1, 2, 3 et 6). Ces propositions posent les bases des résultats établis avec MACAO-II et même de nos choix en matière de construction d'ontologies à partir de textes. Elles présentent donc un intérêt qui dépasse le problème de l'acquisition des connaissances pour les systèmes experts car elles rejoignent les problématiques actuelles de l'ingénierie des connaissances :

- Pour identifier et recueillir des connaissances expertes (question 1), j'ai répertorié, caractérisé et adapté des **techniques et des outils**, en particulier les grilles répertoires, basées sur des entretiens et des simulations. J'ai posé en principe la nécessité de diversifier les techniques au cours d'un même projet et d'en exploiter la complémentarité. Je défends une approche ergonomique, basée sur l'étude des traces de la mise en œuvre des savoir-faire, au moyen de simulations ou d'analyses de l'activité.
- Concernant la nature et le statut des modèles conceptuels (question 2), j'ai choisi de privilégier le rôle de **support au dialogue entre acteurs de la modélisation,** pour apporter une aide véritable au cogniticien. Pour cela, un modèle doit être facilement compris et interprété. Il doit pouvoir être construit de manière incrémentale et être corrigé facilement.
- Pour faciliter la lisibilité des modèles et en garantir une interprétation de bonne qualité, j'ai établi trois principes que je retiendrai dans mes travaux suivants : (i) l'intérêt des **visualisations graphiques** des modèles ; (ii) la définition de **langages au niveau conceptuel**, en amont des représentations formelles ; (iii) la **traçabilité** depuis les « données brutes » (retranscriptions d'entretiens par exemple) jusqu'aux connaissances structurées dans le modèle, et donc la conservation de ces connaissances à différents degrés de structuration ou de formalisation.
- Dans MACAO, la représentation des connaissances proposée au niveau conceptuel est basée sur des structures appelées **schémas**. Elles regroupent les connaissances en fonction des déclencheurs des actions et des objectifs visés.
- Du point de vue méthodologique (question 3), MACAO préconise **l'approche ascendante** comme moyen privilégié pour rendre compte des modes de raisonnements humains. J'aurai par la suite le souci d'étudier principalement les manières d'aider le cogniticien dans toutes les **tâches d'abstraction** pour dégager des éléments conceptuels à partir d'exemples, de données ou de cas particuliers.

Du point de vue de la manière de mener ces recherches, cette période a confirmé le caractère interdisciplinaire de la problématique et donc la nécessité de **collaborer avec d'autres disciplines** comme la psychologie cognitive ou l'ergonomie. J'ai également mesuré combien il est indispensable, pour mener un travail d'ingénierie, de valider des travaux plus théoriques par des retours d'expérience, et pour cela **développer des systèmes en situation d'usage réel**.

MACAO-II a permis de nouvelles avancées à la fois méthodologiques et liées à la représentation des connaissances :

- Au niveau méthodologique (question 6), la notion de **schéma du modèle conceptuel** a été précisée. Dans MACAO-II, ce schéma revient à mieux caractériser le raisonnement ainsi que le rôle des connaissances du domaine utilisées pour ce raisonnement. Cette caractérisation est plus simple en cas de réutilisation de méthode prédéfinie, mais elle nécessite des structures de représentation des connaissances et une démarche poussant à plus d'abstraction. J'ai montré qu'une caractérisation abstraite de ce type assure une meilleure capacité de **maintenance** que des modèles plus descriptifs.
- Toujours pour répondre aux besoins méthodologiques, j'ai mené un travail approfondi sur la manière d'intégrer des propositions méthodologiques, l'utilisation d'une représentation des connaissances et d'outils de recueil ou d'analyse. Deux supports ont été définis : un **guide méthodologique** qui présente le processus et les recommandations de manière claire et le **logiciel MACAO-II**.
- Le **logiciel MACAO-II** a été utilisé par une dizaine de personnes dans le cadre d'études de cas mais surtout de 4 projets avec des entreprises (Ariane 4, SADE, SAMIE et HYPERPLAN) ; suite à des collaborations avec l'IRIN, le LRI et la DER d'EDF, il intègre de nouvelles approches (acquisition d'éléments de méthode à partir du domaine) et de nouveaux éléments de représentation des connaissances (schéma entité association, rôles, opérationnalisation en ZOLA).
- Pour qualifier les modèles obtenus avec MACAO-II (question 5), la confrontation entre **méthode ascendante et réutilisation** a été poussée suffisamment loin pour en montrer la complémentarité, souligner la nécessité de bien étudier les tâches et activités des acteurs du domaine avant de procéder à l'adaptation de modèles génériques. Chacune des approches conduit à des modèles de natures très différentes, plus ou moins performants en matière de résolution de problème et plus ou moins adaptés aux usages prévus du système final. L**'analyse de la tâche**, inspirée de l'ergonomie, conditionne la pertinence du système final auprès des utilisateurs.
- C'est grâce à une représentation des connaissances dissociant tâches et méthodes de résolution de problème que je réponds à la volonté de modéliser des connaissances sur les systèmes coopératifs (question 4). Les structures de Tâche et de Méthode seront d'ailleurs retenues comme briques de base pour la représentation de systèmes coopératifs dans les thèses de l'équipe qui ont suivi (Adj Kacem, 1995). Le **langage MONA** propose des primitives de tâche et méthode pour définir le modèle du raisonnement de manière assez souple. Les buts du système et les méthodes mises en œuvre sont caractérisés séparément. MONA permet de représenter des cas spécifiques autant qu'un modèle conceptuel. Ce langage est un support pour affiner progressivement des descriptions proches du langage naturel, les organiser dans des représentations structurées pour obtenir enfin des structures opérationnelles en LISA.
- Les liens étroits entre méthodes de raisonnement et structures du domaine peuvent guider la représentation des connaissances. Le langage MONA propose la **notion de rôle** pour expliciter les exigences de cette articulation. Des **règles de structuration** du domaine en lien avec le raisonnement constituent une aide précieuse pour organiser les deux niveaux de manière cohérente.
- Enfin, pour compléter les premières approches basées sur des entretiens et des analyses de la tâche, j'ai diversifié les sources de connaissances en m'intéressant aux textes. Les **approches linguistiques et les outils terminologiques** ouvrent de nouvelles perspectives en matière de recueil de connaissances à partir de textes. Les résultats prometteurs obtenus à l'aide de l'extracteur de termes LEXTER soulignent l'importance de pouvoir naviguer dans le réseau de termes extraits en fonction de critères liés à l'usage des termes. Les expériences utilisant COATIS ont confirmé l'intérêt de l'étude des relations pour localiser des contextes riches en information sur les concepts, idée reprise dans la thèse de P. Séguéla (chapitre 6).

4.4.2 Situation de MACAO-II par rapport à d'autres travaux

Plusieurs résultats font référence encore aujourd'hui à la modélisation des connaissances, telle que je l'entends dans ce chapitre : le système Protégé permet de générer des éditeurs pour saisir des instances de modèle ; la méthode CommonKADS est basée sur la réutilisation et l'adaptation de composants génériques ; et enfin les composants de l'expertise de Steels proposent une représentation des connaissances à l'aide de concepts, tâches et méthodes. Les travaux sur MACAO-II y ont une place plus modeste, mais reconnue pour ses spécificités aux niveaux national et international.

La première originalité de MACAO-II se situe avant tout au niveau méthodologique, par la place accordée à l'étude ergonomique de l'activité, aux méthodes de résolution mises en œuvre par les experts et à leur analyse selon une démarche ascendante. La méthode constitue une des alternatives solides à CommonKADS. Une autre particularité de MACAO-II est de proposer le langage MONA pour anticiper, au niveau conceptuel, les processus de sélection dynamique de tâches et méthodes en cours de résolution de problème, dans le modèle opérationnel. MONA se situe en amont de langages comme Lisa, Model-K et Omos ou les langages réflexifs : KARL, Zola puis DSTM. Enfin, un dernier point fort de MACAO-II correspond à sa plate-forme, qui est une des rares propositions intégrant à la fois un environnement de modélisation, des techniques de recueil et d'analyse de connaissances, et un support méthodologique.

4.4.3 Bilan

Finalement, le projet MACAO m'a permis de faire le tour de la plupart des questions liées à la modélisation des connaissances par un cogniticien. Pour assister le processus de modélisation, j'ai constitué un ensemble cohérent de propositions qui aident un cogniticien à accéder aux connaissances requises pour mettre au point un système à base de connaissances répondant aux besoins des utilisateurs. Concrètement, ces propositions sont organisées au sein d'une méthode, MACAO-II, et du support logiciel associé. Cette plate-forme offre des fonctions d'aide au recueil de connaissances, un langage de représentation des connaissances et des moyens de stockage pour gérer les informations recueillies et constituer progressivement un modèle conceptuel. Ces dix années autour de la méthode MACAO ont permis d'identifier des techniques pertinentes et de définir une approche méthodologique générale, indépendante du type de problème à traiter, pour construire des modèles conceptuels. Pour cela, j'ai choisi systématiquement de ne pas retenir le seul point de vue technique de l'informatique et de l'IA pour parvenir à un système opérationnel, mais de prendre la réponse aux besoins des utilisateurs dans toutes ses dimensions, avec une réflexion enrichie d'analyses venant de la psychologie et de l'ergonomie. J'ai également voulu privilégier l'analyse au niveau conceptuel en amont de l'opérationnalisation. Ma contribution est donc plus significative en ce qui concerne l'acquisition et le recueil de connaissances, le repérage et l'organisation du contenu d'un modèle que dans l'aide à la formalisation ou à la validation opérationnelle.

De plus, nos expériences sur l'acquisition de connaissances à partir de textes ont souligné le fort potentiel d'une prise en compte systématique et outillée des connaissances présentes dans les textes. L'identification va bien au-delà du vocabulaire, elle porte sur les concepts et les connaissances associées sous forme de relations ou de propriétés. Pour cela, l'approche terminologique est un premier pas. Les outils linguistiques permettent d'aller plus loin. Il ressort la nécessité de définir des logiciels spécifiques, facilitant l'identification des concepts non seulement à partir de l'étude des termes (leurs déviances par rapport à la langue générale, leur fréquence, etc.) mais aussi à partir de l'analyse de leurs contextes d'usage (cooccurrences, relations grammaticales et sémantiques avec d'autres termes, rôles syntaxiques et analyse des distributions en corpus).

Étudier les connaissances à partir de leur expression en langue soulève des problématiques sur la nature des connaissances et ce que signifie leur opérationnalisation dans un système. J'ai pu rencontrer des linguistes et spécialistes du traitement automatique des langues eux-mêmes

demandeurs d'expériences de validation de leurs propositions en matière de repérage de connaissances dans des textes. Ce contexte très favorable, tant du point de vue de la thématique que des personnes, m'a incitée à retenir cette orientation pour la suite de mes recherches.

4. 5 - Publications sur ces travaux[8]

[LSI-289] AUSSENAC N., SOUBIE J-L. *A knowledge acquisition tool based on a psychological model of reasoning*, Rapport LSI n°289. Janv 1988. 23 p.

[CREIS, 88] AUSSENAC N., MICHEZ B., MACAO : Application d'un modèle psychologique à la réalisation d'un outil d'aide à l'acquisition de connaissances, *Actes du colloque CREIS Représentation du réel et informatisation*, St Etienne, France. 1988.

[EKAW, 88] AUSSENAC N., SOUBIE J-L., FRONTIN J., A knowledge Acquisition Tool for Expertise Transfer, In *Proc. of EKAW'88*, GMD Studien Nr 143, 8.1-8.12. 1988.

[EKAW, 89] AUSSENAC N., SOUBIE J-L, FRONTIN J., RIVIÈRE M-H., A mediating representation to assist knowledge acquisition, *Proceedings of the 3rd European Knowledge Acquisition Workshop (EKAW 89),* Paris (F), July 1989. p 516-529.

[KAW, 89] AUSSENAC N., SOUBIE J-L, FRONTIN J., RIVIÈRE M-H., A mediating representation to assist knowledge acquisition, *Proc. of the 4th Knowledge Acquisition Workshop (KAW 89),* Calgary (CA), Oct 1989.

[Thèse-AUSSENAC, 89] N. AUSSENAC. *Conception d'une méthode et d'un outil d'acquisition des connaissances expertes,* Thèse en Informatique de l'Université Paul Sabatier, Toulouse, n° 523, oct. 1989.

[JAC, 90] AUSSENAC N., SOUBIE J-L., Place d'un outil d'acquisition de connaissances dans la conception des systèmes intelligents, *Actes des Journées d'Acquisition des Connaissances JAC'90*, Lannion, France, p 115-129.

[COGNITIVA, 90] AUSSENAC N., CHABAUD C., La place des savoir-faire dans l'acquisition des connaissances, *Actes du colloque COGNITIVA 90*, vol 2, Madrid (Espagne), Nov. 1990.

[Rapport-RIVIERE, 90] RIVIERE M.-H., *Mémoire d'ingénieur CNAM* - Toulouse, déc. Avril 1990.

[SELF, 90] CHABAUD C., AUSSENAC N., Conception interdisciplinaire d'un système d'aide au diagnostic : le projet SAMIE, *Compte-rendu du XXVIème Congrès de la Société d'Ergonomie de Langue Française : Méthodologie et outils d'intervention et de recherche en ergonomie*, Montréal (Canada), 1990 : 106-109.

[SAMIE, 90] AUSSENAC N., TOZEYRE J., CHABAUD C., VO D.P., CARRE F., *SAMIE : rapport préliminaire à la conception d'un système expert d'aide à la maintenance informatique.* Rapport interne MATRA-Espace (SAMI-RP-0-00-AR), Toulouse (F). Déc. 1990.

[EKAW, 91a] AUSSENAC N., DIENG R., Models of Problem Solving for Knowledge Acquisition : comparison of MACAO and 3DKAT, *Proceedings of the 5th EKAW 91*, SISYPHUS Project, part II, Crieff (Scotland), may 1991.

[EKAW, 91b] AUSSENAC N., Knowledge Acquisition with MACAO : the Sisyphus Case-Study, *Proceedings of the 5th EKAW 91*, SISYPHUS Project, part II, Crieff (Scotland), may 1991.

[KMET, 91] AUSSENAC N., FRONTIN J., SOUBIE J. L., Évolution d'une représentation des connaissances pour l'acquisition, *Knowledge Modelling and Expertise Transfert KMET 91*, HERIN-AIME D., DIENG R., REGOURD J.P., ANGOUJARD J.P. Eds. Amsterdam : IOS Press, 1991 : 135-148.

[IEA, 91] CHABAUD C., AUSSENAC N., SOUBIE J-L., A validation method of the validity of artificial intelligence as a work aid, *Proc. of IEA'91 International Ergonomics Association,* Eds. Y. Queinnec, F. Daniellou.- London : Taylor & Francis, 1991 : 619-621.

[DEA, 91] TESTEMALE G., COURBON, F. Représentation des connaissances. Mémoire de DEA RCFR de l'université Paul Sabatier Toulouse 3. sept. 1991.

[8] Présentation par ordre chronologique.

[Rapport-ROUTABOUL, 91] ROUTABOUL M. *mémoire d'ingénieur CNAM* - Toulouse, déc. 1991.

[GMD-630, 92] AUSSENAC-GILLES N., MATTA N., Making a method of problem solving explicit with MACAO*, Sisyphus 92*, Rapport GMD n° 630. Bonn (G). 1992.

[IRIT-92-24-R, 92] N. AUSSENAC, N. MATTA, *Étude de la notion de méthode de résolution de problème dans l'outil d'acquisition des connaissances MACAO : expérience du projet SISYPHUS, partie 2*. Rapport IRIT/92-24-R. Toulouse, sept. 1992.

[Rapport-SOLER, 92] SOLER C., Système Expert d'aide à l'intégration Ariane, *mémoire d'ingénieur CNAM* - Toulouse, déc. 1992.

[RIA, 92] N. AUSSENAC-GILLES, J.P. KRIVINE et J. SALLANTIN, Éditorial du numéro spécial « Acquisition des connaissances » de la *Revue d'Intelligence Artificielle* (RIA), Ed. N. Aussenac-Gilles, J.P. Krivine et J. Sallantin. Paris : Hermès. 1991/2, Vol 6 N°2. pp 7-18.

[RIS, 92] AUSSENAC N., FRONTIN J., SOUBIE J.-L., RIVIERE M.-H., Le problème de l'extraction des connaissances implicites : apports du système MACAO. *Revue Internationale de Systémique*, Vol 6, N°1-2. 1992. p 167-180.

[JAC, 93] AUSSENAC-GILLES N., MATTA N., Enjeux d'une acquisition des connaissances basée sur l'explication d'un modèle plus générique de l'expertise, *Actes des JAC 93,* St Raphaël, Mars 1993.

[Rapport-BENJAMINS, 93] BENJAMINS R. *Report of work at Aramiihs*, Toulouse, June 1993.

[Rapport-SAURET,93] SAURET M., Méthode MACAO : Manuel Utilisateur, *DESS d'ergonomie,* Toulouse, Juillet 1993.

[Rapport-LEPINE,93] LEPINE P., Contribution à la validation pratique de la méthodologie et de l'outil d'acquisition de connaissances expertes MACAO, *mémoire d'ingénieur CNAM* - Paris, déc 1993.

[JAC, 94a] AUSSENAC-GILLES N., MATTA N., Problèmes méthodologiques liés à la conception d'un modèle conceptuel avec MACAO, *Actes des 5èmes Journées d'Acquisition des Connaissances*, Strasbourg, mars 1994.

[JAC, 94b] LEPINE P., AUSSENAC-GILLES N.,. Modélisation de la résolution de problèmes : comparaison expérimentale de KADS et MACAO. *Actes des 5èmes Journées d'Acquisition des Connaissances*, Strasbourg, mars 1994 : H-1/H-14.

[EKAW, 94] AUSSENAC N., How to combine data abstraction and model refinement: a methodological contribution in MACAO. *A future for Knowledge Acquisition, Proc. of EKAW'94, 8th European Knowledge Acquisition Workshop*. Berlin: Springer Verlag. Series Lecture Notes in AI, N°867. 1994 : 262-282

[IJHCS, 94] AUSSENAC-GILLES N., MATTA N., Making a method of problem solving explicit with MACAO*, International Journal of Human-Computer Studies* . New York : Academic Press. **40**. 193-219. 1994.

[IRIT-94-02-R, 94] MATTA N., AUSSENAC-GILLES N.. *Structures de représentation des connaissances pour MACAO*. Rapport Interne IRIT/94-02-R. Toulouse. Mai 1994

[IRIT-94-17-R, 94] AUSSENAC-GILLES N., PIDOUX C.. *Guide méthodologique MACAO. Version1.1* Rapport interne IRIT/94-17-R. Toulouse. Oct. 1994.

[MEN-92-727, 94] AUSSENAC-GILLES N., SILVIN C., BENJAMINS R., SOUBIE J.L., BREUKER J., *Évolution et maintenance dans les bases de connaissances dans le cadre d'une coopération cognitive entre système humain et artificiel.* Rapport de fin de contrat MEN 92-727 Sciences de la Cognition. Décembre 1994.

[MP-930, 94] AUSSENAC-GILLES N., SILVIN C., BENJAMINS R., SOUBIE J.L., BREUKER J., *Évolution et maintenance dans les bases de connaissances dans le cadre d'une coopération cognitive entre système humain et artificiel*. Rapport de fin de contrat région Midi-Pyrénées 930-0244 . Décembre 1994.

[DEA-SILVIN, 94] SILVIN C., Étude préalable à la conception d'un outil interactif d'aide à la maintenance de systèmes à base de connaissances. *Mémoire de DEA Interaction Homme-Système Multi-modale*, Université Toulouse III, 1994

[Thèse-MATTA, 95] MATTA N. , Méthodes de résolution de problèmes : leur explication et leur représentation dans MACAO-II. *Thèse de l'université Paul Sabatier Toulouse 3*, Oct. 1995.

[IRIT-96-23-R] BEAUBEAU D., AUSSENAC-GILLES N., TCHOUNIKINE P.. *Mona au pays des rôles : opérationalisation de modèles conceptuels MONA en ZOLA*. Rapport Interne 96-23-R, IRIT, Juillet 1996.

[Livre-AC, 96a] AUSSENAC-GILLES N., MATTA N.. Expliciter une méthode de résolution de problèmes avec MACAO : problèmes méthodologiques. in *L'acquisition des connaissances : tendances actuelles* Toulouse : Eds. N. Aussenac-Gilles, P. Laublet, C. Reynaud. Cépaduès-Editions. mai 1996. pp 29-48.

[Livre-AC, 96b] LEPINE P., AUSSENAC-GILLES N.. Modélisation de la résolution de problèmes : comparaison expérimentale de KADS et MACAO. in *L'acquisition des connaissances : tendances actuelles* Eds. N. Aussenac-Gilles, P. Laublet, C. Reynaud. Toulouse : Cépaduès-Editions. mai 1996. pp 131-150.

[Livre-AC, 96c] N. AUSSENAC-GILLES, P. LAUBLET, C. REYNAUD. L'ingénierie des connaissances, composante à part entière de l'informatique du futur. In *L'acquisition des connaissances : tendances actuelles* Eds. N. Aussenac-Gilles, P. Laublet, C. Reynaud. Toulouse : Cépaduès-Editions. mai 1996. 3-25 . 1996.

[JICAA, 97] AUSSENAC-GILLES N., REYNAUD C., TCHOUNIKINE P., AND TRICHET F.. Associer un type de connaissances à un domaine : une question de rôle ? In *Journées Ingénierie des Connaissances et Apprentissage Automatique 1977 (JICAA'97*), mai 1997. IRISA - INRIA, Rennes. 345-361. 1997

[EKAW, 97] REYNAUD C., AUSSENAC-GILLES N., TCHOUNIKINE P., AND TRICHET F.. The notion of role in conceptual modelling. *In Proceedings of EKAW97 - European Knowledge Acquisition Workshop* - R. Benjamins & E. Plaza Eds. Lecture Notes in Artificial Intelligence. Springer Verlag, Heidelberg, 221-236. 1997.

[IMKB, 98] C. REYNAUD, N. AUSSENAC-GILLES, F. TORT. A support to domain knowledge modelling: a case study. In H. Kangassalo and P.J Charrel, editors*, Information Modelling and Knowledge Bases IX*, vol. 45 of Frontiers in AI and Applications, IOS Press, Amsterdam, 35-50. 1998.

CHAPITRE 5 - MODELES POUR LES RESSOURCES TERMINOLOGIQUES ET ONTOLOGIQUES

Dans ce chapitre, je situe tout d'abord le contexte pluridisciplinaire de mes recherches sur la construction de ressources terminologiques et ontologiques à partir de textes, ainsi que les choix retenus (5.1). Je présente ensuite les modèles de données proposés pour les bases de connaissances terminologiques (BCT) et pour les ontologies (5.2). Enfin, je décrirai les plates-formes que j'ai mises au point ou contribué à définir (5.3), l'une pour la construction de BCT et l'autre pour la modélisation d'ontologies à partir de textes.

À partir de l'expérience prometteuse d'une approche terminologique d'analyse de textes pour constituer le modèle du domaine, je me suis orientée vers l'étude des textes comme traces de connaissances. Cette source de connaissances, très différente de l'expertise humaine, suppose de définir et utiliser d'autres techniques d'analyse. Dans la continuité de l'expérience réalisée avec un extracteur de termes, j'ai choisi d'utiliser des logiciels de traitement automatique du langage et les techniques linguistiques. Pour rendre compte du résultat de ces analyses, j'ai ciblé des modèles particuliers, reflétant les connaissances d'un domaine, tout d'abord les bases de connaissances terminologiques puis les ontologies.

L'évolution de mes travaux a devancé un mouvement général du domaine vers une plus grande prise en compte des textes et des documents comme sources de connaissances. Elle est contemporaine à l'émergence des ontologies comme type de modèle conceptuel privilégié pour la gestion des connaissances. De nouvelles problématiques ont redéfini l'orientation de notre travail : (i) nature des modèles conceptuels permettant de rendre compte de connaissances et du vocabulaire utilisé pour les exprimer (ii) méthode d'acquisition de connaissances à partir de textes pour construire des terminologies et des ontologies ; (iii) utilisation et définition de logiciels basés sur des principes linguistiques pour le repérage de connaissances en corpus ; (iv) exploitation des terminologies et des ontologies, pour la gestion documentaire et la gestion des connaissances.

Parmi les questions découlant de cette orientation, j'aborde ici celles qui touchent à la nature de ces ressources, aux modèles de données (représentations) qui les caractérisent et aux environnements informatiques (appelées plates-formes de modélisation) permettant de représenter des connaissances selon ces modèles. Je traiterai des logiciels et techniques utilisés pour l'analyse de textes et des aspects méthodologiques dans le chapitre suivant.

5. 1 - Des experts aux textes, des modèles conceptuels aux ontologies

5.1.1 Nouvelles orientations thématiques

5.1.1.1 Motivations

Afin de diversifier les techniques entrant dans la construction d'un modèle conceptuel, j'ai mesuré l'intérêt d'exploiter les textes associés à un domaine de connaissances et d'utiliser pour cela des logiciels d'analyse du langage favorisant une étude terminologique. Au cours du projet SADE, l'extracteur de termes LEXTER a été utilisé pour étudier la terminologie du domaine. Cette expérience a montré l'atout de ce type de logiciel de traitement automatique des langues (TAL) pour identifier et représenter plus rapidement les concepts du modèle du domaine à partir de textes disponibles dans le domaine. Elle a également souligné la nécessité d'enrichir les capacités de représentation des connaissances pour permettre de gérer les termes (le vocabulaire) comme des entités à part entière. Il était tentant de poursuivre plus avant une recherche dans cette direction. Les deux raisons principales identifiées alors étaient de réduire le coût de la construction des modèles en sollicitant moins les experts et d'aborder, au-delà de la terminologie, la dimension linguistique des connaissances, alors que jusqu'ici, j'en ai étudié essentiellement les aspects cognitifs et ergonomiques. Plus précisément, en matière de modélisation conceptuelle, les avantages attendus étaient les suivants :

- diversifier les sources de connaissances, utiliser de manière complémentaire les documents existants, les échanges écrits entre acteurs du domaine et les traces d'entretiens ;
- exploiter des ressources terminologiques existantes, et proposer un modèle de connaissances cohérent avec les normes, terminologies et descriptions d'objets métiers existants ;
- s'appuyer sur des traces de connaissances plus consensuelles et relativement fiables, diffusées et mises sous forme écrite, de nature autre que des savoir-faire individuels ;
- accélérer le processus de modélisation en automatisant une partie du recueil de connaissances à l'aide d'outil d'analyse de textes ;
- disposer de modèles plus riches, plus faciles à interpréter et à maintenir, en associant aux modèles des textes et fragments de textes qui les documentent ou qui en justifient le contenu.

Les motivations relatives à l'intégration d'une dimension linguistique renvoyaient à des enjeux théoriques et pratiques, toujours d'actualité :

- concernant le processus de modélisation (qui correspond à la construction d'une représentation formelle), il s'agit de bénéficier des recherches sur la représentation formelle du langage naturel, sur le traitement automatique du langage naturel et en linguistique ;
- concernant les primitives de représentation des connaissances, il s'agit d'ajouter aux structures habituelles (concepts ou classes, relations ou rôles, etc.) des structures pour représenter les termes associés, des extraits de textes (corpus), des éléments de caractérisation syntaxique de la présence de ces connaissances dans les textes (patrons).

La problématique abordée est celle du repérage de connaissances à partir de formulations en langage naturel en vue de les utiliser pour construire des modèles conceptuels. Il s'agit non pas d'automatiser la production de représentations (conceptuelles ou formelles) mais bien de fournir un guide pour faciliter l'exploration d'un ensemble de textes afin d'en extraire des représentations conceptuelles. Les modèles conceptuels ciblés, qui étaient des modèles du domaine au début de ce travail, sont devenus des bases de connaissances terminologiques puis des ontologies (au sens donné en ingénierie des connaissances).

Ainsi, ma thématique de recherche demeure la construction de modèles conceptuels. Mais je l'ai traitée par la suite en étudiant en priorité les textes comme traces de connaissances, en définissant d'autres techniques d'analyse, basées sur le traitement automatique du langage et les techniques linguistiques, pour cibler les modèles particuliers que sont les ontologies et les ressources terminologiques. Ces glissements résultent à la fois de l'évolution de la formulation des problématiques dans le domaine, de nouvelles demandes en matière d'applications et en particulier de recherche d'information, et de convergences avec des questions d'autres disciplines. En effet, cette nouvelle orientation de mes travaux est aussi le fruit de collaborations avec des linguistes (en particulier Anne Condamines et Josette Rebeyrolle du laboratoire ERSS[9]) et des spécialistes du traitement automatique des langues (comme Didier Bourigault alors à la DER[10] d'EDF).

5.1.1.2 Collaborations et démarche interdisciplinaires

Avec le développement de la linguistique de corpus, les linguistes se sont tournés vers l'informatique non seulement pour définir ensemble de nouveaux logiciels d'analyse ou d'exploration des textes, mais aussi pour étudier les contributions possibles de la linguistique de corpus à la mise en forme des résultats de ces analyses. Un des premiers supports intéressants pour les linguistes, les terminologues et les lexicographes sont les bases de connaissances terminologiques (BCT). Différentes facettes de la linguistique de corpus peuvent contribuer à la construction de BCT : outiller la mise en œuvre de techniques d'exploration de corpus, définir un modèle de données pour les BCT, mettre en forme une démarche systématique pour leur mise au point ou encore un logiciel permettant de les construire. Plus fondamentalement, la linguistique rejoint ici les préoccupations de la terminologie et s'interroge sur une manière de gérer des termes en tenant compte du sens qu'ils recouvrent en corpus. Or l'ingénierie des connaissances (IC) se pose la question symétrique avec les ontologies : comment gérer le lexique qui désigne les connaissances dans l'ontologie ? L'articulation entre mots, termes, notions et concepts, ou encore le lien entre langage et connaissances, sont au cœur de ces questions proches. Ils justifient une étude tenant compte du regard de chacune de ces disciplines.

C'est donc dans l'esprit de faire avancer des problématiques disciplinaires que j'ai choisi, avec plusieurs chercheurs du laboratoire de linguistique ERSS de Toulouse, un premier objet de recherche commun, les BCT, traité des deux points de vue de l'ingénierie des connaissances et de la linguistique. Cet objet a donné naissance, entre 1996 et 1999, à des études communes plus méthodologiques, la linguistique s'interrogeant sur l'intérêt de son approche de l'analyse de corpus comme moyen de construire des BCT, et l'ingénierie des connaissances sur la prise en compte de ces nouveaux modèles conceptuels. Les outils de construction de BCT ainsi définis répondent aux besoins des linguistes comme à ceux des ingénieurs cogniticiens. Ces travaux ont également porté sur l'intérêt des BCT : si elles représentent un reflet « neutre » des textes, peuvent-elles être adaptées pour construire des ontologies ? à quel coût ?

Par ailleurs, des chercheurs en traitement automatique des langues (TAL) ont mis au point des logiciels, comme l'extracteur de termes LEXTER, pouvant faciliter la modélisation de connaissances à partir de textes. L'apport de LEXTER à la gestion du vocabulaire d'un domaine et à la modélisation de concepts a été évalué dans le projet SAMIE. Avec l'intégration de D. Bourigault au sein de l'ERSS, la collaboration avec les linguistes a été enrichie, devenant « tripartite ». Elle s'est poursuivie pour mieux définir la mise au point et les modes d'utilisation d'outils de TAL en vue d'extraire des connaissances dans un processus de modélisation. Les expérimentations ont porté sur SYNTEX, logiciel d'extraction de termes et d'analyse distributionnelle, et sur un outil d'aide à l'extraction de relations sémantiques, CAMELEON. Plus fondamentalement, l'application de ces logiciels à l'analyse de textes spécialisés, couvrant des domaines précis, pose la question des

[9] Equipe de Recherche en Syntaxe et Sémantique, UMR 5610 du CNRS et université Toulouse 2.
[10] Direction des Études et Recherches

ressources complémentaires utiles à ces logiciels. Doivent-ils s'appuyer sur des connaissances sur le comportement de la langue générale ou seulement sur des connaissances apprises en corpus ? Pour les logiciels utilisés ou définis, le choix retenu a été de s'adapter le plus possible aux corpus et d'utiliser le moins possible de ressources externes.

Enfin, un nouvel élargissement des collaborations a été possible grâce à ma participation au groupe de travail TIA (présenté à la fin de ce chapitre) à partir de 1998. La présence de terminologues dans le groupe m'a permis d'affiner avec eux les aspects méthodologiques sur les BCT, alors que la présence des chercheurs du LIPN[11] travaillant sur la construction d'ontologies à partir de textes a donné un support matériel à mes propositions méthodologiques. Ainsi, à partir des premières versions de la méthode et du logiciel TERMINAE définis par B. Biébow et S. Szulman, notre collaboration a débouché sur une nouvelle version de TERMINAE intégrant des éléments terminologiques, et mettant en œuvre les propositions méthodologiques élaborées pour les BCT.

5.1.1.3 Questions abordées

De nouvelles problématiques communes se sont alors dégagées : (i) acquisition de connaissances à partir de textes pour construire des terminologies et des ontologies ; (ii) utilisation et définitions de logiciels basés sur des principes linguistiques pour le repérage de connaissances en corpus ; (iii) exploitation des terminologies pour la gestion documentaire. Parmi les questions abordées relativement à la construction des ressources terminologiques et ontologiques, j'aborde ici celles qui touchent à la nature de ces ressources, aux modèles de données qui les caractérisent et aux environnements informatiques (*plates-formes de modélisation* dans la suite) permettant de construire des ressources selon ces modèles. Je traiterai des logiciels et techniques utilisés pour l'analyse de textes et des aspects méthodologiques dans le chapitre suivant.

J'expose tout d'abord le contexte interdisciplinaire de ces recherches et les choix que j'ai retenus (§ 5.1). Je présente ensuite les modèles de données proposés pour les bases de connaissances terminologiques (BCT) et pour les ontologies (§ 5.2). Enfin, je décris les plates-formes que j'ai mises au point ou contribué à définir (§ 5.3), l'une pour la construction de BCT et l'autre pour la modélisation d'ontologies à partir de textes.

5.1.2 Textes, ontologies et bases terminologiques

5.1.2.1 Construction d'ontologies à partir de textes

Finalement, la proposition de construire des modèles conceptuels, puis des ontologies, à partir de l'analyse de textes, a vu le jour très progressivement en ingénierie des connaissances, au fur et à mesure que les différentes disciplines concernées ont échangé leurs résultats et regroupé leurs propositions. Ce thème de recherche, tout à fait original en 1995, n'a donc pas été formulé d'emblée ainsi. On peut identifier deux courants de convergences, qui renvoient à deux points de vue différents sur les ontologies :

- pour certains, il s'agissait de constituer, à partir de textes, des modèles de connaissances d'un domaine particulier, à structurer selon des principes ontologiques (B. Bachimont les définira plus tard comme des ontologies régionales). La problématique était plutôt formulée comme celle de la gestion des connaissances extraites de textes spécialisés. Les premiers résultats en la matière sont établis à partir d'approches terminologiques (Meyer et Skuce, 1992) (Enguehard et Pantera, 1995) (Assadi, 1998), de recherche d'information (Martin, 1995) ou de modélisation de connaissances (Biébow et Szulman, 1997). Ces bases sont le fondement du groupe TIA.

[11] Laboratoire d'Informatique de Paris Nord, UMR 7030 du CNRS et Université Paris 13.

- Pour d'autres auteurs, il s'agissait d'une approche plus constructiviste visant à contribuer à l'Ontologie, dans la lignée des travaux sur les concepts universels qui peuvent correspondre aux classes de haut niveau de réseaux sémantiques (Sowa, 1991) ou sur la sémantique formelle. Le problème traité était une analyse du langage (non limité aux textes) pour la définition d'ontologies. L'objectif est alors de retrouver, par une étude de la langue basée sur l'introspection, des notions fondamentales pour définir des concepts de base de l'ontologie ou encore des indices pour situer des concepts plus précis par rapport à ces concepts fondamentaux.

Ce n'est que plus tard qu'une troisième perspective s'est dégagée : définir des modèles de connaissances consensuels et réutilisables adaptés aux documents et aux applications du web, des « modèles universels pour le web ». Cette question diffère des précédentes par son ambition, qui change la nature des problèmes : traiter la langue générale dans sa diversité, telle qu'elle est présente dans les documents du web, viser une grande réutilisabilité des modèles à construire, s'appuyer sur des processus fortement automatisés, … La pression actuelle pour parvenir au web sémantique dynamise ce domaine de recherche. De plus, les perspectives positives offertes par les convergences thématiques entre plusieurs disciplines ainsi que la maturité de leurs résultats conduisent à un essor significatif de ces travaux (Charlet *et al.,* 2005). Aujourd'hui, l'apprentissage automatique associé au TAL y tient une place prédominante.

5.1.2.2 Nouvelles problématiques en terminologie

En terminologie, le terrain théorique a longtemps été occupé par la *Théorie Générale de la Terminologie*. Cette théorie, fondée par Eugène Wüster à la fin des années trente, est née dans le courant positiviste de l'entre-deux guerres et dans la mouvance du Cercle de Vienne. Elle défend une vision unificatrice de la connaissance : le monde de la connaissance est découpé en domaines stables, dont chacun est équivalent à un réseau fixe de concepts, les termes étant les représentants linguistiques de ces concepts (Bachimont, 2004). Au cours des années 80, un rapprochement entre la terminologie et l'informatique s'est opéré avec le développement de la microinformatique. On s'est intéressé à la conception de bases de données terminologiques susceptibles d'aider les traducteurs professionnels dans les tâches de gestion et d'exploitation de lexiques multilingues. Les réflexions ont porté essentiellement sur le format de la fiche terminologique : à l'aide de quels champs décrire un terme dans une base de données qui sera utilisée par un traducteur humain ?

Depuis la fin des années 90, la terminologie classique voit les bases théoriques de sa doctrine ainsi que ses rapports avec l'informatique ébranlés par le renouvellement de la pratique terminologique que suscite le développement des nouvelles applications de la terminologie (Bourigault *et al.,* 1999). Ce renouvellement théorique et méthodologique en terminologie s'est produit de façon concomitante et parallèle aux évolutions de l'intelligence artificielle et de l'ingénierie des connaissances vers d'autres manières d'utiliser les représentations logiques et de définir des systèmes à base de connaissances.

5.1.2.3 Le groupe TIA

La fondation du groupe « Terminologie et Intelligence Artificielle »[12] (TIA) en 1993 concrétise la convergence des recherches françaises dans différences disciplines autour du thème de la terminologie et des modèles de connaissances en IA. Ce groupe de travail rassemble une vingtaine de chercheurs : des informaticiens de l'IA et du TAL, des linguistes terminologues et, depuis courant 2004, des spécialistes en sciences de l'information autour de l'intégration des démarches terminologiques, linguistiques, documentaires et d'ingénierie des connaissances pour

[12] groupe de travail associé à l'AFIA et au GDR-I3, fondé par A. Condamines et D. Bourigault, animé depuis 2002 par A. Condamines et moi-même.

exploiter les connaissances contenues dans des textes. La notion de bases de connaissances terminologiques est un des supports utilisé dans le groupe pour l'étude des problèmes liés à la représentation et au traitement des connaissances identifiées dans des textes. La mise au point et l'expérimentation de techniques linguistiques et de systèmes de traitement automatique des langues pour localiser des connaissances en corpus sont un autre objet d'approfondissements. Enfin, le groupe a cherché à inventorier et formaliser des principes méthodologiques pour mener l'analyse de textes à l'aide de ces logiciels et la structuration de connaissances dans des ontologies ou des terminologies. Les textes fondateurs sont présents sur le site du groupe[13].

Ma participation au groupe « Terminologie et Intelligence Artificielle » à partir de 1998 a fortement influencé mon travail. J'en assure la co-animation avec A. Condamines depuis décembre 2002. Le partage des travaux des différents participants a donné lieu à des collaborations et à des publications : projet Th(IC)² (IC,03), réflexions sur les méthodes de construction d'ontologies (méthodes associées à DOE et TERMINAE [EKAW, 00] [TIA, 01] [TIA, 03], sur l'adaptation des modèles aux usages [GDR-I3, 02] et [RIA, 04] ou, plus récemment, sur les modalités de validations des expériences, des logiciels et des méthodes d'analyse de textes pour la construction de modèles. Le groupe a également eu la volonté de diffuser ses résultats au niveau européen et de les confronter aux courants anglo-saxons considérant les ontologies avant tout comme des constructions formelles, réutilisables et souvent universelles. Ainsi, deux workshops ont été organisés par des membres de TIA « Ontologies and texts » à EKAW2000 et « Natural Language Processing and Machine Learning » à ECAI 2002.

5.1.2.4 L'action spécifique ASSTICCOT

Entre janvier 2002 et mai 2003, l'action spécifique « Constitution de Terminologies à partir de corpus » [14], rattachée au RTP « Documents : création, indexation, navigation » animé par J.M. Salaün (ENSSIB, Lyon), a regroupé environ 30 chercheurs de plusieurs disciplines : recherche d'information et sciences de l'information, linguistique de corpus et terminologie, traitement automatique du langage naturel, ingénierie des connaissances et apprentissage automatique pour la fouille de textes. J'en ai assuré la co-animation avec Anne Condamines.

Le contenu de cette action spécifique découle de plusieurs constats : (i) l'existence de projets collaboratifs autour de la constitution de terminologies à partir de corpus faisant intervenir le plus souvent deux approches ; (ii) le dynamisme de groupes de recherche fonctionnant sur des thématiques proches (comme A3CTE[15] ou TIA) ; (iii) la nécessité d'approfondir les nouvelles convergences possibles entre les sciences de l'information, la terminologie et l'informatique étant donné leurs positions récentes sur la modélisation à partir de textes et pour la gestion documentaire ; (iv) une demande sociétale très importante en matière de ressources terminologiques et des réponses qui se font souvent au coup par coup, en fonction des opportunités mais sans que les compétences propres aux disciplines soient clairement établies.

Il a semblé urgent de dresser un état des lieux des compétences des disciplines concernées et de leurs complémentarités afin de donner une assise aux recherches et de définir les lignes forces de ce qui pourrait constituer la recherche sur ce thème dans les prochaines années. Afin de baliser cette réflexion, quatre thèmes ont été débattus : besoins ciblés, place et nature des corpus, types des ressources utilisées ou produites, méthodes et outils. Le débat au sein du groupe a été diffusé au cours de deux ateliers associés aux conférences CFD[16] et Plate-forme AFIA[17]. Quatre axes de prospectives ont été dégagés à partir des questionnements soulevés [ASSTICCOT, 04] :

[13] http://tia.loria.fr
[14] http://www.irit.fr/ASSTICCOT/
[15] http://www-lipn.univ-paris13.fr/groupes-de-travail/A3CTE/index.html
[16] Conférence Fédérative sur le Document, Hammamet (Tunisie), Octobre 2002.
[17] Plate-forme de conférences de l'Association Française d'Intelligence Artificielle, Laval (F), juillet 2003

1 - Développer et approfondir la notion de « genre textuel » pour rendre compte non seulement des dépendances entre régularités linguistiques et situations de production de langage, mais aussi entre régularités des écrits et situations d'analyse et donc d'interprétation de corpus.

2 - *Prendre en compte les applications* pour comprendre la variabilité des méthodes, des outils ou techniques d'exploration et des ressources terminologiques. À terme, il s'agit aussi de prendre en compte cette variation pour adapter ces méthodes et outils à des objectifs spécifiques.

3 - *Définir des méthodes pour assurer la maintenance des ressources terminologiques*. Il s'agit de répondre ou anticiper le paradoxe dont les ressources terminologiques sont l'objet : elles doivent à la fois « normaliser » des connaissances, c'est-à-dire les figer à un moment donné, et être utilisées pour accéder à des connaissances qui évoluent dans des contextes dynamiques.

4 - *Étudier les problèmes d'évaluation et de validation des résultats*, non seulement par rapport aux contenus des corpus et aux avis d'experts, mais surtout dans l'ensemble d'un projet, voire dans l'ensemble des besoins d'une organisation. Un compromis est à trouver entre d'un côté spécificité, adaptation à un besoin, faible réutilisation et, d'un autre côté généricité, faible adaptation à un besoin particulier, réutilisabilité.

5.1.2.5 Une gamme de ressources terminologiques et ontologiques

La gamme des produits à base terminologique nécessaires pour répondre aux besoins de la gestion documentaire s'élargit considérablement (Bourigault & Jacquemin, 2000). À côté des bases de données terminologiques multilingues classiques pour l'aide à la traduction, on voit apparaître de nouveaux types de ressource terminologique ou ontologique (RTO) adaptés aux nouvelles applications de la terminologie en entreprise : glossaires et liste de termes pour les outils de communication interne et externe, thesaurus pour les systèmes d'indexation automatiques ou assistés, index hypertextuels pour les documentations techniques, terminologies de référence pour les systèmes d'aide à la rédaction, ontologies pour les mémoires d'entreprise, etc. (Fig. 5.1.2).

Une autre tradition, celle des sciences de l'information et de la recherche documentaire, utilise et produit des ressources analogues pour organiser des collections et y retrouver des documents. Le processus central est celui de l'*indexation*, destiné à représenter par les éléments d'un langage documentaire ou naturel des données résultant de l'analyse du contenu d'un document ou d'une question. On désigne également ainsi le résultat de cette opération. Un *index* est une table alphabétique des mots, des termes correspondant aux sujets traités, des noms cités dans un livre. Dans le domaine technique, l'index d'un document est aussi son analyse sommaire présentée sous forme de mots-clés, rubriques, etc. Un *langage documentaire* est un ensemble organisé de termes normalisés, utilisé pour représenter ou indexer le contenu des documents à des fins de mémorisation pour une recherche ultérieure. On distingue essentiellement, dans les langages documentaires, les classifications et les thesaurus. Une *classification* est la répartition systématique en classes, d'êtres, de choses ou de notions ayant des caractères communs notamment afin d'en faciliter l'étude ; c'est aussi le résultat de cette opération. Un *thesaurus* est un langage documentaire fondé sur une structuration hiérarchisée. Organisé alphabétiquement au niveau le plus haut, il répertorie des termes normalisés, reliés à des termes plus précis par des relations hiérarchiques.

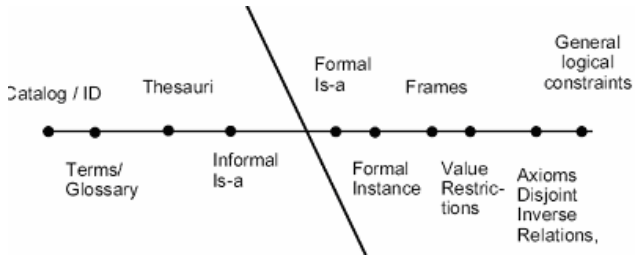

Figure 5.1.2. : Les différentes ressources terminologiques et ontologiques situées sur un axe allant des moins formelles (à gauche) aux plus formelles (à droite) (Lassila & McGuiness, 2001) dans (Gómez-Pérez *et al.*, 2004).

Ces types de structure se différencient par leur rôle plus ou moins normatif, leur conformité à un standard ou une norme existante, la nature des éléments présents (termes et leur description linguistique plus ou moins complète, concepts et leur définition, relations entre concepts, textes, …). Leur degré de formalisation est très variable et impose des contraintes plus ou moins fortes sur l'organisation des termes et des connaissances au sein de ces structures. Il permet aussi de raisonner ou non sur les données de la structure. Parmi les ontologies, on distingue les ontologies « light-weight » qui ne comportent qu'une hiérarchie de concepts et des relations entre concepts, des ontologies « heavy-weight » qui contiennent en plus des axiomes ou des contraintes sur les relations (Gómez-Pérez *et al.*, 2004).

Les critères de structuration à la base de l'organisation des termes ou des concepts conduisent à des formes spécifiques d'organisation, qui vont de la simple liste alphabétique à des réseaux formalisés, en passant par des arborescences. Enfin, les principes et méthodes appliqués pour décider des éléments à retenir dans la structure (introspection, étude de documents, …) ont un rôle fondamental pour en qualifier le contenu. Ainsi, une liste hiérarchisée de termes et de variantes, ou même une terminologie, n'est pas une ontologie : la première ne contient pas de concepts ; dans la deuxième, leur organisation ne respecte pas de principes ontologiques (au sens de Guarino ou de Bachimont par exemple). WordNet, base de données lexicales, est souvent considérée comme une ontologie alors que les « synsets » rendent compte de classes d'équivalences de mots et non vraiment de définitions de concepts. De plus, l'observation fine de Wordnet montre une certaine variabilité dans le sens donné aux relations.

Or la multiplication des types de ressource terminologique met à mal le principe théorique de l'unicité et de la fixité d'une terminologie pour un domaine donné, ainsi que celui de la base de données terminologique comme seul type de ressources informatiques pour la terminologie. Le constat de la variabilité des terminologies s'impose : étant donné un domaine d'activité, il n'y a pas *une* terminologie, qui représenterait le savoir sur le domaine, mais autant de ressources terminologiques ou ontologiques que d'applications dans lesquelles ces ressources sont utilisées. Selon l'application, ces ressources peuvent différer sensiblement par les unités retenues et leur description. L'ensemble de ces constats empiriques entraîne des changements en profondeur de la pratique terminologique, et appelle du même coup à un renouvellement théorique de la terminologie. Ce renouvellement est sollicité aussi sur des bases philosophiques et épistémologiques (Slodzian, 2000).

5.1.2.6 Les bases de connaissances terminologiques

Parmi ces types de ressource, les terminologues et informaticiens de l'IA ont proposé la notion de base de connaissances terminologiques (BCT) (Skuce et Meyer, 1991). Les BCT constituent un enrichissement significatif des terminologies traditionnelles sur papier car elles comportent une trace des informations conceptuelles relevées par le terminologue au moment de l'identification des termes (Meyer *et al.*, 1992). Leur originalité est avant tout leur modèle de structuration des connaissances terminologiques. Ce modèle différencie un niveau linguistique d'un niveau conceptuel : on accède ainsi par les termes du domaine à une modélisation conceptuelle qui donne sens à ces termes. Cette composante s'apparente à la partie descriptive des bases de connaissances, ou encore plus étroitement aux réseaux sémantiques en intelligence artificielle.

L'étude et la définition des BCT sont un travail fondamentalement interdisciplinaire, fruit de résultats en linguistique, en informatique particulièrement en intelligence artificielle. En concrétisant la notion de « modèle du contenu d'un texte » que produirait une analyse linguistique, les BCT sont le lieu d'apports mutuels de ces disciplines. Définir leur modèle de données suppose une analyse conjointe du statut des termes et des représentations conceptuelles. Étudier leur construction et leur utilisation permet de dissocier les problèmes qui relèvent de la linguistique de

ceux qui découlent des contraintes de définition conceptuelle et formelle. Un premier objectif est alors d'améliorer les outils et méthodes à chaque étape, de rendre systématiques les analyses linguistiques ou de mettre en évidence l'impact des contraintes applicatives sur l'organisation du modèle conceptuel. D'autres visées sont d'étudier finement les passages entre ces différentes représentations, du texte à la logique, avec prise en compte progressive de l'application ciblée.

La gestion des BCT requiert des environnements spécifiques, qui soient adaptés aux tâches du terminologue : ils facilitent la définition de fiches terminologiques et surtout l'organisation de concepts au sein d'une hiérarchie ou d'un réseau plus complexe, comportant d'autres relations conceptuelles. Pour un panorama de systèmes de référence en 1995, voir l'état de l'art établi dans [TIA, 97]. Parmi ces environnements, le système précurseur, CODE, proposait de faciliter la gestion des connaissances au sein d'un réseau hiérarchique de frames. CODE permet d'associer des données linguistiques à un réseau conceptuel et de représenter ainsi la terminologie d'un domaine (Skuce, 1991). CODE a été mis au point dans le cadre du projet COGNITERM (Meyer, 1992) afin de gérer des connaissances terminologiques bilingues dans le domaine des technologies de stockage optique. Très vite, ce projet a débouché sur une réflexion méthodologique sur l'utilisation de CODE par les terminologues, ainsi que sur la définition de logiciels pour accélérer l'identification de termes et de concepts. Un outil d'aide à l'extraction de terminologie à partir de textes, TEXTANALYSER, a été mis au point à cet effet (Kavanagh, 1996).

D'autres projets précurseurs sont le projet ILIAD (Toussaint *et al*., 1998) dans le domaine de la géologie et l'étude menée par des linguistes toulousains au sein du laboratoire Aramiihs (Condamines et Amsili, 1993). Ce deuxième projet visait la construction d'une terminologie spécialisée multilingue utile à l'aide à la rédaction. Il a été le lieu d'une confrontation originale entre des analyses linguistiques de corpus, un outillage informatique de type TAL et des objectifs terminologiques. Surtout, il a requis la définition d'un modèle de données (Condamines et Amsili, 1993) qui a été amélioré dans le cadre de notre collaboration avec A. Condamines.

Cette problématique concerne des recherches visant une structuration conceptuelle utile à la gestion documentaire, et contenant donc une composante terminologique comme les systèmes HYTROPES (Euzenat, 1996) et CGKAT (Martin, 1995). D'autres systèmes abordent des questions analogues dans le but de rendre compte de manière formelle ou ontologique de connaissances présentes dans des textes, par exemple les systèmes TERMINÆ[18] (Biébow & Szulman, 1997) ou encore DOCKMAN[19]. Enfin des logiciels de gestion terminologique comme SYSTEM QUIRK[20] (Ahmad, 1995), l'interface HTL de l'extracteur de termes LEXTER (Bourigault, 1996) fournissent des points de comparaison.

Dès le projet CODE, il ressort que la construction de BCT soulève trois types de recherche : la définition d'une représentation des connaissances et d'un modèle de données, des aspects méthodologiques pour la construction à partir de textes, et enfin une dimension ingénierie pour définir des plates-formes utilisables. Ma contribution en matière de BCT porte sur ces trois aspects. Je les présente dans les parties qui suivent, à travers les objectifs scientifiques et les méthodes retenus, et je termine en situant mon travail par rapport aux systèmes mentionnés plus haut.

5.1.3 Partis pris, choix retenus

Mon approche reste dans le cadre des choix énoncés au chapitre 3 : donner un poids fort aux usages ; rendre compte des particularités liées aux domaines dans les modèles ; construire des modèles répondant avant tout aux besoins des utilisateurs. De ce choix, découlent d'autres options

[18] http://www-lipn.univ-paris13.fr/~szulman/TERMINAE.html (valide au 16/12/2005)
[19] http://www.site.uottawa.ca/~jprpic/Dockman/dockmanindex.html (valide au 16/12/2005)
[20] http://www.mcs.surrey.ac.uk/Research/cs/AI/systemQ/ (valide au 16/12/2005)

sur la manière de choisir les corpus, d'en interpréter les contenus, de mener les étapes de la modélisation ou encore sur les techniques à utiliser.

Les textes analysés sont rassemblés en corpus : ils sont choisis en fonction de leur contenu et des objectifs de l'analyse. Il ne s'agit pas de traiter des textes tout-venant dont on ne connaîtrait pas les caractéristiques. La manière de constituer un corpus s'est précisée avec les expérimentations et à travers des échanges avec les linguistes.

Les modèles construits couvrent des domaines précis, délimités et non des connaissances générales. Les concepts rendent compte du sens des termes tel qu'il se définit à partir de l'analyse et l'interprétation de leurs usages accessibles dans le corpus d'une part, et des connaissances que l'on veut présenter aux utilisateurs d'autre part. Les définitions obtenues ne seront ni universelles, ni normalisatrices mais locales, spécifiques et normées en fonction d'un point de vue. Ce point de vue est maîtrisé par la personne qui dépouille les analyses de corpus et se porte garante des besoins auxquels doit répondre la ressource à construire.

Les types de modèle ciblés sont **variés**, plus ou moins formels, par exemple des thésaurus ou des index structurés, des terminologies techniques ou encore des ontologies « régionales ». J'ai peu étudié les problèmes liés à la formalisation et l'opérationnalisation.

Les logiciels de TAL utilisés **produisent des traitements de « haut niveau** », au sens où ils extraient, isolent des éléments de texte qui, a priori, peuvent faire sens. Bien qu'il soit souhaitable de réutiliser des logiciels existants, il était nécessaire de collaborer avec des spécialistes du TAL pour définir ensemble les spécifications de ces logiciels.

Les logiciels de TAL sont considérés comme des **aides à l'interprétation** humaine au cours du dépouillement des textes. Ils ne se substituent pas à l'analyste, comme ce serait le cas en automatisant « l'extraction de connaissances ». En effet, les outils de traitement automatique des langues s'appuient sur la forme (des mots, des structures syntaxiques des phrases ou syntagmes, de leurs combinaisons en corpus), sur des critères numériques et sur les distributions pour faire des propositions. Pour le moment, seul un humain peut leur donner du sens (passer au niveau sémantique) et surtout décider d'une construction, d'une représentation, qui va (ou non) intégrer ce sens. De plus, les outils de TAL utilisés se situent dans une lignée sémantique qui accorde une place importante à l'interprétation humaine. La pertinence de ce choix s'est confirmée au cours des expériences et a trouvé un fondement théorique dans les échanges menés au sein de TIA.

J'ai retenu des **approches linguistiques** plus que statistiques. L'analyse de corpus par l'informatique est avant tout distributionnelle : il s'agit de repérer des régularités de forme, des cooccurrences ou des parentés d'usage de certains mots pour suggérer des termes, des classes de mots ou des phénomènes autres soumis à l'interprétation de l'analyste. Or ces distributions peuvent s'appuyer sur des décomptes plus élaborés que celui de chaînes de caractères, comme des résultats ou des approches établies par la linguistique, ou encore sur des analyses préliminaires (grammaticales, morphologiques, lexicales, etc.). La linguistique propose de nouveaux phénomènes à observer et étudier.

5. 2 - Modèles de données pour les ressources terminologiques et ontologiques

Je présente d'abord dans cette partie un modèle de données pour les bases de connaissances terminologiques (5.2.2) défini en collaboration avec A. Condamines et à partir de ses travaux. Ce modèle a été repris ensuite pour enrichir la représentation des connaissances de la plate-forme TERMINAE (5.2.3) dans le cadre d'une collaboration avec le LIPN. Ces modèles articulent des structures qui rendent compte de termes du domaine ainsi que de connaissances associées, sous forme d'un réseau conceptuel. Leur mise au point a soulevé un travail de clarification sur ce que j'entendais par terme, concept et notion, travail rendu nécessaire par le contexte pluridisciplinaire de mes recherches. J'en rapporte quelques aspects au 5.2.1.

5.2.1 Termes, notions et concepts

Clarifier ce que couvrent les mots *termes*, *notions* et *concepts* est une des composantes de notre recherche (Biébow, 2004). Si le fait que plusieurs disciplines s'appuient sur un vocabulaire « commun » favorise les rapprochements dans un premier temps, très vite, les divergences conceptuelles, liées aux objectifs et à l'histoire de chacune, rendent complexes les échanges et multiplient les malentendus. Les analyses disciplinaires présentées ci-dessous mettent l'accent sur les mutations récentes sur la définition des concepts et des termes. Ces mutations reflètent une convergence de vue et sont, en particulier, le fruit des réflexions collectives de TIA.

5.2.1.1 Concepts et réseaux sémantiques

En intelligence artificielle, seule est utilisée la notion de concept, objet formel relevant de la représentation des connaissances. Savoir ce que représente un concept a été l'objet de débats récurrents parmi les chercheurs travaillant sur les réseaux sémantiques et les formalismes associés (Sowa, 1991). Cette question revient régulièrement parce qu'elle est rapidement évacuée par une vision informatique qui confond la définition de ce que l'on veut représenter avec la solution technique retenue. Par exemple, le concept est souvent assimilé au prédicat de la logique du premier ordre. Aussi, plutôt que de définir ce qu'est un concept, il est plus simple de préciser comment ils sont représentés dans différents formalismes (Gómez-Pérez *et al.,* 2004). Les concepts correspondent à des classes d'objets abstraits ou concrets, que l'on cherche à définir en intension, par des caractéristiques qui sont le plus souvent leurs relations avec d'autres classes, et rarement en extension, par la liste de leurs réalisations ou instances. Elles se définissent donc par leur place au sein d'une hiérarchie et par les relations formelles posées entre elles. Certains formalismes comme les logiques de description différencient des concepts primitifs (donnés par des conditions nécessaires) de concepts définis (par des conditions nécessaires et suffisantes).

