

# **Synergie entre analyse distributionnelle et patrons lexico-syntaxiques pour la construction d'ontologies différentielles**

A. Baneyx<sup>1</sup>, V. Malaisé<sup>1,2</sup>, J. Charlet<sup>1,3</sup>, P. Zweigenbaum<sup>1,3,4</sup> et B. Bachimont<sup>2</sup>

<sup>1</sup>INSERM, U729;

15 rue de l'école de Médecine, 75006 Paris, France;

Fax : 01 53 10 92 01

audrey.baneyx@spim.jussieu.fr

<sup>2</sup>Institut National de l'Audiovisuel, DRE;

<sup>3</sup>AP-HP, DSI, STIM;

<sup>4</sup>INaLCO, CRIM

---

## **Résumé**

La construction d'ontologies dans le domaine médical est un enjeu scientifique significatif. Leur élaboration à partir de corpus peut bénéficier des apports du Traitement automatique des langues. Nous présentons dans cet article une expérimentation mettant en relief la complémentarité de l'analyse distributionnelle et du repérage par patrons lexico-syntaxiques. Nous évaluons et analysons le recouvrement terminologique et la compatibilité de deux hiérarchies obtenues à partir de ces méthodes sur deux corpus de genres différents mais portant sur le même domaine. Nous observons que 8 à 30 % des termes se retrouvent sous une forme plus ou moins fidèle, suivant la taille de la terminologie considérée, et que 83 % des structures hiérarchiques comparables sont compatibles.

**Mots-clés :** Ontologie, ingénierie des connaissances, traitement automatique des langues, analyse distributionnelle, patrons lexico-syntaxiques, sémantique différentielle, corpus textuel.

---

## **1. Introduction**

Le développement de ressources ontologiques facilitant l'usage des terminologies nationales et internationales dans le domaine médical est un enjeu significatif pour le recueil d'information et pour l'accès aux connaissances médicales (base de données bibliographique Medline, base de connaissances Vidal sur les médicaments, etc.). Une ontologie décrit de manière générique les connaissances propres à un domaine donné et offre de celui-ci une compréhension consensuelle. Il s'agit d'une modélisation conceptuelle, non contextuelle et non ambiguë. Dans le cas de l'ontologie différentielle (Bachimont, 2000), le sens de chaque concept est déterminé par sa position dans une hiérarchie, fondée sur une structuration terminologique. Elle se fait suivant des principes différentiels de similarité et différence entre unités terminologiques proches.

Une méthode éprouvée pour la construction d'ontologie repose sur l'exploitation de corpus

appropriés. Elle consiste à construire des ressources termino-ontologiques par analyse distributionnelle (Habert *et al.*, 1996) ou par application de patrons lexico-syntaxiques (Hearst, 1992). Quelques travaux proposent également une approche mixte : Caraballo (1999) regroupe des termes par analyse distributionnelle et type les relations sémantiques qui les lient au moyen de patrons lexico-syntaxiques lorsque c'est possible, alors que Cimiano *et al.* (2004) comparent les structurations issues de l'application d'une analyse distributionnelle avec celles construites au moyen de patrons lexico-syntaxiques sur deux ensembles de corpus. L'expérimentation que nous présentons se place dans cette lignée mais rencontre des problèmes liés à l'adéquation de ces deux méthodes par rapport au genre des corpus à traiter. Notre définition des genres textuels est issue des travaux menés par l'Action Spécifique STIC « Corpus et Terminologie » (Aussenac-Gilles & Condamines, 2003). 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 : dossier patient, enseignement, ressources, publications et oral (*Op. Cit.*).

Nous souhaitons modéliser une ontologie du domaine de la pneumologie<sup>1</sup>, pour laquelle nous disposons de deux corpus : un ensemble de comptes rendus d'hospitalisation (corpus [CRH]) et, en tant que ressource additionnelle, un livre de cours au format numérique<sup>2</sup>(corpus [LIVRE]) : un abrégé de Pneumologie publié chez Masson (Housset, 1999). Dans le cadre de la typologie évoquée précédemment, notre corpus [CRH] fait partie du genre textuel dit « dossier patient » et notre corpus [LIVRE] de la catégorie « enseignement ». L'objectif de cet article est d'étudier dans quelle mesure les informations contenues dans le livre de cours peuvent être mobilisées pour compléter celles extraites à partir des comptes rendus d'hospitalisation. Nous nous sommes limités à la question de la structuration hiérarchique de l'ontologie et nous cherchons à déterminer si les hiérarchies créées à partir des deux corpus sont comparables au niveau terminologique et compatibles au niveau de la structure résultante.

