Bases de connaissances Terminologiques: enjeux pour la consultation documentaire

Nathalie Aussenac-Gilles, Anne Condamines

Bases de connaissances terminologiques : enjeux pour la consultation documentaire

Nathalie Aussenac-Gilles*, Anne Condamines**

** Institut de Recherches en Informatique de Toulouse (IRIT), Toulouse (F), aussenac@irit.fr*

*** Equipe de Recherche en Syntaxe et Sémantique (ERSS), Toulouse (F), acondami@cict.fr*

0 Introduction

L'accès aux connaissances contenues dans les textes est un problème traité par de nombreuses disciplines, aussi bien en sciences humaines — linguistique, ergonomie, psychologie ... — qu'en sciences dites dures — mathématiques, informatique, intelligence artificielle (IA) ... —. L'interaction entre ces disciplines est faible voire inexistante. Dans le cadre d'un projet[1] mettant en oeuvre la terminologie et, en informatique, la modélisation des connaissances et la gestion documentaire, nous avons pu faire l'expérience d'une interdisciplinarité et mesurer ainsi les complémentarités possibles entre ces trois axes pour améliorer l'exploitation des connaissances documentaires. En particulier, nous pouvons à présent commencer à évaluer les difficultés et le coût de la construction d'une Base de Connaissances Terminologiques (BCT) selon la démarche que nous préconisons. Nous expérimentons ainsi les possibilités d'utiliser une BCT pour construire un système d'accès à des données textuelles, et ce en dimension réelle, dans un domaine fourni par notre partenaire industriel, la DER d'EDF-GDF. Ce système[2] s'appuie sur une structuration hypertexte du document et sur la tâche de l'utilisateur pour proposer une indexation structurée. Il exploite aussi une terminologie extraite du document pour guider la recherche par mots-clés dans l'ensemble du texte.

Nous rendons compte ici des résultats de cette expérience et de nos réflexions sur l'utilisation de la BCT que nous avons construite dans le cadre de ce projet. L'article est organisé en quatre parties. La première insiste sur la nécessité d'une collaboration interdisciplinaire pour conduire ce type de recherche. La deuxième décrit la méthode qui a été utilisée pour construire la BCT et le rôle du texte dans cette démarche de construction. La troisième vise à présenter le modèle des données mis au point dans le cadre du projet. Dans la quatrième, nous montrons comment les données de la BCT ont été utilisées pour construire un modèle de tâche et comment on pourrait envisager de les utiliser pour construire un index, voire un thésaurus.

[1] Ce projet de 2 ans est financé en partie par la DER d'EDF-GDF, par le Groupement d'Intérêt Scientifique « Sciences de la Cognition » et par la région Midi-Pyrénées.

[2] Les fonctionnalités de ce système sont indentiques à celles du logiciel Hyperplan, décrit dans (Gros,97).

1 Le contexte : une collaboration interdisciplinaire

1. 1 L'apport de la linguistique

L'ERSS est une équipe de linguistes dont l'un des axes de recherche est l'analyse des corpus spécialisés, en particulier pour étudier leurs caractéristiques par rapport à des corpus généraux, essentiellement du point de vue du lexique, c'est-à-dire de la terminologie. Appliquer ces recherches conduit à des collaborations avec des entreprises (Matra Marconi Space, CNES, EDF, ...) afin de construire des BCT. Ces BCT, rendent compte à la fois de données linguistiques et de données conceptuelles et se veulent le reflet modélisé du corpus étudié : les données implicites sont rendues explicites, les synonymes sont précisés, les sigles explicités, etc. Le texte étudié est, en principe, la référence ultime pour le terminologue-linguiste, ce qui signifie par exemple que les variations de dénomination voire les incohérences sont notées mais non corrigées. De même, la BCT ne garantit pas l'exhaustivité des données sur un domaine.

En revanche, la BCT est une construction solide (partie 2) qui n'est pas élaborée sur des critères intuitifs (comme le sont encore la plupart des bases de données terminologiques) mais grâce à une méthode faisant appel à des critères linguistiques, appliqués à des résultats fournis par des outils d'analyse de texte, critères qui s'affinent au fur et à mesure des expérimentations. Ainsi, la BCT peut être considérée comme une ressource fiable, utilisée soit telle quelle (en consultation directe), soit enrichie, complétée, voire adaptée selon les applications. Dans la mesure où le corpus joue un rôle très important dans la constitution de la BCT, il est évident qu'en retour, la BCT peut être utilisée dans le cadre de la consultation documentaire (partie 3).

1.2 L'apport de l'intelligence artificielle

L'apport de l'IA, assuré par une équipe de l'IRIT, est double : il concerne le choix de la représentation des connaissances dans un outil général de gestion de BCT, et l'utilisation de modèles conceptuels définis en IA pour guider la consultation documentaire. Cette deuxième contribution a été évaluée lors d'un précédent projet, Hyperplan (Gros,96) (Gros,97). A l'origine, les modèles conceptuels, issus du génie logiciel, ont servi à analyser et concevoir des systèmes de résolution de problème. Ils offrent un cadre adapté à la description informelle de connaissances, suffisamment structuré pour en permettre des exploitations simples.

