Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles à partir de corpus textuels

Véronique Malaisé

UINVERSITE PARIS 7 - DENIS DIDEROT **UFR de Linguistique**

Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles à partir de corpus textuels

THÈSE

présentée et soutenue publiquement le 19 octobre 2005

pour l'obtention du

Doctorat de l'Université Paris 7 – Denis Diderot

(Discipline : Linguistique)

par

Véronique Malaisé

Composition du jury

Rapporteurs : Anne Condamines
Adeline Nazarenko

Examinateurs : Pierre Zweigenbaum
Bruno Bachimont
Laurence Danlos
Monique Slodzian

Institut National de l'Audiovisuel – INSERM U729

Mis en page avec la classe thloria.

Remerciements

Je tiens à remercier ici les personnes qui ont contribué, à divers titres, à l'achèvement de cette thèse et à rendre cette recherche intéressante. Tous d'abord, je souhaiterais remercier Laurence Danlos pour avoir présidé le jury de cette thèse, Anne Condamines et Adeline Nazarenko pour leur travail de rapporteur sur ce mémoire.

Je remercie tout particulièrement Pierre Zweigenbaum et Bruno Bachimont pour leur encadrement de qualité et les discussions fructueuses qu'ils m'ont offertes tout au long de cette recherche. Les personnes qui m'ont encadrées et mes collègues, à l'INA comme au STIM et à l'INSERM, m'ont permis d'avoir un environnement de travail stimulant, je tiens à remercier Daniel Teruggi, Philippe Poncin, Vincent Brunie ainsi que les membres de l'équipe *Description des Contenus Audiovisuels*, à la Direction de la Recherche et Experimentation de l'INA, ainsi que Marie-Christine Jaulent, Jean Charlet et mes autres collègues du STIM et de l'INSERM. Une pensée particulière va à Antoine Isaac et Raphaël Troncy, pour m'avoir accueillie dans leur bureau et fait partager leur expérience, leurs informations et leur cafetière, à Younès Hafri, Thomas Drugeon et Jean-Philippe Poli pour leurs conseils et leur aide, à l'INA, et à Natalia Grabar, Audrey Baneyx et Gersende Georg pour leur amitié, au STIM et à l'INSERM.

Un grand merci, enfin, à mes autres amis et amies, à ma famille, pour leur soutien, et, le mot de la fin : merci à Paul dont j'ai mis la patience (pourtant légendaire) à rude épreuve tout au long de ces années.

Table des matières

Partie II Expérimentations et mise au point de la méthodologie

Introduction générale

Il existe, selon le domaine de spécialité considéré, différents points de vue sur ce qu'est une ontologie, et, à bien y regarder, on s'aperçoit que tous les objets désignés par ce terme ne sont pas les mêmes. Nous nous intéressons dans cette thèse aux ontologies différentielles. Comme nous le verrons plus en détail au chapitre 2 (section 2.2.2), il s'agit d'une hiérarchie conceptuelle arborescente, fondée sur une structure terminologique et basée sur des principes linguistiques. Cette terminologie représente une organisation des connaissances propre à un domaine spécifique et à une tâche particulière dans ce domaine car, comme le souligne [Bourigault, 2002, p.83] : « Toute ressource terminologique ou ontologique est construite pour un usage spécifié, et c'est donc au sein de cet usage qu'elle doit être évaluée. ». Cette structure terminologique donne, de ce fait, la structure de base pour une ontologie qualifiée de « régionale » : elle ne cherche pas à modéliser le monde en général, mais un contexte d'usage bien spécifique dans un domaine ciblé. Nous considérons ici la gestion documentaire comme cadre applicatif pour nos ontologies, et nous nous intéressons aux domaines de l'audiovisuel et de la médecine.

En effet, ce mémoire découle d'une thèse CIFRE[1] qui s'est déroulée à l'Institut National de l'Audiovisuel, pour la partie industrielle, et dans les laboratoires de recherche du STIM[2] et de l'INSERM U729, pour le côté universitaire. Une des préoccupations commune à ces deux pôles est le questionnement autour de la gestion de grandes bases documentaires : l'INA doit gérer les archives audiovisuelles et radiophoniques publiques depuis sa création en 1974[3], et la recherche dans le domaine médical s'intéresse, entre autres, à la recherche d'informations dans de grandes bases de documents comme MEDLINE[4] ou le catalogue CISMeF[5]. Il existe d'autres cata-

[1] *Convention Industrielle pour la Formation et la Recherche en Entreprise*

[2] Mission de Recherche en Sciences et Technologies de l'Information Médicale, de l'Assistance Publique - Hôpitaux de Paris (AP-HP)

[3] Il existe deux types d'archives, pour deux types d'usagers distincts : une base documentaire correspondant au Dépôt Légal destinée à des chercheurs et celle des Archives, destinée à des professionnels de l'audiovisuel. Le fonds documentaire atteint un volume de plus de 2,5 millions documents, comprenant plus d'1,5 millions d'heures de radio et de télévision. Il est augmenté quotidiennement, notamment du fait de la captation (un mode d'enregistrement en temps réel) de certains programmes. La base documentaire du Dépôt Légal est mise à jour quotidiennement, avec un différé de 6 mois par rapport à la date de diffusion des programmes.

[4] MEDical Literature Analysis and Retrieval System on LINE, MEDLINE référence plus de 12 millions de documents traitant de tous les domaines biomédicaux : biologie, biochimie, médecine clinique, santé publique, éthique, économie, pharmacologie, psychiatrie, toxicologie, odontologie, médecine vétérinaire. Cette base documentaire est produite par la National Library of Medicine, et est mise à jour quotidiennement.

[5] Catalogue et Index des Sites Médicaux Francophones, CISMeF a pour but « d'assister les professionnels de santé dans leur quête d'informations » et est disponible en ligne. Il s'agit d'un catalogue spécialisé référençant les sites médicaux francophones répondant à un critère de « qualité de l'information de santé sur l'Internet (NetScoring), développé en collaboration avec Centrale Santé et APUIS-Santé ». Ce catalogue est le fruit d'un projet initié par le Centre Hospitalier Universitaire (CHU) de Rouen, débuté en février 1995, date de la création du site Web du CHU de Rouen. Les citations sont extraites du site Web.

logues médicaux, notamment CliniWeb[6], DDRT[7], HON[8], MedWebPlus[9] et OMNI[10], mais nous nous intéressons à MEDLINE parce que c'est la base médicale « la plus employée dans le domaine biomédical » [Zweigenbaum, 1999], et à CISMeF parce qu'il s'agit du catalogue médical francophone le plus fréquemment renvoyé lors de requêtes génériques portant sur le domaine médical, à partir de moteurs de requête eux aussi génériques (pour plus de détails, voir [Zweigenbaum *et al.*, 2002]). Tous ces catalogues médicaux sont indexés par le thésaurus MeSH.

Le traitement documentaire se divise en deux tâches distinctes et interdépendantes : l'indexation des documents et la recherche documentaire. Pour arguer de l'intérêt de construire des ontologies régionales dans ce contexte particulier, nous présentons en introduction les systèmes documentaires classiques et la structure de descripteurs sur laquelle la plupart se fondent : le thésaurus. Il est central à la fois à l'annotation sémantique des documents et à leur recherche dans les bases documentaires. Nous décrivons cet outil au premier chapitre (section 2), les possibilités qu'il offre en termes de traitement documentaire, mais aussi les limites qui lui sont inhérentes, en prenant des exemples tirés du thésaurus de l'INA (section 3.1) et du MeSH[11](section 4.1).

La réponse à ces limites constituent le cahier des charges à remplir par un outil efficace de gestion documentaire, et nous chercherons à vérifier l'adéquation des Systèmes à Base de Connaissance (SBC) pour ce type de tâche. Nous illustrons cette hypothèse dans le cadre du projet OPALES [12], dont un des enjeux était justement de tester l'apport des Systèmes à Base de Connaissance sur les systèmes documentaires, et pour lequel nous avons développé une ontologie différentielle sur le thème de la petite enfance. En effet, les Systèmes à Base de Connaissance ne fondent pas leur vocabulaire sur des thésaurus, mais sur des ontologies. Ce projet nous a donné l'occasion de nous familiariser avec notre objet d'étude, l'ontologie différentielle, et d'en détailler les caractéristiques d'un point de vue sémantique et linguistique. Si l'on veut pouvoir disposer de Systèmes à Base de Connaissances et de leurs avantages pour le traitement documentaire dans des domaines aussi différents et variés que possible, il est nécessaire de bénéficier de méthodologies claires pour aider à la construction d'ontologies. En effet, les domaines ou applications pour lesquels de telles ressources n'existent pas sont encore très nombreux, et leur modélisation manuelle *ex nihilo* est longue et coûteuse. Cet enjeu est précisément l'objet de notre thèse : proposer un ensemble de méthodes à combiner pour aider un modéliseur dans sa tâche de construction ontologique.

Nous intéressant plus spécifiquement aux ontologies différentielles, ontologies à

[6]Oregon Health Sciences University, `www.ohsu.edu/cliniweb`, qui n'est plus en ligne actuellement.

[7]Diseases, Disorders and Related Topics, `www.mic.ki.se/Diseases`

[8]Health on the Net Foundation, `www.hon.ch`

[9]`www.medwebplus.com`

[10]Organizing Medical Networked Information, `omni.ac.uk`

[11]Ainsi que sa variante (principalement sa traduction en français, réalisée par l'INSERM) utilisée pour l'indexation et la recherche d'information dans CISMeF.

[12]`http://opales.ina.fr`

fort fondement linguistique et terminologique, nous pouvons nous baser sur des corpus textuels pour proposer des guides à leur structuration. Nous verrons les différentes méthodologies proposées dans le champ de la construction d'ontologie à partir de corpus en section 3, et nous intéresserons particulièrement aux méthodologies motivées linguistiquement, proposées en Traitement Automatique des Langues (TAL). Nous délaissons les méthodologies fondées majoritairement sur des principes statistiques *et* ne recourrant pas à une analyse linguistique de corpus : ces approches exploitent les forme graphiques des occurrences du corpus et leur redondance pour produire des résultats, et nous verrons qu'elles ne sont pas adaptées à notre matériel textuel initial (à savoir, un corpus spécialisé faiblement redondant). Parmi les méthodologies TAL pour la construction d'ontologie, nous chercherons à montrer en particulier dans quelle mesure l'analyse distributionnelle[13] et l'extraction par patrons lexico-syntaxiques peuvent servir notre objectif particulier, et répondre aux différents besoins soulevés par la construction d'ontologies différentielles.

D'après [Bourigault *et al.*, 2004], il est possible de distinguer deux grands processus dans la construction d'une ontologie de type « linguistique » à partir de corpus : l'acquisition terminologique et la structuration terminologique. Concernant les ontologies différentielles, une troisième tâche émerge : la qualification des axes verticaux (axe liant un père et son fils ontologique, correspondant à un lien entre un terme et son hyponyme selon une approche terminologique) et la structuration et qualification des axes horizontaux (ceux liant des co-hyponymes, soit des termes partageant un même hyperonyme) de l'ontologie. La qualification sémantique de ce qui unit et différencie deux co-hyponymes est une question propre aux ontologies différentielles, comme nous le verrons en section 2.2.2, c'est pourquoi nous nous y intéressons plus spécifiquement dans ce mémoire. Les deux méthodologies de construction d'ontologie mentionnées ci-dessus permettent d'apporter des réponses à ces deux processus de modélisation ontologique et à cette tâche spécifique qu'est la recherche de principes différentiels, mais nous verrons qu'elles présentent un point de vue différent sur le domaine (chapitre 6). Cette différence est une richesse pour le modéliseur, qui a, dès lors, intérêt à avoir recours aux deux types de sources lors de son travail, comme le montrent les travaux de [Caraballo, 1999] et [Cimiano *et al.*, 2004]. Nous dressons un récapitulatif des apports respectifs de ces deux approches, et nous penchons ensuite sur la question de leur compatibilité. En effet, comment est-il possible de combiner les résultats issus de ces deux méthodes pour profiter de leur complémentarité ? Nous proposons un début de réponse dans le cadre de notre collaboration au projet PERTOMed[14], dans le domaine médical, et nous posons des pistes à suivre pour automatiser plus avant cette combinaison en conclusion.

[13] Nous nous intéressons aux résultats de l'analyse distributionnelle lorsqu'elle est calculée à partir de dépendances syntaxiques, et donc d'une analyse linguistique, plutôt qu'à partir de cooccurrences.

[14] `http://spim.jussieu.fr/pertomed/`

Première partie

Contexte de la recherche

Ce travail de thèse concerne l'aide à la modélisation d'ontologies différentielles. Les ontologies sont des structures conceptuelles, modélisées en fonction d'applications dans lesquelles elles seront intégrées. Parmi les différentes utilisations envisagées pour les ontologies dans la littérature, comme le Web Sémantique [Berners-Lee *et al.*, 2001] pour ne citer que la plus en vogue en ce moment, nous nous intéressons plus particulièrement aux applications documentaires. Certaines ontologies servent déjà à la recherche d'information dans des sites Web spécialisés, comme celle modélisée par Guiraude Lame [Bourigault & Lame, 2002] dans le domaine juridique : elle permet de sélectionner un concept de l'ontologie juridique pour faire une requête ou de l'expansion de requête sur la base de cette ontologie, à l'URL `http://ontologie.w3sites.net/`. Nous présentons en introduction les avantages qu'offre un Système à Base de Connaissances fondé sur une ontologie par rapport aux systèmes documentaires classiques. Pour cela nous entreprenons un survol de l'objectif et des moyens « classiques » du traitement documentaire : le thésaurus (chapitre 1, section 2). Nous présenterons notamment ce type de terminologie au travers de deux exemples concrets : le thésaurus utilisé à l'INA (chapitre 1, section 3.1), dans son contexte d'utilisation (section 3.2), et le thésaurus MeSH (section 4.1), ainsi que son emploi dans les systèmes PubMED et CISMeF (section 4.2). Nous en récapitulerons les avantages et les limites (section 5) et nous verrons que les Systèmes à Base de Connaissance permettent d'y répondre au chapitre 2. Nous argumenterons cette position, soutenue notamment dans les travaux de [Troncy, 2004b], à travers une expérimentation réalisée lors du projet OPALES (chapitre 2, section 2.2.1). Ce point une fois posé, nous nous pencherons sur le cœur de notre travail, à savoir exploiter les outils et méthodologies du Traitement Automatique des Langues pour aider un modéliseur humain à construire des ontologies différentielles à partir de corpus textuel (chapitre 3). En effet, très peu de ressources ontologiques sont disponibles à l'heure actuelle, et leur spécificité est telle qu'il n'est souvent pas possible de les utiliser dans une autre application que celle dans le cadre de laquelle elle a été modélisée initialement. Il est donc indispensable d'avoir des outils pour aider à construire ces structures à partir de l'expression des connaissances telle qu'elle est le plus souvent accessible : sous forme textuelle, afin de bénéficier des avantages liés aux Systèmes à Base de Connaissances dans tout nouveau domaine abordé, ou toute nouvelle application.