Ce travail de conception d'ontologies s'inscrit dans le cadre du projet de recherche PER-TOMed (Production et évaluation de ressources terminologiques et ontologiques dans le domaine de la médecine, <http://www.spim.jussieu.fr/Pertomed>), 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 et ontologiques (RTO) dans le domaine médical. Faciliter l'accès aux connaissances médicales est, en effet, un enjeu important pour les professions de santé comme pour le grand public. Les ontologies développées dans le cadre du projet sont construites en collaboration avec des groupes d'utilisateurs chargés d'évaluer 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 des actes et des diagnostics *via* une représentation des connaissances médicales reposant sur le modèle conceptuel d'une ontologie du domaine.

Nous proposons, dans cet article, deux expérimentations en vue de construire une ontologie de la pneumologie dont la structure hiérarchique est organisée selon des principes issus de la sémantique différentielle. Notre première expérimentation consiste à traiter nos deux corpus, d'une part, par l'analyse distributionnelle et, d'autre part, par l'application de patrons lexico-syntaxiques. Notre seconde expérimentation mesure la compatibilité des deux méthodes et l'apport complémentaire de l'une par rapport à l'autre.

---

<sup>1</sup>La pneumologie étant une des spécialités reconnues de la médecine, nous la considérons comme un domaine de connaissances particulier.

<sup>2</sup>Ce livre de cours est rédigé par des pneumologues et destiné à des étudiants en médecine.

Nous commençons par présenter les corpus (section 2). Nous détaillons notre première expérimentation (section 3) : les deux méthodologies mises en œuvre — l'analyse distributionnelle (section 3.1) et l'analyse par patrons lexico-syntaxiques (section 3.3) — et les résultats de cette expérimentation (section 3.2 et section 3.4). Nous décrivons ensuite la seconde expérimentation (section 4) : la procédure de comparaison des deux terminologies modélisées (section 4.1) et les résultats obtenus (section 4.2 et section 4.3). Nous discutons enfin (section 5) de l'intérêt d'un tel travail et concluons cet article par les perspectives qu'il offre.

## **2. Présentation des deux corpus**

La construction d'ontologie à partir de corpus passe par la constitution de corpus textuels sur lesquels appliquer des outils d'analyse du langage naturel. Dans le but de couvrir, avec le plus d'exhaustivité possible, le vocabulaire de référence relatif aux besoins terminologiques des pneumologues en matière d'aide au codage, nous avons collecté des comptes rendus d'hospitalisation (corpus intitulé [CRH]) dans six hôpitaux de l'Assistance Publique-Hôpitaux de Paris<sup>3</sup>. Ce corpus textuel témoigne 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. Au total, nous disposons de 1 038 CRH, sachant qu'il a été établi dans (Le Moigno *et al.*, 2002) que 600 comptes rendus (environ 350 000 mots) semble être un ordre de grandeur satisfaisant pour obtenir de bons résultats. Ce premier corpus [CRH] compte environ 417 000 mots. Le second corpus, intitulé [LIVRE], est construit d'après un ouvrage pédagogique et comprend environ à 823 000 mots. Ces corpus nous parviennent sous des formats inexploitable par les outils d'analyse du langage. Ils sont donc traités<sup>4</sup> 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.

## **3. Analyse distributionnelle et repérage par patrons lexico-syntaxiques**

Nous présentons tour à tour les principes de l'analyse distributionnelle (section 3.1), ceux de l'analyse par patrons lexico-syntaxiques (section 3.3), et leur mise en œuvre sur nos deux corpus (sections 3.2 et 3.4). Nous allons voir que l'adéquation de ces méthodologies varie avec le genre et le type du corpus. Bien que ce constat ait déjà été présenté dans des travaux antérieurs (Condamines, 2003), nous pensons apporter dans cet article des précisions quant à la possible complémentarité des résultats.

### **3.1. Analyse distributionnelle**

L'analyse distributionnelle est un type d'exploration de corpus fondé sur les principes de Harris (1968), et mis en œuvre notamment dans le logiciel LEXICLASS (Assadi & Bourigault, 2000) d'après les résultats fournis par le logiciel LEXTER (Bourigault, 1994), puis dans SYNTAX-UPERY (Bourigault, 2002). Etant donné les relations de dépendances syntaxiques entre mots (ici noms, verbes et adjectifs) dans chaque énoncé d'un corpus, cette analyse pose que les mots qui ont un sens proche se caractérisent par des dépendances similaires. La mise en œuvre de cette méthode se découpe donc classiquement en deux parties : collecte des dépendances syn-

---

<sup>3</sup>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.