Nous nous appuyons sur des modèles conceptuels adaptés de ceux de la méthode d'acquisition des connaissances MACAO que nous avons développée (Matta,96). Une partie du modèle décrit les objets du domaine, leurs propriétés et leurs relations, ainsi que les règles ou contraintes du domaine utilisées pour la résolution de problèmes : c'est le modèle du domaine. Assimilé à un réseau conceptuel, il comporte des connaissances dites descriptives ou statiques. Une deuxième partie du modèle décrit les buts à atteindre pour résoudre ces problèmes et les méthodes de résolution applicables pour réaliser ces buts. Ces méthodes utilisent les connaissances du domaine selon le contrôle qu'elles définissent. Buts et méthodes constituent un modèle de raisonnement, dit dynamique, ou modèle de tâche.

Dans l'objectif de construire un système de résolution de problème, le modèle conceptuel constitue un support au recueil et à l'organisation des connaissances, ainsi qu'un ensemble de spécifications du futur système. En revanche, dans le cadre de la consultation documentaire, cette notion doit être adaptée afin d'assurer des liens vers les textes. Le modèle doit refléter la tâche effectuée par la personne recherchant des informations dans le document.

2 Base de Connaissances Terminologiques : démarche de conception

Afin de garantir sa fiabilité et ainsi son utilisabilité, la BCT doit être construite avec une méthode précise qui laisse le moins de place possible aux choix basés sur la seule intuition. Cet objectif de transparence garantit aussi que cette méthode puisse être vérifiée, reproduite et enseignée. Enfin, en explicitant les critères qui président au choix des données de la BCT, non seulement on enrichit la connaissance théorique qu'on peut avoir de ces données, mais on définit aussi les tâches effectuées par des outils, en totalité ou partiellement, et celles ne pouvant être réalisées que manuellement. Définir la méthode de construction de BCT suppose donc aussi que soient recensés, voire spécifiés, les outils accélérant cette démarche sans altérer la fiabilité des résultats (Condamines,97). Sa mise en place nécessite également d'identifier les connaissances linguistiques pertinentes pour définir des critères de choix des données à conserver. Enfin, il s'agit d'évaluer les possibilités d'intégration des résultats fournis par les outils dans la réflexion linguistique ainsi que le sens à leur donner dans le cadre de la constitution de la BCT.

2.1 Les outils

Les outils qui peuvent être utilisés dans le cadre d'une analyse de corpus pour construire une BCT viennent d'horizons très divers (linguistique informatique, IA, analyse statistique ...). Les outils terminologiques proposent des candidats termes (LEXTER (Bourigault,94), NOMINO (Plante,95)) ou des relations conceptuelles candidates (SEEK (Jouis,94), IKARUS (Kavanagh, 96)). Suivant les cas, ils mettent en oeuvre soit une méthode descendante, qui projette sur les textes des connaissances sur le fonctionnement des phénomènes langagiers, soit une méthode ascendante, qui fait émerger les régularités du texte sans connaître a priori le fonctionnement des données terminologiques.

Prenons l'exemple des termes candidats. NOMINO repère les suites de mots qui obéissent à des patrons syntaxiques dont on présume qu'ils peuvent correspondre à des termes, comme « N de N adj » (*atelier de génie logiciel*) ou « N de N de N adj » (*archivage de l'état de configuration logiciel*). Au contraire, LEXTER recherche les candidats termes "en creux", sans préjuger de la forme syntaxique qu'ils peuvent avoir. Il utilise la notion de barrières entre lesquelles peuvent apparaître des termes (exemple : après un verbe et avant une préposition autre que "de" ou "à").

Les outils d'analyse de corpus, comme SATO (Daoust,92) utilisé dans ce projet, servent à explorer les contextes en détail. Ainsi, on peut valider ou invalider, sur la base de connaissances linguistiques, les résultats proposés par les outils terminologiques. Plus ou moins élaborés, ces systèmes permettent de retrouver une chaîne de caractères donnée littéralement ou par ses caractéristiques (*i.e.* identifiée comme nom ou verbe). Plus rarement, ils s'appuient sur une analyse syntaxique complète pour repérer les termes par leur fonction syntaxique dans la phrase (*i.e.* complément d'objet direct). D'une utilisation très souple, ces outils autorisent une variété importante d'interrogations sans avoir à décider à l'avance le type de résultats attendus.

2.2 Les connaissances linguistiques

La mise en oeuvre de connaissances linguistiques conduit à la fois à justifier les choix qui sont faits et à relier la démarche du terminologue à celle du linguiste. Ainsi, le terminologue bénéficie du développement des connaissances en linguistique.

Pour sélectionner les termes hors contexte, il définit des critères de sélection de termes à partir des listes de candidats-termes proposés. 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 générale* vs *conception détaillée* ou des synonymes comme *petit projet* vs *projet de petite taille).*

La sélection en contexte permet de travailler sur les relations conceptuelles dans le corpus à l'étude. Il s'agit de construire un modèle de ces relations, indépendamment de leur manifestation linguistique. Ainsi, la même relation "a-pour-partie" peut s'exprimer différemment :

X a pour parties (éléments, composants, pièces ...) Y et Z

(la, le) (décomposition, découpage,...) de X en Y et Z

les deux éléments (morceaux, constituants...) Y et Z de X.