Chapitre 1

Indexation et recherche documentaire dans de grandes bases documentaires

Sommaire

1.1 Le traitement documentaire

Le traitement documentaire comporte deux aspects : une annotation des documents, comprenant une description sémantique de leur contenu (opération nécessaire et préliminaire à leur intégration à une base documentaire), et la recherche documentaire dans la base. La première phase a été qualifiée de procédurale par

Muriel Amar [Amar, 2000, p.26] : elle constitue le processus documentaire en tant qu'activité propre ; la deuxième phase est qualifiée d'instrumentale : le processus documentaire précédent est utilisé comme instrument pour retrouver des documents. L'annotation des documents se fait en anticipant sur la manière dont ils vont être recherchés, c'est-à-dire suivant les contraintes matérielles liées à la base documentaire, mais aussi, et surtout pour l'annotation sémantique, en anticipant sur les éléments de sens véhiculés par le document qui seront pertinents pour un utilisateur de la base. Cette contrainte est liée à la spécificité de la base documentaire (spécificité thématique, ou liée au type des documents, par exemple), et aux usagers auxquels elle est destinée. L'annotation « générale » assigne des éléments de catalogage au document (par exemple, la durée de l'émission et la chaîne de diffusion pour les unités documentaires de l'INA, ou bien le genre du document pour CISMeF : recommandation patient, article de périodique, conférence, etc.), alors que l'annotation sémantique consiste en une interprétation du contenu d'un document, suivant une perspective susceptible d'intéresser les futurs usagers de la base. Cette annotation sémantique est un des moyens privilégiés d'accès au document. Elle se fait en général selon deux moyens : la rédaction d'un résumé en texte libre se conformant aux normes rédactionnelles du centre documentaire et à l'exploitation documentaire envisagée, et l'association au document d'un ou plusieurs descripteurs, sélectionnés dans un ensemble fini au libellé normé. Cet ensemble peut être classé de manière alphabétique dans une liste d'autorité « à plat », mais il est souvent structuré hiérarchiquement, notamment lorsqu'il sert à indexer des bases documentaires de grande taille, couvrant un grand nombre de thématiques ou une seule thématique très spécialisée. L'ensemble de descripteurs hiérarchisé le plus courant en traitement documentaire est le thésaurus.

Nous allons à présent décrire plus en détail les caractéristiques d'un thésaurus (section 2), les possibilités qu'il offre en termes de traitement documentaire lorsqu'il est intégré dans un système documentaire (sections 3 et 4.2) et les limites qu'il lui impose, et qui ne sont pas résolues par le sysème lui-même (section 5). Nous verrons pour cela les exemples concrets des systèmes documentaires de l'INA (section 3), d'un côté, et PubMED et Doc'CISMeF pour les bases documentaires médicales, de l'autre (section 4.2).

2 Le thésaurus

[Spark Jones & Willet, 1997, p.111] reprennent la définition du Système Mondial des Sciences de l'Information de l'UNESCO pour décrire un thésaurus :

> Un thésaurus peut être défini selon sa fonction ou selon sa structure. En termes de fonction, un thésaurus est un outil de contrôle terminologique, qui sert à traduire la langue (naturelle) des documents, indexeurs ou usagers en un langage plus contraint, organisé de manière systématique (langage documentaire ou langage d'information). En termes de structure, un thésaurus est un vocabulaire contrôlé et dynamique, composé de termes reliés de manière sémantique et selon des relations de généricité, ces termes couvrant un domaine de connaissance spécifique. [15]

Le thésaurus structure donc des termes, qui sont également appellés descripteurs. Il s'agit d'éléments lexicaux normalisés qui servent à décrire le contenu sémantique de documents. Ces libellés sont isolés de leur contexte d'emploi en langue, et sont donc susceptibles de prendre des sens différents suivant le domaine spécialisé que l'on considère. Par exemple, Java peut référer à une île, une danse, un langage de programmation ou à un café. Pour que la représentation du sens au moyen de ces descripteurs ne soit pas ambigüe, que ce soit lorsqu'on les manipule dans une description ou en recherche d'information, ces termes doivent être associés à une sémantique fixée. Ils ne sont, en général, pas définis explicitement, mais leur sémantique est donnée par leur regroupement en domaines et leur organisation suivant plusieurs relations sémantiques.

2.1 Les relations sémantiques dans le thésaurus

[Grabar & Hamon, 2004] établissent un état de l'art de la place accordée aux différentes relations sémantiques dans les terminologies, selon leur type et selon la théorie suivant laquelle elles ont été réalisées. En effet, ils notent que :

> La place accordée aux relations entre termes dans les terminologies traditionnelles est très variable. Elle varie en fonction des types de relations, mais également en fonction des pratiques terminologiques adoptées. Ainsi, un produit terminologique ne sera pas le même selon que le terminologue travaille conformément aux recommandations ou aux normes définies par les organisations comme le comité TC 37 et l'ISO (cf. [Gouadec, 1990, p.219] et [Cabré, 1999, p.119–120]), ou bien qu'il respecte les usages dans l'entreprise ou ses propres pratiques [Srinivasan, 1992].

Ils reprennent la typologie classique des relations, que nous adoptons également ici, en relations taxinomiques, relations transversales et synonymie. Nous allons décrire

[15] Notre traduction de : *A thesaurus may be defined either in terms of its function or its structure. In terms of function, a thesaurus is a terminological control device used in translating from the natural language of documents, indexers or users into a more constrained « system language » (documentation language, information language). In terms of structure, a thesaurus is a controlled and dynamic vocabulary of semantically and generically related terms which covers a specific domain of knowledge.*

plus précisément ces trois types de relations, en prenant des exemples extraits du thésaurus employé à l'INA.

Les relations taxinomiques

Les relations taxinomiques organisent le thésaurus en un arbre hiérarchique de descripteurs. Wüster en considère deux types ([Wüster, 1981, p.85–101], d'après [Grabar & Hamon, 2004]) : la relation « est-un » et la relation partitive, ou « partie-tout ». Ces relations correspondent aux relations d'hyperonymie et de méronymie. L'hyperonymie associe un descripteur à son super-ordonné, c'est-à-dire un descripteur qui lui est plus générique, mais sans être trop général. Par exemple, dans la branche concernant la médecine du thésaurus utilisé à l'INA, le descripteur OSTÉOPATHIE est classé sous TRAITEMENT MÉDICAL, et non pas directement sous THÉRAPEUTIQUE, qui a comme autre fils MEDICAMENT :

- Thérapeutique
  - Médicament
  - Traitement médical
    - Ostéopathie

La méronymie est la relation partitive, de type « compose » ou « fait partie de » ; elle associe un descripteur à l'ensemble dont il fait partie. Par exemple, elle lie CENTRE HOSPITALIER et CHAMBRE STÉRILE (relation qui peut aussi être interprétée comme de la localisation), ou SANG et GLOBULE BLANC, GLOBULE ROUGE, PLASMA, SERUM.

Ces deux relations structurent de manière indifférenciée la hiérarchie du thésaurus (structuration verticale), représentée par le lien Terme Générique et son inverse, Terme Spécifique (en anglais : Broader term/Narrower term, B.T./N.T.) entre deux descripteurs. Mais ces deux relations ne sont pas les seules à remplir cette fonction. L'observation du cas particulier qu'est le thésaurus à facette de l'INA, où chaque domaine est décliné selon un certain nombre de sous-catégories régulières, mais aussi celle du MeSH, thésaurus à facette[16] plus classique dans sa forme, montre qu'un certain nombre de relations implicites organisent le thésaurus dans sa structure verticale. Le fait que les relations sémantiques de type verticales ne soient ni régulières ni spécifiquement typées dans la majorité des cas fait que le thésaurus n'est pas une structure propre à permettre des raisonnements automatiques poussés, comme nous le verrons en section 5. Un autre type de relations permet également d'associer des descripteurs du thésaurus : celles qui lient des descripteurs de différentes branches hiérarchiques, comme par exemple MÉDECIN MILITAIRE et ARMÉE. Ces relations sont qualifiées de transversales.

Les relations transversales

Les relations transversales lient des descripteurs de manière propres au domaine couvert par le thésaurus dans le cas d'un thésaurus ciblé (comme le MeSH) ou

[16] Un thésaurus à facette est une structure qui organise les descripteurs en fonction de catégories mutuellement exclusives, partitionnant l'espace terminologique du domaine [Tudhope *et al.*, 2002].

propre à une tâche documentaire spécifique, comme le thésaurus de l'INA, conçu pour décrire des émissions télévisuelles. Ces relations ne sont en général pas plus explicitement définies que les relations taxinomiques (verticales), mais sont simplement libellées « Voir Aussi » (V.A., en anglais : Related Terms, R.T.). Elles associent des thématiques proches, dont les descripteurs peuvent être pertinents pour compléter une indexation ou une requête.

La synonymie

Le troisième type de relation que l'on trouve dans un thésaurus est la relation d'équivalence, nommée ici « synonymie ». La synonymie est une relation linguistique à visée normative : elle précise l'ensemble des libellés correspondant potentiellement au sens d'un descripteur qui ne doivent pas être utilisés dans la description documentaire. Elle permet également de vérifier, lors d'une requête, si un mot est un descripteur ou non. La relation liant un descripteur à un libellé non préférentiel est appellée « Utilisé Pour » (U.P., en anglais : Used For, U.F.), et son inverse est appellée « Employer » (Use, symbolisé US ou USE). Elle recouvre également différents cas de figure : elle peut être employée pour témoigner d'une évolution d'ordre linguistique (ou culturelle) autour d'un descripteur (HANDICAPÉ U.P. INFIRME, HYPERLIPIDÉMIE U.P. HYPERLIPÉMIE), éviter un descripteur polysémique (MALADIE U.P. AFFECTION), simplifier la description au moyen d'un acronyme (ORL U.P. OTORHINOLARYNGOLOGIE), ou choisir un niveau de langue plus ou moins spécialisé (MALADIE MENTALE U.P. FOLIE), par exemple[17].

2.2 Thésaurus et systèmes documentaires

Si le thésaurus est une hiérarchie de descripteurs autonome, il est intéressant de l'inclure dans un dispositif informatique global de gestion documentaire. En effet, la forme hiérarchique du thésaurus et ses relations sémantiques présentent des intérêts à la fois en terme d'indexation et en recherche documentaire, et ces possibilités sont avantageusement relayées par les systèmes documentaires qui permettent d'en tenir compte dans des calculs informatiques. Nous allons à présent détailler trois sytèmes documentaires mettant en jeu deux thésaurus, afin de voir toutes les possibilités qu'ils offrent. Il s'agit du thésaurus de l'INA dans le système d'indexation TOTEM et du thésaurus MeSH employé dans PubMed et Doc'CISMeF. Nous présentons dans les prochaines sections ces thésaurus et leurs systèmes de gestion documentaires associés. Cette présentation nous montrera que, malgré les avantages qu'ils présentent, les thésaurus comportent également un certain nombre de limites que les systèmes documentaires dans lesquels ils sont intégrés n'arrivent pas à pallier. A partir de cette analyse, nous dresserons un tableau des spécificités qu'un outil plus performant de gestion documentaire devrait remplir et nous verrons que les ontologies, mobilisées dans des Systèmes à Base de Connaissance en sont des candidats

[17] Ces exemples sont extraits de l'arborescence MÉDECINE du thésaurus de l'INA, choisie pour garder une cohérence thématique entre les deux pôles applicatifs de cette thèse : les documents audiovisuels et la médecine.

intéressants.

3 Indexation et recherche documentaire à l'INA

Le premier thésaurus auquel nous nous intéressons est celui qui est utilisé à l'INA pour l'indexation et la recherche documentaire des archives audiovisuelles, par le biais du système documentaire TOTEM. L'indexation suit des normes documentaires évoluant au fil des années en fonction de l'évolution des technologies, de la mutation des matériels physiques de stockage des documents (passage de différents formats de supports analogiques à des fichiers numériques) et des demandes des clients. Lors de l'indexation, l'annotation sémantique consiste à décrire sous forme textuelle une interprétation du contenu sémantique d'un document audiovisuel. Cette description détaille les événements factuels présents à l'écran, les lieux, les « anecdotes visuelles » (des faits marquants, comme la célèbre séquence du cheval se lançant dans le peloton du Tour de France cycliste) et les sujets filmés (personnes, objets ou animaux), en respectant la structure narrative du document. Les entités apparaissant à l'écran sont les seules à être prises en compte, à la différence de celles qui sont seulement mentionnées dans le document, et ce malgré la dimension sonore du matériel audiovisuel. Les demandes en matière d'audiovisuel concernent en effet plutôt la dimension visuelle du médium. Les requêtes de type commercial, destinées aux professionnels de l'audiovisuel, sont souvent axées sur des « images » de personnes politiques ou célèbres dans le show-biz', d'objets (le bathyscaphe de Cousteau) ou d'animaux : le document audiovisuel est pris dans sa dimension illustrative ou ludique. Des séquences plus longues sont parfois recherchées dans une optique de mémoire, c'est-à-dire pour leur valeur documentaire, comme celles de la crue de 1901, ou de la libération des camps de concentration. Le thésaurus doit donc permettre de décrire des personnes, objets ou animaux, des situations et des actions ou événements, ainsi que de les retrouver avec le maximum de précision et de rapidité. Ce dernier point tient de la gageure sachant l'importance du volume documentaire des archives, volume en accroissement quotidien. Voyons plus en détail la manière dont le thésaurus est construit et les moyens que le système documentaire met à disposition des documentalistes pour indexer et rechercher ces types d'informations.

3.1 Le thésaurus INA

Le thésaurus employé à l'INA est un thésaurus à facettes : il est divisé en neuf catégories mutuellement exclusives (Sciences, Sport et Vie économique, par exemple). Chaque thème principal est ensuite organisé, autant que faire se peut, suivant des catégories pré-établies, que nous qualifions ici également de facettes. Sous la racine figurent les thématiques principales liées à ce thème, puis les types de Lieux, de Matériel, d'Actions/Activités/Evénements, de Personnes, de Produits, de Techniques, de Bâtiments/Infrastructures qui y sont typiquement rattachés. Par exemple, le thème CHIRURGIE, pour rester dans le domaine médical, est structuré suivant les facettes : Personnes-chirurgie, Matériel-réanimation (les facettes peuvent catégoriser des branches à n'importe quel sous-niveau, elles ne sont pas cantonnées à la racine de l'arborescence), qui se retrouvent dans d'autres thèmes, mais aussi Spécialités-Chirurgie, et Accident-chirurgie. En effet, certaines facettes concernent des régula-

rités propres à un seul thème, comme Arbitrage-Judo ou Applications-Atome. Ces sous-thématiques sont le fruit d'un travail sur la normalisation de la pratique documentaire qui a lieu à l'INA en parallèle avec celui d'autres organismes gérant des documents audiovisuels (notamment les travaux de l'Association « Mediadoc-Sciences »[18]), en conformité avec les normes établies en la matière[19]. Dans le cas de ce thésaurus, on peut donc distinguer un ensemble de relations implicites, correspondant en fait à des relations transversales selon notre définition précédente. La sémantique de ces relations est donnée par le type de la facette : « acteur-de », « outil-de », etc. Cependant, le thésaurus étant en évolution permanente (et géré informatiquement depuis relativement peu de temps, ce qui impose une lenteur dans l'harmonisation des mises à jour), certaines incohérences se laissent entrevoir : il y a une certaine redondance des facettes, comme Actions/Activités/Evénements, qui sont ou liées à un emploi syntaxique préférentiel (un événement extraordinaire, une activité économique et une action sportive), ou à une harmonisation lexicale en cours, qui viserait à sélectionner un libellé préférentiel mais n'aurait pas encore été effective dans tout le thésaurus (il compte en effet plus de dix mille descripteurs, dans la version du 8 août 2003, et ne comprend dans sa version actuelle plus que la facette Action). Une autre incohérence est la non-explicitation d'une facette pour une branche de la hiérarchie. Par exemple Accident-chirurgie pourrait être intégré sous Actions/Activités/Evénements-chirurgie, or cette dernière facette n'est pas présente pour le thème Chirurgie ; Anesthésiste figure à la racine de ce même thème de Chirurgie, sans mention de la facette Personnes-chirurgie, et cette facette ne comprend qu'un seul descripteur : le chirurgien.

