Construction d'ontologies médicales à partir de textes : propositions methodologiques

Audrey Baneyx, Jean Charlet

Construction d'ontologies médicales à partir de textes : propositions methodologiques

Audrey Baneyx[1], Jean Charlet[1,2]

[1] INSERM U729 - Laboratoire SPIM
Faculté de Médecine Broussais-Hôtel-Dieu
15 rue de l'Ecole de médecine, 75006 Paris, France
[2] STIM - DSI/AP-HP
{Audrey.Baneyx,Jean.Charlet}@spim.jussieu.fr

Résumé : Dans le contexte du codage des activités médicales, il est nécessaire de construire des représentations conceptuelles des connaissances. Cet article apporte des propositions méthodologiques sur la construction d'ontologies médicales, à partir de textes, à l'adresse d'un ingénieur cogniticien. Cette méthodologie est fondée sur la mise en œuvre des principes de la sémantique différentielle et utilise les outils de traitement automatique de la langue. Notre principale hypothèse de recherche concerne l'utilisation conjointe de deux méthodes : une méthode éprouvée qui consiste à construire des ressources termino-ontologiques par analyse distributionnelle et une méthode fondée sur la recherche de relations sémantiques par l'utilisation de patrons lexico-syntaxiques.

Mots-clés : Ontologie, ingénierie des connaissances, extraction terminologique, sémantique différentielle, analyse distributionnelle, patrons lexico-syntaxiques, pneumologie.

1 Introduction

La réduction des inégalités de ressources entre les établissements de santé figure dans la réforme de l'hospitalisation (ordonnance du 24/04/96). Afin de mesurer l'activité et les ressources des établissements, le gouvernement souhaite disposer d'informations quantifiées et standardisées. Ces informations sont recueillies pour chaque séjour d'un patient sous la forme de résumés standardisés de sortie dans lesquels la codification des diagnostics principaux et secondaires est effectuée à partir de la classification internationale des maladies CIM-10. La procédure de codification appelée Programme de Médicalisation des Systèmes d'Information [1], couramment « codage PMSI », est le plus souvent réalisée ma-

[1] http://www.atih.sante.fr/

nuellement par les praticiens qui s'aident d'un thésaurus de spécialité (Bensadoun, 2001). Ces thésaurus proposés par les sociétés savantes sont construits pour permettre aux médecins de coder à partir de leur terminologie usuelle mais il est aujourd'hui manifeste que les outils d'aide au codage fondés sur ces thésaurus sont inadaptés aux besoins du praticien (Friedman *et al.*, 2004). En effet, les libellés de ces thésaurus se révèlent ambigus (par exemple, à un même code sont associées plusieurs pathologies) et non exhaustifs, le mode de classification choisi est difficile d'utilisation, et le maintien de la consistance ainsi que de la cohérence du thésaurus est impossible. On constate que le sens des libellés de ces nomenclatures médicales (SNOMED, CIM-10,...) repose sur les facultés d'interprétation du lecteur humain et qu'elles ne sont donc pas adaptées à une exploitation par l'ordinateur. Dans ce contexte, il nous semble indispensable de décrire la sémantique et l'organisation des objets du domaine médical afin de se doter de modélisations conceptuelles (non contextuelles et non ambiguës) dont le sens est inscrit dans la structure même du modèle. Une telle modélisation est appelée ontologie (Staab & Studer, 2003). Notre hypothèse est que le développement de telles ressources ontologiques permettra de mettre au point des outils performants, fiables et maintenables pour l'aide au codage. La problématique à laquelle nous nous attaquons est donc la construction de ces ontologies et le terrain d'expérimentation dans lequel nous nous situons est l'aide au codage en pneumologie.

Le terme « ontologie » est utilisé depuis le début des années 90 dans les domaines de l'intelligence artificielle, en particulier de l'ingénierie des connaissances et de la représentation des connaissances. Son champ d'application s'élargit considérablement et il fait désormais partie des objets de recherche courants, notamment dans le secteur de la modélisation des systèmes d'information (Gomez-Pérez *et al.*, 2004). Une ontologie est un système formel dont l'objectif est de représenter les connaissances d'un domaine spécifique au moyen d'éléments de base, les concepts, définis et organisés les uns par rapport aux autres (Rector, 1998). La représentation ontologique des connaissances garantit le maintien de la cohérence des axiomes et de l'intégrité du système, ainsi que l'extensibilité de la représentation sans modification de la structure. Il est cependant difficile de repérer et de classifier les objets d'un domaine. Les critères de classification dépendent des buts poursuivis et n'ont rien d'immuable (Charlet, 2002). Ainsi, nous ne prétendons en aucun cas construire une ontologie universelle de la médecine mais bien une ontologie régionale de la pneumologie (Bachimont, 2000).