Il faut retrouver dans le corpus toutes ces manifestations linguistiques et les identifier comme renvoyant bien à une seule relation conceptuelle. Cette approche suppose que l'on parte d'une batterie de relations les plus générales (et donc les plus fréquentes dans le corpus), actuellement une quinzaine, et de séries de marqueurs linguistiques correspondant à chacune de ces relations. Au fur et à mesure du déroulement du travail en corpus, les marqueurs s'affinent parfois jusqu'à être propres à un corpus donné et de nouvelles relations apparaissent.

Ce travail permet aussi de décider quels termes vont être finalement conservés. En effet, seuls seront conservés les mots ou groupes de mots qui dénomment des concepts, c'est-à-dire ceux qui sont mis en relation, dans le corpus, par une relation conceptuelle. Ainsi, dans *la phase d'architecture est une phase du cycle de vie,* on peut retenir les termes *architecture* et *cycle de vie* car le marqueur *est une phase de* signale la présence d'une relation de partie temporelle.

2.3 Analyse des résultats

Les résultats sont analysés en trois temps, puis stockés dans un logiciel de gestion de BCT :

- Il faut tout d'abord sélectionner, parmi les données obtenues, celles réellement pertinentes pour la BCT : s'assurer de la validité des marqueurs, de la pertinence des relations les unes par rapport aux autres, éliminer les redondances ...
- Il faut ensuite organiser les données selon le modèle : répartir informations linguistiques et conceptuelles et déterminer celles à mentionner dans les liens terme/concept.
- Il faut enfin faire valider les données par des experts du domaine, qui confirment ou non la bonne interprétation des données textuelles. En principe, ils ne doivent pas trop corriger ces données car alors, le texte ne servirait plus de référence.

Ces trois étapes, décrites à plat, sont, en réalité, effectuées en boucles. Ainsi, la validation par l'expert a lieu à différents états d'avancement de la BCT : pour valider les équivalents (synonymes), puis les relations conceptuelles ... A la fin de l'analyse du corpus et de la saisie des données, on obtient un modèle du texte à travers la modélisation terminologique. Une telle BCT peut être utilisée pour différentes applications ; un logiciel d'accès au contenu du texte qui a servi de référence pour la constituer peut être, à l'évidence, une de ces applications.

3 Base de Connaissances Terminologiques : modèle des données

3.1 Présentation générale

Les BCT constituent un enrichissement significatif des terminologies papier traditionnelles car elles comportent une trace des informations conceptuelles relevées par le terminologue au moment de l'identification des termes (Meyer,92). Le modèle de structuration des connaissances terminologiques des BCT différencie un niveau linguistique d'un niveau conceptuel : on accède ainsi par les termes du domaine à une modélisation conceptuelle (Bourigault,95). Cette composante s'apparente à la partie descriptive des bases de connaissances, ou encore plus étroitement aux modèles du domaine et aux ontologies en IA (Bachimont,96).