<sup>4</sup>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.

taxiques entre mots ou syntagmes, puis analyse de leur distribution. Dans le cadre de notre expérimentation, les dépendances syntaxiques sont calculées par le module SYNTAX ; les syntagmes proposés par SYNTAX sont appelés candidats termes et sont composés d'une tête et d'une expansion. Par exemple, dans le syntagme nominal *Opacité dans le poumon gauche*, le terme *opacité* est la tête du syntagme, c'est-à-dire son recteur, et *dans le poumon gauche* est son expansion. L'analyse de leurs distributions est effectuée par le module UPERY. Cette analyse calcule des proximités distributionnelles entre les candidats termes : elle rapproche deux à deux les candidats termes qui partagent les mêmes dépendances (*i.e.* leurs expansions). Ils sont appelés « descendants en Tête et en Expansion » dans l'outil. Comme cette analyse est symétrique, elle rapproche également les dépendants en fonction des recteurs qu'ils partagent (« voisins en Tête et en Expansion »). Les résultats de l'analyse sont visualisables dans TERMONT, l'interface d'accès aux données du logiciel SYNTAX-UPERY. Il faut noter que la plupart des logiciels d'analyse distributionnelle s'appuient sur une approximation des dépendances syntaxiques : les cooccurrences entre termes (Curran & Moens, 2002). Par contraste, SYNTAX est un analyseur syntaxique, ce qui permet de travailler effectivement avec des dépendances syntaxiques.

### 3.2. Résultats de l'analyse distributionnelle

Les deux corpus [CRH] et [LIVRE] sont traités par SYNTAX-UPERY. Le corpus [CRH] produit 36 881 syntagmes nominaux et le corpus [LIVRE] en produit 17 666. Après étude, l'analyse distributionnelle ne donne pas de résultats satisfaisants sur le corpus [LIVRE] :

1. Les termes les plus fréquents extraits par SYNTAX-UPERY ne sont pas pertinents pour construire la hiérarchie des concepts primitifs, essentiels à la représentation du domaine. Par exemple le candidat terme *rapport de vraisemblance* a la plus forte fréquence d'apparition (177) dans le corpus. Or, il est sémantiquement pauvre pour le domaine de la pneumologie, il n'est donc pas caractéristique et ne sera pas normalisé.
2. Par ailleurs, le nombre de voisins en Tête et en Expansion est faible : les candidats termes sont souvent sémantiquement éloignés car le corpus est faiblement redondant. Il est donc difficile pour l'ingénieur de la connaissance de savoir où les placer dans la hiérarchie ontologique sur la base de leur distribution.

Enfin, nous souhaitons construire une ontologie pour l'aide au codage. Dans cette optique, il est important de mettre à disposition du pneumologue un vocabulaire qu'il emploie couramment. Le corpus [LIVRE] est alors moins intéressant car les connaissances qu'il contient sont destinées à une personne non spécialiste du domaine et exprimées de manière pédagogique. Au contraire, les données du corpus [CRH] sont exprimées par des pneumologues dans leur vocabulaire métier et sont donc plus représentatives.

Concernant l'analyse effectuée sur le corpus [CRH], nous décidons de concentrer notre attention, dans un premier temps, sur les 679 syntagmes nominaux dont le nombre d'occurrences en corpus dépasse 12. L'analyse syntaxique permet de dégager trois grands axes particulièrement pertinents : les pathologies, les signes et les traitements/examens. Nous associons des indices numériques compris entre 4 et 6, sur un intervalle allant de 1 à 6, aux syntagmes sémantiquement proches de l'un de ces trois axes. Cet indice est utilisé comme filtre, il est ainsi facile d'écartier ceux qui ne semblent, pour l'instant, pas pertinents (indice fixé à 1). Ce regroupement permet de commencer une première phase de travail sur le rapprochement par contexte et laisse 370 syntagmes nominaux sur lesquels élaborer le cœur de notre ontologie. D'après les résultats de l'analyse distributionnelle, le candidat terme *cure 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 occurrences. Ses voisins en Tête sont : [*Hospitalisation, Examen, Navelbine, Cisplatine, Doxorubicine, Taxotere, Carboplatine, MIP*]. Ses voisins en Expansion sont : [*Traitement, Bilan, Antibiothérapie, Injection, Radiothérapie*]. Ces regroupements par contextes sémantiquement proches permettent de structurer les axes horizontaux (relation frère-frère) et verticaux (relation père-fils) de l'ontologie. Pour cet exemple, nous avons construit la représentation suivante : *ActionMedicale / Traitement / TraitementMedicamenteux / Chimiotherapie*. La chimiothérapie étant considérée comme un traitement médicamenteux, nous retrouvons les médicaments associés (Navelbine, Cisplatine, Doxorubicine, Taxotere, Carboplatine) classés sous *Medicament / Cancerologie*. Nous déduisons du regroupement des voisins en Expansion que les candidats termes *Antibiothérapie* et *Radiothérapie* sont également et potentiellement à placer sous *Traitement*.

Cette méthode de regroupement donne de bons résultats et rend la tâche bien plus facile pour un ingénieur de la connaissance. L'analyse distributionnelle sur le corpus [CRH] donne ainsi des candidats termes pertinents pour la construction de l'ontologie, saisie dans l'éditeur d'ontologie DOE (Troncy & Isaac, 2002) développé à l'INA. Notre ontologie de la pneumologie contient à ce jour 370 concepts primitifs, issus de la première analyse des candidats termes. Les phases de construction étant itératives, nous augmentons très rapidement la représentation en étudiant les candidats termes dont la fréquence d'apparition dans le corpus [CRH] est inférieure à 12. Nous sommes également en train d'améliorer l'ontologie en analysant les termes présents dans les sections concernant la pneumologie des thésaurus CIM-10 et CCAM<sup>5</sup>. De même, nous prévoyons de compléter la hiérarchie en faisant le lien avec le haut de l'ontologie du projet MENELAS. Nous envisageons d'obtenir rapidement une hiérarchie comptant environ 1 200 concepts.