Nous précisons nos objectifs section 2. La section 3 présente le matériel utilisé et la section 4 détaille les différentes étapes de notre méthodologie. Nous donnons les résultats obtenus dans la section 5 et concluons cet article, section 6, en discutant de l'intérêt d'un tel travail et des perspectives qu'il offre.

2 Objectifs

Ce travail de construction d'ontologies s'inscrit dans le cadre, plus large, du projet de recherche PERTOMed [2], financé par le CNRS. Le but de ce projet est de développer une infrastructure proposant un ensemble de méthodes et d'outils opérationnels pour la production et l'utilisation de ressources terminologiques ou ontologiques dans le domaine médical. Les ontologies médicales sont construites en étroite collaboration avec des groupes d'utilisateurs, avec lesquels seront, en particulier, mises en place des procédures d'évaluation de ces ressources dans leur contexte d'usage. Au sein du projet, notre travail consiste à proposer aux médecins pneumologues un environnement d'aide au codage et à la représentation des connaissances médicales reposant sur le modèle conceptuel d'une ontologie du domaine concerné. Nous travaillons en étroite collaboration avec le service de Pneumologie et de Réanimation du groupe hospitalier de la Pitié-Salpêtrière de Paris[3] et avec la Société de pneumologie de langue française[4].

Nous construisons notre ontologie régionale à partir de ressources textuelles sur lesquelles nous appliquons des techniques appartenant au domaine du traitement automatique du langage dans le but de développer les corpus nécessaires à la structuration de l'ontologie. La méthodologie que nous employons a été mise au point au sein du groupe TIA sur, entre autre, les principes de sémantique différentielle (Bachimont, 2000). Notre principale hypothèse de recherche concerne l'utilisation en parallèle de deux méthodes pour enrichir le travail de construction de l'ontologie : *a*) une méthode éprouvée qui consiste à construire des ressources termino-ontologiques par analyse distributionnelle (Bourigault & Lame, 2002), et *b*) une méthode fondée sur la définition *a priori* d'une relation sémantique, puis sur l'observation de séquences en corpus qui véhiculent la relation souhaitée (Séguéla, 2001). Sachant qu'aucune ontologie ne couvre le domaine de la pneumologie, notre objectif est double : il s'agit, d'une part, de construire l'ontologie de la pneumologie et, d'autre part, d'apporter des précisions sur les deux premières étapes de la méthodologie. Nous proposons notre propre expérimentation de construction d'ontologies médicales dans la même optique que le travail de Le Moigno *et al.* (2002). Un point de vue quelque peu différent est adopté ici puisque l'ontologie est construite par un ingénieur cogniticien et non par un expert du domaine médical comme dans ce précédent travail. L'intérêt consiste à mettre au point un processus méthodologique précis destiné à l'ingénieur cogniticien de manière à ne faire appel à l'expert médical que pour des moments précis de validation.

Nous souhaitons, dans ce papier, montrer précisément le déroulement des deux premières étapes de la méthodologie de construction d'une ontologie différentielle de la pneumologie et l'apport conjoint de l'analyse distributionnelle et des patrons lexico-syntaxiques pour la mise en œuvre des principes différentiels.

---

[2] Production et évaluation de ressources terminologiques et ontologiques dans le domaine de la médecine, `http://www.spim.jussieu.fr/Pertomed`

[3] UPRES EA2397

[4] `http://www.splf.org/`

3 Matériel : Corpus et outils

Dans le but de couvrir, avec le plus d'exhaustivité possible, l'ensemble de l'activité de la pneumologie, nous avons collecté des comptes rendus d'hospitalisation (corpus intitulé [CRH]) dans six hôpitaux de l'Assistance Publique-Hôpitaux de Paris[5]. Au total, nous disposons de 1 038 CRH. Ce premier corpus [CRH] compte environ 417 000 mots. Sachant qu'il a été établi dans (Le Moigno *et al.*, 2002) que 350 000 mots est un minimum pour obtenir de bons résultats avec nos outils, le corpus [CRH] semble être une bonne base d'expérimentation. Le second corpus, intitulé [Livre], est construit d'après un ouvrage pédagogique et correspond environ à 823 000 mots.