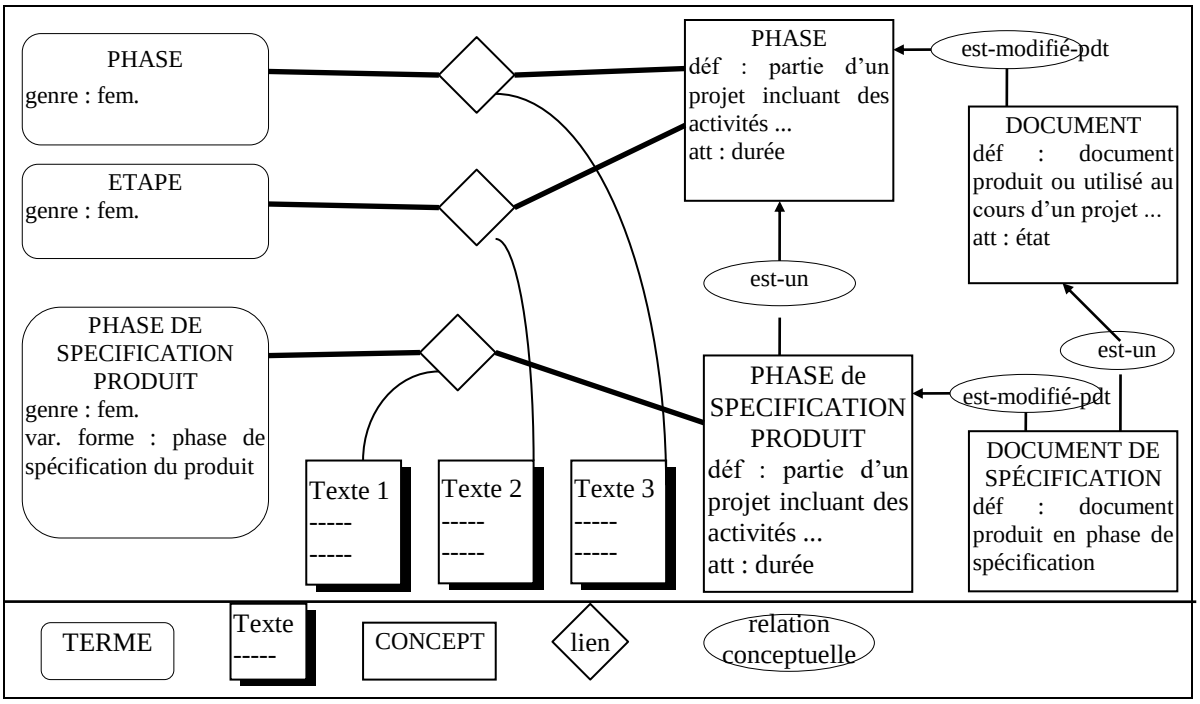

Figure 1 : Organisation des données dans la BCT. Le réseau conceptuel, à droite, est relié aux données terminologiques, à gauche, par des liens qui renvoient à des parties du texte.

Pratiquement, les données d'une BCT sont donc organisées en deux parties : le réseau conceptuel et les données linguistiques, ces dernières étant reliées au corpus d'où elles sont extraites (Séguéla,97). Le modèle de BCT que nous proposons comporte trois types de structures : termes, concepts et textes (Fig. 1). Termes et concepts sont distingués afin de dissocier la manifestation linguistique de la notion qu'elle dénomme. On n'associe donc 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. D'ailleurs, ils sont reliés par des liens illustrant l'usage du terme lorsqu'il désigne ce concept.

3.2 Le corpus

Le choix du corpus soulève un problème fondamental : il conditionne le domaine couvert par la BCT (en général, ne sont retenus dans la BCT que les termes présents dans le corpus), mais,

inversement, le but de l'étude terminologique dans le domaine va déterminer le choix des textes qui le constituent. Dans notre modèle, le corpus est intégré à la BCT, afin de pouvoir y faire directement référence. De manière transparente pour les utilisateurs, il est découpé en unités (la phrase a priori). Ainsi ce sont à ces unités que renvoient les liens terme/concept.

3.3 Les termes

Dans notre modèle de BCT, le terme correspond au signifiant. La structure de terme rassemble donc, en plus du syntagme qui le désigne, uniquement des informations linguistiques : langue, variantes de forme, catégorie grammaticale, genre, ... Les relations entre termes, comme la synonymie ou la polysémie, sont implicites et calculables à partir des liens terme-concept. Pour savoir si deux termes sont proches, assimilables ou différents, on fait appel au concept associé.

3.4 Les concepts et relations conceptuelles

La modélisation conceptuelle de la BCT vise la caractérisation des notions et non leur définition exhaustive. Un concept est repéré par un identifiant, dans notre cas adapté d'un des termes associés qui peuvent le dénommer. Il véhicule un ensemble d'informations sémantiques, sa description, qui fait apparaître des caractéristiques de la notion correspondante, telles que le terminologue les a repérées dans les textes. Cette description doit être suffisamment explicite, puisque son interprétation indique la définition des termes associés (Bachimont,96). En fait, la description des concepts soulève trois problèmes, pour lesquels nous avons retenu des solutions qui privilégient le point de vue du linguiste (choix justifiés dans (Séguéla,97)) :

- Le choix des critères de différenciation des concepts : nous avons décidé qu'un nouveau concept soit créé chaque fois que l'analyse des textes le suggère. Dans sa représentation, ce nouveau concept doit se distinguer d'un concept existant par au moins une relation conceptuelle ou un attribut. Par exemple, la phrase *Le dossier de spécification est construit pendant la phase de spécification* permet de repérer la relation conceptuelle *est-commencé-pendant* entre deux concepts, *phase de spécification* et *dossier de spécification*. Cette relation peut servir à différencier le concept *dossier de spécification* de son père dans la hiérarchie conceptuelle, *document*. Les valeurs d'attributs et relations d'un concept le caractérisent donc de manière unique. Les vérifications associées à la validation des concepts garantissent cette normalisation et ainsi la cohérence du réseau conceptuel.
- La gestion des différentes descriptions possibles d'un concept (ou points de vue) : la notion de point de vue a été retenue au sens linguistique, et s'applique à un terme qui prend plusieurs sens pour un même locuteur. Ainsi, le terme *cycle de vie* prend deux sens différents dans le document pour les mêmes locuteurs (dont le *rédacteur du document*). Il est donc relié à deux concepts (*#cycle en V* et *#cycle en spirale*). Sur le lien vers le concept *#cycle en V* est indiqué que ce point de vue est dominant pour le *rédacteur du document*. Si, dans le texte, on trouve plusieurs descriptions d'un même concept, on distinguera autant de concepts que de descriptions, sans utiliser nécessairement la notion de point de vue.
- L'interprétation de cette description, liée à son degré de formalisation. Nous avons donné priorité à l'interprétation humaine, par le linguiste qui construit la BCT mais aussi par d'autres personnes, en imposant une description seulement structurée et normalisée.