### 3.3. Repérage par patrons lexico-syntaxiques

La deuxième méthodologie est fondée sur la définition *a priori* d'une relation sémantique, par exemple l'hyponymie, puis sur l'observation de séquences en corpus qui véhiculent la relation souhaitée. Cette observation permet de schématiser le contexte lexical et syntaxique des unités lexicales en relation et de construire une synthèse de ce contexte sous la forme d'un patron lexico-syntaxique. Le patron est ensuite comparé aux occurrences en corpus et permet d'en extraire d'autres couples d'unités lexicales correspondant au motif spécifié. L'hypothèse est alors que ces nouvelles unités lexicales sont liées par la relation sémantique souhaitée. Les patrons lexico-syntaxiques se fondent sur un marqueur, ou pivot (une unité linguistique qui peut être un indice d'une relation lexicale, comme *entre autres* pour la relation d'hyponymie) et un ensemble de contraintes que le contexte lexical ou syntaxique de ce pivot doit remplir. Par exemple, dans le cas de l'hyponymie et du marqueur *entre autres*, il faut que la forme syntaxique corresponde au patron : *DET SN, entre autres SN*. Ce patron permet d'extraire une phrase contenant *Les méningites, entre autres pathologies...*, et de mettre en relation *méningites* et *pathologies*. Cette méthodologie a été présentée dans (Hearst, 1992) et mise en œuvre notamment dans (Morin, 1999) et (Séguéla, 2001). Les patrons lexico-syntaxiques liés à l'hyponymie mettent en relation des couples père-fils potentiels qui sont intéressants pour la structuration hiérarchique d'une ontologie. Dans le cadre spécifique de la construction d'ontologies différentielles, nous appliquons cette méthodologie à la recherche d'énoncés définitoires en corpus. Nous nous appuyons pour cela sur les travaux de (Rebeyrolle, 2000). Ces énoncés définitoires sont ensuite mobilisés dans des traitements visant à donner des pistes terminologiques pour la

---

<sup>5</sup>La CIM-10 est la classification internationale des maladies et la CCAM est la classification commune des actes médicaux. Ces deux thésaurus servent au codage médico-économique des patients.

construction des principes différentiels.

### 3.4. Résultats du repérage par patrons lexico-syntaxiques

Nous appliquons sur le corpus [CRH] les patrons lexico-syntaxiques de recherche d'énoncés définitoires développés dans (Malaisé *et al.*, 2004). Le genre textuel n'étant pas adapté à la reformulation ou à l'explicitation du sens des unités lexicales (les textes sont destinés à des personnes de même degré de compétence, et traitent de leur domaine de connaissance, pouvant donc se baser sur tout leur « passif terminologique commun »), nos programmes n'ont extrait que 31 phrases (ou ensembles de phrases) correspondant effectivement à des énoncés définitoires<sup>6</sup>, sur un total de 199 extractions<sup>7</sup>. Il s'agit d'un résultat trop limité pour que cette méthodologie présente un réel intérêt sur ce corpus précis. Les principales erreurs sont les suivantes :

- Concernant les énoncés extraits autour du marqueur de la parenthèse, le patron N(N) supposé renvoyer l'hyperonyme du nom précédant la parenthèse est à l'origine de beaucoup de bruit. En effet, suite à un étiquetage par défaut de notre outil, des énoncés correspondant aux schémas suivants ont été renvoyés : N(HÔPITAL), Dr X (SPÉCIALITÉ), ... ;
- Concernant ceux extraits sur la base de marqueurs métalinguistiques (comme *expression* ou le verbe *définir*), les erreurs sont liées au genre des CRH, comprenant des passages comme *l'expression de mes salutations distinguées*, ou au domaine médical : *définir les modalités d'une opération*, ... ;
- Enfin, concernant les énoncés extraits à partir de marqueurs linguistiques plus génériques (*il s'agit de, indiquer,...*), nous remarquons trois grands types d'erreurs. Tout d'abord, certains patrons associent un diagnostic à une pathologie, association qui est intéressante au niveau de la modélisation du domaine mais qui n'est pas directement définitoire. Ensuite, la structure même des CRH a donné lieu à des extractions erronées car ils associent à un titre de paragraphe (comme *Evolution*) la description d'un patient, en commençant la phrase par *Il s'agit de...* . Nous avons ici un problème de rattachement sémantique, la mention *Il s'agit de* ne se rapportant pas à *l'évolution*. En revanche, dans le corpus [LIVRE], ce patron permet d'associer aux titres de section leurs descriptifs commençant par ce même marqueur. Enfin, le troisième type d'erreurs rencontré rejoint le comportement que nous avons déjà observé dans (Malaisé *et al.*, 2004) sur un corpus de diététique : il semblerait que le marqueur *indiquer* ne soit pas pertinent ou demande des contraintes spécifiques dans le domaine médical.