Les recherches sur les ontologies issues d'une tradition plus philosophique s'inquiètent plus du statut du concept indépendamment de sa représentation, et cherchent à aider l'analyste à déterminer ce que sont les concepts du domaine et comment ils sont reliés entre eux. Ainsi les travaux de N. Guarino (méthode OntoClean) fournissent des propriétés que doivent vérifier les concepts et les relations. Ces propriétés sont le reflet d'une vision des concepts comme renvoyant à des classes universelles d'objets partageant au moins une propriété caractéristique et ayant une unité, ces classes étant disjointes entre elles et stables au cours du temps. Les travaux de B. Bachimont mettent plus l'accent sur le fait que les concepts sont des constructions permettant de rendre compte d'une réalité, d'un sens. En référence aux travaux de F. Rastier, il les définit comme des *signifiés normés*, qui organisent les objets en s'appuyant sur leurs parentés et leurs différences, selon un point de vue fixé par l'analyste.

5.2.1.2 Concepts et ontologies

Il est utile de revenir sur l'analyse de B. Bachimont concernant les différentes approches du concept dans plusieurs courants de la tradition philosophique, à chacune correspondant une manière de les identifier à partir du langage ou d'autres appréhensions du monde (Bachimont, 2004). Revue, cette tradition débouche sur trois facettes des concepts au sein des ontologies :

- **Concept comme essence** : selon cet angle d'analyse, le concept d'un objet est le noyau des propriétés nécessaires vérifiées par un objet, indépendamment des variations qu'il peut subir dans les différents contextes où il se trouve ; il est connu par abstraction ; on considère alors le concept comme une signification normée, qui s'inscrit dans un système et dont la compréhension correspond à sa reformulation à travers d'autres significations conceptuelles ; on rejoint là la terminologie traditionnelle.
- **Concept comme construction synthétique** : le concept d'un objet est vu comme une règle qui permet de rassembler toutes les expériences, les réalisations et d'en tirer un invariant, une

méthode pour passer de la diversité des expériences pour définir, construire une synthèse, une invariance ; selon cette vue, la relation du concept aux autres concepts est moins fondamentale que celle du concept aux objets qui lui correspondent ;

- **Concept comme performatif** : le concept peut être aussi considéré comme une opération, il permet des actions. Alors énoncer un concept, c'est faire quelque chose (Austin, 1991). En informatique, le concept formalisé est interprété par la machine pour produire un comportement effectif. Il est prescription pour le système formel, dont il permet d'exécuter des actions ;

Bachimont considère que ces trois facettes sont autant de niveaux dans la manière de comprendre un concept. Leur articulation est indispensable. Il oppose deux courants actuels dans la conceptualisation : une tradition *nominaliste*, pour laquelle l'ontologie regroupe des essences dont la signification sera explicitée de manière logico-formelle (Smith, 1998) (Guarino, 1995) ; une tradition *essentialiste*, selon laquelle les essences sont étudiées comme des unités linguistiques et conceptuelles, sans préjuger de leur portée ontologique réelle (Bachimont, 2004). Son point de vue, développé dans la méthode Archonte et qui a fortement influencé le groupe TIA, se situe dans une approche à la fois nominaliste et conceptuelle : à partir des manières de parler, on déduit des manières de penser pour aborder des manières d'être.

5.2.1.3 Concepts et notions

En terminologie, la *notion* est l'équivalent du concept (Condamines, 2003) (Biébow, 2004). Selon la terminologie traditionnelle, la notion préexiste aux termes, c'est-à-dire aux unités linguistiques qui la désignent. Les notions d'un domaine scientifique ou technique s'organisent en un ensemble structuré par des relations non linguistiques, le système notionnel, qui peut être « purifié », normé pour définir une norme. La langue serait alors le point d'entrée vers les connaissances. Cette école fait l'hypothèse de la monosémie des termes dans chaque domaine de spécialité. Le but de la terminologie est alors de rassembler, selon une démarche dite onomasiologique, les termes correspondant à une notion donnée par sa définition (normée). Le lien terme-concept est alors un lien univoque de référence.

Les pratiques terminologiques autant que l'évolution de la théorie terminologique remettent en question cette vision. La démarche terminologique partant de textes se déroule dans le sens inverse, selon une démarche sémasiologique : partant de manifestations linguistiques, on identifie et définit des concepts. À ce moment-là, le sens est *construit* à partir de l'interprétation des occurrences du mot en contexte. Puis il est normé (ce qui revient à établir une convention, une norme) et normalisé (restreint pour être adapté à l'application qui va l'utiliser). Rastier parle de *sens normé* et non de *signifié normé* pour bien souligner le caractère contextuel de cette construction, liée aux occurrences observées en corpus, alors que le signifié aurait un statut indépendant du corpus (Biébow, 2004).

Le statut du concept du point de vue de la linguistique est plus complexe : il prend tantôt la place du signifié, tantôt celle d'entités générales, universelles, préexistant aux termes (Condamines, 2004). Le concept permet entre autres de rendre compte de phénomènes comme la polysémie. A. Condamines considère qu'il n'y a concept que si la normaison (définition de la norme) peut s'appuyer sur des régularités d'usage des signes linguistiques et si l'on dispose de critères pour conférer à ces signes un statut particulier. Dans le cadre d'analyse de corpus spécialisés, les concepts ne sont pas préexistants mais construits, et les manifestations linguistiques à partir desquelles ils sont établis ont en commun des caractéristiques liées à la situation d'énonciation. D'une certaine façon, *créer des concepts est un acte de définition, de maîtrise du sens*. Dans ce même esprit, les manifestations linguistiques observées ne prennent le statut de *termes* qu'après avoir effectué cette définition.

5.2.2 Un modèle pour les bases de connaissances terminologiques

Une BCT est avant tout un inventaire des termes d'un domaine enrichi d'informations conceptuelles. Les concepts, organisés en réseau conceptuel, permettent de donner un sens à ces termes, c'est-à-dire de définir les notions qu'ils désignent et de justifier leur place dans la terminologie. Pratiquement, les données d'une BCT sont donc organisées en deux parties : le réseau conceptuel, formé des concepts et relations conceptuelles, et les données linguistiques, les fiches terminologiques, ces dernières étant reliées au corpus d'où elles sont extraites. Le modèle de BCT proposé a été mis au point en collaboration avec A Condamines et J. Rebeyrolle de l'ERSS, dans le cadre du DEA de P. Séguéla [DEA-SEGUELA, 96] et s'inspire du modèle proposé dans (Condamines et Amsili, 1993).

5.2.2.1 Présentation

Le modèle de BCT défini comporte trois types de structure : `termes`, `concepts` et `textes`. La figure 5.2.2.1 illustre l'organisation de ces structures. Dans cet exemple, pris dans le domaine spatial (Condamines, 1993), le terme RELAIS possède deux interprétations. Dans la langue de spécialité des experts en météorologie (point de vue *météorologie*), il désigne le `concept` étiqueté par SATELLITE et il est décrit comme « engin placé sur une orbite autour de la terre ». Le terme RELAIS est aussi utilisé dans la communauté des télécommunications pour désigner le `concept` étiqueté SATELLITE GEOSTATIONNAIRE. D'ailleurs, selon ce point de vue, les termes RELAIS et SATELLITE DE COMMUNICATION sont synonymes.

Ce modèle est décrit en détail dans [TIA, 97]. `Termes` et `concepts` sont deux structures distinctes afin de dissocier la manifestation linguistique de la notion qu'elle dénomme. On n'associe au `terme` aucune information conceptuelle : sa signification découle de ses concepts associés. Son interprétation dépend de la sémantique des structures représentant les concepts et en particulier des relations conceptuelles. `Termes` et `concepts` doivent donc être définis conjointement.

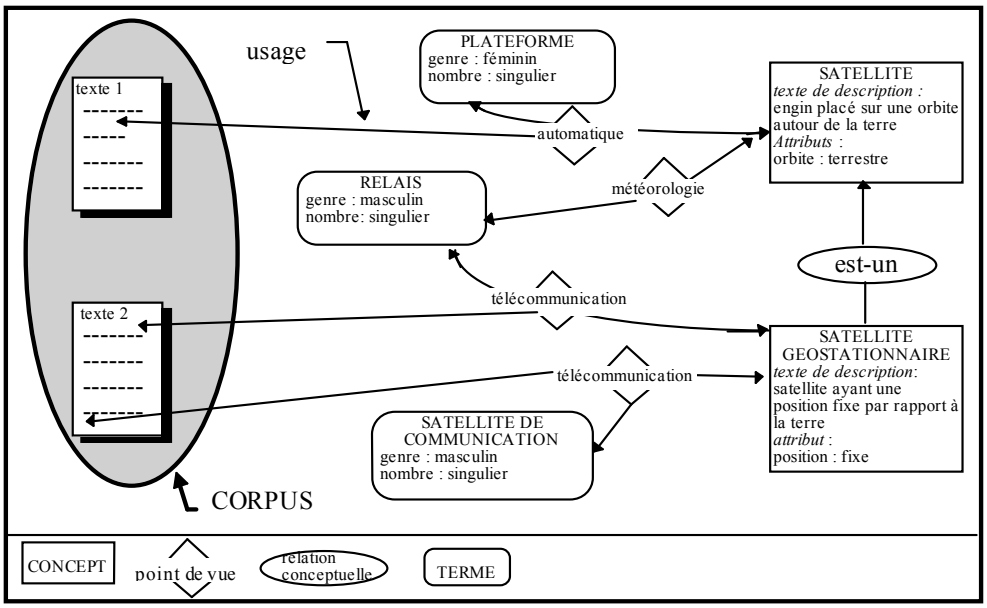

Figure 5.2.2.1 : Organisation des données dans la BCT. Le réseau conceptuel, à droite, est relié aux données terminologiques, à gauche, par des liens qui renvoient à des parties du texte.

D'ailleurs, ils sont reliés par des liens illustrant l'usage du terme lorsqu'il désigne ce concept. Les liens entre `termes`, `concepts` et `textes` véhiculent donc des informations :

- les liens d'usage précisent chacune des occurrences d'un terme d'une langue de spécialité en reliant le lien (`terme`, `concept`) à des textes ;

- les points de vue spécifient la validité d'utilisation de la désignation conceptuelle du terme ; il est une des caractéristiques du lien `terme-concept` ;

- des `relations` sémantiques associent les `concepts` entre eux.

Les `textes` sont stockés dans la BCT sous forme `d'unités textuelles` de manière à faciliter leur gestion. Ce découpage est transparent pour les utilisateurs.

La structure de `terme`, proche de celle présentée dans (Condamines et Amsili, 1993), rassemble, en plus du syntagme qui le désigne, uniquement des informations linguistiques : langue, variantes de forme, décomposition grammaticale, genre, nombre, ... Les relations entre termes, comme la synonymie, la polysémie ou l'anaphore, sont implicites et calculables à partir des liens `terme-concept`.

Enfin, un `concept` correspond à une description structurée et normalisée au sein d'un frame. Les connaissances de différenciation sont exprimées à l'aide de commentaires mais aussi d'attributs et de relations vers d'autres `concepts` : relations assertionnelles (relations étiquetées, sans sémantique formelle) et relations structurelles (« est-un », relation hiérarchique assurant l'héritage des relations assertionnelles). Un `concept` est défini par sa place dans la hiérarchie des concepts et par l'interprétation des commentaires, de ses attributs, de leurs valeurs et de ses relations. La modélisation conceptuelle de la BCT vise la caractérisation des notions selon un point de vue particulier et non leur définition exhaustive. Si besoin, il faut définir autant de concepts que de descriptions de celui-ci.

Ce modèle de données a été implémenté dans deux applications, GEDITERM, logiciel de construction de BCT [Mémoire-FOURNIER, 98] (partie 5.3.1), et CONSULTERM, logiciel de consultation de BCT [Mémoire-LECORGNE, 98] (partie 5.3.2).

5.2.2.2 Intérêt d'un modèle non formel pour les BCT

Pour la représentation des connaissances, j'ai préféré une représentation non formelle, qui rende compte du réseau conceptuel sans permettre de raisonner sur ces connaissances. Ce choix répond à la pratique des linguistes et des bilans faits dans l'état de l'art. Les rapports de S. Simon [Mémoire-SIMON, 98] et P. Séguéla [DEA-SEGUELA, 96] rendent compte de l'évaluation de différentes représentations formelles des connaissances. D'après leurs conclusions et surtout les témoignages avec des linguistes, utiliser un langage formel est trop contraignant pour le linguiste qui construit une BCT. Il l'oblige à s'éloigner de la forme des connaissances telle que la langue permet d'y accéder.

Tout d'abord, la formalisation suppose au préalable une normalisation des définitions (au sens défini dans la partie 5.5) qui contraint le linguiste à répondre à des questions dont la réponse n'est pas forcément dans le texte ou est ambiguë. En effet, la description formelle des connaissances oblige à rendre compte formellement des différences entre concepts à l'aide d'attributs ou de relations (rôle). 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 présentes 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, elle est prévue dans l'environnement d'exploitation de BCT CONSULTERM ainsi que dans l'environnement de modélisation d'ontologies TERMINAE.

5.2.2.3 Comparaison à d'autres modèles de données des BCT

A l'inverse de ce choix, la plupart des formalismes utilisés pour représenter des BCT sont inspirés des réseaux sémantiques, des logiques de descriptions (dans CODE et TERMINAE) ou des graphes conceptuels (CGKAT). HYTROPES utilise des frames (objets du langage TROPES) pour rendre compte de points de vue sur les objets. Le fait d'utiliser une représentation formelle des connaissances terminologiques est perçu comme un avantage, un moyen de réduire les ambiguïtés en obligeant à formuler explicitement des critères de définition et de différenciation, de classer au fur et à mesure les concepts définis, de vérifier leur cohérence, etc.

Malgré cette différence de formalisation, il est intéressant de comparer la richesse du modèle des données des BCT, et en particulier la représentation des relations. Dans mon modèle de BCT, la seule relation formalisée est EST-UN. La signification des autres relations est donnée par l'interprétation de leur nom ainsi que par le type des concepts qu'elle peut associer. Ce même choix est retenu dans HYTROPES ou CODE, où les autres relations sémantiques sont traduites par les attributs des concepts. Un travail plus poussé a été mené dans CGKAT pour proposer un ensemble de relations formelles organisées en une hiérarchie.

Enfin, le corpus est présent ou non dans le modèle des données. Assez caricaturalement, la plupart des systèmes qui visent une formalisation rapide et privilégient les concepts aux termes (CGKAT ou HYTROPES) n'intègrent pas le corpus. Toutefois, TERMINAE et Dockman, parce qu'ils accordent un poids important à l'analyse linguistique et à la justification de la modélisation par les textes, assurent le lien entre termes, concepts et textes. De même, les systèmes centrés sur l'analyse linguistique comme le nôtre privilégient les termes aux concepts et permettent de revenir facilement aux occurrences en corpus. Plus encore, System Quirk et HTL ne se focalisent que sur les termes et leurs occurrences, sans gérer clairement le niveau conceptuel.

5.2.3 Un modèle pour les ontologies régionales à composante terminologique

TERMINAE désigne un logiciel et une méthode d'élaboration de bases de connaissances et d'ontologies à partir de textes, développés par B. Biébow et S. Szulman au LIPN. Ma collaboration avec le LIPN au sujet de TERMINAE s'est concrétisée grâce à ma participation au groupe TIA. Le logiciel TERMINAE présentait alors de nombreuses parentés avec GEDITERM, et s'avérait tout à fait complémentaire puisqu'il prenait mieux en charge la phase de formalisation, débouchant sur de

véritables ontologies en logique de description. Cette collaboration a été motivée par la volonté de mettre à disposition au sein d'une même plate-forme des outils d'exploration de textes et des outils de modélisation de manière à assurer un continuum entre les textes, la composante terminologique et le modèle formel. Cet environnement devait évoluer pour garantir une traçabilité du processus de modélisation et permettre de revenir du modèle aux sources de connaissances qui en justifient en partie le contenu.

J'ai contribué à l'évolution de TERMINAE en tirant profit de l'expérience acquise sur les BCT avec la construction de GEDITERM. Une première étape a été de permettre de représenter des terminologies avec TERMINAE, grâce à la gestion de fiches terminologiques liées aux concepts (TIA, 2001) et (TAL, 2002). Une autre évolution a été de prévoir l'ajout de fonctionnalités pour une meilleure exploitation du contenu des textes à l'aide de logiciels de TAL. Je reviendrai sur ce dernier aspect dans le chapitre suivant.

5.2.3.1 Représentation des connaissances dans TERMINAE

Termes, concepts et rôles

Le modèle de données dans TERMINAE tel qu'il fût défini en 1999 était proche de celui d'une base de connaissances terminologiques, le réseau conceptuel correspondant à une ontologie décrite en une sorte de logique de description (Biébow et Szulman, 1999). Les textes sont conservés comme traces d'occurrence des termes. Plus simplement que dans une BCT, les termes sont des chaînes de caractères associées aux concepts du réseau conceptuel.

Au sein du réseau, les concepts sont organisés selon une hiérarchie de classes, les *concepts génériques* et d'instances*, les concepts individuels* (fig. 5.2.3.1). Les *concepts* sont définis par leur nom, leur place dans la hiérarchie et des *rôles* qui les relient aux autres concepts. Les rôles d'un concept précisent les conditions nécessaires et suffisantes qui caractérisent les éléments de cette classe. De plus, les concepts possèdent des caractéristiques enregistrées pour informer l'analyste : une caractéristique précise si le concept est associé ou non à des termes du corpus (il sera *terminologique, terminologique non attesté* ou *non terminologique*), une autre explique comment ce concept a été introduit (un concept est de *structuration ascendante* s'il est introduit pour structurer les concepts de plus bas niveaux, *descendante* s'il spécifie un concept de plus haut élevé).

Au moment de l'opérationnalisation, les concepts et leurs instances forment un ensemble structuré que TERMINAE gère comme une base de connaissances. Un classifieur permet de tester la validité de l'insertion de tout nouveau concept et informe l'utilisateur de la détection d'incohérences ou de redondances.

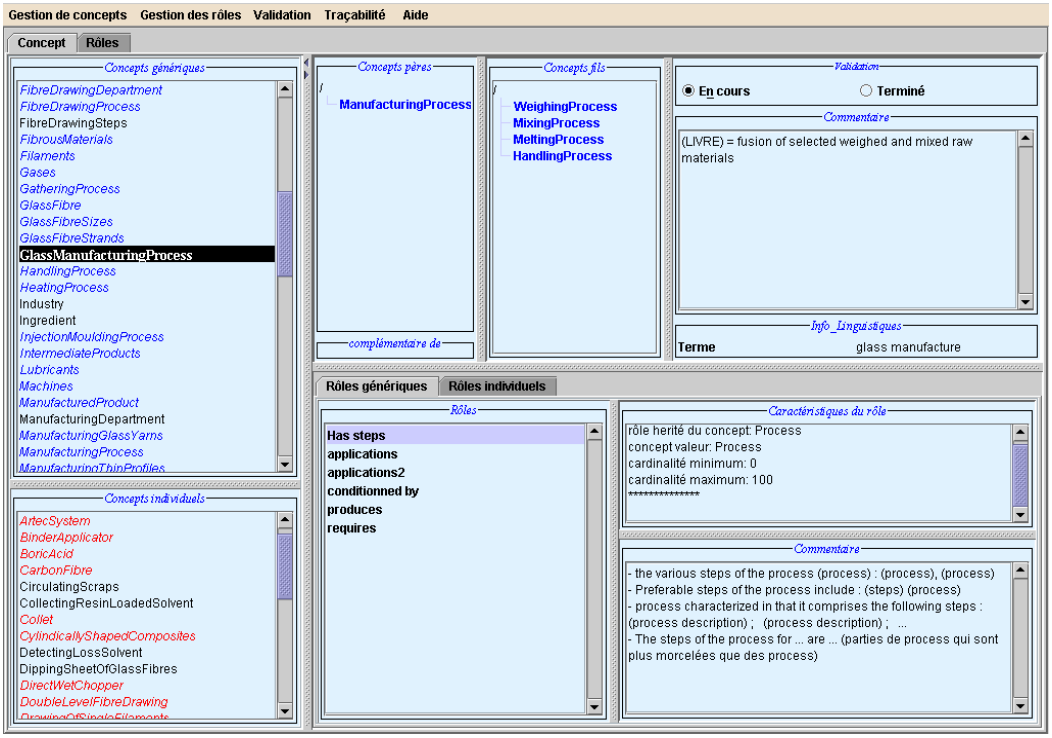

Figure 5.2.3.1 : Modélisation d'un concept dans TERMINAE : détails concernant le concept GlassFiberManufacturingProcess

À côté de la hiérarchie de concepts, une hiérarchie de *rôles* permet de définir les relations entre concepts (fig. 5.2.3.2). Chaque rôle possède un *domaine* et un co-domaine ou *valeur* qui précisent les concepts pouvant être reliés, la relation devant être interprétée comme partant du domaine pour atteindre un concept valeur. Un rôle possède des *cardinalités*, qui précisent le nombre maximum ou minimum de concepts pouvant être reliés à un concept donné par cette relation.

Les rôles sont systématiquement hérités. Ils peuvent être restreints en spécialisant leur valeur. Par exemple, si un `logiciel d'analyse de textes` PRODUIT DES `données lexicales`, si `extracteur de termes` EST-UN `logiciel d'analyse de textes`, `ensemble de termes` EST-UN `données lexicales`, alors on peut spécialiser le rôle PRODUIT : `extracteur de termes` PRODUIT DES `ensemble de termes`. L'organisation des rôles en une hiérarchie offre un double intérêt. Premièrement, l'utilisateur peut affiner une relation entre deux concepts en utilisant tout d'abord un rôle de haut niveau (i.e. « méronymie ») qu'il peut préciser ensuite (i.e. « ingrédience »). Deuxièmement, la place dans la hiérarchie définit une sorte de sémantique des rôles, tant pour l'analyste qui modélise des connaissances (par l'interprétation des étiquettes et de la place dans la hiérarchie) que pour le système (on peut traiter ou présenter à l'utilisateur de manière analogue tous les rôles d'un certain type).

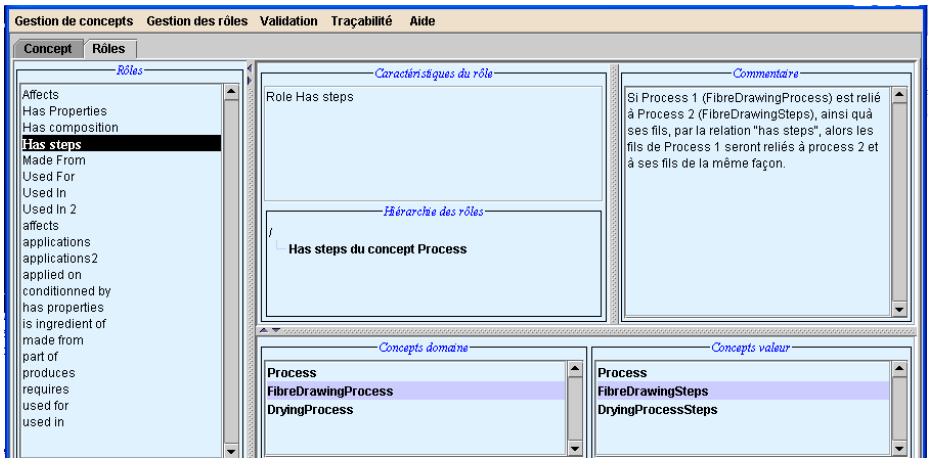

Figure 5.2.3.2 : Gestion des rôles dans TERMINAE : détails concernant le rôle « has-steps » reliant les concepts FiberDrawingProcess et FiberDrawing Steps

Les fiches de modélisation

Dans une première version de TERMINAE, une aide à la formalisation était proposée sous forme de *fiches de modélisation* (Biébow et Szulman, 1999). L'optique de la méthode associée était d'enrichir un noyau générique d'ontologie, présenté à l'utilisateur pour amorcer la modélisation. À ce niveau, l'utilisateur devait choisir parmi un ensemble de relations usuelles prédéfinies. À partir de la définition en langage naturel d'un concept terminologique, l'utilisateur décrivait ce concept dans un langage contraint en utilisant les objets existants dans l'ontologie et les relations prédéfinies. Cette description pouvait être qualifiée de semi-formelle, au sens où les concepts ne sont pas interprétables par le système informatique. Elle était ensuite traduite dans le langage de l'ontologie après des vérifications de cohérence sur les propriétés et une comparaison sous forme normale avec les concepts déjà présents.

5.2.3.2 Vers des ontologies à composante terminologiques

Les fiches terminologiques

Les *fiches terminologiques* sont donc l'autre facette de notre contribution à TERMINAE, conséquence de la volonté de faciliter la construction de terminologies mais aussi de mieux documenter les ontologies [TAL, 02]. Les *fiches terminologiques* permettent de consigner les résultats de l'étude linguistique des termes, ses différents sens (concepts reliés) et ses occurrences en corpus. Concrètement, cette fiche reprend la structure de terme de notre modèle de BCT. Elle comporte des informations lexicales sous forme de rubriques choisies par l'analyste (catégorie grammaticale, langue, normes terminologiques …), une définition en langage naturel, des synonymes et des termes proches (voir aussi). De plus, cette fiche s'articule avec les concepts (appelés « notions » sur la figure 5.2.3.3) qu'elle désigne de manière analogue au lien terme-concept proposé dans GEDITERM : ces liens sont documentés par les occurrences du terme (passages de texte). C'est à partir de l'introduction de cette fiche, la méthode et les fonctionnalités du logiciel TERMINAE ont été revus de manière à définir des terminologies.

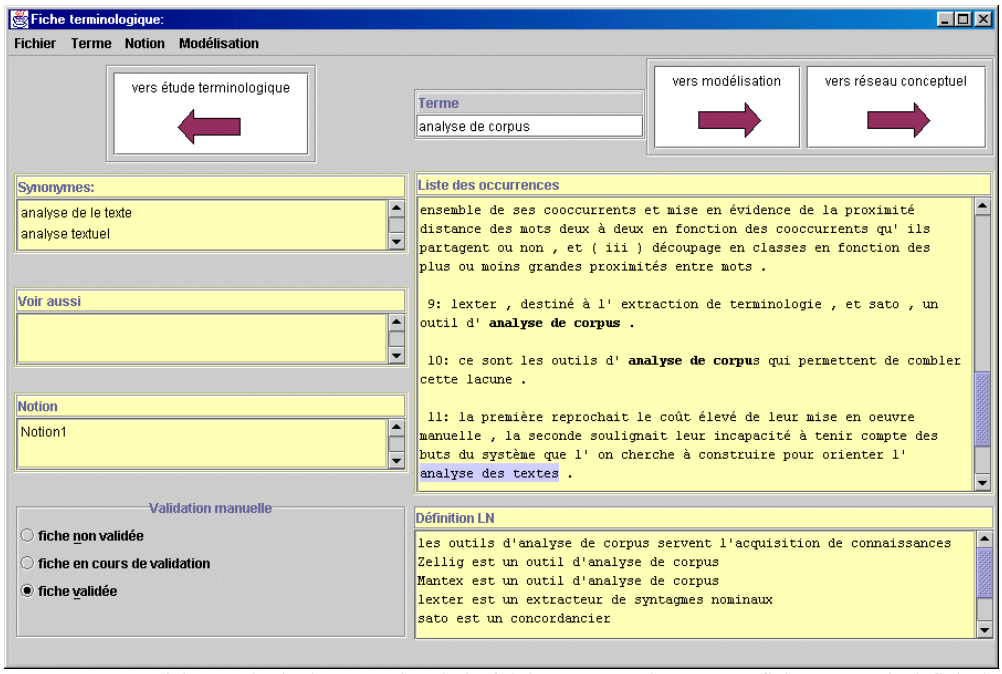

Figure 5.2.3.3 : Fiche terminologique, version du logiciel TERMINAE de 2000. La fiche permet de définir des concepts associés au terme étudié, de visualiser et regrouper les occurrences en fonction des différents sens possibles du terme dans l'application.

Plus que le support à la construction de terminologies, ces fiches ont montré leur intérêt à la fois pour conserver une trace du travail de modélisation à partir des textes et pour matérialiser les choix de modélisation effectués avant la normalisation.

Le réseau conceptuel

Notre collaboration avec B. Biébow et S. Szulman a eu comme premier impact l'introduction d'un niveau de représentation intermédiaire entre les termes bruts et les concepts logiques, un niveau semi-formel, appelé le *réseau conceptuel* [EKAW, 00]. Ce réseau permettait la déclaration de concepts et de relations sans exiger de validité syntaxique ou sémantique globale, ou sans nécessiter de les situer à leur place définitive dans la hiérarchie. La structuration des concepts consistait alors à identifier et poser des relations entre concepts. Le réseau conceptuel correspondait bien au réseau notionnel des terminologues, et se voulait un intermédiaire avant la formalisation de l'ontologie. Avec l'ajout de fiches terminologiques, TERMINAE permettait de définir des terminologies, constituées de fiches et d'un réseau.

La pratique a permis de constater que les deux objets, réseau conceptuel et ontologie, n'avaient pas lieu d'être distingués, le premier n'étant qu'un état non validé de l'autre. Les fenêtres présentées sur les figures 5.2.3.1 et 5.2.3.2 correspondent à ce niveau de modélisation. TERMINAE propose donc désormais des phases de validation et de formalisation progressive, sans donner pour autant un statut au réseau conceptuel. De plus, l'utilisateur peut considérer son modèle terminé avant qu'il ne soit complètement formalisé.

5.2.3.3 Vers des ontologies à composantes terminologiques

Une structure d'ontologie est un quintuplet $\mathcal{O} := \{\mathcal{C}, \mathcal{R}, \mathcal{H}^{\mathcal{C}}, rel, \mathcal{A}^{\mathcal{O}}\}$

- $\mathcal{C}$ et $\mathcal{R}$: ensembles disjoints des **concepts** et des **relations**
- $\mathcal{H}^{\mathcal{C}}$ **hiérarchie** (taxonomie) de concepts : $\mathcal{H}^{\mathcal{C}} \subseteq \mathcal{C} \times \mathcal{C}$, $\mathcal{H}^{\mathcal{C}}(C_1, C_2)$ signifie que C_1 est un sous-concept de C_2 (relation orientée)
- $\mathcal{R}el$: **relation** rel: $\mathcal{R} \rightarrow \mathcal{C} \times \mathcal{C}$ (relations sémantiques non taxonomiques) avec 2 fonctions associées
  - $dom : \mathcal{R} \rightarrow \mathcal{C}$ avec $dom(\mathcal{R}) := \prod 1(rel(\mathcal{R}))$
  - $range : \mathcal{R} \rightarrow \mathcal{C}$ avec $range(\mathcal{R}) := \prod 2(rel(\mathcal{R}))$ co-domaine
  - $rel(\mathcal{R}) = (C_1, C_2)$ s'écrit aussi $\mathcal{R}(C_1, C_2)$
- $\mathcal{A}^{\mathcal{O}}$: ensemble **d'axiomes**, exprimés dans un langage logique adapté (logique de description, logique du 1er ordre)

Table 5.2.3.3 : Définition d'une ontologie dans (Maedche, 2002)

En enrichissant le modèle de données classique à l'aide de fiches terminologiques, le modèle de TERMINAE se situe tout à fait dans l'esprit de la définition donnée par Maedche d'une ontologie (Maedche, 2002). Pour cet auteur, une ontologie comporte un lexique qui permet de gérer les termes désignant les concepts et leurs relations. Il définit une *ontologie à composante lexicale* comme un couple $(\mathcal{O}, \mathcal{L})$ où $\mathcal{O}$ est une ontologie et $\mathcal{L}$ un lexique tels qu'il est défini dans les tables 5.2.3.3 et 5.2.3.4.

Le lexique d'une structure d'ontologie $\mathcal{O} := \{\mathcal{C}, \mathcal{R}, \mathcal{H}^{\mathcal{C}}, rel, \mathcal{A}^{\mathcal{O}}\}$ est un quadruplet $\mathcal{L} := \{\mathcal{L}^{\mathcal{C}}, \mathcal{L}^{\mathcal{R}}, \mathcal{F}, \mathcal{G}\}$

- $\mathcal{L}^{\mathcal{C}}$ et $\mathcal{L}^{\mathcal{R}}$: ensembles disjoints des **entrées lexicales** des concepts et des relations
- $\mathcal{F}, \mathcal{G}$: deux relations appelées **références** $\mathcal{F} \subseteq \mathcal{L}^{\mathcal{C}}$ (pour les concepts), $\mathcal{G} \subseteq \mathcal{L}^{\mathcal{R}} \times \mathcal{R}$ (pour les relations), • Pour $L \in \mathcal{L}^{\mathcal{C}}$: $\mathcal{F}(L) = \{C \in \mathcal{C} / (L,C) \in \mathcal{F}\}$ • $\mathcal{F}^{-1}(L) = \{L \in \mathcal{L} / (L,C) \in \mathcal{F}\}$ • Idem pour $\mathcal{G}$ et $\mathcal{G}^{-1}$

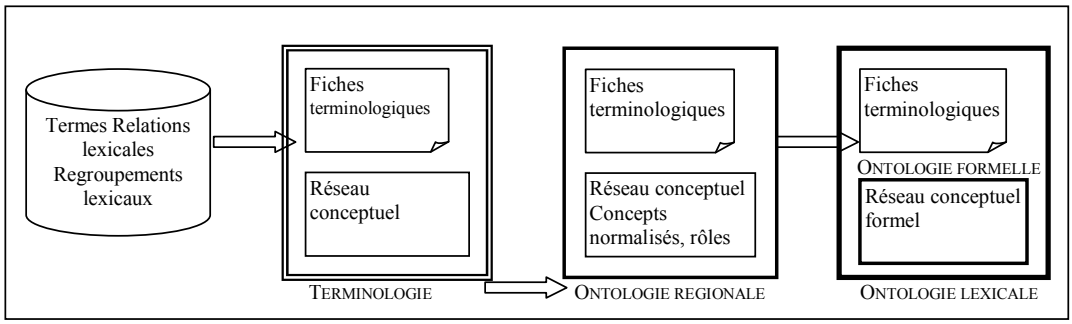

Table 5.2.3.4 : Définition du lexique d'une ontologie dans (Maedche, 2002)Afin de pouvoir exporter en OWL les modèles construits avec TERMINAE selon cette représentation, des aménagements de OWL ont été proposés (Szulman *et al.,* 2004). En OWL, les termes ne sont que des « label » associés au concept. Les aménagements proposés définissent des structures spécifiques pour représenter les termes et synonymes associés à un concept.

Figure 5.2.3.4 : Les différents modèles au cours du processus de modélisation avec TERMINAE, depuis les résultats des outils de TAL jusqu'à l'ontologie formelle et lexicale. Les processus permettant de passer d'une représentation à l'autre seront présentés au 5.5.

TERMINAE permet de produire des modèles plus ou moins formels pouvant aller d'une terminologie à une ontologie formelle, et de retenir ou non la composante terminologique (fig. 5.2.3.4). Le niveau « semi-formel » correspond à celui de l'ontologie régionale, alors que l'ontologie formelle complétée par les fiches terminologiques correspond à une ontologie lexicale au sens où je viens de la définir.

5. 3 - Logiciels pour la structuration et l'exploitation de terminologies

Le modèle de données de BCT s'avère pertinent pour dépasser les limites des terminologies classiques ou encore pour mieux poser la question de modèles conceptuels à partir de textes. Cependant, la quantité de données requises pour leur constitution puis leur exploitation nécessite un support informatique. Tout comme les résultats méthodologiques, les logiciels et leur utilisation dans le cadre de construction de BCT en vraie grandeur font partie de l'outillage nécessaire à l'évaluation expérimentale des BCT. Avec mes collègues linguistes de l'ERSS (A. Condamines et J. Rebeyrolle), j'ai donc mené un projet pluridisciplinaire d'envergure entre 1997 et 2000, basé sur le développement d'une plate-forme informatique et sur plusieurs projets d'expérimentation. Ce projet[21] a débouché sur un rapport très complet qui regroupe nos réflexions sur le modèle de BCT, la présentation des logiciels et leur utilisation dans le cadre de partenariats avec des entreprises [rapport-MOUGLIS, 98].

Deux logiciels complémentaires ont été définis. L'un, GEDITERM, permet à un terminologue de construire et structurer une BCT à partir d'analyses linguistiques de corpus. L'autre, CONSULTERM, permet d'utiliser une BCT, de la consulter et de l'adapter si besoin, dans un contexte applicatif donné. Ces logiciels sont présentés dans les deux parties qui suivent (5.3.1 et 5.3.2). GEDITERM constitue une contribution significative au domaine, dont l'originalité par rapport à d'autres logiciels de l'état de l'art est soulignée dans l'article [IC, 99]. La partie 5.3.3 rapporte l'utilisation de GEDITERM au sein de deux projets de construction de BCT qui ont servi de terrain d'expérimentation et de validation : SGDD et MOUGLIS [rapport-MOUGLIS, 98].

Ce découpage en deux logiciels reflète une vision très particulière du processus de modélisation à partir de textes. Cette vision, échafaudée avec A. Condamines, mise à l'épreuve puis remise en question, fait l'hypothèse que les BCT sont des représentations structurées du contenu des textes, dont elles permettent une lecture différente. Ce choix dissocie des étapes dans la modélisation, et repousse l'interprétation finalisée des données à une deuxième étape, celle de l'exploitation de la BCT :

- construire une BCT à partir d'un texte revient à rendre explicites et structurées des informations plus diffuses et informelles présentes dans les textes ;
- exploiter une BCT revient à en interpréter le contenu pour en extraire une partie, la réorganiser et la représenter en fonction d'une certaine finalité.

Or la construction de plusieurs BCT a souligné l'importance des processus d'abstraction et d'interprétation en jeu dans la construction de modèles à partir de textes. Pour mieux maîtriser ces processus, on peut distinguer des BCT-corpus, qui se voudraient un reflet « neutre » du contenu des textes, et des BCT-Applicatives, qui en seraient l'adaptation finalisée. De nouvelles réflexions, menées en particulier au sein du groupe TIA, ont fait abandonner l'illusion de la neutralité des analyses linguistiques et, avec elles, des BCT-corpus. De ce fait, la parenté des problématiques de

[21] Ce projet a été financé par le Gis Sciences de la Cognition et la région Midi-Pyrénées, ainsi que grâce à deux contrats avec des entreprises, l'un (DDE) avec la Direction Départementale de l'Équipement de la Haute-Garonne, l'autre (MOUGLIS) avec la Direction des Études et Recherches de la Haute-Garonne.

modélisation à partir de textes ressort fortement, quelle que soit la nature du modèle conceptuel visé. Cette réflexion et ses évolutions font l'objet de la partie suivante 5.4.

5.3.1 GEDITERM : gestion de bases de connaissances terminologiques

5.3.1.1 Caractéristiques des modèles construits

Développé par D. Fournier en étroite collaboration avec des linguistes, GEDITERM est un environnement de gestion de BCT [Mémoire-FOURNIER, 98]. Il intègre toutes les composantes de la BCT, y compris le corpus. Ainsi, à partir de l'étude de l'usage des mots en corpus, sont définis des termes, des concepts, des types de relation et des relations entre concepts, ainsi que des liens entre termes et des passages de textes. Chaque structure de donnée peut être documentée de commentaires pour consigner la justification des choix de modélisation et des occurrences.

Les modèles construits avec GEDITERM ne sont pas formels. Il s'agit de réseaux conceptuels formés de concepts et de types de relation dont la cohérence formelle n'est pas vérifiée. En effet, j'ai montré dans la partie précédente qu'une BCT doit s'appuyer sur une structure de données non formelle [Traité-IC, 01]. Seul le type de relation Est-Un est interprété par le logiciel : il assure l'héritage des autres types de relation. La sémantique de ces autres relations est explicitée à l'utilisateur sous forme de commentaires. Enfin, la visualisation graphique du réseau obtenu en facilite la lecture et la construction progressive. Ce réseau conceptuel peut constituer une étape vers une éventuelle formalisation.

5.3.1.2 Fonctionnalités

Les fonctionnalités du logiciel permettent de paramétrer la définition des termes, de stocker et enregistrer des données pour construire un modèle selon les structures de base du modèle (termes, concepts, relations et textes), de guider l'utilisateur pour structurer les connaissances, de les visualiser et enfin de les vérifier.

Aide à la définition de composants du modèle

Chaque type de donnée peut être consulté de manière globale sous forme de liste, ou individuellement sous forme de « cartes » ou fiches (fig. 5.3.1). La carte présente les informations spécifiques à cette donnée et ses composants reliés, dont la sélection ouvre une nouvelle carte. 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 de « vues » sur plusieurs composants ayant une parenté sémantique : triplet concept-relation-concept ou terme-lien-concept. L'accès au texte se fait alors depuis les liens ou les relations.

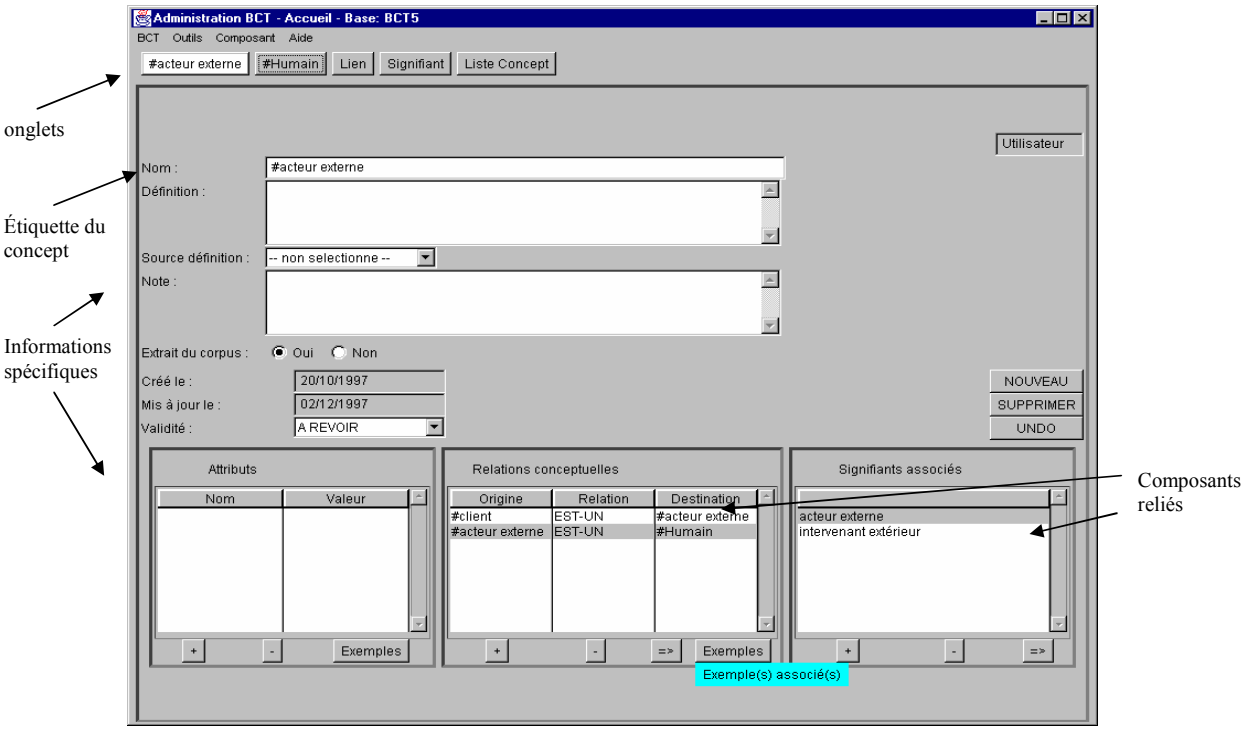

Figure 5.3.1.2 : Exemple de carte de concept active : le concept #acteur interne.

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. GEDITERM 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 (Bourigault, 1994). Cette fonction déclenche un transfert des données systématique puis une rapide validation interactive par le linguiste pour corriger leur organisation.

Aide à la structuration et à l'organisation des composants

Pour guider l'organisation des concepts, les relations conceptuelles sont typées, et l'ensemble des types de relation est répertorié, organisé en hiérarchie et accessible via des listes ou sous forme d'arbre. 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 avant d'être utilisé.

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 (Fig. 5.3.1.3), les critères de sélection sur une liste de termes peuvent être un syntagme nominal, une variante de forme (ellipses, abréviations et formes les plus utilisées d'un terme), un locuteur, un concept (les termes retenus auront au moins un lien avec ce concept) ou un degré de validité.

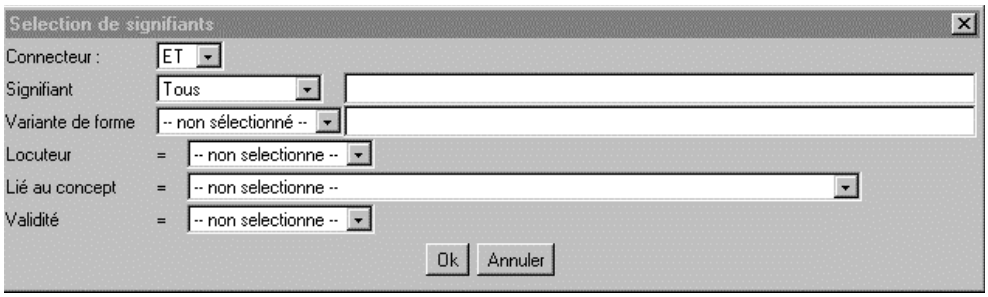

Figure 5.3.1.3. : Fenêtre de définition d'un filtre pour la liste des termes

De même, on peut filtrer des données afin 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 5.3.1.4 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.

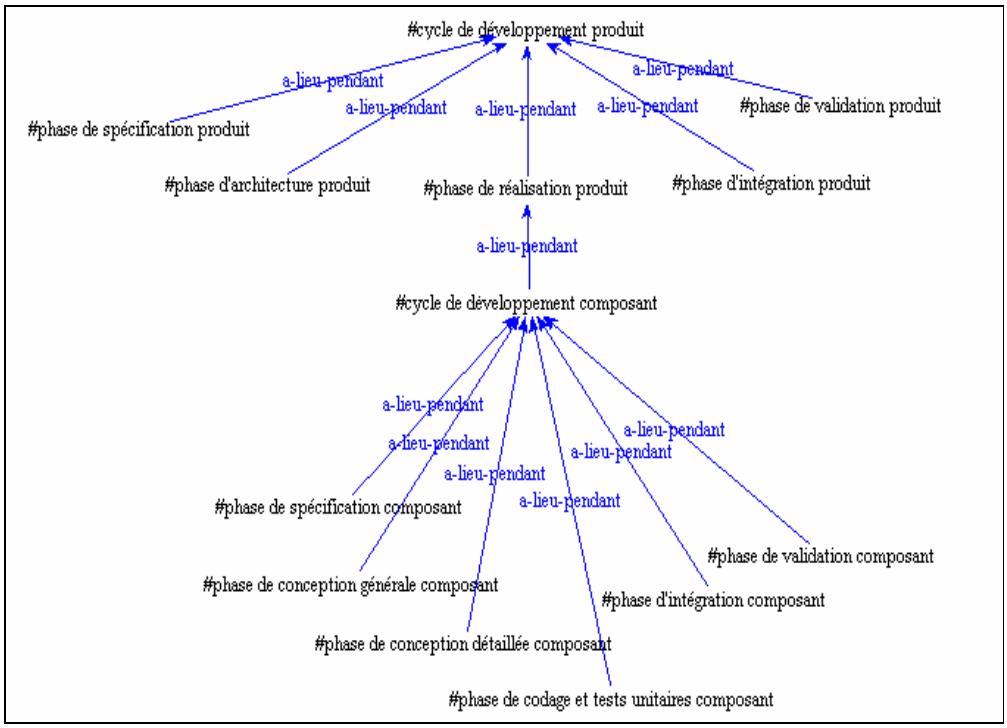

Figure 5.3.1.4 : Extraits du réseau conceptuel de la BCT Mouglis.

Aide à la vérification

Enfin, le logiciel permet de vérifier les données saisies selon des principes prédéfinis et à la demande de l'utilisateur. 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 :

- types des concept reliés par les relations conceptuelles : toute relation spécifique entre deux concepts doit relier des concepts fils des classes indiquées dans la définition de cette relation ;

- place des concepts dans la hiérarchie EST-UN : tout concept doit, en fin de construction de la BCT, être situé dans la hiérarchie EST-UN ;

- fournir des informations minimales pour définir des structures (termes ou concepts), comme son nom, mais aussi un terme relié pour un concept, sa langue pour un terme, etc. .

Une BCT n'est terminée que si ces vérifications ont été effectuées pour tous les termes et concepts, qui doivent avoir le statut de « valides ». Or ce mode de vérification a été remis en question par la pratique des linguistes, qui, à l'utilisation de GEDITERM, 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, l'application de ces règles directives requiert des connaissances pas toujours présentes dans le corpus.

5.3.1.3 Bilan

Le développement de GEDITERM représente un résultat pertinent à double titre, tant par la réflexion approfondie menée sur la représentation des connaissances et le modèle de données (présentés au 5.2) que par les aspects méthodologiques (qui seront abordés au 5.5). En tant que logiciel, il s'avère un des rares logiciels opérationnels de ce type, destiné à des linguistes et facilitant l'exploitation de résultats d'extracteurs de termes.

Une première version de GEDITERM a été présentée à la conférence TIA 99 [TIA, 99]. Son utilisation a été suivie par une ergonome qui a pu en faire une évaluation ergonomique [mémoire-SERRA, 97]. Elle a confirmé la pertinence de choix comme la visualisation graphique l'organisation de l'interface de saisie. Elle a également conduit à enrichir le modèle de données, qui s'est avéré 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 parties, on voudrait pouvoir préciser quels concepts sont complémentaires et forment ensemble l'objet entier. On peut aussi souhaiter rendre compte de schémas de type agent-verbe-objet-moyen. Dans cette perspective, dans l'esprit de DocKMan de D. Skuce (1998), la notion de relation a été enrichie de manière à pouvoir conserver, dans la partie terminologique, des schémas syntactico-sémantiques ou lexico-syntaxiques correspondant à leur expression dans les textes. Une autre perspective, non développée, serait de construire, dans le réseau conceptuel, l'équivalent des relations sous forme de frames rassemblant des sous-ensembles de réseaux.

Parmi les enseignements tirés de ces expériences, il ressort qu'un environnement comme GEDITERM doit intégrer ou être interfacé avec des logiciels automatisant les traitements sur corpus. Ces logiciels 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.

Ensuite, les rubriques des structures de données, en particulier celles qui décrivent les termes, doivent pouvoir être adaptées à chaque type d'application. Initialement, l'hypothèse que ces informations sont indépendantes de l'utilisation prévue des données de la BCT a été mise à mal. En effet, 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.

Enfin, la formalisation des données doit intervenir dans une phase finale, après leur structuration rigoureuse en fonction des besoins de l'application. Dans le modèle conceptuel, la trace des choix de modélisation doit être conservée sous plusieurs formes : à l'aide de commentaires dans chaque structure, grâce au lien vers des occurrences des termes dans les textes et enfin grâce au lien entre les structures elles-mêmes.

5.3.2 CONSULTERM : consultation de bases terminologiques

5.3.2.1 Pourquoi distinguer construction et exploitation de BCT

La BCT constitue un noyau de base de connaissances sur un domaine et se veut réutilisable pour divers types d'application sur ce domaine, comme construire un index ou formaliser des connaissances pour une ontologie. Toutefois, pour chaque application visée, il faut modifier les données de la BCT (restructurer le réseau conceptuel par exemple), ajouter des informations (par exemple ajouter des attributs aux concepts pour mieux les caractériser suivant le point de vue de l'application) ou en supprimer (par exemple si l'on restreint le point de vue). De plus, on peut avoir besoin de sélectionner des sous-ensembles de BCT, en fonction de critères sémantiques ou syntaxiques. Enfin, une application peut nécessiter de recourir à des concepts issus de plusieurs BCT, que l'on souhaite associer pour construire un nouveau modèle.

Le logiciel GEDITERM 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.

Pour répondre aux besoins d'adaptation d'une BCT à un objectif particulier, un prototype d'environnement d'exploitation et de consultation de BCT, CONSULTERM, a été développé. Ce logiciel est destiné à préparer des BCT-Applicatives à partir de BCT-Corpus. Ce prototype propose des fonctions d'aide à la sélection des données dans des BCT existantes, selon les informations linguistiques ou selon le réseau conceptuel, pour les modifier et les réorganiser, en fonction d'un besoin spécifique. Il permet ainsi de construire de nouvelles BCT (dites BCT-Applicatives) en fonction d'objectifs appliqués particuliers à partir de BCT(s) de « référence ». Pour cela, des critères de définition précis doivent être explicités puis appliqués. Ils sont utilisés par des mécanismes de filtrage et de complétion des BCT déjà définies. CONSULTERM produit des données d'un format facilement exportable vers des applications. Une maquette de cet environnement et les analyses associées à sa réalisation sont présentées dans le mémoire d'ingénieur CNAM d'E. Lecorgne [Mémoire-LECORGNE, 98].

5.3.2.2 Lien avec GEDITERM

La connexion entre environnement de gestion et environnement d'exploitation est minimale, mais tout a été fait de manière à autoriser une connexion beaucoup plus étroite. Ainsi, deux principes fondamentaux ont été respectés de manière à faciliter le passage d'un environnement à l'autre : (i) ils partagent le même modèle de données et dialoguent avec les mêmes bases pour sauvegarder les données ; (ii) ils ont été développés avec le même langage et peuvent être accessibles simultanément dans un environnement de travail.

5.3.2.3 Principes de base

CONSULTERM est vu comme un outil qui aide à consulter des bases stables établies par des linguistes (des BCT-Corpus) sans les modifier, pour construire des BCT adaptées à des applications spécifiques, donc des BCT-Applicatives. Plus largement, les fonctionnalités permettent d'utiliser des BCT-Applicatives stables pour construire de nouvelles BCT-Applicatives. On distingue donc plusieurs types de BCT, selon leur état de développement et leur lien avec une application :

- des BCT-Corpus versus des BCT- Applicatives ;

- des BCT stables (qui ne vont plus être modifiées par importation de données) versus des BCT en construction (sujettes à évolution) ;

- des BCT sources qui sont référencées pour développer une BCT de travail.

Ainsi, construire une BCT-Applicative consiste à définir une BCT de travail, à choisir la ou les BCT sources à partir desquelles on va la construire (ce sont une ou plusieurs BCT stables, qu'elles soient Corpus ou Applicatives). Tant que l'on modifie la BCT de travail, elle n'est pas stable. Lorsqu'on l'a terminée, on peut décider de la rendre stable.

5.3.2.4 Fonctionnalités proposées

CONSULTERM permet avant tout d'importer et d'exporter des données d'une BCT vers une autre. Les deux principales fonctionnalités sont la consultation de BCT-Stables afin d'en exporter des sous-ensembles, et la mise à jour de BCT-travail pour construire et enrichir une BCT propre à une application. Pour cela, le système s'appuie sur la notion de buffer, assure une présentation simplifiée du modèle de données auprès de l'utilisateur et autorise différents modes de consultation. Un buffer contient des sous-ensembles de données de la BCT, sélectionnés en fonction de critères sémantiques (concepts reliés par un type de relation par exemple) ou syntaxiques. Enfin, cet environnement permet de visualiser sous forme graphique simple des données ainsi sélectionnées.

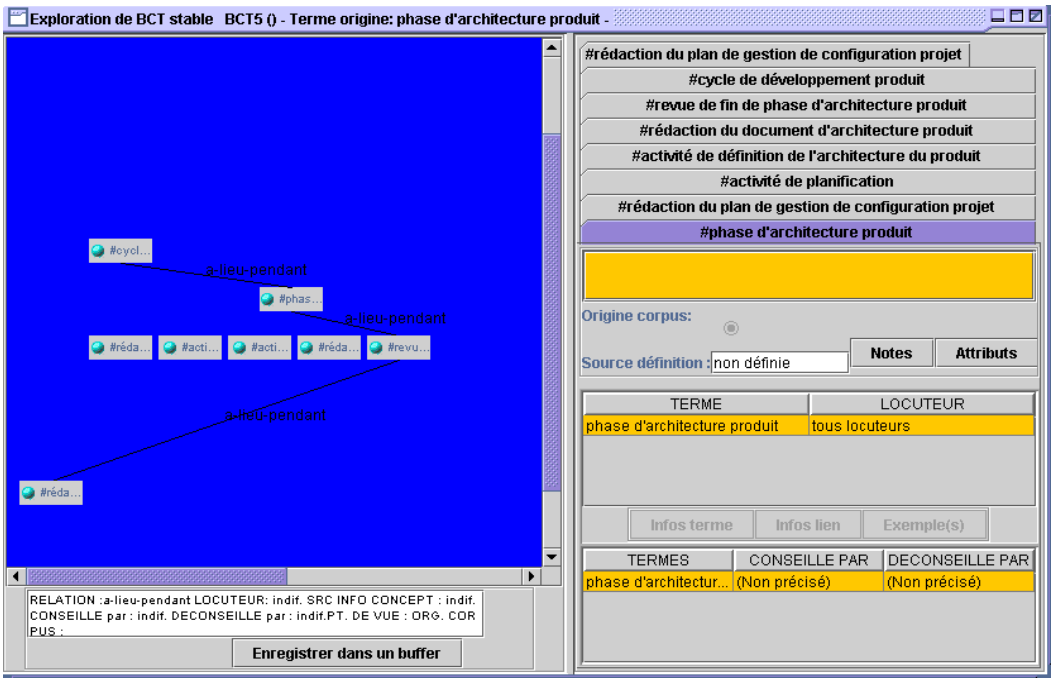

Figure 5.3.2 : Consultation en mode « préférence terminologique » de concepts reliés par *a-lieu-pendant*.

5.3.3 Évaluations expérimentales

L'évaluation de solutions en ingénierie des connaissances passe par la réalisation et surtout par l'utilisation de logiciels. J'ai donc retenu une démarche expérimentale pour évaluer mes hypothèses sur les BCT. Pour mesurer les retombées de choix et de principes fondamentaux, la réalisation d'outils est indispensable. De plus, l'énoncé de l'hypothèse lui-même contient la méthode de son évaluation : évaluer la pertinence d'une BCT pour construire une famille de systèmes (ici, des systèmes de consultation documentaire) suppose que l'on caractérise les types d'utilisation de la BCT dans ce cadre, les transformations des données requises, etc. Cette caractérisation théorique doit s'appuyer sur l'étude empirique d'un cas appliqué au moins.

Dans chacun des deux projets présentés ci-dessous, une BCT a été construite par des linguistes-terminologues à partir d'analyses de corpus avec le logiciel LEXTER. Dans les deux cas, GEDITERM a été utilisé pour rendre compte d'une terminologie et d'un réseau conceptuel unique, formant une BCT, qui mette au premier plan le contenu des textes et leur analyse linguistique. La BCT était alors considérée « comme un modèle du texte, c'est-à-dire comme un produit dont la construction ne fait pas appel à une connaissance extérieure à celle du corpus ».