Nous utilisons le logiciel Syntex-Upery comme outils d'analyse de traitement du langage. Syntex est un module d'analyse syntaxique fondée sur l'hypothèse que les mots qui ont un sens proche se caractérisent par des dépendances similaires (Bourigault & Lame, 2002). Ainsi, ce module permet d'obtenir des relations de dépendances syntaxiques entre mots ou syntagmes (noms *vs* syntagmes nominaux, verbes *vs* syntagmes verbaux et adjectifs *vs* syntagmes adjectivaux). A la fin du traitement nous obtenons un réseau de dépendances syntaxiques – ou réseau terminologique – dont les éléments sont les candidats termes qui vont nous servir pour construire l'ontologie. Le module Upery met ensuite en œuvre le principe de l'analyse distributionnelle « à la Harris » (Harris, 1968) : il calcule des proximités distributionnelles entre les candidats termes du réseau sur la base des contextes syntaxiques partagés. Nous obtenons un réseau de candidats termes, leurs proximités contextuelles et leurs liens avec le corpus source. Les résultats de l'analyse sont visualisables dans Termonto, l'interface d'accès et de traitement des données du logiciel. L'éditeur DOE[6] permet de construire notre ontologie selon la sémantique différentielle. Ce logiciel permet également de compléter l'ontologie en ajoutant à chaque concept sa traduction en anglais ainsi qu'une définition encyclopédique. Il en va de même pour les relations. L'ontologie est exportée en OWL, un langage de représentation des connaissances préconisé par le consortium W3C.

4 Méthode

La méthodologie mise en œuvre permet de décrire les variations des sens des termes considérés en contexte. C'est pourquoi cette méthode considère que le corpus textuel est la source privilégiée permettant de caractériser les notions utiles à la modélisation ontologique et le contenu sémantique qui leur est associé. Nous distinguons quatre étapes : 1) la constitution du corpus des connaissances et son analyse par des outils de traitement automatique du langage, 2) la normalisation sémantique des termes du domaine grâce à la mise en œuvre des

[5] Ils se répartissent comme suit : Créteil : 326 CRH, Hôtel-Dieu : 97 CRH, Kremlin-Bicêtre : 125 CRH, Pitié-Salpêtrière : 57 CRH, Saint Antoine : 372 CRH, Tenon : 61 CRH.

[6] The Differential Ontology Editor, `http://opales.ina.fr/public`

principes différentiels, 3) l'engagement ontologique qui permet de formaliser les concepts, 4) l'opérationnalisation de l'ontologie dans un langage de représentation des connaissances interprétable par l'ordinateur (Bachimont *et al.*, 2002). Notre expérimentation dans le domaine de la pneumologie nous permet d'adapter et de préciser, pour l'ingénieur cogniticien, les deux premières étapes de la méthodologie, ce qui contribue à la réalisation de notre second objectif.

4.1 Traitement des ressources de base

Les deux corpus [CRH] et [Livre] nous parviennent sous des formats inexploitables par les outils d'analyse du langage. Ils sont donc prétraités[7] puis mis sous un format semi structuré par des programmes que nous avons développés. Nous disposons ainsi d'un corpus [CRH] anonyme et d'un corpus [Livre] didactique, tous deux au format XML. Le corpus [CRH] est ensuite traité par le logiciel Syntex-Upery. Le résultat de l'analyse distributionnelle nous permet de construire les éléments de base - *i.e.* primitives - de l'ontologie. Le second corpus [Livre] est analysé par le biais de patrons lexico-syntaxiques prédéfinis qui permettent d'extraire des couples d'unités lexicales correspondant au motif de la relation sémantique recherchée (hypéronymie, synonymie...). Les résultats obtenus nous aident à contrôler et enrichir la hiérarchie de l'ontologie.

4.2 Choix des candidats termes

Les résultats fournis sur la base du corpus [CRH] servent de support dans le choix de candidats termes[8] (CT) représentatifs de la pneumologie en tant qu'activité médicale. Nous distinguons deux étapes dans leur sélection. Le travail est manuel et s'appuit sur les fonctionnalités de Termonto.

1) Nous parcourons l'ensemble des résultats fournis par l'analyse syntaxique et choisissons d'étudier, en premier lieu, les syntagmes nominaux (SN) dont la fréquence d'apparition en corpus est supérieure à 12 (2 % du corpus). Nous repérons les grands axes conceptuels typiques du corpus et donc du domaine représenté. A chaque CT, nous associons un critère de validité (ainsi nommé dans Termonto), compris dans un intervalle allant de 1 à 6, correspondant à l'un de ces axes : 1 (CT non pertinent appartenant à l'axe Autres), 2 (réservé aux CT déjà modellisés dans l'ontologie), 3 (CT appartenant à l'axe Symptômes), 4 (CT appartenant à l'axe Pathologies), 5 (CT appartenant à l'axe Signes) et 6 (CT appartenant à l'axe Traitements/Examens). Par exemple, nous fixons à 6 le critère de validité pour tous les CT de ce dernier axe - *e.g. examen*, *doppler*, *radiographie*, *etc.* Au début de la méthodologie, tous les CT ont un critère de validité égale à 1 et à la fin égale à 2 car ils sont, en principe, tous définis dans

[7] Les fichiers ont été convertis au format texte, « nettoyés », anonymisés, segmentés, associés à des identifiants de section et de phrase, étiquetés et analysés morphosyntaxiquement par Cordial Analyseur de la société Synapse.

