Chapitre 7

Entre textes et ontologies formelles : les bases de connaissances terminologiques

Nathalie Aussenac-Gilles (IRIT – Toulouse)
Anne Condamines (ERSS – Toulouse)

1. Textes et capitalisation des connaissances : place de la linguistique

A travers les documents qu'elles produisent, les entreprises rendent explicite une partie de leur savoir-faire et de leur métier. Qu'il s'agisse de nomenclatures de méthodes et produits, de descriptions de procédures, de notes ou de comptes-rendus de projets, ces documents accompagnent une grande variété d'activités réalisées dans une entreprise. Il est donc naturel, dans une démarche de capitalisation puis de meilleure diffusion des connaissances au sein de l'entreprise, de chercher à exploiter ces documents, de les rendre accessibles en adéquation avec des besoins ciblés ou encore de les gérer de manière cohérente. Cette tendance est encore plus nette depuis que ces documents sont produits sur support informatique, ce qui facilite leur stockage, leur consultation automatique, leur mise à jour et leur diffusion.

Nous allons montrer dans cette introduction que l'exploitation de documents à l'état brut se heurte rapidement à des limites, et qu'un modèle de leur contenu s'avère une aide précieuse pour mieux y accéder. Nous soulignerons alors l'apport d'une analyse linguistique pour construire ces modèles semi-automatiquement à partir des textes. Ce cadre étant posé, nous justifierons dans la partie 2 de l'article la nécessité d'une représentation non formelle de ces modèles. Nous introduirons alors la notion de base de connaissances terminologiques (Bct), notion au centre des

collaborations entre linguistique, terminologie et intelligence artificielle (IA) autour des connaissances dans les textes. Nous présenterons dans la partie 3 une synthèse des travaux relatifs à leur constitution et à leur utilisation. Enfin, dans la partie 4, nous insisterons sur les aspects méthodologiques. Nous nous appuierons sur nos expériences de développement de Bct pour relever les problèmes pratiques et théoriques qui apparaissent lorsque l'on applique cette analyse en deux temps, d'un corpus de textes à un modèle et de ce modèle à une (ou plusieurs) application(s).

1.1. Pourquoi faire des modèles à partir de textes ?

1.1.1. Les textes, sources de connaissances pour construire des modèles

Les textes contenus dans les documents techniques véhiculent une description d'activités importantes pour l'entreprise. La diversité de ces types de textes (supports pédagogiques, description de consignes et procédures, comptes rendus de projets, etc.) en font l'une des premières sources d'information dès qu'il s'agit de mieux connaître l'entreprise et ses activités, de définir de nouveaux outils de travail, et en particulier des logiciels, ou de mettre en place des organisations mieux adaptées à la production. Une analyse précise de leur contenu révèle des connaissances de résolution de problème, des descriptions de produits ou de méthodes. Les fruits de ces analyses peuvent être des représentations structurées, des modèles plus ou moins formels qui caractérisent ou représentent les connaissances contenues dans le texte.

1.1.2. Augmenter la qualité de l'accès au texte par les modèles

La mise sur support informatique des documents et l'utilisation de systèmes de consultation documentaire permettent d'en faciliter l'accès et d'y rechercher plus rapidement des informations. Mais jusqu'à présent, les principes de ces systèmes de consultation, basés sur la recherche mot à mot des termes de la requête dans le texte, en rendaient l'efficacité limitée. Pour devenir plus performants, ces outils sont désormais dotés d'une représentation structurée du vocabulaire du domaine, du plus général au plus spécifique, qui permet d'étendre ou de spécialiser les requêtes. De même, plus un document dispose de modes d'accès variés et tenant compte de son contenu (table des matières, index hiérarchisé, etc.) plus son contenu est accessible. Sa lecture n'est plus linéaire mais guidée par les relations sémantiques du modèle. La modélisation du contenu d'un document offre donc une vue structurée qui permet d'imaginer des outils plus variés pour leur consultation ou leur exploitation.

Finalement, nous avons mis en évidence les deux facettes d'un même enjeu : les documents comme sources de modèles, les modèles comme points d'entrée dans les documents. Plusieurs travaux de recherche font désormais l'hypothèse que, au-delà

d'une lecture intuitive, une analyse linguistique systématique améliore la qualité des modèles. Nous montrons par la suite qu'elle doit être considérée comme une étape, à faire précéder du choix judicieux du corpus et à compléter d'une interprétation des données selon des critères pragmatiques liés à la finalité de la modélisation.

1.2. Intérêt d'une approche linguistique pour analyser le contenu des textes

L'analyse linguistique des textes permet de systématiser la recherche des données conceptuelles dans les corpus en la basant sur des connaissances linguistiques. Elle se fonde sur une interprétation d'éléments de forme (apparition de tel ou tel lexème, ayant tel ou tel rôle syntaxique) auxquels est associé un contenu. Le principal objectif de ces études est le repérage et l'identification de relations conceptuelles à l'aide de marqueurs linguistiques qui peuvent être soit définis *a priori*, soit identifiés à partir des données du corpus. Dans une perspective de sémantique linguistique, l'hypothèse est que l'interprétation peut être faite seulement par rapport au contenu du corpus et aux connaissances sur le fonctionnement linguistique. En ce sens, d'ailleurs, cette linguistique sur corpus constitue un apport très important à la réflexion méthodologique en linguistique. Dans une perspective plus applicative, l'interprétation peut être guidée par la tâche. Dans les deux cas, il s'agit de construire un (ou des) réseau(x) relationnel(s) sensé(s) représenter le contenu du texte.

La recherche de systématicité permet de développer une réflexion sur les possibilités d'automatisation, en tout cas, de mettre en évidence les analyses qui peuvent être faites automatiquement et celles qui, assistées par des outils d'analyse de corpus, exigent une interprétation humaine. La perspective de l'analyse linguistique amène ainsi un point de vue original sur ces outils et peut en constituer une forme de validation (ou d'invalidation). En retour, les résultats obtenus sur corpus peuvent être intégrés aux outils pour en améliorer la couverture.

1.3. Rôle central des corpus en terminologie, linguistique et IA

Ces dernières années, et ce n'est sans doute pas fortuit, on a assisté à une évolution parallèle du rôle des corpus en terminologie, en linguistique et en IA. Cette évolution a conduit à un rapprochement de la terminologie et de la linguistique, et à des collaborations d'un nouveau type entre IA et linguistique.

1.3.1. Linguistique et terminologie

Jusqu'à il y a une dizaine d'années seulement, la terminologie n'était pas considérée comme une discipline linguistique, essentiellement parce que son objectif

était complètement appliqué : il s'agissait le plus souvent d'établir des listes de termes d'un domaine avec leurs équivalents dans une ou plusieurs langues, à des fins de traduction. Dans cette pratique, les textes étaient utilisés, mais de façon non systématique, comme une source d'information parmi d'autres. L'évolution vers l'exploitation de textes en format électronique s'est faite de façon assez naturelle et ils constituent désormais la principale source de données terminologiques. Dans le même temps, surtout dans certains milieux universitaires français, une réflexion a débuté sur les possibilités d'exploiter les corpus ; la terminologie s'est ainsi rapprochée de la linguistique, se cantonnant moins à des applications de traduction.

Pour la linguistique, le travail sur corpus représente une mutation beaucoup plus importante que pour la terminologie. Il ne s'agit plus de décrire le système linguistique d'une langue, en se basant sur une approche introspective, mais de rendre compte d'usages réels, tels qu'ils se manifestent dans un corpus. On ne peut plus faire abstraction des éléments concernant la situation d'énonciation : qui a écrit tel texte, dans quel objectif ? Il s'agit désormais de trouver de nouvelles méthodes d'analyse de textes, qui combinent à la fois la description de phénomènes tels qu'ils apparaissent en corpus et l'interprétation, voire l'extrapolation de ces phénomènes pour évaluer ce qui peut être systématisable dans ces observables. Dans cette prise en compte des corpus, la linguistique rencontre inévitablement la terminologie. En effet, c'est particulièrement dans les entreprises que des corpus sont disponibles et la demande de traitement à leur sujet est très forte. Ainsi, même si l'activité terminologique continue à concerner surtout le domaine de la traduction et constitue une activité appliquée, la jonction avec la linguistique est faite. Ce rapprochement avec une discipline académique a permis à la terminologie de gagner ses lettres de noblesse. Désormais, dans certaines d'entreprises (Aérospatiale, EDF…), les services de terminologie et de documentation participent aux projets de Recherche et Développement. C'est aussi en acquérant ce statut de science, via son intégration à la linguistique, que la terminologie a rencontré l'intelligence artificielle.

