D'une méthode à un guide pratique de modélisation de connaissances à partir de textes

Nathalie Aussenac-Gilles, Brigitte Biébow, Sylvie Szulman

Conférence TIA-2003, Strasbourg, 31 mars et 1 avril 2003

D'une méthode à un guide pratique de modélisation de connaissances à partir de textes

Nathalie Aussenac-Gilles*, Brigitte Biébow** et Sylvie Szulman**

* Université Toulouse 3, Institut de Recherche en Informatique de Toulouse (IRIT)
118, route de Narbonne, F-31062 Toulouse Cedex 4
** Université de Paris-Nord, Laboratoire d'Informatique de Paris-Nord (LIPN)
Av. J.B. Clément, F-93430 Villetaneuse
{Brigitte.Biebow,Sylvie.Szulman}@lipn.univ-paris13.fr

Résumé

Nous proposons dans cet article un guide pratique de modélisation pour construire des ressources ontologiques à partir de textes. Ce guide a été élaboré pour un industriel et mis en oeuvre sur une application. Il détaille l'utilisation de logiciels de traitement automatique des langues et de modélisation, afin d'aider un analyste lors du dépouillement et de la fouille dans les données textuelles. Il précise pas à pas les différentes étapes en mettant en évidence la navigation entre les nombreux éléments de modélisation.

Mots-clés : ontologie, modélisation à partir de textes, ressources terminologiques, méthode

1. Introduction

L'objectif de cet article est de contribuer à un guide méthodologique qui explique "comment" construire un modèle de connaissances (de type ontologie, terminologie, etc.) à partir de textes. La méthode Terminae que nous avons développée fournit un cadre général de construction de produits terminologiques à partir de textes. Elle est fondée sur l'analyse linguistique de textes par des outils de traitement automatique des langues (TAL). Or la pratique montre que le cogniticien chargé de la modélisation se trouve démuni devant la masse d'informations qu'il doit progressivement maîtriser pour créer le produit terminologique adapté à l'application et à ses utilisateurs. Nous nous proposons de décrire des recommandations qui portent sur la phase de normalisation sémantique de notre méthode, c'est-à-dire qui guident la structuration des résultats d'analyse de textes. Ces recommandations sont le fruit de plusieurs expériences, dont la principale répondait à un besoin industriel que nous utilisons pour illustrer cet article.

Dans une première partie, nous rappelons le cadre méthodologique général de Terminae et nous décrivons le contexte de l'application ainsi que les outils de TAL et de représentation utilisés. Puis nous détaillons le guide de réalisation d'un produit terminologique. Nous terminons en situant notre propos par rapport à d'autres travaux similaires.

2. Contexte applicatif

2.1. Cadre méthodologique général de Terminae

Notre méthode, inscrite dans la problématique du groupe TIA [1], propose un cadre pour la construction de produits terminologiques à partir de textes. Ces produits, ou ressources quand ils sont utilisés dans une application, consistent en un ensemble plus ou moins structuré de termes et/ou de concepts : lexique, index, glossaire, terminologie, réseau conceptuel ou thésaurus, ontologie (Aussenac-Gilles *et al.*, 2002). Terminae vise essentiellement la constitution de terminologies, réseaux conceptuels et ontologies. La méthode peut être découpée en quatre étapes, les trois dernières étant mises en oeuvre de manière cyclique. L'importance de chacune dépend du produit terminologique visé et des objectifs d'utilisation de ce dernier.

La *Constitution d'un corpus* vise à choisir documents techniques, comptes rendus, livres de cours, etc. à partir d'une analyse des besoins de l'application.

L' *Étude linguistique* consiste à identifier des termes et des relations lexicales, en utilisant des outils de traitement de la langue naturelle (extracteurs de termes, outils d'analyse distributionnelle, outils d'aide au repérage de relations par des patrons linguistiques, etc.).

La *Normalisation sémantique* conduit à définir dans un langage formel des concepts et des relations sémantiques que nous appelons terminologiques car provenant des termes et relations précédemment étudiés (Biébow & Szulman, 1999). Leur structuration en réseau s'appuie sur les résultats du dépouillement des textes tout en tenant compte de l'objectif d'utilisation de l'ontologie. Elle nécessite l'ajout de nouveaux concepts et relations dits de structuration.

La *Formalisation* permet de préciser, complèter et valider le modèle construit lors de la normalisation. L'analyste indique si les concepts sont primitifs ou définis, vérifie que les relations sont à la bonne place pour favoriser un héritage maximum, etc.

Terminae offre une continuité entre les différentes formes de l'ontologie. Celle-ci passe d'un état proche d'une taxinomie de termes à un réseau conceptuel enrichi de relations et de concepts de structuration pour aboutir à une ontologie formelle. Elle est décrite dans un langage formel masqué à l'analyste par l'interface de Terminae, avec des contraintes de validité minimale. Mais pour que le modèle hiérarchique soit correct, l'analyste doit avoir compris dès le début le mécanisme d'héritage des relations (un concept fils hérite des caractéristiques de ses pères décrites dans l'ontologie, et ce transitivement).

Le logiciel Terminae associé à la méthode fournit des aides pour toutes les étapes de l'analyse des textes à la formalisation. Il offre un support méthodologique qui permet d'évoluer progressivement et en conservant des liens des textes vers les niveaux linguistique et conceptuel.