Il s'agit certes de partis-pris de catégorisation des personnes, événements, etc., mais le manque de rigueur (dans le libellé de la facette) et d'exhaustivité (dans la catégorisation en facettes) font que ce potentiel intéressant ne peut être exploité informatiquement pour des raisonnements automatiques lors de l'indexation ou de la recherche documentaire. En effet, nous verrons que le système documentaire permet d'afficher le contexte d'un nœud thésaural élicité par son libellé (cherché par ordre alphabétique ou par exploration de l'arbre du thésaurus), mais pas de se baser sur les informations que ce nœud contient pour faciliter le travail documentaire.

A côté de ce thésaurus thématique, une liste de noms propres associés à leurs catégories (lieu géographique, homme politique, etc.) est également mobilisable dans le processus documentaire. Mais ces noms propres ne sont pas rattachés à l'arborescence du thésaurus, ce qui limite également les possibilités de raisonnement automatique que l'on peut y associer. Il manque un lien entre ces « instances » et leur catégorie dans le thésaurus, et donc leur lien avec l'arborescence des descripteurs.

Les relations transversales sont présentes sous la forme de renvoi textuel V.A.

[18] Association regroupant 10 médiathèques, centres de documentation et centres de recherche, réunis pour « améliorer la diffusion et l'utilisation des documents audiovisuels en France », publiés notamment das la collection Guides Pratiques : Décrire l'Audiovisuel - manuel méthodologique pour l'analyse de contenu des documents audiovisuels à caractère documentaire.

[19] La norme AFNOR NF Z 44-070, Documentation - Indexation analytique par matière et la norme ISO 999 :1996, Information et documentation - Principes directeurs pour l'élaboration, la structure et la présentation des index.

(VOIR AUSSI), qui peut être suivi manuellement en description ou recherche documentaire. Cependant, ces relations ne sont pas implémentées, dans la version informatique du thésaurus, sous forme de lien physique exploitable automatiquement (pour des expansions de requête, par exemple).

Un certain nombre d'autres informations textuelles sont associées aux descripteurs : des notes historiques, la mention de la date de création du descripteur, des notes d'usage etc. Ces informations suivent une visée prescriptive dans la mesure où elles indiquent les contextes où utiliser ou ne pas utiliser le descripteur en question. Mais elles ne sont associées à aucun outil de vérification de contraintes dans le système documentaire. Elles ont également une valeur descriptive : elles servent à comprendre le contexte d'usage d'un descripteur, mais dans la pratique quotidienne, il n'est pas évident que les documentalistes aient le temps matériel suffisant pour lire les notes d'usage et sélectionner *le* descripteur le plus approprié à la situation. C'est pourquoi un comité de personnes en charge de la réflexion sur l'évolution des pratiques documentaires[20] sélectionne et diffuse un ensemble de descripteurs à mobiliser dans le cadre d'un nouvel événement majeur, comme le raz-de-marée gigantesque qui a eu lieu en Asie en décembre 2004, et qui a été appelé dans les journaux télévisés « Tsunami », d'après l'appellation japonaise d'un tel phénomène. Dans le cadre de cet événement particulier, les mots-clés sélectionnés pour être associés aux notices sont les suivants : RAZ DE MARÉE|TSUNAMI. C'est dans ce contexte de pratique quotidienne que certaines dénominations deviennent préférentielles et d'autres désuètes, faisant évoluer les descripteurs du thésaurus. Celui-ci est mis à jour régulièrement, mais il n'y a pas de procédure automatique de rétroactivité dans les notices déjà créées et stockées dans la base documentaire : il faut rajouter le nouveau libellé du descripteur aux anciennes notices de manière manuelle, ce qui correspond à un travail considérable. C'est pourquoi une indexation plus indépendante du libellé des descripteurs pourrait présenter des avantages : un nouveau libellé serait rattaché à la notion utilisée pour indexer les documents antérieurs, et ils pourraient être retournés quel que soit le libellé du descripteur sélectionné lors d'une requête.

La relation de synonymie, telle que mentionnée plus haut, est également présente dans le thésaurus de l'INA, sous la forme du renvoi U.P., UTILISÉ POUR. Il recouvre, en fait, plusieurs types de rapports de semi-équivalence différents : un niveau de précision différent dans la description, comme dans l'exemple BOUTIQUE U.P. ÉCHOPPE[21] ; un même libellé peut également être préconisé pour regrouper deux notions différentes plus spécifiques, comme dans le cas de CHANGE U.P. COURS DES CHANGES (qui correspond à la *Cotation officielle des valeurs des monnaies étrangères*) *et aussi* U.P. MARCHÉ DES CHANGES (qui signifie *Vente ou achat des devises étrangères.*). On trouve encore des préconisations qui sont liées à l'évolution terminologique d'une notion : TÉLÉVISION LIBRE U.P. TÉLÉVISION PIRATE,

[20]Ce même comité est chargé de valider les suggestions d'ajout de descripteurs au thésaurus et de les intégrer à la structure existante, et de veiller à l'actualisation du libellé des descripteurs en fonction de l'évolution de la langue.

[21]En effet, une échoppe est une *« Petite boutique en appentis, adossée à un mur et faite généralement de planches »* d'après le dictionnaire en ligne TLFi, développé par l'INALF puis par l'ATILF.

BOXE FRANÇAISE U.P. SAVATE -SPORT. Ce dernier exemple présente également un descripteur composé pour désambiguïser le sens de « savate » seul, entre un sport et une pantoufle : la catégorie à laquelle il appartient est accolée au descripteur simple. Les relations sémantiques entre les descripteurs liés par ce même UTILISÉ POUR sont donc différentes et irrégulières, et demanderaient une description sémantique plus détaillée pour permettre une exploitation informatique systématique et uniforme.

En résumé, on peut retenir les points négatifs suivants :

- Le thésaurus est subdivisé en facettes, pour une gestion des descripteurs simplifiée du point de vue du documentaliste en charge de la description ou de la recherche documentaire. Mais différents libellés peuvent renvoyer à une même notion (comme les deux facettes de Action/Activité, qui peuvent sembler concurrentes), et tous les descripteurs correspondant à des facettes ne sont pas organisés selon les catégories correspondantes : il n'y a pas de systématicité dans le découpage en facettes ni dans la lexicalisation des notions sous-jacentes à ces facettes ;
- Les instances (la liste des différents noms propres) ne sont pas rattachées à la hiérarchie des descripteurs, ce qui empêche de faire des raisonnements sur leurs types. Le différents types associés à ces noms propres (personnes politiques, lieux géographiques) ne sont d'ailleurs pas présents dans le thésaurus ;
- Les relations transversales devraient permettre de naviguer entre les différentes branches du thésaurus, mais ce renvoi n'est pas implémenté dans le système, il ne figure que sous la forme d'une indication textuelle ;
- Les informations liées aux descripteurs ne sont pas utilisées à des fins de raisonnement, comme la vérification de la validité d'une annotation, par exemple ;
- Les relations de type transversal ne sont pas typées plus avant, pas plus que celles d'équivalence (de « synonymie »). Elles ne sont pas exploitées à des fins de raisonnement automatique dans le système documentaire, malgré les informations qu'elles véhiculent. Ces informations sémantiques ne sont pas clairement décrites ni assez formelles pour être mobilisables en tant que ressources pour des inférences (c'est-à-dire des raisonnements automatiques).

3.2 Le système d'indexation et de recherche documentaire de l'INA

L'outil documentaire de l'INA se nomme TOTEM, et permet, entre autres tâches de gestion du document, de rédiger une notice descriptive, de l'associer à une unité documentaire sélectionnée[22], et de faire de la recherche dans la base des documents indexés. Son moteur de recherche intègre une implémentation spécifique du moteur de recherche et d'indexation Verity, développé par la société du même nom. Nous allons voir une description plus détaillée de cet outil selon chacun de ces deux aspects :

[22] TOTEM permet de sélectionner une unité documentaire dans le flux, d'y associer une imagette « représentative », et de lier l'unité documentaire isolée à une séquence plus longue à laquelle elle appartient, ce qui permet d'avoir différents niveaux de description documentaire : un niveau fin qualifié d'« extraits », un niveau plus global au niveau d'une émission complète, et un niveau méta, qui relie une émission à une collection, par exemple.

GRILLE D'INDEXATION TYPE POUR LES AFFAIRES	
Première phase	début de la médiatisation des faits (découverte de l'affaire)
Deuxième phase	médiatisation de l'enquête policière / instruction judiciaire
Troisième phase	médiatisation du procès

TAB. 1 – Exemple de grille d'indexation pour un événement de type « récurrent »

la rédaction de la notice documentaire et la recherche de documents.

La rédaction de la notice documentaire

Lors de la création d'une notice documentaire, l'utilisateur du système a accès à l'arborescence du thésaurus : il peut entrer un mot-clé et visualiser le descripteur dans son contexte proche, c'est-à-dire le descripteur qui lui est immédiatement super-ordonné, celui ou ceux qui sont situés au même niveau de hiérarchie que lui, celui ou ceux qui lui sont immédiatement subordonnés et l'ensemble des recommandations textuelles qui lui sont associées. Le fait de visualiser le nœud hiérarchique du thésaurus donne accès à toutes les facettes associées au descripteur sélectionné, puisqu'elles sont distribuées à tous les niveaux de hiérarchie. Cela permet de se positionner au niveau de granularité sémantique correspondant au contenu du document (par exemple, d'indexer une séquence par OSTÉOPATHIE et non pas par son générique : TRAITEMENT MÉDICAL) et de faire une indexation aussi précise que possible, en associant toutes les facettes pertinentes à la description du document en question (CHIRURGIE et CHIRURGIEN, correspondant à la facette personnes-chirurgie, par exemple).

Il est également possible de créer des associations particulières dans le champ *mots clés*, au moyen d'une « précision d'indexation ». Un descripteur est alors associé à du texte libre suivant une relation typée. Par exemple une personne, Gérard Depardieu, peut être associée au type « rôle », et liée par « précision d'indexation » au texte libre « Jean Valjean » pour décrire un document audiovisuel où G. Depardieu interprète le rôle de Jean Valjean. Des termes ou mots-clés qui ne figurent pas dans le thésaurus peuvent également être associés à des descripteurs, comme nous l'avons vu plus haut : RAZ DE MARÉE|TSUNAMI, où TSUNAMI ne figure pas encore dans le thésaurus.

Nous avons déjà vu lors de la description du thésaurus qu'une volonté de normalisation était mise en œuvre dans le processus documentaire. Elle prend également la forme de recommandations pour la description d'événements récurrents, ou d'une sélection terminologique pour celle d'un événement nouveau ou émergeant (indexation des reportages parlant du SRAS, de la grippe aviaire, du Tsunami, etc.). Des schémas spécifiques ont été élaborés pour certains types d'événement : par exemple, dans le cas des « affaires », la grille présentée au tableau 1 a été définie en mai 1997.

TOTEM permet d'accéder aux différentes grilles disponibles, mais uniquement sous la forme d'un texte, dont il faut recopier les éléments lors de la création d'une nouvelle notice. Il n'existe pas de structure de contrôle ou de mécanisme automatique

pour transformer ces grilles en un cadre formel pour la notice, ni pour valider la conformité des notices créées par rapport à la grille. Ce serait pourtant une aide à la rédaction appréciable, et un support de raisonnement intéressant. Nous verrons que d'autres types de structures permettent de créer des schémas prédéfinis, de les utiliser dans la description documentaire et dans des raisonnements automatiques. Dans le cadre du système documentaire actuel de l'INA, ces grilles sont utilisées dans la seule perspective d'aide à la normalisation et à la systématicité dans la rédaction de la notice.

Pour résumer, TOTEM permet de visualiser le thésaurus, et d'en sélectionner des descripteurs lors d'une description documentaire. Il donne également accès à l'historique du descripteur, aux précisions d'usage qui lui sont associées, ainsi qu'aux grilles correpondant à des événements récurrents dans la tâche documentaire à l'INA (comme la grille concernant les affaires que nous avons vue plus haut). Ces derniers types d'information ne permettent toutefois pas de faciliter la tâche rédactionnelle de la notice, car ils ne sont pas exploités informatiquement pour fournir, par exemple, une première structure à instancier dans l'outil documentaire. Ils ne permettent pas non plus de contrôler sa conformité par rapport au modèle. Nous allons à présent nous intéresser aux fonctionnalités de cet outil en recherche documentaire.

La recherche documentaire

La recherche documentaire bénéficie également de la structure hiérarchique du thésaurus. Il est, en effet, possible de sélectionner, par navigation dans les branches, le degré de précision d'une requête, ou associer des sous-thématiques correspondant à des facettes pertinentes à la requête.

L'interface de requête comprend les champs typés de la notice documentaire : genre du document recherché, date de diffusion, etc. et les mots clés « sémantiques », c'est-à-dire ceux liés à une description thématique de son contenu. Une requête peut combiner les informations de plusieurs champs thématiques dans TOTEM, ou différentes informations dans le même champ. Il est, par exemple, possible de chercher une émission de variété présentée par Léon Zitrone en faisant une requête combinant le genre « jeux » dans le champ spécifique de l'interface de recherche et l'association *Zitrone*/PRE dans le champ concernant la description du contenu, pour indiquer que l'on s'intéresse uniquement à Léon Zitrone dans son rôle de présentateur (*Zitrone* associé au rôle PREsentateur). Les combinaisons entre champs thématiques se font au moyen des opérateurs booléens classiques (ET, OU et NON) et/ou des opérateurs dits « de test » suivants :

- « contient » recherche des notices contenant la chaîne de caractères dans l'ordre ;
- « ne contient pas » permet de rechercher des notices ne contenant pas la chaîne spécifiée, dans l'ordre spécifié (pour raffiner des requêtes sur une thématique plus large, par exemple) ;
- « égal à » recherche la chaîne de caractères exacte ;
- « différent de » permet de faire une recherche excluant la chaîne de caractère exacte spécifiée ;

– « contient à proximité » permet de rechercher deux mots avec une proximité de 1 à n mots et dans n'importe quel ordre dans la notice en texte libre ;

Une requête peut également contenir des « jokers » : une étoile * remplace une ou plusieurs lettres, pour permettre de rechercher des mots présents au singulier ou au pluriel dans la notice, ou de contourner le problème des orthographes divergentes, comme celle de François Mitterrand, présent dans la base sous 6 orthographes différentes[23]. TOTEM permet aussi d'ignorer un ou plusieurs mots, pour permettre des recherches de séquences de mots proches mais non immédiatement contigus dans les notices. Cette fonctionnalité est représentée dans l'exemple suivant par la séquence *JOKER : Mit*er*and serr* JOKER la main JOKER Kohl* permet de retrouver des notices comprenant les séquences « Miterrand serre la main de Helmut Kohl », « Miterrand serrant solennellement la main de Kohl » ou « Mitterrand serre la main du chancellier allemand Kohl ».

Le système garde la trace des requêtes, que chaque documentaliste peut archiver. Cela permet de capitaliser les différentes expériences de recherche documentaire et d'exploiter ces sortes de schémas de procédures dans des contextes de requête analogues. Les connaissances obtenues sont également capitalisables : le fait que les requêtes soient mémorisées permet aussi de combiner des résultats de recherches documentaires avec de nouveaux critères.

3.3 Les limites du thésaurus INA et de TOTEM

Nous allons à présent détailler les points faibles de TOTEM, en description documentaire comme en recherche, et nous nous intéresserons à un autre thésaurus, utilisé dans le domaine médical, afin d'oberver si les mêmes limites apparaissent. Nous nous pencherons sur le thésaurus MeSH, et sur les systèmes documentaires PubMed et CISMeF qui l'utilisent. Nous présenterons ensuite une alternative possible pour répondre aux différentes limites soulevées : les Systèmes à Base de Connaissance, fondés non plus sur un thésaurus, mais sur une ontologie.