5.3.3.1 Mouglis : des BCT à la consultation documentaire

Le projet « Terminologie, Modélisation des Connaissances et Systèmes Hypertextuels de Consultation de Documentation Technique » (1996-1998) a soulevé le problème d'exploitation des fonds documentaires techniques par le biais de l'étude sémantique de leur contenu. Traité en partenariat avec l'équipe « sémantique et corpus » de l'ERSS et une entreprise, il a permis d'évaluer l'intérêt des BCT pour constituer des applications dans un domaine, tant du point de vue théorique (modèle de structuration des données, méthodologie de recueil) qu'appliqué (utilisation des données). Ces applications sont ici des outils de consultation documentaire, les SCDT ou Systèmes de Consultation de Documents Techniques. Ils ont en commun avec les BCT un référentiel terminologique relié à un corpus documentaire et à un modèle conceptuel des produits et activités du domaine. Le domaine d'application proposé par le partenaire industriel (la DER d'EDF) concerne le génie logiciel. Le corpus est un document d'aide à la mise en œuvre d'une démarche de génie logiciel scientifique et technique MOUGLIS.

Résultats théoriques

C'est dans le cadre de ce projet qu'un modèle de données et une méthode de construction de BCT ont été élaborés, et avec eux d'autres résultats théoriques sur les BCT :

- une caractérisation des différences sémantiques entre le réseau conceptuel d'une BCT, un index, un modèle du domaine, une ontologie formelle, un modèle de tâche et une « terminologie » pour la consultation documentaire ;
- une étude de l'impact de l'application visée et de la formalisation choisie sur la représentation des connaissances.

Plusieurs modèles ont dû être réalisés pour le SCDT à partir de la BCT construite par les terminologues : un modèle de la tâche décrite dans le document ainsi qu'un modèle du domaine donnant naissance à un index. Cette diversité a permis à la fois d'évaluer l'architecture retenue pour les SCDT, de proposer une démarche pour les construire et en particulier de mesurer l'apport d'une BCT dans cet objectif.

Avec ce projet, j'ai abordé le thème « sémantique et mémoire externe » par le biais de la représentation (qui préfigure ici une mémoire externe) et de l'interprétation des connaissances (qui leur confère une sémantique). En effet, étudier le mode de construction d'une BCT puis son exploitation, c'est s'intéresser au passage du texte à la BCT puis de la BCT à différentes sortes de modèles. Au cours de ce processus, sont construites des abstractions selon des critères différents, des représentations qui devront être interprétées rationnellement par des individus ou par un système formel. Mon travail a consisté à fixer les interprétations possibles des données, à essayer d'expliciter les variations d'interprétation et de prendre en compte la finalité du processus.

Résultats expérimentaux

Les résultats expérimentaux concernent la BCT proprement dite, son utilisation pour construire les différentes entrées vers le texte qui sont proposées dans le SCDT (un modèle de la tâche et un index), et son intérêt pour l'aide à la rédaction.

Mon modèle de structuration de BCT assure une construction pertinente de la BCT et une bonne exploitation de son contenu. Il permet de rendre explicites des phénomènes linguistiques

comme l'homonymie et la synonymie. Enfin, l'expérience confirme l'importance des relations conceptuelles spécifiques au corpus.

Pour construire un modèle de la tâche, les concepts décrivant des tâches et déjà représentés sous forme hiérarchique dans la BCT ont été repris. Les liens entre ces tâches et des concepts décrivant des documents ou des éléments de logiciel ont permis de choisir des attributs décrivant les tâches. Mais la contribution essentielle de la BCT est dans la sélection de renvois associés aux concepts, c'est-à-dire des passages de textes indexés par les concepts et les tâches. Cependant, tous les passages sélectionnés dans la BCT n'ont pas convenu : il s'agit de contextes définitoires dans la BCT alors que, dans le modèle de tâche, on retient les contextes illustrant le rôle du concept dans la tâche.

Pour définir un index, la BCT a fourni une partie (seulement) des entrées, leur organisation en sous-entrées et les renvois vers les textes. Finalement, **la nature des données présentes dans le modèle est en partie remise en question** (non pas leur qualité linguistique, mais leur adéquation au besoin).

Enfin, la BCT s'est avérée un bon support pour une évaluation ergonomique du document car elle en permet une lecture finalisée et thématique. Elle facilite les recoupements et les vérifications de cohérence. Elle permet de repérer rapidement des phénomènes linguistiques déviants enregistrés par les linguistes : paragraphes ayant des titres analogues mais de contenus différents ; hétérogénéité du vocabulaire désignant les types de document et les étapes du cycle de vie. Un mémento a été rédigé pour diffuser la méthode sous un format plus synthétique que le guide complet. La BCT, mais plus encore les échanges avec les linguistes nécessaires à sa définition, ont servi à l'expert du domaine à assurer une meilleure cohérence du contenu.

5.3.3.2 SGDD : Construire une modélisation unique de plusieurs corpus

L'objectif pratique de ce projet était de fournir à quatre entreprises une vue d'ensemble de leurs terminologies respectives et des conceptualisations que celles-ci manifestent pour que ces entreprises puissent communiquer mieux. Il a été décidé de rendre compte des terminologies de chacune, de leurs convergences et de leurs divergences, pour ensuite mettre en place un langage commun, moteur d'un partage de leurs connaissances et de leurs savoir-faire. Du point de vue scientifique, ce projet a offert un nouveau cadre d'évaluation de GEDITERM, avec de nouvelles contraintes : la nécessité de rendre compte de modèles issus de plusieurs corpus, correspondant à plusieurs points de vue, puis de les mettre en correspondance. Ce type d'utilisation n'avait pas vraiment été prévu lors des spécifications de GEDITERM, et posait des problèmes nouveaux de gestion de représentations multiples. Enfin, il a confirmé de façon criante la non-neutralité de l'interprétation linguistique, comme rapporté dans l'article de M.P. Jacques et A.M. Soubeille (Jacques et Soubeille, 2000).

Résultats attendus

L'étude de la terminologie a donc eu pour but de repérer dans quelle mesure des termes identiques employés par plusieurs partenaires renvoyaient aux mêmes concepts ou à des concepts différents. Dans ce dernier cas de figure, un objectif complémentaire était de produire une évaluation de cette différence. Cette étude a donné lieu à la construction d'une BCT dont on espérait que le modèle conceptuel permette de rendre compte de ces différences. Sur un plan méthodologique, elle a été menée en deux temps : (1) l'analyse et la modélisation séparées de la terminologie de chaque partenaire, ce qui a conduit à construire quatre BCT ; (2) la fusion de ces quatre BCT en une seule, qui rend alors compte des différences et des similitudes de conceptualisation de chacun. La représentation des termes et des concepts dans la BCT finale s'appuie sur les textes, les validations d'experts et sur le réseau de chaque base spécifique.

Éclaircir la signification d'un terme implique donc de montrer pour chaque corpus non seulement à quel concept il est lié mais aussi comment ce concept est relié aux autres. L'étude a

fait apparaître des relations qui n'étaient pas présentes pour tous les locuteurs. Ceci tend à montrer l'hétérogénéité de ce que l'on pouvait considérer de prime abord comme étant **un** domaine, si l'on caractérise le domaine par son lien à une pratique sociale, ici une pratique professionnelle.

Restituer des phénomènes linguistiques

La question s'est posée de la complète restitution des divergences de signification. Le point de vue comparatif déterminé par l'objectif de l'étude a impliqué une représentation analogue pour tous. Pour se prononcer sur l'identité ou la différence de deux concepts, il était nécessaire que la sémantique des relations soit constante d'un sous-corpus à l'autre. Structurer les BCT selon une architecture semblable a semblé le moyen de faciliter l'étape ultérieure de « fusion » lors du traitement des concepts dans la BCT finale. Priorité a été donnée aux relations hiérarchiques et de composition. D'autres relations ont été définies à partir de la présence récurrente dans les textes de ce qui a été appelé un *schéma de communication* : X communique Z à Y via W, bien illustré par une phrase comme « *Informations communiquées par la DDE31 : le bulletin prévisionnel est transmis par télécopieur au CIGT31...* ». Ce schéma conduit à créer dans la BCT quatre relations binaires : `envoie-à (émetteur X _ récepteur Y) ; reçoit (récepteur Y _ message Z) ; émet (émetteur X _ message Z) ; médiatise (média W _message Z)`.

Étude des termes communs et des concepts associés

La *polysémie* est manifestée par un traitement unitaire : un terme auquel sont reliés plusieurs concepts. Par exemple, au terme *client* sont associés les trois concepts proches #client-mairie, #client-semvat et #client-smtc. *L'homonymie* est manifestée par un dégroupement : les concepts sont différenciés à partir d'un seul terme. Ainsi, deux concepts recouvrant des réalités très différentes correspondent à *exploitant*, #exploitant-smtc et #exploitant-mairie. Sur les 800 termes étudiés, seuls quatre sont strictement identiques pour les quatre locuteurs. La présence d'éléments de définitions, de relations ou de contextes identiques, a été considérée comme l'indice d'un terme polysémique et de concepts apparentés, et dans le cas contraire, de concepts disjoints et de termes homonymes.

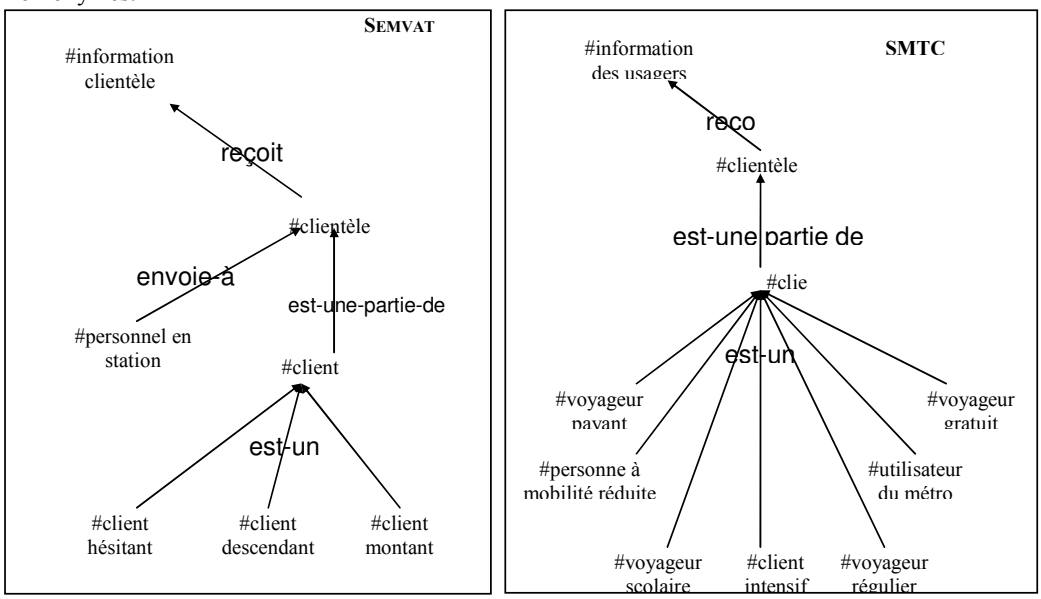

Figure 1 : le concept #client pour deux des locuteurs (SEMVAT et SMTC)

Ceci ouvre sur la question du traitement des termes en usage dans les textes provenant d'univers distincts mais rapprochés par des échanges professionnels ou finalisés entre les locuteurs

ou auteurs. Ces échanges estompent les frontières entre domaines et conduisent à l'utilisation de termes identiques. Il a été choisi, en rupture avec une tradition terminologique de plus en plus remise en question par les faits, de traiter ces cas de figures comme des manifestations de polysémie d'un terme, au regard de la parenté (parfois justifiée par l'opération de référence réalisée par le terme) entre des concepts élaborés dans des systèmes différents. Mais il serait intéressant d'affiner la description des concepts qui semblent ainsi partagés ou éventuellement empruntés à un domaine par un autre et de décider des relations à prendre en compte pour les décrire.

5. 4 - Modèles de ressources terminologiques et ontologiques : bilan

Les modèles et plate-formes de modélisation proposés pour construire des ressources terminologiques (GEDITERM) et ontologiques (TERMINAE) contribuent à répondre aux questions posées au chapitre 3 sur les modèles terminologiques, les ontologies et les textes. C'est ce que montre le bilan qui suit. D'autres éléments de réponse sont fournis au chapitre suivant sur les outils de TAL et leur intérêt pour la construction et l'utilisation d'ontologies en lien avec des documents.

5.4.1 Synthèse sur la notion de BCT : modèle neutre ou modèle lié à une application ?

Finalement, l'hypothèse d'une neutralité de la BCT par rapport aux textes a été remise en question :

- Même proche du texte, la BCT n'est de toute façon pas neutre car le texte lui-même véhicule une intention, celle de ses auteurs ou celle qui a motivé sa rédaction.
- En construisant un réseau conceptuel à partir de textes, les linguistes interprètent les textes et construisent des abstractions. En cela, la représentation résultant de leur analyse est bien une *construction* complètement influencée par leur *interprétation* du langage. De plus, leurs critères de modélisation restant implicites, ils peuvent fluctuer au cours du travail.
- Dès que l'on utilise ce modèle pour une application précise, on est obligé de revenir aux textes pour comprendre les choix de modélisation des linguistes. On refait donc un travail d'interprétation, désormais explicitement guidé par l'application ciblée.

Dans un premier temps [Rapport-MOUGLIS, 98], avec A. Condamines, nous avons défini les notions de BCT-corpus et BCT-Applicative pour apporter une solution à cette double tension présente au sein d'une BCT : la BCT-Corpus, affranchie de l'influence de l'application, serait le fruit de la seule analyse linguistique ; dans un deuxième temps, l'application serait prise en compte pour construire une BCT-Applicative par adaptation de la BCT-Corpus. Or cette proposition s'avère finalement lourde et peu pertinente :

- L'hypothèse que la BCT-Corpus soit « neutre » par rapport aux textes est mise à mal par le constat d'une certaine *interprétation*, inévitable dans l'analyse linguistique. On peut faire l'hypothèse que cette interprétation soit peu variable d'un linguiste à l'autre puisque la situation d'analyse est bien balisée : le corpus et les objectifs de modélisation sont fixés.
- La construction d'une BCT-Applicative requiert bien plus qu'une adaptation, il s'agit plutôt d'une *re-création* : la BCT-Corpus n'étant pas toujours suffisante, un recours au texte s'impose, ce qui revient à faire le travail d'analyse linguistique.
- Au final, la démarche globale nécessite du temps et des ressources coûteuses, en particulier deux types de compétence (linguistiques puis en modélisation des connaissances).
- Un autre risque est de perdre des informations au cours de la modélisation puisqu'elle est effectuée par deux personnes différentes.

Enfin, une dernière restriction est que la vérification formelle est repoussée à la fin du processus, au risque de remettre en question des connaissances incohérentes ou non valides trouvées au début et véhiculées inutilement. De plus, tant qu'on ne dispose pas de critère formel permettant de distinguer une BCT-Applicative d'une BCT-corpus, la frontière n'est pas gérable et semble tout à fait discutable. Il faudrait pouvoir préciser dans quel état doivent être les données en fin de construction de chaque type de BCT.

Dans le cadre de l'ingénierie des connaissances, rendre compte précisément d'un texte par une analyse linguistique sans prendre en compte l'application s'avère inutilement coûteux. Par contre, **tant la structure de données des BCT que les techniques d'analyse linguistique utilisées peuvent être reprises pour construire un modèle ou une terminologie adaptés à des besoins particuliers.**

5.4.2 Originalité des propositions relatives aux BCT

Mes propositions en matière de BCT se situent tant au niveau pratique (modèle de données, environnements de gestion GEDITERM et de consultation CONSULTERM) que théorique (principes de définition des termes et concepts, nature des connaissances dont ce type de modèle rend compte, méthodologie de construction).

Le logiciel GEDITERM est une contribution originale au niveau national (c'était un des seuls **logiciels de gestion de terminologie** inspirés des réseaux sémantiques avant TERMINAE) qu'international, où ses points forts sont la méthodologie de construction associée, le lien vers les textes sources ainsi que la possibilité d'une visualisation graphique. Les logiciels comparables (comme CODE4) (Skuce, 1991) se focalisent plus sur la représentation des connaissances. Depuis, un système analogue a été réalisé à l'université Pompeu Fabra de Barcelona (T. Cabré).

Sur le plan théorique, une première contribution est le **modèle de données** dont l'intérêt dépasse largement celui des terminologies et pour interroger la représentation des ontologies. Il a été repris pour enrichir la représentation des connaissances dans TERMINAE. Ce modèle répond en partie à la volonté de rendre compte de la terminologie d'un domaine et des connaissances qu'elle reflète (question 2 du 3.2.2.). Ce modèle de données intègre des données terminologiques, leur sémantique (à travers un réseau conceptuel) et l'usage des termes à l'aide de liens vers des extraits de textes. De plus, il permet de conserver les outils ou éléments de textes ayant servi à identifier et représenter les parties du domaine. C'est certainement un point fort pour anticiper des utilisations de la BCT pour accéder aux contenus de textes, comme l'ont montré les projets HYPERPLAN et MOUGLIS (le modèle servant à construire des index de consultation). Il s'agit là d'une première manière de tirer profit de ce type de modèle dans des applications de gestion documentaire (question 5 du 3.2.2.)

Un deuxième apport est une **réflexion sur la neutralité des modèles**, leur « distance » par rapport aux textes d'une part, à l'application ciblée d'autre part. En effet, l'hypothèse de neutralité de la BCT par rapport aux textes a été remise en question. Le texte véhicule une intension, qui est éventuellement détournée par les objectifs de la modélisation. En cela, son analyse débouche sur une *construction* complètement influencée par une *interprétation finalisée* du langage. Finalement, **il est illusoire et inutilement coûteux de dissocier une analyse linguistique, qui se voudrait une restitution neutre du texte, de la prise en compte de l'application**. Cette conclusion vaut également pour la construction des ontologies.

Cette analyse contribue à évaluer les possibilités d'utiliser et de réutiliser les modèles construits à partir de textes (question 7 du chapitre 3). Derrière la « neutralité » des BCT-corpus, il y a l'espoir d'une grande réutilisabilité. Or la réutilisabilité des BCT ne peut venir d'une analyse la plus détaillée possible des textes. Il semble donc plus pertinent de privilégier l'utilité des modèles en prenant en compte dès le dépouillement terminologique ce à quoi ils doivent servir. Or ces BCT

s'avèrent alors très spécifiques et peu réutilisables. La réflexion méritait d'être poursuivie, ce qui a été fait dans le cadre de la définition d'une méthode de construction d'ontologies (chapitre 6).

5.4.3 Des BCT aux ontologies

En matière de représentation d'ontologie, le modèle de données proposé reprend les principes du modèle de représentation des BCT. Il a été implémenté dans la plate-forme TERMINAE. Par rapport aux représentations classiques d'ontologies, et au standard OWL, je défends la nécessité de conserver avec l'ontologie une représentation du texte qui a servi de source de connaissances et une représentation des termes servant à désigner les concepts. Ces modèles doivent se situer au niveau conceptuel dans un premier temps, avant d'être formalisés. Ce modèle de données est à la fois plus riche que OWL, et moins précis, puisqu'il ne permet pas de représenter des axiomes ou des règles. Ces résultats constituent le 2e volet de ma réponse à la question sur les représentations posée au chapitre 3 (question 2 du 3.2.2).

La transposition des résultats obtenus pour les BCT aux ontologies n'est pas triviale. Les points communs entre ces types de modèle invitent à utiliser des approches et des outils identiques pour les construire, en ayant des exigences de structuration spécifiques à chaque modèle. La différence de fond entre les deux structures des données n'est pas tant leur format que l'utilisation qui en est prévue, en particulier pour raisonner. Les capacités de raisonnement possibles à l'aide d'une ontologie sont liées au langage utilisé pour sa formalisation et à la richesse des connaissances représentées. Le fait de disposer d'un lexique riche et de liens vers des textes ne modifie en rien ces capacités. En revanche, en enrichissant ainsi une ontologie, on favorise son utilisation pour l'indexation ou l'annotation, son interprétation et son utilisation pour l'interaction homme-machine (chapitre 6).

L'étude successive de ces deux types de modèle fait ressortir qu'ils répondent à des objectifs différents. Une terminologie ou une BCT est construite pour rendre compte de termes jugés pertinents en fonction d'un domaine et du langage utilisé dans des textes. Une ontologie doit permettre une interprétation formelle et un raisonnement conceptuel. L'ontologie peut ne pas faire référence à la langue. Cette analyse amorce la réflexion sur la qualification de la validité des modèles par rapport à des classes d'applications (question 6 du 3.2.2). De fait, d'autres types de ressource, plus simples que les ontologies, comme les hiérarchies de concepts ou les terminologies, sont parfois plus pertinents pour certaines applications de recherches d'information ne faisant pas appel à des raisonnements. Les ontologies se distinguent aussi des BCT car elles sont prévues pour favoriser la réutilisabilité. Les concepts de l'ontologie sont a priori suffisamment consensuels et abstraits pour anticiper différents usages ou raisonnements, pour être compris et interprétés au sein de plusieurs applications. Ce point (question 7 du 3.2.2) reste encore à valider pour le modèle et la méthode définis.

5. 5 - Publications sur ces travaux[22]

[KAW, 95] N. AUSSENAC-GILLES, D. BOURIGAULT, A. CONDAMINES, C. GROS. How can knowledge acquisition benefit from terminology ? *Proceedings of the 9th Knowledge Acquisition Workshop.* Banff, Univ. of Calgary (CA). Feb. 1995.

[JAC, 96] D. GARCIA, N. AUSSENAC-GILLES, AND A. COURCELLE. Exploitation, pour la modélisation, des connaissances causales détectées par COATIS dans les textes. *In Journées Françaises d'Acquisition des Connaissances JAC'96*, pp 123-136, mai 1996. LIRMM, Université de Montpellier.

[22] Présentation par ordre chronologique.

[DEA-SEGUELA, 96] SEGUELA P., Formalisation et modélisation des connaissances terminologiques. *Rapport de DEA Représentation des Connaissances et Formalisation du Raisonnement, Université Toulouse III*. Juin 1996.

[TIA, 97] P. SEGUELA AND N. AUSSENAC-GILLES. Un modèle de base de connaissances terminologiques. In *Deuxièmes rencontres 'Terminologie et Intelligence Artificielle' TIA'97*, avril 1997. Équipe de Recherche en Syntaxe et Sémantique (ERSS), Université Toulouse-Le Mirail. 1997 : 47-68.

[Mémoire-FOURNIER, 98] FOURNIER D., 1998, *Étude et conception d'un système de gestion d'une Base de Connaissances Terminologiques*, Mémoire d'ingénieur CNAM, Toulouse. Mars 1998.

[Mémoire-LECORGNE, 98] LECORGNE E., Étude et conception d'une maquette de consultation de base de connaissances terminologiques. Rapport de stage d'ingénieur CNAM, Toulouse. Déc. 1998.

[Mémoire-SIMON, 98] SIMON S. Intervention dans la réalisation d'un système de gestion de BCT et formalisation des connaissances dans une BCT. Rapport de stage ingénieur ENSEEIHT. Juin 1998

[DEA-SIMON, 98] SIMON S. Représentation formelle des connaissances issues d'une Base de Connaissances Terminologiques. Mémoire de DEA « Représentation de la connaissance et formalisation du raisonnement ». Univ. P. Sabatier, Toulouse. Sept 1998.

[Rapport-MOUGLIS, 98] N. AUSSENAC-GILLES, A. CONDAMINES. Rapport Final de projet GIS Sciences de la cognition - *Terminologie, Modélisation des Connaissances et Systèmes Hypertextuels de Consultation de Documentation Technique*. Rapport Interne. IRIT/98-20-R, IRIT, mai 1998.

[ISKO, 99] AUSSENAC-GILLES N., CONDAMINES A., Bases de connaissances Terminologiques : enjeux pour la consultation documentaire. J. Maniez et W. Mustapha El Hadi (Ed.) *Organisation des connaissances en vue de leur intégration dans les systèmes de représentation et de recherche d'information,* Villeneuve d'Asq : Presses de l'Université Charles de Gaulle (UL3 travaux et Recherches), 1999 : 71-88.

[TIA, 99] AUSSENAC-GILLES N. GEDITERM : un logiciel pour gérer des bases de connaissances terminologiques. *Actes des Journées Terminologie et Intelligence Artificielle TIA'99*. Nantes (F), mai 1999 : 129-150.

[TERMINO, 99] AUSSENAC-GILLES N., GEDITERM : un logiciel pour gérer des bases de connaissances terminologiques. *Terminologies Nouvelles*, 19, Nov. 1999. 111-123.

[Mémoire-GAFFORI 99] GAFFORI H., Évaluation ergonomique d'un logiciel d'administration de bases de connaissances terminologiques : GEDITERM. Rapport de stage de DESS en ergonomie, Université de Toulouse Le Mirail. 1999.

[Livre-IC, 00] GARCIA D., AUSSENAC-GILLES N., Exploitation, pour la modélisation, des connaissances causales détectées par COATIS dans les textes. *Ingénierie des connaissances,* Eds. D. Bourigault, J. Charlet, G. Kassel, M. Zacklad. Paris : Eyrolles. janvier 2000 : 257-274.

[Traité-IC, 01] AUSSENAC-GILLES N., CONDAMINES A., Entre textes et ontologies formelles : les bases de connaissances terminologiques. *Ingénierie et capitalisation des connaissances.* Eds. M. Zacklad, M. Grundstein. Paris : Hermès. Traité IC2. 2001 : 153-177.

[ASSTICCOT, 04] N. AUSSENAC-GILLES AND A. CONDAMINES. *Action spécifique STIC « Corpus et Terminologie » ASSTICCOT (AS 34). Rapport final*. Rapport Interne IRIT/2003-23-R. Oct. 2003. 70 p.

CHAPITRE 6 - DES TEXTES AUX ONTOLOGIES, DES ONTOLOGIES AUX TEXTES

Le chapitre 5 m'a permis d'exposer des modèles conceptuels qui associent terminologie et réseau conceptuel d'un domaine. Or j'ai étudié ces types de modèle dans la perspective de restituer les résultats d'analyses de textes. Je reviens dans ce chapitre sur l'ingénierie de ces modèles, c'est-à-dire à la définition de logiciels et de principes méthodologiques pour conduire le processus qui va de la sélection de textes à l'élaboration d'un modèle. L'évaluation de mes propositions passe par une série d'expériences dans différents contextes, qui ont en commun d'utiliser le modèle pour fournir un accès aux textes.

Cette problématique relève toujours de la modélisation conceptuelle, reformulée comme celle du repérage de connaissances à partir de textes en langage naturel en vue de les utiliser pour construire des modèles conceptuels. Il s'agit non pas d'automatiser la production de représentations (conceptuelles ou formelles) mais bien de fournir un guide pour faciliter l'exploration d'un ensemble de textes afin d'en extraire des indices pour définir des représentations conceptuelles. J'ai choisi d'exploiter en priorité des logiciels de traitement automatique des langues basés sur des principes linguistiques. Du point de vue méthodologique, ma réflexion s'enrichit des débats engagés au sein du groupe TIA sur la construction de ressources terminologiques et ontologiques (RTO par la suite) selon des principes différentiels.

Ce chapitre s'organise donc en trois parties principales. Dans la première partie (§ 6.1), je montre l'atout que représentent les logiciels de traitement automatique des langues pour la modélisation de ressources terminologiques et ontologiques, moyennant la disponibilité d'interfaces d'exploitation de leurs résultats. À partir d'un panorama des types de logiciel applicables, je tire un bilan de plusieurs évaluations d'extracteurs de termes. Je présente alors le logiciel CAMELEON *d'aide à l'identification de relations conceptuelles à l'aide de patrons lexico-syntaxiques.*

Dans une deuxième partie (§ 6.2), je présente ma contribution au niveau méthodologique. La méthode et la plate-forme TERMINAE *intègrent le modèle d'ontologie à composante terminologique présenté au chapitre 5, des logiciels de traitement automatique des langues et des principes de normalisation pour l'élaboration d'ontologies ou de terminologies. Là encore, des expériences d'utilisation viennent illustrer la pertinence de ces résultats. Enfin, une dernière partie (§ 6.3) montre en quoi des modèles conceptuels construits à partir de textes sont particulièrement adaptés pour faciliter l'accès au contenu des textes. À travers des cas précis d'utilisation, j'illustre l'utilisation de ces modèles pour indexer, annoter ou gérer des textes sous forme de documents électroniques.*

6. 1 - Traitement Automatique des Langues et modélisation conceptuelle

6.1.1 Contexte : Traitement du langage naturel pour l'identification de connaissances

6.1.1.1 Rôle des logiciels de TAL dans l'identification de connaissances

Les travaux de recherche sur le développement d'outils d'aide à la construction de RTO à partir de textes peuvent s'organiser selon une typologie fonctionnelle et correspondant à l'état de l'art en 2002. Les outils qui m'intéressent sont ceux qui servent à *construire* des ontologies (*ontology design*), à définir les concepts génériques et les relations qui la composent. D'autres types de logiciel, en général basés sur des principes d'apprentissage, visent à *instancier* l'ontologie (*ontology population*) en cherchant dans les textes des occurrences d'instances de concepts. Ces outils servant à associer des concepts à des endroits précis des textes (les occurrences de termes), ils facilitent aussi l'indexation de documents à l'aide d'ontologies.

Acquisition de termes : Une première classe regroupe les outils dont la visée est l'extraction, à partir du corpus analysé, de *candidats termes*. Sont candidats des mots ou groupes de mots susceptibles d'être retenus comme termes par un analyste, et de fournir des étiquettes de concepts. Ces outils diffèrent principalement quant au type de techniques mises en œuvre (syntaxique, statistique, autres).

Mise en forme : Puces et numéros

Structuration de termes et regroupement conceptuel : Les ressources termino-ontologiques se présentent rarement sous la forme d'une liste à plat. Des outils d'aide à la structuration d'ensembles de termes sont donc nécessaires. Dans cette classe, j'évoque, d'une part, des outils de classification automatique de termes et, d'autre part, des outils de repérage de relation. Signalons que beaucoup d'outils d'extraction proposent déjà une structuration des candidats termes extraits. En fait, la classification des termes est une des méthodes pour l'*identification de concepts*, ou pour l'identification d'instances de concepts par l'association de termes à des classes. En revanche, le *repérage de relations sémantiques* permet une mise en relation des concepts. On distingue souvent les travaux portant sur les relations hiérarchiques, qui occupent une place privilégiée, des outils d'aide au repérage d'autres relations.

6.1.1.2 Typologie des approches pour l'identification de termes

Le repérage de termes correspond au repérage de syntagmes en traitement du langage naturel, et relève donc d'un problème de découpage de phrase en unités englobant plusieurs termes et ayant une unité. Les approches existantes forment trois grands groupes :

- *La morphologie et la syntaxe au service du repérage de termes* : dans ce cas, on s'appuie sur des traitements linguistiques (basés sur la forme et la composition des mots ainsi que sur les constructions syntaxiques des syntagmes dans la langue étudiée). Par exemple, des règles peuvent chercher des groupes nominaux de type *Nom de Nom*.
- *Des approches statistiques* visent le repérage de segments répétés (n-grammes présents plusieurs fois dans le corpus) ou utilisent l'information mutuelle (mots voisins communs aux mots étudiés) pour décider que deux mots ne sont pas à côté par hasard et forment bien un syntagme.
- Enfin, plusieurs systèmes exploitent au contraire les frontières de termes, ces *approches « en négatif »* ayant l'avantage de retenir des formes très variées de syntagmes.

L'outil TERMINO, pionnier de l'acquisition automatique de termes (David et Plante, 1990) se focalise sur le repérage des syntagmes nominaux, appelés "synapsies" d'après les travaux de

Benveniste, seules structures supposées produire des termes. Après une phase d'analyse morphosyntaxique, les synapsies sont générées à partir des dépendances entre têtes et compléments rencontrées dans la structure de syntagme nominal retournée par l'analyseur.

ANA extrait des candidats termes sans effectuer d'analyse linguistique (Enguehard et Pantera, 1995). Les termes sont reconnus au moyen d'égalités approximatives entre mots et d'une observation de répétitions de patrons. Au contraire, ACABIT extrait des candidats termes à partir d'un corpus étiqueté et désambiguïsé (Daille, 1994). L'acquisition terminologique dans ACABIT se déroule en deux étapes : (1) analyse linguistique et regroupement de variantes, au moyen de transducteurs qui analysent le corpus étiqueté pour en extraire des séquences nominales et les ramener à des candidats termes binaires ; (2) filtrage statistique des candidats termes binaires.

À l'instar d'ACABIT, LEXTER extrait des candidats termes à partir d'un corpus préalablement étiqueté et désambiguïsé (Bourigault, 1994). Il effectue une analyse syntaxique de surface pour repérer les syntagmes nominaux maximaux, puis une analyse syntaxique profonde pour analyser et décomposer ces syntagmes. Il est doté de procédures d'apprentissage endogène pour acquérir des informations de sous-catégorisation des noms et adjectifs propres aux corpus. Il organise l'ensemble des candidats termes extraits sous la forme d'un réseau.

FASTR est un analyseur syntaxique robuste dédié à la reconnaissance en corpus de termes appartenant à une liste contrôlée fournie au système (Jacquemin, 1997). Les termes n'ayant pas toujours, en corpus, la même forme linguistique, le principal enjeu est de pouvoir identifier leurs variantes. FASTR est doté d'un ensemble élaboré de métarègles qui lui permettent de repérer les variantes syntaxiques, morpho-syntaxiques et sémantico-syntaxiques. Enfin, dans une optique de modélisation conceptuelle, SYMONTOS (Velardi *et al.,* 2001) propose des outils pour repérer des termes simples et complexes dans des textes et des critères pour décider de définir des concepts à partir de ces termes.

On trouvera un état de l'art plus complet dans (Cabré *et al.,* 2000).

6.1.1.3 Typologie des approches pour la structuration de termes

La gamme des outils d'aide à la structuration de terminologies est très large, couvrant les outils de classification de termes sur la base de cooccurrences dans des textes ou dans des fenêtres, les outils de classification de termes sur la base de distributions syntaxiques et les outils de repérage de relations. Les outils de cooccurrence développés dans le domaine de la recherche d'information rapprochent des termes qui apparaissent fréquemment dans les mêmes (portions de) documents, et qui possèdent donc sans doute une certaine proximité sémantique. La technique de recherche de co-occurrents a été promue très tôt en informatique documentaire pour permettre l'expansion de requêtes (Sparck Jones, 1971). Parmi les applications pour l'acquisition terminologique, on peut citer le projet ILIAD (Toussaint *et al*., 1998) et les travaux de G. Lame (2002). Toujours en informatique documentaire, les travaux visant la construction automatique de thesaurus peuvent être réinvestis dans des applications terminologiques et ontologiques. Par exemple, la chaîne de traitement développée par G. Greffenstette construit automatiquement des classes comportant des noms qui se retrouvent régulièrement comme arguments des mêmes verbes (Greffenstette, 1994). Ce repérage de la position des noms en position d'argument se fait grâce à l'exploitation d'un analyseur syntaxique de surface à large couverture. Ces techniques inspirées de la linguistique harrissienne, qui visent à rapprocher les termes qui ont des distributions syntaxiques analogues, sont à la base de nombreux travaux (Assadi, 1998) (Habert *et al*, 1996) (Faure, 2000).

Dans tous les cas, le système peut se contenter de présenter les phénomènes observés (cas de SYNTEX) ou alors il constitue automatiquement des classes selon des techniques de clustering par exemple. De plus en plus de travaux en TAL et apprentissage automatique (ML) ou extraction d'information (IE) s'intéressent à cette étape. En effet, plus le traitement linguistique amont est élaboré, plus les rapprochements peuvent être nombreux et pertinents. Par exemple, si l'on sait regrouper toutes les variantes et les synonymes d'un terme avant de constituer les clusters, on

pourra calculer les rapprochements en y assimilant les termes et leurs synonymes, ce qui renvoie à plus d'occurrences dans les textes. De même, si les phénomènes d'anaphore sont résolus, même partiellement, de nouveaux contextes des termes peuvent être exploités (Nédellec, 2004).

6.1.1.4 Typologie d'outils pour l'identification de relations

Les outils précédents visent à rapprocher des termes à partir d'une analyse globale de l'ensemble de leurs occurrences. Ils ne touchent que les termes fréquents, et donc le plus souvent des noms simples, et proposent une simple relation d'équivalence (appartenance à une classe). À côté de ces outils qui travaillent sur les types comme regroupement des occurrences, on trouve les outils de repérage de relations, qui travaillent au niveau des occurrences elles-mêmes. Ces outils détectent en corpus des mots ou contextes syntaxiques répertoriés comme susceptibles de " marquer " telle ou telle relation entre deux éléments. L'extraction de relations sémantiques à partir de textes est une problématique récurrente depuis 1995, qui prend de l'ampleur dans la mesure où il se confirme que les relations lexicales correspondent à des contextes riches qui contribuent efficacement à la définition des concepts. Elle se développe avec les perspectives d'automatisation de la construction de modèles de connaissances. Les approches potentielles sont multiples et correspondent à des hypothèses différentes sur le statut des terminologies et des ontologies.

Les travaux de M. Hearst, sur l'extraction automatique des liens d'hyperonymie, font figure de référence (Hearst, 1992). Les recherches sur ce thème se déclinent de multiples façons. L'un des enjeux principaux concerne la généralité des relations et celles des marqueurs de relations. D'un côté, il existe probablement des relations que l'on jugera toujours pertinentes pour décrire un domaine de connaissance, par exemple les relations hiérarchiques ou partitives, et des marqueurs pour ces relations eux aussi généraux (Garcia, 1998). À l'opposé, chaque domaine est aussi structuré par des relations qui lui sont spécifiques, et qu'il convient nécessairement de prendre en compte. De plus, même dans le cas de relations considérées comme générales, il est possible que les marqueurs susceptibles de conduire à les identifier diffèrent d'un corpus à l'autre. Se pose alors le problème de l'apprentissage inductif de ces marqueurs de relation. Un certain nombre de travaux en TAL et en ingénierie des connaissances sont consacrés à ce problème. Ils partent tous du même principe d'une recherche itérative alternée dans le corpus à la fois des marqueurs d'une relation donnée et des couples de termes qui entrent dans cette relation (Rousselot et al., 1996) [IC, 99] (Morin, 1999) (Condamines et Rebeyrolles, 2000) (Maedche et Staab, 2000).

6.1.2 Aide au repérage et à la structuration de termes à partir de textes

Une fois établi, grâce au projet SADE, l'intérêt d'utiliser un extracteur de termes pour la modalisation des connaissances d'un domaine, un de mes sujets de recherche a été de définir précisément les modalités d'utilisation de ces logiciels, à savoir des repères méthodologiques pour leur utilisation dans le cadre de la construction de différents types de ressource terminologique et ontologique. Ce travail a été mené en collaboration avec D. Bourigault, concepteur des logiciels utilisés (LEXTER puis SYNTEX et UPERY) ainsi qu'avec des linguistes utilisateurs de ces outils au même titre que moi. J'ai également cherché à identifier les liens éventuels qui existent entre la nature des corpus analysés, les approches retenues pour l'extraction de termes et la pertinence des résultats pour un type d'application donné. Je présente dans la suite les logiciels utilisés avant d'exposer des propositions méthodologiques.

6.1.2.1 Logiciels utilisés

LEXTER

Le logiciel LEXTER (Logiciel d'EXtraction de TERminologie) est un extracteur de candidats termes (Bourigault, 94). LEXTER utilise en entrée des corpus de textes techniques d'un domaine quelconque qu'il traite au moyen d'une analyse syntaxique automatique partielle. Sur la base de patrons morpho-syntaxiques qui permettent de délimiter les frontières de groupes nominaux, le logiciel fournit en sortie une liste d'unités terminologiques candidates susceptibles de représenter les concepts du domaine étudié. LEXTER s'intéresse essentiellement aux noms et aux groupes nominaux. La liste des candidats termes complexes est structurée en deux composants, sa *Tête* (i.e. RESEAU dans le terme RESEAU REGIONAL) et son *Expansion* (REGIONAL dans le terme RESEAU REGIONAL).

Un *réseau terminologique* est ainsi fourni, très dense, qui relie chaque candidat terme complexe à ses composantes. Ce réseau grammatical s'appuie sur la construction syntaxique des termes et non sur leur signification. À chaque candidat terme sont associées des informations numériques, sur lesquelles l'utilisateur peut se baser pour organiser son dépouillement : sa *fréquence* (nombre d'occurrences du candidat terme détectées par le logiciel dans le corpus) et sa *productivité en Tête (resp. Expansion)*. La productivité correspond au nombre de candidats termes plus complexes qui ont le candidat terme en position tête (resp. expansion).

LEXTER possède une interface hypertextuelle, HTL, qui facilite l'accès à ce réseau et à ses constituants, et qui permet de naviguer de terme à terme suivant les liens *tête* et *expansion*. À partir d'un terme, on peut visualiser des listes paradigmatiques de candidats termes partageant la même tête ou la même expansion, ce qui guide vers la constitution de taxinomies locales. Le système donne également accès aux textes qui contiennent les termes. Cet accès au texte est d'autant plus crucial que l'utilisateur n'est pas un spécialiste du domaine. HLT fournit aussi à l'utilisateur la fréquence d'apparition de chaque candidat dans le corpus et sa productivité dans le réseau. Ces données sont autant de paramètres de tri pour afficher la liste des candidats. HLT facilite pour cela la définition de filtres qui orientent l'accès aux termes candidats et leur visualisation sélective.

Afin d'exploiter plus efficacement le réseau terminologique, trois outils ont été développés par H. Assadi (Assadi et Bourigault, 1996) : typage de termes (en *objet*, *attribut*, *action* ou *valeur*) ; structuration conceptuelle partielle à l'aide de relations prédéfinies entre types ; regroupement en classes d'adjectifs qualifiant les mêmes noms.

Mise en forme : Puces et numéros

Extraction de termes : SYNTEX

SYNTEX (Bourigault et Fabre, 2000) est un analyseur syntaxique de corpus. Il existe actuellement une version pour le français et une version pour l'anglais. Successeur de LEXTER, SYNTEX s'appuie sur une approche différente de l'analyse des textes pour produire des résultats plus riches que ceux de LEXTER, en partie grâce au module UPERY. Après l'analyse syntaxique en dépendance de chacune des phrases du corpus, SYNTEX construit un réseau de mots et de syntagmes (verbaux, nominaux, adjectivaux), dit « réseau terminologique », du même type que celui de LEXTER : chaque syntagme est relié d'une part à sa tête et, d'autre part, à ses expansions. Les éléments du réseau (mots et syntagmes) sont appelés « candidats termes ». A chaque candidat terme sont associées les mêmes informations numériques : fréquence et productivité.

La difficulté essentielle pour l'utilisateur vient de la masse des résultats fournis par l'extraction. Même s'il existe de nombreux travaux fort intéressants sur le filtrage statistique de candidats termes extraits automatiquement de corpus, l'expérience montre qu'aucune mesure statistique ne peut suppléer l'expertise de l'analyste, en particulier parce qu'il y a toujours des candidats termes de fréquence 1 dont l'analyse est intéressante. De façon générale, sachant qu'il ne pourra analyser tous les candidats termes extraits du corpus, l'analyste doit adopter une stratégie optimale qui, étant donné le temps qu'il a choisi de consacrer à la tâche d'analyse textuelle et en

fonction du type de la ressource à construire, lui garantit que, parmi les candidats qui auront échappé à son analyse, la proportion de ceux qui auraient pu être pertinents est faible.

Analyse distributionnelle : UPERY

UPERY (Bourigault, 2002) est un outil d'analyse distributionnelle. Il exploite l'ensemble des données présentes dans le réseau de mots et syntagmes construits par SYNTEX pour effectuer un calcul des proximités distributionnelles entre ces unités. Ce calcul s'effectue sur la base des contextes syntaxiques partagés. Il met en œuvre le principe de l'analyse distributionnelle du linguiste américain Z. S. Harris, dans la lignée des travaux de H. Assadi (Assadi & Bourigault, 1996). L'analyse distributionnelle rapproche d'abord deux à deux des candidats termes qui partagent un grand nombre de contextes syntaxiques. Par exemple, dans le projet VERRE, le candidats terme *glass yarns* est rapproché de *materials*, *filaments, crushed basaltic stone* et *fragments* car ils partagent plusieurs contextes : ils sont sujets du verbe *to be* et compléments du nom *mixture of*. En fait, ces cinq groupes nominaux rapprochés sont des matières premières ou des produits intermédiaires et finaux de la production de la fibre de verre.

Trois mesures permettent d'appréhender la proximité entre deux candidats termes : le nombre de contextes syntaxiques partagés par les deux termes ; pour chaque contexte partagé, l'inverse de sa productivité ; pour chaque couple de candidats termes, le rapport entre le nombre de contextes partagés et le nombre total de contextes dans lesquels il apparaît. UPERY calcule ces coefficients pour chaque couple de candidats termes, et ne sont présentés à l'utilisateur que les couples dont les coefficients dépassent certains seuils fixés empiriquement. L'analyse distributionnelle implémentée dans UPERY est symétrique : on calcule aussi la proximité entre contextes syntaxiques. Deux contextes syntaxiques sont proches si on y trouve les mêmes termes. Par exemple, dans un corpus de médecine, les verbes *montrer* et *mettre en évidence* sont proches car ils partagent en position sujet les termes *échographie*, *bilan infectieux*, *tomodensitométrie*, *artériographie*, etc.

6.1.2.2 Des mots aux termes : analyse des candidats termes

L'usage d'un extracteur de termes a été très tôt envisagé comme un moyen prometteur de faciliter la construction de terminologies ou de modèles conceptuels. J'en ai déjà évoqué plusieurs avantages attendus. Lorsque ce type de logiciel est utilisé, le dépouillement des résultats retournés devient une phase importante dans l'élaboration d'une représentation. L'analyse des candidats termes est une tâche qui peut être déroutante et qui requiert de définir des critères méthodologiques pour éviter qu'elle ne soit fastidieuse et trop longue. Dans mes différents projets en collaboration avec l'ERSS, plusieurs approches ont été expérimentées avant de stabiliser une proposition à la fois matériellement réaliste et théoriquement fondée.

La première approche reflète mes **propositions initiales sur les BCT**. Elle suppose que, dans un premier temps, un terme est défini par l'ensemble de ses occurrences en corpus, qui peuvent être récupérées. Ensuite, il est terme parce qu'il désigne un concept. L'analyse des termes se déroule alors en deux étapes : identification des termes candidats et recherche de relations lexicales.

- Une première *sélection s'opère hors contexte*, le linguiste définissant des critères de sélection de termes à partir des listes de candidats termes. Ces critères peuvent concerner les candidats termes considérés individuellement (termes contenant des éléments trop généraux, trop vagues, etc.) ou les uns par rapport aux autres (termes comportant des éléments opposés comme CONCEPTION GENERALE vs CONCEPTION DETAILLEE ou des synonymes*)*.
- Ensuite, la *sélection en contexte* (par lecture des occurrences) permet de travailler sur les relations conceptuelles dans le corpus. Il s'agit de construire un modèle de ces relations, caractérisant leur manifestation linguistique, puis de l'appliquer au corpus pour relever les occurrences de ces relations.

Cette manière de procéder revient à faire jouer un rôle important aux interprétations « neutres » du linguiste et des experts du domaine, qui sont finalement les garants des différentes sélections *a priori*. Elle correspond à notre première hypothèse sur les BCT, qui seraient vues comme une étape dans la construction d'un MC, préparant le travail du cogniticien. La prise en compte d'un point de vue plus influencé par la représentation des connaissances n'interviendrait que pour utiliser la BCT et l'adapter pour répondre aux objectifs associés à une application donnée. Ce serait supposer qu'il n'y a pas d'interprétation au moment de construire la BCT. Or la BCT est de fait un modèle conceptuel particulier : c'est une construction, artificielle, et le linguiste qui la construit doit choisir entre plusieurs possibilités pour rendre compte sous forme de réseau conceptuel de ce qu'il tire des textes.

Une **deuxième approche** est désormais retenue et sera appliquée aux **ressources terminologiques et ontologiques**. Désormais, un mot ou un syntagme prend le statut de terme à partir du moment où se justifie la définition d'un concept associé à ce terme. L'identification conceptuelle est couplée à la définition terminologique. *L'étude des contextes* riches en connaissances est alors le moyen privilégié de décider simultanément des termes et des concepts d'un domaine. Comme le montrent (Condamines et Rebeyrolles, 00a), l'étude des classes distributionnelles et des relations lexicales (syntaxiques ou sémantiques) est alors un point d'entrée privilégié. C'est par les rapprochements et les relations linguistiques que l'on repère les relations conceptuelles, et que l'on a ainsi des indicateurs pour savoir si on retient ou non un terme (et le concept). Donc il est important de commencer à mettre les termes en réseau, sans même les définir, pour d'abord mieux comprendre s'il est utile ou non de les conserver avant de rédiger une définition cohérente, consensuelle et validée de concept.

6.1.2.3 Dépouillement et fouille : vers une organisation des termes

Quelle que soit l'approche adoptée, l'analyse des termes se fait successivement et alternativement selon deux modes : « dépouillement » et « fouille » (Bourigault, Lépine, 1996). Le dépouillement consiste à balayer les résultats proposés par l'extracteur pour en dégager des données intéressantes. La fouille consiste à rechercher dans ces résultats des données précises pour décrire un concept ou en relation avec un terme, etc. Bien sûr, la part de dépouillement est plus importante dans l'approche « BCT » qui part des termes pour aller vers les concepts. Au fur et à mesure de la structuration d'un modèle, la part de fouille dans les résultats augmente au détriment d'un dépouillement.

L'analyse des candidats termes doit donc déboucher sur la définition de termes et de concepts, dont il faut déterminer une organisation sémantique à partir de leur organisation syntaxique (réseau terminologique) ou de leur seule étiquette. J'appelle *réseau conceptuel* cet ensemble de concepts qui devra être réorganisé avant de produire la ressource terminologique ou ontologique. Or, pour faciliter cette organisation, la lecture d'une liste de termes à plat est largement insuffisante. Les logiciels d'extraction de termes sont d'autant plus pertinents qu'ils fournissent des critères pour se focaliser en priorité sur les candidats qui ont le plus de chances d'être des termes d'une part, et qu'ils amorcent ce travail en suggérant des regroupements ou des liens entre termes d'autre part.

La focalisation sur les candidats les plus « importants » peut s'appuyer sur des différents types de critère :

- des critères linguistiques, comme ceux mis en évidence au paragraphe précédent,
- des critères numériques (fréquence et productivité),

- des éléments statistiques (xhi²) ou de distribution (présence dans un ou plusieurs sous-corpus, répartition mesurée par tf.idf[23], information mutuelle des mots composant les syntagmes)
- des éléments de forme (les syntagmes versus les mots simples, les nominalisations de verbes, etc.).

De plus en plus de logiciels, comme celui de (Gillam *et al.,* 2005) ou du projet Ontobasis (Reinberger *et al.,* 2004), se servent d'une implémentation de ces heuristiques pour filtrer automatiquement la liste des candidats ou pour en ordonner la présentation à l'utilisateur. Les logiciels utilisés, LEXTER et SYNTEX, présentent des critères numériques et laissent à l'utilisateur la tâche de les utiliser pour filtrer ou réorganiser les listes de candidats.

L'étape suivante de la plupart des approches de modélisation à partir de textes est de regrouper les termes. Certains de ces systèmes ont été présentés dans la typologie du 5.4.1.3. Cette étape est d'autant plus pertinente qu'elle s'appuie sur une liste de termes filtrés. Cependant, dans LEXTER et SYNTEX par exemple, ces regroupements s'opèrent sur les listes brutes. Une fois identifiées des classes, décider de les retenir et leur donner du sens est encore une tâche de plus haut niveau. Comme la sélection des termes, elle dépend de l'objectif d'utilisation de la ressource.

Au final, l'organisation des concepts doit prendre en compte les objectifs de la ressource à construire. Cette étape correspond à la *normalisation* des concepts de la méthodologie de construction de RTO (partie 6.2.2). Elle peut s'appuyer sur les propositions de regroupement que peuvent faire les logiciels, auxquelles l'analyste donne du sens. Elle justifie parfois de revenir vers les textes (ce qui correspond à une fouille des résultats).

Ce va et vient des textes vers les modèles et inversement est matérialisé sur la figure 6.1.2.3. Nous avons mis en évidence un ensemble de principes récapitulés dans la manière de parcourir les modèles et de recherche des informations dans les textes.

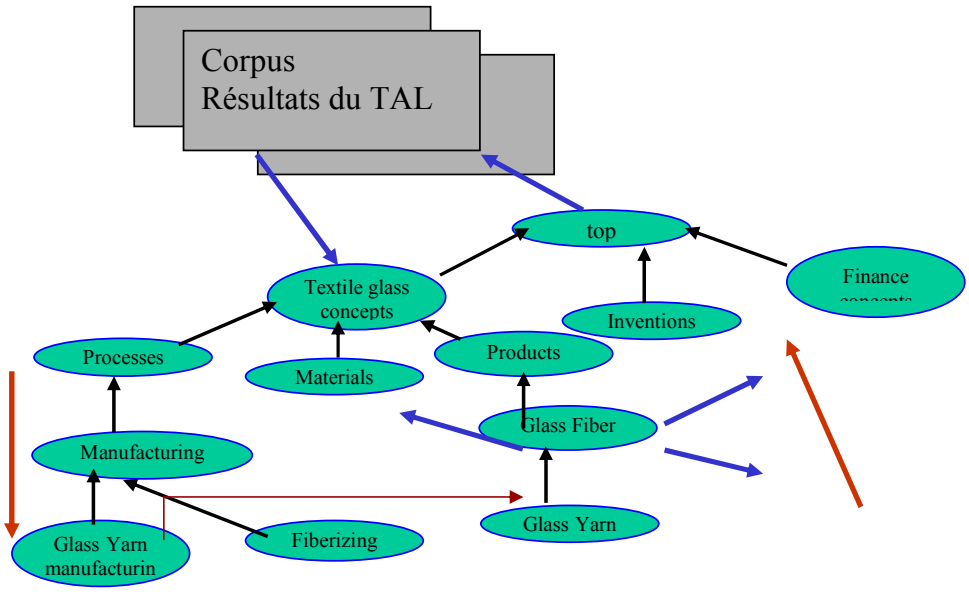

Figure 6.1.2.3 : Mise en évidence des différents types de parcours pour enrichir un modèle : L'analyse des textes peut se faire de manière ascendante, en fonction du contenu des textes, ou guidée par l'état courant du modèle à compléter. L'interprétation des informations obtenues mais aussi les choix de l'analyse pour compléter le modèle peuvent aussi correspondre à différents parcours au sein du modèle : processus d'abstraction, de généralisation ou de focalisation à partir d'un concept.

[23] Fréquence du terme dans le document divisée par la fréquence totale, coefficient utilisé en recherche d'information.

6.1.2.4 Expérimentations

Expériences d'utilisation de LEXTER

J'ai utilisé LEXTER dans le cadre de trois projets pour répondre à des besoins d'entreprises : SADE (Bourigault, Lépine, 1996), HYPERPLAN [EKAW, 96] et MOUGLIS [rapport-MOUGLIS, 98]. Je l'ai également testé sur une étude de cas, le projet Th(IC)² [OT-a, 00] [IC, 03]. Dans tous les cas, la taille des corpus était comprise entre 80 000 et 150 000 mots. De plus, l'objectif de la modélisation était bien identifié. En revanche, les caractéristiques de ces corpus étaient très différentes, ce qui a permis de diversifier les contextes d'évaluation : dossiers hétérogènes, d'auteurs et contenus différents pour Sade, documents regroupant plusieurs documents techniques très normalisés, écrits par quelques auteurs dans le cas de MOUGLIS et HYPERPLAN, enfin, collection d'articles scientifiques de deux périodes différentes dans le cas de Th(IC)².

Pour l'ingénieur des connaissances, le point de départ est bien sûr l'exploitation du réseau terminologique. Soulignons que ce réseau, de nature grammaticale, est éloigné d'un réseau conceptuel. Même si certaines relations de décomposition permettent de retrouver des concepts plus spécifiques ou génériques (par exemple RESEAU REGIONAL, RESEAU NATIONAL et RESEAU HT sont des concepts spécifiques de leur tête RESEAU), on ne peut pas interpréter systématiquement ces relations. Dans ces trois projets, la même approche a été adoptée pour conduire l'analyse du réseau terminologique en deux temps :

(1) une phase de sélection-validation a consisté à retenir, parmi les candidats, les termes considérés comme pertinents dans le domaine ;

(2) une phase de modélisation a porté sur la définition et la structuration de concepts à partir de l'étude fine de certains termes.

Dans le cadre D'HYPERPLAN, le corpus était constitué d'un seul document technique (un *Guide de planification des réseaux régionaux*). Le réseau terminologique obtenu comportait 13 000 candidats termes. La mise au point du modèle du domaine s'est appuyée sur l'étude directe de la tâche de planification décrite dans le guide et validée par des experts, puis sur l'interprétation du réseau terminologique et des résultats de ces trois modules. La tâche de planification a été modélisée sous forme d'un arbre de tâches élémentaires, chacune faisant appel à des concepts du modèle du domaine.

Dans le projet MOUGLIS, le corpus était également un guide de procédure, dans le domaine du génie logiciel (le *guide MOUGLIS*). Face à près de 6000 propositions de termes (résultat fourni par LEXTER) une analyse linguistique (manuelle) a permis de rejeter des termes sur des critères de forme ou sur des critères sémantiques. Ensuite, d'autres critères linguistiques (morpho-syntaxiques ou sémantiques) ont servi à retenir des termes [rapport-MOUGLIS, 98]. La liste initiale de LEXTER a été ainsi réduite de 75 %. Son analyse a permis d'acquérir ou d'ordonner des connaissances comme des variantes de termes (sigles, ellipses), des termes équivalents ou des relations hiérarchiques.

Bilan sur LEXTER

LEXTER fournit automatiquement des listes de termes reliés à leurs occurrences qui servent de point d'entrée pour élaborer une BCT ou un modèle du domaine. L'interface HTL constitue une aidé précieuse : elle permet une lecture du texte en fonction d'informations recherchées, et accélère la découverte et la compréhension du domaine ; elle met en évidence des informations sur l'utilisation des termes qui facilitent le repérage de termes synonymes, de concepts du domaine et de relations entre termes ; enfin, elle fournit un support riche pour conduire des entretiens avec les experts

Les modules proposés par H. Assadi sont peu productifs mais fournissent des résultats de qualité. La liste des termes typés fournis par le module de typage oriente vers la définition de

concepts ou de termes synonymes. L'analyse des propositions de structuration facilite la mise en relation des concepts. Enfin, les classes d'adjectifs servent à définir des concepts attributs.

Cependant, il est ressorti plusieurs difficultés de sa mise en œuvre : le coût de la préparation technique du corpus lorsque les documents ne sont pas sur support informatique (ce problème sera de moins en moins pénalisant avec la généralisation du support informatique) ; le volume des termes générés à partir du corpus (cette difficulté invite à diversifier les critères de consultation, ceux présents dans l'interface HTL étant déjà très pertinents).

Du point de vue de la construction d'ontologie, ce type de logiciel est d'un grand apport, et remis en question par la suite la phase de validation, inutilement coûteuse.

Exemple d'utilisation de SYNTEX

SYNTEX a permis d'analyser les termes de corpus dans le cadre de plusieurs projets terminés – IndexWeb [Rapport-INDEXWEB, 04], VERRE [Rapport-VERRE, 02] – ou en cours – Arkeotek [Rapport-ARKEOTEK, 04] –. Le bilan présenté ici s'appuie sur son utilisation dans le projet Verre, et en particulier son intérêt pour le repérage de synonymes. Je reviendrai sur son utilisation dans le projet IndexWeb dans la partie sur l'étude des relations, où les contextes verbaux trouvés par SYNTEX ont facilité le repérage de nouveaux marqueurs de relations sémantiques.