[8] Un candidat terme est un syntagme nominal composé d'une tête et d'une expansion. Par exemple, dans le SN *Opacité dans le poumon gauche*, le terme *Opacité* est la tête du syntagme et *dans le poumon gauche* est son expansion.

Descendants en tête	**Descendants en expansion**	**Voisins en tête**	**Voisins en expansion**
Épanchement pleural droit	Lame d'épanchement pleural	Lésions	Liquide
Épanchement pleural liquidien	Récidive d'épanchement pleural	Infection	Infiltrats
Épanchement pleural de la grande cavité	Lier la dyspnée à l'épanchement pleural	Signes	Décompensation

TAB. 1 – Exemple de résultats du rapprochement contextuel pour le SN *Epanchement pleural.*

l'ontologie. Les critères de validité 3, 4, 5 et 6 sont utilisés temporairement durant la phase de construction. Ces regroupements permettent une première phase de travail sur les rapprochements des CT par contexte. La sélection par critère de validité laisse 35 % des CT sur lesquels élaborer le cœur de notre ontologie.

2) L'analyse distributionnelle rapproche deux à deux les termes partageant les mêmes contextes (descendants en tête et en expansion). Comme cette analyse est symétrique, elle rapproche également les contextes en fonction des termes qu'ils partagent (voisins en tête et en expansion). Sur le tableau 1 *épanchement* est la tête du SN *épanchement pleural* et *pleural* est son expansion. Les descendants en tête donnent des informations sur ce qui pourrait être des concepts fils ou des concepts définis. Les descendants en expansion donnent des informations sur la place du concept dans la hiérarchie, sur le concept père. Les voisins en tête et en expansion nous permettent de constituer des regroupements de candidats termes sémantiquement proches du candidat terme étudié ici, *épanchement pleural.* Ces regroupements sont d'une grande aide pour élaborer la structure hiérarchique de l'ontologie, aussi bien l'axe horizontal que vertical. L'exemple ci-dessous montre un premier rapprochement possible : nous pouvons mettre en rapport le groupe A *{épanchement, lésion, infection, décompensation}* avec *{signes}*. Les CT du groupe A partagent un même contexte sémantique, la première hypothèse est donc qu'il peut s'agir de concepts frères dont *signes* est possiblement le concept père.

4.3 Mise en œuvre des principes différentiels

Pour élaborer cette hiérarchie, il convient d'articuler les CT choisis dans la précédente étape (*cf.* section 4.2) en précisant les principes différentiels qui les définissent. Par exemple, le concept *Ultrasonographie* et le concept *ExamenIsotopique* sont des concepts frères dont le concept père est *ImagerieParRayonnement* (*cf.* figure 1). Le principe de communauté avec le concept père est la projection ou l'injection d'une substance artificielle dans le but d'effectuer des mesures. Le principe de communauté entre les concepts frères est lié au média d'injection. Le principe différentiel entre les concepts frères est relatif au type de media artificiel

mis en œuvre : un isotope dans le cas de l'*examen isotopique* (la scintigraphie est un exemple d'examen isotopique) et les ultrasons pour l'*ultrasonographie*. Les candidats termes des 4 axes conceptuels (3, 4, 5 et 6) sont définis selon ces principes.