2.2. Axes d'analyse des connaissances dans la modélisation à partir de textes

Le travail de modélisation à partir de textes nécessite de nombreux allers-retours entre les étapes d'étude linguistique et de normalisation, c'est-à-dire entre texte et modèle. Il va au delà d'une étude linguistique : une expression linguistique, terme ou relation, est validée non pas uniquement par rapport au corpus, mais aussi en fonction de sa pertinence par rapport au modèle en cours et à l'application visée.

[1] Le groupe de recherche français TIA (Terminologie et IA) correspond au thème 6.2 du GDR I3. http://www.biomath.jussieu.fr/TIA/

L'étude des textes offre un point de départ pour trouver des éléments de modélisation qui mènent aux concepts centraux du modèle. Nous appelons *concept central* un des concepts clés du domaine, qui peut facilement être demandé aux experts ou trouvé dans les textes. La tâche menée s'apparente à un *dépouillement* : des données vers le modèle. Dans un mouvement inverse, des critères de "bonne structuration du modèle" poussent à l'enrichir à l'aide d'éléments particuliers, répondant à des besoins de modélisation précis. Un des moyens d'identifier ces éléments est de retourner du modèle vers les textes, dans un mode *fouille*. Le cheminement classique du cogniticien consiste à donner priorité au dépouillement au début de son travail, pour privilégier la fouille des données dès qu'il dispose d'un modèle suffisamment structuré. Mais les deux types de tâches continuent d'être effectuées.

Au sein du modèle même, la progression suit différents axes qui sont explorés à tour de rôle et autant de fois que nécessaire. *L'axe ascendant* consiste à regrouper des concepts isolés ou des parties de hiérarchies indépendantes. Il faut donc identifier et expliciter un ou plusieurs points communs entre des concepts, trouver ou définir leur père, le situer dans la hiérarchie, etc. *L'axe descendant* vise à détailler la hiérarchie : il s'agit de trouver tous les fils d'un concept, de veiller à leur homogénéïté vis à vis d'un ou plusieurs critères. Le choix des critères et du niveau de détail auquel s'arrêter sont à préciser. *L'axe centrifuge* revient à se focaliser sur un concept et à chercher toutes les relations qui le concernent. Cette démarche est une bonne amorce de la modélisation, ou de la vérification d'un modèle, alors que l'analyse du modèle selon les deux autres axes est plus pertinente pour le compléter. Ces différentes dimensions d'analyse permettent de mieux caractériser les tâches effectuées par le cogniticien et éclairent deux aspects importants de son travail : sa diversité, qui se traduit par la nécessité de changer souvent la manière d'aborder les connaissances ; sa complexité, due aux divergences, voire aux contradictions, entre sources de connaissances et objectifs de modélisation.

2.3. L'application : la demande, le corpus

L'expérience relatée provient d'un contrat industriel (Saint Gobain Recherche) demandant la création d'une ontologie, pour favoriser la communication et la recherche d'information dans l'industrie de la fibre de verre. Cet objectif général a été décliné dans un premier temps en une évaluation de l'intérêt d'une ontologie au sein d'un système de classification documentaire pour la veille technologique. Le corpus utilisé est hétérogène, de langue anglaise et comporte différents types de documents : un livre technique très complet et relativement pédagogique, faisant référence dans le domaine, des textes de brevets déposés par l'entreprise et des articles de la presse économique relatifs à l'industrie concernée. Chacun de ces documents couvre plusieurs aspects de l'industrie de la fabrication de la fibre de verre : des aspects techniques classiques ou innovants, des aspects concurrentiels et économiques. Les différentes composantes du corpus ont été traitées séparément et comparées pour tenir compte de leurs spécificités.

2.4. Les outils d'analyse de la langue utilisés

Trois logiciels de TAL utilisés sont Syntex, Upery et Yakwa. Ces outils ont en commun d'aider à repérer des régularités d'usage de la langue, et de faire l'hypothèse que ces régularités sont révélatrices de sens. Ils fournissent des résultats bruts à interpréter via une interface de consultation et en fonction des objectifs de leur utilisation. Syntex et Upery disposent d'une interface de dépouillement et d'exploitation des résultats commune.

Syntex est un analyseur syntaxique de corpus en français ou en anglais (Bourigault & Fabre, 2000). Syntex identifie non seulement les noms et les syntagmes nominaux, mais aussi les

verbes et syntagmes verbaux. A partir d'un corpus étiqueté, il construit un réseau de dépendance syntaxique, dit "réseau terminologique", dans lequel chaque syntagme est relié d'une part à sa tête (ou recteur, lien T) et d'autre part à son expansion (ou régi, lien E). Les liens sont étiquetés par le nom de la relation syntaxique de dépendance. Par exemple, le syntagme *E glass composition* a pour recteur *composition*, pour régi *E glass*, les deux étant reliés par une relation "complément de nom" (NN). Cette relation est représentée par le triplet (*composition*, NN, *E glass*). Les noeuds du réseau (mots et syntagmes) sont appelés *candidats termes*. Syntex donne pour chaque candidat terme, sa fréquence dans le corpus, sa productivité en tête et en expansion.