Dans ce projet, le corpus d'étude initial comportait trois sous-ensembles de documents : un livre technique et didactique, des dépôts de brevet et des dépêches économiques et financières se rapportant au domaine industriel concerné. Afin de mieux comparer les résultats, les trois sous-corpus ont été regroupés pour être traités ensemble de manière différenciée. Les résultats de SYNTEX ont présenté de grands écarts entre chaque sous-corpus, qui confirment l'hétérogénéité de leurs vocabulaires, très spécifiques. À partir de là, le corpus a été réduit au livre. SYNTEX offre en effet la possibilité très intéressante d'organiser le corpus en sous-corpus, et de les étudier chacun séparément comme des corpus à part entière, ou bien de les comparer (visualisation des termes spécifiques ou communs, des termes reliés au sein de chaque corpus ou globalement, etc.).

Ce projet a permis d'approfondir des axes particuliers de dépouillement des résultats de SYNTEX :

- l'intérêt d'un dépouillement préliminaire rapide pour ajuster le corpus ;
- l'utilisation des contextes verbaux comme indicateurs de relations lexicales (voir 5.4.3) ;
- la recherche systématique de relations des synonymie, récapitulée ici.

Plusieurs types d'études complémentaires, répertoriées en linguistique et en TAL, sont possibles pour la recherche de synonymies. Les unes recherchent des candidats synonymes (repérage de mots déviants, d'ellipses ou de sigles, de variantes de formes), les autres exploitent-vérifient-confirment ces hypothèses (utilisation de marques explicites « comme » « autrement dit » ; propagation des synonymies dans des groupes nominaux plus complexes). Les résultats de SYNTEX s'avèrent de très bons points de départ pour vérifier des synonymies :

1. étude de termes voisins : les termes appartenant à une même classe sont parfois des sous-classes d'une unique classe plus générique, parfois des synonymes, ou encore à regrouper en 2 ou 3 classes homogènes ;
2. étude des co-occurrents syntagmatiques de termes synonymes. Ainsi, *glass* et *glass fiber* ayant les mêmes termes co-occurrents, leur synonymie a été confirmée.

Premier bilan sur SYNTEX

L'exploration du réseau tête-expansion confère à SYNTEX les mêmes avantages que LEXTER, augmentés du fait que les verbes, les syntagmes verbaux et adjectivaux sont aussi pris en compte. Il s'avère que les rapprochements effectués par SYNTEX-UPERY sont extrêmement utiles et pertinents pour la construction de classes conceptuelles. Le nombre de rapprochements effectués dépend de la

redondance du corpus. Dans un corpus très régulier, contenant des descriptions répétées d'objets ou d'événements proches, et donc dans lesquels les mêmes structures syntaxiques reviennent régulièrement, SYNTEX identifie de nombreux rapprochements. A l'opposé, dans des textes où les redondances, répétitions, reformulations sont évitées, comme les textes de loi, les documents pédagogiques, les résultats sont faibles. Le phénomène est encore plus accentué dans un corpus de petite taille.

6.1.3 Étude des relations sémantiques : CAMELEON

Les expériences d'utilisation de COATIS dans le cadre de la modélisation conceptuelle (4.3.2) ont confirmé la complémentarité de l'étude des relations et de l'extraction de termes. Toutes deux contribuent à une lecture focalisée des textes qui facilitent l'accès rapide à des points importants. Les relations lexicales présentes en corpus sont de bons indicateurs pour repérer des termes et des relations sémantiques. L'étude des relations fournit un éclairage sur des zones de textes portant sur des concepts importants, sur des phrases suffisamment riches en connaissances pour amorcer une description de concepts, conforme à ce que les acteurs du domaine expriment dans les textes.

Par ailleurs, le centre d'études nucléaires (CEN) du CEA à Cadarache m'a proposé une collaboration pour définir une méthode et un logiciel qui permettent de renseigner plus rapidement des modèles conceptuels. Définis pour le système REX de gestion des connaissances, ces modèles organisent des termes et des concepts pour guider la rechercher d'informations au sein d'une base documentaire composée de fiches de retour d'expérience. Ces fiches, élaborées selon la méthode REX, sont consultées en formulant des requêtes qui sont étendues en exploitant les relations au sein de plusieurs modèles conceptuels qui organisent des concepts du domaine selon plusieurs points de vue. Dans ces réseaux, les relations ne sont pas étiquetées mais pondérées par des valeurs proportionnelles à la force du lien entre concepts.

C'est donc à la fois pour des raisons théoriques et pour répondre à la demande du CEA que je me suis focalisée sur la notion de relation sémantique et sur l'identification de ces relations à partir des relations lexicales en corpus. Ce travail a représenté une part importante de mes recherches depuis 1998. En particulier, il a été l'objet de la thèse de P. Séguéla (menée au CEA) que j'ai encadrée de 1997 à 2001. Grâce à une collaboration avec D. Bourigault et A. Condamines, un système d'aide au repérage de relations par analyse de corpus a été développé, s'appuyant sur une approche par patrons. La mise au point de ce système, CAMELEON, a induit une étude approfondie du processus qui va du repérage et de l'observation de relations lexicales à la définition de relations sémantiques.

6.1.3.1 Choix d'une approche par marqueur

Dans la continuité de la démarche suivie dans COATIS, pour définir une aide informatique à l'extraction de relations, le choix retenu a été de s'appuyer sur des travaux de linguistique. Alors que l'atout des approches statistiques est plutôt l'automatisation et l'identification des termes reliés, la référence linguistique présente la particularité de tenir compte de la sémantique des relations. Lorsqu'une relation lexicale est repérée, on connaît a priori sa sémantique. L'approche linguistique autorise également de traiter des corpus peu volumineux et moins homogènes.

Étudier la modélisation de relations conceptuelles à partir de l'observation de relations lexicales en corpus revient à distinguer deux sous-problèmes : (1) repérer et caractériser des relations en corpus (2) utiliser celles jugées stables et consensuelles à des fins de modélisation conceptuelle. Pour conduire chaque étape, des éléments méthodologiques et un support informatique, le logiciel CAMELEON, ont été fournis. L'approche retenue est une approche supervisée par marqueur.

Deux hypothèses sous-tendent cette approche : (1) la possibilité, à travers des éléments de forme, d'accéder à du contenu ; (2) le contenu identifié est une relation binaire entre concepts. Un

marqueur correspond à une *formule linguistique* dont l'interprétation définit régulièrement le même rapport de sens entre des termes. La *relation lexicale* renvoie au type de relation entre ces deux termes.

Le *marqueur* est un moyen de décrire précisément des fonctionnements lexicaux et de leur associer une interprétation sémantique systématique. La définition de patrons de fouille permettant de trouver ces marqueurs et ainsi des relations lexicales est l'objet de recherches en linguistique (Condamines, 2004). Les travaux les plus récents mettent l'accent sur la variabilité de ces marqueurs en fonction du type de corpus, du domaine et du type de la relation. Inversement, le marqueur caractérise la relation à laquelle il est associé et, en poussant à l'extrême cette idée, on peut dire que l'ensemble des marqueurs d'un type de relation est une manière très explicite de lui donner du sens, plus que sa seule étiquette. En effet, les marqueurs mettent en évidence tout ce qui est commun à toutes les occurrences de la relation.

Le *patron de recherche* d'un marqueur est une abstraction basée sur l'observation des marqueurs en corpus : il caractérise des comportements linguistiques réguliers en repérant des formes linguistiques faisant partie de catégories prédéfinies (grammaticales, lexicales, syntaxiques ou sémantiques). Les patrons lexico-syntaxiques sont composés de mots de la langue et de symboles renvoyant à des catégories grammaticales (SN pour Syntagme Nominal, Nom, Dét. pour Déterminant, Adj. pour Adjectif, etc.), à des classes sémantiques (i.e. *NomOutil* pour retrouver *logiciel, atelier, outil, plate-forme*, ...) ou, plus rarement, aux rôles grammaticaux (agent, patient, objet, …). Par exemple, la formule linguistique suivante est un des marqueurs de la relation d'hyponymie (SN2 EST-UN SN1) :

SN1 ET ADVERBE_DE SPECIFICATION SN2
où SN1 et SN2 sont des syntagmes nominaux, ET renvoie au mot "et", et ADVERBE_DE SPECIFICATION à une des expressions adverbiales suivantes : "plus précisément", "tout particulièrement", "plus particulièrement", "en particulier", "surtout", "principalement", "essentiellement", "notamment"

Ce patron permet de retrouver entre autres la phrase suivante :

On calcule la thermique de la partie haute du réacteur, et en particulier la température du toit.

Cette phrase révèle une relation d'hyponymie entre les syntagmes nominaux *thermique de la partie haute du réacteur* et *température du toit*.

L'étude des relations par les linguistes entre dans une perspective sémantique. Il s'agit de rendre compte de régularités de la langue, de proposer des méthodes pour trouver de nouveaux marqueurs et leurs patrons pour une nouvelle relation et de caractériser les corpus dans lesquels on les trouve. La pertinence des patrons est validée en fonction de leur efficacité pour retrouver des contextes pouvant contenir des relations d'un certain type. L'utilisation systématique de patrons de fouille ne peut se faire qu'à l'aide d'un logiciel qui guide leur définition et automatise au moins leur projection sur un corpus, comme par exemple un concordancier.

6.1.3.2 Des relations lexicales aux relations conceptuelles

En tant que mode déclaratif de représentation des connaissances, les réseaux sémantiques mettent l'accent sur la représentation des concepts ou des actions d'un domaine, plus que sur l'explicitation de l'algorithme qui permet de raisonner à l'aide de ces connaissances. De ce fait, ils sont alternativement considérés sous deux facettes. Tantôt ils sont vus comme une notation commode pour représenter certains types de connaissance et les manipuler ; tantôt ils sont utilisés comme un langage formel, prenant un sens précis et permettant une interprétation logique par un programme. Cette dichotomie des réseaux sémantiques est pertinente dans la perspective de modélisation à partir de texte. Il est commode de les utiliser comme de simples notations graphiques à vocation d'interprétation humaine dans un premier temps, les contraintes liées à la formalisation n'étant prises en compte que dans un deuxième temps.

La perspective est ici d'enrichir un modèle conceptuel de type réseau sémantique, non formel, à l'aide de *relations conceptuelles*, en tant que notations qui prennent du sens pour un

individu. La représentation formelle de ces relations n'est donc pas prise en charge dans CAMELEON et relèverait d'une étape de formalisation où il s'agirait d'en affiner la sémantique.

Un des résultats établis à l'aide de CAMELEON a été d'identifier quatre étapes significatives dans le processus qui va des observations en langue à un modèle formel (fig. 6.1.3.2) :

(a) caractériser les comportements linguistiques stables associés à des relations lexicales, et utiliser ensuite cette caractérisation pour repérer des types de relation spécifiques en corpus ;

(b) à partir de l'expression lexicale brute d'une relation particulière (d'une occurrence de relation lexicale), décider d'une relation ayant une étiquette donnée entre deux termes ; le fait de la nommer revient à la classer dans la typologie de relations ;

(c) passer ensuite du niveau linguistique au niveau conceptuel, en décidant du devenir de cette relation au sein d'un réseau sémantique en cours de construction ; cette relation contribue ainsi à la définition des concepts et à leur différenciation par rapport à d'autres ;

(d) représenter ces relations conceptuelles selon une représentation qui en garantisse une interprétation unique et non ambiguë par un système formel.

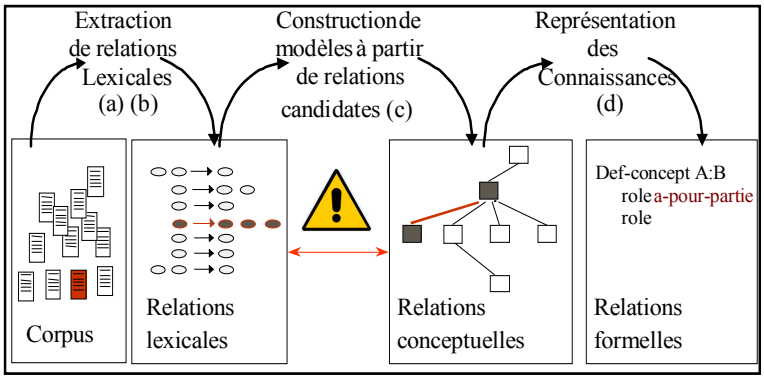

Figure 6.1.3.2 : Les relations sémantiques : du corpus au modèle conceptuel

L'étape (a) comporte en fait deux facettes : un travail purement linguistique conduit à définir les marqueurs à l'aide de primitives sémantiques, syntaxiques et lexicales ; les utiliser pour repérer des relations en corpus suppose ensuite l'opérationnalisation de ces marqueurs et l'utilisation d'un logiciel adapté. Or trouver la bonne forme informatique d'un marqueur est loin d'être trivial (Rebeyrolle, Tanguy, 2000). Sa qualité conditionne la capacité du marqueur à retrouver un type de relation dans un corpus.

6.1.3.3 Opérationnalisation des marqueurs

L'exploitation de marqueurs pour la recherche de relation en corpus revient à s'appuyer fortement sur l'hypothèse de régularité pour retrouver des connaissances, plus précisément des relations conceptuelles et les concepts reliés. La terminologie et l'informatique détournent ainsi les marqueurs de leur rôle initial pour en faire des outils de recherche de connaissances spécifiques et non pour rendre compte de l'expression de ces connaissances.

Dans la suite, la *formule linguistique*, le patron, est distinguée du *schéma informatique* qui correspond à son opérationnalisation. Comme le soulignent J. Rebeyrolle et L. Tanguy (2000) ou P. Séguéla (1999), l'opérationnalisation directe de patrons donne des schémas peu précis qui fournissent des résultats peu satisfaisants en corpus (en termes de rappel et de précision). Outre l'ajustement du patron au corpus, il faut trouver le schéma le plus efficace de ce patron. En effet, un même patron sera opérationnalisé différemment en fonction des capacités du logiciel utilisé et du corpus. Le type d'outil de traitement automatique des langues et la richesse du langage

informatique dont on se dote ont un impact significatif sur le type de recherche possible sur le corpus. Par exemple, l'analyse syntaxique préalable du corpus peut permettre de définir des marqueurs utilisant des indications syntaxiques.

Pour un langage et un outil donnés, la mise au point d'un patron sur un corpus consiste à fixer le schéma qui obtient les meilleurs résultats après évaluation et à décider du type de relation qu'il révèle. Projeter un patron revient ensuite à lancer un programme qui recherche toutes les occurrences du schéma dans le corpus, les enregistre ou les présente à l'utilisateur. La définition de patrons de fouille vraiment efficaces requiert des choix, méthodologiques et techniques, et représente un travail complexe associant informatique et linguistique (Rebeyrolle, Tanguy, 2000).

6.1.3.4 Un logiciel de modélisation de relations à partir de corpus : CAMELEON

Dans le cadre de sa thèse, P. Séguéla a développé le système CAMELEON qui s'appuie sur l'utilisation de marqueurs pour rechercher des relations lexicales en corpus puis guider leur intégration dans un modèle. L'analyste utilisant CAMELEON intervient pour ajuster les patrons de recherche des marqueurs, interpréter les relations lexicales et décider des relations conceptuelles à définir, en fonction du domaine et de l'usage prévu de la terminologie. Un patron est de la forme A X B Y C, X et Y étant les termes mis en relation, et A, B et C des caractérisations du contexte dans lequel apparaissent ces termes. CAMELEON suggère de suivre une méthode selon les quatre étapes définies plus haut, auxquelles le logiciel apporte un support :

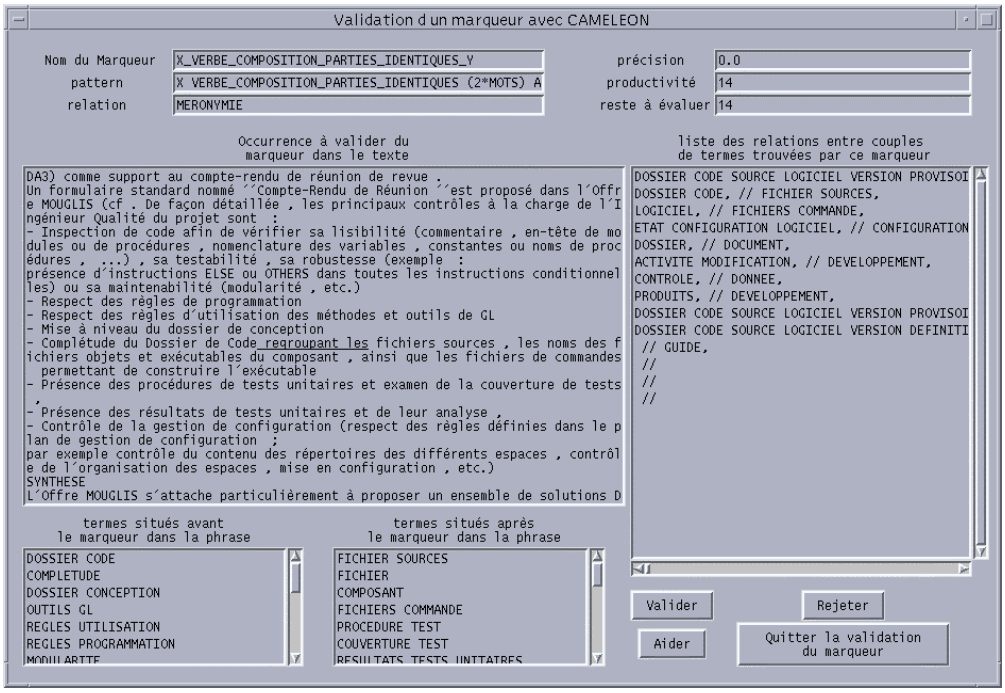

Figure 6.1.3.4 : Interface de validation de marqueur dans CAMELEON

- la mise au point d'une base de marqueurs et de types de relation adaptés à un corpus à partir d'une base de marqueurs génériques et de leur étude en corpus ; les marqueurs spécifiques sont soit totalement nouveaux soit adaptés des marqueurs génériques; cette mise au point passe par la projection des patrons sur le corpus puis l'observation des phrases retournées pour accepter, modifier ou rejeter ces marqueurs (fig. 6.1.3.4) ;

- leur projection sur le corpus pour localiser des hypothèses de relations lexicales entre termes ; ces hypothèses sont riches sémantiquement car elles comportent des propositions sur l'étiquette de la relation et sur les termes arguments, grâce à l'approche par marqueurs ;
- la sélection et l'interprétation en contexte des hypothèses de relation pour enrichir un modèle conceptuel au moyen d'une interface de modélisation ;
- la projection de couples de termes reliés pour l'identification de nouveaux patrons à partir de la lecture des contextes contenant ces couples de termes.

Les marqueurs génériques sont tirés de travaux de linguistique sur les relations d'hyperonymie et de méronymie, alors que les marqueurs spécifiques sont définis à l'aide de patrons trouvés par l'utilisateur à partir de l'étude des contextes de co-occurrence de termes en relation (fig. 6.1.3.5). En effet, la démarche prévoit de travailler sur des textes techniques, dans lesquels on s'attend à trouver des phénomènes linguistiques typiques : (a) des interactions spécifiques entre les objets du domaine qui se traduisent par des types de relation propres à ce corpus ; (b) des usages spécifiques en corpus, qui fixent le sens de termes polysémiques, et aussi qui permettent de trouver des marqueurs spécifiques de relations génériques. Ces phénomènes justifient la phase de mise au point de marqueurs.

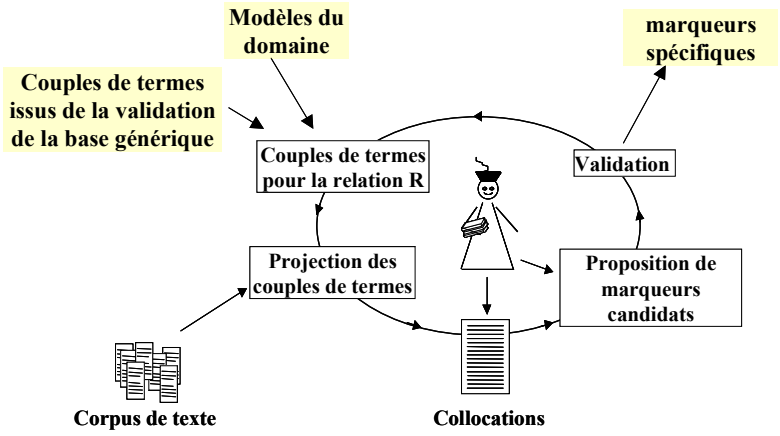

Figure 6.1.3.5 : Processus cyclique d'identification de marqueurs spécifiques

Pour faciliter la modélisation, CAMELEON permet ensuite de reprendre les hypothèses de relations lexicales en contexte grâce à une interface d'enrichissement de modèle conceptuel. Dans sa version initiale, ce module permettait de compléter des modèles de la méthode REX. Le passage de la relation lexicale à une relation conceptuelle concerne deux concepts au plus. Pour un des concepts concernés, on prend en compte sa place dans le modèle existant et les relations qu'il entretient déjà avec d'autres concepts. Une relation conceptuelle n'est ajoutée que si elle permet de mieux définir le concept de manière cohérente avec les concepts qui lui sont directement associés et avec l'objectif pour lequel le modèle est conçu. La personne chargée de la modélisation donne ainsi à l'indice linguistique une pertinence sémantique.

La décision d'intégrer une relation conceptuelle dans un modèle à partir de l'observation en corpus d'une relation lexicale ne peut être prise directement. La complexité de la décision est due autant à l'hétérogénéité du corpus qu'à la distance qui sépare le corpus de l'application visée. Pour bien comprendre la nature de cette décision, décomposons-la en fonction des contextes, parfois contradictoires, qui peuvent être pris en compte. En fait, il s'agit de filtres qui viennent réduire les chances de conserver une relation observée. P. Séguéla différencie (fig. 6.1.3.6) l'*énoncé*, c'est-à-dire la phrase ou le paragraphe englobant l'occurrence, le *texte* auquel elle appartient, et le *corpus* qui regroupe différents textes. Le texte est supposé homogène et cohérent, ce qui le démarque du

corpus, qui, lui, l'est moins. L'*application* vient se rajouter comme filtre pragmatique, combiné à la contrainte de conserver la *cohérence* du modèle.

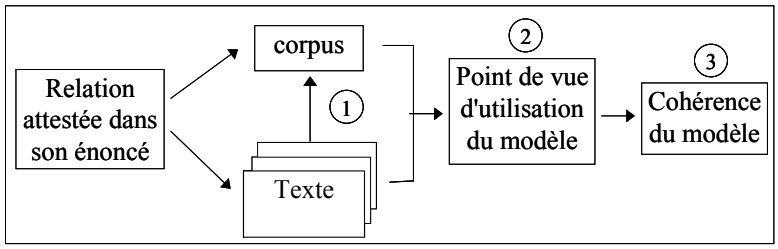

Figure 6.1.3.6 : Niveaux d'interprétation d'une hypothèse de relation

Pratiquement, pour décider de retenir ou non une relation conceptuelle à partir d'une proposition trouvée par projection de marqueur, CAMELEON oriente le cogniticien en lui fournissant le contexte d'interprétation. Ce contexte correspond (i) à l'état courant du modèle et, s'il existe déjà, à la place du concept dans la hiérarchie des concepts et (ii) aux phrases contenant la relation lexicale correspondant au marqueur. Le modèle est présenté de manière à appliquer le principe de différenciation homogène du concept (Bachimont, 1995). Le concept étudié doit posséder un trait commun avec son père, un autre avec ses frères et au moins un trait le différenciant de son père et d'autres de ses frères. Pour cela, le concept est présenté avec ses frères et son père selon le type de relation étudié dans l'interface principale du processus d'enrichissement (fig. 6.1.3.7).

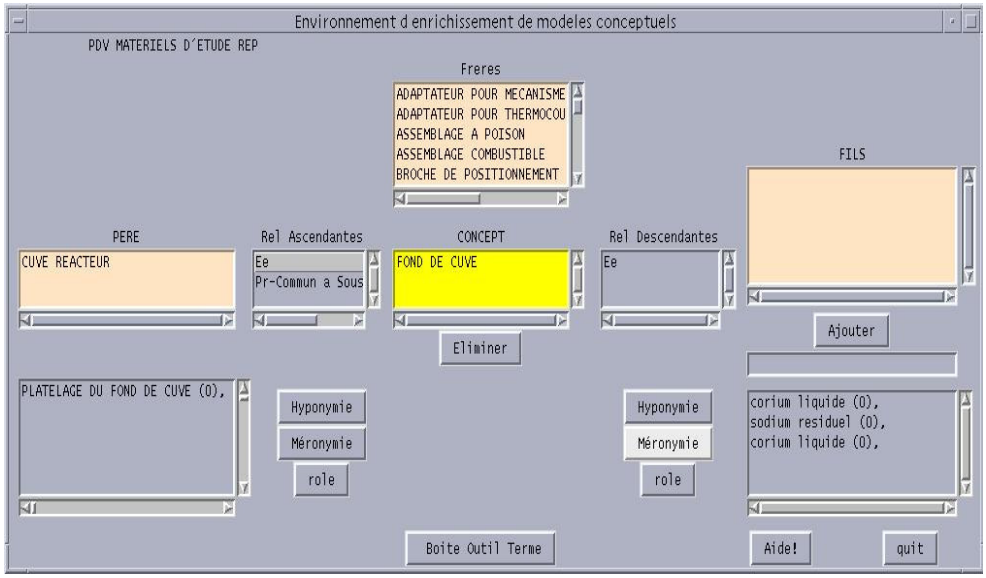

Figure 6.1.3.7 : CAMELEON : validation d'une relation dans le contexte du modèle

La fenêtre présente aussi d'autres concepts candidats à être des fils de ce concept pour cette relation : ce sont les hypothèses trouvées pour ce terme et cette relation par CAMELEON (liste en bas à droite sur la figure 6.1.3.7). Le cogniticien consulte une à une ces propositions pour les valider ou les rejeter, ce qui conduit à définir des relations ou de nouveaux concepts. Il poursuit ensuite la structuration du modèle à partir d'un autre concept, ou par l'étude d'un autre type de relation. Le modèle peut aussi être enrichi de concepts et de relations non trouvés par CAMELEON.

6.1.3.5 Contributions et évaluation

Une des originalités essentielles de CAMELEON est de couvrir le processus de modélisation depuis l'analyse de textes techniques jusqu'à l'intégration dans un modèle des connaissances tirées de ces textes. En effet, en 1998, les travaux de traitement automatique des langues et d'ingénierie des connaissances étaient fortement découplés. D'un côté, on trouve les systèmes d'extraction de termes ou d'aide à l'extraction de relation. D'un autre côté, on dispose de plates-formes ou de langages de modélisation de réseaux conceptuels. Or le fait de viser une application finalisée oblige à mieux intégrer ces deux composantes, à anticiper le fait que les résultats servent dans le cadre d'une modélisation dès l'analyse de textes. Les contributions de CAMELEON sont donc autant d'ordre fondamental que pratique :

1) Capitaliser des résultats en linguistique sur les relations d'hyperonymie et de méronymie dans une base de marqueurs génériques, applicables sur tout corpus.

2) Proposer un langage simple pour l'expression des marqueurs, à base d'expressions régulières.

3) Fournir un support continu au processus de repérage de relations conceptuelles à partir de la projection de patrons sur des textes

marqueurs trouvés sur le corpus technique du CEA	Précis	Prod.
`Y irradi(é|és|ée|ées) DANS X`	1	181
`X équip(é|ée|és|ées) (2*MOTS) DE Y`	1	134
`(stockages?|piégeages?|remplissages?|pénétrations?|introductions?) (1*MOTS) DE Y DANS X`	0,83	114
`(traces?|masses?) (3*MOTS) DE Y DANS X`	1	90
`(transit|passage|mise en place) (2*MOTS) DE Y DANS X`	1	84
`DE Y (stock(é|ée|és|ées)|piég(é...)| rempl(i|...)|introdui(t|...)) (2*MOTS) DANS X`	0,92	67
`Y ETRE (2*MOTS) (stock(é|ée|és|ées)|piég(é|... )|rempl(i...)|introdui(t|...)) (2*MOTS) DANS X`	0,95	45
`X (enrich(i|is|ie|ies)|contamin(é..)) (5*MOTS) en Y`	0,88	28
`(stocker|piéger|remplir|pénétrer|introduire) (2*MOTS) ARTICLE Y DANS X`	0,84	25
`(enrichissements?|contaminations?|teneur) DE X en Y`	0,8	23
`(transférer|décharger) (1*MOTS) ARTICLE_DEFINI Y DANS X`	1	4
`Y se dépose (3*MOTS) sur X`	1	2

Table 6.1.3.5 : marqueurs spécifiques au corpus du CEA. La précision indique le nombre de phrases valides trouvées par rapport au nombre de phrases retournées (productivité).

CAMELEON a été appliqué sur le corpus Rex, ensemble de fiches de retour d'expérience établies dans le cadre du projet SuperPhoenix au CEA. Ce premier projet a été le support à la définition de la base de marqueurs génériques, qui contient près de 190 marqueurs. Des marqueurs spécifiques ont été trouvés, comme ceux présentés dans la table 6.1.3.5 pour la relation de méronymie (composition). Leur pertinence a été évaluée en jugeant de la validité des phrases retournées.

Marqueur	Tech. oral	Tech. écrit	Textes de contraintes	Textes universitaires
Hyponymie base générique	2,9%	1,8%	6,8%	9,6%
Méronymie base générique	5,9%	1,8%	4,9%	2,2%
Hyponymie trouvés	0,6%			
Méronymie trouvés	1,1%			
Relations spécifiques trouvés	0,7%			

Table 6.1.3.6 : précision (en %) des types de marqueur sur les différents corpus du CEA

Le corpus du CEA étant composé de textes techniques de natures différentes, il a été découpé en sous-corpus, ce qui a permis de tester la pertinence des marqueurs en fonction du type de texte. Il en ressort une grande variabilité, chaque corpus contenant plus ou moins de relations hiérarchiques (table 6.1.3.6). On remarque le grand intérêt de définir des marqueurs spécifiques, qui permettent d'augmenter le nombre de relations pertinentes trouvées.

CAMELEON a fait l'objet d'une nouvelle expérimentation par Natalia Grabar sur un corpus de documents relatifs à des analyses économiques et écologiques liées à l'énergie nucléaire, projet SAFIR en lien avec EDF. Ce travail a montré l'intérêt de l'approche, en particulier de la possibilité de mettre au point de nouveaux marqueurs. Il a également fait ressortir la faible portabilité du logiciel, qui dépend de l'installation d'un gestionnaire de BD objet très lourd, Matisse.

Une autre restriction à l'utilisation de CAMELEON est le mauvais enchaînement des différentes tâches, l'utilisateur devant prendre en charge la mise à jour des fichiers de marqueurs à partir de la base générique puis lancer la projection avant d'utiliser l'interface de validation des hypothèses. L'intégration des relations dans un modèle conceptuel correspond à une autre application indépendante. Cette architecture correspond à celle d'un prototype dont chaque fonctionnalité a été mise au point séparément. Une fois validés les choix retenus, un environnement plus homogène, facilitant l'accès à ces fonctionnalités et aux différents résultats depuis une interface unique, serait préférable.

Les évaluations de CAMELEON ont montré que la notion de marqueur ou de patron de recherche peut être très large, et conduire à définir des formes très simples ou très élaborées. CAMELEON a été utilisé parfois pour rechercher deux mots co-occurrents dans un fenêtre donnée ou d'autres informations qui ne correspondent pas à des relations lexicales. Dans tous ces cas, un concordancier aurait été plus approprié. De plus, un concordancier présente une partie des fonctions requises pour la recherche de relations : la définition et la projection de patrons.

Enfin, CAMELEON a été défini pour le français. Or le nombre croissant de textes disponibles en anglais tout autant que la volonté de comparer ces travaux à d'autres au niveau international imposent de pouvoir travailler aussi sur des textes en anglais.

6.1.3.6 Modélisation de relations à partir de corpus étiquetés CAMELEONIII

Intégration d'un concordancier

Deux nouvelles versions de CAMELEON ont donc été développées en 2002 et 2003, chacune correspondant à une meilleure intégration des fonctionnalités recherchées et à une plus grande modularité. De plus, ces nouvelles versions ont bénéficié de notre collaboration avec les chercheurs de l'ERSS ayant développé le concordancier YAKWA (L. Tanguy) et l'extracteur de terme SYNTEX (D. Bourigault). Ces deux logiciels sont parvenus à une certaine maturité au moment où il semblait opportun d'améliorer CAMELEON. YAKWA est destiné à la mise au point de patrons de recherche, que ce soit pour identifier des relations lexicales ou tout autre schéma linguistique. Par contre, YAKWA ne permet pas de capitaliser les marqueurs dans une base, ne prévoit pas de noter la manière d'interpréter le résultat de recherches associées à un patron, etc. YAKWA est un outil de TAL plutôt conçu pour des linguistes, complètement dissocié de l'activité de modélisation ; il suppose une bonne connaissance de la grammaire ; il présente l'avantage de permettre de travailler sur différentes langues, toutes celles pour lesquelles on dispose d'un étiqueteur de textes.

Les nouveaux choix scientifiques et techniques retenus tirent donc parti des expériences acquises autour de ces 3 logiciels.

- analyser des corpus étiquetés par un analyseur syntaxique de manière à définir des marqueurs plus pertinents (comme YAKWA) ; ceci permet de définir des patrons plus précis au pouvoir de recherche élevé ;
- s'appuyer sur un concordancier parce que ce type de logiciel permet de définir des patrons de recherche, il réalise donc une partie des fonctions requises pour le repérage de relations (gestion des corpus, création et projection de patrons lexico-syntaxiques) ;
- ne pas refaire un concordancier et utiliser YAKWA puisque ce logiciel existe en licence libre ;
- reprendre les principes de CAMELEON en séparant la mise au point des marqueurs sur un corpus de leur utilisation pour enrichir un modèle conceptuel ;

Ces nouvelles versions reprennent donc exactement le scénario de CAMELEON et proposent deux types de fonctionnalité au sein d'un même environnement : la mise au point de marqueurs (évaluation de marqueurs génériques, mise au point de marqueurs spécifiques et détermination des types de relation associés) puis l'ajout de relations conceptuelles au sein d'un modèle. L'environnement de modélisation a été simplifié. La base de marqueurs génériques identifiés pour CAMELEON, adaptée au nouveau langage d'expression des marqueurs, sert toujours d'amorce à la définition des marqueurs d'un nouveau projet pour un corpus donné. Le logiciel suppose les corpus étiquetés à l'aide de Tree-tagger pour l'anglais et le français ou Cordial Université pour le français. Il est paramétrable en fonction de la langue du corpus. L'appel de l'étiqueteur YAKWA constitue le cœur de la mise au point et de l'évaluation des marqueurs et des patrons.

Dans toutes les versions de CAMELEON, il s'agit d'aider à repérer les termes en relation par le marqueur. CAMELEON supposait d'avoir recours aux résultats de l'extracteur de termes LEXTER, CAMELEONII utilise les résultats de SYNTEX. Dans CAMELEONIII, une autre option a été retenue, qui permet de s'affranchir d'un extracteur de terme : l'utilisateur peut spécifier la place des termes reliés dans le patron de fouille.

Logiciels développés

La première maquette, CAMELEONII, a été développée en juin 2002 en Python et en Perl. Elle s'appuie sur la même architecture client-serveur que YAKWA et utilise une base de données MySql pour gérer les données relatives à chaque projet (corpus, marqueurs, types de relation, termes, relations lexicales).

Une version mieux structurée et plus facilement portable, CAMELEONIII a été développée en 2003 en Java. Cette version s'est affranchie de la dépendance de SYNTEX : les termes en relation au sein d'une phrase retournée par un marqueur ne sont pas recoupés avec les sorties de SYNTEX. Ce choix assure au logiciel plus d'autonomie, et vient également du fait que les termes que l'analyste veut mettre en relation ne se trouvent pas forcément exactement dans la phrase contenant le marqueur. L'intégration de CAMELEON III et de TERMINAE comme environnement de modélisation est prévue. De plus, cette nouvelle version de CAMELEON sera accessible sur un serveur de l'IRIT au sein de la plate-forme RFIEC, afin de faciliter sa visibilité.

Aucune publication n'a encore mis en valeur ce travail récent. Il serait indispensable d'abord évaluer sur des exemples le gain apporté par cet enrichissement du logiciel par rapport à la version initiale CAMELEON. Un travail prévu début 2006 sera de porter tous les marqueurs identifiés par P. Séguéla ainsi que des marqueurs pour l'anglais dans le nouveau langage associé à CAMELEONIII. Une autre étude porte sur la spécification du cycle d'identification de nouveaux marqueurs par projection de termes dont on sait qu'ils sont en relation, ou même leur apprentissage à partir des contextes partagés par ces termes. Enfin, un dernier enjeux concerne la qualité et les performances du concordancier. Il tient une place clé dans le logiciel. Or YAKWA présente certains points faibles, dont celui de ne plus être maintenu. Un autre concordancier a été programmé courant 2005, pour mieux spécifier la position des termes en relation, au-delà de la fenêtre de la phrase. Il permet d'explorer des textes étiquetés de manière plus riche (sémantiquement par les termes trouvés par exemple, ou tel que l'analyseur de SYNTEX les étiquette).

Bilan sur l'extraction de relations

A titre de bilan, soulignons quelques particularités de CAMELEON par rapport aux logiciels de ce type. On assiste en effet à une multiplication des projets de ce type depuis 2002. Ces logiciels traitent en général différemment les relations hiérarchiques des autres relations : pour les premières, des techniques statistiques de regroupement en classes (clustering) ou des ressources externes (comme Wordnet) sont utilisées alors que les secondes sont souvent abordées à l'aide de patrons. Or *l'approche par patron a des limites reconnues par tous ses utilisateurs : elle est rarement très productive, et l'ajustement des patrons exige à la fois une certaine compétence linguistique et du temps*. La tendance actuelle étant à l'automatisation du processus de construction

d'un noyau d'ontologie à partir de textes, plusieurs solutions sont proposées pour rattraper ces inconvénients : soit les patrons ne sont pas ajustés, et seuls quelques patrons très robustes sont utilisés ; soit le système ne prend pas en charge le fait de donner un nom, une « signification » à la relation, ce qui autorise d'utiliser des marqueurs peu précis. Les patrons gérés par ces logiciels sont simples et sont supposés relier les groupes nominaux reconnus par un extracteur de termes et les plus proches autour du marqueur.

L'approche retenue dans CAMELEON, supervisée, suppose effectivement un travail approfondi de la part de l'analyste. *Un des points forts de CAMELEON est bien la base de marqueurs génériques* proposée comme point de départ, qui peut fournir un premier ensemble intéressant d'exemples de patrons pour comprendre comment en définir de nouveaux, et dont la projection retourne des occurrences tirées des textes, ce qui facilite leur ajustement au corpus. *La définition des marqueurs peut être riche et permet de préciser la place des termes en relation.* Ceci est un autre point fort de CAMELEON.

Pour terminer, il semble indispensable de mieux coupler ce logiciel avec d'autres logiciels d'analyse de textes d'une part, avec un véritable atelier de modélisation d'autre part. Ce couplage doit être un interfaçage modulaire, permettant d'utiliser ou non un extracteur de termes par exemple, de valider les hypothèses de relation trouvées à l'aide des marqueurs dans divers environnements de modélisation ou dans celui prévu par défaut. La programmation actuelle de CAMELEONIII autorise cette souplesse.

6.1.3.7 Apport de l'analyse distributionnelle et des approches statistiques

Deux expérimentations nous ont permis de proposer une approche pour l'identification de relations entre concepts à partir des résultats de l'analyse distributionnelle : le projet IndexWeb [rapport-INDEXWEB, 04] et Verre [Rapport-VERRE, 02]. Ces éléments méthodologiques encore empiriques confirment l'intérêt de considérer les verbes comme des candidats termes, dans la mesure où ils peuvent indiquer des relations propres à un domaine. Trois types de connaissance ont été identifiés à partir de l'étude des classes proposées par l'analyse distributionnelle d'UPERY et des contextes partagés : des relations entre termes, des marqueurs de relation et des indicateurs d'activité.

a- Étude des relations entre termes

L'objectif est de constituer des classes sémantiques, c'est-à-dire de regrouper les termes en familles ayant le même sens dans le corpus. Pour cela, on identifie les relations entre termes et on note les termes qui se comportent de manière identique dans ces relations. L'hypothèse est que des termes reliés à un même ensemble d'autres termes ont des chances d'appartenir à la même classe sémantique, d'avoir un sens proche. Deux types de relation ont été étudiés à l'aide des résultats de SYNTEX-UPERY, et ce pour les termes composés les plus fréquents : les relations de Voisin (au sens de l'analyse des distributions) les plus proches, les relations de compositions entre mots (relations Tête et Expansion). On peut ainsi aboutir à la constitution de deux types de classe :

- l'un part des couples de Noms (ex. : « satellite » + « plate-forme ») et aboutit à des groupes de Noms-Expansions rassemblés par les Têtes repérées (ex. : « bouée » et « balise » rassemblés par « DEPLACEMENT »),

- l'autre part des Têtes ainsi repérées (ex. : « fournir », lui-même décomposé en « fournir_OBJ », « fournir_SUJ » et « fournir par »), et aboutit à des classes de Noms-Expansions rassemblées par les couples de Noms de départ (ex. : « position », « bouée », « service », « localisation », « référence », « information », « niveau », rassemblés par « MESSAGE + DONNEE »).

Dans l'idéal, on peut approfondir la modélisation ainsi obtenue jusqu'à rassembler ces résultats en des structures du type « Classe de SUJETS – VERBE – Classes d'OBJETS » ou du type « Classe d'OBJETS – VERBE PASSIF – Classe de SUJETS-agents ». Mais cette démarche

suppose un travail long et précis avant d'aboutir à des regroupements de Sujets-Objets autour d'un même Verbe, du type : (Corpus CLS)

Sujet X		Objet Y
Système Argos Station Autorités	SURVEILLER	Volcan Océan Activité des bateaux Inclinométrie Niveau des fleuves Balises

Mise en forme : Puces et numéros

ou à des regroupements de Verbes-Objets autour d'un même Sujet, du type : (Corpus CEDOM)

sujets	verbes	objets
Système d'alarme	ALERTER	Vous
	PREVENIR	Centre de télésurveillance
		Personnes qualifiées
		Vous
	AVERTIR	Vous
	TRAITER	Incident

Mise en forme : Puces et numéros

b- Étude des marqueurs de relations

Pour enrichir ces classes sémantiques, on peut dégager des marqueurs de relations en combinant l'exploitation des résultats de SYNTEX (en particulier la notion de contexte partagé) et l'utilisation de YAKWA. Pour des syntagmes nominaux, ces contextes correspondent à des groupes nominaux ou verbaux pour lesquels plusieurs de ces syntagmes jouent le même rôle (complément, sujet ou objet). Par exemple, dans le corpus Verre, « Glasses » et « Polymer » partagent trois contextes dont un seul est verbal (V_Say_Prop) et correspond bien à un marqueur de relation hiérarchique.

Mis en forme

Mis en forme

Parmi les contextes trouvés, en général peu nombreux dans des corpus hétérogènes mais riches dans les corpus didactiques, tous ne correspondent pas à des relations entre concepts, mais cela est pertinent pour beaucoup d'entre eux. Le verbe pivot des contextes est un bon candidat à faire partie du marqueur. Lorsqu'un marqueur de relation peut ainsi être mis en forme à partir d'un verbe et des rôles des termes reliés, sa projection en corpus avec YAKWA permet de dégager d'autres couples de concepts reliés. L'observation des autres contextes ainsi trouvés peut conduire à affiner le marqueur.

Mis en forme

c- Utilisation des Verbes-supports et Structures-supports

Ces structures, quand elles introduisent des verbes importants du domaine (comme « déclencher », « dissuader » … du corpus CEDOM), sont du type « permettre de », « il vous suffit de », « de façon à », « est en mesure de », « chargé de ». Elles sont très représentatives des structures propres aux corpus de sites d'entreprises : elles servent à introduire des Prédicats (Noms ou Verbes) qui décrivent (avec leurs Arguments) les différentes actions propres au domaine étudié et aux techniques utilisées. Pour les repérer, une amorce classique est de repérer les verbes et les déverbaux d'action introduits par « permettre (de)» ou et « assurer ». Ensuite, la projection de ces déverbaux permet de repérer d'autres verbes introducteurs, et ainsi de suite. Ainsi, de nouveaux termes représentatifs peuvent être ajoutés à la terminologie.

Cette étude souligne l'importance de l'extraction des verbes pour l'analyse terminologique, en particulier pour repérer des actions et des activités.

6. 2 - Méthodes et plates-formes pour construire des ressources terminologiques et ontologiques

Dans la continuité naturelle de mon travail sur les BCT, j'ai donc abordé la construction à partir de textes de ressources du même type, comportant à la fois une dimension conceptuelle et une composante terminologique. Les terminologies et les ontologies en sont des cas particuliers. Dans ce cadre, le passage du texte au modèle suppose de s'appuyer sur un cadre méthodologique, au sein duquel sont définies l'utilisation des logiciels d'analyse de textes, mais aussi les tâches de l'analyste, la prise en compte des objectifs d'utilisation du modèle ainsi que les propriétés que doit vérifier le modèle. Dans une approche d'ingénierie des connaissances, la dimension méthodologique est fondamentale. Elle est indissociable des outils, logiciels et techniques choisis pour apporter un support au processus de modélisation.

TERMINAE, la proposition méthodologique pour les ontologies présentée dans cette partie, ne m'est pas propre. Elle est le fruit d'échanges avec les collègues du groupe TIA, en particulier sur les ontologies régionales de B. Bachimont, et surtout d'une collaboration étroite avec B. Biébow et S. Szulman du LIPN. Cette méthode intègre aussi notre expérience relative à la construction des BCT. Par rapport à GEDITERM, le logiciel TERMINAE, développé au LIPN, va au-delà de la structuration de concepts, jusqu'à la formalisation d'une ontologie. Par rapport aux travaux existants en 2000, TERMINAE vient répondre à un besoin non traité par les outils de construction d'ontologie classiques : la prise en compte de la dimension terminologique, de critères de normalisation ontologique ainsi que l'adaptation du modèle à une tâche particulière.

Cette partie rappelle donc les limites des méthodes et logiciels de construction d'ontologies existant en 1999 (§ 6.2.1) que l'on veut dépasser avec TERMINAE. Les principes qui fondent la méthode proposée sont ensuite présentés (§ 6.2.2), en insistant sur la part de l'interprétation humaine et des principes de structuration ontologique. Les étapes de la méthode sont alors détaillées (§ 6.2.3) ainsi que le logiciel TERMINAE (§ 6.2.4) en soulignant ma contribution dans l'ensemble du projet.

6.2.1 Contexte

6.2.1.1 Logiciels de construction d'ontologies

Initialement, le problème de la construction des ontologies a été abordé dans l'optique de favoriser leur réutilisation, et, étudié par des informaticiens, il a été ramené à un problème d'interopérabilité, de langage et de format d'échange. C'est ainsi qu'Ontolingua, historiquement le premier outil dédié à la construction et à l'échange d'ontologies, est orienté réutilisation, par fusion et extension, d'ontologies existantes disponibles dans une bibliothèque, et autorise l'exportation d'ontologies dans différents formats. Il permet à un utilisateur, ou groupe d'utilisateurs, de visualiser des ontologies existantes et de construire coopérativement de nouvelles ontologies. Il s'appuie sur un langage compatible avec le format d'échange KIF, ce qui est supposé assurer une facilité de réutilisation des ontologies, pour lequel il offre des interfaces de saisie et d'organisation des éléments de l'ontologie.

À partir de1995, dans ce même esprit, toute une lignée de langages et d'environnements assistent la création d'ontologies avec la même gamme de propositions : (i) un langage formel plus ou moins expressif et en général d'autant moins puissant pour réaliser des inférences qu'il est plus expressif ; (ii) des interfaces ou éditeurs pour écrire plus facilement des connaissances (concepts, relations, axiomes, etc..) selon le langage ; (iii) des moyens de vérifier l'organisation des connaissances au sein de l'ontologie : visualisation graphique du réseau conceptuel ou des concepts en hiérarchie ; gestion des listes par type d'objets, etc. ; (iv) des moyens pour vérifier formellement la définition de l'ontologie conformément à la sémantique du langage : par classification de concepts, par vérification de l'unicité des concepts ou de la complétude du modèle.

La plupart de ces éditeurs sont accessibles publiquement sur le web et non commercialisés. Ces logiciels prennent peu en compte la dimension terminologique, l'association d'un lexique au réseau conceptuel qui constitue l'ontologie. Ces outils se préoccupent rarement de la manière de trouver, identifier, décrire les éléments qui doivent former le modèle à construire. La seule aide apportée peut être d'utiliser la structure de l'ontologie construite pour guider la définition d'instances de concepts, comme le propose le logiciel CUE (Van Eijst, 1995). Dans cette première génération d'outils, on trouve OntoSaurus (Swartout *et al.*, 1997) pour le langage Ontolingua, WebOnto (Domingue, 1998) au dessus du langage OCML ou OilEd pour les langages OIL puis DAML+OIL (Bechhofer *et al.,* 2001).

Pour un état de l'art exhaustif, voir (Gómez-Pérez *et al.,* 2004), un ouvrage qui intègre un inventaire des outils et méthodes pour la construction d'ontologies conçus dans le cadre de plusieurs projets européens (OntoWeb, Espéronto, MKBEEM).

En se focalisant sur la réutilisation, ces outils semblent mettre de côté les vrais problèmes relatifs à la construction d'une ontologie. Avant de savoir comment les formaliser, le cogniticien doit d'abord trouver les bons concepts à conserver dans l'ontologie et leurs propriétés ou relations, et cela à partir des entretiens avec les experts et utilisateurs, ou, de manière complémentaire, par l'étude de documents. Il doit pouvoir vérifier qu'ils ont du sens et établissent un consensus auprès des utilisateurs ou des experts du domaine. Enfin, il doit pouvoir juger de la bonne adéquation de cette ontologie à l'application visée. Or le choix des concepts et leur description s'effectue au niveau des connaissances. La formalisation n'est ici d'aucune aide.

Dans cet esprit, des environnements plus larges, intégrant une série d'outils, ont vu le jour à partir de 2000. Ces environnements sont modulaires et extensibles, le module central étant toujours un éditeur d'ontologie associé à plusieurs langages de représentation. Leurs composants traitent de problèmes particuliers comme la représentation des axiomes, l'alignement ou la fusion d'ontologies, ou l'apprentissage de nouvelles connaissances. Protégé-2000 (Fridman-Noy *et al*., 2000) est ainsi devenu l'environnement de modélisation d'ontologies le plus utilisé. L'intérêt de ce logiciel est que le noyau, un éditeur d'ontologie, est défini au sein d'une architecture extensible grâce à des plugg-in qui viennent enrichir les aides offertes à l'utilisateur. WebODE (Arpiréz *et al.,* 2003) et OntoEdit (Sure *et al.,* 2002) reposent sur des principes analogues.

6.2.1.2 Méthodes pour la construction d'ontologies

Les méthodes de construction d'ontologie prévoient effectivement de prendre en charge la spécification de ce modèle avant sa formalisation. Il serait ambitieux de dresser un inventaire des méthodes de construction d'ontologies disponibles avant 2000. Le lecteur peut consulter l'ouvrage de synthèse de (Gómez-Pérez *et al.,* 2004). Peu d'auteurs fournissent des repères pour mener la tâche d'identification et de structuration des concepts. De plus, très peu de ces méthodes utilisent les textes comme source de connaissances privilégiée.

La méthode de Ushold et King (1995) reprise par Fox et Grüninger (1996) évalue l'impact d'une démarche ascendante (partant des instances ou classes spécifiques pour les abstraire), descendante (spécialisation de classes génériques) ou combinant les deux approches sur la nature du modèle obtenu. Les travaux plus récents, comme la méthode On-To-Knowledge (Staab *et al.,* 2001), en ont repris l'idée de formuler des questions de compétence à partir des scénarios d'usage de l'ontologie. Ces questions doivent aider à repérer les concepts clés utiles pour l'objectif visé, leurs propriétés et les contraintes sur ces propriétés.

La méthode METHONTOLOGY s'intéresse aux ontologies au niveau conceptuel (Gómez-Pérez *et al.*, 1998). Cette méthode distingue les tâches nécessaires à l'organisation de l'ontologie en fonction des éléments de modèle à identifier (concepts, relations hiérarchiques, autres relations binaires puis instances et règles). Le point de départ est l'inventaire des termes du domaine et leur définition. Un peu comme Protégé, l'environnement associé, WebODE, offre une interface présentant les concepts et leurs propriétés sous forme de tables, puis permettant de saisir des

instances de concepts selon ce modèle. La méthode met l'accent sur la qualité syntaxique du modèle construit plus que sur sa pertinence sémantique par rapport à l'usage prévu de l'ontologie.

À côté de ces approches privilégiant soit les objectifs d'utilisation, soit la forme de l'ontologie, des travaux théoriques issus de la philosophie se sont interrogés plus fondamentalement sur ce que devait être leur contenu et comment le déterminer. La méthode OntoClean (Guarino et Welty, 2001), même si elle est présentée comme une méthode de validation, peut s'appliquer aussi au moment de la construction d'une ontologie. Elle fournit des propriétés que doivent vérifier les concepts et leurs relations, et des contraintes sur ces propriétés et leur présence ou non sur des concepts reliés. Ainsi, G. Kassel a repris les principes d'OntoClean pour les appliquer en phase de construction d'ontologie. La méthode OntoSpec qu'il propose avec S. Bruaux (Bruaux *et al*., 2005) s'appuie sur la vérification de ces méta-propriétés pour guider pas à pas la définition de concepts. Selon ce courant, le haut niveau des ontologies génériques pouvant être réutilisé, une aide à la construction est justement d'affiner et valider les choix ontologiques de plus haut niveau. Ces deux méthodes s'appuient sur la top Ontology DOLCE (Masolo *et al.,* 2003).

6.2.1.3 De l'interprétation à la normalisation

L'analyse des résultats des logiciels de TAL comme les extracteurs de termes et de relations correspond à un processus *d'interprétation* et non d'extraction ou de mise au jour de connaissances qui seraient présentes telles quelles dans les textes. Cette interprétation se justifie par le statut des termes (et de leurs occurrences) par rapport aux concepts autant que par la nécessité d'adopter un point de vue local et lié aux usages pour décrire des concepts. Pratiquement, le dépouillement des résultats des extracteurs de termes ne peut prétendre conduire à une représentation « neutre » des connaissances contenues dans les textes. De plus, il ne fait pas uniquement appel à une compétence sur la langue. Ce dépouillement est indissociable d'une interprétation qui prend en compte le contexte d'usage du modèle et l'état courant de son contenu. Il relève d'une *normalisation*, au sens où le concept, représenté en dehors du contexte où il est identifié, est une réduction (normée) de ce qui est observé.

L'étude des termes peut conduire à définir des éléments de modèle conceptuel à plusieurs niveaux : de nouveaux termes désignant des concepts, de nouveaux concepts, des relations entre concepts (hiérarchiques ou autres), des propriétés de concepts ou encore des instances de concepts, des axiomes ou des contraintes sur les relations. L'analyste prend en compte l'état courant du modèle, la représentation des connaissances et l'utilisation prévue du modèle. Expliciter les critères d'organisation retenus a pour but d'éviter que la manière de représenter l'information interprétée ne relève d'un choix arbitraire. Ces critères précisent pourquoi chaque concept et relation a été défini. Ils fixent un engagement ontologique qui assure une cohérence globale au modèle construit. Pour N. Guarino, cet engagement se traduit par une fonction d'interprétation des termes du domaine. Pour nous, il s'agit juste de fixer un point de vue.

L'héritage historique de travaux en philosophie sur l'ontologie puis en sciences de la vie sur les classifications et les hiérarchies a permis de dégager des principes de structuration. Ces principes visent à expliciter la manière dont une réalité peut être « conceptualisée », sur quels aspects de cette réalité on va se focaliser pour que les concepts définis soient pertinents dans l'application visée (Studer, 1998). Ces principes guident également la manière de définir des propriétés, de définir et de formaliser les concepts (Guarino, 1998). Deux groupes de travaux de la communauté internationale se sont traduits par des propositions pratiques : ceux de N. Guarino, relevant de l'ontologie formelle (Guarino, 1995), sur les méta-propriétés des relations conceptuelles (Guarino et Welty, 2001) et ceux de B. Bachimont sur les ontologies différentielles (Bachimont, 2004).

Je me place dans la lignée des propositions de B. Bachimont, qui ont fortement marqué les travaux des membres du groupe TIA et en particulier la méthode TERMINAE. B. Bachimont se situe dans le cadre d'une sémantique différentielle au niveau des termes (Rastier, 1997), qu'il propose de traduire par une différenciation explicite au niveau conceptuel. Chaque unité d'un modèle est alors

décrite par ce qui la rapproche et ce qui la différencie d'autres unités voisines. Ces analyses, qui remontent à Aristote et Porphyre, débouchent sur l'organisation (hiérarchique) des genres et des espèces pour décrire des catégories d'objets du monde. B. Bachimont a repris ces principes et en fournit une lecture critique (Bachimont, 2004). Il en tire des propositions méthodologiques pour guider l'interprétation de termes et la définition des concepts d'une ontologie, qui constitue le cœur de la méthode ARCHONTE (Bachimont, 2004). Cette méthode propose en particulier une phase importante de normalisation basée sur ces principes différentiels, suivie d'une formalisation et d'une opérationnalisation. La normalisation comporte deux facettes successives :

- une organisation globale des concepts sous forme arborescente, évitant les treillis et l'héritage multiple au sein du réseau sémantique intensionnel ; les concepts sont « définis » en intension ;
- une structuration locale basée sur les principes différentiels, tout concept rajouté devant posséder des points communs et des différences explicites avec son père et ses frères.

Pratiquement, la plate-forme d'aide à la construction d'ontologies DOE a été définie pour mettre en œuvre ces principes (Troncy et Isaac, 2001).

6.2.1.4 Sémantique textuelle, sémantique différentielle : nécessité de l'interprétation

Analyser le contenu de textes, c'est tenter d'accéder au sens de ces textes. Cela suppose une analyse sémantique. Or les approches retenues et les logiciels utilisés peuvent renvoyer à différentes sémantiques : référentielle, grammaticale, lexicale, … En effet, chacun propose un type différent d'analyse de corpus. Ainsi, la sémantique lexicale repose sur l'étude des usages des mots et du statut des termes, en faisant l'hypothèse de redondances significatives. Étudier l'analyse de textes suppose de se situer par rapport aux différentes sémantiques proposées en linguistique, en IA et en traitement automatique des langues. Cette question, fondamentale du point de vue de la linguistique de corpus, est largement étudiée par A. Condamines (Condamines, 2003). Les courants qui retiennent le corpus comme objet d'étude correspondent à une sémantique textuelle. Ils englobent entre autres la sociolinguistique, la récente terminologie textuelle, l'analyse de discours et la sémantique interprétative. Dans ce cadre, les termes ne sont pas étudiés dans leur rapport aux objets ou concepts qu'ils peuvent désigner ou référencer, mais essentiellement à travers les liens qu'ils entretiennent entre eux dans les corpus étudiés. L'interprétation de leurs différents usages tente de déboucher sur des régularités révélatrices d'un ou plusieurs sens. A. Condamines qualifie donc cette sémantique de textuelle et de doublement située, les sens attribués aux termes étant influencés autant par la situation de production des textes que par celle de leur interprétation.

Ainsi, un concept issu de l'interprétation de l'usage de termes et présent dans le modèle conceptuel est avant tout une construction (artificielle) et non une représentation qui serait fidèle, complète ou exhaustive d'une réalité matérielle ou cognitive. Or dans le cadre de la modélisation conceptuelle, rendre compte de ce sens (et des liens entre termes observés en corpus) revient à définir des concepts, et soulève à nouveau la nécessité de dégager la sémantique. Parce que le processus de normalisation retenu correspond à la différenciation dans la méthode Archonte (Bachimont, 2004), la définition des concepts relève d'une sémantique différentielle. Mais cette différentiation n'est pas mise en avant du côté de la langue et de l'usage des termes, même s'ils en sont le point de départ, elle est conceptuelle et fait suite à une interprétation humaine.