Les résultats de l'analyse par patrons lexico-syntaxiques sur le corpus [LIVRE] nous aident à définir les principes différentiels. Les patrons lexico-syntaxiques représentent des motifs de relations sémantiques spécifiques. Ils sont construits autour d'un marqueur, également appelé pivot, qui est l'indice d'une relation lexicale, comme le marqueur *entre autres* pour la relation d'hyperonymie. Ainsi, un patron de la forme *DET SN, entre autres SN* permet d'extraire l'unité lexicale *Les méningites, entre autres pathologies ...* et de mettre en relation d'hyperonymie *méningite* et *pathologie*. Cette méthode a été présentée dans (Hearst, 1992) et expérimentée dans plusieurs travaux, notamment dans (Caraballo, 1999). Les patrons lexico-syntaxiques liés à l'hyperonymie mettent en relation des couples père-fils potentiels intéressants pour contrôler et enrichir la structure hiérarchique de l'ontologie. Dans le cadre de la construction d'ontologies différentielles, nous appliquons cette méthode à la recherche d'énoncés définitoires en corpus pour aider à renseigner les principes différentiels. Les patrons que nous employons ont été développés par Malaisé *et al.* (2004). Le corpus [LIVRE], d'un genre pédagogique, à la particularité d'être très structuré et se révèle particulièrement propice à ce type de recherche. Par exemple, les patrons utilisant le marqueur *Il s'agit de* ont pu associer un titre de section (par exemple « Asthme ») à sa description dans la première ligne du texte (« Il s'agit d'une maladie inflammatoire des voies aériennes »). Les unités lexicales extraites sont validées manuellement et les hiérarchies créées sont visualisables sous DOE. Il est alors facile de comparer les deux structures terminologique obtenues : la hiérarchie issue de l'analyse distributionnelle du corpus [CRH] et celle issue du repérage par patrons lexico-syntaxiques sur le corpus [LIVRE], et d'en tirer les informations nécessaires à l'enrichissement et au raffinement de l'ontologie. La comparaison des structures obtenues est détaillée dans (Baneyx *et al.*, 2005).

Enfin, il paraît important de souligner que l'ingénieur cogniticien doit, tout au long du travail de construction, décider si le concept doit être primitif ou défini, c'est-à-dire s'il est essentiel ou non par rapport aux buts poursuivis. Un concept défini est construit à partir d'un ou plusieurs concepts primitifs et d'une ou plusieurs relations. Le SN *douleur thoracique gauche* est modélisé par le concept défini suivant : *[Douleur](au niveau du)[thorax](localisée à)[gauche]*[9]. A la fin de cette étape, nous avons normalisé le sens des termes du domaine et représenté la hiérarchie des concepts primitifs et des relations avec DOE.

[9]Le crochet indique un concept primitif et la parenthèse indique une relation. Dans notre méthodologie, les relations sont traitées de la même manière que les concepts.

5 Résultats

Après l'utilisation de SYNTEX, le corpus [CRH] donne 36 881 SN. D'après les résultats de l'analyse syntaxique et de l'analyse distributionnelle, le SN *chimiothérapie de <nom>* a le plus grand nombre de voisins en tête , soit 28, et sa fréquence d'apparition dans le corpus s'élève à 190, le SN *<nom> de chimiothérapie* a le plus grand nombre de voisins en expansion, soit 52, et sa fréquence d'apparition est également la plus haute, soit 454. Nous pouvons vérifier la pertinence des rapprochements par groupe de candidats termes dont les contextes d'apparition sont sémantiquement proches. Pour *cure de chimiothérapie* par exemple : *[Hospitalisation, Examen, Navelbine, Cisplatine, Doxorubicine, Taxotere, Carboplatine, MIP]* sont ses voisins en tête et *[Traitement, Bilan, Antibiothérapie, Injection, Radiothérapie]* sont ses voisins en expansion. Ces résultats sont examinés et mis sous une forme ontologique visualisable avec DOE. Nous construisons la hiérarchie suivante : *ActionMedicale/Traitement/TraitementMedicamenteux/Chimiotherapie*. La chimiothérapie étant considérée comme un traitement médicamenteux, nous retrouvons les principes médicamenteux suivant classés sous *Medicament/PrincipesMedicamenteux/Navelbine, Cisplatine, Doxorubicine...* Les candidats termes *Antibiothérapie* et *Radiothérapie* sont également placés sous *Traitement*. Cette méthode de regroupement semble donner de bons résultats pour construire l'ontologie et rend la tâche bien plus facile pour un ingénieur cogniticien non spécialiste du domaine médical modélisé.

Le repérage de définition par patrons lexico-syntaxiques sur le corpus [LIVRE] permet d'extraire 799 unités lexicales. Nous en validons 119, ce qui représente près de 15 % des extractions. La comparaison des deux structures terminologiques obtenues (*cf.* tableau 2) permet de distinguer s'il s'agit de hiérarchies complémentaires, identiques ou comparables, ou bien divergentes. Ces informations permettent de préciser et de corriger la structure.

Notre ontologie (*cf.* figure 1) contient à ce jour 500 concepts primitifs, issus de la première analyse des candidats termes. Les étapes de construction 1 et 2 étant itératives, nous augmentons très rapidement la représentation en examinant les candidats termes dont la fréquence d'apparition dans le corpus [CRH] est inférieure à 12 et dont le critère de validité vaut 1.