L'analyse des résultats soulève qu'il est, d'une part, toujours problématique de contraindre des patrons lexico-syntaxiques de peur d'induire du silence informationnel, et, d'autre part, que le fonctionnement de certains patrons est fortement lié à des différences de genres textuels.

Nous appliquons ensuite ces patrons au corpus [LIVRE]. C'est un corpus d'enseignement, un genre textuel particulièrement propice à la découverte d'énoncés définitoires. Nos programmes ont extrait 799 phrases ou groupes de phrases, nous en avons validé 119<sup>8</sup>.

Nous suivons la méthode de (Malaisé *et al.*, 2004) pour exploiter ces énoncés définitoires. Les groupes extraits sont présentés à l'ingénieur de la connaissance dans une interface HTML :

---

<sup>6</sup>Parmi ces énoncés, 5 correspondent également à des paradigmes, relation que nous avons jugée intéressante dans la mesure où elle permet de proposer des « candidats co-hyponymes ».

<sup>7</sup>Pour un aperçu plus détaillé de la dépendance entre genre textuel et patrons lexico-syntaxiques, voir les travaux de (Condamines, 2003)

<sup>8</sup>Ce qui représente une précision de 15 %.

Type	Nb	Exemple
Termes identiques, à la normalisation terminologique près	3	<i>Asthme</i> [CRH] vs <i>asthme</i> [LIVRE], <i>Emphyseme</i> [CRH] vs <i>emphysème</i> [LIVRE]
Variantes lexicales comparables	3	<i>SaturationEnAir</i> [CRH] vs <i>saturation en oxygène</i> [LIVRE]
Niveaux de granularité différents	18	<i>Adenopathie</i> [CRH] vs <i>Adénopathies médiastinales</i> [LIVRE]

TAB. 1 – Comparaison des termes des deux hiérarchies terminologiques.

un formulaire où il est possible de les modifier, de valider les relations sémantiques<sup>9</sup> et les énoncés pertinents pour la construction d'ontologie. Les données sont ensuite insérées dans une base de données MySQL avec un export au format d'ontologie OWL préconisé par le consortium W3C. Les hiérarchies créées sont visualisables dans un éditeur d'ontologie comme DOE.

## 4. Comparaison des deux méthodes

### 4.1. Procédure de comparaison

Nous comparons les terminologies structurées suivant les deux méthodes précédentes (analyse distributionnelle sur le corpus [CRH] et patrons sur le corpus [LIVRE]) pour voir dans quelle mesure elles sont compatibles ou complémentaires. Pour cela, nous comparons tout d'abord manuellement les 370 termes de la future ontologie (arborescence [CRH]) avec la structuration à un niveau de profondeur issue de la validation des extractions d'énoncés défini-toires (arborescence [LIVRE]), comprenant 119 candidats termes ou groupes syntaxiques plus larges. Nous trouvons 24 ensembles de termes comparables<sup>10</sup> à différents titres. Ces variantes sont détaillées au tableau 1.

Nous comparons ensuite les structures autour de ces termes communs ou similaires. Là encore, il y a plusieurs cas de figure : dans les cas où la structuration terminologique liée à la validation du formulaire correspond à une hiérarchie (c'est-à-dire lorsque l'énoncé défini-toire ne met pas en relation deux « synonymes » au sens large), nous observons que les deux structures peuvent être identiques, complémentaires ou divergentes (voir tableau 2).

Les derniers exemples de ce tableau 2, concernant les hiérarchies divergentes, montrent l'intérêt d'avoir deux hiérarchies à confronter pour pré-valider les données issues d'extraction à partir de corpus avant de les proposer aux experts. Dans la majorité des cas, les hiérarchies sont soit équivalentes, soit complémentaires (seulement quatre contre-exemples sur 24 termes, incluant deux cas où la hiérarchie [LIVRE] n'est pas juste : elle associe un terme à ses caractéristiques et non pas à son hyperonyme). Nous constatons que les termes potentiels hiérarchisés à partir du corpus [LIVRE] sont souvent plus spécifiques que les candidats termes issus du corpus [CRH]. Cette différence peut être due au fait que les termes de base ne sont pas définis dans le livre de cours à l'origine du corpus [LIVRE], car ils correspondent à des notions supposées acquises. Elle peut également être due au mode d'extraction de ces termes et marquer une des spécificités de l'extraction par patrons lexico-syntaxiques.

<sup>9</sup>Il s'agit d'un pluriel car nous ne pensons pas, d'une manière générale et en dehors du cadre précis de cet article, que la seule relation sémantique pertinente pour la construction d'ontologie soit l'hyponymie.