Un concept est donc représenté par un frame où sont explicitées sa place dans la hiérarchie des concepts (selon la relation « est-un »), les connaissances de différenciation (attributs et relations

conceptuelles, hérités ou non par les concepts fils), selon un format précis, mais non interprétable par la machine. D'après ce choix, les relations conceptuelles ont, en plus du rôle d'organisation du modèle, un rôle de différenciation linguistique des concepts reliés. Leurs types sont soit prédéfinis, soit propres au domaine. Ces relations n'ont pas une sémantique formelle, mais elles peuvent être interprétées sans erreur grâce à un texte de définition.

Le lien terme-concept comporte des informations qui caractérisent les conditions dans lesquelles ce terme désigne ce concept : des extraits du corpus illustrant des usages du terme ; les locuteurs pour qui ce lien est valide ; éventuellement, des marqueurs syntaxico-sémantiques servant à repérer dans les textes des relations conceptuelles associées au concept.

4 Des données de la BCT à un système de consultation documentaire

Dans le domaine applicatif étudié, le génie logiciel scientifique et technique, la documentation concernée est le guide MOUGLIS[3]. Ce classeur, d'environ 350 pages, contient plusieurs types de documents techniques : des recommandations méthodologiques, des plans types ou des guides de rédaction. Une BCT a donc été construite à partir de l'analyse terminologique du Guide, selon la démarche présentée en partie 2, et exploitée pour mettre au point un système de consultation en ligne du Guide, qui est pour le moment sur support papier.

4.1 Le système de consultation visé

Le système de consultation documentaire visé s'appuie sur une structure hypertexte pour offrir différents types d'accès au texte, choisis pour leur complémentarité : table des matières ; index de termes, hiérarchisé et précis ; index des activités ; recherche plein texte avec possibilité d'étendre les requêtes en parcourant une hiérarchie de concepts. Une de ses originalités est de proposer au lecteur des accès prédéfinis, choisis pour leur pertinence dans différents contextes (étapes de la tâche au cours de laquelle il utilise le document).

4.2 Utilisation de la BCT pour mettre au point le modèle de la tâche

4.2.1 Nature du modèle de la tâche

Dans un contexte de consultation documentaire, un modèle de tâche doit refléter la tâche de celui qui consultera la documentation. Ainsi, il sert à préciser la partie de la tâche que l'utilisateur est en train de réaliser lorsqu'il recherche des informations sur un certain concept. De même, le modèle du domaine est particulier : il est constitué des concepts utiles à la description de la tâche, mais aussi de références au texte. Pour chaque concept et dans chaque tâche où il est mentionné, ces références sont choisies de manière à présenter le passage du document expliquant le mieux comment ce concept est utilisé pour réaliser cette tâche, en quoi il consiste ou à quoi il sert. Le modèle conceptuel permet donc de définir avec les utilisateurs un ensemble de contextes de consultation, et de fixer des parties de texte pertinentes dans ces contextes.

Pour obtenir ce modèle, nous exploitons les données de la BCT, et surtout, nous suivons une démarche basée sur une approche ergonomique. L'idéal serait d'étudier l'activité réelle des utilisateurs. Or dans ce projet, nous nous sommes heurtés à deux difficultés : la tâche étudiée concerne l'ensemble d'un projet, soit plusieurs types d'intervenants ; et surtout, nous n'avons pu

[3] Méthodes et Outils pour le Génie Logiciel en Informatique Scientifique

procéder à des observations mais seulement à des entretiens auprès d'un expert, l'auteur des documents. En même temps, le modèle doit demeurer cohérent avec le contenu du document, également plus proche de la tâche prescrite que de l'activité. Nous avons donc constitué un modèle de la tâche prescrite, et ceci en trois temps : une première ébauche décrivait les tâches avec du texte ; puis un modèle plus structuré les a décrites à l'aide de concepts regroupés en 'champs' ; enfin, dans le modèle final, les concepts renvoient à des occurrences dans le texte.

4.2.2 Décomposition de la tâche sous forme d'un arbre de tâches

La BCT reflète la description de la tâche prescrite que présente le document. Par exemple, tous les concepts ayant pour étiquette 'phase de ...' correspondent aux phases d'un projet et donc aux tâches du modèle. Ils sont reliés entre eux par des relations : 'est-une-phase-de' indique qu'une phase entre dans la décomposition d'une autre et 'précède' indique la chronologie de leur déroulement. Enfin, les concepts de phases sont reliés à des activités plus élémentaires par la relation 'est-une-étape-de'. Ces informations ont été utilisées pour choisir une dénomination des tâches cohérente avec le texte et pour vérifier les résultats de l'analyse faite par le cogniticien du document et des entretiens avec l'expert. Elles ont permis de s'assurer de la décomposition des tâches, de leur ordre et de leur position dans l'arbre. Finalement, après validation par l'expert, le degré de décomposition de la tâche n'est pas le même dans la BCT et dans le modèle de tâche. Certaines *activités* ont pris le statut de *tâches* dans le modèle afin d'en faciliter la lecture.