1.3.2. L'intelligence artificielle et les connaissances en corpus

Si les collaborations entre intelligence artificielle (IA) et linguistique remontent aux fondements de l'IA et sont un des piliers des travaux en sciences cognitives, ce n'est en effet que depuis peu que la terminologie a trouvé un terrain d'étude commun avec l'IA. Par ce biais, les rapports entre linguistique et IA eux-mêmes se sont trouvés renouvelés. En effet, les collaborations de longue date entre ces deux disciplines se focalisaient essentiellement sur la formalisation des théories linguistiques d'interprétation de la langue. Des résultats pointus sont désormais disponibles pour des types d'énoncés spécialisés, comme les raisonnements spatiaux ou temporels. Malheureusement, il existe peu de formalisations applicables de manière générale ou rendant compte de discours complexes.

En revanche, la rencontre entre l'ingénierie des connaissances, en tant que composante de l'IA, et la linguistique, par le biais de la terminologie, semble très prometteuse [CHARLET 00]. Il s'agit de passer d'un niveau linguistique à un niveau conceptuel, de représenter les connaissances véhiculées dans un corpus constitué pour une application donnée, à partir d'énoncés, et ceci pour une application particulière : les modèles produits doivent être pertinents pour développer une certaine application. On ne cherche pas à rendre compte systématiquement de tous les énoncés, mais des connaissances qu'ils révèlent et qui sont utiles à l'application visée. Ici, la séparation claire entre termes et concepts, entre niveaux linguistique et conceptuel, est indispensable. Ainsi, l'IA a repris des structures de données et des méthodes d'analyse des terminologues et des linguistes pour déboucher sur des propositions d'acquisition de connaissances à partir de textes. En retour, les contraintes de structuration et d'automatisation de certains traitements ont conduit linguistes et terminologues à mieux structurer leur démarche, à la systématiser et à préciser la nature et la validité de leurs résultats [CONDAMINES 00b].

1.3.3. Originalité française

La France est particulièrement bien placée dans l'identification des convergences entre ces trois disciplines : linguistique, terminologie, IA. Depuis 1993, le groupe de recherche TIA (Terminologie et Intelligence Artificielle[1]) a beaucoup contribué à cette identification. Il a permis aux chercheurs de ces différentes communautés de se rencontrer régulièrement pour approfondir la connaissance de leurs travaux et amorcer des collaborations précises. Surtout, chacun a pu énoncer les fondements et hypothèses théoriques à la base de son approche. Ainsi, après plusieurs années de fonctionnement, ces membres ont convergé vers une position originale dans la communauté internationale au sujet des ontologies et des modèles de connaissances. Sur ce point, ce groupe se démarque par rapport à ses collègues européens ou nord-américains, lorsqu'il s'agit de mesurer la réutilisabilité et la validité des ontologies. En effet, nos diverses expériences, tant en linguistique qu'en IC, convergent pour souligner la nécessité de prendre en compte très tôt le domaine étudié et l'application qui va utiliser les modèles et les ontologies à produire. Nos travaux soulignent la validité locale (" régionale ") de ces ontologies, difficiles à utiliser pour d'autres applications que celles pour lesquelles elles ont été construites. En revanche, nos méthodes et outils sont tout à fait réutilisables.

[1] http://www.biomath.jussieu.fr/TIA/

2. Quels modèles pour représenter le contenu des textes ?

2.1. Formel versus informel : justification des Bct

2.1.1. Ontologies et modèles formels

Plaçons-nous dans le cas où le modèle visé serait une ontologie formelle. Ce type de modèle se veut indépendant de toute utilisation pour une application visée, et contient donc *a priori* des descriptions consensuelles au moins pour un groupe de personnes de ce qui " existe " dans ce domaine. On retrouve ainsi la notion d'ontologie, dont la définition, en intelligence artificielle et en particulier en ingénierie des connaissances, est l'objet de discussions [CHARLET 00].

Vue comme une spécification de la conceptualisation des connaissances d'un domaine, une ontologie est tantôt définie comme un vocabulaire désignant les entités et relations d'un domaine de connaissances, tantôt comme une description conceptuelle formelle de ces entités et de leurs relations. Les auteurs s'opposent aussi sur la couverture de ce domaine : est-il restreint ou non par l'application qui est envisagée à partir de ces connaissances ? Enfin, pour tout un courant de l'IA (Gruber, Lenat, Studer, Neches), une ontologie doit avoir pour qualités d'être universelle, facile à enrichir, à maintenir et à exploiter : la formalisation est, pour ces auteurs, la clé de ces ambitions [REYNAUD 94] [GOMEZ-PEREZ 99]. Les connaissances y sont représentées sous forme de réseaux sémantiques, de graphes conceptuels, de frames ou en logique (logique de description ou frame logics). Mais concrètement, la réutilisation d'une ontologie pour une application donnée se limite souvent au vocabulaire. En effet, plusieurs travaux montrent que cette formalisation est aussi une limite, en particulier à l'universalité ou à une certaine réutilisabilité [LEROUX 96] [CHARLET 96] [BACHIMONT 00].

Nous prendrons donc *ontologie* dans son sens le moins ambitieux, comme une description structurée des concepts et relations d'un domaine, représentés formellement et indépendamment du raisonnement, même si ces connaissances, elles, ont été choisies pour leur adéquation aux besoins de ce raisonnement ou au point de vue d'une classe de personnes [LEROUX 96]. Ce cadre permet d'envisager tout à fait raisonnablement l'exploitation de modèles issus du texte. Construire l'ontologie requiert un travail d'inventaire, de repérage et de structuration de concepts et de relations, qu'il est fastidieux de réaliser manuellement ou avec un expert. L'exploitation de documents, quand ils existent, constitue intuitivement un gain de temps que des expériences récentes ont confirmées [AUSSENAC-GILLES 99a]. Les textes fournissent des connaissances stabilisées qui garantissent plus de qualité au modèle réalisé. Bien sûr, il convient d'être réaliste et de ne pas prétendre trouver directement dans les textes des pans entiers d'ontologies.

2.1.2. Entre textes et ontologies formelles : les Bct

Les textes sont une des sources de connaissances pour faire des ontologies formelles, mais pas la seule. De fait, un passage direct du texte à une ontologie n'est pas réaliste. Plusieurs expériences montrent qu'une analyse linguistique purement syntaxique et sémantique ne peut suffire à produire une ontologie [CONDAMINES 00a] [MEYER 00] [ASSADI 99]. En effet, tout d'abord, l'organisation formelle dans l'ontologie met l'accent sur la classification et la différenciation des concepts entre eux, au sein d'une hiérarchie de types ou de classes. Or ce type de critère n'est pas présent pour décider de l'organisation des concepts par les linguistes, tout simplement parce que le texte lui-même ne reflète pas systématiquement une telle hiérarchie. Ensuite, l'intervention d'un expert du domaine s'impose, la prise en compte de la finalité de cette ontologie et de points de vue spécifiques obligent à s'éloigner du texte et du seul résultat de l'analyse linguistique. Ceci revient à prendre en compte une dimension pragmatique forte. A l'inverse, tous les exemples illustrant plus haut l'intérêt de construire des modèles du contenu de textes montrent que les résultats de l'analyse linguistique d'un texte peuvent servir à d'autres applications que la construction d'ontologies. Pour cela, il est intéressant de les matérialiser, de les considérer comme des résultats à part entière.

Ce rôle clé est joué par les Bases de Connaissances Terminologiques (Bct), structures de données initialement définies pour renforcer les liens entre réseaux conceptuels et textes en les doublant d'une couche d'informations terminologiques. Leur contenu et leur degré de formalisation ont évolué depuis les premières réalisations [SKUCE 94], [CAPPONI 95], qui correspondaient à des ontologies formelles enrichies de données terminologiques. Notre définition rapproche le contenu des Bct vers les textes. Nous parlerons de Bct-Corpus.