6.2.2 Modélisation d'ontologies à partir de textes : une méthode

6.2.2.1 La méthode TERMINAE : principes et originalité

La méthode TERMINAE a été définie au LIPN sous une première forme en 1999 (Biébow, Szulman, 1999) (Biébow, Szulman, 2000). Cette méthode repose sur des principes tout à fait analogues à ceux retenus pour la construction de BCT avec GEDITERM. Elle met en oeuvre les

principes élaborés par le groupe TIA. Grâce aux expériences relatives à GEDITERM mais aussi à la mise au point de CAMELEON, j'ai acquis une compétence et des propositions méthodologiques précises sur les phases initiales de la construction de modèles à partir de textes : construction de corpus, exploitation des résultats d'outils d'analyse de textes et surtout prise en compte de la dimension terminologique des modèles. De manière tout à fait complémentaire, les points forts de TERMINAE se situent dans les phases suivantes de structuration des concepts et de leur formalisation logique. Afin de réunir ces points forts au sein d'une même plate-forme et d'une méthode unique, nous avons collaboré avec B. Biébow et S. Szulman à partir de 2000, la méthode et le logiciel TERMINAE étant le support de ces nouveaux résultats méthodologiques et logiciels.

Le premier fruit de ce travail a été une nouvelle version de TERMINAE (méthode et outil) en tant qu'aide à la construction d'ontologies présentée dans [EKAW, 00]. Dans ce cadre, les fiches de modélisation ont été remplacées par un réseau conceptuel semi-formel. Une deuxième série de modifications, en particulier la définition de fiches terminologiques, a permis d'adapter la méthode et le logiciel à la construction de terminologies [TAL, 02]. Le logiciel a fait l'objet d'évolutions plus récentes auxquelles je n'ai pas participé : exportation des résultats en OWL (Szulman et Biébow, 2004) et intégration de SYNOTERM, un logiciel de calcul de synonymies (Hamon, 2000).

Par rapport aux méthodes d'Ingénierie des ontologies répertoriées par exemple dans (Gómez-Pérez *et al.,* 2004), TERMINAE présente plusieurs originalités :

- **Partir des textes comme sources de connaissances** : Ils constituent un support tangible, rassemblant des connaissances stabilisées qui servent de référence. Leur utilisation améliore la qualité du modèle final. Cette méthode a été une des pionnières en la matière. Ce type d'approche se trouve aujourd'hui largement développé dans différents projets, en particulier pour préparer l'annotation automatique de documents à l'aide des concepts identifiés : projets AKT[24] (Ciravegna *et al.,* 2002), OntoLearn (Navigli *et al.,* 2004) ou encore Kaon (Maedche *et al.,* 2000).
- **Enrichir le modèle d'une composante linguistique** : L'accès aux termes et textes qui justifient la définition des concepts garantit une meilleure compréhension du modèle. Ce type de résultat est encore très novateur. Encore aujourd'hui, malgré l'intérêt de ce type de choix, mis en évidence pas exemple dans (Maedche, 2002), le standard OWL ne permet pas de définir les termes comme des entités à part entière. Les propositions faites dans ce sens sont toujours timides. Par exemple, dans le modèle de données du projet Crossmarc (Paliouras, 2004), les termes sont réduits à un ensemble de variantes, ou encore, dans la proposition d'évaluation d'ontologie du projet PASCAL, une confusion gênante est maintenue entre termes et concepts.
- **utiliser des techniques et outils de TAL basés sur des travaux linguistiques** : ces outils permettent l'exploitation systématique des textes et leurs résultats facilitent la modélisation. Ces outils sont utilisés de manière supervisée, et leurs résultats consultés au sein d'interfaces qui en facilitent l'exploitation. Il ne s'agit pas d'automatiser le processus, dans la mesure où l'interprétation humaine est fondamentale pour décider de la manière d'identifier et organiser concepts et relations.

Les paragraphes qui suivent reprennent la méthode telle qu'elle est décrite dans [EKAW, 00] et (Biébow, 2004), puis précisent les aspects qui ont évolué suite à ma collaboration avec le LIPN.

[24] Http://www.aktors.org

6.2.2.2 La méthode TERMINAE : étapes

Des textes à un modèle formel

Je présente d'abord ici une vision axée sur les objets mis en jeu. La méthode part des textes constituant la documentation technique pour aboutir à une modélisation formelle du domaine. C'est l'étude des contextes d'apparition des mots et de leurs relations de dépendances syntaxiques qui guide la conceptualisation. La méthode distingue les termes des concepts et les relations lexicales des relations sémantiques. Les termes et les relations lexicales correspondent à des syntagmes présents dans le corpus et considérés comme caractéristiques du domaine. Ils sont identifiés en appliquant des méthodes et outils linguistiques afin d'aider l'utilisateur (outils d'extraction de termes, de relations, recherche de synonymes, regroupements de termes pour former des "classes" conceptuelles ...). Les regroupements lexicaux rassemblent des syntagmes apparaissant dans des contextes analogues. Les syntagmes sont interprétés en contexte local (la phrase ou le paragraphe) puis global (le texte ou le corpus). Lorsqu'ils sont attestés, ils donnent lieu à la création de concepts et relations sémantiques, dont ils sont les étiquettes. L'ensemble des concepts et relations forme un réseau conceptuel, non formel mais compréhensible par le concepteur. Les concepts et relations étant extraits du corpus et contraints par l'application, ce réseau forme une ontologie régionale au sens de (Bachimont, 1995).

Ensuite, dans le modèle formel, concepts et relations sont formalisés dans un langage terminologique assimilable à une logique de descriptions, sous forme de concepts et de rôles organisés en hiérarchie. Les concepts sont caractérisés selon deux dimensions, l'une linguistique exprimant s'ils correspondent ou non à un syntagme du corpus, l'autre de structuration indiquant la motivation ayant conduit à les intégrer dans le modèle formel.

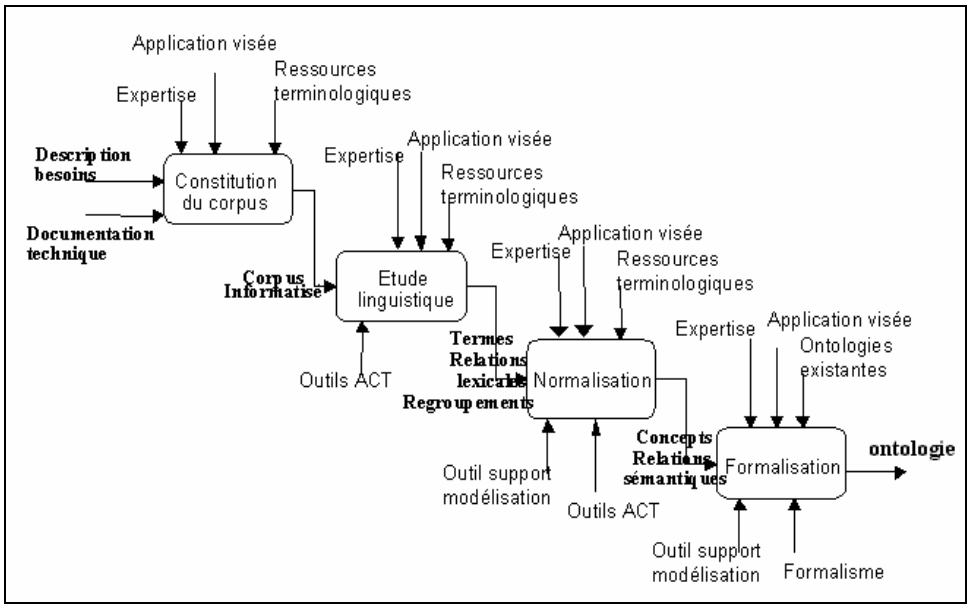

Fig. 6.2.2.2 : Étapes du processus de modélisation à partir de textes selon TERMINAE

Les quatre principales étapes de la méthode peuvent être présentées selon un enchaînement linéaire comme sur la figure 6.2.2.2. Cet enchaînement correspond bien au changement de nature des données manipulées. Il reflète aussi la vision plus simple retenue pour la construction des BCT, où l'étude linguistique aurait pu avoir une finalité en elle-même, sans lien avec la normalisation. Toutefois, ce schéma ne doit pas être interprété comme purement séquentiel, ce qui serait réducteur sur la complexité de la mise en œuvre réelle de la méthode. Implicitement, le déroulement de

TERMINAE correspond plus à un processus cyclique. En particulier, comme l'ont confirmé toutes les analyses citées jusque-là, l'étude linguistique et la normalisation sont étroitement imbriquées.

Constitution du corpus

À partir de la description des besoins et des objectifs de développement du modèle, le cogniticien choisit dans la documentation technique à sa disposition les textes à inclure dans le corpus. Il peut s'agir de textes didactiques, de spécifications techniques, de normes, de comptes rendus d'expériences, d'articles scientifiques … Le corpus doit couvrir complètement le domaine requis par l'application. Le choix nécessite une expertise des textes du domaine afin de caractériser leur type et la couverture du domaine. Un glossaire sur le domaine est utile pour déterminer les sous-domaines à explorer et vérifier qu'ils sont tous couverts. Le corpus est ensuite mis sur support informatique s'il ne l'était pas. Le début de la modélisation, en particulier le dépouillement rapide des résultats des logiciels de TAL, peut conduire à revoir le contenu du corpus, à le réorganiser pour traiter séparément certaines parties, à le compléter pour combler des lacunes ou à éliminer des textes qui s'avèrent peu adaptés.

Utilisation d'outils de TAL

L'étude linguistique, menée à l'aide d'outils de TAL, cherche à déterminer les termes et les relations lexicales qui seront éventuellement modélisés. La méthode TERMINAE recommande d'utiliser des extracteurs de candidats termes, des extracteurs de relations, des concordanciers et des outils de regroupement conceptuel. Elle met en avant les limites à l'utilisation de ces différents types d'outils de manière indépendante : le dépouillement de leurs résultats, souvent très volumineux, est fastidieux et difficile à organiser. Au contraire, l'utilisation conjointe de différents logiciels permettrait de disposer rapidement de plusieurs éléments d'information sur les termes et leur usage, de mieux les interpréter, de s'orienter vers des concepts importants du domaine. Une autre difficulté à recommander des logiciels est la compétence linguistique requise pour les mettre en œuvre. Par exemple, les extracteurs de relation selon une approche par patrons lexico-syntaxique nécessitent de bonnes compétences grammaticales et un savoir faire informatique.

C'est pour cela que la méthode oriente vers le choix d'outils simples, auxquels sont associés des aides à la consultation et à la validation des résultats, comme les extracteurs de termes LEXTER et SYNTEX, ou des ressources, comme la base de marqueurs dans CAMELEON. C'est aussi pour cette raison que le logiciel support TERMINAE propose ce type d'interface de navigation dans les résultats de ces extracteurs et d'un extracteur de relations intégré, Linguae.

Normalisation

La normalisation consiste en deux parties : la première reste dans le domaine du traitement lexical et exploite les données retenues par l'étape antérieure ; la seconde partie porte sur l'interprétation sémantique et la structuration des concepts et des relations sémantiques. Au cours de la normalisation, la masse de données à considérer est peu à peu restreinte.

Les termes et les relations lexicales déterminés à l'aide des outils précédents sont associés à leurs occurrences dans le corpus. Parmi l'ensemble des termes et relations lexicales, le cogniticien choisit ceux dont il va poursuivre l'analyse. Ce sont les termes qui à la fois ont du sens en corpus et qui présentent un intérêt par rapport aux objectifs du modèle. Puis, il étudie chaque syntagme d'après ses contextes d'occurrence afin d'en donner une définition en langage naturel qui rende compte du contenu des textes. En cas de polysémie, il décide quels sens parmi ceux présents dans le corpus sont à retenir car pertinents pour la modélisation.

La deuxième étape de la normalisation consiste à définir des concepts et des relations sémantiques à partir des termes et des relations lexicales précédentes. Le cogniticien doit en donner une description normalisée, reprenant les étiquettes de concepts et de relations déjà définies, et pertinente par rapport à la tâche pour laquelle le modèle est construit. L'interprétation de la description est contrainte par le corpus dont elle est issue et l'application. Ces descriptions amorcent une structuration du domaine sous forme de réseau conceptuel, non formel.

En pratique, ces deux étapes sont étroitement mêlées, l'utilisateur organisant le plus tôt possible quelques concepts dans l'ontologie. Ces concepts sont dits centraux [TIA, 02], d'abord parce qu'ils correspondent aux termes semblant les plus pertinents, ensuite parce que l'ontologie va être élaborée à partir d'eux. La modélisation va du texte à l'ontologie, mais la définition d'un concept et son insertion dans l'ontologie renvoie l'utilisateur à étudier l'usage de termes proches et donc aux résultats de l'analyse linguistique. L'analyse des termes proches du concept étudié peut conduire à définir des concepts qui lui sont reliés. Ainsi, l'utilisateur "tire le fil" d'un concept et élabore les autres autour de lui, à partir des relations non hiérarchiques dans un mouvement transversal ou en suivant les relations hiérarchiques dans un mouvement vertical. Peu à peu, plusieurs réseaux indépendants peuvent être dégagés, qui se rejoignent pour former l'ontologie.

Formalisation

La formalisation comprend l'élaboration et la validation de la base de connaissances. Des ontologies existantes, générales ou proches du domaine, ou même un glossaire, peuvent faciliter la découpe des couches hautes de la base de connaissances, c'est à dire les plus générales, en larges sous-domaines. Ensuite, le cogniticien traduit les concepts et relations sémantiques provenant de l'étude linguistique en concepts formels et rôles dans le langage de la base de connaissances, puis il les insère dans le modèle. Cette insertion des concepts et rôles terminologiques nécessite parfois une remise en question de la structure existante, car elle doit prendre en compte la correction de l'héritage des caractéristiques (rôles) des concepts formels. Le cogniticien doit souvent rajouter des concepts pour améliorer la structuration de la base, des concepts de structuration. Lors de l'insertion d'un nouveau concept, l'outil support effectue une vérification locale, qui garantit la correction syntaxique de la description ajoutée. Une validation complète du modèle doit être réalisée lorsque la base atteint un état stable, pour vérifier la cohérence du modèle.

6.2.2.3 Contributions à TERMINAE

La question des corpus

Ces travaux, bénéficiant d'une réflexion commune avec A. Condamines, ont permis d'affiner ce que recouvrait un corpus dans le contexte de la construction de modèles conceptuels. Je reprend donc la définition qu'elle propose (Condamines, 2004) :

« Un corpus est une collection de textes (éventuellement un seul texte) constituée à partir de critères linguistiques ou extra-linguistiques pour évaluer une hypothèse linguistique ou répondre à un besoin applicatif. »

Cette définition met en avant le besoin d'expliciter des caractéristiques (comme la taille, le sujet, les auteurs, la diversité ou l'homogénéité des documents le composant, la langue utilisée, son niveau de correction, l'étendue du domaine couvert, le genre des documents, etc.) qui délimitent ce que peut être un corpus pour un projet et une méthode donnés. En effet, les moyens choisis pour l'analyser autant que la finalité de l'analyse orientent le choix des documents formant un corpus « pertinent ». La démarche sera d'autant mieux reproduite ou adaptée à un autre contexte que l'on aura explicité les caractéristiques qui permettent de décider de cette validité. Cette définition pose ainsi la constitution du corpus comme un tâche à part entière dans une démarche comme TERMINAE, et souligne l'ajustement nécessaire entre corpus et outils d'analyse.

Dans TERMINAE, parce que l'on s'intéresse à des modèles pour des tâches et des domaines, spécifiques, les corpus sont choisis avec les interlocuteurs du domaine concernés. Ils peuvent être de petite taille puisque l'on n'utilise pas de méthode statistique. Les résultats des logiciels d'analyse comme SYNTEX et CAMELEON seront plus ou moins pertinents suivant la nature du corpus. Par exemple, CAMELEON produit plus de résultats sur des textes pédagogiques, écrits avec une langue grammaticalement correcte. SYNTEX requiert des textes contenant des régularités, des

formes répétées et un peu de redondance. Lorsque le corpus ne peut être modifié, il peut donc être nécessaire de chercher des approches alternatives et d'autres logiciels.

Cette analyse se démarque de certaines approches actuelles appliquant les techniques d'apprentissage automatique sur les textes. La plupart de ces travaux considèrent le corpus comme donné, imposé, auquel il faut s'adapter pour en tirer le plus d'information possible. Dans cette perspective, le Web peut être un corpus. Cependant, sans disposer d'un *a priori* sur l'information recherchée ou sur les textes composant le corpus, il n'est pas réaliste de dégager d'un ensemble très hétérogène des régularités et d'apprendre à partir de là des éléments conceptuels. Donc ces travaux se rapprochent plus d'une problématique d'extraction d'information : le système recherche des informations prédéfinies, dont on a établi les caractéristiques par une première analyse du corpus ; de plus, finalement, seul un sous-ensemble du corpus est exploité.

Axes d'analyse des textes

Le processus cyclique qui unit les étapes d'analyse linguistique et la normalisation correspond pratiquement à des allers-retours entre texte et modèle [TIA, 03]. Pour mieux guider la mise en œuvre de la méthode, j'ai explicité la progression de la construction du modèle selon deux directions :

- *un axe texte-modèle*, qui permet de rendre compte de deux types de tâche : des tâches de *dépouillement* qui vont du texte au modèle, l'enrichissement du modèle étant alors orienté par les données (éléments de textes ou résultats d'analyse de textes) ; des tâches de *fouille* au sein des textes ou des résultats d'analyse, qui correspondent à une recherche ciblée pour affiner ou compléter des parties spécifiques du modèle.
- *des axes de parcours au sein du modèle* : l'organisation de concepts dans le modèle peut alternativement être menée de manière *ascendante* (trouver des concepts pères des concepts existants par regroupement et abstraction d'éléments spécifiques, d'exemples, d'instances, etc.), de manière *descendante* (définir des concepts fils par spécialisation, décomposition ou raffinement des concepts existants) ou encore centrifuge (étudier toutes les relations concernant un concept donné qui devient le centre de l'étude).

Ces dimensions d'analyse soulignent la diversité des tâches effectuées autant que leur complexité, due aux divergences ou au manque de précisions des sources de connaissances et des objectifs de modélisation.

Questions pour guider la normalisation des concepts

Ces expériences de construction d'ontologies, en particulier le projet VERRE, ont permis de dégager des éléments méthodologiques pour mener à bien les deux facettes de la normalisation : la structuration des concepts puis l'application des principes de différentiation [TIA, 03] et [rapport-VERRE, 02].

1. *Repérage de concepts centraux et étude des termes associés* : ce repérage peut s'appuyer sur des critères statistiques (répartition …) ou numériques (fréquence, productivité …), structurels et grammaticaux, sur la richesse des contextes d'apparition ; l'étude des relations de synonymies et des variantes fait partie de l'étude des termes.

2. *Organisation hiérarchique* : il s'agit d'organiser des hiérarchies locales autour des concepts identifiés, en cherchant des concepts plus spécifiques (fils) ou plus génériques (pères) des concepts centraux ; le réseau tête-expansion de SYNTEX peut ici être précieux, suggérant des regroupements de termes en classes ou des relations générique-spécifique entre termes composés ; CAMELEON permet de confirmer ou trouver ces relations par la projection de patrons propres à la relation de hiérarchie.

3. *Étude des autres types de relation* associés à chaque concept : je propose d'appuyer cette étape sur l'étude des verbes (qui peuvent indiquer des relations), des termes reconnus comme voisins, l'analyse des distributions et le réseau tête-expansion de SYNTEX, ainsi que l'interprétation

(manuelle) des séquences associées aux termes d'une part, et d'appliquer l'approche par patron de CAMELEON en cherchant des marqueurs de relations propres au corpus.

4. *Enregistrement des résultats* dans TERMINAE.

Il s'avère que les trois premières étapes de cette progression dans l'organisation de concepts sont désormais classiques dans la plupart des systèmes « d'ontology learning » (Cimiano, 2004) et (Reinberger, 2004). A l'issue des tâches précédentes, le modèle est souvent composé de sous-ensembles non homogènes, pas toujours reliés entre eux, redondants ou incomplets. La normalisation consiste à vérifier le modèle en fonction de critères ontologiques, syntaxiques ou de connaissances du domaine. De nouvelles tâches visent alors à justifier que chaque élément est nécessaire au sein de l'ontologies, qu'il est pertinent à cet endroit et défini conformément à l'objet de modélisation. Plusieurs points sont à contrôler : (a) unicité de définition ; (b) homogénéité de point de vue ; (c) cohérence des descriptions. TERMINAE suggère les critères de différenciation de concepts de la méthode Archonte pour atteindre ces objectifs. Pour chaque concept, on doit expliciter sous forme de commentaires ou de rôles les points communs et les différences entre ce concept et son père puis entre ce concept et ses frères.**Modélisation d'ontologies à partir de textes : plate-forme de modélisation TERMINAE**

Le logiciel TERMINAE a été développé dès 1998 par B. Biébow et S. Szulman au LIPN. Cette partie présente les derniers développements introduits dans TERMINAE suite à ma coopération avec l'équipe du LIPN [TAL, 02]. TERMINAE a évolué afin de prendre en compte les résultats établis au sujet des BCT et mes expériences de construction de terminologies. Cette nouvelle version du logiciel a permis de décrire des terminologies aussi bien que des ontologies. Suite à mes travaux sur l'étude des relations sémantiques à partir de patrons (CAMELEON), TERMINAE intègre des outils linguistiques en offrant quelques mécanismes simples de traitement de corpus (LINGUAE) et de dépouillement de résultats d'extracteurs de termes (LEXTER et SYNTEX). Enfin, la structuration des concepts est proposée selon des modalités moins contraignantes, dans un réseau conceptuel. Il s'agit là d'un des intérêts de TERMINAE par rapport aux éditeurs qui se limitent à des interfaces de saisie en amont de langages de représentation formelle d'ontologies comme OilEd ou OntoEdit.

Les utilisateurs ciblés, à savoir cogniticien, linguiste-terminologue ou informaticien, peuvent utiliser différemment cette plate-forme en fonction de leurs compétences et objectifs. Selon son profil, l'utilisateur profitera d'avantage des fonctionnalités d'analyse des textes, de structuration conceptuelle ou encore de formalisation. Il pourra également faire appel à des ressources autres que les textes (experts du domaine et terminologies ou ontologies existantes) pour compléter ou valider les informations dégagées des textes et interprétées comme des connaissances. Le modèle ciblé sera plus ou moins structuré et plus ou moins proche du contenu des textes (Biébow, 2004).

6.2.3.1 Des textes aux modèles

Par rapport à d'autres environnements de construction d'ontologies, TERMINAE a été une des premières plates-formes à assurer un support à la modélisation à partir de textes. Une des originalités de TERMINAE est de maintenir une continuité des textes vers les modèles et inversement.

Dans un premier temps, l'utilisateur se place dans un environnement d'étude terminologique, qui lui permet de consulter le corpus et surtout les résultats d'un des extracteurs de termes SYNTEX ou LEXTER (Fig. 5.5.3.1). Il peut les filtrer et les valider. Des fonctionnalités d'exploration du corpus à partir de patrons de relations sémantiques sont disponibles dans le module Linguae pour contribuer à repérer des concepts et des relations entre concepts. Enfin, depuis peu, l'outil SYNOTERM (Hamon, 2000) peut être utilisé depuis TERMINAE pour aider au repérage de relations.

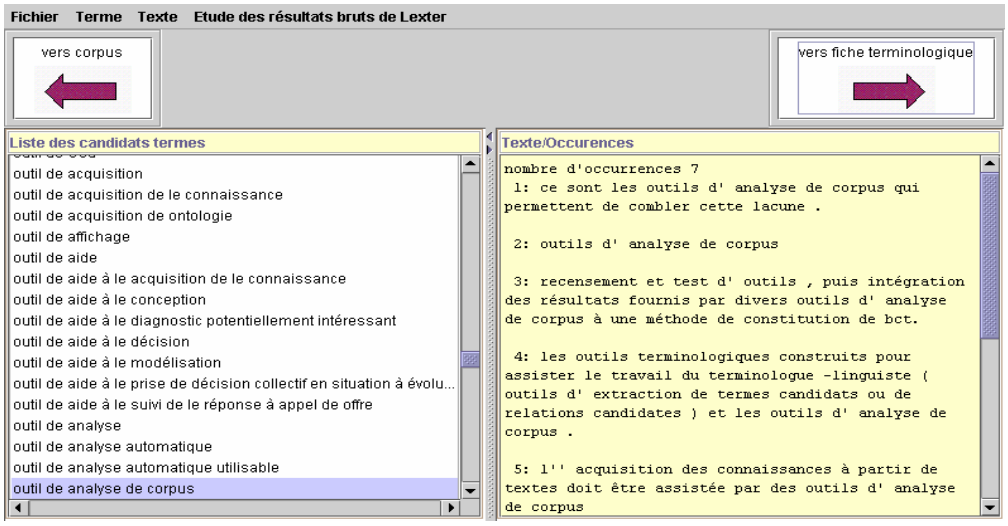

Figure 5.5.3.1 : Fenêtre d'étude des termes, en amont de la création de fiches terminologiques.

Ensuite, le travail de conceptualisation se poursuit dans un environnement d'aide à la modélisation dont le support essentiel est la fiche terminologique en lien avec l'interface du modèle conceptuel. Le format de cette fiche est décrit dans la partie suivante. Pour chaque terme du domaine, il y a autant de fiches que de concepts associés, chacun correspondant à un des sens de ce terme. À partir de cette fiche, en sélectionnant le concept, l'utilisateur accède à l'interface du modèle conceptuel qui lui présente la description du concept dans l'état courant de l'ontologie.

Pour terminer, le réseau conceptuel de TERMINAE permet d'ajuster la définition des concepts et de leurs relations sans respecter d'emblée toutes les exigences syntaxiques ou sémantiques de la représentation des connaissances. Ce modèle peut être considéré comme un résultat à part entière (livré dans des fichiers XML) ou bien comme une étape avant la formalisation. La formalisation se place comme une dernière étape qui permet de vérifier la syntaxe du modèle et en assure la livraison selon un langage plus formel, comme OWL (Szulman et Biébow, 2004).

L'utilisation de ces fonctionnalités est décrite dans la contribution présentée au workshop EON 2002. La fiche terminologique et l'interface de gestion du réseau conceptuel sont présentés dans la suite.

6.2.3.2 Originalité du modèle de données : une ontologie à composante terminologique

Fiches terminologiques

La notion de fiche terminologique est présente dans TERMINAE dès sa première version (Biébow et Szulman, 1999). Un mot ou un syntagme ne prend le statut de terme qu'à partir de la création de sa fiche. La fiche permet de distinguer les différents sens d'un terme dans ses diverses occurrences en corpus. Grâce à l'expérience acquise avec GEDITERM et la construction de BCT, j'ai proposé des évolutions de cette fiche pour l'adapter à la construction de terminologies. Une terminologie rassemble alors un ensemble de fiches terminologiques et le réseau des concepts associés à ces termes.

Les fiches terminologiques ont été enrichies de manière à pouvoir décrire différents types d'informations grammaticales, linguistiques, normatives ou d'usage qui sont associées à chaque terme (fig. 5.5.3.2). C'est l'utilisateur qui paramètre le logiciel pour décider des rubriques retenues (langue, sigle, …). La fiche comporte également des données classiques pour ce genre de fiche : des informations sur la création et les mises à jour, sur l'avancement de la validation. Enfin, la fiche prépare la définition du ou des concepts associés à ce terme. Pour cela, elle regroupe les

occurrences du termes regroupées en fonction des différents sens du terme, et donc des concepts associés. Pour chacun de ces concepts, elle contient aussi une définition en langage naturel, des termes synonymes et des termes proches.

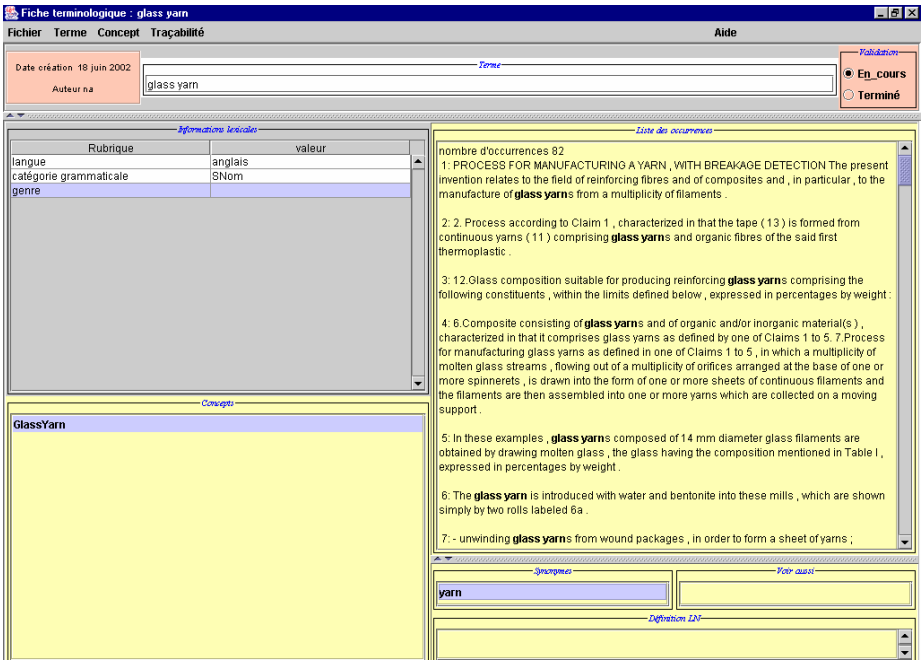

Figure 5.5.3.2 : Fiche Terminologique dans TERMINAE

Pour la construction d'une ontologie, ces fiches présentent l'avantage d'assurer une traçabilité des textes vers les termes (via les occurrences) puis vers les concepts et le modèle. Elles permettent de consigner le travail de définition du concept, et de mieux justifier les choix relatifs à l'organisation de l'ontologie.

Réseau conceptuel

La première version de TERMINAE proposait une aide à la formalisation via des fiches de modélisation ainsi qu'un noyau de base de connaissances (concepts et relations génériques) qui pouvait être enrichi pour construire une ontologie particulière. À partir de ces fiches, tout nouveau concept était formalisé pour être intégré dans l'ontologie après vérification de sa syntaxe et de sa cohérence avec les concepts existants. Ce type de vérification alourdit significativement la mise au point du modèle et ne convient que si l'on est sûr des concepts à définir. Ce n'est pas l'esprit de la phase de normalisation, au cours de laquelle le modèle est modifié et réorganisé à plusieurs reprises, tenant lieu de « brouillon » avant l'ontologie formelle. Ces fiches ont donc été abandonnées au profit d'une structure plus souple, moins contraignante et ne conduisant pas immédiatement à une représentation formelle : le réseau conceptuel (fig. 5.5.3.3).

Dans le même esprit que le réseau conceptuel des BCT, ce réseau permet de définir des concepts « localement », sans devoir vérifier leur cohérence globale. Ils peuvent par exemple être rapidement définis à partir de leur nom, et sous un niveau élevé de l'ontologie. Cependant, tout nouveau concept doit être situé dans la hiérarchie des concepts et, en ce sens, il n'est pas totalement indépendant des autres. Ce réseau n'a de sens que pour l'interprétation humaine, mais il peut facilement être traduit vers le langage formel de TERMINAE, afin de le rendre « calculé » par le système, validé formellement. Les principes de représentation des concepts dans GEDITERM ont donc introduits dans TERMINAE.

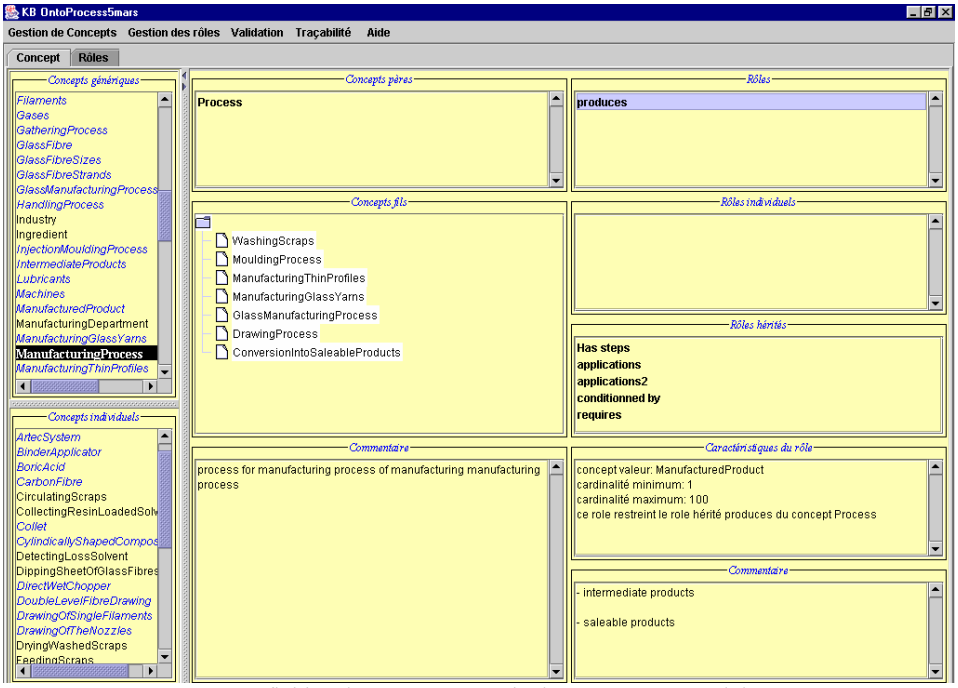

Figure 5.5.3.3 : Définition d'un concept au sein du réseau conceptuel de TERMINAE.

Ce choix vise deux objectifs : (i) mieux marquer la différence de l'analyse au niveau linguistique du niveau conceptuel ; (ii) élargir la gamme des utilisations de TERMINAE à la construction de terminologies, le réseau et des fiches terminologiques forment des BCT.

Le niveau conceptuel prépare la formalisation sans en avoir les contraintes. Il permet de traiter des questions comme : faut-il représenter une information sous forme de concept, d'instance ou de relation ? s'il s'agit d'un concept, est-il primitif ou défini ? comment le définir ? s'il s'agit d'une relation, à quels concepts de la hiérarchie l'associer ?

Après plusieurs expérimentations, en particulier le projet Verre, il ressort que ce réseau conceptuel présente toutes les caractéristiques d'une ontologie qui ne serait ni formelle ni validée, et ne comporterait ni axiome ni règle. La représentation des connaissances choisie dans TERMINAE ne permet pas de représenter simplement des axiomes, et donc ce type de connaissances est complètement mis de côté. Le réseau conceptuel a ainsi une structure identique à une ontologie au sens de TERMINAE, et peut en être considéré comme une version préliminaire de celle-ci. La validation syntaxique est lancée à la demande de l'utilisateur lors de la formalisation.

6.2.4 Expérimentations et validations

6.2.4.1 Utilisation pour la construction de terminologie

Les premières mises en oeuvre de la nouvelle version de TERMINAE correspondent à la construction de thésaurus structurés selon un modèle conceptuel : le projet Th(IC)² portait sur le domaine de l'ingénierie des connaissances (IC), et nous avons réduit l'étude à celle des logiciels utilisés en IC [IC, 00] et [EKAW, 00] ; une étude de cas simple dans le domaine de la fabrication de la fibre de verre visait l'applicabilité de l'approche à ce domaine [TAL, 02].

Projet Th(IC)²

Le groupe TIA a voulu confronter ses propositions théoriques, techniques et méthodologiques dans le cadre d'un projet, Th(IC)². L'objectif de ce projet est l'élaboration d'un thesaurus en français du domaine de l'ingénierie des connaissances pour permettre d'indexer les pages Web des chercheurs. Au sein de ce projet, avec le LIPN et l'ERSS, nous avons décidé de construire une ontologie des outils de l'IC, qui permette de situer de nouveaux outils par rapport à ceux qui existent déjà. Finalement, le résultat obtenu est plus proche d'une terminologie.

Le corpus, initialement constitué d'articles scientifiques, a été enrichi de documents présentant le domaine pour disposer de plus de définitions et de connaissances explicites sur les concepts de base du domaine. L'analyse des textes a été menée avec LEXTER pour le repérage de termes et CAMELEON pour l'identification de relations entre concepts. TERMINAE a servi à valider et visualiser les résultats de LEXTER puis à construire le réseau conceptuel.

Pour tirer des connaissances des résultats extraits, la démarche privilégiée est centrée sur les termes et non sur les relations lexicales : étude des noms propres (noms de logiciels), des syntagmes commençant par « outil de » ou « logiciel de ». L'étude des termes voisins des noms propres conduit à rechercher également « système », « projet », « algorithme », « méthode » comme pouvant introduire des noms d'outils. L'étude des relations avec CAMELEON s'est focalisée sur les relations hiérarchiques, à base de marqueurs standard. La hiérarchie ainsi obtenue décrit une cinquantaine de concepts, regroupés en deux classes : Outils Conceptuels et Outils Logiciels.

C'est au cours de cette expérience qu'est ressortie la place toute relative des textes, qui ne sont qu'un des éléments de décision du cogniticien. Plusieurs critères sont pris en compte conjointement pour retenir les connaissances à modéliser : (1) notre expertise sur le domaine ; (2) l'application visée ; (3) les textes (les contextes des informations, les auteurs et la nature des textes) ; (4) d'autres informations apportées par les outils linguistiques : termes co-occurrents, termes en relations lexicales, fréquence de certains termes.

Projet verre-français

Une deuxième terminologie a été construite à partir d'un texte portant sur la fabrication du verre. Ce texte court (3 pages, 1 425 mots) est extrait d'un manuel pédagogique destiné à des élèves du secondaire. Le manuel a été rédigé par un expert d'une entreprise verrière pour présenter les bases technologiques de l'industrie et des applications du verre. L'étude du texte sert de point de départ pour organiser en une terminologie structurée des connaissances sur la nature des matières vitreuses et sur la fabrication du verre.

LEXTER et le module LINGUAE de TERMINAE ont été utilisés pour l'analyse de ce corpus. Plus que la couverture du domaine, le modèle doit refléter une organisation précise et justifiée des concepts. L'étude visait à mettre en évidence des éléments méthodologiques propres à l'application prévue des données. Les relations sémantiques étudiées sont celles de composition « compose » ou « entre dans la composition de » (qui relie *verre* et *oxydes*) ou la synonymie marquée par *appelé* dans le corpus.

La phase de normalisation a permis de mieux définir certains concepts, comme les concepts de *corps liquide* (comme le *verre*) par rapport à un *corps solide* (comme le *cristal*). Elle a conduit à introduire des concepts ne correspondant pas à des termes (comme *processus chimique* pour regrouper *fusion, solidification*). Enfin, elle incite à associer des termes très différents qui correspondent au même concept (comme *viscosité* et *processus de solidification*).

Ces deux expériences font ressortir l'intérêt d'une modélisation conceptuelle associée à une terminologie pour parvenir à des définitions fondées et systématiques des termes.

6.2.4.2 Utilisation pour construire une ontologie

Projet Verre : une ontologie sur la fabrication et l'utilisation de la fibre de verre

Le projet Verre vient répondre à une demande du centre de recherche du groupe Saint-Gobain. Au sein des différentes filiales du groupe, les activités de veille documentaire et technologique, parmi lesquelles le repérage de nouveaux documents sur le web, jouent alors un rôle crucial pour garantir une avance technologique et industrielle. L'objectif du projet était donc de tester la faisabilité du développement d'une ontologie utile pour guider le classement de documents en fonction des profils des utilisateurs.

Dans ce projet, les aspects méthodologiques étaient tout aussi importants que l'ontologie elle-même. Cette étude a débouché sur un début d'ontologie (50 concepts, 20 relations) ainsi que sur un ensemble de recommandations méthodologiques pour ce contexte applicatif particulier [rapport-VERRE, 02]. L'ontologie a été structurée à l'aide de TERMINAE, à partir de l'analyse d'un corpus de langue anglaise composé de différents types de document sur le domaine. SYNTEX, UPERY et YAKWA ont été utilisés pour le dépouillement de ces corpus.

Ce projet a contribué à faire progresser le logiciel TERMINAE, en particulier en définissant des formats d'entrée et sortie des données qui soient standard ; à affiner la méthode, en précisant des recommandations pratiques pour mener des tâches réalisées jusque-là de manière empirique ; et enfin à mettre en forme les aspects « gestion de projet » jusque-là ignorés. Ce projet a donc débouché sur une proposition pratique de gestion de projet et sur des techniques pour réaliser les différentes tâches de modélisation [rapport-VERRE, 02].

Mise en forme : Puces et numéros

EON : amorce d'une ontologie du tourisme

L'expérience associée au workshop EOT associé à EKAW 2002 visait l'évaluation de logiciels de construction d'ontologies et surtout de leur représentation des connaissances [EOT, 02]. Bien que l'originalité de TERMINAE soit plus la méthode que la représentation des connaissances, la participation à cette évaluation a permis de bien situer la méthode et le logiciel par rapport à l'état de l'art. Un texte très court (1 page) servait de source de connaissances. Dans ce cas, l'analyse automatique de textes était triviale. En revanche, l'exigence demandée aux auteurs de justifier pas à pas la composition de l'ontologie résultante s'avère très riche.

Tout d'abord, elle a révélé l'originalité et la qualité de notre contribution, la plupart des autres démarches étant complètement orientées par le format de la représentation et non par des hypothèses sur la nature de la source de connaissances et de ce qu'est une ontologie. Ensuite, elle a permis de décrire exhaustivement une étude de cas illustrant la mise en œuvre de la méthode [EOT, 02]. Des critères jusque-là implicites pour décider des concepts à représenter et de la manière de les représenter ont été mis au jour : classe ou instance, concept ou relation, situation dans la hiérarchie des concepts, présence ou non d'une relation entre deux concepts, nature de la relation sémantique reliant deux concepts, etc.

Ce travail, tout à fait complémentaire du précédent, a été synthétisé sous forme d'un ensemble de recommandations dans [TIA, 03].

6.2.5 Impact de l'application ciblée sur le processus de construction de RTO

De ces expériences et de plusieurs projets menés au sein du groupe TIA, il ressort que l'application ciblée (dans laquelle doit être intégrée la ressource) détermine les choix méthodologiques à tous les stades de la modélisation. Cette analyse a été menée en deux temps avec différents collègues [GDR I3, 02] [RIA, 04]. L'impact de l'application cible sur le processus de modélisation a été examiné à travers les points suivants : profil du « constructeur », construction du corpus, choix et manière d'utiliser les outils de TAL, choix et utilisation des outils de modélisation. Sur chacune de ces dimensions, les points communs et les divergences entre

plusieurs projets, les propositions et les limites des travaux actuels ont été mis en évidence. Les conclusions de ce travail concernent plusieurs aspects de la recherche :

- la méthode d'étude de cet impact ;
- les enjeux en matière de disponibilité et d'intégration des logiciels ;
- les évolutions nécessaires des structures de représentation des ressources terminologiques.

Au niveau de la méthode à retenir pour ce type d'étude, un chemin encore long reste à parcourir pour passer de la mise en parallèle de quelques retours d'expérience à la définition d'une méthodologie générique en acquisition de connaissances à partir de textes et à la présentation de résultats illustrant de manière convaincante les retombées pratiques des recherches dans ce domaine. On ne peut que constater ici les limites d'une démarche expérimentale : beaucoup de recherches doivent encore être effectuées pour affiner des critères d'évaluation. Le travail considérable à mettre en œuvre pour élaborer un produit terminologique et le peu de réutilisations envisageables ne rendent pas facile la reproduction d'expériences. Néanmoins, les progrès ne pourront désormais venir que de la confrontation des problèmes rencontrés et des solutions choisies dans des expériences concrètes, face à des communautés d'utilisateurs.

Le verrou se situe désormais au niveau de l'*intégration* de ces résultats dans des tâches réelles de construction de ressources terminologiques ou ontologiques à partir de textes, et les progrès ne peuvent venir qu'en mettant ces résultats à l'épreuve de la pratique. Il faut alors sortir d'un cercle vicieux puisque des expérimentations en vraie grandeur ne peuvent être menées à bien que si les outils et interfaces sont arrivés à un stade de maturité assez élevé. L'une des principales difficultés est liée à la largeur du spectre des types de ressource terminologique ou ontologique envisageables. Une autre difficulté importante est le manque de modularité de certains logiciels. Il semble prometteur de prévoir des outils très simples, dont on sait clairement quelles sont les entrées et les sorties, quelle peut être leur localisation possible et leur apport dans une chaîne de traitements. Après avoir misé sur la possibilité de réaliser des outils génériques, il faut maintenant étudier comment la pertinence et le mode d'utilisation de ces outils sont conditionnés par la structure de la ressource à construire et donc *in fine* par l'usage prévu de cette ressource. Quel que soit le contexte, le problème le plus important à gérer restera celui de la masse (la quantité d'information que les outils d'analyse peuvent d'extraire des corpus est parfois énorme) qui génère un problème de temps. La construction d'une ressource exige du temps expert. Il faudra alors bien arriver à montrer que le coût de ce temps expert est largement compensé par une augmentation importante de la qualité et de l'efficacité des systèmes exploitant les ressources.

Le coût de développement de ressources terminologiques et ontologiques pose également la question de leur évolution. La nature même de ces structures doit être revue pour mieux prendre en compte de nouveaux besoins, tels que la facilité de mise à jour et le lien étroit avec les textes. Il semble indispensable de se tourner à l'avenir vers des structures de données plus souples, dont la mise à jour serait plus dynamique, qui permettent de revenir aux sources de connaissances (les textes) et de reconstruire le sens, voire les modèles, en fonction des usages. L'idée est de réduire les coûts en évitant de reconstruire une ressource pour chaque application dans un domaine donné. Au lieu de repartir uniquement de corpus pour faire une ressource complètement nouvelle, il s'agit de favoriser la réutilisation par l'adaptation de ressources existantes. Ces structures devraient donc laisser une part active à l'utilisateur et ne figent pas le sens définitivement.

6. 3 - Modèles conceptuels comme accès au contenu de documents

Je viens de montrer que les ontologies et les BCT étaient de « bons modèles » pour rendre compte des connaissances d'un domaine à partir d'analyse de textes. J'ai également montré que les logiciels de traitement automatique des langues permettaient d'assister efficacement la modélisation, et d'assurer une traçabilité des textes vers les modèles, moyennant d'adopter une

démarche méthodologique et fondée. Dans cette partie, je me focalise sur le rôle des ontologies et des modèles conceptuels construits à partir de textes, comme moyens d'accès au contenu de textes dans le cadre d'activités ciblées. Il s'agit d'une première réponse à une des questions relatives à la problématique exposée au 3.2.2 : *Comment tirer profit du parallèle évident entre l'extraction d'éléments de modèles à partir de textes et l'annotation des textes à l'aide d'éléments de modèles* ?

Dans les applications visées, les modèles doivent permettre de (pré)définir des parcours de lecture. Les connaissances modélisées, liées au raisonnement (modèles de tâches) ou décrivant le domaine (ontologies), servent à guider la consultation via des interfaces où l'utilisateur décide de son parcours, selon une logique éventuellement différente de celle de l'auteur. Pour moi, la question est de savoir comment construire les modèles pertinents pour ces applications. Je la reformule en prenant le parti d'une modélisation à partir de textes et de la prise en compte de l'activité dans laquelle va s'intégrer cette consultation. Ma contribution fait référence à trois expériences pour lesquelles le modèle conceptuel joue un rôle différent : point d'entrée direct vers des passages du document, représentation intermédiaire pour définir un index et ressource fournissant des méta-données d'indexation. Ces travaux, menés entre 1995 et 2000, révèlent le potentiel qu'offrent les modèles conceptuels pour faciliter la consultation de documents textuels. Ces trois utilisations des modèles anticipent les rôles joués par les ontologies dans le projet d'un Web Sémantique. Ils posent les bases de notre projet de recherche actuel.

6.3.1 Motivations

Les besoins des entreprises relatifs aux ressources terminologiques correspondent en fait soit à la terminologie, soit à la gestion documentaire, soit à la gestion des connaissances. Concernant les documents, les entreprises se heurtent à de multiples difficultés comme l'existence encore importante du support papier, de gros volumes documentaires, de leur gestion dans le temps et de leur accès, de la capacité à en interpréter ou exploiter le contenu, etc. Les besoins exprimés sur les vocabulaires portent sur la définition de méthodes et d'outils pour contrôler les défaillances langagières, repérer des ambiguïtés ou des polysémies, etc. En matière de connaissances, les entreprises souhaitent mieux la véhiculer, la retrouver, la localiser, la rendre accessible et la ressource terminologique est tantôt perçue comme un médiateur renvoyant à la source que sont les documents, tantôt comme une nouvelle source de connaissances à exploiter directement (Dieng *et al.,* 2001). L'enjeux est la pérennisation des techniques, des savoir faire et des compétences.

Au moment d'évaluer l'intérêt d'un modèle conceptuel pour orienter la consultation de documents, autour de 1995, il s'agissait de pousser plus avant le potentiel d'une consultation hypertextuelle de documents techniques. Ce mode de consultation peut faire appel à des index. Dans ce cas, les modèles de connaissances mis en jeu sont relativement simples et très proches du document (des listes de termes extraits des textes). Des parcours de lecture peuvent être proposés, en général sous forme d'historiques enregistrés à partir des usages et des habitudes des lecteurs. Ces solutions prennent plus en compte les éléments documentaires, les mots et la structure syntaxique du document, plus que des éléments conceptuels. Du point de vue de l'IC, peu de recherches avaient été menées sur l'utilisation de modèles conceptuels dans le cadre d'applications documentaires, alors que des besoins étaient exprimés par la communauté travaillant sur les hypertextes (Nanard *et al.*, 1996).

Dans ce contexte, j'ai envisagé une contribution de l'ingénierie des connaissances via les modèles conceptuels et les approches terminologiques. J'ai voulu évaluer l'apport des modèles pour mieux assurer l'adéquation entre les modalités de consultation, les parties de document présentées et les contextes d'usages (tâche et intérêts des utilisateurs). De plus, je voulais profiter de l'opportunité offerte par les approches terminologiques à partir de textes et ainsi les valider sur de nouveaux types d'applications. Finalement, en m'engageant vers cette voie nouvelle pour moi, j'ai élargi la gamme des applications pour lesquelles la modélisation de connaissances peut être un atout. Ceci permet de tirer des enseignements sur la nature des modèles adaptés à chaque type

d'application, sur les modalités de construction de ces modèles et la pertinence des approches à partir de textes.

6.3.2 Apport de modèles construits à partir de textes à la consultation documentaire

6.3.2.1 Modèle de tâches et consultation de documents

Dans le cadre de la conception d'un système de consultation de document technique, la prise en compte de la tâche dans laquelle s'inscrit la consultation des documents permet d'anticiper sur les contextes de consultation et de proposer des accès aux textes adaptés aux besoins, en fonction du déroulement de la tâche. C'est là un des objectifs du projet HYPERPLAN. Le document concerné, le *Guide de Planification des réseaux régionaux*, prescrit les tâches du planificateur des réseaux régionaux [EKAW, 96]. Le système de consultation du *Guide* doit s'appuyer sur un modèle de la tâche du planificateur des réseaux régionaux telle qu'elle est décrite dans le document. La démarche retenue comporte deux phases : d'abord, l'élaboration d'un modèle conceptuel de la tâche et du domaine concernés, puis l'association des composants de ce modèle à des fragments du document. Le *Guide* est un document assez volumineux. Cette approche a été possible grâce à l'utilisation d'un extracteur de termes, LEXTER, qui permettait de disposer directement de liens entre des termes du domaine et les parties de textes où ils sont utilisés. Ce projet a d'ailleurs été présenté lors de la présentations des systèmes LEXTER et COATIS.

La tâche décrite dans le document a été modélisée en suivant l'approche MACAO-II. Le but de cette modélisation était inhabituel pour moi : orienter la consultation d'un document et non contribuer à l'automatisation d'un processus de résolution de problème. Cette finalité nouvelle modifie certaines caractéristiques du modèle, donnant par exemple un poids plus important aux liens entre modèle et textes, alors que le degré de décomposition de la tâche est, lui, moins fin [EKAW, 96]. La mise en place des liens entre tâches et parties de texte passe par les termes associés aux tâches et aux concepts qu'elles utilisent en entrée ou produisent en résultat. Ce travail revient à sélectionner, parmi toutes les occurrences d'un terme, celles qui illustrent ou expliquent la mise en œuvre d'une tâche du modèle. De ce fait, l'exploitation du réseau terminologique tiré du *Guide* par l'analyse de LEXTER joue également un rôle important. De plus, le modèle de tâche doit être défini en premier et constitue une grille pour orienter le choix de passage de texte pertinents.

Le mode de consultation ainsi obtenu n'est qu'une des composantes du système final, qui comporte aussi un index, également construit à partir des résultats de LEXTER, et une table des matières classique. L'utilisateur dispose donc de points d'entrée complémentaires dans le document, dont l'apport a été validé auprès d'utilisateurs.

Cette première expérience a confirmé les atouts des résultats de l'extracteur de termes, mais aussi le coût élevé d'une sélection manuelle des renvois, des parties de textes associées aux entrées de l'index ou bien aux tâches du modèle. Elle a permis d'établir des résultats méthodologiques sur la manière de déterminer le contenu d'un modèle conceptuel (ici, le modèle de tâche) en fonction des objectifs de l'application. L'ensemble de ces travaux en amont de la diffusion d'un document sont autant de moyens pour outiller un meilleur accès à son contenu.

6.3.2.2 Modèle de tâches et construction d'index

Dans la cadre d'un autre projet ayant un objectif analogue, le projet MOUGLIS, une orientation différente a été retenue. L'hypothèse était qu'une base de connaissances terminologiques pourrait faciliter la mise au point de différents types de modèle, chacun servant ensuite de point d'entrée pour consulter les textes. Ce projet, déjà mentionné, a permis de mettre au point la notion de base de connaissances terminologiques (BCT), le logiciel GEDITERM et d'éprouver la (non) neutralité des BCT comme représentations du contenu de textes.

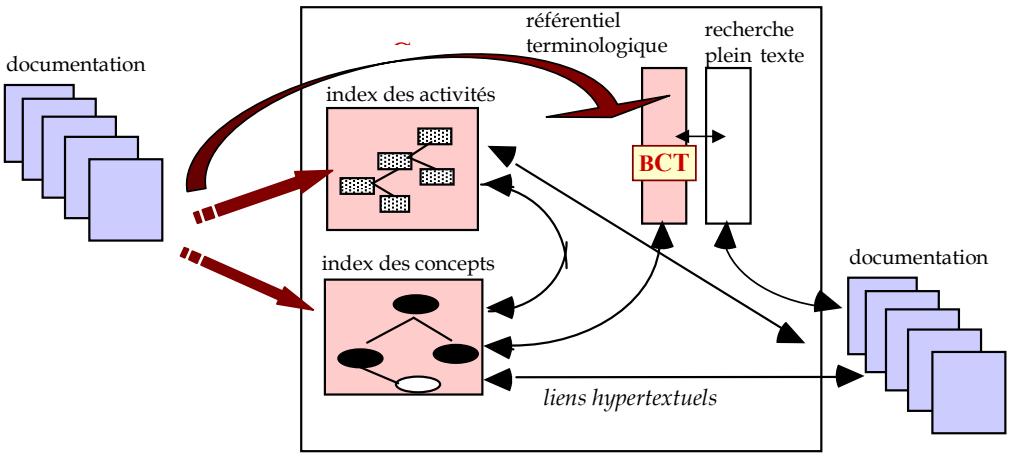

Figure 6.3.2 : Construction de différents modèles à partir de textes et place d'un référentiel terminologique dans le projet MOUGLIS

Dans ce cas, la BCT prépare la sélection de parties de textes associées aux termes et aux concepts. En intégrant ces concepts dans le modèle de tâche et dans l'index construits pour consulter le document, on pose donc des liens directs vers des parties de textes choisies a priori (fig. 6.3.2). Il s'avère que suivant la tâche utilisant un concept, ce ne sont pas les mêmes parties du document qu'il faut présenter. De ce fait, un travail d'adaptation du contenu de la BCT est nécessaire, soit pour identifier des concepts mieux adaptés à la modélisation des tâches, soit pour modifier la liste des parties de texte associées. La BCT permet de gagner un temps considérable pour mettre au point ces modèles de consultation, mais le coût total de la construction de la BCT et de ces modèles est à peu près équivalent.

Une autre originalité du projet a été de mieux mesurer la complémentarité des différents modèles comme points d'entrée dans les documents. Une réflexion a été menée pour savoir par exemple si les entrées choisies pour constituer l'index devaient imposer des contraintes sur le choix des concepts dans le modèle de tâches, ou inversement. Il ressort que la consultation ne bénéficie d'une vraie complémentarité des modes de consultation que s'ils renvoient au texte chacun avec sa propre logique, et donc éventuellement vers des passages différents.

6.3.2.3 Apport des modèles conceptuels au référencement de sites Web

Le référencement d'un site web renvoie à l'enrichissement de ce site de manière à augmenter les chances de le retrouver à l'aide de mots clés et des moteurs de recherche du marché. Au centre du projet se trouve le logiciel IndexWeb, développé par Synapse Développement, qui aide à évaluer un site et suggère des améliorations pour qu'il soit mieux référencé : ajout de mots-clés dans les méta-données des pages, modification de leur titre ou même de leur contenu.

Une base terminologique pour l'indexation de pages web

L'idée originale du projet est de constituer un thesaurus (c'est-à-dire un lexique de l'entreprise dont les termes sont définis par des phrases et quelques relations entre eux) à partir de documents techniques de l'entreprise pour faciliter le choix de mots clés et la rédaction du contenu des pages. Trois sites de PME ont été étudiés (CLS Argos, OKTAL et CEDOM). Pour chacun d'eux, un corpus a été constitué à partir des sites eux-mêmes mais aussi de documents techniques ou de communications de l'entreprise. Chaque corpus a été dépouillé à l'aide des logiciels SYNTEX et CAMELEON, de manière à constituer des thesaurus. Cet ensemble de termes reliés par des relations de hiérarchie et sélectionnés comme étant des termes utilisés pour décrire ou retrouver l'entreprise via son activité ont été rattachés à des pages particulières du site et ont permis d'en constituer un

index. Il comporte les concepts principaux du domaine du point de vue des personnes consultant le site, et renvoie vers une sélection de pages traitant de ces concepts.

Plusieurs hypothèses relatives à l'analyse de document pour mieux référencer des sites web ont été validées à travers ce projet. Tout d'abord, cette analyse doit s'appuyer sur les documents de communication des auteurs du site (par exemple une entreprise) et de leur public cible (par exemple ses clients). Ensuite, cette analyse gagne à être menée selon des techniques linguistiques précises, qui aillent jusqu'à la caractérisation des relations entre les concepts rattachés aux termes. Enfin, les termes et concepts produits à l'issue de cette analyse sont de bons candidats à être utilisés pour rédiger les pages du site ou en caractériser le contenu par des mots-clés. Le thésaurus résultat de cette analyse se matérialise sous la forme d'une base de terminologique [Rapport-CANNESSON, 01]. Si on inclut dans les documents analysés les pages du site, les termes de la base sont en quelque sorte des index du site. Le fait que la base terminologique soit structurée autour de concepts permet alors d'organiser l'index de consultation en fonction de champs sémantiques et de concepts liés à l'activité ou au domaine que le site met en valeur [Rapport-BUSNEL, 01].

Index ou moteur de recherche : une question de coût ?

Notre expérience vise à rajouter des connaissances structurées à des pages web n'ayant initialement pas d'autre sémantique que leur contenu en langage naturel. A ce titre, l'étude entre dans la problématique dite « du web sémantique ». Dans leur forme initiale, ces pages ont du sens à la lecture par un humain, mais elles restent mal exploitées par les moteurs de recherche. Or les logiciels d'analyse terminologique et d'extraction de termes permettent d'améliorer le contenu et la description des pages sans un sur-coût trop important. Leur apport correspond à l'ajout de termes plus pertinents et mieux classés dans les méta-données de la page, mais aussi au choix de ces termes dans la rédaction de pages elles-mêmes. Enfin, la lisibilité du site peut se trouver enrichie par un index formé des éléments essentiels de la terminologie et renvoyant aux pages du site. L'avantage de cet index est d'offrir en un coup d'œil une vue synthétique du contenu du site.