Upery (Bourigault, 2002) applique des principes d'analyse distributionnelle aux résultats de Syntex pour proposer des regroupements de syntagmes à partir des relations de dépendance trouvées autour des candidats termes. Un triplet (recteur, relation, régi) du réseau de dépendance de Syntex donne lieu à un couple (contexte, terme) où le contexte est le couple (recteur, relation) et le terme est le régi. L'analyse distributionnelle rapproche les candidats termes qui partagent un grand nombre de contextes syntaxiques (par exemple, les termes *strand, rod, Fibre* partagent deux contextes *Nom_lengths-Prep_of* et *V_make_OBJ*), et de manière symétrique les contextes qui contiennent les mêmes termes (ainsi, *Nom_parameters_NN V_be_SUJ Nom_bushings-Prep_of* ont en commun un terme *designs*). Upery peut ainsi aider l'analyste à construire des classes conceptuelles ou à trouver des relations.

Yakwa (Rebeyrolles & Tanguy, 2000) est un concordancier travaillant sur un corpus étiqueté. Il permet de construire des patrons lexico-syntaxiques et d'afficher toutes les parties d'un corpus contenant ce patron pour, éventuellement, leur donner du sens. Dans notre contexte, c'est donc un bon support pour la recherche de relations sémantiques à partir de leurs formes lexicales en corpus, qu'elles soient générales (adaptation de patrons généraux) ou spécifiques au corpus étudié (construction de patrons nouveaux). Yakwa nous sert aussi à vérifier des résultats trouvés par ailleurs ou à affiner des patrons de recherche autres que les marqueurs de relations binaires.

2.5. Outils de représentation

Nous tenons à souligner l'importance de pouvoir utiliser conjointement un outil de modélisation et des outils de TAL, de pouvoir définir aisément des concepts à partir de termes jugés « à retenir » ou de pouvoir revenir à des résultats bruts ou aux textes à partir d'un modèle. Le travail de modélisation n'est ni linéaire ni direct. Il nécessite de naviguer dans le réseau terminologique autant que de parcourir le modèle, et de pouvoir passer de l'un à l'autre en conservant les contextes en cours d'étude. Terminae permet de naviguer en partie dans le réseau terminologique Syntex, de définir des fiches terminologiques puis des concepts et des relations et, inversement, de revenir des concepts vers les termes puis vers les textes.

2.5.1. Représentation au niveau terminologique

A l'aide de Terminae, l'analyse des termes et des expressions lexicales débouche soit uniquement sur la définition de termes et de leurs synonymes dans des fiches terminologiques, soit sur une fiche et sur la création de structures au niveau conceptuel. La continuité entre l'observation de phénomènes en langue et la représentation au sein du modèle conceptuel de Terminae incite à choisir le plus rapidement possible la manière dont on va représenter une connaissance au niveau conceptuel. Or on n'a pas toujours les réponses pour faire ces choix.

Ceci a justifié d'utiliser, dans le cas de notre étude, une représentation, externe à Terminae, permettant de conserver l'information sous une forme très proche de son expression dans les

textes. Un tableur (Excell) a été utilisé pour représenter le réseau terminologique sous forme de fiches et de relations hypertextuelles. Les tableaux ont été intuitivement organisés soit par concept au sens large (une fiche pour chaque type de PROCESS) soit par type de relation (une fiche décrit toutes les instances d'une relation d'un type donné comme PRODUCES). Les liens hypertextes permettent de passer d'une vue à l'autre et de naviguer au sein du réseau ainsi constitué. Le passage de ce niveau informel à la représentation dans Terminae est manuel.

2.5.2. Représentation au niveau conceptuel

La définition de concepts et de relations dans Terminae est étroitement liée aux primitives offertes par le langage de formalisation, qui est de la famille $\mathcal{ALN}$ des logiques de description (Woods & Schmolze, 1992), maintenant à la base de la plupart des langages de représentation d'ontologie (DAML, OIL, OWL, ...). De ce fait, la représentation au niveau conceptuel prépare la formalisation et anticipe certaines des questions qu'elle pose.

(a) Concept, instance ou relation ? La même connaissance du domaine peut être représentée par une classe (Concept Générique dans Terminae) ou par une instance (Concept Individuel dans Terminae) ou par une relation (rôle dans Terminae). Ainsi, on peut utiliser une relation PARTIE_DE particulière, HAS_STEPS, pour relier un procédé et les étapes qui le composent, les étapes étant elles-mêmes des sous-classes de PROCESS. On peut aussi décider de créer un concept STEP qui aurait comme sous-classes toutes les étapes possibles, et de relier par HAS_STEPS chaque procédé et ses propres étapes. Des éléments complémentaires peuvent être consignés sous forme de commentaires.

(b) Concepts primitifs ou définis ? La distinction entre concept défini et concept primitif rend compte du fait qu'un concept soit totalement défini ou non par l'ensemble de ses rôles. Comme les concepts primitifs bloquent les inférences de classification, il est souhaitable de définir le plus possible les concepts, donc de les différencier par au moins un rôle, tant que cela ne débouche par sur une trop grande complexité. Terminae permet de vérifier que l'ontologie est correcte (tous les concepts définis doivent être différenciés deux à deux), que l'héritage des propriétés est valide et propose une classification optimale des concepts définis.

(c) Organisation des rôles et structuration des relations : Dans Terminae, les relations autres que hiérarchiques sont exprimées par des rôles. Un rôle est caractérisé par un nom, un concept domaine (départ de la relation) et un concept valeur (cible de la relation). Un rôle a également une cardinalité, qui permet de préciser, pour une instance de concept domaine particulière, combien d'instances de concept valeur peuvent lui être associées. Choisir le concept domaine et le concept valeur d'un rôle revient à définir la signature de la relation qu'il représente. Ces concepts correspondent aux classes les plus génériques qu'il peut associer. Spécialiser le rôle revient à le redéfinir pour qu'il relie des concepts plus spécifiques.