<sup>10</sup>Il s'agit de termes ou d'ensembles de termes comparables. Par exemple, au concept de *Tumeur* de l'ontologie [CRH] correspond un ensemble de différentes tumeurs dans la terminologie [LIVRE].

Type	Exemple	Commentaire
Identique ou comparable	<i>Broncho pneumopathie / Asthme [CRH] vs Bronchopathie / Asthme à dyspnée continue [LIVRE]</i>	Pour comparer ces deux hiérarchies, nous avons regardé comment ces trois notions étaient organisées dans le MeSH. Les deux premières sont classifiées sous <i>Poumon, maladie</i> , alors qu' <i>Asthme</i> est une notion plus spécifique dans la même branche hiérarchique. Ce qui tend à valider la cohérence et la compatibilité des deux hiérarchies terminologiques trouvées.
Complémentaire	<i>Signe / [...] / Signe-Respiratoire / Insuffisance-Ventriculaire [CRH] vs Signe / Insuffisance ventriculaire droite [LIVRE]</i>	La deuxième arborescence vient confirmer la première et permet de la compléter d'un niveau, celui de <i>Insuffisance ventriculaire droite</i> .
Divergente	<i>EtatPathologique / Maladie-Respiratoire / Bronchite ET Signe / Toux [CRH] vs Toux avec expectoration / Bronchite chronique [LIVRE]</i>  <i>EtatMorphologique / AnomalieMorphologique / Lésion / Atelectasie – Adénopathie [CRH] vs Opacité médiastinale / Adénopathies médiastinales ET Opacité dense arrondie / Atélectasie par enroulement... [LIVRE]</i>	Dans le MeSH, la toux est classifiée à la fois comme <i>signe-symptôme</i> et comme <i>pathologie</i> : les deux sources textuelles illustrent chacune un de ces aspects. Nous avons choisi de privilégier le point de vue abordé dans le corpus [CRH].  Dans les deux cas, <i>Atélectasie</i> et <i>Adénopathie</i> sont co-hyponymes, mais leurs hyperonymes sont contradictoires : dans l'arborescence [CRH], les <i>Opacités</i> sont classifiées sous <i>Signes</i> . Il s'agit d'une erreur d'interprétation de la relation sémantique dans les extractions à partir du corpus [LIVRE]. Une <i>Opacité</i> est bien un élément d'une image médicale qui est interprétée comme le <i>signe</i> d'une <i>adénopathie</i> . L'extrait du corpus était toutefois ambigu : <i>Adénopathies médiastinales. Il s'agit des opacités médiastinales les plus fréquentes...</i>

TAB. 2 – Comparaison des deux structures terminologiques.

Enfin, nos deux dernières expérimentations concernent les « termes propres » des deux structures terminologiques : nous cherchons à comprendre pourquoi certains termes ne se retrouvent pas dans les deux arborescences.

#### 4.2. Comparaison des termes du corpus [LIVRE]

Une comparaison automatisée des deux hiérarchies nous permet d'isoler les termes qui n'apparaissent que dans la structure terminologique issue du corpus [LIVRE]. Nous cherchons à comprendre pourquoi ces termes ne se retrouvent pas dans l'ontologie [CRH]. Nous étudions, pour cela la présence d'une ou plusieurs occurrences dans le corpus [CRH], auquel cas l'analyse distributionnelle aurait dû les isoler. Il y a 83<sup>11</sup> termes « propres » à la terminologie construite à partir du corpus [LIVRE], et les résultats de l'observation de leurs occurrences dans le corpus [CRH] sont présentés dans le tableau 3.

Les termes qui ne figurent pas dans le corpus [CRH] ne peuvent pas être proposés comme candidats termes par l'analyse distributionnelle. Ceux qui sont présents sous la même forme en

<sup>11</sup>Il y a 83 et non 95 termes « propres » (119 termes - 24 termes communs) dans cette terminologie parce que le décompte des 24 termes communs regroupe des termes composés comparables, comme nous l'avons vu plus haut.

Type	Nb
Termes non présents dans le corpus [CRH], ou sous une forme non identifiée	20
Termes présents sous la même forme dans [CRH] mais à faible nombre d'occurrences	15
Termes correspondant à une même racine dans [CRH] ( <i>spondylarthrite</i> [LIVRE] vs <i>spondylarthropathie</i> [CRH]), ou présence de la tête du syntagme complexe ( <i>mésothéliome malin diffus</i> [LIVRE] vs <i>mésothéliome</i> ou <i>mésothéliome pleural</i> [CRH])	42
Termes correspondant à des énoncés définitoires, validés comme paraphrases synonymiques ou par erreur en tant qu'hyperonymes, et n'ayant donc aucun équivalent terminologique dans [CRH]	6

TAB. 3 – Comparaison des termes propres à la terminologie construite à partir du corpus [LIVRE] avec le corpus [CRH].