Les entreprises ont convenu de la pertinence et de l'intérêt évident des résultats de l'analyse terminologique, ainsi que de l'ajout d'un index structuré sur leur site. Cependant, la démarche de mise au point de la base terminologique leur a semblé encore trop coûteuse. Enfin, le bilan de l'étude a permis de proposer des évolutions du logiciel d'aide au référencement.

Une alternative à ce genre d'approche est de s'appuyer directement sur un réseau terminologique correspondant aux résultats de l'analyse des pages du site. En effet, ce réseau peut être généré automatiquement, à partir des résultats d'un extracteur de termes (comme SYNTEX) et d'heuristiques de sélection de termes. Ensuite, pour éviter de fixer manuellement les liens des termes vers les textes, on peut mettre en place un moteur de recherche local au site qui s'appuie sur cette ressource. Pour toute recherche formulée par un utilisateur, le système exploite la terminologie comme un index pour retourner des pages du site pertinentes, et propose des termes proches de ceux de la requête pour l'affiner ou la reformuler. Dans ce cas, les auteurs du site doivent bien maîtriser le contenu des pages, qui seul guide le lecteur. Et le lecteur doit avoir une idée des informations disponibles.

6.3.3 Premières conclusions

Ces trois premiers projets ont été menés systématiquement en partenariat avec des entreprises pour répondre à des besoins précis. Ils ont permis de tirer des enseignements méthodologiques sur la construction de modèles conceptuels pertinents pour l'accès au contenu de documents, et même pour définir avec les auteurs des supports rendant plus accessible ce contenu. Parmi ces enseignements, certains choix initiaux ont été mis à l'épreuve. Les grandes lignes en sont les suivantes :

- l'élaboration d'un thésaurus ou plus, d'une ontologie, ne se justifie que si l'entreprise exprime des besoins stratégiques en matière de gestion de sa terminologie ou de ses documents. Le coût de l'élaboration d'une telle structure de donnée, même à partir de textes et à l'aide de logiciels, ne peut pas être justifié par une seule application (comme l'accès au site web, s'il ne s'agit pas là d'un moyen de communication clé pour l'entreprise) ;
- le lecteur ne bénéficie d'une vraie complémentarité des modes de consultation que s'ils renvoient au texte chacun avec sa propre logique, et donc éventuellement vers des passages différents ; en ce sens, un modèle utilisé comme support à l'accès au texte fixe un cadre pour définir des parcours de lecture ;
- ce type d'analyse doit porter sur un corpus plus large que le site web de l'entreprise, et trouve sa place dans une réflexion plus globale sur la terminologie de l'entreprise et sa stratégie de communication avec ses partenaires et ses clients.

Les leçons tirées du point de vue de l'utilisation des logiciels de traitement automatique pour la consultation documentaire sont développées dans le bilan de ce chapitre.

6. 4 - Bilan

6.4.1 Traitement automatique des langues pour la construction de RTO

6.4.1.1 Synthèse de notre contribution

Pratiquement, le logiciel CAMELEON [Thèse-SEGUELA, 01] a été un des premiers extracteurs de relations sémantiques avec PROMETHEE (Morin, 1999) et LIKES (Rousselot *et al.,* 1996). **L'approche par patron** avait déjà été expérimentée en linguistique de corpus, essentiellement à l'aide de concordanciers comme SATO ou de logiciels d'exploration de textes comme SEEK (Jouis, 1993). Les points forts de CAMELEON sont de proposer une **base de marqueurs génériques**, qui peuvent être adaptés au corpus étudié, et surtout de permettre la définition de marqueurs spécifiques. Le logiciel est également un des rares à permettre ensuite d'intégrer les résultats des recherches au sein d'un modèle sous forme de relations sémantiques entre concepts. En assurant la continuité du processus, CAMELEON se place vraiment dans une perspective d'ingénierie des connaissances. Dans le même esprit d'intégration au sein d'un processus, le lien établi avec SYNTEX a permis de produire des résultats complets en aidant à **identifier les concepts pouvant être en relation**. Enfin, CAMELEON est un logiciel maintenu et actualisé. En traitant un corpus étiqueté, la dernière version s'appuie sur les avancées récentes du TAL pour fournir des résultats plus précis. Une des limites du logiciel est de ne pas proposer encore une base de marqueurs pour l'anglais.

Au-delà de la fourniture de ce logiciel, ma contribution est à la fois d'ordre théorique et méthodologique. De mes différentes études, j'ai retenu des principes qui fondent mon point de vue sur l'analyse de textes, en particulier pour l'identification de termes, de concepts et de relations :

(i) la nécessité d'une approche traitant simultanément les niveaux linguistique et conceptuel, tout en identifiant bien la part de chacun ;

(ii) l'intérêt d'étudier ensemble les termes et les relations lexicales, l'identification des termes et des relations étant complémentaires et presque indissociables ;

(iii) l'apport des interfaces de navigation pour parcourir l'ensemble des résultats en fonction des relations syntaxiques entre termes ;

(iv) l'intérêt de multiplier les critères (numériques, statistiques ou heuristiques) de sélection et de présentation des candidats termes ;

(v) les possibilités (non encore complètement exploitées) d'automatiser le rapprochement des termes en fonction de critères syntaxiques et sémantiques ;

(vi) la nécessité de combiner l'utilisation d'approches complémentaires afin de combler les limites de chacune (étude linguistique et statistique de distributions, utilisation de patrons) ; chaque approche convient mieux à certains types de corpus (i.e. les approches par patrons pour les documents pédagogiques, les approches distributionnelles pour les documents redondants et présentant des régularités) ;

(vii) l'importance de donner aux termes étudiés un statut particulier, et au réseau de termes une place à côté du réseau conceptuel. On peut considérer que **l'ensemble des termes associés aux concepts de l'ontologie constitue un lexique.** Ce lexique n'est pas construit avant l'ontologie mais en même temps : **il en est une composante**, car c'est un des résultats du processus de modélisation.

Les cinq premières affirmations constituent autant d'éléments de réponse à nos questions sur la nature des logiciels de TAL utiles à la construction de RTO, et à la manière d'utiliser leur complémentarité pour couvrir la diversité des connaissances à recueillir et représenter (question 3 du 3.2.2). Le dernier point confirme le travail présenté au chapitre 5 sur les modèles de données pour les RTO (question 2 du 3.2.2).

6.4.1.2 Situation par rapport aux travaux nationaux et internationaux

Nos travaux d'utilisation et de définition de logiciels de TAL pour la construction de RTO ont été précurseurs de ce type d'approche. Avec l'utilisation de Lexter pour la modélisation de connaissances en 1994, au moment de la création du groupe TIA, notre collaboration avec D. Bourigault et A. Condamines a ouvert de nouvelles perspectives pour l'acquisition de connaissances à partir de textes, tant au niveau national (d'autres extracteurs de termes comme NOMINO ont été utilisés en IC, par exemple par R. Dieng, et d'autres logiciels de TAL ont été développés dans une finalité IC, comme l'extracteur de termes Likes de F. Rousselot) qu'au niveau international, où les travaux comparables sont d'abord venus de constructeurs de terminologies (le System Quirk de Ahmad par exemple ou Text Analyser de J. Kavanagh associé à Code4). D'autres contributions venues de chercheurs du TAL (comme U. Hahn, pour la compréhension de textes, ou P. Zweigenbaum, au sein du projet Ménélas) ont ensuite été transposées à la construction d'ontologies. Ce n'est que depuis 1998 environ que des approches utilisant le TAL pour la modélisation conceptuelle se sont développées au sein de l'ingénierie des connaissances en Allemagne (au DFKI, R. Studer, A. Maedche et S. Staab qui ont défini la suite de logiciels KAON), en Italie (R. Basili, université de Rome, ou P. Velardi, univ de Parme) et plus récemment en Grande-Bretagne (extraction d'information à Sheffield, F. Ciravegna et MnM pour l'annotation sémantique au KMI de l'Open University, (Vargas-Vera *et al.,* 2002).

Les travaux actuellement apparentés à CAMELEON sont, au plan national, ceux de (Baneyx et Malaizé, 2004) ou (Grabar et Hamon, 2002) et, au plan international, Text-to-Onto (Maedche et Staab, 2004) ou (Samdja *et al.,* 2005). La particularité de CAMELEON est de constituer un logiciel indépendant, pouvant être adapté à différentes langues à condition qu'un étiqueteur soit disponible. Une autre spécificité est de gérer uniformément tous les types de relation, alors que les travaux récents proposent deux approches différentes pour les relations hiérarchies (est-un) et les autres relations sémantiques. Pour les relations hiérarchiques, ils utilisent soit des ressources existantes (WordNet par exemple) soit des outils de regroupement (clustering sur la base de critères de co-occurrence, de distributions, etc.) (Cimiano *et al*., 2004) (Cimiano *et al.,* 2005).

6.4.2 Propositions méthodologiques

6.4.2.1 La méthode TERMINAE

Il est difficile de faire la part de ma contribution à TERMINAE par rapport à celle de mes collègues du LIPN. Je mentionne ici les propositions intégrées à la méthode suite à ma participation à ce projet, et élaborées en commun avec B. Biébow et S. Szulman. Ces propositions portent essentiellement sur l'étape d'analyse terminologique en lien avec la normalisation. Elles viennent enrichir celles développées au sujet des bases de connaissances terminologiques et mentionnées au chapitre 5. Elles constituent des réponses à l'objectif méthodologique, formulé au chapitre 3 (question 4 du 3.2.2). La méthode proposée suggère l'utilisation des logiciels d'analyse de textes (comme CAMELEON) que j'ai développés ou évalués (comme SYNTEX) ainsi que d'une plate-forme de modélisation comme TERMINAE.

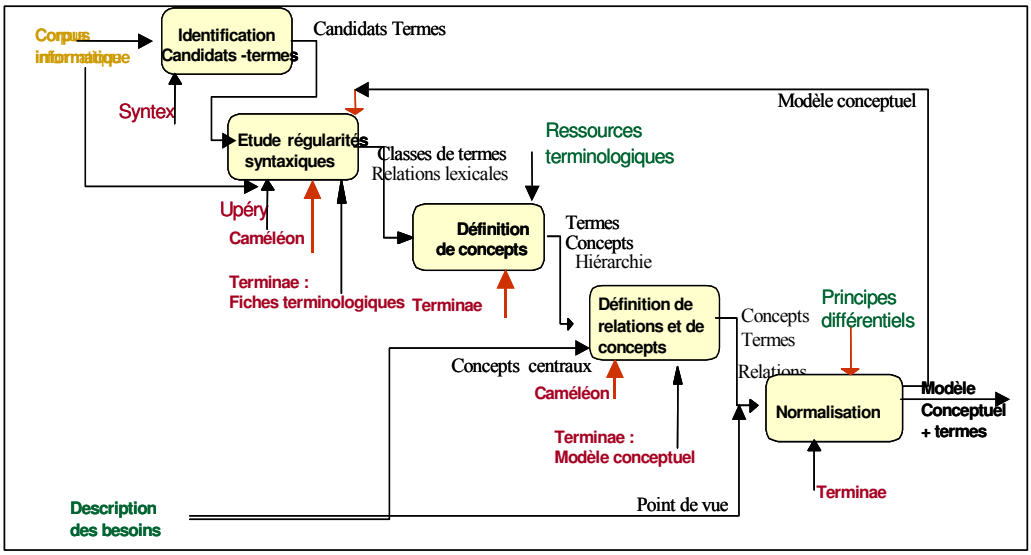

Figure 6.4.2 : Détail de l'analyse terminologique suite à notre participation à la méthode TERMINAE : l'étape d'analyse terminologique vise désormais l'identification de termes et de concepts à partir des résultats d'analyse de textes et en fonction des objectifs de modélisation. On distingue deux temps dans la définition de concepts : un étape où les concepts sont définis à partir des informations trouvées dans les textes, et la normalisation, qui, par appalication de principes différentiels, conduit à réorganiser les concepts ou à en définir de nouveaux.

(i) Identifier des concepts et des relations entre concepts sont des problèmes complexes qui peuvent être décomposés en sous-problèmes plus simples (fig. 6.4.2) : chacun peut être pris en charge par le système ou laissé à l'utilisateur. Je défends la **nécessité d'une intervention humaine au sein d'un processus supervisé pour parvenir à des modèles fins et pertinents**. Seule l'interprétation humaine permet de décider quels seront les concepts avant même leur normalisation, de juger si une relation est présente à partir d'une phrase, de déterminer le type de relation et les concepts reliés. L'interprétation humaine prend place à deux moments du processus de modélisation : (1) au moment de parcourir les textes et d'observer l'usage des termes à partir des résultats de l'analyse de corpus (étapes « identification de candidats-termes » et « étude des régularités syntaxiques » dans la figure 6.4.3) puis (2) au moment de construire et organiser des concepts dans le réseau conceptuel (étapes « définition de concepts et de relations » et « normalisation » figure 6.4.2). Dans le premier temps, c'est la linguistique qui montre son incapacité à trancher dans le débat du sens des termes sans faire appel à une perspective de lecture, en s'appuyant seulement sur l'introspection ou sur les usages manifestés dans les textes. Dans le deuxième temps, c'est la tradition philosophique de l'ontologie reprise dans une perspective opérationnelle par B. Bachimont qui souligne que les critères de différentiation conceptuelle ne

sont pas le seul reflet de relations entre termes : au-delà de l'usage de la langue, l'analyste est porteur de critères structurants liés à des avis d'experts et à la prise en compte des traitements, des raisonnements qui seront faits sur les concepts définis.

(ii) **La modélisation est un processus itératif** : chacune des étapes valide, complète et enrichit les précédentes. Mais elle peut aussi les remettre en question et conduire à les reprendre avec de nouveaux paramètres. Par construction, un modèle ne peut être construit de manière linéaire après une analyse exhaustive de tous les textes, un dépouillement systématique de tous les résultats qui peuvent en être extraits et une normalisation de l'ensemble. L'état courant du modèle sert de grille pour rechercher d'autres connaissances à partir de nouvelles analyses ou de nouvelles interprétations de résultats de logiciels de TAL (flèche de retour du modèle conceptuel vers l'étape « étude des régularités syntaxiques » sur la figure 6.4.2). On retrouve là un résultat classique de la modélisation conceptuelle actualisé dans le contexte de la modélisation à partir de textes.

(iii) La **constitution du corpus** est une étape cruciale, qui doit pouvoir être validée ou corrigée par les premières analyses terminologiques à l'aide de logiciels de TAL. La nature du corpus conditionne la qualité des résultats obtenus par les différents logiciels utilisés. Le choix des outils d'analyse doit tenir compte des caractéristiques du corpus ou inversement, suivant les contraintes du projet. Les résultats disponibles pour **ajuster logiciels et corpus** sont encore peu nombreux. Il ressort que l'analyse distributionnelle s'applique bien à des textes rédigés avec des formes redondantes mais pas forcément grammaticalement correctes ou habituelles. Les approches par marqueurs supposent des textes explicites et bien écrits, comme les textes pédagogiques, les livres, etc. Ce résultat est conforme à celui de (Baneyx *et al*., 2005)

(iv) Un environnement de construction d'ontologies à partir de textes doit permettre de mettre au point un **modèle au niveau conceptuel** avant de disposer d'une version formelle. Ce modèle se présente sous un format peu contraignant, autorisant des états incohérents, incomplets du modèle.

(v) La **normalisation** conditionne le fait d'obtenir une ontologie et marque la différence avec d'autres RTO. Les **spécificités des ontologies** ne se situent pas uniquement dans les primitives de représentation des connaissances (concepts, relations, etc.) et leur degré de formalisation (elles doivent être manipulables par un système informatique). Elles résident essentiellement **dans la nature de leurs contenus et dans la manière de les déterminer.** Or ces contenus, les « connaissances », n'émergent pas de l'analyse de textes ou autres sources de connaissances : ils sont des constructions révélant un point de vue, une interprétation particulière du domaine, que seul l'analyste peut produire [DAGSTUHL , 05]]. A l'heure où une plus grande automatisation du processus est envisagée, il me semble important de rappeler cette caractéristique, quitte éventuellement à y renoncer pour des structures plus simples, moins lourdes et mieux adaptées aux objectifs ciblés.

(vi) En suivant la méthode TERMINAE, les ontologies construites prennent en compte les contraintes liées à l'usage qui en est prévu au sein du système informatique à concevoir. Elles devraient être utiles et pertinentes, mais sans aucune garantie de **réutilisabilité**. La recherche de la réutilisabilité n'est ni systématiquement possible ni souhaitable pour certaines applications. Ces ontologies comportent une composante terminologique et même des liens vers les textes à partir desquels elles ont été construite, ce qui facilite leur lisibilité et devrait favoriser leur **maintenance**.

6.4.2.2 Situation par rapport aux travaux nationaux et internationaux

Au niveau méthodologique, notre approche est une émanation des réflexions du groupe TIA, qui se démarquent des travaux internationaux sur la notion d'ontologie.

Le choix de partir des usages et de définir des ontologies répondant à d'autres usages particuliers s'avère orthogonal avec toutes les approches faisant l'hypothèse d'ontologies génériques et universelles (ontologies formelles en philosophie ou ontologies générales pour l'interopérabilité des systèmes). Les usages anticipés des modèles sont de plus en plus pris en compte depuis quelques années pour des applications ciblées, y compris sur le Web comme les

Web Service. Le caractère consensuel des ontologies demeure, mais leur neutralité par rapport aux utilisations est remise en question. Parce que les connaissances couvrent des domaines cernés, spécialisés, les ontologies telles que je propose de les construire se démarquent des celles d'approches abordant les connaissances et la langue dans toute leur généralité (comme Wordnet, Cyc, MicroCosmos). Je tire parti du fait que les usages de la langue et les connaissances dans ces domaines peuvent être anticipés et observés avec une relative régularité. En partant des textes, le point fort est de s'appuyer sur le vocabulaire en usage. Le prix à payer est parfois le manque de définitions dans les textes ou leur mauvaise couverture du domaine. Dans la plupart des cas, l'analyse de leur seul contenu permet de produire facilement des réseaux sémantiques, mais parvenir à des ontologies respectant des principes différentiels requiert de faire appel aux experts ou à des connaissances de sens commun.

En privilégiant les textes comme sources de connaissances, et les logiciels du traitement du langage comme méthode d'approche de leur contenu, le groupe TIA a contribué à un renouvellement des travaux sur les ontologies. Les propositions méthodologiques correspondantes sont venues donner des méthodes et des indications précises et pratiques pour mettre en oeuvre des recommandations vagues du type « trouver les concepts ; les organiser en hiérarchies, … » telles qu'on les trouve dans (Fridman-Noy *et al.,* 1997) ou (Fernández-López *et al*., 1999). Plus radicalement, ces propositions ouvrent la porte à une certaine automatisation qui permettrait de parvenir enfin à une maîtrise des coûts et des délais pour ce type de projet. Cette automatisation favoriserait la multiplication des ontologies disponibles. Or il s'agit là d'un des enjeux du Web Sémantique, largement mis en avant par le W3C : plus il y aura d'ontologies dans des domaines divers, plus il sera facile d'annoter ses pages web. Là encore, nos travaux invitent à la prudence : toute ontologie d'un domaine ne convient pas forcément pour un site donné ; au moins autant que l'ontologie, la capacité à associer des termes, des passages de textes aux concepts de l'ontologie conditionne son utilisation. Ce point de vue converge avec ceux des chercheurs en terminologie et sciences de l'information [AO, 05]. C'est dans la perspective d'une construction plus rapide des ontologies que se situe le développement croissant de l'apprentissage automatique pour la construction et l'utilisation des ontologies : par apprentissage, on peu associer automatiquement des « instances » identifiées dans des documents et des concepts d'une ontologie.

6.4.2.3 Bilan sur la plate-forme TERMINAE

La plate-forme TERMINAE permet d'intégrer dans un même environnement le dépouillement des résultats d'analyses de textes et la modélisation conceptuelle. La mise à disposition d'outils de traitement du langage au sein d'un environnement de modélisation, originale en 1999, est aujourd'hui proposée dans plusieurs plates-formes : en amont de DOE (Malaisé *et al.,* 2004), avec Text-To-Onto au sein de KAON (Maedche & Staab, 2000), dans WebODE (Corcho *et al.,* 2002), sous forme d'outils d'apprentissage à partir de texte (Ciravegna *et al*., 2002), ou de plug-ins dans Protégé-2000 (Buitelaar *et al.,* 2004).

Mon expérience fait ressortir la nécessité de pouvoir ajouter ou faire évoluer les outils d'analyse indépendamment de la méthode, de diversifier les modalités d'usage des différents modules. Par exemple, il doit être possible d'utiliser ou non un extracteur de termes, de valider les hypothèses de relation trouvées à l'aide des marqueurs dans différents environnements de modélisation ou dans celui prévu par défaut, etc. Pour cela, l'architecture de la plate-forme de modélisation doit garantir cette modularité et ne pas figer un enchaînement optimal.

Les originalités de TERMINAE se situent dans l'importance accordée aux termes et aux textes. Les extraits de textes associés aux termes ainsi que les fiches terminologiques, qui forment la composante terminologique de l'ontologie, enrichissent le modèle produit. La traçabilité du modèle vers les textes et inversement favorise à la fois la maintenance et l'utilisation des résultats pour l'indexation sémantique.

6.4.3 Modèles conceptuels et consultation documentaire

6.4.3.1 Intérêt du TAL pour construire des modèles de consultation documentaire

À partir des objectifs initiaux d'exploiter au mieux la relation entre modèles conceptuels et textes puis d'évaluer l'intérêt des logiciels de TAL pour associer ces modèles à de nouveaux textes, j'ai établi le bilan suivant :

- Parce qu'ils sont construits en fonction de la tâche des futurs utilisateurs, les modèles conceptuels permettent de mieux organiser la consultation de documents au cours de la réalisation de cette tâche ; c'est une des contributions originales de l'ingénierie des connaissances par rapport aux approches classiques de recherche d'information.
- Les modèles construits à partir d'analyse de textes sont tout à fait adaptés pour indexer ou consulter ces mêmes textes ; la composante terminologique des modèles permet de retrouver plus facilement des éléments de textes relatifs à certains concepts.
- L'utilisation des modèles pour la consultation des documents à partir desquels ils sont construits doit être prise en compte très tôt dans le choix du corpus, des logiciels d'analyse et dans le choix du contenu du modèle.
- Les outils d'analyse de textes tels que ceux utilisés pour construire des ontologies, comme les extracteurs de termes, ceux qui suggèrent des familles de termes ou des relations entre termes, sont de bons outils pour échafauder automatiquement un réseau terminologique simple qui s'avère un bon index pour la consultation des sites. Ces principes sont à la base de certains outils de recherche sur les sites.
- Pour aller au-delà d'un réseau terminologique, c'est-à-dire construire un modèle conceptuel (comme un modèle de tâche ou un thésaurus), les résultats d'un extracteur de termes constituent un point de départ utile. Mais la sélection manuelle des renvois, des parties de textes associées aux entrées de l'index ou bien aux tâches du modèle représente un coût parfois dissuasif. Du point de vue méthodologique, le contenu du modèle conceptuel est clairement fixé en fonction des objectifs de la consultation.

L'ensemble de ces travaux, réalisés en amont de la diffusion d'un document, sont autant de moyens pour outiller un meilleur accès à son contenu. En cela, ils sont originaux et j'ai décidé de poursuivre dans ce sens en étudiant d'autres types de modèle et d'autres contextes d'application.

6.4.3.2 Ce que les ressources terminologiques apprennent des entreprises

Les expériences menées montrent qu'un projet de construction d'une ressource terminologique produit un impact qui dépasse la structure de données proprement dite. Tous les échanges et les interrogations nécessaires sont d'excellents révélateurs de la nature des connaissances détenues par les différents acteurs de l'entreprise à différents moments de son existence [CIFT, 04]. L'étude des cohérences ou divergences terminologiques, voire conceptuelles, conduit à en repérer les causes ou les répercussions éventuelles, et à expliquer ou à anticiper les dysfonctionnements entre domaines spécialisés, métiers ou équipes. La mise en évidence de la part des connaissances partagées par rapport aux connaissances implicites constitue un bon indicateur des communications et des collaborations. Ainsi, les analyses terminologiques contribuent à mieux caractériser l'identité de l'entreprise, ses composantes et l'image qu'elle offre. Enfin, elle révèle des savoir-faire et des connaissances techniques parfois jusque-là implicite ou éparpillées. En conclusion, elles n'ont pas seulement un intérêt technique (modélisation de connaissances) ou organisationnel (meilleure coordination du travail, meilleure caractérisation des « cultures » des divisions ou encore support à la réalisation de tâches), elles ont également un intérêt pour le marketing (image interne et externe de l'entreprise) et stratégique (meilleure connaissance d'elle-même).

6.4.3.3 Nouvelles perspectives

L'atout principal de ces recherches a été d'élargir notre perspective d'utilisation des modèles conceptuels. À travers ces projets, j'ai confirmé que l'ingénierie des connaissances pouvait cibler des applications nécessitant des modèles de connaissances sans les rendre opérationnels. Le système de consultation du document est vu comme une application à base de connaissances pour laquelle le modèle conceptuel permet de restituer auprès de l'utilisateur les connaissances de manière plus pertinente.

Ces premiers travaux ont soulevé des limites et la nécessité d'approfondir de nouvelles questions ou des axes de recherche. Je mentionne ceux qui font l'objet de mon projet de recherche actuel, qui seront développés au chapitre suivant avec les approches choisies pour y répondre.

1. J'ai développé ou contribué à la spécification de plusieurs logiciels pour faciliter la construction d'ontologies à partir de textes. Or pratiquement, et malgré les efforts entrepris, la chaîne des traitements automatiques du langage en amont de la modélisation n'est pas continue. Ce verrou s'explique par des raisons plus techniques que méthodologiques, autant pour CAMELEON que TERMINAE. De plus, le logiciel ne conduit pas à expliciter clairement des principes différentiels. Enfin, et c'est là un enjeux de taille, le logiciel n'intègre pas assez d'éléments méthodologiques pour orienter son utilisation et le processus de modélisation en fonction du type d'application visé. **L'enrichissement de ces logiciels** s'impose.
2. La sélection des entrées et passages de textes associés à des éléments de modèles se fait soit à la main, sur la base d'une lecture des occurrences des termes associés, soit en fonction de la présence d'indices, comme des éléments de définition, etc. Ce travail, long et coûteux, rend rédhibitoire l'utilisation des ontologies dans ce contexte. Des recherches existent sur la manière de l'outiller, qui tiennent peu compte des aspects terminologiques. Il faudrait désormais **tirer profit des logiciels de TAL pour indexer des documents avec une ontologie**.
3. La structure des documents ainsi que leur mise en forme matérielle constituent des indices sémantiques aujourd'hui reconnus comme très importants. Or les analyses menées avec les logiciels de TAL ne les prennent pas en compte. Pire, ces analyses sont menées souvent uniquement au niveau de la phrase, et cassent l'unité du texte qui contribue à lui donner du sens. **Comment exploiter la mise en forme des textes pour repérer des éléments de connaissance** ?
4. J'ai expérimenté la manière **d'utiliser de ressources terminologiques et ontologiques pour accéder au contenu de documents** dans des contextes très proches. Or les applications de recherche d'information invitant à utiliser des ressources de ce type sont de plus en plus nombreuses : classification de documents, recherche d'information à l'aide de moteur, au sein de collections ciblées, fermées ou sur le web. Mes conclusions sur l'apport des RTO doivent être revues dans ces autres contextes.
5. Utiliser des ontologies pour des applications documentaires est envisagé jusqu'ici d'un point de vue statique : l'ontologie correspond à l'état des connaissances d'un domaine à un instant donné. Elle est définie en accord avec le contenu des documents à annoter ou à indexer. Mais on peut s'interroger sur sa pertinence à moyen terme si jamais la collection de documents et le contexte d'utilisation évoluent. La question de la **maintenance cohérente** des modèles et des documents se pose, ainsi que celle de l'intérêt des modèles pour repérer des changements dans les connaissances ou la terminologie d'un domaine.

6. 5 - Publications sur ces travaux[25]

6.5.1 Publications sur les outils de TAL et les méthodes de construction de RTO

[IC, 99] P. SEGUELA, N. AUSSENAC-GILLES, Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. *Actes de la conférence IC'99 (Ingénierie des connaissances) - Plate-forme AFIA*. Palaiseau (F), 14-18 Juin 1999. 79-88.

[TIA-Séguéla, 99] SEGUELA P., Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés, in *Actes de TIA'99 (Terminologie et Intelligence Artificielle),* Nantes, *Terminologies Nouvelles* n°19, pp 52-60, 1999.

[CG, 00] AUSSENAC N., SEGUELA P., Les relations sémantiques : du linguistique au formel. *Cahiers de grammaire*, Numéro spécial sur la linguistique de corpus. A. CONDAMINES (Ed.). Toulouse : Presse de l'UTM. Vol. 25. 175-198. Déc. 2000.

[IC, 00] AUSSENAC-GILLES N., BIEBOW B., SZULMAN S., Modélisation du domaine par une méthode fondée sur l'analyse de corpus. *Actes de la conférence d'Ingénierie des Connaissances IC'2000*, Toulouse, mai 2000.

[EKAW, 00] AUSSENAC-GILLES N., BIÉBOW B., SZULMAN N., Revisiting Ontology Design: a method based on corpus analysis. *Knowledge engineering and knowledge management: methods, models and tools, Proc. of the 12th International Conference on Knowledge Engineering and Knowledge Management*. Juan-Les-Pins (F). Oct 2000. R Dieng and O. Corby (Eds). Lecture Notes in Artificial Intelligence, Vol. 1937. Berlin: Springer Verlag. 2000. 172-188.

[OT-a, 00] AUSSENAC-GILLES N ., BOURIGAULT D., The Th(IC)2 Initiative : Corpus-Based Thesaurus Construction for Indexing WWW Documents. *in Proc. of the EKAW'2000 workshop « Ontologies and texts »*. Juan-Les-Pins (F). Oct. 2, 2000. 71-78. http://www.ceur-ws.org/Vol-51/

[OT-b, 00] AUSSENAC-GILLES N., BIÉBOW B., SZULMAN S., Corpus Analysis for Conceptual Modelling. *in Proc. of the EKAW'2000 workshop « Ontologies and texts »*. Juan-Les-Pins (F). Oct. 2, 2000. http://www.ceur-ws.org/Vol-51/

[TIA, 01] SZULMAN S., BIEBOW B., AUSSENAC-GILLES N., Vers un environnement intégré pour la structuration de terminologies : TERMINAE. *4° rencontres internationales Terminologie et Intelligence Artificielle.* Nancy, Mai 2001. Nancy : Unité de recherche et Innovation-INIST-CNRS. 98-108.

[Thèse-SEGUELA, 01] SEGUELA P., Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques. *Mémoire de Thèse en Informatique de l'Université Toulouse III*. Mars 2001.

[TAL, 02] SZULMAN, S., BIEBOW B., AUSSENAC-GILLES N., Structuration de Terminologies à l'aide d'outils d'analyse de textes avec TERMINAE. *Traitement Automatique de la Langue (TAL).* Numéro spécial « Structuration de Terminologie ». Eds A. Nazarenko, T. Hammon. (43). N°1. Hermès : Paris. 2002. pp 103-128.

[EON, 02] AUSSENAC-GILLES N., TERMINAE, an experiment contribution. EON2002, workshop on Evaluation of Ontology-based Tools associated to EKAW2002, Sigüenza, Spain. Oct. 2002. pp 112-128. http://www.CEUR-WS.org/Vol-62

[GDR I3, 02] N. AUSSENAC-GILLES, A. CONDAMINES, SZULMAN S., Prise en compte de l'application dans la constitution de produits terminologiques. *Actes des 2e Assises Nationales du GDR I3,* Nancy (F), Déc. 2002. Toulouse : Cépaduès Editions. Pp 289-302.

[25] Présentations par ordre chronologique.

[OLT, 02] N. AUSSENAC-GILLES, A. MAEDCHE (Eds.). *Proceedings of the ECAI 2002 Workshop about Natural Language Processing and Machine Learning for Ontology Engineering*. Lyon. July 22-23 2002. http://www.inria.fr/acacia/OLT2002

[Rapport-STGOBAIN, 02] N. AUSSENAC-GILLES AND A. BUSNEL. *Méthode de construction à partir de textes d'une ontologie du domaine de l'industrie de la fibre de verre.* Rapport intermédiaire, contrat de recherche entre IRIT et Saint-Gobain Recherche. Rapport Interne IRIT/2002-11-R. Avril 2002. 129 p.

[Rapport-VERRE, 02] N. AUSSENAC-GILLES AND A. BUSNEL. *Méthode de construction à partir de textes d'une ontologie du domaine de l'industrie de la fibre de verre.* Rapport final, contrat de recherche entre IRIT et Saint-Gobain Recherche. Rapport Interne IRIT/2002-28-R. Sept. 2002. 190 p.

[IC, 03] AUSSENAC-GILLES N., BOURIGAULT D., TEULIER R., Analyse comparative de corpus : cas de l'ingénierie des connaissances. *Actes de IC2003 (14° journées Francophones d'Ingénierie des Connaissances)*. Présidente : R. Dieng-Kuntz. Laval (F), 1-3 Juillet 2003. Presses Universitaires de Grenoble. pp 67-84

[TALN, 03] D.BOURIGAULT, N. AUSSENAC-GILLES, Construction d'ontologies à partir de textes. *Actes de TALN 2003*, Batz, Juin 2003.

[TIA, 03] AUSSENAC-GILLES N., BIEBOW B., SZULMAN S., D'une méthode à un guide pratique de modélisation de connaissances à partir de textes. *5e rencontres Terminologie et IA, TIA 2003*. Ed. F. Rousselot. Strasbourg (F), ENSSAIS, Avril 2003. pp 41-53.

[RIA, 04] BOURIGAULT D., AUSSENAC-GILLES N., CHARLET J. Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle (RIA*). Numéro spécial sur les Techniques Informatiques et Structuration de Terminologies. PIERREL J.M. et SLODZIAN M. (Ed.). Paris : Hermès. 18 (1) : 87–110. 2004.

[Livre-IC, 04] AUSSENAC-GILLES N., BIEBOW B., SZULMAN S., Modélisation du domaine par une méthode fondée sur l'analyse de corpus. *In Ingénierie des Connaissances*. R. Teulier, P. Tchounikine et J. Charlet Eds. Paris : L'harmattan. 2004.

{DAGSTUHL, 05] AUSSENAC-GILLES N., Supervised text Analysis for Ontology and Terminology Engineering. In *Machine Learning for the Semantic Web*, Dagstuhl Seminar 05071. Dagstuhl (Germany), 13-18 feb. 2005.

[AO, 05] AUSSENAC-GILLES N., SOERGEL D., Text Analysis for Ontology and Terminology Engineering *Applied Ontology*. Amsterdam : IOS Press. 1(1). 2005.

6.5.2 Publications sur modèles conceptuels et accès au contenu de documents

[EKAW, 96] C. GROS, H. ASSADI, N. AUSSENAC-GILLES, AND A. COURCELLE. Task Models for Technical Documentation Accessing. In *European Knowledge Acquisition Workshop, EKAW'96 Complement to the proceedings*, mai 1996. University of Nottingham, Nottingham (UK).

[Rapport-MOUGLIS, 96] AUSSENAC-GILLES AND A. CONDAMINES. *Rapport d'avancement Terminologie, Modélisation des Connaissances et Systèmes Hypertextuels de Consultation de Documentation Technique*. Rapport Interne 96-48-R, IRIT, novembre 1996.

[ISKO, 97] N. AUSSENAC-GILLES AND A. CONDAMINES. Bases de connaissances Terminologiques: enjeux pour la consultation documentaire. In *Actes des 1ères journées du Chapitre Français de l'ISKO*, oct. 1997. Presses Universitaires de Lille. 1997 : 71-98.

[Rapport-MOUGLIS, 97] AUSSENAC-GILLES. Rapport d'avancement - *Mise au point d'un Système de Consultation de Documentation Technique.* Rapport Interne 97-38-R, IRIT, juillet 1997.

[Mémoire-ZERGUINI, 97] ZERGUINI F. Modélisation de la tâche d'un utilisateur en amont d'un système de consultation de documents techniques. Rapport de stage de DESS de Sciences Cognitives, Univ. Toulouse le Mirail. Sept. 1997.

[Mémoire-SERRA, 97] SERRA, H., *Contribution ergonomique à la conception et à l'évaluation d'un outil servant à gérer une BCT (Base de Connaissances Terminologiques)*, Mémoire de DESS de Psychologie Sociale et du Travail, Toulouse. 1997.

[Rapport-MOUGLIS, 98] N. AUSSENAC-GILLES, A. CONDAMINES. Rapport Final de projet GIS Sciences de la cognition - *Terminologie, Modélisation des Connaissances et Systèmes Hypertextuels de Consultation de Documentation Technique*. Rapport Interne. IRIT/98-20-R, IRIT, mai 1998.

[Mémoire-TOURNAIRE, 98] TOURNAIRE A. Rapport d'intervention ergonomique à EDF-GDF : *Modélisation de tâches prescrites pour la construction d'un index destiné à la consultation de documents techniques et évaluation du guide papier MOUGLIS*. Rapport Interne IRIT/98-19-R. Janv. 1998.

[Mémoire-BUSNEL, 01] BUSNEL A. *Rôle des Verbes et Structures Support dans l'élaboration de terminologies spécialisées*. Mémoire de DESS de Traductique et Gestion de l'Information. CRIM-INALCO, Paris. Déc. 2001.

[Rapport-BUSNEL, 01] BUSNEL A. *Analyses Linguistiques de Corpus Techniques pour constituer une Terminologie aidant au Référencement de Sites Web*. Rapport de stage de DESS de Traductique et Gestion de l'Information. Sept. 2001.

[Rapport-CANNESSON, 01] CANNESSON E. *Évaluation et analyse terminologique du site de CLS Argos*. Rapport de stage de DESS en Lexicographie et Terminographie. Université de Lille III, Villeneuve d'Ascq. Juin 2001.

[Rapport-INDEXWEB, 04] N. AUSSENAC-GILLES AND A. CONDAMINES. *Outils et méthodes linguistiques pour l'amélioration de la visibilité des PME sur le Web*. Rapport final des projets financés par le Conseil Régional de Midi-Pyrénées (1999-2004) 99008562 et 01002676. Rapport IRIT/2004-28-R. Nov. 2004. 27 p.

CHAPITRE 7 - SYNTHESE ET PERSPECTIVES : ONTOLOGIES ET DOCUMENTS

Le parcours de recherche que j'ai retracé jusque-là reflète la diversité et la cohérence de mes contributions au domaine de l'ingénierie des connaissances. Ce chapitre les rappelle tout d'abord de manière synthétique (§ 7.1). Il présente ensuite les questions qui structurent mon projet de recherche. Pour chacun des axes ainsi définis (§ 7.2 à 7.4), je mentionne les travaux en cours, les choix retenus pour aborder cette question et le programme envisagé. Je situe enfin ce projet de recherche dans les perspectives du domaine.

7. 1 - Nature de mes contributions

7.1.1 Cohérence et évolution de mes travaux

Tout au long de mes travaux, j'ai choisi de m'appuyer sur la manière dont les connaissances sont exprimées et utilisées par les acteurs d'un domaine, à travers leurs écrits, leur activité ou ce qu'ils en disent. Ce choix conduit à définir des *approches ascendantes* pour construire des modèles de connaissances au sein d'applications informatiques. Pour cela, j'ai proposé et développé des méthodes, des modes de représentation des connaissances adaptés à leur explicitation et à leur gestion, et des logiciels d'aide à l'analyste chargé la modélisation. Je défends en effet la nécessité d'avoir recours à un individu (le cogniticien) qui supervise le processus de construction de modèle et de faire participer les personnes impliquées dans le projet. Par rapport aux acteurs du projet et à ce cogniticien, le modèle joue plusieurs rôles, dont celui de support au dialogue entre humains.

La continuité de mes recherches porte sur la *facette des modèles qui m'intéresse* : il s'agit avant tout de *leur contenu, de leur sémantique* et non leur syntaxe ou les propriétés (expressivité, calculabilité) des formalismes de représentation des connaissances. De ce fait, la validation de ces modèles renvoie à leur utilisation et non à leur seule validité syntaxique ou à leur cohérence formelle. J'ai fait le choix que le contenu des modèles reflète les besoins des utilisateurs et les usages faits des connaissances. Pour cela, leur construction et leur validation supposent la prise en compte des aspects sociologiques, ergonomiques ou organisationnels de la mise au point d'une nouvelle application à base de connaissances, et non seulement des apports technologiques.

Au-delà du contenu des modèles, j'ai mis en avant que *la nature de l'application visée* et le contexte de son utilisation *conditionnent aussi la représentation des connaissances* à choisir, *la méthode et les outils de construction* eux-mêmes. Afin de permettre une adaptation des méthodes et des outils, j'ai défini plusieurs techniques et modes de structuration parmi lesquels choisir. La méthode fournit des recommandations pour les sélectionner et les adapter en fonction des

caractéristiques des connaissances à modéliser et des usages prévus des modèles. La bonne maîtrise de ce paramétrage passe par une série d'expérimentations par type d'application pour mieux définir les choix optimaux dans chaque contexte et travailler en tenant compte des spécificités.

J'ai ciblé successivement plusieurs types d'applications à base de connaissances, ce qui m'a conduite à étudier des modèles conceptuels et des méthodes de construction différents. Ces changements de point de vue vont de pair avec une évolution du domaine de l'ingénierie des connaissances. Après l'espoir mis dans les systèmes experts, et dans la formalisation logique de toutes les connaissances requises pour réaliser une tâche complexe, les limites de ces applications ont conduit à des choix plus nuancés. Les applications actuelles tiennent compte de la richesse et de la diversité des connaissances, de l'impossibilité de retrouver dans un système toutes les facultés humaines, comme l'adaptation par exemple, l'ambiguïté et la souplesse du langage, etc. . Ces applications comportent toujours des modèles plus ou moins formels, en complément de connaissances sous leur forme d'origine lorsque cela est plus pertinent. La formalisation ne se substitue plus aux autres formes de connaissances à la disposition des utilisateurs, dans leurs documents ou outils de travail. Elle vient les compléter, permettre d'y accéder plus rapidement, de les gérer plus efficacement, ou prendre en charge une partie des tâches à réaliser.

Enfin, la continuité de ma recherche est aussi méthodologique. Tous mes travaux ont été menés en collaboration avec des chercheurs d'autres disciplines (psychologues cognitivistes, puis ergonomes et enfin linguistes et spécialistes du traitement automatique des langues). De plus, ils s'appuient systématiquement sur des retours d'expérience, et donc ils supposent de développer des systèmes destinés à un usage réel.

7.1.2 Résultats établis

Mes contributions concernent successivement la modélisation conceptuelle dans son ensemble, à partir de l'analyse de l'expertise, puis l'analyse des textes pour la construction et l'utilisation de ressources terminologiques et ontologiques. Pratiquement, mes travaux comprennent des résultats de trois ordres : des propositions sur la nature des modèles de connaissances et de langages de modélisation conceptuelle ; des techniques et outils de recueil, d'analyse ou de structuration de connaissances ; des méthodes et plate-formes de modélisation. Ces différentes aides à la modélisation étant étroitement liées, je les ai élaborées en cohérence et dans la perspective de proposer des approches unifiées.

Plus précisément, la nature des avancées que constituent ces résultats est analysée à la fin de chacun des chapitres 4, 5 et 6. Le lecteur peut se référer aux bilans de ces chapitres.

7.1.2.1 Modèles et structures conceptuelles

Pour faciliter la lisibilité des modèles et en garantir une interprétation de bonne qualité, j'ai retenu trois principes retenus communs à mes différents travaux : (i) l'intérêt des *visualisations graphiques* des modèles ; (ii) la définition de *langages au niveau conceptuel*, en amont des représentations formelles ; (iii) la *traçabilité* depuis les « données brutes » (retranscriptions d'entretiens par exemple) jusqu'aux connaissances structurées dans le modèle, et donc la conservation de ces connaissances à différents degrés de structuration ou de formalisation.

Parmi les différents rôles que joue un modèle conceptuel, mes travaux ont privilégié les rôles de *représentation partagée* par les acteurs concernés par la modélisation (experts, cogniticien, utilisateur) et représentation *permettant d'exprimer des connaissances*, au détriment de celui de représentation facilitant l'opérationnalisation (langage compréhensible par l'artéfact) ou de métalangage. En effet, que ce soit en général ou pour les ontologies, la richesse d'expression et le potentiel d'interprétation par l'analyste conduisent à une formalisation moins efficace. L'écart entre ces points de vue est suffisamment important pour justifier de les dissocier et de les prendre en compte successivement dans le temps sous forme de plusieurs vues sur le modèle. De ce fait,

dans les ontologies, j'ai laissé de côté le repérage des axiomes et des heuristiques associées aux concepts. Les ontologies que je considère ne comportent que des concepts et des relations (hiérarchiques et autres) entre concepts. Elles sont appelées « light weight ontologies » dans la littérature actuelle (Gómez-Pérez *et al.,* 2003).

La continuité entre ces différents degrés de formalisation est assurée à l'aide de deux représentations reliées. Ainsi, pour les modèles conceptuels, le langage MONA a été proposé en amont de langages opérationnels existants (LISA et ZOLA). Pour les bases de connaissances terminologiques et les ontologies, les modèles définis respectivement dans les environnements GEDITERM et TERMINAE, préparent la structuration mais ne permettent pas d'autre traitement que la classification de concepts. Ma contribution à TERMINAE a été de proposer la mise en forme d'un modèle conceptuel avant la formalisation en logique de description déjà présente dans la version développée par B. Biébow et S. Szulman. L'originalité de ces langages par rapport aux représentations des ontologies dans l'état de l'art est d'intégrer une *composante terminologique* associée à un réseau conceptuel. Dans les deux cas, la *traçabilité* qui va des entretiens retranscrits ou des textes vers les modèles conceptuels puis formels est assurée par le biais des différentes structures de connaissances qui sont mises en correspondance : des tâches opérationnelles aux tâches conceptuelles puis aux cas dans MACAO-II ; des concepts aux termes puis aux textes dans TERMINAE.

Partant de l'étude de la construction des modèles conceptuels, où s'articulent modèle du domaine et modèle du raisonnement, je me suis sommes centre sur le domaine avec des modèles comme les bases de connaissances terminologiques et les ontologies. Au cours de ces deux périodes, le *critère de pertinence du modèle a évolué* et ce, de manière analogue. À la vue cognitiviste adoptée dans MACAO a succédé une vue constructiviste dans MACAO-II, où le modèle est une construction artificielle validée par le bon fonctionnement du système. Ce changement est à rapprocher d'un glissement de point de vue sur les ressources terminologiques et ontologiques (RTO). Alors que les bases de connaissances terminologiques étaient posées a priori comme des modèles du contenu de textes, avec TERMINAE, les RTO sont reconnues comme le fruit d'une interprétation finalisée.

Ainsi, les modèles construits, qu'ils proviennent d'analyse de textes ou d'activités d'experts, ne sont pas des représentations neutres et objectives d'un domaine ou d'une tâche, mais des constructions venant répondre à des besoins précis. Leur contenu et la manière de les construire sont liés à l'application visée.

7.1.2.2 Techniques et logiciels pour le recueil et l'analyse de connaissances

Pour identifier le contenu des modèles conceptuels, j'ai étudié successivement deux types de source de connaissances liées au système à développer : les activités d'experts et les textes relatifs à la tâche et au domaine du futur système.

Dans le cadre de la méthode MACAO, j'ai inventorié et organisé plusieurs techniques pour le recueil d'expertise : entretiens centrés, simulations d'études de cas suivies de demandes de justification, grilles répertoires, analyse de l'activité et de la tâche. Ces différentes techniques sont décrites dans le guide méthodologique de MACAO-II, qui indique comment les mettre en œuvre.

Pour la construction de modèles à partir d'analyse de textes, plusieurs techniques et logiciels ont été évalués ou définis. La multiplication d'expériences et la collaboration avec l'auteur des extracteurs de termes LEXTER puis SYNTEX montre la richesse de ces logiciels. Leur intérêt réside autant dans la diversité et la qualité des résultats obtenus que dans la pertinence de l'interface de consultation, qui permet de dépasser le problème de l'analyse de la grande quantité de données tirées des textes. Un des enseignements tirés de leur utilisation est l'atout de *traiter simultanément l'identification de termes et de relations lexicales* pour aller vers la définition de concepts. Les concepts ne sont pas définis uniquement sur la base d'éléments quantitatifs (fréquence,

productivité) ou qualitatifs (termes voisins, classes sémantiques) mais aussi sur le fait qu'ils sont en relations syntaxiques et sémantiques avec d'autres termes.

Le logiciel CAMELEON apporte une aide à l'extraction de relations sémantiques selon une approche par patrons lexico-syntaxiques. Les points forts de ce logiciel sont sa *base de marqueurs génériques*, qui peut être enrichie à partir de chaque projet, la possibilité *d'ajuster des marqueurs au corpus étudié,* celle de définir de nouveaux marqueurs propres à un corpus, l'aide au repérage des concepts mis en relation à partir des termes présents dans le contexte de chaque occurrence du marqueur. Enfin, CAMELEON dissocie clairement la mise au point et la validation des marqueurs de l'évaluation de leurs occurrences en corpus pour décider ou non de définir des relations conceptuelles. Du point de vue des ontologies elles-mêmes, ce travail sur les relations contribue à mieux expliciter leur sémantique, et permet de disposer d'indices pour les repérer dans la langue.

7.1.2.3 Méthodes ascendantes et plates-formes de modélisation de connaissances

Mes propositions relèvent toutes d'*approches ascendantes* pour rendre compte des modes de raisonnement humain, spécifier le raisonnement du système final, rendre compte des connaissances révélées par l'usage du langage et en dégager des ontologies ou des terminologies.

Ce processus de modélisation est considéré nécessairement comme *itératif et supervisé par un cogniticien*. Il comporte des tâches de caractérisation de haut niveau des éléments constitutifs du modèle qui guident ensuite le recueil de nouvelles connaissances. Ces tâches correspondent à la mise en œuvre des techniques proposées par la méthode. La méthode intègre des principes de « bonne structuration » comme la normalisation dans le cas des ontologies, qui rendent explicites les choix de modélisation.

La méthode MACAO-II propose d'organiser la modélisation conceptuelle à partir d'études de cas et d'analyse de la tâche. Une première caractérisation des connaissances de résolution de problème débouche sur une représentation abstraite de la résolution de problème et du domaine, appelée *schéma du modèle conceptuel*. Ce schéma guide ensuite le recueil et la modélisation de connaissances. La méthode oriente le cogniticien pour parvenir à une organisation cohérente du modèle du domaine et du modèle du raisonnement via la *notion de rôle*. Les plates-formes MACAO et MACAO-II fournissent un support pour guider des entretiens (grilles répertoires par exemple), en gérer les retranscriptions puis construire progressivement un modèle.

La méthode proposée pour la construction de bases terminologiques (BCT) à partir de textes a fortement influencé mes propositions dans TERMINAE. Le support matériel en est le logiciel GEDITERM, un des rares éditeurs de BCT, qui est avant tout une interface pour définir le modèle selon les structures de représentation des connaissances prévues.

Les résultats produits en matière de modélisation d'ontologie à partir de textes font partie de la méthode TERMINAE. Un des points originaux en est l'utilisation de logiciels de TAL, à considérer comme une boîte à outils à ajuster en fonction de la nature des modèles à construire, et non comme un enchaînement systématique de fonctions automatisées. Avec TERMINAE, l'accent est mis sur la *constitution, l'évaluation et l'organisation du corpus* analysé et les moyens fournis pour assurer la normalisation. La *normalisation* conditionne le fait d'obtenir une ontologie et marque la différence avec d'autres ressources terminologiques. Les spécificités des ontologies ne se situent pas uniquement dans les primitives de représentation des connaissances (concepts, relations, etc.) et leur degré de formalisation (elles doivent être manipulables par un système informatique). Elles résident essentiellement dans la nature de leurs contenus et dans la manière de les déterminer. J'ai ainsi contribué au logiciel associé à cette méthode, développé par S. Szulman et B. Biébow, par trois apports essentiels : l'introduction d'un niveau conceptuel, l'ajout de fiches terminologiques et l'adaptation à la construction de terminologies.

7.1.3 Programme de recherche

À l'issue de ces premiers travaux, mon programme de recherche se trouve à un tournant : après avoir mis essentiellement l'accent sur la construction de modèles en faisant l'hypothèse d'une grande homogénéité des applications visées, je mesure que la récente diversification de ces objectifs oblige à mieux *étudier les liens entre démarche et outils de construction d'une part, et classe d'application ciblée d'autre part*. L'autre tendance, amorcée avec l'étude de la construction de modèles par analyse de textes, est la *prise en compte croissante des documents comme supports de connaissances*, complémentaires des modèles mais ne s'y substituant pas. Enfin, étudiés conjointement, ces deux problèmes conduisent à celui de la *maintenance des modèles* de type ontologie en cohérence avec les documents auxquels ils sont associés, parce qu'ils en modélisent en partie le contenu, servent à les indexer ou à les caractériser.

Cela soulève deux premières questions : celle de *la légitimité des modèles* par rapport aux documents qui servent à les construire, de leur valeur en tant que connaissances ; celle des *applications ciblées par l'ingénierie des connaissances*, qui manipulent des modèles de connaissances seulement. Au-delà de la construction de modèles (ressources termino-ontologiques), l'association entre ces modèles et des documents (essentiellement des textes) renvoie non seulement à l'indexation, l'annotation, mais aussi à l'accès au contenu ou sa représentation. Il s'agit d'exploiter au mieux les complémentarités du texte et d'une représentation structurée ou formelle, tant du point de vue de leur interprétation que des capacités de traitement par le système et par l'utilisateur. L'objectif est toujours de guider, orienter, préparer des manipulations informatiques, des parcours de lecture ou même des interprétations. Une troisième question concerne le *degré d'opérationnalisation des mo*dèles pour un type d'application donné. Faut-il systématiquement aller jusqu'à la construction d'une ontologie pour répondre à ces besoins documentaires ou d'échange d'information ? Ne peut-on pas jouer autrement de la complémentarité entre des logiciels de recherche d'information, de repérage de connaissances à partir d'indices liés à la langue, à la mise en forme, etc. d'une part, et des représentations structurées ou formelles d'autre part ? À l'heure du web sémantique, cette question n'est pas triviale. Le fait même de la poser remet en question le discours actuellement dominant qui met en avant la formalisation et le raisonnement logique.

La continuité de mon projet de recherche réside dans les types de solution étudiés : méthodes, logiciels et modèles de données. Elle concerne aussi les objets étudiés : modèles de connaissances (ressources terminologiques ou ontologiques) et textes. Elle se situe enfin au niveau de ma démarche scientifique : privilégier une approche pluridisciplinaire, faire appel particulièrement au traitement automatique des langues, à la linguistique et à l'ergonomie, s'appuyer sur l'étude des connaissances en usage et les restituer, et surtout capitaliser des expériences pour mettre à l'épreuve et valider les propositions faites.

J'identifie plusieurs points originaux dans ce projet de recherche :

- la volonté de **mettre à profit tout ce qui a été appris** ou développé pour la construction d'ontologies (ou de RTO) à partir de textes **au moment de les utiliser pour consulter, annoter des textes et y rechercher des informations** ; jusqu'ici, les solutions présentées dans l'état de l'art dissocient complètement ces deux phases du cycle de vie du modèle ;
- considérer l'annotation de documents comme **aide à une meilleure appréhension de leur contenu,** des connaissances qu'ils véhiculent ; un modèle est alors une sorte de reformulation (figée et plus rigide que le texte lui-même) du texte. Selon les applications, un ou plusieurs de ces modèles peuvent faciliter l'exploitation de ces documents, ou constituer un complément offrant des possibilités d'interprétation différentes.
- aller vers **plus d'intégration et de continuité** : intégration des logiciels d'analyse de sources de connaissances et de construction de modèle au sein de plates-formes ; meilleure continuité dans les étapes du cycle de vie d'une ressource (construction, utilisation et

maintenance) ; exploiter la complémentarité entre la dynamique autorisée par les logiciels d'analyse de textes d'une part et la qualité des ontologies ou des modèles formels.

- **tester d'autres techniques** (agents émergents, statistiques, apprentissage) pour construire des modèles et surtout **anticiper les besoins de maintenance** ou bien, quand cela est possible, pour envisager de nouvelles manières de répondre aux mêmes besoins sans passer par la construction manuelle de modèles (qui sont coûteux, longs à construire et difficiles à maintenir).

L'intérêt de ce cadre est à la fois le foisonnement de questions qu'il rassemble et le renouvellement qu'il autorise sur des questions présentes depuis les débuts de la modélisation conceptuelle. Pratiquement, j'aborderai ces points selon trois axes de recherche qui répondent aux questions soulevées en conclusion du chapitre 6 :

1. **valorisation** et évolution de logiciels et méthodes pour la modélisation à partir de textes (CAMELEON et TERMINAE) ; il s'agit là de finaliser les travaux précédents ;
2. **ontologies et gestion de documents** : apport des ontologies à différents types de recherche d'information, que ce soit à l'aide d'un moteur général, au sein de documents structurés ou utilisation du TAL pour **l'annotation et l'indexation sémantique de documents** ; je traite ces questions depuis les deux dernières années à travers de nouvelles collaborations avec des équipes en recherche d'information ;
3. **maintenance d'ontologies et de terminologies en environnement dynamique**, c'est-à-dire dans lesquels les textes, la terminologie ou les connaissances du domaine changent.

Au cœur de mes perspectives se trouvent ces deux derniers axes : indexation sémantique de documents d'une part, maintenance d'ontologies et de terminologies d'autre part. L'originalité de mon projet est de traiter ces deux questions conjointement, et d'en chercher des solutions cohérentes s'appuyant sur les outils et méthodes utilisés pour la construction d'ontologies à partir de textes.

7. 2 - Construction d'ontologies à partir de textes : valorisation des logiciels

7.2.1 De CAMELEON à l'exploration de corpus étiquetés

7.2.1.1 Travail en cours ou à court terme

Un des axes prioritaires de mes travaux actuels est la finalisation d'une version de CAMELEON qui soit facile à diffuser, accompagnée d'une base de marqueurs valides et adaptés au fait de s'appliquer à des corpus étiquetés. Actuellement, cet objectif se heurte à plusieurs obstacles :

- Les limites du concordancier YAKWA utilisé pour définir puis projeter des patrons de fouille. YAKWA ne permet pas de retrouver facilement les termes mis en relation par un patron dans le cas où celui-ci contiendrait un « joker », c'est-à-dire une indication de mot d'un type quelconque.
- L'absence d'une base de marqueurs au format requis par la nouvelle version du logiciel. En effet, la base de marqueurs mise au point par P. Séguéla dans sa thèse s'applique à des corpus non étiquetés, et son format ne convient plus.

Pour dépasser ces difficultés, un nouveau concordancier est en cours de développement à l'aide de logiciels libres. La bibliothèque de programmes d'exploration de textes MDROS a été

utilisée à cet effet. Par ailleurs, une terminologue reformulera les marqueurs inventoriés par P. Séguéla selon ce nouveau format. Enfin, afin de favoriser l'information sur ce logiciel, il sera présenté sur la plate-forme « recherche d'information » RFIEC de l'IRIT comme un des outils possibles d'exploration de textes.