3. Des heuristiques pour guider la démarche de normalisation

Notre contribution porte avant tout sur la phase de normalisation située au coeur du processus de modélisation. Il se dégage de notre pratique une démarche itérative, que nous présentons selon deux types de tâches étroitement liés : structuration des données tirées des textes, où le travail de dépouillement l'emporte sur le travail de représentation ; normalisation du modèle et application de critères de "bonne modélisation", selon une démarche privilégiant la fouille.

3.1. Tâches de structuration de données tirées des textes

Les tâches de structuration visent à organiser au niveau terminologique, voire au niveau conceptuel, des informations repérées dans les textes ou parmi les résultats des outils de TAL. Ces tâches viennent d'abord répondre au problème de la gestion de la masse des données produites par ces outils, et aiguiller le cogniticien qui se demande "par où commencer ?" ou "comment exploiter ces résultats ?" avant d'en guider l'organisation dans un modèle. Conduites de manière cyclique, elles permettent d'obtenir des parties de modèles, des hiérarchies partielles, éventuellement indépendantes, pas toujours homogènes ni cohérentes entre elles. La structuration occupe entièrement le début du processus de modélisation, puis sa part diminue au profit de la représentation dans le modèle et de sa vérification. Nous suggérons dans la suite des critères pour décider des termes puis des concepts à retenir, puis nous fournissons des repères pour les décrire au niveau lexical et conceptuel.

3.1.1. Repérage des termes à analyser

A l'amorce du processus, une démarche centrifuge consiste à décider quels termes étudier pour définir des concepts centraux. Plus tard, lorsqu'un début de modèle existe, les termes sont des indicateurs pour décider des sous-ensembles du modèle à développer. Parmi l'ensemble des candidats termes proposés par Syntex, ceux présentant une ou plusieurs des caractéristiques suivantes peuvent être étudiés en priorité :

(a) Termes complexes les plus fréquents ou productifs dans le corpus : dans notre étude, *fibre size, glass fibre, E glass, raw materials, bushing position, glass composition* sont les SN les plus fréquents et productifs ; FIBRE SIZE sera un concept à part entière, qui désigne la gaine couvrant la fibre de verre pour la rendre plus résistante, et non la taille de la fibre.

(b) Termes simples ayant de nombreux contextes partagés : on consulte la liste des termes triée en fonction du nombre de co-occurrents syntagmatiques, c'est-à-dire les termes qui sont souvent associés à ces termes par des relations grammaticales (12 termes simples dont *strand, fibre, bushings, glass, sizes, temperature, materials, processing* ont plus de 40 termes voisins).

(c) Nom ou SN composés de plusieurs termes très fréquents et productifs : *fibre size*, *fibre glass* et *bushing position* ressortent nettement car *fibre, glass, bushing, sizes* et *position* sont fréquents et très productifs.

(d) Termes apparaissant dans les titres : le fait qu'une des occurrences d'un terme se trouve dans un titre peut confirmer qu'il est à retenir (cf. *rovings* ou *fibre size*); inversement, la lecture systématique des titres fait ressortir des termes (cf. *foreheads, raw materials, process*).

(e) Termes choisis grâce à une connaissance superficielle du corpus et du domaine : le terme *process* a été étudié bien avant des termes plus productifs pour cette raison.

(f) Termes indiqués par d'autres ressources (ici, le lexique du domaine présent dans le livre) ou par les experts au cours de réunions de validation.

Ces différents critères se renforcent et permettent ici de dégager entre autres *glass fibre*, *fibre size*, des composés de *product* et de *process* ou *processing*. On a donc fait ressortir très nettement le vocabulaire au coeur du domaine industriel étudié : les procédés de fabrication et les produits, intermédiaires ou finaux, qui en découlent, sans disposer de connaissances préalables sur le domaine.

3.1.2. Première analyse linguistique

L'étude porte ensuite sur les termes retenus et s'appuie sur les logiciels Syntex, Upery, Yakwa, la partie linguistique de Terminae, avec lesquels l'analyste va pouvoir :

(a) Consulter les termes complexes produits par ces termes : ainsi, *glass fibre* est en tête de 31 termes (*continuous glass fibre, waste glass fibre*, ...) et forme l'expansion de 140 termes comme *glass fibre strands* ou *glass fibre product* ;

(b) Lire plusieurs contextes dans lesquels ces termes complexes sont utilisés ;

(c) Formuler des requêtes pour des recherches ciblées à l'aide d'un concordancier ; par exemple, vérifier sur les occurrences fournies par Yakwa qu'un terme est toujours pris dans le même sens, lister les autres termes avec lesquels il est souvent conjointement utilisé, etc. ; dans le projet, nous avons comparé l'usage du verbe *to process* et celui des noms *process* et *processing* pour constater que c'est le nom qui est presque toujours utilisé ;

(d) Etudier les termes qui les constituent. Ainsi, l'étude de *glass* a permis de voir que ce nom est souvent utilisé comme ellipse de *glass fibre*, et donc les deux termes (considérés ensuite comme synonymes) ont été étudiés pour décrire le concept de GLASS FIBRE.