2.1.3. Structures de données, méthodes et enjeux théoriques

Notre analyse a un double impact, l'un sur la méthode de constitution et d'utilisation des Bct, l'autre sur leur contenu (tant la représentation des connaissances que les connaissances représentées). Du point de vue méthodologique, nous avons donc ressenti la nécessité de différencier deux phases, une première consacrée à l'exploitation d'outils et de principes linguistiques pour dégager une représentation du contenu du texte relativement neutre, une deuxième dédiée à la définition plus formelle d'un réseau conceptuel pertinent et valide pour une utilisation ciblée. Ce choix présente l'avantage de bien situer l'analyse linguistique en amont de la formalisation. En tant que structure de données, nous considérons que les Bct contiennent le résultat de la première étape, structuré sous forme d'un réseau sémantique auxquels sont associés des termes et leurs descriptions, en lien avec leurs occurrences dans le corpus. Nous montrerons que les Bct sont de préférence non

formelles, en particulier pour élargir le champ des applications envisageables, selon une analyse analogue à celle de B. Bachimont [BACHIMONT 00].

L'étude et la définition des Bct est 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 " tel que le révèle l'analyse linguistique, les Bct sont le lieu d'apports mutuels de ces disciplines. Etudier leur construction et leur utilisation permet de dissocier les problèmes qui relèvent du linguistique de ceux découlant 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 d'une ontologie. 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.

Afin d'évaluer la pertinence du choix théorique des Bct, nous l'avons éprouvée sur un corpus à partir duquel une Bct a été construite puis nous avons évalué l'utilisabilité de cette Bct sur différents types d'applications. Les parties 3 et 4 rendent compte de cette expérience en séparant le travail d'extraction et de modélisation de connaissances (partie 3) du travail de formalisation (partie 4).

2.2. Quelques systèmes de gestion de Bct

Plusieurs logiciels ou langages existent aujourd'hui pour constituer des Bct. Nous en présentons quelques-uns, essentiellement des prototypes universitaires aux potentiels très différents, en insistant sur leurs caractéristiques, leur utilisation potentielle et leur contribution à la modélisation de connaissances à partir de textes.

2.2.1. Les utilisateurs ciblés

L'utilisateur peut être un linguiste-terminologue, qui, au fur et à mesure qu'il analyse un corpus, enregistre les termes et les connaissances du domaine qu'il identifie. Dans ce cas, la production intermédiaire d'un réseau notionnel associé à des termes est primordiale, de même que les liens étroits avec le corpus pour justifier des choix de modélisation et des relations sémantiques. En vue de maintenir ces données ou d'appliquer la démarche à un autre corpus, le système doit aussi permettre d'enregistrer les connaissances linguistiques (marqueurs, critères définitoires, etc.) utilisées. Géditerm [AUSSENAC-GILLES 99b], l'interface HTL de

Lexter [BOURIGAULT 96] ou System Quirk [AHMAD 95][2] rentrent dans cette perspective.

Un cogniticien est plus préoccupé de construire le modèle d'une application que de rendre compte fidèlement des données textuelles, même si le lien vers le texte est important pour lui permettre de tracer ses choix de modélisation. Il accorde donc moins d'intérêt à la production d'un réseau notionnel relié aux termes et privilégie sans doute soit une représentation formelle des données, soit leur vérification systématique (au moyen de la formalisation). Par ailleurs, sa démarche va plus directement des données linguistiques vers leur formalisation. C'est ce que proposent, Terminæ [BIEBOW 00][3] ou Dockman[4], qui favorisent une formalisation rapide, en logique de description, des notions indiquées dans les textes.

2.2.2. Prise en compte de l'utilisation qui sera faite de la Bct

Certains outils visent à développer des Bct en sachant à quoi elles vont servir (ces Bct constitueront une partie d'application) alors que d'autres permettent de développer une représentation structurée en vue d'utilisations potentielles non définies. Les premiers privilégient en général la base de connaissances de la Bct, sa formalisation et son utilisation, alors que les seconds s'intéressent plus au processus de construction des Bct et à la composante terminologique.

Ainsi, CODE [SKUCE 94] prévoit que les données modélisées contribuent à la formation d'une base de connaissances ; DockMan suppose que les connaissances modélisées à partir des textes permettront d'y rechercher des connaissances par requêtes sur le modèle. De même, Hytropes [EUZENAT 96] permet de classer, retrouver et sélectionner des objets formels en fonction de critères sémantiques (leurs attributs) et de points de vue. Tous ces modèles sont formalisés en fonction des traitements requis par l'utilisation envisagée. System Quirk, HTL, Terminae et Géditerm, au contraire, sont *a priori* ouverts à plusieurs types d'applications, surtout si l'on s'arrête avant la formalisation. Ils privilégient l'analyse linguistique. L'hypothèse sous-jacente est que les données rendent compte du texte uniquement, de manière « neutre », et ne sont pas choisies en fonction d'une application. Paradoxalement, en étudiant de près les structures de données et les informations associées aux termes et aux concepts, on peut quand même anticiper le type d'application ciblé. Ainsi, System Quirk suppose que les banques de termes produites serviront à la traduction de textes depuis ou vers l'anglais.

---

[2] http://www.mcs.surrey.ac.uk/Research/cs/AI/systemQ/

[3] http://www-lipn.univ-paris13.fr/membres/szulman/TERMINAE.html

[4] http://dkm.csi.uottawa.ca

2.2.3. Couverture du processus de modélisation et degré de formalisation du résultat

Il est important de savoir quelles phases du cycle de modélisation couvrent ces logiciels : identification des termes (analyse linguistique), normalisation (organisation des notions associées), formalisation (représentation formelle des connaissances). Certains systèmes assurent seulement la première partie du processus (extraction et normalisation de notions à partir des données linguistiques), comme Géditerm ou System Quirk. Le réseau notionnel non formalisé et les termes associés forment alors un résultat à part entière, que nous appelons la Bct-corpus. Ces outils facilitent la description des notions à l'aide de relations ou de textes en langage naturel, de manière à traduire leur organisation hiérarchique et leurs critères de différentiation.

D'autres systèmes ne se focalisent que sur la dernière partie du processus, la formalisation, autant pour aider à structurer qu'à exploiter les concepts du modèle. Ainsi, Hytropes est avant tout une interface de saisie de frames. De même, dans le travail de N. Capponi, un modèle terminologique est obtenu par une analyse linguistique du corpus (candidats termes trouvés par Lexter et étude de relations lexicales). Puis, dans ce modèle, sont recherchés des critères de différenciation pour organiser les concepts et les représenter en logique de description. On peut reprocher à ce type de travaux de négliger l'étape de normalisation, au sens que lui donne [BACHIMONT 00] en n'incitant pas l'utilisateur à la traiter en tant que telle.

Enfin, certains systèmes cherchent à couvrir l'ensemble du processus. Ainsi Terminae ou, d'une autre manière, la plate-forme associée à la démarche ACI [ASSADI 98] assurent un suivi depuis l'analyse du contenu des textes jusqu'à la formalisation des informations qui y sont trouvées. Les deux plates-formes ont en commun de s'appuyer sur les résultats de Lexter, à savoir une liste de mots du corpus, simples et composés, pouvant être des termes. Terminae, de manière analogue à System Quirk ou Géditerm, permet de décrire les termes et les notions identifiés de manière structurée sous forme de fiches. Plus actif, ACI intègre des outils comme Lexiclass qui s'appuie sur des analyses statistiques pour suggérer des classes conceptuelles ou des champs sémantiques. Le réseau conceptuel résultat, représenté à l'aide de frames, est donc une ontologie formelle, régionale (par sa validité) et documentée (grâce aux liens vers les textes).