La question de la validation en ingénierie ontologique n'est pas résolue. Bien que certaines pistes semblent se faire jour, il n'existe pas de méthode unanime pour évaluer ce type de travail. Nous construisons une ontologie de la pneumologie dans le but de servir de support à un outil d'aide au codage, aussi il nous parait important d'axer la question de la validation sur l'aide que nous serons en mesure d'apporter aux pneumologues, c'est-à-dire sur l'usage qu'ils feront de l'outil. Cependant, la tache n'étant pas terminée, nous validons des étapes intermédiaires, en terme de qualité et de complétude. Dès à présent, l'avancement de la hiérarchie conceptuelle est corrigé et validé périodiquement par des médecins pneumologues de la Société de pneumologie de langue fran-

Type	Exemple	Commentaire
Identique ou comparable	*Broncho pneumopathie/Asthme*[CRH] **vs** *Bronchopathie/Asthme à dyspnée continue*[LIVRE]	Les deux premières hiérarchies sont classifiées dans le MeSH sous *Poumon, maladie*, alors qu'*Asthme* est une notion plus spécifique dans la même branche hiérarchique. Cela valide la cohérence et la compatibilité des deux hiérarchies.
Complémentaire	*Signe/[...]/SigneRespiratoire/Insuffisance-Ventriculaire*[CRH] **vs** *Signe/Insuffisance ventriculaire droite*[LIVRE]	La deuxième arborescence vient confirmer la première et permet de la compléter d'un niveau, celui d'*Insuffisance ventriculaire droite*.
Divergente	*EtatPathologique/Maladie-Respiratoire/Bronchite* ET *Signe/Toux*[CRH] **vs** *Toux avec expectoration/Bronchite chronique*[LIVRE]	Dans le MeSH, la toux est classée à la fois comme *signe-symptôme* et comme *pathologie* : les deux sources textuelles illustrent chacune un de ces aspects.

TAB. 2 – Comparaison des deux structures terminologiques.

çaise avec laquelle nous collaborons. Le soin que nous apportons à la définition des principes différentiels et leur nature normalisatrice sont un gage de qualité et d'évolutivité pour notre ontologie. Ces principes rigoureusement appliqués assurent la cohérence et la robustesse de notre modélisation. Ainsi, la place de chaque concept étant bien définie dans la hiérarchie, il est plus facile de calculer la position de nouveaux concepts venant enrichir l'ontologie. Dans un deuxième temps, l'évaluation se fera également en testant la couverture de l'ontologie par rapport au thésaurus de spécialité. Cette validation garantit, autant que faire se peut, que l'ontologie développée sera adéquate à un outil d'aide au codage fondé sur le thésaurus de spécialité. Pour estimer cette couverture, nous allons vérifier la possibilité de construire une représentation conceptuelle des connaissances médicales en combinant les concepts primitifs et les relations dont nous disposerons dans l'ontologie. Pour l'instant, nous pouvons construire quelques concepts définis comme par exemple celui de *Chimiothérapie intra-pleurale* qui se trouve dans le chapitre « Plèvre » du thésaurus : *[ActionMedicale/Traitement/TraitementMedicamenteux/Chimiotherapie](RelationActe/ModeRealisation/RealiseAuNiveau/intra)[Anatomie/AppareilRespiratoire/Plevre].*

6 Discussion et conclusion

Nous avons présenté un ensemble de principes méthodologiques pour la construction, à partir de textes, d'ontologies médicales différentielles. Nous sommes en train d'améliorer l'ontologie de la pneumologie en analysant les termes présents dans les thésaurus de spécialité CIM-10 et CCAM. Cela étoffe les branches et