La description documentaire et la recherche d'information actuelles à l'INA se basent sur la structure arborescente du thésaurus, présentant un descripteur dans le contexte de son thème général et des différentes facettes selon lesquelles il se décline, et sur un ensemble de connecteurs logiques ou *ad hoc* pour associer les descripteurs sélectionnés lors d'une requête.

La structure du thésaurus est certes arborescente, mais cette arborescence se décline suivant différentes relations sémantiques, dont la sémantique, justement, n'a pas été définie de manière à être interprétable par le système informatique de traitement documentaire. La structuration verticale mélange le rapport hypéronymique (relation sémantique de type « est-un ») entre un descripteur et celui qui est classé directement au-dessus lui, la méronymie (rapport entre une partie et un tout), ou encore les différentes facettes qui lui sont liées : lieux, infrastructures, personnes, matériels, etc. Les facettes sont le plus souvent isolées par un libellé ex-

[23]Ce nom propre est en effet orthographié : MITERAND, MITERRAND, MITTERRAND, Mitterand, MITTERAN et MITTERRAN.

plicite, mais pas systématiquement, et aucun mécanisme de raisonnement ne prend en compte ces informations au niveau informatique. La valeur sémantique des autres relations présentes entre descripteurs n'est pas libellée et les seules relations transversales à même d'exprimer des connaissances au moyen de descripteurs de différentes branches, comme « Opération chirurgicale à cœur ouvert réalisée par le Professeur XX » sont limitées à des renvois du type VOIR AUSSI entre thèmes connexes : MÉDECINE - CHIRURGIE - ANATOMIE, par exemple. Il n'est pas possible d'exprimer explicitement des relations de localisation, ou des informations temporelles, par exemple, ce qui pourrait être intéressant dans la mesure où les descriptions qui nous intéressent sont centrées sur les activités de personnes, d'animaux ou la description d'objets physiques de manière générale, et sachant que les documents audiovisuels ont comme caractéristique première d'être des images en mouvement, donc à forte composante temporelle et spatiale.

Le fait de pouvoir faire des requêtes concernant un acteur jouant un rôle particulier, ou une personne en tant que présentateur dans une émission est liée à la fonctionnalité de la « précision d'indexation », que l'on peut interroger dans TOTEM. Cette précision d'indexation confère une utilisation du thésaurus proche des terminologies post-coordonnées, telles que la SNOMeD dans le domaine médical (voir [Zweigenbaum, 1999]), ou dans une moindre mesure le MeSH. Le MeSH est un thésaurus à hiérarchies multiples (un descripteur peut se trouver classifié dans différentes hiérarchies) qui comporte des descripteurs pré-coordonnés, c'est-à-dire à mobiliser en tant qu'entités autonomes (INTERVENTIONS CHIRURGICALES VOIES BILIAIRES, pour citer un exemple de [Dailland *et al.*, 2004]), mais comporte également des descripteurs combinables pour créer de nouvelles entités : *Chirurgie du Foie* est par exemple exprimée par l'association du qualificatif CHIRURGIE au descripteur principal FOIE. Ce type de terminologie permet de construire des termes complexes, que ce soit en associant des termes primitifs de différentes facettes dans la SNOMED, ou des qualificatifs appropriés à des descripteurs majeurs dans le MeSH. Cependant, les combinaisons de type « précision d'indexation » dans le système de requêtes de l'INA ne concernent qu'un nombre restreint de descripteurs. Ce type d'association, même limité, ayant été perçu comme une avancée significative dans le système de recherche, on peut imaginer que l'introduction d'autres relations transversales typées dans le système documentaire pourrait être considéré avec intérêt.

La structure du thésaurus est tout à fait pertinente pour une manipulation humaine (que ce soit lors de la description ou dans la recherche d'information), dans la mesure où c'est un acteur humain qui interprète les différentes relations et fait la sélection terminologique adéquate ; elle n'est cependant pas assez formelle pour qu'un ordinateur s'en serve à des fins de raisonnement. Pourtant ce type de raisonnement informatique pourrait aider à la fois la pratique documentaire en tant que telle (i.e. la description sémantique d'un document), mais également la recherche d'information. En effet, Raphaël Troncy a montré dans sa thèse [Troncy, 2004a] que rajouter des raisonnements au format documentaire permet d'accroître le rappel des requêtes. Par exemple, si une requête concernant une interview de Sandy Casar permettait bien de retrouver une notice documentaire pertinente, la réponse fournie par le système documentaire avait plusieurs défauts : l'interview était entrecoupée

d'autres reportages, ce qui signifie que la réponse était bruitée (elle contient du bruit informationnel : une partie non désirée lors de la requête). Il s'agissait en fait de la *suite* d'une interview, dont la première partie était indexée au moyen d'autres mots clés, et n'était pas renvoyée lors de la requête : la réponse comportait donc également du silence informationnel (une partie de la réponse n'est pas fournie par le système). De plus, une requête concernant plus génériquement un *coureur cycliste* ne renvoie pas cette notice puisque l'information que Sandy Casar est un coureur cycliste ne figure nulle part (l'instance n'est reliée à aucun type sémantique sur lequel faire des raisonnements automatiques).

Il y a également un aspect lié aux recommandations de descriptions qui n'est pas pris en compte dans la requête, ou alors uniquement incidemment, du fait que ce sont les mêmes documentalistes qui indexent et qui recherchent des documents : la régularité des descriptions en fonction des types d'événements pourrait être prise explicitement comme modèle ou grille de recherche lors de requêtes concernant ces événement typés. Ce n'est pas le cas actuellement.

Et enfin, on peut penser à un problème générique lié à l'utilisation d'un outil *ad hoc* comme TOTEM, qui est la limitation en termes d'échange de données et d'interopérabilité entre différents centres documentaires, ou pour mise à disposition de contenus sur Internet (voir [Turner *et al.*, 2000]).

En résumé, on peut dégager les besoins suivants dans la tâche documentaire : représenter le contenu sémantique d'un document de manière non ambigüe et de manière à ce que cette représentation soit opérationnelle pour une gestion informatique du système (nécessaire dans le cas de grosses bases documentaires). L'interprétation du contenu du document et la représentation de sa sémantique suivent des principes de normalisation qui gagneraient à être mieux pris en compte dans le système informatique. Et enfin, la représentation du contenu d'un document dans une notice se faisant au moyen de descripteurs, certes normés mais toutefois issus de la langue naturelle, il est important de disposer de méthodes efficaces pour faire le lien entre une représentation de la sémantique d'un document et celle d'une requête, la correspondance n'étant pas triviale.

4 Indexation et recherche d'information dans MEDLINE et CISMeF

Nous allons à présent nous pencher sur un autre thésaurus à facette, le MeSH, spécialisé dans le domaine médical. Le domaine est cette fois-ci plus restreint (le thésaurus de l'INA est conçu pour pouvoir décrire toutes les émissions télévisuelles, ce qui représente un nombre de descripteurs ouvert et en constante évolution), il est de type scientifique et semble bénéficier d'une structure stable et reconnue au niveau international. Nous verrons que l'indexation au moyen de thésaurus dans le domaine médical pose également un certain nombre de problèmes. Nous chercherons à voir si les difficultés reprennent celles évoquées dans le cadre de l'INA, et nous examinerons si une des réponses possibles qu'est l'utilisation d'ontologies pourrait convenir aux

deux types de situations. Le thésaurus MeSH servant à indexer, entre autres, les bases documentaires médicales de MEDLINE et CISMeF[24], nous allons les présenter brièvement avant de décrire le thésaurus plus avant, et de nous intéresser aux deux systèmes documentaires correspondant à MEDLINE et CISMeF qui l'implémentent.

La base MEDLINE et le catalogue CISMeF regroupent des articles médicaux dont la qualité est évaluée par un ensemble de spécialistes. Ces bases documentaires sont indexées manuellement : à chaque document est associé un certain nombre de descripteurs concernant son genre (article de conférence, séminaire, etc.), son URL d'accès et les sujets abordés dans son contenu. Ces derniers sont extraits du thésaurus MeSH, ou de sa version française [25]. MEDLINE se consulte notamment par l'interface de recherche Pub'Med, et CISMeF par l'intermédiaire de Doc'CISMeF. Nous présentons les grandes lignes du MeSH en section 4.1 et la manière dont il est intégré dans les outils de recherche en 4.2. Nous dresserons un bilan des limites liées à ces associations et présenterons une alternative qui a été envisagée également dans le domaine médical : l'utilisation d'une ontologie[26] plutôt que d'un thésaurus, et un Système à Base de Connaissances plutôt qu'un système documentaire.

4.1 Le MeSH

Le MeSH (Medical Subject Headings) est le thésaurus développé par la National Library of Medicine (NLM)[27], aux Etats-Unis. Il s'agit d'une structure hiérarchique comprenant plus de 22 000 descripteurs, déclinés selon 9 niveaux de spécificité. Le niveau le plus général est organisé en 15 catégories :

- Anatomie,
- Organismes,
- Maladies,
- Produits chimiques, biologiques et pharmaceutiques,
- Équipements et techniques analytiques, diagnostiques et thérapeutiques,
- Psychiatrie et psychologie,
- Sciences biologiques,
- Sciences physiques,
- Anthropologie,
- Technologie aliments et boissons,
- Arts et sciences humaines,
- Sciences information,
- Individus,

[24] Voir l'Intoduction Générale pour une justification de ce choix.

[25] Traduction de l'ensemble des termes et de certaines définitions effectuée par l'INSERM.

[26] Voir les projets Ménélas et GALEN pour des exemples d'utilisation d'pntologies dans le domaine médical.

[27] La National Library of Medicine, située sur le campus des National Institutes of Health (Instituts Nationaux de Santé, équivalents, toutes proportions gardées, de l'INSERM en France) à Bethesda, dans le Maryland, est la plus grande bibliothèque médicale au monde, selon les informations fournies par son site Internet : `http://www.nlm.nih.gov/about/index.html`. La bibliothèque rassemble des documents et fournit des informations et des services de recherche dans tous les secteurs / toutes les disciplines de la biomédecine et de la santé.

– Santé (administration des soins),
– Emplacements géographiques [28].

Les descripteurs plus spécifiques sont principalement classés selon la relation d'hyperonymie ou de meronymie. Par exemple, la catégorie ACCIDENT se décline notamment suivant les sous-catégories de CHUTE ACCIDENTELLE, ou de NOYADE, plus les différents types d'accidents : ACCIDENT AVION, ACCIDENT CIRCULATION, ACCIDENT DOMESTIQUE, etc. Mais on peut également trouver d'autres types de relations entre un descripteur et son niveau subordonné. Par exemple, sous ACCIDENT, en dehors de la liste mentionnée ci-dessus (autrement dit, des descripteurs qui entretiennent bien une relation d'hyponymie avec ACCIDENT), on trouve également le descripteur PRÉVENTION DES ACCIDENTS. Cette fois-ci, la relation sémantique n'est plus de l'hyponymie, mais peut être interprétée comme une relation transversale, malgré le fait qu'elle soit représentée dans le thésaurus sous la forme d'un lien de type hiérarchique.

Les relations transversales explicitement prévues dans le MeSH (et non pas celles qui se trouvent de manière implicite dans la hiérarchie de descripteurs) visent à permettre une description précise des concepts spécifiques au domaine médical. Elles se présentent de manière classique sous la forme de renvois dans la fiche décrivant chaque descripteur (voir le tableau 2), ou sous la forme de descripteurs à rattacher aux différents termes « principaux » du thésaurus, pour en préciser la sémantique. Il est possible d'associer un total de 84 qualificatifs aux descripteurs appropriés appartenant aux 15 catégories générales mentionnées ci-dessus. On ne peut associer qu'un qualificatif à la fois à un descripteur MeSH. Le qualificatif peut contraindre l'interprétation d'un descripteur à un seul aspect (comme, par exemple lors de l'association d'un descripteur avec le qualificatif *diagnostic*), induire l'interprétation d'une relation sémantique particulière (MALADIE/ÉTIOLOGIE marque une relation de cause à effet), ou peut servir à construire de nouveaux descripteurs post-coordonnés, comme par exemple FOIE/CHIRURGIE pour décrire un document traitant de chirurgie du foie.

La synonymie est également gérée dans le MeSH par la fiche associée à chaque entrée que nous avons mentionnée plus haut, comprenant des indications textuelles sur son emploi, une définition pour la plupart des descripteurs et un ensemble de libellés qui correspondent au descripteur : des variantes lexicales ou des notions spécialisées qui ne correspondent à aucun descripteur propre dans le MeSH. Un exemple d'extrait de la fiche correspondant au descripteur anglais PAIN (DOULEUR en français) est représenté tableau 2. La première ligne correspond au libellé du descripteur, la deuxième détaille un ensemble de recommandations aux indexeurs, la troisième donne une explication détaillée du concept ; s'ensuivent les différents synonymes et termes à ne pas employer en indexation, des renvois, la liste des qualificatifs associables à cette entrée et une combinaison post-coordonnée figurant comme entrée du MeSH.

[28] Traduction présente sur le site de CISMeF, à l'URL `http://www.chu-rouen.fr/ssf/arborescences.html`

MeSH Heading	Pain
Annotation	IM GEN only ; prefer precoord locational terms like ABDOMINAL PAIN, CHEST PAIN, etc. [. . .]
Scope Note	An unpleasant sensation induced by noxious stimuli and generally received by specialized nerve endings.
Entry Term	Suffering, physical
Entry Term	Ache
Entry Term	Pain, Burning [. . .]
See Also	Analgesia
See Also	Analgesic
See Also	Hyperalgesia
See Also	pain insensitivity, Congenital
See Also	Palliative Care
Allowable Qualifiers	BL, CF, CI, CL, CN, DO,. . .
Entry Combination	drug effects :Pain
Unique ID	D010146

TAB. 2 – Extrait de la notice décrivant le descripteur anglais PAIN, avec mention des synonymes et des renvois

4.2 Les systèmes d'indexation et de recherche d'information médicale PubMED et Doc'CISMeF

Les descripteurs MeSH définissent un vocabulaire contrôlé qui permet de décrire les articles indexés, entre autres, dans le catalogue médical MEDLINE. L'intérêt d'un tel vocabulaire est notamment d'éviter les problèmes de synonymie à l'indexation mais également lors de l'interrogation de la base de données, *via* l'interface PubMED. La description des articles peut comporter jusqu'à 15 descripteurs MeSH. Ces termes sont assignés manuellement par des indexeurs et déclinent la ou les thématique(s) de l'article entier ou d'une de ses parties, en complément d'une indexation automatique effectuée sur la base des mots du titre et du résumé, lorsqu'il y en a un.

Les termes MeSH utilisés lors de l'indexation pour décrire un concept de façon univoque ont différents statuts :

- les termes majeurs (précédés de * dans les notices) reflètent la totalité de l'article (syntaxe d'interrogation [MAJR]) ;
- les termes non majeurs sont eux même de 2 types : « obligatoires » (concernant des informations de catalogage génériques sur le document : liés à son sujet — animal, humain, tranche d'âge,. . .— et à son genre —type de publication—) et « non obligatoires », qui concernent seulement une partie de l'article (syntaxe d'interrogation [MH]).

Le MeSH organise également le vocabulaire de référence pour l'indexation dans CISMeF, indexation qui est également réalisée sur la base d'une expertise humaine. Comme nous l'avons vu plus haut, il n'est possible d'associer qu'un qualificatif à la fois à un descripteur MeSH, lors de la description documentaire. Pour exprimer

plusiseurs spécifications, il faut redoubler les associations Descripteur/Qualificatif. Par exemple, pour décrire une maladie induite par une substance chimique, il faut associer les descripteurs et qualificatifs suivants : MALADIE/INDUIT-CHIMIQUEMENT SUBSTANCE-CHIMIQUE / EFFETS-INDÉSIRABLES. Le lien enre INDUIT-CHIMIQUEMENT et son agent SUBSTANCE-CHIMIQUE n'est ni marqué, ni typé, et est laissé à l'interpétation humaine des juxtapositions de Descripteurs/Qualificatifs.