Ces éléments sont autant d'indices pour repérer le ou les sens d'un terme et pour décider de définir dans le modèle un ou plusieurs concepts associés. L'analyse permet aussi de savoir si le(s) concept(s) que désigne le terme est sujet ou objet de nombreuses actions, s'il interfère avec d'autres concepts, s'il est décrit explicitement dans les textes comme important, etc. Toute information jugée utile et pertinente pour le modèle doit pouvoir être consignée au fur et à mesure selon une des propositions faites en 2.5.1.

3.1.3. Structurer une hiérarchie locale autour des concepts déjà identifiés

Ensuite, le choix des parties de modèle à développer se fait en priorité à partir des concepts déjà définis. Pour un concept choisi, on recherche, toujours à partir du réseau tête-expansion de Syntex et des textes, des termes synonymes de celui qui le désigne ainsi que des termes désignant des concepts spécifiques et génériques. Dans le modèle, ces concepts seront reliés au concept étudié par la relation EST_UN, ce qui le situe au sein de la hiérarchie, l'ossature de l'ontologie dans Terminae.

Nous illustrons cette étude avec le concept PROCESS, repéré par les termes synonymes *process* et *processing*. PROCESS a été placé dans un premier temps au plus haut niveau de la hiérarchie à côté des concepts PRODUCT et MATERIALS. Situer un concept dans la hiérarchie EST_UN s'appuie sur quatre types de données tirées des textes :

(a) le réseau tête-expansion qui relie les groupes nominaux. Les constructions linguistiques révélant des types particuliers de PROCESS sont de l'un des types suivants : *Process of manufacturing the glass yarns, process for manufacturing thin profiles, the manufacturing process of components*. *Manufacturing process* ressort comme un nom de concept important qui serait une sous-classe de PROCESS. La plupart des autres termes comportant *process* en tête sont autant de types de PROCESS. Il convient ensuite d'étudier leurs occurrences pour savoir comment les regrouper en sous-classes à différents niveaux sous PROCESS.

(b) le réseau des relations syntagmatiques qui permettent de mettre en vis à vis autour d'un verbe pivot des groupes nominaux sujet et objet de ce verbe. Pour le concept PROCESS, les co-

occurrents syntagmatiques du verbe *to process* contiennent presque systématiquement les noms *manufacture* ou *manufacturing*. De ce fait, on voit que tous les PROCESS dont on parle sont des MANUFACTURINGPROCESS, ce qui ne justifie pas de créer 2 concepts différents.

(c) les contextes partagés de Upery, où les verbes formant les contextes sont souvent des indicateurs de relations ; dans ce corpus, peu de termes sont riches en contextes parce que le texte présente peu de redondances et de régularités. Pourtant, de l'étude des contextes partagés de *glass* (qui a 2 voisins : *roving* et *flue gases*) ressort la relation lexicale *nom_manufacture_of* qui montre que les 3 termes étudiés pourraient être regroupés. C'est un premier indice pour créer un concept PRODUCT.

(d) l'utilisation de Yakwa, à l'aide de marqueurs mis en forme à partir des résultats de Syntex ou à l'aide des marqueurs génériques. Parmi les marqueurs de la relation EST_UN très utilisés dans ce corpus, mentionnons *is said*, *is referred to as*.

3.1.4. Etude des autres relations concernant ces concepts

Pour un concept choisi, on recherche des termes correspondant à des concepts reliés, à ses propriétés ou à ses attributs à partir des énoncés des textes. L'approche par marqueurs (2.4), s'applique ici, avec une majorité de marqueurs spécifiques car nous disposions de peu de marqueurs génériques pour l'anglais au début du projet.

(a) Identification de relations et de marqueurs propres au domaine. Nous avons mis en évidence trois manières de trouver des relations spécifiques au domaine et des marqueurs associés : la lecture d'occurrences, de phrases du corpus ; la lecture des rapprochements entre termes au sein du réseau trouvé par Syntex et Upery ; la projection avec Yakwa de couples de termes associés à des concepts dont on sait qu'ils sont en relation. Ainsi, le contexte *Nom_part-Prep_of* permet de relever des relations de type PARTIE_DE (avec des sens différents) pour les trois termes qui partagent ce contexte *furnace, direct melt processing et processing*. La consultation du modèle vient confirmer ou non l'intérêt de modéliser ce type de relation. Par exemple, ont été retenues les relations IS_PRODUCT_OF entre un PRODUCT et un PROCESS si ce produit est le résultat du procédé et la relation REQUIRES entre un procédé et les produits nécessaires à sa mise en oeuvre.

(b) Mise au point de marqueurs à partir de ces observations : Il s'agit de trouver, à partir de formes lexicales, une forme générique pertinente et adaptée au langage d'expression des marqueurs dans Yakwa. Les marqueurs sont ajustés après projection sur le corpus et observation d'un échantillon de phrases retournées. Par exemple, plusieurs marqueurs très différents ont été identifiés pour la relation REQUIRES : *needs* ; *calls_for* ; *depends_on* ; *conditions* ; *requires*.

(c) Exploitation des résultats de la projection des marqueurs : Ces résultats sont évalués systématiquement en fonction du contenu courant du modèle pour décider si chacune des occurrences du marqueur doit ou non donner naissance à une relation dans le modèle. Par exemple, on peut se demander s'il est utile de décrire finement tous les procédés dans le modèle, y compris les étapes terminales de certains procédés très particuliers.