7.2.1.2 Questions et programme de recherche

L'approche par patron s'avère tout à fait intéressante lorsque l'on maîtrise les informations recherchées, lorsque le corpus contient des formulations explicites de définitions ou d'autres relations. On sait désormais qu'elle convient bien pour exploiter des livres [rapport-VERRE, 02] ou des documents pédagogiques, mais aussi des documents comportant des régularités même s'ils ne sont pas grammaticalement corrects. Dans le premier cas, des patrons génériques peuvent être utilisés après adaptation. Dans le deuxième cas, les patrons sont essentiellement propres au corpus.

Cependant, cette approche présente l'inconvénient d'être coûteuse en temps et finalement d'une productivité très inégale, souvent faible. Pour aller vers plus d'efficacité, certains systèmes appliquent automatiquement des patrons robustes, sans retouche de la part de l'analyste. Je souhaite continuer à explorer cette approche car elle est précise et permet une étude fine des textes en lien avec un modèle à construire ou existant.

Plutôt que l'automatisation, mes différentes expériences soulignent d'autres pistes pour améliorer une recherche par patrons, qui pourraient conduire à des évolutions dans le logiciel CAMELEON. Je les liste ici, en précisant celles qui seront explorées en priorité :

- *Favoriser une interaction coopérative pour ajuster et réviser des connaissances trouvées* par des patrons avec CAMELEON. À ce jour, les processus de mise au point des patrons et d'évaluation des résultats de leur projection sont laissés à l'initiative de l'utilisateur. Une aide coopérative pour l'analyste serait de lui fournir des repères pour mieux interpréter les formes lexicales retournées (et y trouver les concepts concernés), pour l'orienter vers des résultats plus pertinents ou plus précis en priorité, ou encore pour « simuler » les impacts de l'ajout d'une relation dans le modèle conceptuel. En collaboration avec l'équipe SMAC de l'IRIT, ce module, présenté au chapitre 7 (maintenance des ressources termino-ontologiques) sera développé selon une approche par émergence, à l'aide d'un système multi-agent adaptatif.
- *Prendre en compte des facteurs autres que la syntaxe et le lexique pour définir des patrons*. Parmi ces autres indicateurs, je pense à la mise en forme et la structure des documents (Virbel, 2002). Ainsi, dans le projet Arkeotek, des relations explicitées par le titre et la première phrase d'un paragraphe ont été identifiées. Pour cela, il est possible de faire évoluer CAMELEON pour que le format des patrons permette d'exprimer ce genre de caractérisation.
- *Étendre l'approche par patron pour rechercher sur des corpus étiquetés d'autres types de connaissances*, par exemple des relations autres que binaires, des connaissances concernant plusieurs concepts et pouvant se représenter sous forme d'axiomes dans une ontologie. Cette perspective à moyen terme suppose une collaboration avec l'ERSS pour orienter dans cette direction les modes d'utilisation d'un environnement d'exploration de corpus étiqueté qui est en cours de développement par ce laboratoire.
- *Combiner des approches linguistiques et statistiques pour l'identification de concepts, d'instances et de relations sémantiques*. J'ai déjà expérimenté l'intérêt des verbes extraits par SYNTEX et des résultats de l'analyse distributionnelle comme amorces de patrons de relations lexicales. Il s'agirait de mieux associer ces deux analyses au sein d'une plate-forme de modélisation. À plus long terme, je voudrais explorer deux autres types d'approches qui sont aujourd'hui les plus innovantes en matière de construction d'ontologies à partir de textes : l'extraction d'information et l'apprentissage automatique.

7.2.2 De TERMINAE à une plate-forme de modélisation à partir de textes

7.2.2.1 Programme

À ce jour, un des points faibles des propositions en matière de modélisation à partir de textes est le manque d'intégration des différents logiciels de TAL et de représentation des modèles. Les difficultés d'échange de résultats d'un logiciel à l'autre, le fait qu'il s'agisse soit de prototypes universitaires soit de produits propriétaires, rend leur utilisation conjointe peu commode et réduit la possibilité de les diffuser. Avec plusieurs équipes de recherche françaises ayant développé des outils complémentaires (DOE, TERMINAE, SYNTEX et CAMELEON), je souhaite les intégrer au sein d'une plate-forme unique et facile à diffuser, dans l'esprit de la solution offerte par TERMINAE. Il s'agit de faciliter l'enchaînement de leur utilisation en reprenant les principes méthodologiques de TERMINAE et de DOE pour expliciter les choix de différentiation entre concepts. Enfin, l'utilisation de ces logiciels doit pouvoir être paramétrée en fonction du type de modèle à construire ou du type d'application visé. Pour ce dernier point, il faut poursuivre un effort de capitalisation d'expériences, de mise en forme, d'analyse critique et de restitution de la manière de procéder pour des projets particuliers. Une demande de financement pour ce projet a été déposée en mars 2005, en collaboration avec l'INA, le LIPN, l'ERSS et l'AP-HP.

7.2.2.2 Questions de recherche

Dans l'optique de la mise en place de cette plate-forme, les questions de recherche couvrent des aspects méthodologiques liés à l'utilisation des logiciels d'analyse du langage.

Ce cadre interroge sur la manière de mieux exploiter la complémentarité entre les compétences d'experts, l'analyse de textes et la réutilisation de modèles existants. Jusqu'ici, j'ai envisagé des solutions ad hoc, et la méthode TERMINAE ne propose rien en matière de réutilisation d'ontologie ou de terminologie pour un nouveau projet. Or la réutilisation et l'adaptation de modèles existants sont quelques-uns des moyens de réduire les coûts, de s'adapter aux normes en cours ou encore de bénéficier d'études déjà validées. Ce problème est proche de celui de la maintenance d'un modèle par l'ajout de connaissances extraites d'un corpus : les connaissances tirées de l'analyse de textes et le nouvel objectif d'utilisation obligent à modifier la ressource qui est réutilisée.

À l'articulation entre méthode et logiciels, une des questions est de proposer une gamme plus large d'outils de TAL, ce qui soulève le problème de leur intégration dans des chaînes de traitement. On peut considérer que chaque logiciel est indépendant et que l'ensemble est modulaire, dans l'esprit des plug-in qui font le succès du logiciel Protégé. Dans ce cas, le cogniticien sélectionne les logiciels qui lui conviennent en fonction de ses besoins et il les organise dans une chaîne de traitement. Une autre possibilité serait d'abandonner la perspective de plates-formes générales pour s'orienter vers des solutions adaptées par type d'application ou de modèle à construire. Pour le moment, dans la lignée de TERMINAE, j'envisage une méthode générale.

7.2.2.3 Travaux concernés : lien entre méthode, type de modèle et application visée

Du point de vue méthodologique, je prévois de poursuivre la mise en forme d'indications méthodologiques propres à des classes d'applications à partir de chacune des expériences de modélisation. L'approche retenue consiste à capitaliser des retours d'expérience pour en dégager des indications sur la nature du modèle adapté à un type d'application, et sur la manière de le construire. Quand cela sera possible, ces propositions seront ensuite intégrées dans l'approche TERMINAE pour la faire évoluer.

De plus, une réflexion sera menée sur l'articulation dynamique entre connaissances du domaine et modèles de tâche. Ce sujet était laissé de côté depuis nos travaux sur MACAO. Or il s'agit là d'expliciter un engagement ontologique et de rendre compte de l'impact de la tâche

réalisée par le système sur la manière de structurer les connaissances du domaine dans l'ontologie. Une perspective d'étude porte sur les modèles de tâches tels qu'ils sont définis dans notre équipe pour programmer les systèmes coopératifs (Camilleri, 2003). L'autre se situe dans le cadre du laboratoire Autodiag, pour une application d'aide au diagnostic de pannes électroniques, l'ontologie étant utile au raisonnement de diagnostic (thèse de A. Reymonet).

Enfin, toujours avec la volonté de capitaliser des éléments méthodologiques à partir des expériences menées, je cherche à mettre en évidence des critères pour qualifier la nature des modèles (ontologies, réseaux terminologiques …) requis par type d'application. Je conduis actuellement quatre expériences qui donnent lieu à des applications suffisamment différentes pour permettre déjà d'en tirer des indices significatifs. Je les présente dans les parties qui suivent.

7.2.3 Situation par rapport aux perspectives du domaine

Il n'existe pas de plate-forme traitant de construction d'ontologie à partir de textes pour le français couvrant tout le processus et permettant d'enchaîner traitements linguistiques, analyse de textes, et modélisation conceptuelle. Il existe une plate-forme de ce type pour l'anglais, KAON, et des projets moins complets mais de plus en plus sophistiqués : plate-forme de K. Ahmad par exemple pour l'anglais, plug-ins associés à Protégé-2000 développé par P. Buitelaar. Ceci est d'autant plus regrettable que la position des chercheurs français, grâce aux activités du groupe TIA, a été originale et perçue comme innovante entre 1995 et 2002. Cette plate-forme est attendue tant au niveau pratique, pour des projets opérationnels ou de recherche, qu'au niveau scientifique national et international, comme preuve tangible de l'efficacité des propositions de TIA.

Aujourd'hui, le groupe TIA continue de défendre un point de vue encore marginal en portant un regard critique sur le point de vue classique. Ainsi, l'approche classique défend une représentation rigide, fixe et très stable de concepts et de termes pour traduire les connaissances et la terminologie d'un domaine. A contrario, le groupe TIA défend une sémantique textuelle et différentielle, qui pose d'emblée la notion de concept et le sens des termes comme des éléments fluctuants en fonction des interprétations, des contextes, de la dynamique d'un domaine.

En développant une plate-forme complète qui permette d'assurer une suite de traitements continus, et en poussant plus loin la validation et l'intégration de CAMELEON dans cette plate-forme, je m'oriente vers la valorisation de cet acquis du groupe TIA, vers l'expérimentation systématique des propositions méthodologiques du groupe et en particulier de TERMINAE. Mais cette plate-forme pourrait être aussi un support pour continuer à innover et mieux restituer la dynamique liée à l'identification de la terminologie et des connaissances d'un domaine. En effet, une fois défini l'enchaînement de briques d'analyse élémentaire, il sera plus facile d'imaginer différents paramétrages :

- Des alternatives aux approches actuelles, par exemple en introduisant plus d'apprentissage ; au niveau européen, l'utilisation de méthodes d'apprentissage à partir de textes pour compléter des ontologies (*ontology learning*) est devenue depuis 2002 un des points les plus nouveaux en matière de construction d'ontologies ; l'apprentissage sert également à retrouver dans des textes des traces linguistiques de la présence de concepts ou d'instances de concepts (*ontology population*) ; ce procédé est utilisé pour apprendre à annoter automatiquement ou indexer des textes à l'aide de concepts (Buitelaar *et al.,* 2005).
- La prise en compte des caractéristiques du corpus et de l'objectif visé dans la manière d'utiliser les logiciels ; plusieurs retours d'expériences sont capitalisés au niveau national : modélisation d'index à partir de textes (Ait el Mekki *et al*., 2002), modélisation d'ontologies formelles, de terminologies etc. qui soulignent la nécessité d'ajuster les outils et les principes à chaque classe d'application.
- L'évaluation précise du rôle de chaque composant, de leur contribution ainsi que de leur coût. La question de l'évaluation des modèles et des outils utilisés pour leur construction a été

retenue comme une des perspectives importantes dans le domaine par l'AS ASSTICCOT. La mise à disposition d'une plate-forme unificatrice peut favoriser ces évaluations.

À ce jour, la tendance au niveau international est de rechercher une automatisation complète du processus de construction d'ontologies à partir de textes, en faisant une part de plus en plus importante à des traitements automatiques et à l'apprentissage automatique (Buitelaar *et al.,* 2005). Malgré le degré élevé de pertinence de ces logiciels, la position que je défends est de maintenir un humain pour superviser ce processus, faire des choix et intégrer la prise en compte de la finalité d'utilisation du modèle au moment de le construire.

Enfin, ce projet de valorisation et intégration des outils de construction d'ontologie à partir de textes tient peu compte des travaux récents sur les ontologies formelles et les fondements théoriques de l'ontologie. L'IRIT ayant constitué un laboratoire international ILIKS avec plusieurs laboratoires italiens et particulièrement avec l'équipe de N. Guarino, je serei amenée à me situer par rapport à ses travaux. Il serait important de parvenir à intégrer dans ma proposition des travaux théoriques sur l'ontologie formelle, par exemple sur les aspects méthodologiques de construction et de vérification des ontologies formelles ; et inversement, poser au niveau théorique les questions qui surgissent à partir de la construction pratique d'ontologies pour des applications.

7. 3 - Ontologies, documents et recherche d'information

Pour confirmer l'intérêt réel des ontologies dans différents contextes applicatifs, il faut préciser les conditions de leur pertinence, les modalités de leur utilisation mais aussi les contraintes que cela impose sur leur construction. Mon choix est d'étudier conjointement ces deux facettes, construction et utilisation, dans un double objectif :

- Un objectif méthodologique : comment construire des ontologies possédant les bonnes caractéristiques pour des applications particulières ? comment associer les problématiques de la construction et de l'utilisation ? en quoi les outils de traitement automatique des langues et les approches adoptés pour tirer des ontologies à partir des textes permettent aussi d'associer ces ontologies à de nouveaux textes pour les annoter, les indexer, etc. ?
- Un objectif appliqué : confirmer les modalités d'utilisation des ontologies pour différents types d'applications : quel doit être leur degré de formalisation, leur couverture du domaine, leur degré de précision, leur richesse terminologique ? Il s'agit là de questions de recherche fondamentales pour l'étude des ontologies en tant que modèles utiles pour des applications. Je m'attends bien sûr à des réponses diverses en fonction des applications visées.

Pour atteindre ces deux objectifs, j'ai choisi de me focaliser sur les applications documentaires et la recherche d'information dans des textes. En effet, pour ces types d'applications, la construction de modèles à partir de textes se justifie tout à fait (TIA, 2002). Cela se confirme par le très fort attrait de la communauté recherche d'informations pour les ressources terminologiques depuis 1995 et pour les ontologies depuis 1998. L'intérêt des modèles conceptuels, des ontologies ou des bases lexicales pour améliorer les réponses fournies par ces systèmes a été discuté et évalué dans différents travaux. Mon objectif est d'étudier différents cas particuliers avec l'originalité d'insister sur l'analyse de la langue, la dimension terminologique et les logiciels de TAL. J'ai commencé à aborder ces questions dans plusieurs projets, dont certains menés en collaboration avec l'équipe SIG[26] de l'IRIT et l'ERSS, spécialisée en recherche d'information, et dans le cadre des plates-formes régionale RFIEC[27] et nationale PLEXIR[28]. Mon programme s'organise en trois parties complémentaires :

[26] Systèmes d'Information Généralisés

[27] http://www.irit.fr/RFIEC/ Ce groupe de travail s'est mis en place depuis 2001, faisant intervenir les équipes SIG (recherche d'information), CSC (construction d'ontologies), SMAC (Systèmes Multi-agents Coopératifs) et LilAC

1. Évaluer *l'apport des ontologies à deux types d'applications de recherche d'information* : la recherche au sein d'une collection) l'aide d'un moteur généraliste et la classification documentaire selon des préférences d'utilisateurs.
2. Étudier les modalités d'*utilisation des modèles (ontologies et terminologies) et des logiciels de TAL utilisés pour les construire dans le but d'annoter ou indexer sémantiquement des documents*. Le problème est d'outiller efficacement l'association de concepts à des textes ou à des passages de textes.
3. Étudier cet apport et ces modalités dans le *cas particulier de l'analyse de documents structurés* ; au-delà du lexique, on envisage de tirer profit d'autres types d'informations disponibles pour donner du sens au contenu des textes : la structure des documents, leur mise en forme matérielle, ou encore le typage sémantique (XML) des paragraphes.

7.3.1 Ingénierie des connaissances et recherche d'information : convergences

Les préoccupations des communautés de recherche d'information et d'ingénierie des connaissances se recoupent actuellement autour d'un ensemble de problèmes liés à l'accès au contenu de documents sur support numérique, à leur gestion, leur caractérisation et leur représentation symbolique et formelle afin de faciliter les manipulations informatiques. Cette convergence se matérialise autour des applications documentaires (moteurs de recherche d'information, systèmes de question-réponse, classification de documents, recherche au sein de bases documentaires thématiques, indexation, etc.) locales ou sur le web. Au cœur de la convergence, se trouve la notion d'ontologie, qui se décline sous diverses formes pour faire référence à des ressources plus ou moins terminologiques ou ontologiques.

Dans ce cadre, on attend des ontologies et des modèles assimilés de permettre de caractériser le contenu (la sémantique) des documents et des besoins des utilisateurs. Ainsi, les ontologies peuvent fournir de nouvelles représentations plus riches, plus précises et plus efficaces des documents et des besoins des utilisateurs que selon les approches classiques en recherche d'information. La dissociation entre termes et concepts autant que l'organisation des concepts en un réseau sémantique permet d'étendre ou de reformuler automatiquement des requêtes pour retrouver des documents n'utilisant pas exactement les termes de l'utilisateur mais répondant à sa recherche d'information. On espère ainsi retrouver plus efficacement les documents pertinents ou localiser l'information pertinente précisément au sein de documents structurés. L'ontologie est envisagée aussi pour mieux expliciter les centres d'intérêt des utilisateurs ou encore les paramètres d'utilisation des systèmes (support matériel, contexte, type d'utilisateur, etc.).

7.3.2 RTO pour optimiser les résultats d'un moteur de recherche

La recherche d'information s'intéresse à l'apport de ressources sémantiques (ou lexicales) aux moteurs de recherche généraux. Dans ce cas, les ressources utilisées doivent couvrir la langue générale. On attend de leur utilisation une amélioration des performances (rappel et précision) des moteurs. Plusieurs travaux ont montré les limites de cette hypothèse : l'utilisation non contrôlée de ressources comme la base de données lexicales WordNet peut au contraire dégrader les résultats car la multiplication des concepts génère du bruit. À partir d'une étude précise des raisons de ces

(Ontologies formelles) ainsi que l'opération « sémantique et corpus » du laboratoire de linguistique ERSS. L'objectif est de mener diverses expérimentations afin d'identifier les apports éventuels des logiciels de TAL ainsi que des ressources ontologiques et terminologiques à différentes applications de recherche d'information documentaire

[28] http://www.irit.fr/PLEXIR/

échecs, il a semblé intéressant de mener d'autres expériences basées sur d'autres utilisations de WordNet.

Afin d'évaluer l'apport de modèles conceptuels à la recherche d'information en amont d'un moteur de recherche, deux études ont été menées successivement. L'une porte sur la reformulation de requêtes en exploitant les relations entre concepts, l'autre sur la représentation de documents sous forme d'un réseau de concepts. Ce travail correspond au DEA puis à la thèse de M. Baziz, que je co-encadre avec M. Boughanem (équipe SIG de l'IRIT) depuis septembre 2002.

Malgré certains défauts mis en avant par les terminologues (qualité du réseau sémantique, vocation essentiellement lexicale et non conceptuelle), la base de données lexicales WordNet a été choisie pour des raisons pratiques. Elle offre l'avantage d'être disponible facilement et de couvrir la langue anglaise de manière générale. Cette base se prête bien à l'évaluation de l'apport de ressources lexicales pour explorer avec un moteur de recherche général des corpus de référence (qui sont en anglais) comme ceux des programmes TREC. De plus, la structure de WordNet s'avère pertinente pour les besoins de la recherche d'information : il s'agit d'un réseau de nœuds lexicaux, appelés Synsets, reliés par des relations rendant compte des liens entre termes dans la langue (liens d'hyperonymie, de méronymie, etc.). Ce réseau est riche grâce à la diversité des relations présentes et au vocabulaire associé à chacun des nœuds. Pourtant, WordNet n'est pas une ontologie, entre autres à cause de la sémantique peu précise des relations, que ce soit la synonymie au sein d'un nœud ou entre plusieurs nœuds. De plus, la validité de ce réseau est tout à fait discutable dès que l'on s'intéresse à des domaines de spécialité pointus. Ce qui pourrait sembler un inconvénient ne gêne pas l'utilisation en recherche d'information en langue générale, où la présence d'une relation est aussi importante que sa sémantique.

7.3.2.1 Reformulation de requêtes

Une première expérience a donc consisté à utiliser une base de données lexicale pour la reformulation des requêtes utilisateurs. Pour assurer un gain dans les résultats retournés, un processus d'"expansion prudente" a été défini en amont d'un moteur de recherche. Ce processus, transparent à l'utilisateur, exploite d'abord la notion de concepts multi-termes pour désambiguïser les mots de la requête (au sens de WordNet). Il s'appuie ensuite sur les relations sémantiques entre concepts pour élargir la requête. Différents tests ont été effectués pour évaluer ce processus qui conduit à une amélioration significative de la pertinence des réponses fournies par le moteur. Les expérimentations ont été réalisées en utilisant le moteur Mercure développé à l'IRIT, WordNet comme base de données lexicales et Clef2001 comme collection de test [INFORSID, 03].

Le travail précis mené au cours du DEA de M. Baziz montre que la nature des relations entre les nœuds du modèle conceptuel a une influence significative sur l'expansion de requête [ISI, 03]. Les requêtes étendues, sous certaines conditions, avec des concepts reliés à ceux de la requête initiale par la relation « est-un », permettent d'améliorer les résultats par rapport à la requête d'origine. L'utilisation des relations de méronymie ou d'antinomie au contraire n'améliore pas les résultats. Ces résultats sont originaux car jusque-là, les recherches sur l'apport des ontologies avaient conclu à des apports minimes voire à la dégradation des résultats. La conclusion positive de M. Baziz s'explique par l'étude séparée des différents types de relation ainsi que par la volonté de minimiser le nombre de concepts utilisés pour représenter une requête.

Les modules permettant l'expansion de requête en choisissant le type de relation pris en compte ont été intégrés à la plate-forme de recherche d'information RFIEC. Ainsi, il est possible de reproduire cette expérience sur d'autres corpus.

7.3.2.2 Représentation sémantique de documents

De manière symétrique à l'enrichissement de la requête, un moyen d'améliorer la recherche d'information par des connaissances est de construire une représentation sémantique des documents à parcourir. Dans sa thèse, M. Baziz a choisi de représenter le contenu sémantique de

documents sous la forme d'un réseau de concepts jugés représentatifs du document [VSST, 04]. La difficulté est donc d'identifier automatiquement ces concepts à partir d'une ressource générale, et de déterminer des critères de « représentativité ». L'approche consiste à projeter les documents sur une « ontologie » linguistique générale (ici WordNet). Pour chaque document, les concepts de l'ontologie le *représentant* sont choisis en fonction de critères de co-occurrence (CF.IDF) dérivés de ceux qui sont utilisés pour les termes simples (tf.idf) [SWIR, 04]. Un autre critère de sélection, la proximité sémantique, permet de désambiguïser les concepts candidats via le réseau sémantique de l'ontologie. En effet, un même groupe de mots du texte peut correspondre à plusieurs entrées dans WordNet, et donc à plusieurs concepts. Le choix du concept le plus plausible s'appuie sur les termes voisins en corpus et sur les différents mots entrant dans la définition du concept.

Le résultat de ce "matching" entre le document et l'ontologie est un ensemble de concepts désambiguïsés (appelés aussi concepts-sens ou nœuds) avec des liens pondérés entre eux, formant ce qui a été appelé le *noyau sémantique du document*. Ce noyau est supposé représenter au mieux le contenu sémantique du document. Dans ce noyau, les nœuds représentent des concepts désambiguïsés et les arcs des liens de similarité sémantique calculés à partir de relations présentes dans WordNet. Le calcul de ce noyau, long et coûteux, est fait une fois pour toutes pour une distance donnée. Ainsi, la collection interrogée est représentée par l'ensemble des noyaux sémantiques des différents documents qui la composent.

Lors de la recherche d'information, la requête est traduite sous forme de concepts et étendue selon les principes d'expansion prudente. Elle est ensuite comparée aux différents noyaux sémantiques pour identifier les documents les plus pertinents.

Six distances ont été utilisées pour mesurer la proximité sémantique entre concepts. On a comparé et évalué leur apport sur un jeu de test. Il ressort que la mesure de Resnik est la plus efficace sur la collection utilisée, combinée au calcul du C_Score [SAC-TIAR, 05]. L'ajustement des poids associés aux concepts s'avère d'un impact presque aussi important sur la qualité des résultats que le choix des concepts eux-mêmes. En effet, en recherche d'information, la représentativité des concepts (explicitée par leur pondération) est exploitée pour classer et comparer des documents répondant à une requête.

7.3.2.3 Perspectives

L'approche proposée présente une originalité par rapport aux propositions venant des chercheurs du Web Sémantique, dans la mesure où elle combine à la fois la richesse d'une représentation à base de concepts et les principes de la recherche d'information. En effet, aux nœuds du réseau sont affectés des poids (C_Scores) rendant compte de leur importance par rapport au document. Elle peut être considérée comme une première étape vers l'objectif à long terme qui est l'indexation intelligente et la recherche sémantique du Système de Recherche d'Information (SRI). Le noyau sémantique peut tenir lieu de représentation support pour ce type de recherche.

Le fait de constituer ce réseau de manière automatique est à la fois un atout : il est produit efficacement et de façon transparente pour l'utilisateur. C'est aussi une limite, et il serait utile de vérifier sa représentativité par rapport au contenu des documents. Une représentation graphique de ce noyau montre qu'il contient en général des concepts clés des documents.

Cette étude constitue un premier pas pour explorer l'apport des ontologies à la recherche d'information. Elle permet de bien formuler la question : il ne s'agit pas de savoir si « les ontologies améliorent la recherche d'information (en général) », mais bien « dans quelles conditions les ontologies peuvent améliorer la recherche d'information ? ». Ceci suppose de préciser leur contenu, leur degré de formalisation ou encore leur couverture du domaine. Comme dans l'expérience avec WordNet, il faut aussi définir des heuristiques pour exploiter au mieux les relations entre concepts. Une deuxième étape prévue est de s'intéresser à des domaines spécialisés, avec des corpus et des ontologies spécifiques, pour voir dans quelles conditions l'ontologie constitue une aide.

7.3.3 Classification de documents pour la veille technologique

Une autre classe d'applications en recherche d'information peut bénéficier de l'utilisation d'une ontologie : la classification de documents dans le contexte de la veille technologique. À l'inverse de l'utilisation d'un moteur de recherche, l'activité de veille suppose que l'utilisateur soit un spécialiste du domaine sachant formuler précisément ses centres d'intérêt. Le domaine à couvrir est généralement bien ciblé, ainsi que la nature ou les caractéristiques des documents recherchés. Les approches classiques se basent sur des thésaurus ou des listes de mots-clés du domaine. Les propositions les plus élaborées gèrent des syntagmes et des mots simples, organisés le plus souvent en hiérarchies qui reflètent une spécialisation thématique. Les mots-clés (entrées du thésaurus) renvoient à des sous-entrées sur un ou deux niveaux de profondeur, la relation entre ces termes ayant une sémantique fluctuante et peu précise : « voir aussi », exemples (instances) ou sous-ensembles (relation est-un), etc.

La masse croissante de documents disponibles sous format électronique a permis à ce domaine de se développer. Les recherches en cours définissent des solutions pour faciliter l'expression des préférences et des centres d'intérêt des utilisateurs. Elles proposent des interfaces de navigation dans les collections de documents ou encore des algorithmes pour mieux fouiller leur contenu, le caractériser puis le représenter afin de retrouver rapidement l'information. Pour un domaine donné, il semble indispensable de mettre à plat une représentation sous forme de réseau sémantique autant que d'en identifier la terminologie. Les ontologies sont alors une solution de plus en plus explorée, leur structure permettant de répondre à ces deux besoins.

7.3.3.1 Questions de recherche

Dans ce cadre, les hypothèses de recherche suivantes doivent être évaluées :

- *Tester la pertinence de construire ces ontologies ou hiérarchies à partir de textes*. Ici, on peut se demander si la collection des textes à classer ou indexer forme à elle seule le corpus d'étude. De plus, les textes à classer ne suffisent pas pour obtenir le vocabulaire des utilisateurs, pour savoir comment ils expriment leurs centres d'intérêt et les précisent progressivement. D'autres sources de connaissances doivent être exploitées en complément.
- *Déterminer le type de modèle de données le mieux adapté pour ce type d'application*, parmi la gamme des RTO. D'une part, je veux mesurer l'apport d'une composante terminologique riche associée au réseau des concepts. Plus cette composante est riche, plus le repérage de concepts dans les textes pour les indexer peut être efficace, et donc plus l'analyse des textes en vue de l'indexation est simple. Pour la formulation des préférences des utilisateurs, une terminologie riche peut permettre plus de précision. D'autre part, il semble que cette représentation doit faire cohabiter plusieurs points de vue de focalisation sur le domaine, sous la forme de plusieurs représentations hiérarchiques.
- *Évaluer comment les logiciels de traitement automatique des langues utilisés pour leur construction peuvent faciliter l'indexation des documents par les concepts* ; la manière de définir une indexation par les concepts n'est pas triviale ; en plus de la présence dans un document des termes associés au concept, il faut déterminer comment exploiter les relations entre concepts ou des heuristiques liées à la nature des documents du domaine.

J'aborde les deux premières questions à travers deux projets, l'un terminé (Verre) et l'autre en cours (classification de documents). Je reviendrai sur la troisième question dans la partie suivante, car elle s'avère une problématique majeure de l'utilisation actuelle des ontologies, que j'aborde dans deux autres projets.

7.3.3.2 Problématique de la construction d'ontologies pour la classification documentaire

Suite à la demande d'une entreprise (Saint-Gobain Recherche) désirant mettre en place un système de veille, j'ai été sollicitée pour étudier la faisabilité d'une approche basée sur une ontologie. Ma contribution consistait à définir une méthode de construction d'ontologie qui soit peu coûteuse tout en permettant à l'entreprise de l'appliquer pour assurer ensuite la maintenance de cette ontologie. L'activité concernée par cette application était la veille économique, scientifique et technique dans le domaine de la fabrication de la fibre de verre. Le système utilisant l'ontologie devait assurer le routage de documents électroniques extraits du web vers des profils d'utilisateurs. À plus long terme, l'ontologie devait aussi servir de support à une meilleure communication au sein de l'entreprise sur les thèmes relatifs à ce domaine. En particulier, cette ontologie devrait aider tout nouvel ingénieur à comprendre le vocabulaire des métiers de l'entreprise.

Comme je l'ai montré au chapitre 6, ce projet VERRE a permis d'affiner et de mettre en forme au sein de la méthode TERMINAE des heuristiques utiles au dépouillement terminologique, à l'identification des synonymes ou encore à la définition des concepts par leur mise en relation. À côté de cet aspect méthodologique, ce projet m'a permis de fournir les premières réponses aux questions de recherche que je viens de mentionner [Rapport-VERRE, 02] :

- Concernant le corpus, l'analyse des textes à classer s'avère insuffisante. En effet, ces textes (des brevets et des dépêches de presse) ne contiennent pas assez de termes du domaine. Le corpus a été enrichi de manière à couvrir plusieurs aspects de l'industrie de la fabrication de la fibre de verre : des aspects techniques, des aspects concurrentiels et économiques. Le corpus étudié était donc hétérogène de sorte qu'il a été traité de manière différenciée.
- Le point de vue et la terminologie des utilisateurs ont été peu pris en compte. Les besoins liés à l'activité de veille technologique et économique ont été intégrés via des séances de travail avec des experts de ce domaine et de cette activité. Cela a eu finalement peu d'impact sur la nature du modèle. Cette question restait à approfondir.
- J'ai précisé la nature des modèles utiles dans ce contexte. De cette expérience, il ressort qu'une ontologie dite « light-weigh » (sans axiome) et même non formelle convient à condition de comporter des liens hiérarchiques pouvant être traités par le système opérationnel et surtout de comporter une terminologie riche. Cette composante est un des points forts d'un travail à partir de corpus bien sélectionné.

La manière d'utiliser cette structure de données pour indexer et classer les documents ou pour formuler des profils utilisateurs a été définie par l'entreprise. J'ai eu trop peu de retour de cette expérience pour en tirer des conclusions.

7.3.3.3 Navigation dans des hiérarchies de concepts pour réorganiser des classifications

Cette même classe d'application est étudiée par N. Hernandez dans sa thèse (que je co-encadre avec sa directrice de recherche J. Mothe). Comme dans le projet VERRE, un modèle proche d'une ontologie sert à inventorier et structurer des concepts du domaine sur lesquels les utilisateurs vont interroger un système de veille. Cependant, la place et le rôle de cette « ontologie » y sont envisagés sous un angle assez différent : c'est l'organisation hiérarchique des concepts selon un ou plusieurs points de vue qui joue un rôle privilégié. Par exemple, dans le domaine de l'astronomie, des articles scientifiques peuvent être rassemblés en fonction de critères comme les objets astronomiques dont ils parlent, des instruments de mesure mentionnés, des stations observatoires, des journaux dans lesquels ils sont parus, de leur date ou de leurs auteurs.

Chaque critère définit un point de vue qui organise un ensemble de concepts plus précis. Le choix de plusieurs critères permet de constituer des groupes de documents traitant de sujets plus ou moins précis, d'affiner les classes de documents en fonction des concepts caractérisant leur contenu. Les réponses aux requêtes des utilisateurs sont établies à partir d'une analyse multidimensionnelle (à partir des points de vue exprimés dans les hiérarchies). Ensuite, un

environnement de visualisation, DocCube, présente plusieurs hiérarchies pour faciliter la focalisation sur des documents particuliers et en assurer la consultation rapide. Les hiérarchies de concepts sont vues comme un guide pour naviguer dans l'espace d'information que constitue la collection de documents [INNO, 04]. Dans ce projet, une des questions posées est l'intérêt d'une ontologie offrant un cadre unifié pour associer ces différentes hiérarchies.

À la différence du projet VERRE, le modèle de données de la structure ontologique a été défini de manière ad hoc pour améliorer l'indexation des documents. Ce modèle se veut une avancée par rapport à une représentation classique sous la forme de « sac de mots » [RIAO, 04]. L'intuition derrière ce choix est de réduire les phénomènes classiques qui dégradent les résultats en recherche d'information : ambiguïté de mots polysémiques, différence de vocabulaire entre les textes et les formulations des utilisateurs ou encore mauvaise gestion des variations de forme, des ellipses, etc. Les concepts et le vocabulaire associé servent à décrire l'espace d'information de manière structurée. Ils servent donc aussi de langage pour exprimer le besoin en information. La recherche d'information via ces concepts revient alors à parcourir des vues tirées du modèle (correspondant à un ou plusieurs points de vue), à explorer la collection réorganisée en fonction de concepts choisis par l'utilisateur puis à naviguer entre les documents ou groupes de documents.

Ce travail en cours répond en partie seulement aux questions qui m'intéressent, de manière très complémentaire au projet VERRE :

- Pour élaborer les hiérarchies de concepts, les experts du domaine jouent un rôle primordial. Ce sont eux qui sélectionnent ces points de vue, choisissent les concepts qui les intéressent et les organisent en hiérarchies. En revanche, une analyse d'un corpus représentatif des textes à explorer peut assurer un enrichissement terminologique de ces hiérarchies. Ce point reste à étudier.
- Il est simpliste d'imaginer que le modèle unique qui organise et associe entre elles ces hiérarchies puisse être bâti seulement de manière ascendante, en assemblant les concepts des hiérarchies. En effet, la cohérence de ce modèle requiert une bonne maîtrise du domaine et de prendre un parti permettant d'unifier les différentes dimensions. Ce modèle ne peut être une ontologie que si des principes de structuration ontologique sont appliqués.
- Inversement, j'en retiens qu'une ontologie constitue une représentation du domaine trop complexe et trop riche pour être présentée directement pour guider la formulation de besoins en recherche d'information. En revanche, des hiérarchies de concepts sont des représentations plus pertinentes que des listes de mots car elles guident mieux les utilisateurs. Chaque hiérarchie correspond à une dimension d'analyse ou encore à un point de vue sur le domaine. Elle est représentée par un arbre de concepts reliés selon une relation unique (est-un ou une autre relation). Même si plusieurs points de vue peuvent être combinés par l'utilisateur ensuite, ils doivent d'abord être extraits de l'ontologie pour constituer une aide à la navigation dans l'espace des documents.
- Une des manières d'utiliser l'ontologie pour guider la classification des documents et la consultation des classes consiste à proposer une visualisation graphique. Or il est plus facile de « naviguer » dans des hiérarchies (arbres) extraites de l'ontologie pour restreindre plus ou moins l'ensemble des documents recherchés. La présentation de l'ensemble des concepts des hiérarchies ou de l'ontologie donne à l'utilisateur une idée du contenu des documents de la collection.
- L'indexation à l'aide des concepts de l'ontologie est assez immédiate si l'on exploite les termes associés aux concepts et des traitements linguistiques élémentaires comme la lemmatisation. Améliorer cette indexation reste une question ouverte.

Ce travail en cours doit faire l'objet d'expérimentations pour évaluer la pertinence des différents choix effectués. Ce genre d'approche suppose de s'intéresser à des domaines stables, dans lesquels les connaissances évoluent peu. Ou alors il faut définir un protocole de maintenance

de l'ontologie et de définition de nouveaux points de vue au fur et à mesure que les connaissances du domaine évoluent ou que la collection s'enrichit de nouveaux documents.

7.3.4 Ontologies pour l'annotation sémantique de documents structurés

L'indexation de documents à l'aide d'ontologies et l'association automatique de méta-données à des documents ressortent comme deux questions d'actualité au cœur de nombreux projets de recherche. La maîtrise de l'indexation constitue, avec la disponibilité d'un plus grand nombre d'ontologies, une des clés du succès de la mise en place du Web Sémantique. L'indexation de textes est souvent reformulée comme le problème du repérage d'instances de concepts de l'ontologie dans les textes alors que l'association de méta-données relève d'une caractérisation de l'information à un niveau différent. Pour repérer des instances de concepts dans des textes, l'analyse du langage naturel offre des perspectives prometteuses : si on arrive à caractériser les contextes (lexicaux ou grammaticaux) d'apparition d'un concept dans un texte, soit manuellement soit par apprentissage à partir d'un échantillon de textes étiquetés à la main, les principes de l'extraction d'information peuvent être appliqués pour localiser des phrases où se trouvent des concepts ou des instances de concepts.

La représentation de documents à l'aide des concepts d'une ontologie est supposée augmenter les possibilités de recherche de connaissances dans ces documents. Les requêtes des utilisateurs doivent alors suivre un traitement analogue et déboucher sur une expression du besoin en information sous forme d'un ensemble de concepts.

Parce qu'ils s'appuient sur les liens entre termes et concepts, les processus d'indexation pour la recherche automatique ou d'annotation à l'aide de mots-clés sont les symétriques de l'analyse de textes pour construire les modèles qui servent à les indexer. Partant de ce constat, j'ai relevé plusieurs questions de recherche proches de celles de la construction d'ontologies, et choisi de les traiter avec mon expérience en matière d'utilisation de logiciels d'analyse terminologique pour construire ces ressources. Le coût de la construction de l'ontologie et de l'indexation ne semble justifié que pour des applications dans des domaines spécialisés et visant la recherche d'informations précises dans les textes. J'ai donc choisi de me focaliser sur des documents structurés dont les paragraphes constituent les unités d'information.

Dans le cas de documents structurés à l'aide de balises XML, l'indexation par les concepts et les balises constituent deux manières de caractériser le contenu des documents. Les balises rendent compte de la sémantique introduite par l'auteur du document à l'aide d'une DTD, chaque fragment de texte étant caractérisé par son rôle dans l'ensemble du document. À cela, l'indexation conceptuelle des fragments de texte précise la nature des connaissances du domaine qu'ils contiennent. L'idée est d'exploiter la complémentarité de ces deux caractérisations des textes pour améliorer la recherche d'information.

L'annotation sémantique de documents structurés et balisés en XML pose des questions de recherche en recherche d'information autant qu'en matière de construction, de maintenance et d'utilisation d'ontologies. Je listons les questions à aborder dans le paragraphe suivant, avant de présenter deux nouveaux projets sur l'annotation de documents structurés : Arkeotek et Autodiag.

7.3.4.1 Questions de recherche soulevées

Dans le cas particulier des documents étudiés, le balisage XML reflète le rôle de chaque fragment de texte dans un raisonnement particulier. Ce type de caractérisation suppose que l'on dispose d'une ontologie du domaine (ou d'un autre type de modèle proche) d'une part, et d'une DTD liée à la manière de structurer les textes pour expliciter le raisonnement d'autre part.

Je fais l'hypothèse que *si cette ontologie est construite à partir de corpus, son utilisation pour l'indexation sera plus efficace.*

- Une première question est alors de savoir *comment exploiter la structure des documents pour guider l'analyse terminologique et l'organisation des concepts*. Cette question ouvre des perspectives très riches pour la construction de modèles à partir de textes, car jusqu'ici, la structure et la mise en forme des documents sont rarement exploitées.
- Une deuxième question est de savoir *quelles informations conserver avec les termes et les concepts représentés pour anticiper le problème de l'indexation*, c'est-à-dire du repérage dans les textes de la présence de concepts. Par exemple, si l'on conserve dans le modèle les paragraphes dans lesquels les termes sont utilisés, le retour aux textes est facilité.

Pour indexer des textes structurés par des concepts, je souhaite aussi exploiter la structure des documents : il s'agit de *représenter chaque paragraphe à l'aide de concepts* et, pour cela, de *tenir compte de la nature des paragraphes (de leur balisage) pour décider des concepts à leur associer*. Je prévois de définir un processus supervisé dans lequel l'analyste ajuste les concepts à associer parce qu'il maîtrise suffisamment le domaine et le corpus. Enfin, je voudrais tester une solution faisant peu appel à l'analyse du langage au moment de l'indexation elle-même, et s'appuyant plutôt sur la richesse de l'ontologie construite à l'aide de cette analyse. Ce choix soulève de nombreuses questions, comme celle de la nécessité ou non d'une pondération des concepts, de la manière de l'évaluer, et surtout la manière *d'exploiter les relations entre paragraphes et entre concepts pour améliorer la caractérisation sémantique*. On peut imaginer d'associer des concepts reliés aux concepts présents ou de propager les concepts d'un paragraphe à des paragraphes proches au sens de la DTD.

La recherche d'informations dans des documents passe par une représentation des requêtes selon un format comparable à celui des documents, ainsi que par la définition d'une métrique pour associer des documents pertinents à une requête donnée. Selon l'hypothèse retenue, il est assez simple d'imaginer comment parvenir à une représentation de la requête sous forme de concepts. Plus le processus d'association de concepts à des paragraphes sera élaboré, moins il sera utile de reformuler la requête. En revanche, il est plus difficile de *spécifier le mode de calcul de distance entre les concepts des documents et des requêtes*. Dans l'esprit de la proposition de M Baziz (partie 7.3.2), comment adapter les distances actuellement définies entre des mots à des concepts ? pour aller plus loin, comment tenir compte du type des paragraphes dans la sélection des réponses fournies à l'utilisateur ?

Un dernier ensemble de questions, plus orientées recherche d'information, porte sur la présentation des résultats dans ce contexte particulier. *L'interrogation d'une base de documents structurés revient à définir un nouveau mode de consultation basé sur la recomposition de documents*. Grâce à l'indexation sémantique, des portions de documents peuvent être sélectionnées et présentées conjointement. Cela revient à rompre avec la logique de l'auteur, à ne pas respecter la linéarité de son écriture pour privilégier celle de l'utilisateur. Une réponse peut alors réorganiser plusieurs documents selon le point de vue associé aux concepts d'une requête. Cette hypothèse doit être testée. Elle suppose de définir un mode de présentation (et de réagencement) des paragraphes pertinents et son évaluation par les utilisateurs.

Je pose ces différentes questions dans le cadre des deux projets présentés ci-dessous. Pour y répondre, je poursuis ma collaboration avec l'équipe SIG de l'IRIT, spécialisée en recherche d'information (J. Mothe et M. Boughanem).

7.3.4.2 Ontologies pour l'accès à des publications scientifiques

Contexte

Le projet Arkeotek[29] vise une meilleure gestion de l'archivage de monographies, d'articles multilingues et de données (dessins, photographies, …) liés à l'archéologie des techniques (Gardin & Roux, 2004). Afin de constituer des bases de connaissances, les documents sont structurés selon les principes du logicisme et les travaux de Gardin (Gardin, 1991), en identifiant les différentes briques élémentaires de raisonnement, sous forme de propositions, et la contribution des unes aux autres pour produire des inférences. Les documents sont ensuite transformés en documents électroniques par le biais du format SCD (Roux, Blasco, 2004). Ce format se caractérise par l'édition structurée des textes scientifiques en fragments hiérarchisés (propositions interprétatives et antécédentes). Il se veut avant tout une grille d'organisation des écrits scientifiques afin d'en consolider la rigueur argumentaire et ainsi la qualité scientifique. En utilisant ce format pour rédiger, on reste au plus près d'énoncés informatifs dont l'ensemble forme une construction scientifique. Le format peut se traduire par une DTD XML où les balises précisent le rôle des paragraphes dans l'expression d'une argumentation scientifique dans le domaine de l'archéologie.

Cette fragmentation permet, d'une part, un accès plus facile et rapide aux constructions scientifiques et, d'autre part, une indexation raisonnée des bases de données documentaires en fonction des raisonnements qu'elles présentent, répondant par là aux questions d'archivage de masses de données scientifiques. À terme, les bases ainsi constituées favorisent le cumul des connaissances au sein des SHS, à condition toutefois de prévoir des outils pour les interroger [rapport-ARKEOTEK, 04].

Approche retenue et travaux développés

Pour répondre à ce besoin, dans le cadre d'un premier projet financé par le programme Société de l'Information du CNRS, j'ai défini des outils de requête permettant d'accéder rapidement aux résultats disséminés au sein de nombreuses publications. Le programme de ce projet entre tout à fait dans la problématique de l'indexation sémantique telle que je pense l'étudier.

L'indexation s'appuie sur une double caractérisation du contenu des documents préalable à l'archivage : l'une porte sur l'argumentation scientifique à laquelle correspond chaque fragment de texte, et correspond à la structuration selon le format d'édition scientifique SCD ; l'autre vient enrichir les documents structurés par une représentation explicite des connaissances du domaine sur lesquelles porte chaque fragment de document, sous la forme de concepts. Ces concepts, tirés d'une ontologie du domaine concerné, constituent un index sémantique des documents.

L'ontologie est construite à partir d'une analyse linguistique du contenu des documents à indexer. Ainsi, en retour, on utilise les liens identifiés entre termes et textes pour établir des liens entre les concepts de l'ontologie et les textes. Cette association entre concepts et fragments de textes est supervisée. Elle tient compte du typage SCD des paragraphes. La disponibilité de ces liens facilite l'indexation.

Un prototype du système d'interrogation de la base de documents est en cours de développement. Il comprend deux parties. *L'environnement auteur* est destiné à construire la représentation enrichie des textes de la base. Il comprend deux modules, l'un pour guider la construction et la maintenance de l'ontologie à partir de documents structurés, l'autre pour gérer l'indexation supervisée de nouveaux documents à l'aide de l'ontologie. *L'environnement utilisateur* permet d'interroger la base documentaire ainsi indexée. L'utilisateur peut formuler une

[29] www.arkeotek.org

requête, que le système projette sur la base documentaire et pour laquelle il présente des paragraphes pertinents.

Premières réponses aux questions posées

Nature de l'ontologie : L'application visée détermine fortement le niveau de détail de l'ontologie. À partir du moment où l'on estime que les utilisateurs ne feront pas de différence entre deux termes au moment d'interroger la base documentaire, un seul concept est présent dans l'ontologie, auquel tous les termes plus précis ou proches sont associés. De ce fait, il a semblé plus pertinent que l'ontologie comporte une composante terminologique riche, mais soit peu détaillée au sens où certains concepts ne sont pas différenciés. Cette ontologie n'est pas formalisée en logique.

Mode d'indexation : La solution envisagée privilégie une analyse automatique des textes et un travail de modélisation poussé au moment de construire l'ontologie. Dans ce modèle, les liens des termes vers leurs occurrences sont conservés. Ils sont exploités lors de l'indexation. Le processus d'indexation est supervisé : le système propose un ensemble de concepts à associer à chaque paragraphe en fonction des termes présents et d'heuristiques liées au format SCD ; l'utilisateur vient ensuite les modifier ou les valider. Le système implémente des solutions pragmatiques aux questions soulevées qui n'ont pas encore été évaluées.

7.3.4.3 Ontologie pour l'exploitation d'une base d'incidents

Un deuxième projet sur l'indexation sémantique est mené au sein du laboratoire Autodiag[30]. La thèse d'A. Reymonet (que j'encadre depuis septembre 2004) porte sur l'apport du traitement automatique du langage naturel et des ontologies à la définition d'un système d'aide au diagnostic. Un des modes de diagnostic envisagés s'appuie sur une base d'expériences en langage naturel, en complément de modules réalisant un diagnostic à base de modèles ou encore à base de reconnaissance des formes.

Dans ce contexte, la collection de documents à explorer est une base d'expérience constituée de fiches de résolution de pannes sur les calculateurs automobiles. Ces fiches, rédigées en langage naturel, contiennent la description des véhicules sur lesquels le problème est susceptible de survenir, les symptômes qui se manifestent lors de l'apparition de ce problème et la réparation associée. Elles sont structurées en paragraphes étiquetés selon leur rôle dans le raisonnement de diagnostic. L'application visée doit aider à rechercher, à partir de symptômes et de la description du véhicule formulés en langage naturel par l'utilisateur, la ou les fiches qui correspondent au cas courant. À partir de cette fiche, le système présente des indications sur la manière de diagnostiquer et de réparer la panne. Si une telle fiche n'existe pas ou ne permet pas de résoudre le problème, le système doit transmettre une représentation du symptôme qui soit utilisable par les autres méthodes de diagnostic.

L'approche retenue est de s'appuyer sur une ontologie, avec une composante terminologique permettant de gérer le multilinguisme, et couvrant au plus près les concepts mentionnés dans les fiches. On attend de cette ontologie trois types de contribution :

- réduire et simplifier le travail de saisie des informations relatives à un cas, et ceci dans plusieurs langues, grâce aux termes associés aux concepts de l'ontologie ;
- assurer l'indexation de la base d'expériences ;
- faciliter la communication entre les différents modules de raisonnement, ce qui revient à lui faire jouer le rôle de représentation partagée par ces modules.

[30] AutoDiag est un laboratoire mixte fondé en septembre 2004. Il associe deux unités de recherche en informatique du CNRS, l'IRIT et le LAAS, et la société ACTIA, spécialisée dans le diagnostic de pannes électroniques et de calculateurs automobiles.

Or ces utilisations orientent de manière contradictoire le contenu de l'ontologie. Il se dégage des premières études qu'une ontologie utile pour l'exploitation des fiches d'incident et l'aide à la formulation des pannes serait assez simple, comporterait peu de concepts représentant des fonctions de haut niveau. On retrouve là les conclusions du projet Arkeotek. En revanche, si cette ontologie sert aussi au raisonnement à base de modèles, elle doit être beaucoup plus riche, complexe et précise pour décrire toutes les fonctions et tous les composants pouvant être incriminés. Une première version de l'ontologie est en cours de construction à partir de l'analyse du contenu des fiches. Le processus d'indexation n'a pas encore été défini.

7.3.5 Situation par rapport aux perspectives du domaine

Les thématiques de l'indexation sémantique de documents et celle, assez proche, de l'annotation de documents à l'aide de méta-données sont aujourd'hui deux des pôles très actifs des recherches dans le domaine. Cette activité se justifie en partie par la perspective du web sémantique, pour lequel les pages ou les données disponibles sur les sites du web doivent être enrichies à l'aide des concepts formels. Ces concepts doivent pouvoir être manipulés par des agents logiciels et respecter la sémantique que leur attribuent des utilisateurs. Or cette thématique est de plus en plus abordée sous l'angle du traitement automatique des langues, en particulier de l'apprentissage (Ciravegna, 2005) : des pages de référence, annotées manuellement à l'aide de concepts, servent de base d'apprentissage. À partir de ces exemples, le système apprend des patrons pour annoter de nouvelles pages automatiquement. Cette approche suppose des corpus volumineux et suffisamment réguliers.

Mes choix et hypothèses se démarquent de ces approches en donnant une place importante à l'humain qui supervise le processus d'indexation, et en proposant de s'appuyer non pas sur une ontologie formelle mais sur une ontologie ayant une composante terminologique riche. Enfin, une autre particularité des questions abordées est de se centrer sur des domaines restreints, dans lesquels les textes sont structurés de manière explicite. La caractérisation du contenu des documents par les connaissances qu'ils abordent (représentation par concepts) et par le rôle des paragraphes dans un raisonnement est innovante et ouvre la voie à différentes manipulations du « contenu » informatif des documents. Ensuite, le rapprochement entre représentations de documents et requêtes impose de définir des outils et algorithmes pour comparer des ensembles de concepts ou encore agréger des modèles pour mieux les réutiliser. Les recherches sur les modes de calcul de distances sémantiques entre ontologies sont d'ailleurs un thème très actif du domaine, ces calculs s'appuyant aussi sur des comparaisons syntaxiques et sémantiques.

En associant ressources terminologiques, modèles de connaissances et documents, je pose l'enjeu du web sémantique sous une forme différente et plus souple. L'intérêt en serait d'autoriser des traitements plus puissants des informations disponibles dans les documents à travers la mise à disposition de représentations associées traduisant leur structure et leur contenu. Le choix de ces représentations suppose d'anticiper les traitements et l'interprétation qui en sera faite. La diversité des ressources possibles est une chance, que ne doit pas cacher l'usage systématique du mot ontologie. Enfin, utiliser des ontologies pour des agents logiciels ou pour une recherche d'information n'est pas équivalent. Cela ne couvre ni les mêmes besoins ni les mêmes objectifs en matière de ressource de connaissances. Or l'amalgame est souvent fait, au risque d'occulter des différences fondamentales. J'ai choisi de situer nos travaux dans le seul cadre d'applications pour lesquelles les modèles servent de médiateur avec l'utilisateur, sont au service de son interaction avec l'application.

7. 4 - Mise à jour et maintenance de modèles conceptuels

La question de l'évolution et de la maintenance des modèles conceptuels dans des environnements dynamiques a été jusqu'ici peu abordée. Or les ontologies sont des représentations

de connaissances figées, faites pour fonctionner dans des applications où les utilisateurs, les connaissances et les terminologies autant que les textes de référence évoluent sans cesse. La maintenance de ces modèles en cohérence avec l'évolution des sources documentaires ou des besoins constitue un problème rarement anticipé au moment de leur construction. Cette cohérence est d'autant plus nécessaire dans le contexte de la recherche d'information où les modèles sont à la fois les résultats d'analyses de textes et des ressources pour les indexer, comme dans le cadre de l'indexation par des concepts.

La maintenance et l'évolution des ressources terminologiques et ontologiques ont été identifiés comme un axe de prospective à approfondir et étudier de manière interdisciplinaire par le rapport de l'action spécifique « corpus et terminologies » [ASSTICCOT, 04]. Cette question est aussi ressortie comme cruciale à partir de mes différents travaux. Pour l'étudier, je me place toujours dans le contexte de domaines spécialisés et d'ontologies pour une gamme d'applications précises.

7.4.1 Observation des évolutions de l'expression de connaissances dans le temps

7.4.1.1 Questions et programme de recherche

L'objectif ici n'est pas vraiment de savoir comment gérer le modèle en cohérence avec un contexte qui change, une terminologie qui évolue, de nouveaux besoins exprimés pas les utilisateurs ou de nouveaux textes à référencer ou indexer. Il s'agit plutôt de juger *en quoi une terminologie ou une ontologie peut servir à l'étude diachronique d'un domaine*, de ses concepts ou de sa terminologie. Implicitement, ces études font l'hypothèse que la représentation normalisée ou tout au moins structurée que constitue un modèle serait mieux adaptée que les documents bruts ou d'autres sources de connaissances pour mettre au jour des évolutions, des glissements ou des innovations au niveau terminologique ou conceptuel.

Pour l'instant, je n'ai envisagé cette question que dans une optique d'archivage et d'observation, de restitution de phénomènes passés. Mon travail s'appuie toujours sur une étude privilégiée des textes, qui deviennent alors les traces à analyser pour restituer des modèles à différentes périodes. Plusieurs approches sont alors possibles :

- constituer autant de modèles ou d'analyses que de périodes temporelles à comparer : cas du projet Th(IC)² (travail réalisé en 2003), que je présente ci-dessous ;
- constituer une ressource unique et voir comment elle se projette à différentes périodes : cas du projet CNES (projet qui commencera en septembre 2005). Ce projet invite à étudier aussi comment le modèle peut anticiper des usages et faciliter le repérage futur d'éléments dont les auteurs avaient indiqué qu'ils pourraient devenir critiques.

7.4.1.2 Analyse terminologique et évolution d'un domaine : projet Th(IC)²

Le projet Th(IC)² a eu pour objectif d'appliquer les méthodes et logiciels de l'ingénierie des connaissances à des documents de ce domaine pour confronter différents travaux [IC, 03]. Une première tâche avait pour objectif de construire une ontologie du domaine à partir de l'analyse d'un corpus d'articles scientifiques représentatifs. Une deuxième tâche[31] voulait montrer qu'une analyse lexicale contrastive de corpus chronologiquement successifs pouvait souligner l'évolution des thématiques et des concepts importants du domaine.

[31] Travail mené en collaboration avec D. Bourigault et R. Teulier

Dans ce projet, aucune ressource spécifique n'a été construite pour être conservée ensuite en tant que *modèle* ou *terminologie* du domaine. Les supports à l'analyse ont été les résultats des logiciels SYNTEX et UPERY, et le réseau terminologique ainsi constitué. Pour dépouiller et analyser ces résultats, des listes de termes extraits par sous-corpus ont été constituées. Même sans la mise en forme d'un modèle, il s'agit bien d'une expérience d'ingénierie des connaissances : les logiciels utilisés ont permis une analyse et une interprétation en termes de connaissances et d'évolution thématique du domaine au cours du temps. Il manque un support à la restitution de cette analyse, qui aurait pu, dans ce cas, prendre la forme d'un modèle ou d'un index renvoyant aux articles.

J'en retiens ici l'intérêt et les limites de l'approche et des logiciels utilisés. À partir des résultats de SYNTEX, ont été repérés les termes dont les propriétés (fréquence, termes reliés, mais aussi environnements textuels comme les voisins ou les mots co-occurrents) sont très nettement différentes d'un corpus à l'autre. L'interface de consultation organise les termes en classes thématiques qui révèlent l'émergence, la disparition ou la reformulation de certains thèmes de recherche ou de notions. SYNTEX favorise l'exploitation séparée et comparée de chaque document au sein des corpus, et donc l'analyse fine des comportements des termes.

Certains points forts ou faibles de l'approche s'expliquent par les caractéristiques des corpus. Ce type d'analyse semble plus pertinent lorsque les textes étudiés peuvent être regroupés en différents sous-corpus, tous étant relativement homogènes dans la forme, comparables entre eux (même genre textuel par exemple) et suffisamment volumineux et représentatifs du domaine. Dans cette expérience, les corpus étaient trop peu volumineux, de telle sorte que les contextes des termes communs à plusieurs documents d'un corpus étaient rares, et encore plus rares ceux pouvant être comparés d'un corpus à l'autre. De plus, l'analyste étant un expert du domaine, l'exploitation des résultats est très marquée par cette expertise, et rend difficile l'évaluation de l'apport des textes.

7.4.2 Maintenance des ressources terminologiques en lien avec des textes

Étroitement liée au mode de conception de l'ontologie, la maintenance se heurte aux mêmes obstacles : difficulté pour trouver les connaissances à modifier pour que l'ontologie réponde mieux à un besoin particulier, caractère laborieux de la réorganisation d'un modèle, difficulté à évaluer l'impact et la pertinence d'une modification sur la structure conceptuelle. Cette maintenance peut être facilitée si l'ontologie a été documentée, et surtout si les choix de sélection et d'organisation des concepts sont explicitement notés dans l'ontologie. Dans le cas d'ontologies construites à partir de textes, on constate que la possibilité de revenir aux textes sources constitue une aide à l'interprétation et donc à la maintenance. Je défends cette idée depuis la construction de GEDITERM et envisageons d'en poursuivre l'étude sous de deux angles originaux.