Le thésaurus est intégré dans les deux systèmes de recherche documentaires liés à ces catalogues : PubMed et Doc'CISMeF. Nous les décrivons tour à tour dans la section suivante.

La recherche documentaire dans MEDLINE avec PubMED

Les informations présentées dans cette section sont détaillées dans le tutoriel *Savoir interroger MEDLINE, l'interface PubMED*[29], où elles sont associées à des exercices interactifs.

Le moteur de recherche de PubMED s'appelle « Entrez », et il est commun à un ensemble de bases documentaires de la National Library of Medicine, comme, par exemple, la banque de données sur la génomique GenBank. Il permet de construire une requête avec des termes du MeSH, la sélection terminologique s'effectuant ici aussi par navigation dans les arborescences : celles des 15 catégories de haut niveau détaillées plus haut (les facettes, présentées en section 4.1), celle des lieux géographiques et celle des qualificatifs. Le thésaurus est en effet accessible *via* l'interface PubMED, sous la forme du MeSH Database[30]. Il faut toutefois noter que tous les documents accessibles par PubMED ne sont pas indexés avec des termes MeSH, soit parce qu'ils sont en attente d'indexation (le délai peut varier de quelques jours à plusieurs mois selon les périodiques indexés), soit qu'il s'agisse de documents très courts (par exemple les « letters to the editors ») qui ne sont pas indexés du tout.

Une recherche faite avec un terme MeSH (majeur ou pas) ne renverra pas uniquement les documents indexés avec ce terme, mais aussi tous ceux indexées avec les termes plus spécifiques que ce terme (termes situés « en dessous » de lui dans l'arborescence), par défaut. Cette expansion automatique de requête, appellée « explosion », est désactivable en accolant la spécification « :noexp » au terme sélectionné. Elle peut générer du bruit informationnel, du fait de la non-régularité de la structure arborescente du MeSH. En effet, comme nous l'avons vu, l'arborescence du MeSH est organisée *principalement* selon deux relations sémantiques : l'hyperonymie et la méronymie, mais également selon d'autres types de relations implicites. Rajouter à la requête l'ensemble des termes plus spécifiques revient à la compléter par l'ensemble des hyponymes et meronymes de chacun de ses termes[31], mais aussi à la compléter par un ensemble de termes connexes. Lors d'une requête à propos d'ACCIDENTS, il

[29] `http://web.ccr.jussieu.fr/urfist/biolo/bioguide2/medline/medline.htm`

[30] `http://www.ncbi.nlm.nih.gov/entrez/meshbrowser.cgi`

[31] Cela peut également poser des problèmes, si l'on pose une requête à un niveau spécifique de granularité. Par exemple, si l'on recherche des documents traitant de la vacularisation de la TÊTE, le système retourne également des documents traitant de l'OREILLE ou du CUIR CHEVELU, qui sont les termes subordonnés à TÊTE, et cela peut être considéré comme du bruit.

n'est pas forcément à propos de renvoyer des articles concernant leurs causes. De même PRÉVENTION ET CONTRÔLE se trouvent subordonnés à THÉRAPEUTIQUE, alors que les premières notions visent à l'évitement de la maladie, et la deuxième à son traitement *une fois que la maladie s'est déclarée*. Ces deux descripteurs sont liés à des conceptions symétriques, mais est-il toujours pertinent de les traiter comme une expansion l'une de l'autre ?

On ne peut, en principe, associer qu'un qualificatif à la fois à un descripteur MeSH, mais le moteur de requête est doté d'opérateurs booléens permettant de combiner des associations Terme/descripteur. Il est donc possible de composer les requêtes suivantes : *descripteur/qualificatif1 OU descripteur/qualificatif2*, ou *descripteur/qualificatif1 ET descripteur/qualificatif2*, suivant le sens que l'on veut donner à l'association *descripteur / qualificatif1 / qualificatif2*. La sémantique de cette association est donnée (à la lecture, par un utilisateur humain) par le type des deux descripteurs mis en relation. Si l'on associe plusieurs qualificatifs à un descripteur dans l'interface de requête, ou si l'on « force » une association non prévue par le système entre un qualificatif et un terme MeSH (par la syntaxe : terme [MESH] AND qualificatif [SH]), le système recherchera l'ensemble des documents contenant le descripteur et le qualificatif, mais sans vérifier qu'un lien *entre le descripteur et le qualificatif en question* a été posé lors de l'indexation. Autrement dit, le qualificatif pourra s'appliquer à n'importe quel descripteur indexant le document. Ceci pose un problème pour l'interprétation informatique de la requête, et notamment la comparaison du sens entre la requête et les représentations sémantiques des documents de la base.

Un système a été prévu pour réduire la polysémie lors d'une requête : le « mapping ». Il s'agit du remplacement automatique d'un terme non-MeSH par un terme MeSH lors d'une saisie dans l'interface de recherche. Quand l'utilisateur inscrit un terme qui ne fait pas partie du MeSH, un mécanisme de correspondance lance une recherche d'équivalent dans le thésaurus. Ce « mapping » est également un mécanisme par défaut, que l'on peut désactiver en mettant les termes de la requête entre guillemets. Lorsque la recherche de correspondance ne trouve pas d'équivalent dans le MeSH, elle est lancée itérativement avec les sous-parties du terme de la requête : elle se fait d'abord sur tous les mots et, si aucune correspondance n'est trouvée, le terme de droite est éliminé, et le mécanisme est relancé sur la base des mots restants. Si le « mapping » ne donne aucun résultat, l'utilisateur se voit proposer une liste de termes MeSH proches de ceux de sa requête par leur orthographe, parmi lesquels il peut choisir le terme qui lui convient.

Si le terme de la requête est un terme MeSH, ou si le mapping a permi d'en trouver une correspondance MeSH, son descriptif s'affiche. S'il y a plusieurs correspondances avec le terme saisi, l'utilisateur se voit proposer une liste de termes dans laquelle il peut faire son choix. L'utilisateur a ensuite différentes possiblités :

- attacher un qualificatif adéquat au terme MeSH : les seuls qualificatifs proposés par l'interface sont ceux qu'il est possible, par défaut, d'apposer au terme MeSH en question ;
- restreindre sa recherche aux termes MeSH majeurs ou empêcher l'explosion si le terme possède des termes « fils » (c'est-à-dire de ne pas faire de l'expansion

automatique de requête avec les termes plus spécifiques que ceux de la requête dans le MeSH) ;
- introduire le terme MeSH, avec toutes ses options et restrictions de requête, dans une recherche complexe en sélectionnant l'opérateur booléen désiré : AND, OR ou NOT.

Différents mécanismes permettent également de faciliter la rédaction de la requête ou d'élargir celle-ci. Il existe notamment un système de jokers (*) qui permettent de rechercher des chaînes de caractères commençant par les lettres indiquées dans la requête. Par exemple, *DNA** permet de retrouver tous les mots commençant par *DNA* (ADN en français). Cependant cette requête ne renverra pas *DNA ploidy* : il s'agit d'une expression complexe commençant par *DNA*. Pour retrouver cette séquence, il faut utiliser l'opérateur booléen *ET* dans la requête, et la composer comme suit : *DNA AND ploidy*. Pour forcer la recherche sur l'expression *DNA ploidy*, il faut mettre l'expression entre guillemets et donc saisir « DNA ploidy ». Le fait de saisir l'expression exacte désactive le mapping et l'explosion automatique, comme nous l'avons vu plus haut.

Il est possible d'utiliser des parenthèses pour composer une requête complexe. Sans parenthèses, les opérateurs booléens sont interprétés de gauche à droite. Par exemple *cold AND (vitamin c OR zinc)* spécifie une sémantique différente de *cold AND vitamin c OR zinc*.

Certains mots sont automatiquement exclus de la recherche : les mots-vides ou mots-outils (« Stopwords » en anglais). Il s'agit des mots ne présentant pas d'intérêt dans la construction du sens de la requête (d'après les concepteurs de l'outil) : les déterminants, par exemple. Une liste exhaustive de ces mots-vides est disponible sur le site.

Il existe un certain nombre d'options visant à cibler sa requête. L'outil permet de limiter à recherche à :
- un champ d'indexation particulier
- un type de publication particulier[32] ;
- une tranche d'age spécifique[33] ;
- une certaine antériorité (de 30 jours à 10 ans) ;
- une date ou à un intervalle de dates (soit celle/s de publication de l'article, soit celle/s d'entrée de l'article dans PubMed) ;
- aux références ayant un résumé ;
- une langue de publication[34] ;
- des documents traitant de sujets en rapport avec l'homme ou l'animal, ou traitant d'individus mâles ou femelles ;
- un sujet donné (l'exobiologie par exemple...) ;

[32]Clinical Trial, Editorial, Letter, Meta-Analysis, Practice Guideline, Randomized Controlled Trial, Review

[33]All Infant : birth-23 months, All Child : 0-18 years, All Adult :19+ years, Newborn : birth-1 month, Infant : 1-23 months, Preschool Child : 2-5 years, Child : 6-12 years, Adolescent : 13-18 years, Adult : 19-44 years, Middle Aged : 45-64 years, Middle Aged + Aged : 45+ years, Aged : 65+ years, 80 and over : 80+ years

[34]English, French, German, Italian, Japanese, Russian, Spanish

– des articles anciens appartenant à la base OLDMEDLINE (avant 1966)[35]

L'historique de la recherche est conservé, comme dans TOTEM, et cette fonctionnalité permet de capitaliser une expertise en recherche d'information spécialisée.

La recherche documentaire dans CISMeF avec Doc'CISMeF

Catalogue et Index des Sites Médicaux Francophones, CISMeF a pour but « d'assister les professionnels de santé dans leur quête d'informations, disponibles sur Internet ». Il s'agit d'un catalogue spécialisé référençant les sites médicaux francophones répondant à un critère de « qualité de l'information de santé sur l'Internet (NetScoring), développé en collaboration avec Centrale Santé et APUIS-Santé »[36]. Ce catalogue est le fruit d'un projet initié par le Centre Hospitalier Univesitaire (CHU) de Rouen, débuté en février 1995, date de la création du site Web du CHU de Rouen. Le recensement des sites et des documents se fait sur la base d'une veille quotidienne (de différentes sources de référence et d'Internet en général) et d'une évaluation humaine de la qualité et de la fiabilité des informations véhiculées. Les documents sont catalogués et indexés par des documentalistes, aidées de l'avis de médecins experts pour les cas les plus difficiles. Le catalogage détaille les neuf éléments suivants (au moyen de balises de métadonnées du Dublin Core) : auteur, date, description, éditeur, format, identifiant de la ressource, langue, titre, type de ressource (recommandations pour bonne pratique clinique, conférences de consensus, matériels d'enseignement, rapports techniques), et l'indexation se fait au moyen de mots clés extraits de la traduction française de MeSH réalisée par l'INSERM, toujours au moyen de balises Dublin Core (10 balises sur le jeu des 15 proposées par la norme sont utilisées dans le catalogage et l'indexation). Le catalogue contenait 12 300 documents en juin 2003. Le site est également une source d'information pour le grand public, des formations pratiques à l'utilisation du catalogue ont été données à des associations de malades, par exemple. La notion de genre du document qui figure dans son catalogage permet de sélectionner les contenus adaptés au niveau de spécialité de la personne effectuant une requête. En résumé, « CISMeF utilise deux outils standard pour organiser l'information : le thésaurus MeSH (Medical Subject Heading)[. . .] et le format de métadonnées Dublin Core. ».

4.3 Les limites du MeSH et des systèmes documentaires dans lesquels il est utilisé

Le MeSH est organisé en une hiérarchie mais, tout comme le thésaurus de l'INA, cette structure verticale se décline suivant différentes relations sémantiques. Cette hiérarchie, si elle est, ici encore, tout à fait pertinente pour une utilisation humaine,

[35] A savoir : AIDS, Bioethics, Cancer, Complementary Medicine, Core clinical journals, Dental journals, History of Medicine, MEDLINE, Nursing journals, OLDMEDLINE, PubMed Central, Space Life Sciences, Toxicology.

[36] Cette citation et les suivantes sont extraites de la page de présentation en ligne de CISMeF : *CISMeF, Catalogue et Index des Sites Médicaux Francophones : pourquoi, comment ?*, disponible à l'URL `http://www.chu-rouen.fr/cismef/cismef.html##3`.

peut être problématique pour les traitements automatiques des requêtes. L'expansion automatique de requête avec les termes subsumés peut engendrer du bruit informationnel. La structure « compartimentée » du MeSH permet d'associer explicitement des descripteurs et des qualificatifs, mais la sémantique de leur association n'est pas validée par un système informatique et laisse la porte ouverte à une interprétation large de cette association : les types des qualificatifs à associer aux descripteurs ne sont pas contraints dans le système informatique où ils sont mobilisés, ou plutôt cette contrainte est contournable par une association « forcée ». La sémantique de leur association dépend de l'interprétation humaine des libellés des descripteurs et des qualificatifs, ce qui peut introduire une distance préjudiciable entre la formulation d'une requête et d'une indexation, et entraîner du silence informationnel.

5 Avantages et limitations inhérents au thésaurus et aux systèmes documentaires classiques

Un système de gestion documentaire doit permettre de représenter le contenu sémantique d'un document, ou du moins une interprétation de son sens suivant un point de vue susceptible d'intéresser les usagers du système, de manière aussi peu ambigüe que possible. Un système de gestion documentaire informatisé doit, en plus de l'indexation, permettre de représenter le sens d'une requête de manière à pouvoir l'apparier avec celui des documents de la base.

Les systèmes actuels se fondent sur la structure arborescente des descripteurs pour faire de l'expansion de requête, mais comme le thésaurus n'est pas organisé hiérarchiquement selon une relation sémantique unique et univoque, cela peut poser des problèmes. Il faudrait donc pouvoir baser le système documentaire sur une structure arborescente plus stricte, ou du moins une structure où la sémantique de la relation entre deux descripteurs est définie d'une manière calculable par un ordinateur.

La structure arborescente du thésaurus est un moyen de typer les descripteurs, surtout dans des cas comme le MeSH, où ils sont organisés selon des facettes (les 15 catégories principales listées en section 4.1). Cependant, les instances créées lors de la rédaction des notices, par exemple dans le cadre de la gestion documentaire à l'INA, ne sont pas rattachées à des types exploitables par le système : dans l'exemple que nous avons présenté plus haut, Sandy Casar n'est pas relié à un descripteur COUREUR CYCLISTE, auquel pourrait être attaché l'information sémantique *participe à une course cycliste*, par exemple. Il faudrait donc envisager une description permettrant de typer *tous* les éléments de la description documentaire, tout en les liant par des relations explicites et exploitables par le système. Il est, en effet, particulièrement dommage de ne pas pouvoir associer un nom de personne à un type du thésaurus, à l'INA, comme le montre notamment [Troncy, 2004b] (impossibilité de composer des requêtes génériques sur les types des personnes, comme COUREUR CYCLISTE, par exemple), ou de ne pas pouvoir expliciter de succession

temporelle ou d'agencement spatial (sur lesquels faire des raisonnements) dans les documents audiovisuels.

Certains éléments de description sémantique des entrées du thésaurus sont visualisables *via* le système documentaire, mais ils ne sont pas mobilisables dans des raisonnements informatiques : ils ne figurent que sous la forme de notes textuelles. Les relations sémantiques au libellé spécifié ne sont pas non plus exploitées par le système (en dehors, parfois, des préférences de sélection d'un terme parmi l'ensemble de ses synonymes en contexte) : les liens VOIR AUSSI ne permettent pas une navigation transversale dans le thésaurus.