Type	Nb
Termes définis, classifiés ou caractérisés, plus ou moins précisément dans le corpus [LIVRE]	68
Termes sans occurrence dans le corpus [LIVRE]	60
Termes non définis ou caractérisés dans le corpus [LIVRE]	49
Termes de haut niveau (comme <i>Inanime</i> , <i>SigneFonctionnel</i> , ...), ne correspondant pas forcément à une occurrence dans le corpus [CRH]	48
Termes exprimant une caractéristique : des qualificatifs ( <i>Gauche</i> , <i>Positif</i> , ...)	21

TAB. 4 – Comparaison des termes propres à l'ontologie basée sur les corpus [CRH] et [LIVRE].

corpus n'ont pas un nombre d'occurrences supérieur ou égal à 12 et ne sont donc pas encore pris en compte dans notre analyse des résultats de SYNTAX. Enfin, la moitié des termes extraits par les patrons lexico-syntaxiques ont des « homologues » ou termes proches dans le corpus [CRH]. Pour les rapprocher, il faut disposer de connaissances morphologiques sur la dérivation et la composition impliquées dans ces termes (Namer & Zweigenbaum, 2004).

#### 4.3. Comparaison des termes de l'ontologie basée sur le corpus [CRH]

Nous extrayons ensuite les termes qui ne sont présents que dans l'arborescence [CRH], et regardons s'ils sont définis ou caractérisés dans le corpus [LIVRE]. Cette comparaison étant manuelle, nous passons outre la normalisation terminologique des termes [CRH] (majuscule initiale et désaccentuation) et vérifions également la présence d'éventuelles variantes terminologiques. Le résultat de la comparaison est détaillé dans le tableau 4.

Soixante-huit termes correspondent à des énoncés pouvant être interprétés comme définitoires ; nous voulons comprendre pourquoi nos patrons lexico-syntaxiques ne les ont pas renvoyés. L'analyse de ces énoncés nous donne plusieurs explications quantifiées dans le tableau 5.

Nous détaillons ensuite l'analyse des énoncés à intérêt définitoire semblant pouvoir être extraits au moyen de patrons lexico-syntaxiques (sachant que les deux autres types de contextes ne peuvent pas être trouvés par ce genre de méthode, mais demandent plutôt des solutions apparentées à la résolution d'anaphore), afin de savoir s'il faut augmenter notre système de nouveaux patrons, en adapter certains ou relâcher des contraintes (tableau 6). Une partie des 47 énoncés analysés sont extraits par notre système (23), mais ne sont pas validés lors de l'analyse des réponses. En effet, ils donnent une définition partielle ou associent un terme à un hyperonyme inattendu, jugé non pertinent lors de la validation des formulaires. Les autres cas de figure nous donnent des pistes pour compléter les patrons existants.

Type	Nb d'énoncés
Contextes sur plusieurs phrases pouvant être interprétés comme donnant des éléments de définition, mais étant relativement vagues	11
Contextes étant des définitions plus claires, mais ne pouvant pas être retrouvés au moyen de patrons lexico-syntaxiques	9
Contextes étant des définitions et pouvant, ou semblant pouvoir, être extraits au moyen de patrons lexico-syntaxiques	47

TAB. 5 – Évaluation des énoncés à intérêt définitoire ou assimilés du corpus [LIVRE] non renvoyés par nos patrons.

Type de patron	Exemple
Des patrons envisagés mais pas encore implémentés	Liste, virgule, double points
Des patrons classiques pas encore implémentés	de NOM tel que le NOM, des NOM et d'autres NOM
Des patrons implémentés, mais pour lesquels il faudrait pousser l'analyse, ou voir si des modifications sont envisageables	parenthèse, ou/et
Des patrons correspondant à la méronymie	comportent, consistent en
Des patrons correspondant à des relations spécifiques à la médecine	Le traitement des formes cryptogéniques <i>repose sur</i> la corticothérapie...

TAB. 6 – Évaluation des énoncés à intérêt définitoire pouvant être renvoyés par des patrons lexico-syntaxiques.

## 5. Conclusion et perspectives

Nous avons présenté un ensemble de traitements pour la construction de hiérarchies terminologiques fondées sur deux méthodologies TALN, adaptées chacune à un type et genre de corpus : l'analyse distributionnelle sur un corpus redondant et riche en termes spécialisés, et l'extraction par patrons lexico-syntaxiques sur un corpus didactique à la structure régulière. Nos résultats ont montré : *a*) l'intérêt de chacune des méthodes en fonction du type et du genre de corpus, *b*) la nature des connaissances pouvant être identifiées par chaque approche, et *c*) la complémentarité et la différence des résultats de chacune. Bien que les traitements aient porté sur des corpus différents, il est intéressant d'observer la relative compatibilité entre les deux ensembles terminologiques extraits et leur apport mutuel. Comme nous l'avons signalé dans la section 3, la question des genres textuels n'est pas nouvelle. Cependant, nous pensons que l'approche comparative de ce travail est plus rarement exploitée.