4.2.3 Choix des champs décrivant les tâches

La description des tâches consiste à associer à chacune un ensemble de concepts pertinents pour sa réalisation. Pour en faciliter la lecture, ces concepts sont organisés selon un plan découpé en « champs », et ceci en fonction de leur « rôle » ou utilisation pour réaliser la tâche. A l'origine, nous avons repris les champs proposés dans le formalisme de MACAO : *entrées, sorties, connaissances associées et méthodes*. Or dans le document, les tâches sont décrites selon un autre format : *documents en entrée/finalisés/initialisés, activités principales, éléments logiciels en entrée/sortie*. Afin de garantir la cohérence entre document et modèle, et de rester au plus près de la terminologie familière aux lecteurs, nous avons retenu les mêmes champs que dans le document, ajoutant une *description générale* à chaque tâche.

Des traces de ces champs se retrouvent dans les étiquettes de certaines relations associées aux concepts de phases dans la BCT. Les relations 'conditionne-le-debut-de' (qui précise les résultats requis au début d'une phase) ou 'conditionne-la-fin-de' (qui précise quand une phase est jugée terminée) renvoient aux conditions de réalisation (début et fin) des phases. Elles correspondent aux champs *documents en entrée* ou *document finalisé*. De même, la relation 'est-initialisé-pendant', reliant un *document* et une *phase* correspond au champ *documents initialisés*.

4.2.4 Choix des concepts associés aux tâches et contribuant à leur description

Les concepts associés aux tâches ont été choisis tout d'abord à partir des mots utilisés dans la description en langage naturel de la tâche. Ensuite, cette description a été affinée grâce au réseau conceptuel de la BCT, et c'est là sans doute l'apport le plus significatif de la BCT au modèle. Pour cela, nous avons exploité non pas le nom des relations entre concepts de haut niveau, comme à l'étape précédente, mais les relations spécifiques d'un type donné entre des concepts plus spécifiques. Ainsi, l'ensemble des relations de type 'est-initialisé-pendant' entre les concepts

d'étiquette *phases-de-...* et des fils de *document* indique les concepts à associer au champ *initialisé pendant* des tâches correspondantes. D'autres relations présentes dans la BCT ont permis de rajouter à la description de tâches des concepts parce qu'ils étaient en relation avec des concepts déjà mentionnés dans ces tâches. De plus, la confrontation à la BCT a permis de garantir un choix d'étiquettes de concepts cohérent avec le vocabulaire utilisé dans le document.

4.2.5 Choix des références associées aux concepts

Il est sans doute surprenant a priori de noter que la BCT joue une rôle moindre dans la sélection des références associées aux concepts du modèle. Mais cela l'est moins si l'on rappelle les critères appliqués pour retenir une unité textuelle dans les deux cas : critères définitoires, relatifs à l'identification des relations conceptuelles et à la présence du terme associé, dans la BCT (cf. 3.1) ; critères contextuels, relatifs à l'utilisation du concept dans une tâche, dans le modèle (cf. 4.2.1). Les occurrences citées dans la BCT peuvent donc être reprises dans le cas où l'on veut renvoyer à une définition du concept. Mais dans les autres cas, le choix d'une référence suppose un retour vers le texte, via un outil comme SATO, pour rechercher les plus pertinentes parmi les occurrences du terme, ou parmi les paragraphes les englobant.

4.3 Des données de la BCT à une indexation

4.3.1 Index ou thésaurus ?

Le terme d'indexation est ambigu. Il peut en effet renvoyer à la détermination soit d'indexeurs, soit de mots-clés. Les indexeurs sont des mots ou groupes de mots extraits du document et recensés en une liste, l'index. Pour chaque indexeur, l'index donne les références des passages (tous ou ceux jugés les plus pertinents) où il apparaît. L'index est ou non hiérarchisé mais la structure qui organise cette hiérarchie (la nature des relations) n'est pas explicitée. Les mots-clés, eux, sont définis a priori pour un domaine, souvent avec l'aide d'experts, et organisés en thésaurus. Ce thésaurus est également hiérarchisé avec des relations très rarement explicitées. Les documentalistes y choisissent les mots-clés adéquats pour caractériser un document.

En résumé, les mots-clés servent à retrouver un document pertinent à l'intérieur d'un ensemble de documents et les indexeurs servent à retrouver un passage pertinent à l'intérieur d'un document : dans les deux cas, il s'agit de retrouver du texte pertinent à l'intérieur d'un volume de données textuelles largement supérieur à la partie pertinente. De plus, les relations logico-sémantiques implicites qui structurent un index sont très probablement de la même nature que celles qui structurent un thésaurus. Enfin, il n'est pas impossible de trouver, à l'intérieur des index, des mots-clés, c'est-à-dire des mots qui ne sont pas utilisés dans le document mais qui sont pertinents pour accéder à des passages du texte.

Dans ce double mouvement : des mots-clés au texte et du texte aux indexeurs, comment peut-être utilisée la BCT ?