Il existe des systèmes qui permettent de raisonner à partir de la définition des concepts constituant le vocabulaire d'une description, d'effectuer des raisonnements sur la base des relations sémantiques explicitées entre ces concepts, et d'effectuer des raisonnements sur la base de la structure hiérarchique des concepts. Ces systèmes comprennent notamment les Systèmes à Base de Connaissance, et les structures hiérarchiques conceptuelles en question sont des ontologies.

Nous allons à présent faire une rapide présentation du principe de fonctionnement des Systèmes à Base de Connaissance, des principaux systèmes de Représentation des Connaissances qu'ils intègrent et montrer leur intérêt concret dans le projet OPALES. Nous nous pencherons ensuite plus précisément sur le cœur de ces systèmes : les ontologies, et sur la manière de les construire à partir de corpus. Puis nous détaillerons les spécificités de la construction d'ontologies différentielles pour la description de documents audiovisuels, dans le cadre de ce même projet, qui fut notre premier terrain d'expérimentation. Nous aborderons ensuite les différents éléments sémantiques à mobiliser en corpus pour aider à leur structuration et définition et nous verrons en particulier dans quelle mesure les méthodologies de construction d'ontologies en général peuvent s'appliquer à la construction d'ontologies différentielles. Dans un souci de généricité concernant nos outils, nous appliquons ensuite les traitements envisagés dans le cadre de l'audiovisuel à des corpus spécialisés dans le domaine médical.

Chapitre 2

Systèmes à Base de Connaissance et ontologies

Sommaire

2.1 Systèmes à Base de Connaissance : généralités et application au projet OPALES

Les Systèmes à Base de Connaissance sont des systèmes informatiques qui permettent de représenter des informations à propos d'un domaine de connaissance d'une manière calculables par un ordinateur, et d'effectuer des raisonnements automatiquement sur la base des informations implémentées. Ils constituent la deuxième génération des Systèmes Experts, les premiers ayant pour but de « "capturer" l'expertise d'un expert et de la représenter au sein du système expert de façon à ce que le système se comporte comme l'expert sollicité dans la même situation. Les difficultés à capturer précisément cette expertise, ont amené à parler de goulet d'étranglement. »[Charlet, 2002, p.1]. Les Systèmes à Base de Connaissance proposent alors de séparer les connaissances liées au domaine des raisonnements et des instances particulières [Newell, 1982] (référence extraite de [Bachimont, 2004]). D'autres types d'acquisition des connaissances que la sollicitation directe et exclusive des experts peuvent alors êtres mises en place, comme l'exploitation de corpus textuels par exemple. Les connaissances du domaine sont alors construites et modélisées *en collaboration* avec un ou plusieurs expert(s) du domaine considéré, leur tâche étant allégée par les propositions (tant en termes de « candidats concepts » que de

propositions de structuration) faites par un ingénieur de la connaissance. Ce type de méthode(s) permet notamment à un non spécialiste du domaine[37] de construire une première modélisation, à faire valider par un expert du domaine représentatif des utilisateurs futurs du système.

D'une manière générale, on distingue trois types de connaissances dans les SBC : les connaissances du domaine, les tâches et les méthodes. Tout d'abord, les connaissances du domaine : « *Les connaissances du domaine d'un SBC sont les connaissances relatives au domaine de l'application et nécessaires pour que les méthodes de raisonnement puissent s'exécuter.* »[Charlet, 2002, p.3]. Ces connaissances sont modélisées de manière structurée, et « *exprimées à l'aide de langages ayant une sémantique bien définie* »[Charlet, 2002, p.3]. Il s'agit d'une ontologie : « *Les ontologies sont des systèmes conceptuels destinés à fournir les notions élémentaires à la formulation des connaissances dont on dispose sur un sujet donné.* » [Bachimont, 2004, p.128], et exprimée à l'aide de langages opérationnels standards, par exemple comme ceux définis dans le cadre du Web Sémantique. Nous reviendrons plus en détail sur ce qu'est une ontologie en section 2.2, mais pour la compréhension de cette section-ci, il est utile d'en avoir une idée générale. Une ontologie[38] organise les concepts et les relations pertinents et consensuels dans un domaine pour une application donnée. Un concept renvoie à une classe d'objets d'un domaine, une relation à la manière dont ces classes d'objets sont liées dans le domaine. Par exemple, dans le domaine de l'anthropologie (domaine sélectionné dans le projet OPALES, que nous présentons plus avant en section 2.2.1) et dans le cadre d'une application documentaire, le concept de TOILETTE permet de caractériser tous les documents présentant des bains d'enfant et autres savonnages énergiques, c'est-à-dire qu'il permet de regrouper toutes les instances particulières de TOILETTE dans le domaine considéré (une instance revient ici à un document traitant de l'information représentée par le concept). La relation *est-pratiqué-avec-instrument* permet de lier une TOILETTE au type d'instrument ou de matériel au moyen duquel elle est pratiquée : TOILETTE/*est-pratiqué-avec-instrument*/GANT DE TOILETTE. Les concepts et relations sont organisés de manière arborescente ou hiérarchique[39], suivant la seule relation d'inclusion logique : un concept est subsumé par un autre si et seulement si l'ensemble des instances du premier est compris dans l'ensemble des instances du second (TOILETTE/*est-un*/PRATIQUE D'HYGIÈNE et PRATIQUE D'HYGIÈNE/*est-un*/PRATIQUE), selon la conception extensionnelle largement répandue dans la communauté de l'Ingénierie des Connaissances. Notre travail de thèse se concentre sur la modélisation de ce type de structure.

Après les connaissances du domaine, il nous faut donc définir les tâches et les

[37]Mais dont les compétences se situent du côté de la modélisation conceptuelle, ce qui n'est en général pas le cas des experts consultés.

[38]Nous parlons ici d'une ontologie *régionale*, et non pas d'une structure conceptuelle universelle : elle est univoque et consensuelle dans un cadre applicatif et interprétatif donné.

[39]Le type de modélisation selon lequel on choisit de structurer l'ontologie peut induire une modélisation mutuellement exclusive des concepts, comme dans le cas de l'ontologie différentielle, et donc créer des arbres disjoints, ou peut permettre de créer une structure conceptuelle hiérarchique sous forme de treillis, comme c'est le plus souvent le cas dans la littérature.

méthodes, les deux autres composantes des Systèmes à Base de Connaissance. Ces deux éléments font partie du modèle de raisonnement :

> Les modèles de raisonnement décrivent de façon abstraite le processus de résolution à mettre en œuvre dans un SBC en termes de tâches et de méthodes. Une tâche est une description de ce qui doit être fait dans l'application en termes de buts et de sous-buts. Elle se définit par des connaissances de sortie obtenues à partir des connaissances d'entrée, et ce en fonction des contraintes et ressources disponibles. [...] les méthodes décrivent comment un but peut être atteint en termes d'une série d'opérations et d'un ordre de réalisation.[Charlet, 2002, p.4]

En résumé, les connaissances du domaine sont modélisées au moyen de primitives conceptuelles représentant les objets et relations du domaine, les tâches représentent les buts à atteindre par le système, et les méthodes la manière de les atteindre.

Les Systèmes à Base de Connaissance comprennent donc deux modules principaux : une base de connaissances du domaine dans laquelle sont implémentées les informations propres au domaine, en fonction de l'application choisie (concepts, attributs, relations), et une structure de raisonnement (comprenant les tâches et méthodes). A ces deux modules sont associés une interface d'accès au contenu et une base de faits. L'interface d'accès au contenu permet de mobiliser les primitives sémantiques du domaine pour construire des connaissances et de lancer les méthodes correspondant aux différentes tâches envisagées dans le système, alors que la base de faits permet de stocker les instances créées. Par exemple, pour OPALES, les primitives sémantiques correspondant aux connaissances du domaine (l'ontologie) devaient servir à décrire des documents audiovisuels traitant de la petite enfance d'un point de vue anthropologique. Les descriptions sont réalisées dans l'interface de travail, et elles associent un ensemble de concepts et de relations à une séquence audiovisuelle délimitée dans le flux (à laquelle est donc assignée une pertinence documentaire). Chaque description est stockée dans la base de faits, et il est possible d'y accéder via une interface de requête. Cette interface de requête fait également partie du module « interface », qui permet à l'utilisateur d'interagir avec le système. Elle est liée au module de raisonnement, qui calcule la manière opératoire de rechercher des documents, soit, pour nous, des annotations stockées dans la base. Les raisonnements implémentés vont parcourir l'ensemble des annotations, les comparer à la requête et renvoyer (dans l'interface des résultats) celles qui lui correspondent suivant les modalités de raisonnement définies. Ces différents constituants sont présentés en fonction de leurs interactions (et de manière indépendante à leur implémentation) en figure 2.1.

Ces modalités de raisonnement sont le principal intérêt des SBC, et elles peuvent se décliner en trois types, en fonction de la source du raisonnement.

Trois types de raisonnement dans un SBC Tout d'abord, la relation de subsomption unique qui organise les concepts (auxquels sont liés les instances) et relations de l'ontologie permet de faire un premier type de raisonnement, équivalent à celui proposé à partir du thésaurus : l'association à un concept de l'ensemble des

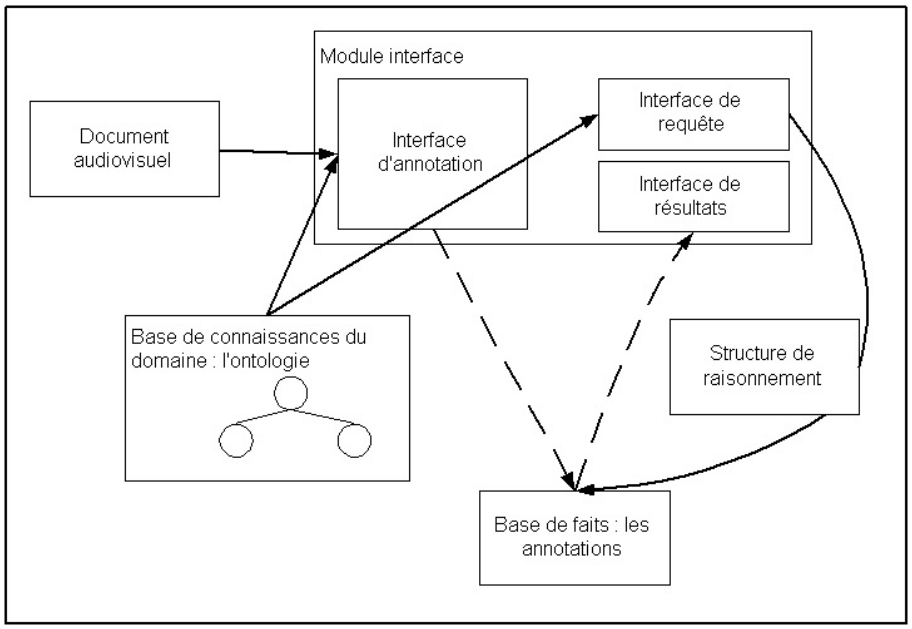


FIG. 2.1 – Interaction des différents constituants du SBC d'OPALES

concepts plus spécifiques, pour faire de l'expansion de requête dans le cas des raisonnements de type documentaire. Comme la relation sémantique entre deux concepts liés de manière hérarchique est unique et univoque, le problème de bruit informationnel que nous avons évoqué à propos du thésaurus n'apparaît pas ici. Les relations sont également organisées selon cette relation d'inclusion : une relation est définie formellement par l'ensemble des concepts qu'elle permet de lier, créant un produit cartésien de concepts (c'est ce que l'on appelle la signature de la relation). Il est donc possible de classifier les relations de la plus générique aux plus spécifiques en tenant compte de la hiérarchie des concepts mis en relation. Par exemple, la relation PRATIQUE/*est-pratiqué-au-moyen-de*/INSTRUMENT est plus générique que celle de TOILETTE/*est-pratiqué-avec-instrument*/GANT DE TOILETTE, vue précédemment.

Le deuxième type de raisonnement est lié à la définition formelle des concepts et des relations. La définition formelle des concepts est donnée « par conditions nécessaires et suffisantes avec des combinaisons booléennes, ou encore par l'expression de contraintes relationnelles » [Troncy & Isaac, 2002]. Ces conditions ou contraintes relationnelles créent un treillis relationnel entre les concepts, et cette propriété permet par exemple de classifier automatiquement de nouvelles données, des concepts définis par rapport aux données primitives entrées dans l'ontologie, dans une hiérarchie de concepts existants[40]. Par exemple, à partir des primitives *personne*, *sexe* et de la valeur *féminin*, on peut définir la classe des femmes et la classe des hommes en tant que, respectivement *Personne pour qui l'attribut sexe a la valeur féminin* et *Personne pour qui l'attribut sexe a la valeur masculin*, et en déduire automatique-

[40]Cette fonctionnalité est notamment implémentée dans des classifieurs, comme celui disponible avec l'éditeur d'ontologie Protégé 2000 REF

ment que le concept de personne devra subsumer à la fois femme et homme. A partir de la définition formelle de FEMME, et de sa place dans la hiérarchie conceptuelle, il est possible de classer automatiquement le concept d'HOMME, sur la base de sa définition formelle propre. Il est également possible de définir que ces deux classes (FEMME et HOMME) sont mutuellement exclusives en associant à la propriété *sexe* que les deux valeurs qu'elle peut prendre, féminin et masculin, forment une partition disjointe. Ce type d'information peut également être mobilisé dans des raisonnement automatiques, notamment pour vérifier la validité des expressions créées au moyen des concepts, relations et attributs.

Enfin le troisième type de raisonnement automatique possible dans un SBC est lié à des règles *ad hoc*, implémentées dans le système et liées au domaine ou à l'application envisagée pour l'ontologie. Par exemple le côté transitif de la localisation spatiale peut être « codé en dur » et servir de base à des raisonnements : si un enfant est dans les bras d'une personne, et que cette personne est dans une pièce donnée, alors l'annotation devra être renvoyée lors d'une requête portant sur un enfant *dans une pièce donnée*, cette information étant calculée par raisonnement à partir des informations initiales et de la règle spécifique (et n'a pas besoin d'être explicitement détaillée dans la description documentaire).

2.1.1 Les formalismes de Représentation des Connaissance

Les Systèmes à Base de Connaissance sont associés à un formalisme de Représentation des Connaissances, qui constitue le paradigme selon lequel les informations peuvent être représentées, comme son nom l'indique, mais également selon lequel les informations sont implémentées dans la base, définissant l'ensemble des opérations possibles sur celles-ci, et enfin selon lequel les systèmes de raisonnement sont opératoires. Il existe deux principaux types de formalismes de représentation des connaissance :

> Dans une approche orientée « logique », le langage de représentation est souvent une variante de la logique du premier ordre, et le raisonnement revient à la vérification de la conséquence logique. Dans les approches non-logiques, souvent basées sur des interfaces graphiques, les connaissances sont représentées au moyen de structures ad hoc, et le raisonnement se fait par des processus également ad hoc manipulant ces structures.[Baader, 2003, p.2] [41]

Les formalismes principaux en sont respectivement les Logiques de Descriptions et les Graphes Conceptuels [Sowa, 1984], [Sowa, 1994]. Nous présentons un survol des principes et modes de fonctionnement des Logiques de Description et des Graphes Conceptuels en section 2.1.1, et développerons ensuite un exemple concret : l'uti-

[41] Notre taduction de : In a logic-based approach, the representation language is usually a variant of the first-order predicate calculus, and reasoning amounts to verifying logical consequence. In the non-logical approaches, often based on the use of graphical interfaces, knowledge is represented by the means of some ad hoc data structure, and reasoning is accomplished by similarly ad hoc procedures that manipulate the structure.

lisation de Graphes Conceptuels et d'un Système à Base de Connaissance pour la gestion documentaire, dans le cadre des expérimentations liées au projet OPALES.