Le fait d'utiliser une représentation formelle des connaissances terminologiques permet, lors de la construction d'une Bct, 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. La contribution de l'Intelligence Artificielle est ici significative : la plupart des formalismes utilisés

sont inspirés des réseaux sémantiques, comme les logiques de descriptions (Terminae et application de N. Capponi) ou les graphes conceptuels (CGKAT [MARTIN 95]). Hytropes utilise des frames (objets du langage Tropes) et présente l'originalité de rendre compte de points de vue sur les objets. Par contre, la formalisation impose des contraintes, oblige à faire des choix en tenant compte de la façon dont les connaissances seront utilisées et comment le formalisme sera interprété.

On constate deux approches possibles : soit le système favorise l'analyse, la structuration puis la formalisation rapide des données par sous-ensembles (démarche que nous qualifierons d'« horizontale ») ; soit la démarche (dite « verticale ») considère chaque étape de la modélisation de l'ensemble des données comme produisant un résultat intermédiaire et accessible. Ainsi, le réseau notionnel peut être manipulé en tant que tel dans Géditerm, mais pas dans ACI ou Hytrope, où ces modèles ne sont pas exploitables pour d'autres finalités.

2.2.4. Complexité et richesse du modèle de données, prise en compte du corpus

Les outils de constitution de Bct se différencient également par les structures de données produites, leur degré de formalisation et la richesse de leur représentation des connaissances. Un des indicateurs de la richesse du modèle des données est la représentation des relations. Même en amont de la formalisation, la structuration des concepts suppose leur connexion par des relations sémantiques. La sémantique se traduit par le nom de la relation, parfois par le type des concepts qu'elle peut associer, comme dans Terminae et Géditerm, où les types des relations utilisables peuvent être complétés et modifiés en fonction du corpus. Un travail très poussé a été mené dans CGKAT pour proposer un ensemble de relations formelles organisées en une hiérarchie. Dans Hytropes ou Code, la seule relation formalisée est EST-UN alors que les autres relations lexicales sont traduites par les attributs des concepts.

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 (comme le système de N. Capponi, CGKAT ou Hytropes) n'intègrent pas le corpus. Toutefois, Terminae et Dockman, accordant 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. A l'inverse, les systèmes centrés sur l'analyse linguistique (alors bien dissociée de la formalisation) privilégient d'abord les termes plutôt que les concepts et permettent de revenir facilement aux occurrences en corpus. Ainsi, System Quirk se focalise sur les termes et leurs occurrences, sans gérer clairement le niveau conceptuel. De même, HTL permet avant tout de parcourir les candidats termes trouvés par Lexter et de revenir à leurs occurrences. Les fonctions relatives à la mise en évidence de concepts et de

relations sont moins avancées. Seul Géditerm permet d'organiser concepts et relations (sans aller jusqu'à la formalisation) après avoir mis l'accent sur l'analyse linguistique.

3. D'un corpus à une Bct : expérience

Cette première étape, du passage d'un corpus à une Bct, pose plusieurs questions méthodologiques. Toutes n'ont pas été résolues, beaucoup nécessitant encore une réflexion de type linguistique (passage d'une occurrence à un type, identification et traitement de la polysémie vs de l'homonymie, nature linguistique des interprétations à effectuer …). Cependant, les grandes lignes d'une méthode se dégagent désormais.

3.1. Les outils de l'analyse linguistique

La méthode de constitution des Bct repose sur l'exploration d'un corpus afin d'en modéliser le contenu. Le volume des données à examiner fait que cette exploration ne peut se faire sans l'assistance d'outils. En outre, l'utilisation d'outils permet de mettre en évidence des points de vue plus riches et diversifiés que la seule lecture « manuelle » . Les outils pertinents pour ce type de méthode peuvent provenir des différentes disciplines qui visent à repérer automatiquement des connaissances dans les textes : recherche d'information, documentation électronique (thésaurus), acquisition de connaissances à partir de textes, terminologie … Tous les outils permettant de faire du TAL (Traitement Automatique de la Langue) peuvent avoir un intérêt et c'est d'ailleurs une étude à part entière que d'identifier tous les outils pertinents pour la constitution de Bct ainsi que la façon dont les résultats produits par ces méthodes peuvent être intégrés et interprétés. De façon générale, on peut dire que les outils les mieux adaptés sont d'une part, ceux qui permettent d'explorer des textes et d'autre part, ceux qui sont dédiés à la terminologie : extracteurs de termes-candidats (comme Lexter [BOURIGAULT 96] et Nomino) et de relations-candidates (comme Caméléon [SEGUELA 99] ou Prométhée [MORIN 99]).

3.1.1. Outils d'analyse de textes

Les outils d'analyse de textes ne sont pas conçus pour une application particulière mais proposent des fonctionnalités qui permettent d'explorer le texte : recherche de concordances, calcul de fréquences, possibilité d'attribuer des caractéristiques (syntaxique ou sémantique) aux mots… Nous avons choisi d'utiliser Sato, développé au centre ATO de Montréal[5], qui a pour avantage de disposer d'une

[5] http://www.ling.uqam.ca/sato/outils/sato.htm

base de données lexicales, contenant toutes les catégories grammaticales auxquelles peut appartenir une forme (par exemple, *ferme* peut être un nom, un adjectif ou un verbe). Il n'y a pas d'analyseur syntaxique, ces formes ne sont donc pas désambiguïsées dans les corpus. Malgré cette limite, importante, les catégories grammaticales sont très utiles pour prendre en compte dans les interrogations des contraintes syntaxiques (par exemple, rechercher dans les textes tous les noms précédés d'une préposition).

3.1.2. Outils dédiés à la terminologie

Que ce soit pour les extracteurs de termes-candidats ou pour les extracteurs de relations-candidates, deux types de méthodes peuvent être mises en œuvre par ces outils : une méthode descendante ou une méthode ascendante.

La *méthode descendante* considère que les corpus spécialisés ne sont que des usages du système linguistique d'une langue et que si l'on peut définir *a priori* les fonctionnements de ce système, on peut les utiliser pour identifier certains éléments dans les corpus. Ainsi, certains extracteurs de termes-candidats (par exemple Nomino[6]) partent de l'idée que les termes sont de la forme Nom + adjectif (ou Nom + préposition + déterminant + nom etc.) et recherchent dans les corpus tous les syntagmes de cette forme ; l'utilisateur doit alors faire le tri des syntagmes qu'il va retenir comme termes. Certains outils d'extraction de relations-candidates mettent en œuvre une méthode descendante en recherchant dans les corpus des marqueurs linguistiques de relations pré-identifiées. Par exemple, on sait que des formes comme *tous les Y…sauf X* ou *X est un Y qui* permettent de repérer une relation générique/spécifique entre Y et X. Seek [JOUIS 93] fonctionne sur cette base pour repérer les contextes contenant potentiellement une relation de générique à spécifique ; Coatis [GARCIA 98] fonctionne sur la même base pour repérer des contextes contenant une relation de cause.

La *méthode ascendante* considère, elle, qu'il est très difficile de décider à l'avance ce que l'on va trouver dans les corpus à l'étude et qu'il vaut mieux faire remonter du texte les éléments recherchés. Les extracteurs de termes-candidats qui fonctionnent sur cette base, comme Ana [ENGUEHARD 95], mettent en œuvre une approche statistique dite des segments répétés. Sont ainsi retenus comme termes-candidats les suites de mots qui apparaissent plus d'un certain nombre de fois. Les extracteurs de relations-candidates qui mettent en œuvre une méthode ascendante recherchent les contextes dans lesquels apparaissent des termes dont on sait qu'ils entretiennent telle ou telle relation ou dont on présume qu'ils peuvent entretenir une relation. Sont alors recherchés les éléments qui, récurrents dans ces contextes,

[6] http://www.ling.uqam.ca/nomino/

peuvent servir à exprimer cette relation. Il faut alors interpréter les résultats obtenus pour décider finalement quelle est la relation qui unit les termes et quelle est la structure linguistique qui la marque le mieux. Des outils comme Likes [ROUSSELOT 96], Prométhée [MORIN 99] fonctionnent sur cette base ; Caméléon [SEGUELA 99] utilise les deux approches.