La divergence des structures terminologiques obtenues est également un point intéressant car elle dénote la possibilité d'organisations conceptuelles différentes au sein du domaine considéré, connaissance précieuse pour un ingénieur cognitif. Il est reconnu depuis quelques années en Traitement automatique des langues qu'il est utopique, voire inutile, de chercher à modéliser une ontologie unique du domaine. En effet, bien qu'une ontologie soit une modélisation conceptuelle, non contextuelle et non ambiguë, nous sommes convaincus qu'il existe de nombreuses modélisations possibles pour un domaine donné, en fonction de la tâche à réaliser. Ainsi, le travail de structuration ontologique vise à expliciter les choix faits parmi l'ensemble des modélisations potentielles. Proposer à un ingénieur cognitif, non spécialiste du domaine, des vues complémentaires ou divergentes lui donne des arguments critiques pour faire ses choix

et les valider.

Nous avons toutefois cerné certaines limites liées à cette comparaison : un rapprochement plus automatique de ces deux ensembles terminologiques nécessiterait, d'une part, de mettre en œuvre des techniques plus sophistiquées d'appariement, et, d'autre part, d'améliorer la précision (et probablement le rappel) des patrons lexico-syntaxiques, ce qui implique leur adaptation aux spécificités du domaine médical. Toutefois, ces mêmes patrons permettent déjà de repérer d'autres relations propres au domaine médical, qu'il serait alors intéressant d'isoler et de caractériser plus précisément. Il serait également pertinent de comparer les hiérarchies non redondantes entre les deux structures modélisées avec une terminologie ou un thésaurus de référence comme le MeSH, pour vérifier leur cohérence et validité propres, ou, le cas échéant, proposer des suggestions de compléments au MeSH.

## Références

- ASSADI H. & BOURIGAULT D. (2000). Analyses syntaxique et statistique pour la construction d'ontologies à partir de textes. In J. CHARLET, M. ZACKLAD, G. KASSEL & D. BOURIGAULT, Eds., *Ingénierie des connaissances - Evolutions récentes et nouveaux défis*. Eyrolles.
- 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, disponible sur le site <http://rtp-doc.enssib.fr/archiveas.html>.
- 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 - Evolutions récentes et nouveaux défis*, p. 305–336. Paris: Eyrolles.
- BOURIGAULT D. (1994). Extraction et structuration automatiques de terminologie pour l'aide à l'acquisition de connaissances Ō partir de textes. In *Actes du 9<sup>e</sup> congrès Reconnaissance des Formes et Intelligence Artificielle - AFCET*, Paris.
- BOURIGAULT D. (2002). Analyse distributionnelle étendue. In *Actes de la 9<sup>e</sup> conférence sur le traitement automatique des langues*, Nancy.
- 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.
- CIMIANO P., PIVK A., SCHMIDT-THIEME L. & STAAB S. (2004). Learning taxonomic relations from heterogenous evidence. In *ECAI-2004 Workshop on Ontology Learning and Population*, Valencia, Spain.
- CONDAMINES A. (2003). *Sémantique et corpus spécialisé : constitution de bases de connaissances terminologiques*. Habilitation à diriger des recherches, Université de Toulouse Le Mirail.
- CURRAN J. R. & MOENS M. (2002). Scaling context space. In *Proc 38<sup>th</sup> ACL*, p. 231–238, USA.
- HABERT B., NAULLEAU E. & NAZARENKO A. (1996). Symbolic word clustering for medium-size corpora. In J.-I. TSUJII, Ed., *Proceedings of the 16<sup>th</sup> COLING*, p. 490–495, Copenhagen, Denmark.
- HARRIS Z. (1968). *Mathematical Structures of Language*. New-York: John Wiley and Sons.
- HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In A. ZAMPOLLI, Ed., *Proceedings of the 14<sup>th</sup> COLING*, p. 539–545, Nantes, France.
- HOUSSET B. (1999). *Abrégé de Pneumologie*. Abrégés. Paris: Masson.
- 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<sup>es</sup> 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.

MORIN E. (1999). Des patrons lexico-syntaxiques pour aider au dépouillement terminologique. *Traitement Automatique des Langues*, **40(1)**, 143–166.

NAMER F. & ZWEIGENBAUM P. (2004). Acquiring meaning for French medical terminology: contribution of morphosemantics. In M. FIESCHI, E. COIERA & Y.-C. J. LI, Eds., *Actes 10<sup>th</sup> World Congress on Medical Informatics*, p. 535–539, San Francisco, Ca.

REBEYROLLE J. (2000). *Forme et fonction de la définition en discours*. Université de Toulouse II - Le Mirail: Thèse de doctorat.

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

TRONCY R. & ISAAC A. (2002). DOE : une mise en oeuvre d'une méthode de structuration différentielle pour les ontologies. In *13<sup>th</sup> Journées Francophones d'Ingénierie des Connaissances (IC'02)*, p. 63–74, Rouen, France.