(d) Exploitation de marqueurs génériques : De la même manière, les marqueurs génériques sont évalués, ajustés puis projetés, et leur résultats intégrés ou non dans le modèle. Cette tâche est habituellement la première lorsqu'on dispose d'une gamme de marqueurs génériques, ce qui n'était pas notre cas pour l'anglais. Elle devient peu pertinente si les marqueurs sont peu productifs sur le corpus étudié.

Une relation lexicale identifiée en corpus ne donnera pas forcément lieu à une relation entre concepts dans le modèle (Aussenac-Gilles & Séguéla, 2000). Avant de représenter une relation par un rôle, les questions suivantes se posent :

- plusieurs relations au niveau linguistique vont-elles correspondre à un ou plusieurs rôles?
- une relation associée à un terme ou reliant des termes doit-elle être ou non représentée dans le modèle? Doit-elle relier les concepts directement associés à ces termes ou des concepts plus spécifiques voire plus génériques ?

A ce moment de l'analyse, la manière de représenter les résultats est très incertaine. Ceci justifie le fait de passer par une représentation très souple, facilitant la navigation et suffisamment proche du niveau terminologique pour laisser ouverts les choix de représentation conceptuelle par la suite, comme nous l'avons décrit en 2.5.1.

3.2. Tâches de normalisation

Il s'agit de reprendre les différentes parties du modèle, de le vérifier et de le compléter. Des vérifications individuelles portent sur les concepts pris un à un, alors que d'autres, plus globales, portent sur les rôles. Ces tâches consistent à justifier que chaque élément est nécessaire dans l'ontologie, pertinent à cet endroit et défini conformément à l'objectif de modélisation. Nous énonçons quelques principes généraux avant de lister les différents points à contrôler.

3.2.1. Principes généraux

(a) Unicité de définition : L'unicité peut d'abord être exprimée au niveau terminologique (par les étiquettes des concepts par exemple), et une justification commentée des définitions des concepts et des rôles suffit. Elle s'appuie sur des connaissances trouvées dans les textes ou sur des arguments de modélisation (regroupement de concepts ayant des propriétés communes, frères de concepts absents des textes mais repérés dans d'autres sources, etc.). Terminae contrôle l'unicité de définition des concepts et signale les concepts définis par les mêmes propriétés. Pour l'instant, la plupart des fils de MANUFACTURINGPROCESS ne se différencient les uns des autres que par leurs propres fils et sont considérés comme identiques.

Inversement, on peut se demander quelle(s) différence(s) justifie(nt) de définir des concepts distincts alors qu'ils sont sémantiquement proches du point de vue de celui qui modélise. Il arrive souvent de définir 2 concepts différents parce que 2 termes sont trouvés dans les textes, et de constater plus tard qu'ils renvoient aux mêmes connaissances dans le modèle. C'est le cas de MANUFACTURINGPROCESS et PROCESS, qui doivent ne former qu'un seul concept.

(b) Homogénéité de point de vue et cohérence des descriptions : A chaque niveau de la hiérarchie des concepts, les relations associées à un concept doivent être choisies en cohérence avec un même point de vue. Par exemple, la relation hiérarchique convient bien pour décrire les procédés plus spécifiques d'un procédé : EGLASSMELTINGPROCESS, REMELTPROCESS, DIRECTMELTPROCESS sont bien des cas particuliers (donc des fils) de MELTINGPROCESS. Mais MELTINGPROCESSSTEPS ne peut pas être un frère de ces concepts, car sa relation à MELTINGPROCESS est une relation PARTIE_DE qui n'a pas à être confondue avec la relation d'héritage EST_UN.

3.2.2. Vérification d'un concept

Pour chaque concept de l'ontologie, les réponses à ces questions peuvent être trouvées soit dans les textes, soit demandées à un expert, soit jugées secondaires et laissées sans réponse

volontairement.

(a) Valider son étiquette (terme associé), et les mots synonymes de cette étiquette

(b) Confirmer sa place dans la hiérarchie, en appliquant les principes de différenciation de (Bachimont, 2000) : traits sémantiques communs et différents entre le concept et son père puis ses frères ; ces connaissances peuvent être explicitées par un commentaire ou par un rôle, mais elles ne sont pas toujours accessibles dans les textes. Ainsi, les fils de MANUFACTURINGPROCESS ne sont pas explicitement différenciés autrement que par leur nom.

(c) Inversement, confirmer la complétude et l'unicité de point de vue de la spécialisation du concept père : selon ces mêmes principes, on doit se demander si le concept étudié a d'autres frères et vérifier s'ils sont de même niveau et s'ils sont comparables selon le trait sémantique commun. Si plusieurs points de vue sont entre-mêlés au niveau des fils du concept (plusieurs traits sémantiques cohabitent et ne sont pas partagés par tous), il faut pouvoir regrouper les fils par point de vue. Cela oblige soit à créer des concepts intermédiaires (comme FIBERDRAWINGSTEPS sous FIBERDRAWINGPROCESS), soit à privilégier un des points de vue dans la hiérarchie, et à refléter les autres à l'aide d'autres types de relations (comme HAS_STEPS).

(d) Vérifier l'ensemble des relations héritées, depuis la racine vers ce concept ; juger de leur validité pour ce concept mais aussi qu'elles sont définies au bon niveau de la hiérarchis.

3.2.3. Vérification d'un rôle

Cette vérification nécessite de lister l'ensemble des concepts reliés par un même rôle, ainsi que l'ensemble des rôles. Vérifier la définition d'une relation au niveau conceptuel revient à :

(a) s'assurer que le nom du rôle sera bien interprété avec le sens donné au moment de sa création, que ce sens est le même pour tous les rôles de ce type, que les concepts domaine et valeur sont bien choisis et que la cardinalité est valide.