Malheureusement, dans tous les cas, ces outils constituent une aide certes appréciable mais qui ne remplace pas le travail d'analyse approfondie de contextes. Pour cela, les outils d'analyse de textes sont utilisés. Pour une description plus précise de ces outils, on pourra consulter dans la même collection, [BOURIGAULT 00]. Notons enfin que la plupart de ces outils (hormis Sato et Seek) sont des prototypes de laboratoire, ce qui met une sérieuse limite à leur utilisation : problèmes de bogues, mais aussi de disponibilité, en particulier pour les entreprises.

3.2. Méthode de constitution de la Bct

Une vision strictement linguistique des Bct permet de la définir 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. Une Bct est ainsi élaborée sur la base d'une interprétation linguistique de contextes : rapprochement sur la base d'une identification de contenu, repérage de familles de contextes, repérage de contenus différents pour une même forme ...

Pour constituer une Bct, l'analyse linguistique est guidée par le modèle de Bct qui a été défini *a priori*. On connaît ainsi la nature des éléments qui doivent être recherchés dans le texte ; pour résumer, il s'agit des termes et des relations conceptuelles. Divers projets et outils ont été constitués autour de la recherche de ces deux types d'éléments, le plus souvent autour d'un de ces deux types d'éléments. La particularité des travaux toulousains tient à ce que le processus complet d'élaboration d'une Bct à partir d'un corpus a été défini : étapes à suivre, outils à utiliser, connaissances linguistiques à convoquer. D'un corpus à une Bct, trois étapes majeures ont ainsi été décrites : la première concerne le repérage des termes, les deux autres, le repérage des relations conceptuelles.

3.2.1. Repérage des termes

Repérer des candidats-termes revient à trouver dans le texte des syntagmes potentiellement aptes à devenir des termes. De nombreux outils sont dédiés à cette tâche, dont les résultats, bruités ou beaucoup trop volumineux**,** doivent être retriés. Par exemple, à partir d'un corpus fourni par EDF (corpus Mouglis, manuel de génie logiciel) d'environ 50000 mots, 5878 termes-candidats ont été proposés par le

logiciel Lexter qui combine une approche ascendante et une approche descendante [BOURIGAULT 96]. Il est donc nécessaire de réduire ce nombre. Cela est possible sur des bases intuitives. On peut aussi systématiser cette intuition et la fonder sur des bases linguistiques. Ainsi, le nombre de candidats-termes proposés est passé de 5878 à 1516 en appliquant des règles du type :

- Supprimer les candidats-termes contenant des déictiques (*présent document*) ou des anaphoriques (*phase suivante du projet*), ou dont la tête est un nom vague qui joue plutôt le rôle de déterminant (*ensemble de projets*) ;

- Conserver les candidats-termes qui forment des familles (même tête (*test de qualification*, tes*t d'acceptation*…), les candidats-termes contenant des adjectifs équivalents ou opposés (*petit projet/grand projet*…) [CONDAMINES 00a].

3.2.2. Repérage des relations conceptuelles dans le corpus

Que ce soit de manière intuitive ou systématique, le repérage des relations conceptuelles en corpus se fait par l'utilisation de marqueurs linguistiques, c'est-à-dire de formes lexicales, ayant éventuellement une fonction syntaxique, qui, bien que différentes, renvoient à une seule relation conceptuelle. Certaines relations : hyperonymie (ou générique/spécifique), méronymie (ou partie de) sont bien connues et leurs marqueurs assez bien recensés. La difficulté vient de ce que certains corpus font apparaître des marqueurs de relation (ou même des relations), qui leur sont propres et qu'il est souvent difficile de détecter. La méthode mise au point par l'Equipe de Recherche en Syntaxe et Sémantique s'organise en deux temps : constitution de taxinomies et identification de relations propres au corpus.

3.2.3. Constitution de taxinomies.

Les taxinomies sont constituées grâce à la relation générique/spécifique. Cette relation a un fonctionnement tout à fait particulier :

- on la retrouve dans la grande majorité des corpus, elle joue en effet un rôle important de structuration de la connaissance ;

- ses marqueurs les plus courants, qui ont en outre le mérite d'être indépendants d'un domaine ou d'un genre textuel particulier, sont désormais bien connus [BORILLO 97], [MORIN 99], [MEYER 00].

Ces marqueurs d'hyperonymie sont recherchés systématiquement dans le corpus, en combinaison avec les candidats-termes, à l'aide de l'outil Sato. Les portions de texte récupérées sont analysés afin de sélectionner les contextes pertinents, i.e. ceux qui expriment bien une relation d'hyperonymie; des couples de termes sont constitués puis combinés afin de former des arbres. Dans un second temps, les couples de termes retenus sont projetés sur le corpus afin de faire éventuellement

émerger des marqueurs de la relation d'hyperonymie qui seraient propres au corpus [CONDAMINES 00b].

Cette méthode permet de constituer des ensembles de termes dont on sait qu'ils ont au moins un élément sémantique commun (un trait, dans la terminologie des ontologies, un sème dans la terminologie linguistique). Il est très rare qu'un corpus fasse apparaître une taxinomie unique et complète, c'est-à-dire un seul arbre construit grâce à la seule relation générique/spécifique. Le plus souvent, le corpus exprime des portions d'arbre : des taxinomies. Ainsi, dans le corpus Mouglis, 204 paires de termes ont été retenues après application des marqueurs d'hyperonymie, ces 204 termes s'organisant dans quatre taxinomies : activité (*conception détaillée*, document (*dossier de conception détaillée*), humain (*chef de projet*), portion de temps (*phase de conception*).

3.2.4. Identification de relations propres au domaine

Il s'agit ici de croiser les taxinomies afin de faire émerger les relations horizontales entre les termes. L'hypothèse est qu'à l'intérieur du corpus, ces relations se mettent en œuvre entre des éléments de taxinomies. A ce niveau de l'analyse en effet, nous ne pouvons plus que nous appuyer sur les candidats-termes pour faire la recherche de relations puisque nous ne savons rien des relations qui sont utilisées dans ce corpus. Cette façon de procéder permet d'éviter de tester en vrac tous les couples de candidats-termes qui ont été retenus lors de l'étape de sélection. Il s'agit en quelque sorte d'une consultation du corpus organisée, qui permet aussi de ne pas avoir à lire tout le corpus ni surtout à le lire de manière linéaire. Dans le cas du corpus Mouglis, 16 combinaisons ont ainsi été examinées (par exemple candidat-terme humain X candidat-terme document). Il s'est avéré que pour certaines de ces combinaisons, on pouvait donner un sens**,** identifier une relation et un marqueur de cette relation. Au total, 13 relations ont été identifiées de cette façon. Si l'on rajoute la relation d'hyperonymie, ce sont au total 14 relations qui ont été retenues. 526 paires de termes ont été repérées de cette façon.

4. D'une Bct à des applications

Afin d'illustrer l'intérêt de disposer d'une Bct dans un domaine donné pour développer ensuite des applications dans ce domaine, nous allons présenter deux exemples d'utilisation de la Bct décrite dans le paragraphe précédent (modélisation dans le domaine du génie logiciel scientifique et technique, corpus Mouglis). Dans le premier, il s'agit de construire un index du document formant le corpus (le guide Mouglis) à partir de cette Bct. Dans le second, la Bct sert de source à la construction d'un modèle formel des connaissances de ce domaine. En fait, ces exemples sont

tirés d'une expérience menée en collaboration avec la DER d'EDF pour évaluer la pertinence de la notion de Bct [AUSSENAC-GILLES 98]. Ce travail nous a permis de mesurer l'impact de l'application visée sur l'organisation des connaissances, et donc de confirmer le besoin de prendre en compte très tôt cette application dans le processus de modélisation des données tirées d'un texte ou d'une Bct. Lorsqu'elle existe, la Bct permet d'atteindre plus rapidement l'objectif final, mais sa constitution, assez coûteuse, ne se justifie que si plusieurs applications sont envisageables.

4.1. Une application non-formelle : construction d'un index[7]

Conçue en lien étroit avec les textes, la Bct est une source de connaissances disponible pour constituer un index de ces textes. Même s'il n'est pas judicieux de conserver toutes les relations conceptuelles de la Bct, si certains termes de la Bct ne sont pas pertinents pour l'index ou encore si la Bct ne fournit pas vraiment de mots clés, elle offre un ensemble de données riches et utiles pour aider à construire un index. Bien sûr, l'apport de la Bct dépend aussi de la nature de l'index désiré, des critères retenus pour sélectionner les entrées et de leur organisation dans l'index.