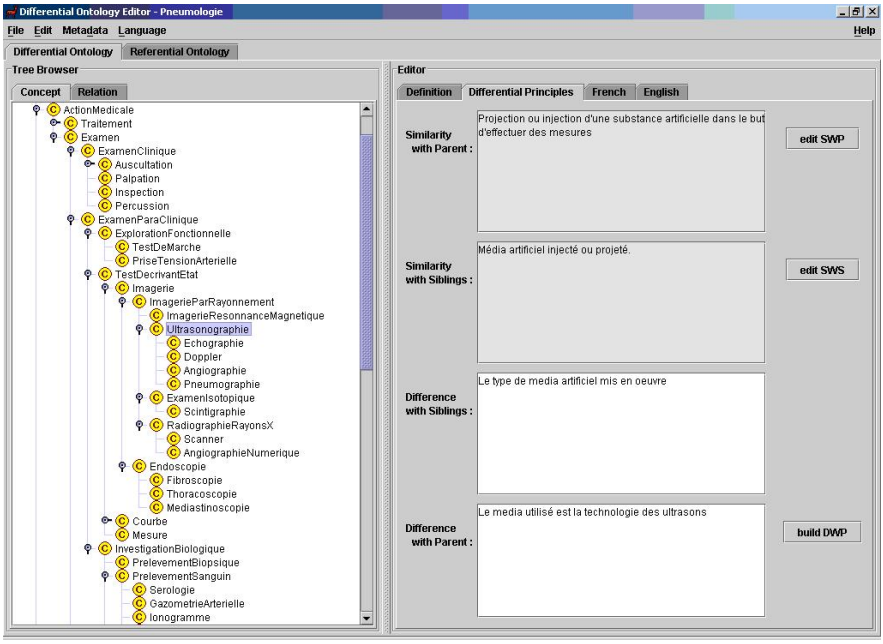

FIG. 1 – Un extrait de l'ontologie de la Pneumologie visualisable sous DOE.

augmente le nombre des feuilles de notre hiérarchie. De même, nous prévoyons de compléter l'ontologie en faisant le lien avec le haut de l'ontologie du projet Ménélas[10]. Ce travail nous permettra de vérifier s'il existe un haut niveau conceptuel commun au domaine médical, sachant que notre ontologie est régionale, c'est-à-dire spécifique au domaine de la pneumologie. Nous envisageons d'obtenir rapidement une ontologie de la pneumologie comptant environ 1 200 concepts. D'autre part, un des enjeux de ce travail est de montrer que la méthodologie mise au point permet à un ingénieur cogniticien, non spécialiste du domaine modélisé, de construire une ontologie à partir de textes à l'aide d'outils de traitement automatique des langues. Un travail récent dans le domaine de la réanimation chirurgicale (Le Moigno *et al.*, 2002) ainsi que les premiers résultats de notre recherche permettent de penser que nous allons dans la bonne direction.

Nous avons, en outre, présenté l'utilisation conjointe de deux méthodes adaptées chacune à un genre de corpus particulier. En prenant les textes comme source de connaissances, nous reposons la question des genres textuels développée dans (Aussenac-Gilles & Condamines, 2003). L'utilisation des corpus en Ingénierie des connaissances se veut une réponse au problème de l'accès aux connaissances d'un domaine, pour un objectif particulier, lié à une application informatique, dans notre cas une ontologie. *A priori*, les corpus textuels, comme ceux que nous

[10] `http://www.biomath.jussieu.fr/Menelas/Ontologie`

utilisons, témoignent d'un vocabulaire métier, fixé par l'écrit, consensuel car diffusé et partagé à l'intérieur du corps médical. Cela offre une garantie de fiabilité et de stabilité à notre modélisation. Dans le domaine particulier du traitement automatique de la langue appliqué au domaine médical, Pierre Zweigenbaum propose cinq catégories de mots-clés visant à caractériser les genres textuels (*Op. Cit.*) : dossier patient, enseignement, ressources, publications et oral. Dans ce cadre, notre corpus [CRH] fait partie du genre textuel dit « dossier patient » et notre corpus [LIVRE] de la catégorie « enseignement ». Nous avons également montré qu'il existe une relative compatibilité entre les deux hiérarchies terminologiques obtenues par l'utilisation de l'analyse distributionnelle puis la mise en oeuvre des principes différentiels et par l'emploi des patrons lexico-syntaxiques. La complémentarité de ces structures est intéressante car elle résulte de l'emploi de méthodes différentes appliquées à des corpus de genres textuels également différents. L'analyse de la divergence des résultats est riche en information et apporte un point de vue critique à l'ingénieur cogniticien et à l'expert sur la manière de modéliser le domaine. De plus, cette expérimentation montre qu'il existe des organisations conceptuelles différentes au sein d'un même domaine, ce qui appuit le fait qu'il s'agit bien de construire des ontologies régionales et non pas universelles car toute modélisation n'est jamais qu'un point de vue sur le monde. Soulignons, qu'il n'est bien évidement pas question ici de remplacer l'expert du domaine qui intervient à plusieurs moments clés de l'élaboration de l'ontologie : consultations préalables pour cerner aux mieux les besoins de la communauté, phases périodiques de validation des résultats, analyse des divergences, phases de test...