4.3.2 De la BCT à un index

Conçue en lien étroit avec le texte, il est évident que la BCT peut être directement utilisée pour constituer l'index. Trois éléments doivent cependant être examinés dans cette perspective :

- Il n'est pas certain qu'il soit judicieux de conserver toutes les relations conceptuelles de la BCT. D'une part, certaines relations sont sans doute moins parlantes, moins immédiates pour

l'utilisateur. Ainsi, dans le projet auquel nous participons, la relation "conditionne le début de", considérée comme très importante par le principal rédacteur du document, n'est sans doute pas à retenir dans l'index. D'autre part, il convient peut-être de ne pas surcharger l'index avec de multiples relations qui risquent de noyer les relations les plus structurantes. En réalité, seule l'expérimentation réelle, qui permettra d'évaluer la réaction des utilisateurs, permettra d'apprécier les choix qui auront été faits.

- Les termes de la BCT ne sont peut-être pas tous pertinents pour l'index ; il se peut que le nombre de termes retenus pour la BCT (près de 1500 pour le projet concerné) soit beaucoup trop élevé. Mais, là encore, il est difficile de décider du nombre idéal d'indexeurs, surtout tant qu'une expérimentation n'a pas été réalisée. Le fait de disposer de l'index et du document sur support informatique et non sur papier change en effet considérablement le nombre d'indexeurs tolérés.
- Le rajout d'éventuels mots-clés, méta-termes qui synthétiseraient un ensemble de termes, supposerait qu'on fasse appel à une "méta-BCT", ce qui nous amène à examiner les rapports entre BCT et thésaurus.

4.3.3 De la BCT à un thésaurus

Si on regarde de près les contenus d'un thésaurus et d'une BCT, il semble qu'il y ait assez peu de différences. Tous deux présentent des concepts structurés, des informations linguistiques et des informations d'usage. Les principales différences sont au nombre de trois :

- Dans une BCT, les informations linguistiques et conceptuelles sont nettement identifiées et distinguées. Dans un thésaurus, cette distinction est aplatie : on ne distingue pas, par exemple, les liens de synonymie ou d'antonymie des liens de généricité ou de partie-à-tout.
- Dans une BCT, les relations conceptuelles sont très définies, leur sémantique est clairement explicitée, on tend même vers leur formalisation. Dans un thésaurus, les relations conceptuelles sont peu précises et souvent même pas explicitées ; la relation "voir aussi", très vague, autorise toutes les interprétations.
- Une BCT est toujours construite à partir de corpus alors qu'un thésaurus est construit à partir de la connaissance d'un domaine.

Un thésaurus se distingue d'un index par le fait que l'index est construit en référence à un texte alors que le thésaurus est construit par introspection. Une BCT, elle, comme l'index, est toujours construite en référence à un texte. Cependant, une BCT pourrait être utilisée comme thésaurus si elle est construite à partir d'un très important volume de textes qui couvrent un domaine entier, en limitant alors le nombre des relations. Enfin, soulignons qu'un des intérêts majeurs d'une BCT est son mode de construction systématique (par opposition à la manière souvent intuitive dont sont construits index et thésaurus). La BCT serait donc plus fiable puisqu'on peut vérifier l'origine des données et la manière dont elles ont été organisées.

4.4 Intérêt de la BCT pour guider la consultation documentaire.

La consultation documentaire «plein texte» met souvent en oeuvre des relations sémantiques pour élargir le champ des requêtes. Lorsque, par exemple, on ne trouve pas un mot dans un texte, il est tout à fait légitime de chercher le mot hyperonyme ou celui qui dénomme une partie du concept dénommé par le mot. Ainsi, si à la question "le satellite a-t-il été endommagé?", le logiciel de

recherche ne trouve que la phrase "la plate-forme a été endommagée", il doit pouvoir répondre positivement, sachant que la plate-forme est une partie du satellite.

Il existe désormais des projets de grande envergure qui visent à recenser, pour une langue, les relations sémantiques qui structurent le lexique. C'est le cas du projet Wordnet, et maintenant du projet Eurowordnet pour les langues européennes. Pour chaque nom (mais les verbes, adjectifs et adverbes sont aussi traités), on peut obtenir ses synonymes, antonymes, hyperonyme, ainsi que les noms qui dénomment l'objet dont le nom traité est une partie ou bien les objets qui sont des parties de l'objet dénommé par le nom traité. Il existe aussi une relation "terme coordonné" qui correspond au "voir aussi" des thésaurus.

Pour les domaines spécialisés, qui relèvent d'une connaissance experte, ce sont les terminologues qui, traditionnellement, travaillent sur le lexique. Mais alors les relations conceptuelles sont peu étudiées et lorsqu'elles sont notées, c'est souvent de manière intuitive. Dans le cas des corpus spécialisés, la BCT est certainement une source de données très fiable pour constituer le réseau des relations à retenir pour la recherche «plein texte». La seule difficulté consiste à déterminer celles des relations conceptuelles qui vont être retenues. Tout comme pour la constitution de l'index, il convient sans doute, en effet, de ne pas laisser ouverts trop de chemins possibles sous peine d'aboutir à une explosion combinatoire.