Ainsi, à partir du corpus Mouglis, un index a été constitué. Cet index devait être hiérarchisé, contenant des entrées et des sous-entrées plus précises, chacune renvoyant vers des passages de texte. Construire cet index revient à décider des termes à retenir comme entrées et sous-entrées, à les organiser puis à sélectionner les occurrences associées. Nous avions plusieurs hypothèses sur les aides que pouvait apporter la Bct au cours de ces différentes étapes :

- pour décider des entrées de l'index : il s'avère que les critères de sélection et de structuration des termes de la Bct sont différents de ceux de l'indexation ;

- pour organiser la hiérarchie des entrées de l'index, à l'aide des relations conceptuelles dans la Bct ; le travail d'adaptation du réseau sémantique est considérable, mais le gain en qualité de cet index est aussi élevé ;

- pour décider des occurrences associées aux entrées, la Bct offre une sélection de textes par terme qui permet de gagner du temps par rapport à la lecture du texte.

L'exploitation du réseau conceptuel de la Bct contribue davantage au niveau méthodologique et théorique qu'au niveau opérationnel pour le moment. Ainsi, l'organisation générale de l'index a émergé précisément en étudiant une à une les différences entre l'indexation et l'analyse terminologique. Par exemple, la gestion des points de vue multiples, non systématique dans la Bct (les points de vue associés

[7] Cette expérience est rapportée en détail dans [AUSSENAC-GILLES 99a].

aux relations ne sont pas mis en évidence clairement), est très restrictive dans l'index : un point de vue est privilégié, les autres étant traduits par des liens au sein de la hiérarchie des entrées. Cette remarque suggère une meilleure définition des points de vue dans la Bct. D'autres exemples, comme la gestion des relations, montrent au contraire que les résultats de la Bct ont contribué à décider de choix liés à l'index.

Le modèle des données de la Bct semble adapté et pertinent : à part la gestion des synonymes et de leurs occurrences, la plupart des choix de représentation des connaissances dans le modèle de données facilitent l'exportation des données pour construire un index. La richesse du modèle permet de structurer cet index à la fois syntaxiquement et sémantiquement, en repérant les informations les plus importantes.

C'est *la nature des données présentes dans le modèle qui est en partie remise en question* (non pas leur qualité linguistique, mais leur adéquation au besoin). L'impression qui se dégage du travail réalisé est une perte d'énergie, liée aux redondances inutiles constatées dans la démarche. De plus, les linguistes mettent l'accent sur la rigueur de leur analyse en amont, alors que les données de la Bct sont jugées en aval via leur cohérence et la pertinence de la modélisation.

En fait, *c'est l'idée même d'une Bct-Corpus qui trouve ici ses limites*. Le réseau conceptuel étant réalisé en fonction du contenu du texte et non en vue d'une application précise, il est moins pertinent. Son adaptation requiert de nombreux retours au texte et des validations par l'expert et les utilisateurs. Ce travail recoupe celui déjà effectué pour la Bct-Corpus. Après avoir conduit une analyse linguistique 'neutre', il semble indispensable de prendre en compte la finalité de l'application pour rendre utilisables les données. L'équipe qui a développé l'index souhaite que cette prise en compte ait lieu le plus tôt possible, au cours de la mise au point de la Bct. La démarche suivie pourrait être très proche de celle adoptée, avec en plus le souci de choisir des occurrences pertinentes pour l'indexation, de retenir ou organiser les relations hiérarchiques afin de faciliter la lisibilité de l'index et de le simplifier.

4.2. Une application formelle

A titre d'expérience, nous avons représenté une partie de la Bct établie à partir du corpus Mouglis à l'aide d'une logique de description, en appliquant les principes requis par ce formalisme [SIMON 98]. Les conclusions de l'étude montrent que l'étape de formalisation rend plus systématique la description des connaissances, permet de mettre en évidence des anomalies de modélisation et oblige à recueillir

des connaissances supplémentaires, pas toujours disponibles dans les textes d'origine, pour répondre aux besoins de la différenciation des concepts d'une part, et de leur utilisation dans un certain raisonnement d'autre part. Cette étude justifie la séparation entre Bct et modèle conceptuel, chaque structure de données répondant à des contraintes différentes que ce soit pour la définition ou l'organisation des concepts ou pour la sémantique des relations ou encore les critères de modélisation.

4.2.1. Présentation de l'exemple : un sous-ensemble de la Bct Mouglis

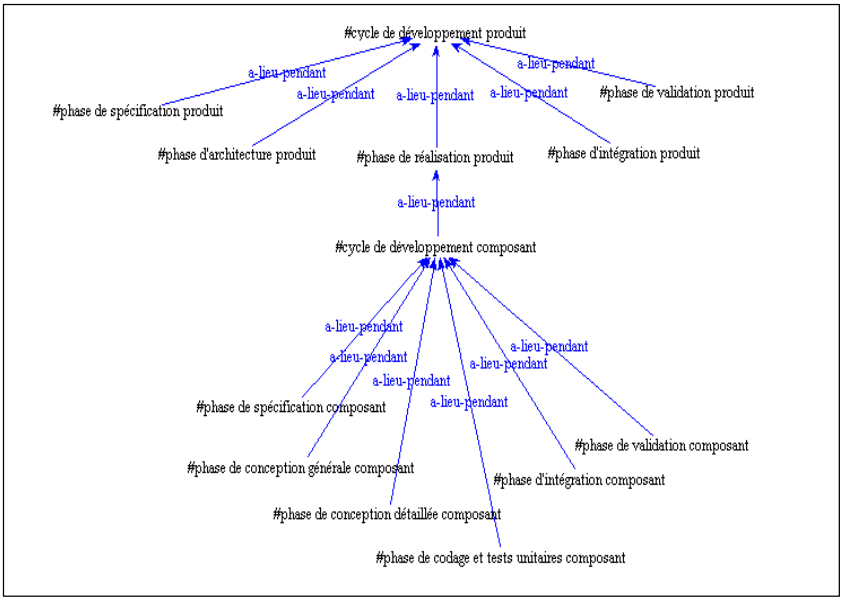

Figure 1 : **Extraits du réseau conceptuel de la Bct Mouglis** pour la relation « a-lieu-pendant » autour du concept #cycle de développement produit.

Le sous-ensemble du réseau conceptuel de la Bct étudié est relatif au cycle de développement d'un produit logiciel et à son découpage en phases et activités (Figure 1). Nous nous sommes aussi aidés de la description de la tâche (développer un produit logiciel scientifique et technique) dans le document. Les connaissances sont formalisées en vue de décrire *« le cycle de développement d'un produit logiciel »*, ce qui marque le *point de vue* retenu. La Bct, elle, a été faite sans intention particulière ni prévision d'utilisation, simplement avec l'objectif de couvrir l'ensemble du corpus. Nous avons ainsi évalué si, sur un point précis, la Bct répondait à nos attentes.

4.2.2. Le langage de formalisation retenu

Le langage utilisé pour structurer formellement les connaissances est basé sur les primitives, opérateurs et règles des logiques de descriptions (LD). Le modèle obtenu peut donc être opérationnalisé directement à l'aide d'une logique de ce type.

- Les primitives de base, concept et rôle, permettent de désigner des types d'entités et de relations (ou rôles) du domaine, qui sont déclarés à partir de leur position dans une hiérarchie (pour les concepts et rôles primitifs) et/ou de leurs rôles associés (conditions nécessaires et suffisantes pour les concepts et rôles définis). Les opérateurs de base permettent aussi de restreindre la cardinalité des rôles et d'ajouter des propriétés non définitoires aux concepts.

- Des informations terminologiques et conceptuelles sont associées aux concepts formels en indiquant les structures de la Bct à l'origine de leur définition.