(b) contrôler la signature du rôle, ce qui revient à se poser des questions sur les regroupements possibles des rôles et sur le niveau où les définir dans la hiérarchie. Des rôles associés à des concepts ne pourraient-ils pas être associés à leur père et hérités également par les frères de ce concept ? Par exemple, les rôles PRODUCES et REQUIRES ont d'abord été identifiés (sous divers noms) pour des cas particuliers, donc des fils, de MANUFACTURINGPROCESS avant d'être associés à ce concept.

4. Discussion - Etat de l'art

S'il existe quelques travaux sur les méthodes générales de construction d'ontologies (Fernandez *et al.*, 1997), les réflexions générales sur des critères de définition des concepts (Bachimont, 2000), (Guarino & Welty, 2000) et les conseils de bonne pratique comme vérifier la consistance entre les termes et éviter la circularité (Ushold & Gruninger, 1996) sont rares. Les auteurs s'intéressant à la construction d'ontologie à partir de textes avec des outils de TAL ont pour la plupart une approche aussi automatisée que possible et ne s'attardent pas sur l'inévitable partie manuelle (Maedche & Staab, 2000), (Kang & Lee, 2001). Le praticien reste donc dépourvu face aux résultats des outils de TAL à partir desquels il doit élaborer un modèle.

Seuls, à notre connaissance, (LeMoigno *et al.*, 2002) ont expliqué leur démarche de construction d'une ontologie à partir de corpus. Utilisant Syntex et Upery, leur travail est très proche du nôtre. Tout d'abord, grâce à l'analyse distributionnelle, ils commencent par dégager les grandes

classes de concepts qui conduiront à des sous-ontologies. Ils analysent les couples de contextes verbaux les plus proches pour déterminer les principales classes de verbes, puis considèrent les expansions autour de ces verbes pour déterminer d'autres classes, puis les expansions des expansions et ce de manière itérative. Par exemple, un verbe d'action a pour expansion une action qui a pour expansion une localisation : cette analyse mène au concept ANATOMIE.

Dans un deuxième temps, ils complètent et raffinent l'ontologie en identifiant des couples de contexte "clés" produisant de nouvelles classes. Le travail de complétion s'effectue alors sur un mode centrifuge, en analysant les contextes partagés, les termes de ces contextes, puis en itérant sur les contextes des termes pris deux à deux. Les relations entre concepts sont définies au fur et à mesure de l'analyse des contextes partagés, par exemple la relation LOCALISATION, lorsqu'il faut relier les concepts de classes partageant des contextes. Les notions de classes, concepts et termes sont un peu confuses, ce qui est sans doute dû à l'outil qui met ensemble des termes et des contextes de termes sans que ces regroupements aient un statut autre que celui donné par l'analyste. De plus, l'ontologie n'étant pas opérationnalisée, c'est-à-dire décrite dans un langage de représentation des connaissances, elle reste très proche de la langue avec la facilité de lecture et les inconvénients qui en découlent.

Le travail précédent émane comme le nôtre du groupe TIA. Il se réfère également aux travaux de Bachimont sur une approche différentielle de construction d'ontologie. Ce dernier préconise l'élaboration d'une ontologie régionale, c'est-à-dire valable pour un contexte d'interprétation unique (lié entre autres à l'application), dans lequel le sens des termes est normalisé et défini par rapport aux autres. Il construit un arbre de concepts linguistiques dont les différences et les ressemblances entre père et fils ou entre frères sont explicitées sous forme de commentaires à partir de l'expression linguistique de ces critères. Un arbre de relations est également élaboré de la même manière. Cette étape permet de figer linguistiquement la signification des étiquettes de concepts. L'étape suivante transforme l'ontologie régionale en ontologie référentielle, par l'ajout de nouveaux concepts et relations, en changeant de paradigme : les concepts sont maintenant des concepts formels. Enfin, le passage à un langage de représentation des connaissances aboutit à une ontologie opérationnelle. Pour un exposé complet, voir (Troncy & Isaac, 2002). Cette approche a été utilisée dans le projet Menelas pour une application de compréhension de comptes rendus d'hospitalisation [2].

Comme LeMoigno et al., nous retenons de ce travail l'application de principes différentiels sans chercher à créer une première ontologie purement linguistique. Celle-ci peut être justifiée dans des applications de compréhension ou génération du LN, mais elle est parfois très artificielle. Ainsi, dans notre étude de cas, il n'a pas été possible de trouver dans le corpus une relation pour différencier des concepts frères de niveau élevé dans la hiérarchie comme PRODUCT et PROCESS, ou comme tous les fils de MANUFACTURINGPROCESS. Il faudrait faire appel aux experts pour "forcer" la différenciation. Par contre, le texte seul permet de repérer les termes synonymes de MANUFACTURINGPROCESS : *process of manufacturing, manufacturing of, processing the manufacture of*. Les principes différentiels conduisent à distinguer les fils de MELTINGPROCESS, sans forcément créer de relation pour expliciter leurs différences. Nous considérons l'application de ces principes comme un guide utile à la modélisation, mais sans qu'on en retrouve forcément les résultats dans l'ontologie. Le lien avec le texte est pour nous plus important pour rendre l'ontologie intelligible que le passage par une ontologie linguistique intermédiaire qui ne correspond pas toujours aux usages mis en évidence dans le corpus.