Le formalisme de Représentation des Connaissances des Logiques de Description

La représentation des connaissance suivant le formalisme des Logiques de Description se fait au moyen d'expressions logiques, qui sont souvent intégrées dans un logiciel permettant à un utilisateur non spécialiste de ne pas être rebuté par l'aridité de cette représentation. Un exemple de ce type d'intégration a été réalisé dans le projet GALEN [Rodrigues *et al.*, 1999], un environnement générique de traitement de l'information dans le domaine médical comportant une ontologie représentée selon le formalisme de référence GRAIL. Pour rendre la sémantique des connaissances représentées univoques, les Logiques de Description permettent de définir un vocabulaire au moyen de prédicats primitifs et de constructeurs, dans une syntaxe proche de la logique du premier ordre. Il existe deux types de prédicats primitifs : les prédicats unaires qui représentent des concepts et les prédicats binaires qui représentent des rôles (notion qui correspond à une relation entre concepts). Les concepts correspondent à des ensembles d'individus vérifiant les caractéristiques ou propriétés définies comme nécessaires et suffisantes au concept. Les rôles sont construit par ce que l'on appelle des restrictions de valeur : la sélection du type de concept auquel l'individu lié par la relation doit appartenir. Il est également possible de quantifier la relation entre une instance et une autre. Les concepts dérivés, ou *concepts définis* sont construits et définis à partir des concepts primitifs et de constructeurs. Les constructeurs principaux[42] sont l'intersection de concepts (correspondant à la conjonction de concepts), l'union (correspondant à la disjonction) et le complément (correspondant à la négation). L'intersection permet de définir un ensemble d'instances communes à plusieurs concepts. Par exemple, étant donné un prédicat Personne et un prédicat Féminin, on peut créer le concept défini de « personne féminine » par la construction : Personne INTERSECTION Féminin.

Les inférences, ou raisonnements automatiques, sont basées sur cette définition. Elles calculent principalement la relation de subsomption entre les différents éléments du vocabulaire « primitif » et du vocabulaire construits à partir des concepts primitifs (sur la base des restrictions qui ont permi de les construire). Un autre type de calcul concerne les rôles, et vérifie l'adéquation des types d'instances mis en relation. L'ensemble du vocabulaire est défini au moyen des prédicats et constructeurs dans le composant appelé la TBox, pour Terminological Box. Il y a une définition par terme, et les définitions doivent veiller à être acycliques (un concept ne doit pas être défini au moyen de concepts qui sont eux-mêmes définis par le premier). La TBox contient toutes les données génériques (ou informations de type « universels » : pour tout X correspondant au concept Y, il se définit en tant que...) du domaine et sert de base pour les raisonnements, et notamment pour la vérification

[42] L'inventaire complet des constructeur dépend des variétés de Logiques de Description et de leur implémentation. VERIF

de la consistance dans la hiérarchie du vocabulaire. L'autre composante des systèmes de Représentation des Connaissances basés sur les Logiques de Description est la ABox. Celle-ci contient toutes les connaissances « individuelles »[43] ou de type existentiels, c'est-à-dire toutes les informations liées aux instances particulières du domaine considéré : elle permet de lier les instances aux concepts de la TBox.

Le formalisme de Représentation des Connaissances des Graphes Conceptuels

Selon leur fondateur, John F. Sowa,

> Les Graphes Conceptuels (GC) sont un système de logique basé sur les graphes existentiels de Charles Sanders Peirce et sur les réseaux sémantiques de l'intelligence artificielle. Il expriment le sens sous une forme précise d'un point de vue logique, lisible par un humain et manipulable par un outil informatique.[...] Avec leur représentation graphique, ils servent de langage de modélisation et de spécification lisible tout en étant formel.[44]

Il s'agit donc d'un formalisme visant à représenter des sens d'énoncés, ou plus spécifiquement des concepts, relations et attributs liés à un domaine d'intérêt, de manière à la fois graphique (facilement manipulable par un acteur humain) et formelle (pour être manipulés dans un système informatique).

La représentation des connaissances se fait au moyen de graphes étiquetés et orientés comprenant des nœuds et des arcs. Les nœuds représentent des concepts et les arcs des relations, qui sont tous deux définis dans une ontologie. Cette ontologie est alors qualifiée de « support » pour le graphe. Les concepts sont classiquement représentés sous forme de rectangles, les relations sous forme de flèches orientées, dont le libellé figure dans un ovale (voir figure 2.2). Ce formalisme autorise des relations binaires et N-aires, permet de spécifier l'orientation de la relation ainsi que ses domaines et co-domaines, c'est-à-dire le type des concepts source et cible de la relation (sa signature).

Le processus de raisonnement spécifique aux Graphes Conceptuels est appelé projection, et est notamment détaillé dans les travaux de Michel Chein (par exemple ceux menés en collaboration avec Marie-Laure Mugnier [Mugnier & Chein, 1998]) et David Genest [Genest, 2000], dans le cadre de la théorie des graphes. De manière simplifiée, il s'agit d'une comparaison des concepts et des relations de deux représentations sémantiques : celle d'une requête et celle d'un ensemble de graphes (des annotations décrivant des documents auidiovisuels, dans le projet OPALES). La

[43] Assertional Knowledge, d'où le libellé de ABox.

[44] Notre traduction de « Conceptual graphs (CGs) are a system of logic based on the existential graphs of Charles Sanders Peirce and the semantic networks of artificial intelligence. They express meaning in a form that is logically precise, humanly readable, and computationally tractable. [...] With their graphic representation, they serve as a readable, but formal design and specification language. », extraite du site de John F. Sowa dédié aux Graphes Conceptuels, à l'URL `http://www.jfsowa.com/cg/`.

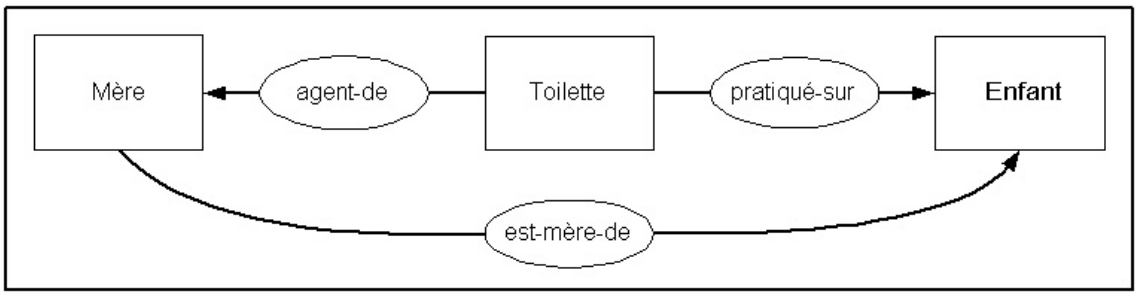


FIG. 2.2 – Exemple de Graphe Conceptuel représentant une mère faisant la toilette de son enfant

comparaison se fonde sur la structure hiérarchique de l'ontologie pour faire l'appariement : s'il n'y a pas de correspondance exacte entre les concepts et relations de la requête et un graphe existant, le moteur d'inférence étend le graphe-requête avec les concepts et relations plus spécifiques que ceux employés dans la structure hiérarchique de l'ontologie. D'autres types d'inférences sont liées aux relations transversales de l'ontologie (celles propres au domaine sélectionné et à l'application envisagée) et sont implémentées dans des règles spécifiques. Nous allons présenter leur intérêt dans le cadre du projet OPALES, dont un des buts était de démontrer les avantages de ce type de systèmes par rapport aux systèmes documentaires classiques, et pour lequel le formalisme choisi a été celui des Graphes Conceptuels.

2.2 Ontologies et ontologies différentielles

2.2.1 Le projet OPALES et la gestion documentaire au moyen de Graphes Conceptuels

Présentation du projet OPALES

Nous allons brièvement présenter le contexte du projet OPALES, puis les règles qui ont été implémentées à cette occasion par Antoine Isaac [Isaac *et al.*, 2004]. Elles montreront clairement l'intérêt de disposer de systèmes de RC pour la gestion documentaire, et donc de la nécessité de construire des ontologies dédiées à ce besoin particulier.

Outils pour des Portails AudiovisueLs Educatifs et Scientifiques, OPALES est un projet[45] centré sur la mutualisation et le partage des connaissances autour de bases documentaires, notamment audiovisuelles. Ce que l'on entend par connaissance consiste en un ensemble d'annotations représentant autant d'interprétations à propos

[45]Projet financé par le Ministère de l'Industrie dans le cadre du projet RIAM : Recherche et Innovation en Audiovisuel et Multimédia, `http://opales.ina.fr`, qui a duré de 2001 à 2003 et réuni l'INA, le LIRMM de l'Université Montpellier II -équipes IHM et Graphes Conceptuels-, le CNDP (Centre National de Documentation Pédagogique), la MSH (Maison des Sciences de l'Homme), la vidéothèque du CNRS-Bellevue et CS-Systèmes d'Information.

d'un document ou d'une partie d'un document audiovisuel. Le fait de segmenter une unité documentaire globale est déjà une indication de sens en soi : il n'existe, en effet, pas d'unité minimale de sens dans un document audiovisuel, et tout découpage revient donc à assigner une cohérence et une pertinence à un segment.

Tous les membres de la communauté d'expérimentation, en l'occurrence un groupe de travail d'anthropologues intéressés par l'analyse des différences dans les types de maternage pratiqués dans différentes cultures et utilisant des vidéos comme documents de travail, peuvent accéder aux différents contenus *via* une interface de visualisation de la vidéo et des annotations. Chaque membre de la communauté peut également créer des annotations propres, complétant ou reprenant les informations déjà présentes dans la base d'annotations, en divergeant aussi potentiellement. Ce processus crée de la valeur ajoutée aux documents « bruts » et permet aux utilisateurs de profiter de l'expérience des autres membres, et de la fédérer autour d'une base de contenus.

Il existe trois manières de faire des annotations dans le système : la classique édition de texte libre, des formulaires dénotant un point de vue sur le document et contraints par cette adhésion à un point de vue partagé, et enfin des Graphes Conceptuels. Le moteur de graphe qui a été choisi pour l'expérimentation est le logiciel CoGITaNT[46], développé par David Genest [Genest & Salvat, 1998]. CoGITaNT permet d'écrire des graphes à partir d'une ontologie support et de faire des raisonnements sur ces graphes, sur la base de la structure (strictement) hiérarchique de l'ensemble du vocabulaire (*i.e.* de l'ontologie support), mais aussi à partir de règles de raisonnement spécifiques à cette application particulière ; dans le projet OPALES, elles ont été développées par Antoine Isaac [Isaac *et al.*, 2004].

Nous allons à présent voir des exemples de raisonnements (ou inférences), simples et plus complexes, qui justifient l'intérêt d'utiliser des Systèmes à Base de Connaissances pour la gestion documentaire avancée. Puis nous nous pencherons sur le support conceptuel fournissant le vocabulaire nécessaire à la construction des connaissances dans ces Systèmes à Base de Connaissances : les ontologies.

Les avantages du Système à Base de Connaissance dans la gestion documentaire

Lorsqu'un utilisateur délimite une séquence audiovisuelle et souhaite y rattacher une annotation sous la forme d'un Graphe Conceptuel, le système propose de naviguer dans l'ontologie pour sélectionner les concepts et relations adéquats, fonc-

[46]Disponible gratuitement sur le site `http://cogitant.sourceforge.net/`, CoGITaNT est un ensemble d'outils logiciels permettant le développement d'applications basées sur le modèle des graphes conceptuels. CoGITaNT est construit autour d'une bibliothèque de classes C++ permettant de développer facilement des logiciels manipulant des graphes conceptuels, et qui offre un grand nombre de fonctionnalités sur les objets du modèle (création, modification, projection, règles, entrées/sorties, etc.). Cette bibliothèque a été utilisée pour construire une architecture client/serveur au sein de laquelle une interface graphique de saisie de graphes conceptuels a été développée. CoGITaNT est une extension de la plate-forme CoGITo développée depuis 1994 dans l'équipe *Représentation de connaissances par des graphes* (anciennement équipe « Graphes conceptuels ») du LIRMM.

tionnalité analogue à celle d'un système de gestion documentaire donnant accès à un thésaurus. Si le graphe ne contient que des concepts, l'annotation est comparable à une annotation par mots-clés d'un thésaurus à ceci près que les concepts regroupent différents libellés d'une même notion, et facilitent la gestion de la variation terminologique. Si le graphe contient des relations, la sémantique de l'association des mots-clés est clairement définie, ce qui donne une indexation plus précise qu'avec une association de mots-clés, et une indexation sur laquelle des raisonnement automatiques vont pouvoir être effectués. Par exemple, une séquence de toilette d'enfant peut être annotée par ses seuls participants : MÈRE et NOURRISSON. Cette annotation revient à associer les différentes lexicalisations des concepts au document : *mère*, *maman*, *nourisson*, *bébé*, et leurs hyponymes éventuels. Une annotation plus complète, MÈRE | *agent-de* | TOILETTE | *pratiqué-sur* | NOURISSON, avec la relation complémentaire MÈRE | *est-la-mère-de* | NOURISSON, explicitant le lien familial entre les deux protagonistes de l'action, permet de renvoyer la séquence annotée à partir des requêtes suivantes :

- MÈRE|*agent-de*|ACTION|*pratiqué-sur*|ENFANT : ACTION est un concept plus générique que la toilette, ENFANT est plus générique que NOURISSON ;
- PERSONNE|*en-relation-familiale-avec*|ENFANT, sur lequel est pratiqué une TOILETTE : *en-relation-familiale-avec* est plus générique que la relation *est-la-mère-de* ;
- ENFANT|*est-patient-de*|ACTION|*pratiqué-par*|MEMBRE DE FAMILLE : l'inverse de la relation *en-relation-familiale-avec*, orientée d'une PERSONNE vers le NOURISSON est rajoutée grâce à une règle implémentée dans la système, et la nouvelle relation crée un lien familial entre le NOURISSON et la MÈRE, qui est un MEMBRE DE FAMILLE.

Les inférences facilitent la gestion documentaire de deux manières : elles permettent de faire des descriptions sémantiques moins exhaustives (ce qui permet de simplifier le travail de l'indexeur, qui peut se concentrer sur l'essentiel de l'information, et non sur la forme ou le contenu lexical de la description effectuée), et de faire des requêtes plus précises, et plus faciles à apparier avec les descriptions sémantiques des documents (ce qui améliore à la fois la précision et le rappel) en faisant de l'expansion de requête basée sur la structure hiérarchique de l'ontologie-support et sur des règles spécifiques. Les travaux de [Schreiber *et al.*, 2001] à propos de l'annotation sémantique de collections de photos (au moyen d'outils basés sur RDF et des schémas RDFs) vont dans le même sens et concluent sur l'intérêt des annotations structurées pour la précision et le rappel en recherche documentaire.