4.2.3. Le processus de normalisation

La représentation en LD d'un modèle conceptuel se fait de façon descendante : tout concept utilisé doit d'abord avoir été défini. Il est impossible de modifier la définition d'un concept, mais sa description peut être complétée à l'aide de rôles non définitoires au fur et à mesure de l'exploitation de la Bct. Donc, avant de commencer la formalisation, il faut décider précisément de l'ensemble des entités à représenter et de leur définition. Nous avons suivi les étapes suivantes pour mener la formalisation :

(a) Identification des primitives conceptuelles à partir de la Bct ;

(b) Classification des primitives en concepts et rôles au sein de hiérarchies ;

(c) Expression de concepts définis à partir des propriétés nécessaires et suffisantes qui les caractérisent et selon une sémantique compositionnelle ;

(d) Ajout de connaissances non définitoires, sous forme de règles attachées aux concepts et transmises par héritage aux concepts subsumés ;

(e) Implémentation dans le formalisme choisi ;

(f) Ajout de Traces des choix de modélisation à l'aide de commentaires justifiant les éléments de la Bct à l'origine de la modélisation, la dénomination des concepts ou la représentation d'une primitive conceptuelle par un concept ou un rôle.

4.2.4. Mise en œuvre sur un exemple

Ainsi, les concepts primitifs identifiés à partir du sous-ensemble de la Bct Mouglis montré en figure 1 sont les notions structurantes (et absentes de la Bct) d'ENTITE-TERMPORELLE et d'ENTITE-PHYSIQUE. Parmi les ENTITES-TEMPORELLES, on a différentié les concepts d'ACTIVITE, de CYCLE et de PHASE, qui sont eux présents dans la Bct. ENTITE-LOGICIELLE et DOCUMENT sont deux

sous classes d'ENTITE-PHYSIQUE. Comme types d'ENTITE-LOGICIELLE, on distingue aussi le PRODUIT (global) de ses COMPOSANTS.

Plusieurs rôles primitifs ont été déclarés, dont le rôle « objet-de-l'activité » pour préciser l'objet auquel se rapporte une activité. A partir des concepts et rôles primitifs, plusieurs concepts et rôles ont été définis par composition, comme le concept PHASE-DE-SPECIFICATION-PRODUIT (tiré du concept de la Bct *#phase de spécification produit*) qui fait intervenir les concepts primitifs de PHASE et de PRODUIT.

4.2.5. Choix en cours de modélisation : exemple de la modélisation des activités

La formalisation à partir des données lexicales soulève deux types de problèmes que nous allons illustrer sur un exemple, et qui soulignent le caractère non automatique de cette étape : la gestion des écarts d'interprétation possible entre la Bct et le modèle formel ; le choix entre plusieurs représentations possibles.

Ainsi, dans la Bct, beaucoup de concepts de type #*activité* ont été recensés et organisés en taxinomie. Leur organisation ne fait pas bien apparaître que les types d'activités sont différenciés par la nature des objets sur lesquels elles portent. Par exemple, *#activité de rédaction* porte sur un document, alors que *#activité de préparation* a pour objet une phase. Pour rendre compte de cette connaissance utile au raisonnement dans le modèle formel, nous avions le choix entre deux solutions :

- *différencier par les catégories conceptuelles* : on définit autant de concepts que de types d'activités de base (REDACTION, GESTION, CODAGE, … qui sont des concepts fils de ACTIVITE) ; on définit aussi le rôle *a-pour-activité* qui relie une ENTITE-TEMPORELLE à une ACTIVITE.
- *différencier par les rôles* : on définit autant de rôles que de types d'activités particuliers, comme *rédaction* entre une PHASE et un DOCUMENT ou *codage* entre une PHASE et un LOGICIEL.

La première solution rend explicites les différents concepts mais elle multiplie les structures de données. L'avantage de la seconde solution est de permettre d'associer plusieurs rôles à une même phase si besoin, ou encore de spécialiser ces rôles (par exemple, préciser que la rédaction d'un document est achevée, initialisée ou poursuivie au cours d'une phase). Les taxinomies obtenues dans les deux cas sont très différentes.

4.2.6. Apports de la formalisation par rapport à la normalisation dans la Bct

La formalisation permet de réaliser des opérations liées au langage formel et de vérifier ou interroger les connaissances modélisées, qui constituent un réel avantage pour mieux organiser le modèle en fonction d'une application particulière :

- assurer la cohérence du modèle (respect des contraintes de typage, des cardinalités, domaines et co-domaines des rôles, etc.) ;
- interroger cette base sur l'état des concepts et de leurs propriétés (ex : Pendant quelle(s) phase(s) se déroule l'activité de DEFINITION-ARCHITECTURE-PRODUIT ?) ;
- vérifier la bonne représentation des connaissances, comme la connexité de la taxinomie, l'expression plus précise de la sémantique des relations, le repérage et l'élimination de données redondantes, pour obtenir un modèle plus homogène et systématique ;
- guider le recueil de connaissances supplémentaires, de manière à compléter les taxinomies, mieux différencier des concepts, etc.

Le résultat obtenu est un arbre de concepts homogène et systématique (au moins deux fils par nœud de l'arbre), cohérent (respect des contraintes de typage) et le plus complet possible (un concept défini est différencié de ses frères et de son père). Ainsi la masse des connaissances est mieux organisée que dans la Bct (les entités de base -primitives- sont différenciées des entités plus complexes -définies- sur lesquelles on va raisonner), et tient compte de l'utilisation qui sera faite des données (donner des conseils sur l'activité de production de logiciel). Par contre, le passage à la formalisation est prématuré au sein de la Bct puisqu'il requiert le recueil de données supplémentaires (pour différencier) et impose de faire des choix relatifs aux rôles des concepts dans le raisonnement.

5. Conclusion

Cet article nous a permis de montrer l'intérêt des Bct, modèle de données construit à partir de textes en préalable à différents types d'applications formalisées ou non. Nous avons situé les travaux actuels qui, suivant les perspectives, centrent leur intérêt plutôt sur le lien avec le texte ou plutôt sur la formalisation : les outils de gestion de Bct sont construits en privilégiant un ou l'autre point de vue. Le degré de prise en compte de l'application permet aussi de distinguer ces outils et les méthodes de constitution. Lorsqu'ils sont orientés sur la modélisation du texte, méthodes et outils sont, en principe, indépendants d'une application ; ils sont au contraire très dépendants de l'application lorsqu'ils sont proches de la formalisation. Les expériences que nous avons menées, dont les résultats sont présentés dans cet article, se sont faites en distinguant deux étapes, l'une concernant la stricte modélisation du texte (et qui consiste pour nous en la construction d'une Bct), l'autre concernant la formalisation, et nous ont permis d'évaluer la pertinence des Bct. Ces expériences ont mis en évidence le besoin de cadrer le plus possible une représentation, de

fournir avec elle une grille de lecture, les objectifs et les choix qui ont guidé sa construction, pour mieux maîtriser l'interprétation qui en sera faite.

Le passage à un modèle formel utile à une application donnée suppose plusieurs étapes pour lesquelles un environnement logiciel doit faciliter la traçabilité des choix effectués et des points de vue retenus. La constitution d'une plate-forme d'outils proposant toutes sortes de fonctionnalités, depuis l'étiquetage des corpus, jusqu'aux interfaces d'interrogation des Bct, devient un des défis importants pour les années à venir. Du point de vue méthodologique, deux pistes de travail devront être explorées. L'une concerne les corpus qui jouent un rôle fondamental dans la constitution des Bct. Or, leur constitution, de façon à ce qu'ils soient équilibrés, représentatifs d'une application et en même temps riches en « contextes définitoires » n'a pas encore été suffisamment travaillée. L'autre piste de réflexion consistera à approfondir la prise en compte de l'application dans l'adaptation des méthodes aux besoins. On perçoit intuitivement que construire un index ne demande pas le même type d'approche que capitaliser l'ensemble des connaissances d'une documentation technique. Il faudra travailler cette intuition afin d'adapter les outils et les interprétations, en un mot de rendre plus efficace le travail d'analyse de corpus. Ainsi, derrière ce terme de « Base de connaissances terminologiques », on voit poindre tout un courant de recherches qui prend les textes comme point de départ et construit des modèles, de natures très diverses. L'interdisciplinarité est au cœur de ces approches tout comme la relation entre recherche théorique et recherche appliquée ; on peut donc espérer des résultats très prometteurs dans un proche avenir.