Nous avons également cerné certaines des limites liées à la comparaison de ces deux méthodes : un rapprochement semi-automatisé des hiérarchies nécessiterait, d'une part, la mise en œuvre de techniques plus sophistiquées d'appariement et, d'autre part, d'améliorer la précision des patrons lexico-syntaxiques et des marqueurs de relation définitoire en les adaptant spécifiquement au domaine médical. Ainsi, des marqueurs comme *indiquer* et *définir* sont à spécifier plus finement : *indiqué* est souvent utilisé dans le cadre de « traitements *indiqués* pour soigner une pathologie », *définir* intervient surtout dans des phrases telle que « Ces résultats ont permis d'acquérir certaines connaissances et de *définir* les meilleurs traitements pour soigner les 30 patientes ».

Ce travail de modélisation des connaissances à partir de textes nous a également permis de mesurer la nécessité d'utiliser conjointement des outils de traitement automatique du langage (SYNTEX-UPERY) et de modélisation (DOE). Il semblerait intéressant d'intégrer ces deux outils pour faciliter le passage des candidats termes à la représentation des concepts, tout en assurant de pouvoir revenir aux textes (Szulman & Biébow, 2004).

Pour conclure, nous précisons que l'étape finale de la validation de l'ontologie se fera par l'usage et nous tenterons de quantifier et de qualifier l'aide que notre travail apporte aux pneumologues. La méthodologie de construction d'ontologies différentielles utilisée est constructive, elle permet de placer de manière précise chaque concept dans la structure hiérarchique. L'ontologie doit être mise à dis-

position du corps médical au travers d'un environnement d'aide au codage des actes et des diagnostics et à la représentation des connaissances médicales, dans le cadre de la plateforme terminologique du projet PERTOMed.

Références

AUSSENAC-GILLES N. & CONDAMINES A. (2003). *Rapport de l'action spécifique ASTICCOT*. Rapport interne IRIT/2003-23-R, CNRS. Rapport de l'action spécifique ASTICCOT, « Terminologie et corpus » rattachée au RTP-DOC (RTP-33) du CNRS.

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*, chapter 19. Paris : Eyrolles.

BACHIMONT B., ISAAC A. & TRONCY R. (2002). Semantic commitment for designing ontologies : A proposal. In *Proceedings of EKAW*, p. 114–121, Sigüenza, Espagne : Springer.

BANEYX A., MALAISÉ V., CHARLET J., ZWEIGENBAUM P. & BACHIMONT B. (2005). Synergie entre analyse distributionnelle et patrons lexico-syntaxiques pour la construction d'ontologies différentielles. In *Actes de la conférences Terminologie et Intelligence artificielle*, p. 31–42, Rouen, France.

BENSADOUN H. (2001). Pmsi et chirurgiens : pourquoi les chirurgiens doivent-ils coder, comment bien coder ? *Journal de Chirurgie, Masson*, **138**(1).

BOURIGAULT D. & LAME G. (2002). Analyse distributionnelle et structuration de terminologie. application à la construction d'une ontologie documentaire du droit. *Traitement automatique des langues*, **43**(1).

CARABALLO S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Meeting of the Association for Computational Linguistics (ACL'99)*, p. 120–126, Maryland, USA.

CHARLET J. (2002). *L'Ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales*. Habilitation à diriger des recherches, Université Paris 6.

FRIEDMAN C., SHANIGA L., LUSSIER Y. & HRIPSACK G. (2004). Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, **11**, 392–402.

GOMEZ-PÉREZ A., FERNANDEZ-LOPEZ M. & CORCHO O. (2004). Ontological engineering. In *Advanced Information and Knowledge Processing*. Madrid, Spain : Springer.

HARRIS Z. (1968). *Mathematical Structures of Language*. New-York, USA : John Wiley and Sons.

HEARST M. (1992). Automatic acquisition of hyponyms from large text corpora. In A. ZAMPOLLI, Ed., *Proceedings of the 14 th COLING*, p. 539–545, Nantes, France.

LE MOIGNO 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 B. BACHIMONT, Ed., *Actes des 6 ^{es} Journées Ingénierie des Connaissances*, p. 229–38, Rouen, France.

MALAISÉ V., ZWEIGENBAUM P. & BACHIMONT B. (2004). Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologie. In P. BLACHE, Ed., *Actes de TALN 2004 (Traitement automatique des langues naturelles)*, p. 269–278, Fès, Maroc : ATALA LPL.

RECTOR A. (1998). Thesauri and formal classifications : Terminologies for people and machines. *Methods of Information in Medicine*, **37**(4–5), 501–509.

SÉGUÉLA P. (2001). *Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques*. Thèse de doctorat, Université Toulouse III.

STAAB S. & STUDER R. (2003). *Handbook on Ontologies*. Berlin, Germany : Springer, 1 edition.

SZULMAN S. & BIÉBOW B. (2004). OWL et Terminae. In *14e journées francophones d'Ingénierie des Connaissances (IC 2004)*.