5 Conclusion

Pratiquement, notre projet est en phase finale. Les bilans tirés de ce travail, qu'ils concernent la pertinence de nos choix méthodologiques, de la représentation des connaissances proposée ou l'intérêt de constituer des BCT pour définir des applications documentaires, ne sont encore que partiels. En tant que source de connaissances et de données associées à un corpus, les BCT semblent des supports tout à fait prometteurs et des bases de travail intéressantes. Par contre, il nous reste encore à mieux évaluer le coût de leur mise au point, l'adéquation de notre démarche, et sa validité sur un autre type de corpus (plus volumineux, plus éloigné de la tâche des utilisateurs par exemple). D'autres pistes d'évaluation de nos hypothèses restent à explorer. Maintenant que nous avons développé un environnement informatique de gestion de BCT, nous allons mesurer son adéquation aux besoins des terminologues. Nous envisageons aussi d'autres types d'utilisation des BCT conçues selon notre approche et notre modèle, comme la construction d'ontologies ou de modèles formels (Napoli,93), voire la consultation directe par des acteurs de différents métiers amenés à coopérer.

L'étude du passage des BCT aux ontologies est d'un enjeux majeur en IA car il peut donner une validité nouvelle à ces ontologies (Aussenac,95). Dans une perspective plus linguistique, les BCT soulèvent la question de la continuité entre thésaurus et BCT. On peut se demander si les BCT, assimilables à une version informatisée des thésaurus, en sont ou non l'avenir.

Bibliographie

AUSSENAC-GILLES N., BOURIGAULT D., CONDAMINES A., GROS C. (1995) : « How Can Knowledge Acquisition Benefit from Terminology ? », Proc. 9th Knowledge Acquisition Workshop, Banff. pp. 1/1-1/19.

BACHIMONT B. (1996) : « Engagement sémantique et engagement onotoloqique : propositions méthodologiques et problèmes théoriques à propos des ontologies en IA ». Actes des JAC. Séte. Mai 1996.

BOURIGAULT D. (1994) : « Extraction et structuration automatique de terminologie pour l'aide à l'acquisition de connaissances à partir de textes », Actes du 9ème congré RFIA'94. Paris. pp 397-408.

BOURIGAULT D., CONDAMINES A., (1995) : « Réflexions sur le concept de base de connaissances terminologiques ». Actes des Journées du PRC-GDR-IA, Nancy. TEKNEA, Toulouse. pp 425-444.

CONDAMINES A. (1996) : "Analyse de textes pour l'acquisition de données terminologiques". Terminologies Nouvelles (Bruxelles) n° 14 , pp 35-42.

CONDAMINES A. et REBEYROLLE J. (1997) : "Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode". Actes d'IC'97, Roscoff (F), pp. 191-206.

DAOUST F. (1992) : « SATO (Système d'Analyse de Texte sur Ordinateur) V 3.6 : Manuel de référence ». Centre ATO - Univ. de Québec à Montreal (UQAM).

GROS C., ASSADI H., AUSSENAC-GILLES N., COURCELLE A. (1996) : « Task models for Technical Documentation Accessing ». Compl. to the Proc. of IXth EKAW'96. Nottigham (UK).

GROS C., ASSADI H., (1997) : « Les systèmes de Consultation de Documentation technique ». 1ères journées du chapitres français de l'ISKO, Lille (F).

JOUIS C (1994) : "SEEK, un logiciel d'acquisition de connaissances utilisant un savoir linguistique sans employer de connaissances sur le monde externe". Actes des JAC'94, Grenoble (F), pp.159-172.

KAVANAGH J. (1996) : "The Text Analyzer: a tool for extracting knowledge from text". Master thesis, Univ. d'Ottawa.

MATTA N., AUSSENAC-GILLES N. (1996) : « Le schéma du modèle conceptuel, étape dans la modélisation des connaissances ». Acquisition et Ingénierie des Connaissances : tendances actuelles. Toulouse : Cépaduès.

MEYER I., SKUCE D., BOWKE L., ECK K. (1992) : « Toward a New Generation of Terminological Ressources : An Experiment in Building a Terminological Knowledge Base ». COLING. pp 956-960. Nantes, F.

NAPOLI A., VOLLE P. (1993) : « Une introduction aux logiques terminologiques ». Rapport de recherche 93-R-033, Centre de Recherche en Informatique de Nancy, Vandoeuvre Les Nancy.

PLANTE P., DUMAS L. (1995) : « Manuel utilisateur de NOMINO-SIGNET ». Atelier FX 6.0. ATO - Univ. de Québec à Montreal (UQAM).

RIVIER A. (1990) : "Construction des langages d'indexation, aspects théoriques". Documentaliste, vol. 27, n°6, pp. 263-279.

SEGUELA P., AUSSENAC-GILLES N. (1997) : « Un modèle de base de connaissances terminologiques ». Actes des Journées Terminologie et Intelligence Artificielle TIA'97. Toulouse : ERSS. pp 47-68.