[2] http://www.biomath.jussieu.fr

5. Conclusion

Notre pratique nous oriente vers l'élaboration itérative et la plus rapide possible d'ontologie, quitte à la modifier souvent au début avant d'obtenir un état stable. En effet, ce sont les concepts centraux qui structurent l'ontologie et l'étude de leurs relations avec d'autres concepts qui dirige l'étude linguistique. La signification de ces concepts s'exprime par leur place dans la hiérarchie, par leurs relations avec d'autres concepts, et par des commentaires. Cette signification découle de l'usage des termes associés.

Nos expériences soulignent l'intérêt de passer par une représentation très souple au début de la modélisation, puis une représentation formelle dès la structuration de concepts pour garantir une validité minimale du modèle. En début de modélisation, il faut faciliter la navigation entre textes et représentation conceptuelle et rester suffisamment proche du niveau terminologique pour laisser ouverts les choix de représentation conceptuelle par la suite. Ensuite, l'outil de modélisation doit offrir une continuité entre les différents états, valides ou non, de l'ontologie.

Ce travail est bien sûr loin d'être terminé. Il mériterait d'être validé par de nouvelles expériences, pour voir si les heuristiques et la démarche proposées sont encore pertinents dans de nouveaux contextes. Ensuite, il serait intéressant de faire évoluer les outils en fonction des besoins mis au jour. De plus, il faudrait reprendre les tâches et principes proposés pour les formuler non plus en fonction des capacités des logiciels utilisés, mais en terme de moyens et d'objectifs plus génériques, seulement liés à la structure de la ressource ontologique à construire et aux éléments de la langue trouvés dans les textes. Enfin, nous devons mesurer l'impact des outils choisis et théories linguistiques sous-jacentes sur la démarche proposée.

Références

AUSSENAC-GILLES N., CONDAMINES A. & SZULMAN S. (2002). Prise en compte de l'application dans la constitution de produits terminologiques. In J. L. MAITRE, Ed., *Information, Interaction, Intelligence : Actes des 2e Assises Nationales du GDR I3*, p. 289–303: Cépaduès Editions.

AUSSENAC-GILLES N. & SÉGUÉLA N. (2000). Les relations sémantiques: du linguistique au formel. *Cahiers de Grammaire , Numéro spécial "sémantique et corpus"*, **25**, 175–198.

BACHIMONT B. (2000). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. In J. CHARLET, M. ZACKLAD, G. KASSEL & D. BOURIGAULT, Eds., *Ingénierie des Connaissances, évolutions récentes et nouveaux défis*, p. 305–323, Paris: Eyrolles.

BIÉBOW B. & SZULMAN S. (1999). TERMINAE: A linguistics-based tool for building of a domain ontology. In D. FENSEL & R. STUDER, Eds., *Proc. of the 11th European Workshop (EKAW'99)*, LNAI 1621, p. 49–66: Springer-Verlag.

BOURIGAULT D. (2002). Upery : un outil d'analyse distributionnelle étendu pour la construction d'ontologies à partir de corpus. In *Actes de la 9ième conférence annuelle sur le Traitement Automatique des Langues (TALN 2002)*, p. 75–84, Nancy, France.

BOURIGAULT D. & FABRE C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire , Numéro spécial "sémantique et corpus"*, **25**, 131–151.

FERNANDEZ M., GOMEZ-PEREZ A. & JURISTO N. (1997). Methontology : From ontological arts towards ontological engineering. In *AAAI 97 Springs Symposium Series on Ontological Engineering*, p. 33–40, Stanford USA.

GUARINO N. & WELTY C. (2000). Identity, unity and individuality : Towards a formal toolkit for ontological analysis. In *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI'2000)*, p. 219–223, Berlin: IOS Press.

KANG S.-J. & LEE J.-H. (2001). Semi-automatic practical ontology construction by using a thesaurus, computational dictionaries and large corpora. In *ACL workshop on Human Language technologies and Knowledge Management*, p. 29–36.

LEMOIGNO S., CHARLET J., BOURIGAULT D. & JAULENT M.-C. (2002). Construction d'une ontologie à partir de corpus: expérimentation et validation dans le domaine de la réanimation chirurgicale. In *Actes des 13ièmes journées francophones d'Ingénierie des Connaissances (IC 2002)*, p. 63–74.

MAEDCHE A. & STAAB S. (2000). Semi-automatic engineering of ontologies from text. In *Proceedings of the Twelfth Inernational Conference on Software Engineering and Knowledge Engineering (SEKE'2000).*

REBEYROLLES J. & TANGUY L. (2000). Repérage automatique de structures linguistiques en corpus: le cas des énoncés définitoires. *Cahiers de Grammaire , Numéro spécial "sémantique et corpus"*, **25**, 153–174.

TRONCY R. & ISAAC A. (2002). Doe: une mise en oeuvre d'une méthode de structuration différentielle pour les ontologies. In *Actes des 13ièmes journées francophones d'Ingénierie des Connaissances (IC 2002)*, p. 229–238.

USHOLD M. & GRUNINGER M. (1996). Ontologies : Principles, methods and applications. *Knowledge Engineering Review*, **11:2**, 93–136.

WOODS W. A. & SCHMOLZE J. G. (1992). The KL-ONE family. *Computers Mathematical Applications*, **23**, 133–177.