Une première direction est d'anticiper la maintenance au moment de la construction des modèles en prévoyant *de gérer, avec l'ontologie, des outils qui ont permis de la construire*, comme des patrons linguistiques ou des critères statistiques. J'ai commencé dans GEDITERM à enregistrer avec les relations entre concepts les patrons qui ont permis de trouver ces relations. Il s'agirait de faire de même pour les ontologies et d'enregistrer des éléments équivalents pour certains concepts. Je l'ai évoquée dans les perspectives d'évolution d'une plate-forme de modélisation comme TERMINAE.

Une autre est de *tirer parti des approches à base de traitement automatique des langues* pour enrichir de manière incrémentale des modèles au fur et à mesure de l'évolution des corpus indexés. J'ai commencé à aborder ces questions dans les projets en cours :

- À travers les projets Arkeotek et Autodiag liés à la recherche d'information dans des documents spécialisés, j'abordons cette question sous l'angle classique d'un *cycle d'ajustement du modèle au fur et à mesure de l'évolution du corpus* ; ces évolutions se font par intégration des résultats de logiciels de traitement automatique des langues appliqués aux nouveaux textes ; la difficulté est d'aider à intégrer dans l'ontologie une information jugée pertinente à

partir de l'analyse des textes. Ce processus est coûteux et risque de mettre à mal la cohérence du modèle.

- À travers le projet DynamO, j'explore une piste plus innovante basée sur *l'utilisation d'agents adaptatifs associés à des logiciels de traitement automatique des langues* (Fig. 7.4.1) ; le système d'agents doit permettre de réviser et ajuster rapidement le contenu des ontologies ou terminologies en fonction de nouveaux documents ;
- à travers un projet de gestion des connaissances (projet CNES), j'étudie manière de repérer des changements dans l'expression des connaissances au cours du temps.

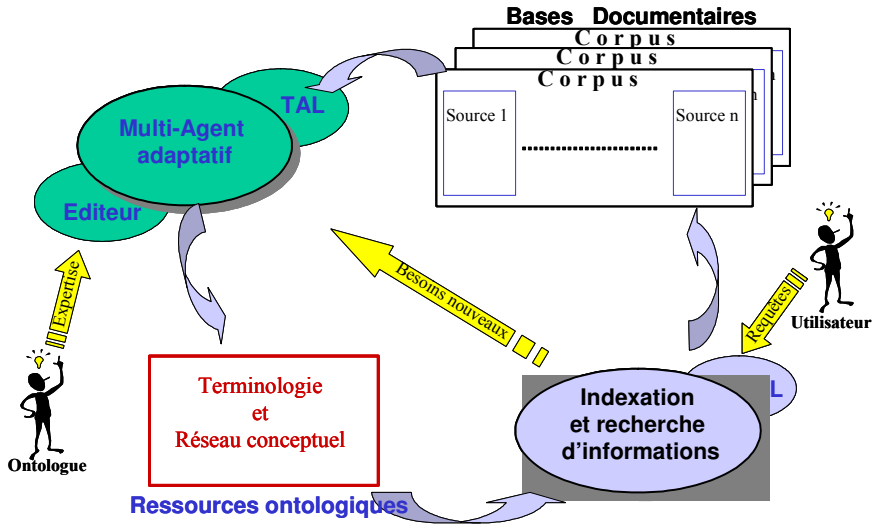

Figure 7.4.1 : Le lien entre construction d'ontologies à partir de textes et indexation sémantique de documents souligne la nécessité de gérer la maintenance en continu**.**

7.4.3 Maintenance classique par enrichissement manuel

Projet Arkeotek

Dans ce cadre, il est prévu une mise à jour régulière de la base de données documentaire dans laquelle sont recherchées les informations. Cela suppose une révision de l'ontologie pour l'adapter aux connaissances couvertes par les nouveaux documents, et une nouvelle définition de l'index pour les documents ajoutés.

Le système de recherche d'information dans ces textes comporte donc un module de maintenance de l'ontologie en fonction de l'ajout de nouveaux documents. Contrairement à la plupart des éditeurs d'ontologie actuels, l'environnement de construction de l'ontologie permet d'intégrer des textes de manière incrémentale, sans avoir à relancer l'analyse des textes déjà traités. Il met en valeur les nouveaux termes identifiés ainsi que les concepts à actualiser. Ce processus anticipe de manière coordonnée la maintenance de l'ontologie et l'indexation des documents à partir de l'ontologie. En effet, les termes et concepts ajoutés à l'ontologie suite à l'analyse des nouveaux textes peuvent ensuite être utilisés pour indexer l'ensemble du corpus, et surtout les nouveaux documents.

Les solutions envisagées, y compris pour l'indexation, sont pragmatiques. Je ne prétends pas innover en matière de recherche d'information. Elles seront le point de départ pour une étude plus approfondie dans le cadre du projet DYNAMO.

Projet Autodiag

Une des particularités de ces bases d'incidents est leur perpétuelle évolution suite à l'introduction de nouveaux types de véhicule dans la gamme des constructeurs, à l'identification de nouvelles pannes ou de solutions à des pannes connues. Une équipe de rédacteurs assure la mise à jour régulière de cette base d'expérience. Un processus de maintenance de l'ontologie s'impose donc pour garantir la couverture par l'ontologie du domaine défini par les fiches. En retour, cette ontologie, une fois enrichie, doit permettre l'indexation des nouvelles fiches ajoutées à la base d'expérience. Le cadre est donc celui du processus cyclique étudié dans ce projet. La solution envisagée actuellement est une approche classique basée sur l'analyse du langage naturel et sur un effort de modélisation manuelle de la part de l'ontologue. Une nouvelle solution à base d'agents pourrait être mise en œuvre et évaluée si le projet Dynamo se poursuit. Une autre étape du projet visera la confrontation de cette approche à une approche à base d'agents adaptatifs.

7.4.4 Agents adaptatifs pour construire et maintenir une ontologie dynamique

Je pose conjointement les questions de construction et de maintenance d'ontologies à partir de textes afin de fournir une solution qui favorise les mises à jour simples et fréquentes. L'approche retenue, en collaboration avec M. P. Gleizes (équipe SMAC de l'IRIT), s'appuie sur un Système Multi-Agents (SMA) adaptatifs. Les agents sont chargés de s'appuyer sur une analyse des textes pour repérer des fragments de textes (candidats termes, classes sémantiques, relations lexicales), en faciliter l'exploitation pour définir des éléments d'ontologie puis la validation par le cogniticien.

Dans un premier temps, les agents définis géraient une analyse lexicale et syntaxique. Ils étaient ensuite présentés à l'analyste pour être révisés (DEA-OTTENS, 2004). Ce travail a permis de dresser un état de l'art des rares travaux faisant appel aux systèmes multi-agents pour faire du traitement automatique des langues. Il a confirmé la nécessité d'avoir recours à un minimum de ressources sur la langue analysée comme dans le système HACTAR (Lebarbé, 2002).

La suite plus ambitieuse de ce travail vise la construction et la maintenance d'ontologies en contexte dynamique en s'appuyant sur des textes. Les domaines d'expérimentation de ce projet - DynamO (Dynamic Ontologies) – sont ceux des Sciences de l'Information et de la Communication (collaboration avec l'URFIST de Nice) et ceux des projets Arkeotek et Autodiag. Cette étude fait l'objet de la thèse de K. Ottens depuis septembre 2004.

Dans la continuité de l'étude de l'indexation à l'aide de concepts, l'objectif principal est de concevoir une approche méthodologique et un ensemble d'outils qui permettent de prendre en compte de manière unifiée à la fois la construction de ressources ontologiques à partir de documents et l'utilisation de ces ressources pour la recherche d'informations (fig. 7.4.1). Cela correspondra à deux groupes d'outils indépendants dans une application globale. Pour cela, je traiterai conjointement la gestion de l'ontologie et celle de l'indexation. J'ai déjà abordé la question de l'indexation. Je m'intéresse ici à la partie maintenance, gérée par un module de construction et de maintenance d'ontologies. Ce module est composé de deux outils qui permettront de comparer les approches associées : l'approche classique de création et de maintenance d'ontologie est une approche basée sur un système multi-agent adaptatif.

L'approche classique est adaptée de celle du projet Arkeotek. Le fonctionnement envisagé pour l'approche à base d'agents est la suivante. Le système multi-agents construit de manière automatique un premier réseau conceptuel à partir de son analyse des textes ou des résultats d'un logiciel comme SYNTEX. Ensuite, le système présente à l'analyste le réseau des agents obtenus à partir des unités linguistiques et de leurs interactions. En les validant ou les corrigeant, l'analyste permet au réseau de se modifier, le réseau apprenant à partir de l'observation des modifications. Le processus se poursuit jusqu'à l'obtention d'un état stable. En phase de maintenance, le logiciel

assure un support coopératif à l'analyste pour orienter l'évolution de l'ontologie en fonction de son utilisation. Plus le système d'agents sera entraîné sur le domaine concerné, plus le rôle de l'analyste sera minimisé.

L'apport scientifique repose essentiellement sur l'articulation originale entre le domaine multi-agent, le traitement de la langue et les ontologies. Les propriétés de coopération, d'auto-organisation et d'émergence des systèmes multi-agents adaptatifs seront exploitées de manière innovante pour ajouter aux ontologies une dimension dynamique qui correspond à sa capacité d'évolution (définition de nouveaux concepts, termes, relations, ou réorganisation de concepts existants) en fonction des situations. Cette approche vise aussi à sortir de la séquentialité des traitements linguistiques en intégrant l'interaction entre syntaxe et structures conceptuelles.

7.4.5 Situation par rapport aux perspectives du domaine

Autant avec le web sémantique que pour des applications « locales » d'aide à des utilisateurs ou des groupes d'utilisateurs, les applications à base de connaissances n'apportent une assistance réelle que si elles évoluent, s'adaptent à leur contexte d'utilisation. Cette condition n'est pas toujours suffisamment mise en avant en ingénierie des connaissances. Or les contextes dans lesquels des connaissances s'élaborent, se mettent en œuvre, se partagent ou se stabilisent sont de plus en plus changeants. Parmi ces contextes, j'ai mentionné ceux de la recherche d'information (évolution des documents de la collection, de la terminologie) ou de la gestion des connaissances documentaires liées aux activités en entreprise. Or pour développer des applications, l'ingénierie des connaissances s'est donné comme outil les modèles conceptuels. Ces représentations figent les connaissances utiles à l'application à un moment donné, avec une ambition forte : le modèle peut être considéré comme valide une fois pour toute, de façon définitive et universelle.

Or, dans de nombreuses applications comme l'indexation sémantique ou l'annotation par des méta-données, les ontologies sont souvent destinées à renvoyer à des documents peu stables, qui évoluent ou sont modifiés régulièrement. Les travaux sur la gestion des connaissances ont également montré la limite qu'il y a à fixer des modèles de connaissances rigides dans des contextes où les utilisateurs et leurs activités évoluent. De nombreux auteurs en IA, y compris J. Sowa[32], soulignent les limites des modèles dès lors que l'on s'intéresse à la réalité d'usage ou de mise en œuvre de connaissances. La plupart des systèmes informatiques construits pour faciliter la réalisation d'une tâche cognitive s'appuient sur des représentations qui prétendent reproduire des connaissances. Or si les connaissances sont cruciales, c'est parce qu'elles sont transmises, enseignées, échangées, qu'elles circulent. J. Sowa parle de « flux de connaissances ». Cette analyse souligne la dynamique qui anime les connaissances, et le paradoxe qu'il y a à les figer dans des modèles. Parmi les perspectives de recherche qui se dégagent de cette analyse, il y a la volonté de rendre compte de la dynamique des connaissances, de mieux intégrer les utilisateurs et les évolutions de leurs besoins en offrant des supports aux échanges et à la reconstruction de connaissances. Cet objectif suppose une meilleure intégration des dimensions sociale, ergonomique et cognitive à côté d'une analyse purement technique.

La question de la maintenance rejoint cette problématique. Je fais en effet l'hypothèse que la construction et la maintenance des modèles à partir de textes sont deux des moyens de faire vivre le modèle. Une solution à la maintenance permettrait de contribuer à répondre à cette volonté que le modèle participe d'un mouvement d'échange et d'interprétation des connaissances. Plus fondamentalement, on peut se demander si une ontologie pourra jamais anticiper les usages et avec eux les points de vue sur le monde. Les concepts doivent régulièrement être adaptés et évoluer, ce qui nécessite de définir des scénarios d'évolution autant que de travailler sur les représentations

[32] http://users.bestweb.net/~sowa/

elles-mêmes. Le projet d'étudier ce processus de maintenance pour une utilisation ciblée entre dans ce cadre.

7. 5 - Conclusion

Ce mémoire fait état de mes recherches et contributions en ingénierie des connaissances. Pour le terminer, il me semble intéressant de souligner la nécessité de développer des recherches pluridisciplinaires, qui fournissent des propositions à une ingénierie. Ces recherches ne peuvent pas seulement s'intéresser à des aspects informatiques de représentation et de formalisation. Dès lors qu'il faut outiller et instrumenter les activités cognitives par des systèmes informatiques « à base de connaissances », la prise en compte des usages, des activités, de la mise en œuvre des connaissances s'impose.

Pour répondre à cette question, j'ai d'abord proposé une méthode et des outils de modélisation conceptuelle qui étaient généraux. J'avais l'ambition de fournir un cadre unique pour la gamme des applications et des modèles possibles qui pourraient servir dans des applications à base de connaissances. Parmi les originalités de cette méthode et de la plate-forme associée, je souligne la diversité des techniques de recueil et d'analyse des savoir-faire proposées, inspirées de la psychologie cognitive et de l'ergonomie. Cette méthode et toutes mes approches défendent une analyse ascendante des connaissances, partant des savoir-faire en usage pour dégager des modèles plus abstraits, concis et opérationnalisables.

Pour être plus efficace dans la construction des modèles, les textes s'avèrent être des sources de connaissances très riches, utiles surtout à la modélisation des domaines. Le traitement automatique des langues, basé sur des principes linguistiques ou statistiques, offre des outils très pertinents dans cet objectif, moyennant la définition d'interfaces de navigation et d'exploitation de leurs résultats. L'approche à partir de textes est particulièrement adaptée à la modélisation de terminologies et d'ontologies. Elle permet de leur associer une composante terminologique utile pour ensuite revenir vers les textes pour les indexer.

Par rapport aux entretiens d'experts, la construction de modèles à partir de textes fait ressortir de manière plus saillante que les modèles reflètent des points de vue particuliers, un parti pris sur le domaine essentiellement lié à l'utilisation qui sera faite du modèle. L'analyse de textes souligne aussi la difficulté qu'il y a à fixer un sens aux termes du domaine, remet en question l'hypothèse que les concepts « existent » et doivent être identifiés. Ce qui existe, ce sont des traces d'usage des termes, qui permettent d'en reconstruire un sens à partir de leur interprétation générale. Seul l'analyste ayant en tête les besoins des utilisateurs peut être porteur du point de vue qui conduira à la définition d'un concept. L'analyste est également garant d'une certaine cohérence et de rigueur dans l'organisation des concepts selon des principes ontologiques. C'est pourquoi je défends une approche supervisée et non une analyse complètement automatisée des textes.

Afin de mieux comprendre l'impact de la prise en compte des besoins d'utilisation des modèles sur leur contenu mais aussi leur structure, leur degré de formalisation ou leur construction, j'ai approfondi un cadre d'utilisation des ontologies qui se développe de plus en plus : la recherche d'information. Parce que ces applications gèrent des collections de textes ou de documents, en fonction de leur contenu, elles constituent un cadre privilégié pour des ontologies construites à partir de textes et ayant une composante terminologique riche. Les modèles comme les bases de connaissances terminologiques ou les ontologies améliorent la recherche d'information, en particulier dans des domaines spécialisés. Les ambitions du Web Sémantique sont justement de les utiliser comme des sources de méta-données. Les concepts associés aux textes constituent des représentations manipulables par des applications comme des agents logiciels sur le Web. Mais leur coût et leur complexité ainsi que leur maintenance peuvent être un obstacle à la généralisation de leur diffusion.

La construction d'ontologies à partir de textes et l'utilisation du traitement automatique des langues (en particulier l'extraction d'information) et de l'apprentissage génèrent aujourd'hui beaucoup d'espoirs pour réduire le coût et anticiper les problèmes de maintenance. Ils ne doivent pas cacher les limites qu'il y a à figer dans des représentations formelles très sophistiquées des bribes de connaissances utilisées dans des contextes mouvants et sans cesse renouvelés. Une direction de recherche à approfondir pour l'ingénierie des connaissances est donc de bâtir des applications respectant cette dynamique et facilitant les adaptations à des nouveaux contextes.

7. 6 - Publications sur ces travaux

[INFORSID, 03] BAZIZ M., AUSSENAC-GILLES N., BOUGHANEM M., Exploitation des Liens Sémantiques pour l'Expansion de Requêtes dans un Système de Recherche d'Information. *XXI° Congrès INFORSID 2003*. Nancy, 3-6 Juin 2003. INFORSID. Inforsid, 20 rue Axel Duboul - 31000 Toulouse. 124-131.

[ISI, 03] BAZIZ M., AUSSENAC-GILLES N., BOUGHANEM M., Exploitation des Liens Sémantiques pour l'Expansion de Requêtes dans un Système de Recherche d'Information. *Revue Ingénierie des Systèmes d'Information (ISI)*. J. Le Maître (Ed.). Paris : Hermès. Vol 8. N°4/2003. 113-136.

[INNO, 04] HERNANDEZ N. AUSSENAC-GILLES N., OntoExplo : Ontologies pour l'aide à une activité de veille ou d'exploration d'un domaine. In *actes du VI° Colloque International Sur l'Innovation « Structuration des Ressources Profils des Usages »*. Foix (F). 28 au 30 janvier 2004.

[RIAO, 04] AUSSENAC-GILLES N., MOTHE J., Ontologies as background knowledge to explore document collections. *In proc. Of RIAO 2004*. Avignon, April 2004. 129-142.

[CIFT, 04] AUSSENAC-GILLES N., Représentation sémantisée des textes: terminologies et dimensions pragmatiques (qui-quand-où). *Actes du 2e Colloque International sur la Fouille de textes (CIFT 2004*). Eds. Antoni M.-H., Yvon F., La Rochelle (F), 22-24 juin 2004. p - .

[SWIR, 04] BAZIZ M., BOUGHANEM M., AUSSENAC-GILLES N., Semantic representation of Documents by Ontology-Document Mapping. In Proceedings of the 2nd ACM SIGIR Workshop *on Semantic Web and Information Retrieval (SWIR 2004)*. Sheffield (UK), July 25-29th, 2004.

[VSST, 04] BAZIZ M., BOUGHANEM M., AUSSENAC-GILLES N., Vers une représentation sémantique de documents. Dans *Actes de V.S.S.T 2004 (Veille Stratégique Scientifique et Technologique)*. Toulouse (F). 25-29 octobre 2004.

[RI3, 04] AUSSENAC-GILLES N., CONDAMINES A. Documents électroniques et constitution de ressources terminologiques ou ontologiques. *Revue Information, Interaction, Intelligence I3*. Numéro spécial sur le document numérique. Eds CHARLET J. et SALAÜN J.-M. 4(1):75-94. 2004.

[ASSTICCOT, 04] N. AUSSENAC-GILLES AND A. CONDAMINES. *Action spécifique STIC « Corpus et Terminologie » ASSTICCOT (AS 34). Rapport final*. Rapport Interne IRIT/2003-23-R. Oct. 2003. 70 p.

[AGENTAL, 04] OTTENS K., AUSSENAC-GILLES N., GLEIZES M.-P., GLIZE P., Systèmes multi-agents pour la construction d'ontologies à partir de textes : revue de questions. *AGENTAL : Agents et Langue, actes de la journée d'étude de l'ATALA*. Paris, 13 Mars 2004. pp 15-22.

[Rapport-ARKEOTEK, 04] N. AUSSENAC-GILLES, V. ROUX, P. BLASCO, *Rapport intermédiaire du projet Arkeotek : Utilisation d'ontologies pour la définition de vues dynamiques permettant la lecture par niveaux d'une base de connaissances de textes et de données scientifiques structurés selon le format SC*. Rapport IRIT/2004-20-R. Nov. 2004. 27 p.

[Rapport-IRIX, 04] N. AUSSENAC-GILLES, M. BOUGHANEM, MOTHE J., *Réseau Recherche et Filtrage Sémantique d'Information. Rapport final du contrat région 102R*. Rapport interne IRIT/2004-30-R. Décembre 2004. 36 p.

[SAC-TIAR, 05] BAZIZ M., BOUGHANEM M., AUSSENAC-GILLES N., CHRISMENT C., Semantic Cores for Representing Documents in IR. *Proceedings of SAC-IAR'05, the ACM SAC Track on Information Access and Retrieval*. Santa Fe (NM, USA). 2005

[STI, 05] AUSSENAC-GILLES N., ROUX V, de SAIZIEU B., BLASCO P., Ontologies dédiées à la consultation de documents structures selon un modèle logico-sémantique. In *Actes du colloque de clôture du programme Société de l'Information*. Lyon (F), mai 2005.

CHAPITRE 8 - BIBLIOGRAPHIE

8. 1 - Acquisition et Ingénierie des connaissances, modèles conceptuels

AUSSENAC-GILLES N., KRIVINE J.-P., SALLANTIN J., Introduction, *Revue d'Intelligence Artificielle*, numéro spécial sur l'Acquisition des Connaissances, Eds : N. Aussenac-Gilles, J.P. Krivine, J. Sallantin, Paris : Hermès, Vol 12 (1/2) , 1992.

BACHIMONT B. Pourquoi n'y a-t-il pas d'expérience en ingénierie des connaissances ? *Actes de la conférence Ingénierie des Connaissances IC2004*, Lyon (F), 5-7 mai 2004. 53-64. 2004.

BENJAMINS R., Problem Solving methods of diagnosis, *Thesis Universiteit van Amsterdam,* With index ref. ISBN 90-9005877-X, Amsterdam, 1993.

BERRY, 1988

BOOSE J., Personal Construct theory and the transfer of human expertise. *Proceedings of the 6th European Conference on Artificial Intelligence ECAI'84*, Advances in Artificial Intelligence. T. O'Shea Ed. : 51-60. 1984.

BOOSE J.H., *Expertise Transfer for Expert System Design*. Elsevier. Series "Advances in Human factors and ergonomics". 1986 : 321p

BOOSE J.H., BRADSHAW J. Expertise Transfer and Complex problems : using AQUINAS as a knowledge-acquisition workbench for knowledge based systems. *International Journal of Man Machine Studies* (IJHCS). Academic Press. **26**. 1987 : 3-28.

BREUKER J., VAN DE VELDE W., *The CommonKADS library for expertise modelling : reusable problem solving components,* Amsterdam : IOS Press. 1994.

CAUSSE K., Heuristic control knowledge. *Knowledge Acquisition for Knowledge Based Systems*. Proc. of EKAW'93. Aussenac N., Boy G., Gaines B., Linster M., Ganascia J.G., Kodratoff Y. Eds. LNAI 723. Heidelberg: Springer Verlag. pp. 183-199. 1993.

CAVERNI J.-P., Les protocoles verbaux comme observables des processus cognitifs. *Colloque annuel SFP* sur « activités cognitives : modèles de processus et niveau d'observation ». Aix-en-Provence : 26 p. mars 1986.

CHABAUD C. SOUBIE J.L., SPERANDIO J.C. , Repères pour une démarche et une méthodologie de validation des systèmes à base de connaissances. *2° conférence internationale TEV89 Travail à l'écran de Visualisation*, Montréal (Can.), Sept. 1989.

CHANDRASEKARAN B., Generic Tasks in Knowledge based reasoning : High-level building blocks for Expert System Design, *IEEE Expert,* Autumn 1986 : 23-30.

CHARLET J., REYNAUD C., KRIVINE J.-P., Causal model-based knowledge acquisition, *International Journal of Human Computer Studies*, **44,** 629-652, 1996.

CHARLET J., ACTE : acquisition des connaissances par interprétation d'un modèle causal. *Revue d'Intelligence Artificielle*, **6**, 99-129. Paris : Hermès. 1992

CLANCEY W., The Knowledge Level Reinterpreted: Modelling Socio-Technical Systems. In *International Journal of Intelligent Systems*, Special Issue on Knowledge Acquisition as Modelling, Part 1. Eds K. Ford & J. Bradshaw. Vol**. 8**, 1 : 33-50. 1993.

CLANCEY, W. J. (1997). *Situated Cognition: On Human Knowledge and computer Representation*. Cambridge, UK: Cambridge University Press.

CORBY O., FARON-ZUCKER C., Corese: A corporate Semantic Web Engine. *WWW 11th Workshop on Real World RDF and Semantic Web Applications*. Hawaï (USA), 2002.

DAVID J.-M., KRIVINE J.-P., Augmenting experience-based diagnosis with causal models. *International Journal of Intelligent Systems*, **5**, 83-124. 1989.

DAVID J.-M., KRIVINE J.-P. ET SIMMONS R. (Eds.) *Second Generation Expert Systems*. Berlin : Springer Verlag. 1993

DAVIS R., Interactive transfer of expertise : acquisition of new inference rules. *Artificial Intelligence*. **12** (2) : 121-157, 1979.

DELOUIS I., LISA : Un langage réflexif pour la modélisation du contrôle dans les systèmes à base de connaissances. Application à la planification des réseaux électriques, *Thèse de l'Université de Paris Sud, Centre d'Orsay,* Paris, 1993.

DELOUIS I., KRIVINE J.P., LISA, un langage réflexif pour opérationnaliser les modèles d'expertise. *Revue d'Intelligence Artificielle*. **9** (1) : 53-88, 1995.

de TERSSAC G., SOUBIE J-L., NEVEU J-P., Systèmes experts et transfert d'expertise, *Sociologie du travail*, n° 3-88. 1988

GAINES, B. R., & SHAW, M. L. G. (1980). New directions in the analysis and interactive elicitation of personal construct systems. *International Journal Man-Machine Studies, 13*, 81-116.

GALLOUÏN J.-F., *Systèmes experts*. Paris : Eyrolles, 1988 : 168 p.

GARBAY C. Les sciences du traitement de l'information comme pivot de l'interdisciplinarité : une vision systémique. *Revue I3 : Information, Interaction, Intelligence*. Toulouse : Cépaduès Éditions. **3**(1). 2003

GREBOVAL-BARRY C., KASSEL G., Opérationnalisation de modèles conceptuels : Le Générateur AIDE. In A*cquisition et ingénierie des connaissances, tendances actuelles,* Ed. by N. Aussenac-Gilles, P. Laublet et C. Reynaud. Toulouse : Cépaduès-Editions, 167-184. 1996

GREBOVAL C. , KASSEL G., An Approach to Operationalize Conceptual Models: The Shell Aide. *Proceedings of EKAW 1992*: 37-54. 1992.

HADJ KACEM A., Systèmes à base de connaissances coopératifs : modélisation des connaissances et étude du contrôle. *Thèse de l'Université Toulouse 3*, Janvier 1995.

HADJ KACEM A., FRONTIN J., SOUBIE J.-L., Acquérir des connaissances et structurer le système pour coopérer. *Actes des Journées Acquisition Validation et Apprentissage, JAVA'93*. Saint Raphaël (F). Avril 1993. 1993a.

HADJ KACEM A., SOUBIE J.-L., FRONTIN J., A software architecture for cooperative Knowledge Based Systems, in *Proceedings of HCI International '93, August 2003*, M. J. Smith & G. Salvendy Ed., Elsevier, Orlando, Florida (USA), Vol 2, 303-308, 1993b.

ISTENES Z., TCHOUNIKINE P., Zola: a language to Operationalize Conceptual Models of Reasoning, *Journal of computing and information* (numéro spécial ICCI'96), **2**(1):689-706, 1996.

KARBACH W. LINSTER M., VOSS A., Models, methods, roles and tasks : many labels – one idea ? *Knowledge Acquisition*, **2** : 279-299. 1990

KASSEL G., GREBOVAL-BARRY C., ABEL M.-H., Programmer au niveau connaissance en def*. *Ingénierie des Connaissances, évolutions récentes et nouveaux défis,* J. Charlet, M. Zacklad, G. Kassel & D. Bourigault, éds. Paris : Eyrolles, 145-160. 2000

KASSEL G., PHYSICIAN is a role played by an object, whereas SIGN is a role played by a concept (1999). *Proceedings of the IJCAI workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends*. http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-18/6-kassel.pdf

KRIVINE J.P., DAVID J.M., L'acquisition des connaissances vue comme un processus de modélisation; méthodes et outils, *Intellectica, (***12***),* Paris dec.1991.

LEROUX B., O'HARA K., OUTTANDY S., SHADBOLT N., LAUBLET P., MOTTA E., VITAL: a methodology based workbench for KBS Life Cycle Support, *Rapport final T2.1.5. du projet ESPRIT P5365*, Déc. 1993.

LINSTER M., KRITON: A Knowledge Elicitation Tool for Expert Systems, *Proceedings of EKAW'88*, GMD Studien N° 143, Bonn (RFA). June 1988 : 4.1-4.9.

LINSTER, Marc: Towards a second generation knowledge acquisition tool. In: *Knowledge Acquisition*, **1**: 163-183. 1989.

LINSTER M., L'ingénierie des connaissances, une symbiose de deux perspectives sur le développement de modèles. *Actes des 3e Journées d'Acquisition des Connaissances (JAC)*, Dourdan, 1992.

LINSTER M., Closing the gap between Modelling to make sense and Modelling to implement Systems. *International Journal of Intelligent Systems*, **8**, 209-230, 1993.

MARCUS S. Ed., *Automated Knowledge Acquisition for Expert Systems*, Boston : Kluwer Academic Publishers, 1988, 270 p.

MUSEN M., FAGAN L., COMBS D., SHORTLIFFE E., Using a domain model to drive an interactive knowledge tool. *International Journal of man-Machine Studies*, **26**/1. 105-121. 1987.

NEWELL A., The knowledge Level, *Artificial Intelligence,* Elsevier : Amsterdam. North-Holland. 18. 1982 : 87-127.

NONAKA I., TAKEUCHI H., La connaissance créatrice. La dynamique de l'entreprise apprenante. De Bloeck University. 300 p. 1997.

PEDAUQUE R. T., Le document : forme, signe et medium les re-formulations du numérique. STIC-CNRS - 2003. http://archivesic.ccsd.cnrs.fr/sic_00000511.html

PEDAUQUE R. T., Le texte en jeu, Permanence et transformations du document, STIC-SHS-CNRS – article de travail. 2005. http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/14/01/index_fr.html

PUERTA A., SAMSON W., MUSEN M. Modelling Tasks with Mechanisms. *International Journal of Intelligent Systems*, **8**, 128-152, 1993.

RASMUSSEN J., The role of hierarchical knowledge representation in decision making and system management. *IEEE Trans. On System, Man and Cybernetics*. Vol SMC-**15**, n°2, 1985 : 234-243.

REINDERS M., E. VINKHUYZEN, A. VO, J. M. AKKERMANS, J. R. BALDER, B. BARTSCH-SPORL, B. BREDEWEG, U. DROUVEN, F. VAN HARMELEN, W. KARBACH, Z. KARSSEN, A. TH. SCHREIBER, AND B. J. WIELINGA. A conceptual modelling framework for knowledge-level reflection. *AI Communications*, **4**(2-3):74-87, 1991.

REYNAUD C., Une aide à l'acquisition de connaissances de surface à partir des connaissances profondes. Actes des *9èmes journées d'Avignon sur les Systèmes Experts et leurs applications*. Conférence sur les « systèmes experts de deuxième génération ». Juin 1989.

REYNAUD C., TORT F., Using Explicit Ontologies to Create Problem Solving Methods, *International Journal of Human Computer Studies*, **46**, 339-364, 1997.

REYNAUD C., *L'exploitation de modèles de connaissances du domaine dans le processus de développement d'un système à base de connaissances.* Mémoire d'Habilitation à diriger des recherches. Université d'Orsay. Rapport de recherche 1201 du LRI. 02/1999.

SCHREIBER G A.Th., WIELINGA B.J., (eds.); *KADS, a principled approach to knowledge based system development.* London: Academic Press. 1993

SCHREIBER G., WIELINGA B., AKKERMANS H., VAN DE WELDE W., ANJEWIERDEN A., CML: The CommonKADS Conceptual Modelling Language, in *Proceedings of EKAW'94, A future for Knowledge Acquisition*. Steels L., Schreiber G., Van de Velde W. (Eds). Lecture Notes in AI n°867, Berlin : Springer Verlag : 1-25, 1994.

SCHREIBER G., AKKERMANS H., ANJEWIERDEN A., de HOOG K., SHADBOLT N., VAN DE WELDE W., WIELINGA B., *Knowledge Engineering and Management: the CommonKADS methodology*. MIT Press, Cambridge (Ma). 2000.

SOUBIE J.L., *Coopération et systèmes à base de connaissances*. Mémoire d'habilitation à diriger des recherches en Informatique de l'université de Toulouse 3. Novembre 1996.

SPERANDIO J.C., L'homme face aux changements du travail contemporain. Réflexion sur le rôle des ergonomes. *in L'ergonomie face aux changements technologiques et organisationnels du travail humain*. J.C. Spérandio Ed. Octarès : 3-8. 1996.

STEELS L., Components of expertise, *AI magazine,* **11(2)**: 28-49. 1990.

STEELS L., Reusability and configuration of applications by non-programmers, *VUB-AI memo 92-4*. Brussels (B.). 1992

M. STEFIK. *Introduction to knowledge systems.* Morgan Kaufman. 1995

TCHOUNIKINE P., *Mapcar, une approche pour l'élaboration du modèle conceptuel de raisonnement d'un système à base de connaissances.* Mémoire d'habilitation à diriger des recherches en Informatique de l'université de Nantes. Janvier 1998.

TCHOUNIKINE P., Pour une ingénierie des Environnements Informatiques pour l'Apprentissage Humain. *Revue I3. Information, Interaction, Intelligence*. **2**(1). Toulouse : Cépaduès-Editions, 59-95. 2002.

TRICHET F., TCHOUNIKINE P., DSTM : une approche de l'opérationnalisation fondée sur la réutilisation d'un noyau opérationnel, *Actes des journées Ingénierie des Connaissances,* Roscoff (F), Mai 1997.

VOGEL C. *Génie cognitif.* Paris : Masson. 1988 : 196 p.

WIELINGA, B.J., BREUKER J.A., Interpretation of verbal data for knowledge acquisition, *Proceedings of the 6th European Conference on Artificial Intelligence ECAI'84*, Advances in Artificial Intelligence. T. O'Shea Ed., 1984 : 41-50.

WIELINGA, B.J., BREUKER J.A., Models of expertise, *Proceedings of the 7th European Conference on Artificial Intelligence ECAI'86. Brighton (UK), July 1986, Vol. **1**, 306-318.*

WIELINGA B., SCHREIBER G., BREUKER J., KADS : a modelling approach to knowledge acquisition, Special Issue, *Knowledge Acquisition*, **4**, (1), 1992. 5-54

WIELINGA B.J., VAN DE VELDE W., SCHREIBER G., AKKERMANS H., The CommonKADS Framework for Knowledge Modelling, Proceedings of *KAW'92*, Banff (CAN), Oct 1992.

ZACKLAD M. *Ingénierie des connaissances appliquée aux systèmes d'information pour la coopération et la gestion des connaissances*. Mémoire d'habilitation à diriger des recherches. Université Paris 6. 2000.

8. 2 - Ontologies, Web sémantique, terminologie et linguistique

AIT EL MEKKI T., NAZARENKO A., « Comment aider un auteur à construire l'index d'un ouvrage ? *», actes du Colloque International sur la Fouille de Texte CIFT'2002*, Y. Toussaint et C. Nedellec Eds., oct. 2002, 141-158.

AHMAD K., HOLMES-HIGGIN P.R., SystemQuirk : a unified approach to text and terminology, in *Terminology in advanced Microcomputer Applications.* Proc. of the 3rd TermNet Symposium : recent advances and user reports, Vienna, Austria, 181-194, 1995.

ASSADI H., BOURIGAULT D., Acquisition et modélisation de connaissances à partir de textes : outils informatiques et éléments méthodologiques, *11ème Congrès Reconnaissance des Formes et Intelligence Artificielle,* Paris, 1996.

BACHIMONT B., Ontologie régionale et terminologie : quelques remarques méthodologiques et critiques. *La Banque des Mots*, 7/95, 65--86. 1995.

BACHIMONT B. *Art et sciences du numérique : ingénierie des connaissances et critique de la raison computationnelle.* Mémoire d'habilitation à diriger des recherches de l'Université Technologique de Compiègne. Janvier 2004.

BANEYX A., MALAISE V., CHARLET J., ZWEIGENBAUM P., BACHIMONT B., Synergie entre analyse distributionnelle et patrons lexico-syntaxiques pour la construction d'ontologies différentielles. In *actes de la conférence TIA 2005 (Terminologie et Intelligence Artificielle*), Rouen (F), Avril 2005.

BECHHOFER S. HORROCKS I., GOBLE C., STEVENS R., OILed : a reasonable ontology editor for the Semantic Web. In *Joint German/Austrian Conference on Artificial Intelligence (KI'01*), LNAI volume 2174, Viennes : Springer Verlag, 396-408. 2001.

BIEBOW B., *Élaboration d'ontologies à partir de textes*. Résumé des travaux en vue de l'obtention d'une habilitation à diriger des recherches. Université de Paris 13. Juillet 2004.

BIEBOW B., SZULMAN S., Méthodologie de création d'un noyau de base de connaissances terminologique à partir de textes, Actes des *2° journées Terminologies et IA TIA'97*, Toulouse (F) : Université Toulouse-Le Mirail, ERSS, 69-84. 1997.

BIEBOW B., SZULMAN S., TERMINAE : A linguistic-based tool for the building of a domain ontology, in D. Fensel and R. Studer (Eds.) *Proceedings of 11th European Workshop, Knowledge Acquisition, Modelling and Management (EKAW 99)*, Springer Verlag, LNAI 1621 : 49-66. 1999.

BIEBOW B. & SZULMAN S., TERMINAE : une approche terminologique pour la construction d'ontologies du domaine à partir de textes. *Actes de RFIA 2000, Reconnaissances des Formes et Intelligence Artificielle*, Paris (F). 81-90. 2000.

BOURIGAULT D., LEXTER*, Un logiciel d'extraction de terminologies. Application à l'acquisition de connaissances à partir de textes*. Thèse de doctorat, Ecole des Hautes Études en Sciences Sociales. Paris. 1994a.

BOURIGAULT D., Extraction et structuration automatiques de terminologie pour l'aide à l'acquisition de connaissances à partir de textes. *XVe congrès Reconnaissance des Formes et Intelligence Artificielle RFIA* 1994b.

BOURIGAULT D. , LEXTER, a Natural Language Processing Tool for Terminology Extraction . *Proceedings of Euralex'96*, Göteborg University, Department of Swedish : 771-779. 1996.

BOURIGAULT D. , UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, Actes de *la 9ème conférence annuelle sur le Traitement Automatique des Langues* (TALN 2002), Nancy, 75-84, 2002.

BOURIGAULT, D. & JACQUEMIN, C., Construction de ressources terminologiques, in J.-M. Pierrel (éd), *Ingénierie des langues*, Traité I2C, Paris, Hermes. 2000.

BOURIGAULT D., LEPINE P. Utilisation d'un logiciel d'extraction de terminologie (LEXTER) en acquisition des connaissances, *Acquisition et ingénierie des connaissances : tendances actuelles.* Eds. N. Aussenac-Gilles, P. Laublet, C. Reynaud. Toulouse : Cépaduès-Edition. 267-283, 1996.

BOURIGAULT D., SLODZIAN M. Pour une terminologie textuelle. *Terminologies Nouvelles*, 19 : 29-32, 1999.

BREUKER J., HOEKSTRA R., Core concepts of Law: taking Common Sense seriously. In *Formal Ontology in Information Systems.* A.C. Varzi and L. Vieu (Eds.). IOS Press. 210-221. 2004

BRUAUX S., KASSEL G. and MOREL G., An ontological approach to the construction of problem-solving models. LaRIA Research Report 2005-03, University of Picardie Jules Verne. Available at http://hal.ccsd.cnrs.fr/ccsd-00005019 , 2005.

BUITELAAR P., OLEJNIK D., SINTEK M., A Protégé Plug-In for Ontology Extraction from Text based on Linguistic Analysis. In *Proceedings of the 1st European Semantic Web Symposium*. Héraklion (Greece), May 2004.

BUITELAAR P., CIMIANO P. and MAGNINI B.*, Ontology Learning from Text: Methods, Evaluation and Applications*. Volume 123, Frontiers in Artificial Intelligence and Applications. IOS Press, 2005.

CABRÉ M.T., ESTOPA R., VIVALDI J., Automatic Term Detection: a review of current systems. In *Recent Advances in Computational Terminology*. D. Bourigault, C. Jacquemin, MC. L'Homme Eds. John Benjamins. 53-87. 2000.

CAHIER J.P., ZACKLAD M., MONCEAUX A., Une application du Web socio-sémantique à la définition d'una nnuaire métier en ingénierie. *Actes de la 15e conférence d'Ingénierie des Connaissances IC 2004,* Lyon (F), mai 2004. Grenoble : PUG. 29-40. 2004.

CAMILLERI G.. *Une approche, basée sur les plans, de la communication dans les systèmes à base de connaissances coopératif.* Thèse de doctorat, Université Paul Sabatier, école doctorale d'Informatique et Télécommunications., 118, route de Narbonne. 31062 Toulouse Cedex, décembre 2000. Accès: http://www.irit.fr/recherches/MODEL/CSC/Publis/TheseCamilleri.pdf

CHARLET J., *L'ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales.* Mémoire d'habilitation à diriger des recherches en Informatique de l'université de Pierre et Marie Curie. Décembre 2002.

CHARLET J., LAUBLET P., REYNAUD C., *Web sémantique*, rapport final de l'action spécifique 32 du RTP-DOC du CNRS/STIC. Déc. 2003.

CHARLET J., LAUBLET P., REYNAUD C., (Eds.) Numéro spécial « Web sémantique », *Revue I3*, Cépadues-Editions, 2005.

CIMIANO P., PIVK A., SCHMIDT-THIEME L., STAAB S., Learning taxonomic relations from heterogeneous sources. In *Proceedings of the ECAI 2004 Ontology Learning and Population Workshop*. 2004

CIMIANO P., VÖLKER J., Text2Onto, a framework for Ontology Learning and data-driven Change Discovery. Submitted. 2005.

CIRAVEGNA F., DINGLI A., PETRELLI D., WILKS Y., User-system cooperation in document annotation based on information extraction. *In* Gómez-Pérez A., Benjamins V.R., (Eds), *proceedings of the 13th International conference in Knowledge Engineering and Knowledge Management, EKAW 2002*. LNAI 2473. Berlin : Springer Verlag, 2002

CIRAVEGNA F. & CHAPMAN, Mining the Semantic Web: Requirements for machine learning. In *Machine Learning for the Semantic Web*, Dagstuhl Seminar 05071. Dagstuhl (Germany), 13-18 feb. 2005.

CONDAMINES A. *Sémantique et corpus spécialisés : Constitution de bases de connaissances terminologiques.* Mémoire d'habilitation à diriger des recherches en Linguistique de l'université de Toulouse 2. Carnets de grammaire de l'ERSS N°13 - Octobre 2003.

CONDAMINES A. , AMSILI P., Terminology between language and Knowledge : an example of Terminological Knowledge Base. In *Proceedings of TKE'93 (Terminology and Knowledge Engineering).* Schmitz K. (Ed.). Franckfurt : Indeks Verlag : 316-323. 1993.

CONDAMINES A., REBEYROLLES J. , Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode. In Charlet J, Zacklad M., Kassel G. & Bourigault D. (eds.) *Ingénierie des connaissances. Tendances actuelles et nouveaux défis.* Paris : Eyrolles/France Telecom. 2000.

CORCHO O. , FERNÁNDEZ-LÓPEZ M., GÓMEZ-PEREZ A., VICENTE O., WebODE : an Integrated Workbench for Ontology Representation, Reasoning and Exchange. In Gómez-Pérez A., Benjamins V.R., (Eds), *Proceedings of the 13th International Conference in Knowledge Engineering and Knowledge Management (EKAW'02*). Sigüenza, Spain. LNAI 2473. Berlin : Springer Verlag. 138-153. 2003.

DAILLE B., *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. Thèse en Informatique Fondamentale, Université de Paris 7, Paris. 1994.

DAOUST, F., *SATO (Système d'analyse de texte par ordinateur), Version 4.0, Manuel de référence*, Service d'analyse de texte par ordinateur (ATO), Université du Québec à Montréal, 256 p. http://www.ling.uqam.ca/sato 1996

DAVID, S., PLANTE, P., *Termino version 1.0*. Rapport de recherche du Service d'Analyse de Textes par Ordinateurs (ATO), Université du Québec à Montréal. http://www.ling.uqam.ca/nomino , 1990.

DESCLES J.-P., MINEL J.-L., L'exploration contextuelle, *Le résumé par exploration contextuelle*, recueil des communications effectuées aux rencontres Cogniscience-Est, 25 nov. 1994. Nancy. Rapport interne du CAMS n°95/1. 3-17. 1994

DOMINGUE, J., Tadzebao and WebOnto: Discussing, Browsing, and Editing Ontologies on the Web. In B. Gaines and M. Musen (eds), *Proceedings of the 11th Knowledge Acquisition for Knowledge-Based Systems Workshop*, April 18th-23th 1998, Banff, Canada. 1998. http://kmi.open.ac.uk/people/domingue/banff98-paper/domingue.html.

ENGUEHARD C., PANTERA L., Automatic natural acquisition of terminology. *Journal of quantitative linguistics*. Vol. 2, N°1 : 27-32, 1995.

EUZENAT J., Hytropes : a www front-end to an object knowledge management System. In *proceedings of the 9th Knowledge Acquisition Workshop, KAW'96*, Fiche démonstration, Banff, Canada, 1996.

FAURE D., *Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM*, thèse de Doctorat Université de Paris Sud. 2000.

FAURE D., POIBEAU T., First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX. In proc. *of Ontology Learning ECAI-2000 Workshop*. 7-12. 2000.

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

FENSEL D., Ontologies : a silver bullet for knowledge management and electronic commerce . Eds: Springer , , 2001

FERNÁNDEZ-LÓPEZ M., GÓMEZ-PEREZ A., JURISTO N., METHONTOLOGY: From Ontological Arts Towards Ontological Engineering, *AAAI 97 Springs Symposium Series on Ontological Engineering*, Stanford USA. 33-40, 1997.

FRIDMAN-NOY N., HAFNER C., The state of the Art in Ontology Design : a Survey and Comparative Review, *Artificial Intelligence Magazine*, Fall, 53-74, 1997.

FRIDMAN-NOY N., FERGESON R., MUSEN M., The knowledge model of Protege-2000 : combining interoperability and flexibility. In Dieng R., Corby O. (Eds.) *Proceedings of the 1éth International Conference in Knowledge Engineering and Knowledge Management (EKAW'00*). Juan-Les-Pins (France). LNAI 1937. Berlin : Springer Verlag. 17-32. 2000.

GARCIA D. *Analyse automatique des textes pour l'organisation causale des actions, Réalisation du système COATIS*. Thèse d'informatique, Université Paris IV-Sorbonne, 1998.

GARCIA D., JACKIEWICZ A., Aide à l'acquisition des connaissances causales à partir de textes, *6èmes Journées d'Acquisition des Connaissances*, Grenoble, avril 1995.

GARDIN, J.C.. Le calcul et la raison. Essais sur la formalisation du discours savant. Paris : Editions de l'EHESS. 1991.

GARDIN, J-C. & ROUX, V., The Arkeotek project : a European network of knowledge bases in the archaeology of techniques. *Archeologia e Calcolatori*, 15, 25-40. 2004.

GILLAM L., TARIQ M., K. A. SMADJA, Terminology and the construction of ontology. *Terminology*. 2005.

GÓMEZ-PÉREZ A., FERNÁNDEZ-LÓPEZ M, CORCHO O., *Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Verlag, London. 2004.

GREFENSTETTE G. , *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher, Boston, MA. 1994

GRUBER T.R., The role of common ontology in achieving sharable, reusable knowledge bases. *Proc. Of the 2nd Int. Conference on the Principles of Knowledge Representation and reasoning :* Morgan Kaufmann, San Mateo (CA, USA) : 601-602, 1991.

GRUBER T. R, Translation approach to portable ontology specifications. *Knowledge Acquisition,* **5**, 199-220. 1993

GRUBER T.R., Toward principles for the design of ontologies used for knowledge sharing. In N. Guarino and R. Poli (Eds.) *Int. WS on Formal Ontology in Conceptual Analysis and Knowledge Representation*. Padova (I.) Kluwer : 1993.

GUARINO N. Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human-Computer Studies.* **43** (5/6) : 625-640. 1995

GUARINO N. Formal ontology in Information Systems. In *proceedings of the 1st Formal Ontology in Information Systems.* Trento (It.), N. Guarino (Ed). Amsterdam: IOS Press. 3-15. 1998

GUARINO N. GARRIETTA P., Ontologies and Knowledge Bases: Towards a Terminological Clarification. In Mars N. (Ed.) *Towards very large Knowledge Bases (KBKS'95)*. Amsterdam : IOS Press. 25-32. 1995.

GUARINO N., WELTY C., A formal Ontology of Properties. In Dieng R., Corby O. (Eds.) 12th *International Conference in Knowledge Engineering and Knowledge Management* (EKAW'00). Juan les Pins (F). LNAI 1937. Springer-Verlag. 97-112. 2000.

HAHN U., MARKO K.G. Ontology and Lexicon Evolution by Text Understanding In *proceedings of the ECAI 2002 Workshop 16 Natural Language Processing and Machine Learning for Ontology Engineering* . Lyon (F), July 2002.

HAMON T., *Vérification sémantique en corpus spécialisé : acquisition de relations de synonymie à partir de ressources lexicales.* Thèse d'informatique de l'université Paris 13, Villetaneuse (F.), 2000.

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

HEARST M., Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proc. of the 15th international Conference On Computational Linguistics (COLING-92),* Nantes (F). 539-545. 1992.

JACQUES M.-P., SOUBEILLE A.-M. : Partage des termes, partage des connaissances ? Construire une modélisation unique de plusieurs corpus. In *Actes de la conférence francophone d'Ingénierie des connaissances IC 2000*, Toulouse (F), 313-314 . 2000.

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

JOUIS C., *Contributions à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes,* Thèse de doctorat de l'université Paris 3 Sorbonne, Paris : EHESS, 1993.

KAVANAGH, J., *The Text Analyzer : a Tool for Extracting Knowledge from Text*. Master's of computer science Thesis, Univ. of Ottawa (Can). 1996

LAME G., *Construction d'ontologie à partir de textes. Une ontologie du Droit français dédiée à la recherche d'information sur le Web*, thèse de l'école des Mines de Paris, 2002.

LEBARBE T. (2002), Hiérarchie Inclusive des Unités Linguistiques en Analyse Syntaxique Coopérative, *Thèse de* Doctorat, Université de Caen.

LEMAIRE F., RECHENMANN F. 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.** 103- 112, 1995.

NECHES R., FIKES R.E., FININ T., GRUBER T.R., SWARTOUT W., Enabling technology for knowledge sharing. *Artificial Intelligence*, 12(3): 36-56, 1991.

MAEDCHE A. , *Ontology learning for the Semantic Web*. Kluwer Academic Publisher. 2002.

MAEDCHE A. & STAAB S., Semi-automatic Engineering of Ontologies from Text, in *Proceedings of the Twelfth Inernational Conference on Software Engineering and Knowledge Engineering (SEKE'2000).* 2000.

MAEDCHE A. & STAAB S., Ontology learning. In S. Staab and R. Studer (eds) *Handbook on Ontologies*. Springer Verlag, 173-189, 2004.

MALAISE V., ZWEIGENBAUM P. & BACHIMONT B., Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologie. In P. Blache (Ed.), *Actes de TALN 2004 (Traitement automatique des langues naturelles),* 269–278, Fès (Maroc). ATALA LPL . 2004

MARTIN P., Knowledge Acquisition using Documents, Conceptual Graphs and a Semantically Structured Dictionary *Proc. of KAW95, Knowledge Acquisition for Knowledge-Based Systems Workshop.* Banff (Can). 1995.

MASOLO C., BORGO S., GANGEMI A., GUARINO N., OLTRAMARI A. and SCHNEIDER L., *The WonderWeb Library of Foundational Ontologies and the DOLCE ontology*. WonderWeb Deliverable D18, Final Report (vr. 1.0, 31- 12-2003). 2003.

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

MEYER, I., SKUCE, D., BOWKER, L. & ECK, K., Toward a new generation of terminological resources: an experiment in building a terminological knowledge base. In *Proc. 13th International Conference on Computational Linguistics*. Nantes : 956-960. 1992.

A. MIKHEEV & S. FINCH, A Workbench for Acquisition of Ontological Knowledge from Natural Language, in Proc. of the *9th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop* (KAW'95). 1995.

MORIN E. , Des patrons lexico-syntaxiques pour aider au dépouillement terminologiques, *Traitement Automatique des Langues*, **40**, Numéro 1, 143-166. 1999.

NANARD M., NANARD J., MASSOTTE A. M., CHAUCHE J., DJEMAA A., JOUBERT A. et BETAILLE H., La métaphore du généraliste : Acquisition et utilisation de la connaissance macroscopique sur une base de documents techniques, in *Acquisition et ingénierie des connaissances : tendances actuelles*, Cépaduès Editions, Toulouse. 285-305, 1996.

NAZARENKO A., ZWEIGENBAUM P., HABERT B. & BOUAUD J. , Corpus-based extension of a terminological semantic lexicon, in Bourigault D., Jacquemin C. & L'Homme M.-C., *Recent advances in computational terminology*, John Benjamins Publishing, Amsterdam, pp 327-352. 2001.

NAVIGLI R., VELARDI P., Learning domain ontology from document warehouses and dedicated web sites. *Computational Linguistics* (**50**)2, 2004.

NAVIGLI R. VELARDI P. CUCCHIARELLI, A., NERI F. Quantitative and Qualitative evaluation of the OntoLearn Ontology Learning System. In *WS on Ontology Learning and Population (OLP) at ECAI 2004.* Valence (Esp.) 2004.

NEDELLEC C., (2004), Machine Learning for Information Extraction in Genomics – State of the art and perspective. In *Text Mining and its Applications: Results of the NEMIS launch Conference.* Series : Studies in Fuzziness and Soft Computing, Ed. Sirmakessis and Spiros, Springer Verlag.

POIBEAU T. (1999) , Repérage des entités nommées, enjeux pour les systèmes de veille. *Terminologies Nouvelles*, 19 : 43-51.

RASTIER F. (1995), Le terme, entre ontologie et linguistique. Dans *Banque des mots. N° spécial Terminologie et Intelligence Artificielle.* **7.** 35 – 64.

REBEYROLLE J., TANGUY L. (2000), Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire*, n°25, Université Toulouse - Le Mirail : 153-174.

REIMER U. (1990), Automatic knowledge acquisition from texts: Learning terminological knowledge via text understanding and inductive generalization. In *Proc. of the Workshop on Knowledge Acquisition for Knowledge-Based Systems (KAW'90) :* 27.1-27.16.

REINBERGER M.-L., SPYNS P. (2004), Discovering Knowledge in Texts for the Learning of DOGMA-inspired ontologies. In *proceedings of OLP04 (workshop on Ontology learning and Population at ECAI04),* August 2004, Valencia (E.), 19-24.

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

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

ROUX V., BLASCO P. (2004). Faciliter la consultation de textes scientifiques. Nouvelles pratiques éditoriales. *Hermès, Critique de la raison numérique*, CNRS éditions, 39, 151-159.

SKUCE D., A frame-like knowledge representation integrating abstract data-types and logic. In J. Sowa Ed. *Principles of Semantic Networks*. Morgan Kaufmann Publishers. 543-563. 1991.

SKUCE D., Intelligent Knowledge Management: Integration of Documents, Knowledge Bases, Databases and Linguistic Knowledge, *Proc. of the 10th Knowledge Acquisition and Management Workshop (KAW'98*), Univ. of Calgary, Banff, Canada, 1998.

SKUCE D., IKARUS: Intelligent Knowledge Acquisition and Retrieval Universal System. Technical reports. http://www.csi.ottawa.ca/~kavanagh/Ikarus/Ikarus4.html

SKUCE, D, MEYER, I., Terminology and knowledge acquisition: exploring a symbiotic relationship. In *Proc. 6th Knowledge Acquisition for Knowledge-Based Systems Workshop*, Banff. 29/1-29/21. 1991.

SLODZIAN M., Comment revisiter la doctrine terminologique aujourd'hui ? *La banque des mots*, **7**/95 :11-18, 1995.

SLODZIAN M. , L'émergence d'une terminologie textuelle et le retour du sens, in *Le sens en terminologie*, publication du Centre de Recherche en Terminologie et Traduction de l'Université Lyon 2. 2000.

SMITH B., Basic Concepts of Formal Ontology. In *Formal Ontology in Information Systems*. N. Guarino (ed.). IOS Press. 19-28. 1998.

SOWA J. (Ed.), *Principles of Semantic Networks*. Morgan Kaufmann Publishers. 1991.

SOWA J. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Col Publishing Co. Pacific grove, CA, 2000.

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

STUDER R., BENJAMINS R., FENSEL D. Knowledge Engineering: principles and methods. *IEEE transactions on Data and Knowledge Engineering.* 25 (1-2): 161-197. 1998.

SARTOUT B., PATIL R., KNIGHT and RUSS T. Towards Distributed use of large-scale Ontologies. Spring Syposium Series on Ontological Engineering. Stanford University, CA. 138-148. 1997.

SZULMAN S., BIEBOW B., OWL et TERMINAE. *Actes des 15° journées francophones d'Ingénierie des Connaissances (IC 2004)* Lyon (F.), Presses Universitaires de Grenoble : 41-52, 2004.

TOUSSAINT Y., NAMER F., DAILLE B., JACQUEMIN C., ROYAUTE J., HATHOUT N. , Une approche linguistique et statistique pour l'analyse de l'information en corpus. Actes *de la 5ème conférence annuelle sur le Traitement Automatiques des Langues Naturelles (TALN'98)*, Paris (F), 182-191, 1998.

TRONCY R., ISAAC A., DOE: une mise en œuvre d'une méthode de structuration différentielle pour les ontologies*, Actes des 13ièmes journées francophones d'Ingénierie des Connaissances IC 2002*, Rouen (F), 63-74. 2002.

USCHOLD M. M., GRUNINGER M. Ontologies : principles, methods and applications. *Knowledge Engineering Review*. 1996

VALENTE A., BREUKER J., Towards principled Core Ontologies. In *Proceedings of KAW'96*, Calgary, University of Calgary. 1996.

VAN HEIJST G. *The role of Ontologies in Knowledge Engineering.* Thesis Universiteit van Amsterdam. 1995.

VARGAS-VERA M., MOTTA E., DOMINGUE J., LANZONI M., STUTT A., CIRAVEGNA F., MnM: Ontology driven semi-automatic or automatic support for semantic markup. In *Proceedings of the 13th International Conference onKnwoledge Engineering and Knowledge Management* EKAW'02. Springer Verlag. LNAI 2002.

VELARDI P., MISSIKOFF M., BASILI R., Identification of relevant terms to support the construction of domain ontologies. *ACL WS on Human Language Technologies and Knowledge Management*. Toulouse (F), 18-28, 2001.

VIRBEL J. (2002) *Éléments d'analyse du titre*. Dans : *Inscription Spatiale du Langage*, *Toulouse*, 29 - 30 janvier 2002. IRIT, 123-134.

WOODS D. (1975), What's in a link. In D.G. Brobow and A. Collins (eds.) *Representation and Understanding*, Academic Pres, 35-82.