Bibliographie

[AHMAD 95] AHMAD K. & HOLMES-HIGGIN P.R., System Quirk : a unified approach to text and terminology. In *terminology in advanced Microcomputer Applications. Proc. Of the 3rd TermNet Symposium : recent advances and User Reports.* TermNet : Vienna (Austria) pp 181-194. (ISBN 3-901010-12-2).

[ASSADI 98] ASSADI H. « Construction d'ontologies à partir de textes techniques : Application aux systèmes documentaires ». Thèse de l'université Paris 6. Oct. (1998).

[AUSSENAC-GILLES 98] AUSSENAC-GILLES N., CONDAMINES A., « Terminologie, Modélisation des Connaissances et Systèmes Hypertextuels de Consultation de Document Technique ». Rapport Interne IRIT/98-20-R. Toulouse : IRIT, Univ. P. Sabatier.106 p. (1998)

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

[AUSSENAC-GILLES 99b] AUSSENAC-GILLES N., « GEDITERM, un logiciel de gestion de bases de connaissances terminologiques », in ENGUEHARD C. et CONDAMINES A. (Eds.) : actes des 3es Rencontres "Terminologie et intelligence artificielle" (Nantes, 10 et 11 mai 1999), dans Terminologies Nouvelles. (19). Déc. 1998 et juin 1999. Bruxelles : Agence de la Francophonie et de la Communauté Française, pp. 111-123.

[BACHIMONT 00] BACHIMONT B., « Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances », J. CHARLET, M. ZACKLAD, G. KASSEL & D. BOURIGAULT, (eds). *: Ingénierie des Connaissances, évolutions récentes et nouveaux défis*. Paris : Eyrolles, 2000.

[BIEBOW 97] 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.* Ed. ERSS. Toulouse (F) : Université Toulouse-Le Mirail. pp. 69-84. (1997).

[BIEBOW 00] BIEBOW B. & SZULMAN S., « Terminae : une approche terminologique pour la construction d'ontologies du domaine à partir de textes ». Actes de RFIA2000, Reconnaissances des Formes et Intelligence Artificielle, Paris (F), février 2000.

[BORILLO 97] BORILLO A. : « Exploration automatisée de textes de spécialité : repérage et identification automatique de la relation lexicale d'hyperonymie ». *Linx,* n°34-35, pp.113-121.

[BOURIGAULT 96] BOURIGAULT D. : « Lexter, a Natural Language Processing Tool for Terminology Extraction ». *Proceedings of Euralex'96*, Göteborg University, Department of Swedish, 1996, pp. 771-779.

[BOURIGAULT 00] BOURIGAULT D., JACQUEMIN C., « Construction de ressources terminologiques », In J.M. PIERREL (ed) : *Ingénierie des langues*, Traité I2C, Paris : Hermes. (2000).

[CAPPONI 95] CAPPONI N. Modélisation d'une base de connaissances terminologiques, DEA de l'Université de Nancy 1. CRIN/LORIA, Nancy. 1995.

[CHARLET 96] CHARLET J., BACHIMONT B., BOUAUD J., ZWEIGENBAUM P., « Ontologie et réutilisabilité : expérience et discussion », N. AUSSENAC-GILLES, P. LAUBLET, C. REYNAUD (eds) : Acquisition et Ingénierie des Connaissances, Toulouse : Cépaduès-Editions, 1996.

[CHARLET 00] CHARLET J., REYNAUD C., « Ingénierie des connaissances pour les systèmes d'information »in *Conception des Systèmes d'Information*, C. Cauvet (ed), Traité IC2, Paris : Hermès. 2000.

[CONDAMINES 00a] CONDAMINES A., REBEYROLLE J. « Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode » J. CHARLET, M. ZACKLAD, G. KASSEL & D. BOURIGAULT, (eds). *: Ingénierie des Connaissances, évolutions récentes et nouveaux défis*. Paris : Eyrolles, 2000.

[CONDAMINES 00b] CONDAMINES A., REBEYROLLE J., « Searching for and Identifying Conceptual Relasionships via a corpus-based approach to a Terminological Knowledge Base (CTKB) : method and results ». D.BOURIGAULT, M.C. L'HOMME, C.JACQUEMIN (eds) : *Recent Advances in Computational Terminology*, John Benjamins, 2000.

[ENGUEHARD 95] ENGUEHARD C., PANTÉRA L., « Automatic natural acquisition of terminology » *Journal of Quantitative Linguistics*, vol.2, n°1, pp. 27-32, 1995.

[EUZENAT, 96] EUZENAT J., « Hytropes : a www front-end to an object knowledge managment System » Knowledge Acquisition Workshop, KAW'96, Fiche démonstration, Banff, Canada, 1996.

[GARCIA 98] GARCIA D., Analyse automatique des textes pour l'organisation causale des actions, Réalisation du système Coatis, Thèse d'informatique, Université Paris IV, 1998.

[GOMEZ-PEREZ 99]. GOMEZ-PEREZ A. "Développements récents en matières de conception, de maintenance et d'utilisation des ontologies". in ENGUEHARD C. et CONDAMINES A. (Eds.) : actes des 3es Rencontres "Terminologie et intelligence artificielle" (Nantes, 10 et 11 mai 1999), dans Terminologies Nouvelles. (19). Déc. 1998 et juin 1999. Bruxelles : Agence de la Francophonie et de la Communauté Française. pp. 9-20.

[GROS 99] GROS C., ASSADI H. « Les systèmes de consultation de documentation technique ». J.Maniez et W.Mustapha El Hadi (eds), *Organisation des connaissances en vue de leur intégration dans les systèmes de représentation et de recherche d'information*, Villeneuve d'Asq : Université Charles de Gaulle, (UL3 travaux et recherches), 1999.

[JOUIS 93] JOUIS C., Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse de textes. Réalisation d'un prototype : le système Seek. Thèse en informatique, EHESS, Paris, 1993.

[LEMAIRE 95] LEMAIRE F., RECHENMANN F., « Intégration de connaissances terminologiques dans les grandes bases d'objets, exemples en biologie moléculaire », Actes de TIA'95, *La Banque des mots* n°spécial 7, 1995, pp.103-112.

[LEROUX 96] LEROUX B., « Eléments d'une classification des approches ontologiques ». *Actes des Journées d'Acquisition des Connaissances JAC'96*, Sète, mai 1996. pp 109-122. 1996.

[MARTIN 95] 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).

[MEYER 00] MEYER I., « Extracting Knowledge-rich Contexts for Terminography : A Conceptual and methodological Framework ». D.BOURIGAULT, M.C. L'HOMME, C.JACQUEMIN (eds) : *Recent Advances in Computational Terminology*, John Benjamins, 2000.

[MORIN 99] MORIN E., « Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique » , *TAL* (Traitement Automatique des Langues), vol.40, n°1, Paris : Université Paris VII, pp.143-166, 1999.

[REYNAUD 94] REYNAUD C., TORT F., « Connaissances du domaine d'un SBC et ontologies : discussion » *Actes des Journées d'Acquisition des Connaissances JAC'94*, Strasbourg, mars 1994. pp B1-B13, 1994.

[ROUSSELOT 96] ROUSSELOT F., FRATH P., OUESLATI R. 1996 : « Extracting Concepts and relations from corpora ». *Proceedings ECAI'96, 12th European Conference on Artificial Intelligence,* 1996.

[SEGUELA 99] SEGUELA P., AUSSENAC-GILLES N., « Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine », *Actes de IC'99 (Ingénierie des Connaissances)*, pp 79-88, Paris, 1999.

[SIMON 98]. SIMON S. « Représentation formelle des connaissances issues d'une Base de Connaissances Terminologiques ». Mémoire de DEA Représentation des Connaissances et Formalisation du Raisonnement. Univ. P. Sabatier, Toulouse (F). Sept. 1998.

[SKUCE 94] SKUCE D., LETHBRIDGE T.C., « CODE4 : A multifunctional Knowledge Managment System » Proceedings of the 8th Knowledge Acquisition Workshop, KAW'94, Banff, Canada, 1994.

R

S

T