Une autre idée intéressante mise en oeuvre dans OPALES est celle reprenant les grilles d'indexation canoniques présentes dans des systèmes documentaires comme TOTEM (les grilles d'indexation liées aux événements récurrents), et implémentée sous la forme de graphes patrons. Le graphe patron reprend le schéma d'indexation le plus récurrent, c'est-à-dire les concepts les plus génériques les plus fréquemment utiles aux annotations, liés par les relations appropriées (voir figure 2.3). Il est directement accessible pour créer une instance d'annotation. Le graphe patron a été modélisé à partir du corpus de documents audiovisuels, et d'une vidéo particulière qui nous a servi de référence. En effet, cette vidéo est un montage des principaux

documents du corpus audiovisuel, et délimite donc des unités documentaires pertinentes dans ces différents documents. Cette pertinence a été définie en fonction de l'appréciation de spécialistes du domaine puisque la cassette vidéo de référence a été montée en collaboration avec les différents anthropologues auteurs des documents de notre corpus. Il s'agit d'un document de vulgarisation, mais il couvre l'ensemble des thématiques liées à la petite enfance abordées dans notre corpus : la toilette, l'alimentation, les soins, l'habillement, le jeu, l'apprentissage et l'endormissement. Il présente également des pratiques issues des zones géographiques de notre corpus, et insiste sur les moyens mis en œuvre dans les pratiques en question : toilette de l'enfant sur les genoux de la mère ou dans une bassine, au moyen de savon ou d'autres substances, endormissement de bas en haut ou de gauche à droite, dans les bras, dans un pagne accroché sur le dos ou dans un hamac, etc. L'analyse de cette vidéo nous a permis de dégager les éléments récurrents suivants : les personnes susceptibles d'être protagonistes d'une action, les actions (toilette, endormissement, etc.), les types de lieux (public/privé, couvert/ouvert), les types d'objets potentiellement mis en oeuvre dans l'action (hamac, baignoire, biberon, etc.) et les relations associées : est-protagoniste-de, est-patient-de, est-en-interaction-avec, se-passe-dans, se-passe-sur, pratiqué-au-moyen-de, etc... Ces différents éléments ont été proposés aux experts et jugés pertinents (validés) par eux.

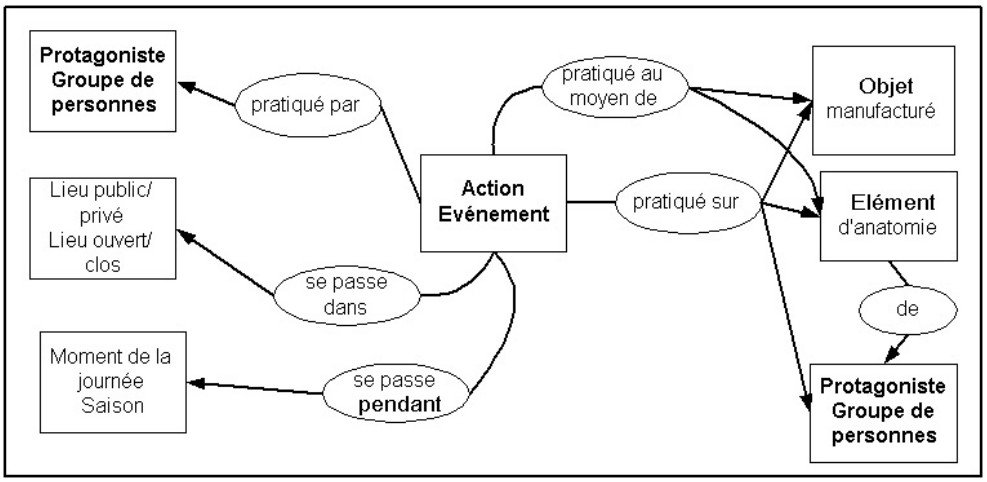


FIG. 2.3 – Le graphe patron Petite Enfance

Le graphe patron peut être complété ou simplifié lors de l'instanciation d'une annotation, mais surtout le système permet d'accéder à la branche de l'ontologie où est classé un concept en cliquant sur sa représentation dans un graphe. Cette fonctionnalité permet de spécialiser tous les concepts mis en relation dans le graphe patron et de créer une instance d'annotation plus spécifique au moindre coût. Elle permet aussi une navigation dans l'ontologie à un niveau de granularité pertinent pour des usagers du systèmes qui ne sont pas experts cogniticiens : la navigation commence à un niveau médian (avec des concepts comme PERSONNE, ACTION, OBJET MANUFACTURÉ), plutôt qu'à la racine, où sont organisées les notions les

plus abstraites, à vocation classificatoire (OBJET TEMPOREL ou OBJET SPATIAL, OBJET ABSTRAIT ou OBJET CONCRET). La navigation depuis la racine de l'ontologie est toujours possible, mais elle peut perturber l'utilisateur en lui demandant de tracer un chemin sémantique parmi des notions qui ne lui sont pas familières, ou lui sembler fastidieuse car les concepts pertinents à l'annotation se trouvent souvent à un niveau de profondeur hiérarchique important.

CoGITaNT permet de valider les annotations créées par rapport au modèle fixé par l'ontologie. Par exemple, les relations devront lier les types de concepts qui ont été prévus à cet effet dans l'ontologie sous peine de générer une erreur. Ainsi, des libellés de relations polysémiques seront désambiguïsés par leurs « contextes conceptuels » : leurs signatures (domaines et co-domaines).

Il est donc intéressant d'utiliser des Systèmes à Base de Connaissance pour des applications ou des domaines spécifiques, comme la gestion documentaire. Pour pouvoir en bénéficier, il faut disposer de structures ontologiques dédiées à ces tâches et domaines. Nous allons donc nous intéresser à la constrution d'ontologies liées à un domaine (des ontologies dites « régionales »), dans le but applicatif de la gestion documentaire. Nous nous intéresserons aux documents audiovisuels dans un premier temps, et ensuite, nous testerons la généricité de notre méthode de construction d'ontologie en l'appliquant au domaine médical. Mais tout d'abord, nous allons développer les notions d'ontologie et d'ontologie différentielle, notamment dans le cadre de ce projet OPALES.

2.2.2 Ontologies et ontologies différentielles

Notre participation au projet OPALES concernait la construction d'ontologies différentielles adaptées aux besoins des usagers du système. Cette expérience a constitué une première approche de notre objet d'étude : les ontologies différentielles, et leur construction à partir de corpus[47]. Nous allons présenter dans cette section un survol de la notion d'ontologie, et un aperçu plus détaillé de celle d'ontologie différentielle, puis nous développerons notre recherche pour leur construction dans OPALES : le matériel textuel et audiovisuel sur lequel nous avons bâti notre modélisation, les demandes et besoins des usagers. Nous présenterons ensuite les méthodes que nous avons testées sur ce corpus pour l'aide à la construction d'ontologies différentielles et les résultats qu'elles nous ont permis d'avoir. Ces expérimentations nous ont aidé à mettre au point notre méthodologie globale.

Comme nous l'avons vu plus haut, une ontologie fournit le vocabulaire nécessaire à l'expression des connaissances dans un Système à Base de Connaissance. En effet,

> Ce n'est qu'une fois l'ontologie définie qu'il est possible d'associer une

[47]Dans le cadre du projet même, nous avons construit une ontologie sur la thème de la petite enfance, du point de vue anthropologique, et une ontologie sur le thème de l'eau pour un groupe d'enseignants s'intéressant au médium de transmission de la connaissance qu'est le document audiovisuel. Nous avons construit ces ontologies à partir de corpus textuels, et cette première expérience nous a servi à fonder nos hypothèses de travail, mais la modélisation a majoritairement été un travail manuel. Nous nous sommes cependant basée sur le corpus textuel construit à cette occasion pour tester nos hypothèses et mettre au point notre chaîne de traitement lors de notre thèse.

> sémantique dans le domaine aux constructions syntaxiques du langage. Il n'est possible de représenter des connaissances, c'est-à-dire d'exprimer dans un langage formel les connaissances du domaine, que si les formules et les primitives non logiques qu'elles contiennent sont pourvues d'une sémantique qui permette de savoir quelle connaissance est assumée par cette formule.[Bachimont, 2000, p.308]

En effet, le système formel permettant d'exprimer les connaissances définit la sémantique des connecteurs du système, et ce qu'il est valide de faire, mais ne donne pas la sémantique des primitives non logiques du langage : le vocabulaire lui-même. Celui-ci doit être structuré et défini de manière indépendante du système, et c'est le rôle de l'ontologie. Nous allons développer ce dernier aspect plus bas : comment définir le vocabulaire conceptuel pour le rendre calculable informatiquement ? Ce vocabulaire, ou les primitives sémantiques du langage de représentation des connaissances, se divisent en deux catégories : des concepts, représentant des types (ou classes) plus ou moins spécialisés d'instances du domaine, et des relations[48] permettant de les associer suivant une syntaxe propre au système de Représentation des Connaissances employé, afin de construire des connaissances complexes. Par exemple, à partir des concepts : MÈRE, ENFANT et TOILETTE, associés aux relations *est-mère-de*, *agent-de* et *pratiqué-sur*, on peut construire une annotation à laquelle est rattachée la sémantique « une mère allaite son enfant », sur laquelle le système peut effectuer des raisonnements (cette annotation correspond à l'exemple de graphe présenté plus haut, il est repris ici figure 2.4).

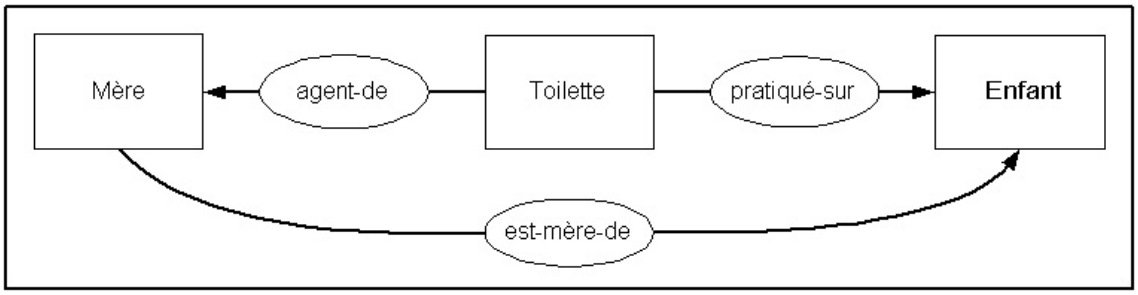


FIG. 2.4 – Un graphe conceptuel interprété comme « une mère faisant la toilette de son enfant »

Définition des termes et des relations dans l'ontologie Nous allons aborder dans cette sous-section les différentes manières d'envisager la définition des concepts et des relations, en rapport avec le type de l'ontologie dans laquelle ils sont intégrés. En effet, il existe différents types d'ontologies, catégorisés dans la liste ci-dessous par [Gomez-Perez *et al.*, 2004, p.9] suivant le degré de formalisation de leur structure et les modalités de définitions de leurs concepts :

[48] Ou des prédicats en Logique de Description. Nous adoptons ici la terminologie propre aux Graphes Conceptuels, le système de Représentation des Connaissances qui a été choisi dans la cadre du projet.

- hautement informelle (exprimée en langue naturelle) ;
- semi-informelle (exprimée selon une forme restreinte et contrainte du langage naturel) ;
- semi-formelle (exprimée dans un langage artificiel et définie formellement) ;
- rigoureusement formelle (donnant une définition méticuleuse des termes, selon une sémantique formelle, des théorèmes et des preuves de propriétés telles que la correction et la complétude).

Les concepts sont définis de manière formelle dans une ontologie semi ou rigoureusement formelle ; ils sont associés à une définition d'ordre plus terminologique dans une ontologie semi-formelle ou informelle. Ils sont également, d'une certaine manière, définis par le treillis de relations qui les mettent en jeu : d'après Ryle [Ryle, 1949] (cité dans [Sowa, 1984, p.15]), « *Le type logique ou la catégorie à laquelle appartient un concept est constitué par l'ensemble des opérations qui lui sont légitimement associées* »[49]. Autrement dit, le type du concept se déduit de l'ensemble des relations qui le lient aux autres objets du domaine.

En dehors de ces trois types de définition, formelle, textuelle et « fonctionnelle » (par les associations sémantiques permises entre un concept et les autres du domaine, soit sa manière de *fonctionner* dans le domaine) pour la définition des concepts d'une ontologie, [Bachimont, 2004, p.139] distingue les trois manières d'appréhender un concept, et donc de concevoir une ontologie :

- *Le concept comme signification* : le concept est une signification normée dont la compréhension correspond à sa reformulation à travers d'autres significations conceptuelles. Par conséquent, le concept s'inscrit dans un système de significations normées. Le comprendre, c'est le distinguer, le différencier des autres concepts en fonction de sa position dans le système de signification.
  A ce niveau, définir une ontologie, c'est définir un ensemble de significations normées.
- *Le concept comme construction* : le concept ne correspond plus à la signification dans la langue, mais à une méthode de construction d'un objet à partir de la donnée d'un divers de sensation. Le concept ne se détermine donc pas à partir d'un système de significations, mais à partir d'un donné de l'expérience, sans relation particulière aux autres concepts possibles. Le comprendre c'est construire l'objet dont il est le concept.
  A ce niveau, définir une ontologie, c'est associer à chaque concept les objets qui lui correspondent et déterminer leur méthode de construction ou de désignation.
- *Le concept comme prescription* : le concept n'est ni une position dans un système de significations, ni une méthode de construction à partir d'un donné de l'expérience, mais il correspond à une suite d'actions à entreprendre. Le concept est alors une prescription : le comprendre c'est l'exécuter.
  A ce niveau, définir une ontologie, c'est associer à chaque concept les actions qu'il faut entreprendre lorsqu'on a compris le concept.

[49] Notre traduction de : *The logical type or category to which a concept belongs is the set of ways in which it is logically legitimate to operate with it.*

Ces trois points de vue sur le concept constituent les trois étapes du processus de construction d'ontologie développé dans [Bachimont, 2000] : la méthodologie pose les jalons pour passer d'une conception conceptuelle à une autre, la dernière étape étant opératoire dans un système informatique. Nous reviendrons plus en détail sur cette méthodologie, et notamment sur l'ontologie différentielle qui en constitue la première étape, en section 2.2.2.

Les relations, quant à elles, sont définies par un domaine et un codomaine, c'est-à-dire par les types des concepts permettant d'être associés par cette relation. Cette association orientée donne une sémantique plus précise à la relation que son libellé textuel, qui est toujours sujet à interprétation. Par exemple la relation *associé-à* peut donner lieu à une interprétation dans le monde du travail (au sens large : entreprises ou collaboration agricole) qui est assez différente de celle qu'inspire la combinaison : RITE|*associé-à*|ETHNOTHÉORIE, et évite les autres emplois envisageables, comme OBJET (de culte/profane)|*associé-à*|RITE (où *associé-à* a la valeur sémantique de *pratiqué-au-moyen-de*). Comme les documents audiovisuels de la base documentaire d'OPALES sont susceptibles de traiter de ces différents aspects, le fait d'ancrer les relations à des concepts permet d'en évacuer une certaine polysémie. Cependant cela n'est pas encore suffisant : la signature d'une relation donne un produit cartésien de concepts, alors qu'une relation spécifique a une sémantique plus spécifique, qui n'est pas donnée par la signature générique. Cette signature permet toutefois de faire des « restrictions de sélections » et d'éviter un trop grand écart sémantique entre une annotation et une requête du fait de l'interprétation humaine de l'association conceptuelle.

En résumé, on peut dire qu'une ontologie regroupe, organise et définit (suivant différentes modalités, mais de manière assez précise pour que la modélisation puisse être opératoire dans un système informatique) les concepts et relations d'un domaine. De quelles ressources dispose-t-on pour trouver ces concepts et relations d'un domaine afin d'en élaborer une ontologie ? C'est à cette question que nous apportons des éléments de réponse dans la section suivante.

Principes de construction d'ontologie

Puisque nous cherchons à construire une ontologie dans le domaine de l'anthropologie, pour fournir le vocabulaire nécessaire à la description, par les utilisateurs du système, d'un ensemble de documents audiovisuels traitant de la petite enfance, nous devons regrouper, organiser et définir les concepts et relations pertinentes à cette tâche, dans le cadre de ce domaine particulier. La question est alors de savoir où trouver ces éléments de vocabulaire conceptuel ? Sur quelle base les sélectionner et suivant quels principes les organiser ?

Plusieurs réponses ont été proposées à la première question, qui oscillent entre l'interrogation des experts et l'utilisation de corpus textuels regroupant les traces linguistiques des principaux éléments de connaissance des experts (solution notamment proposée par le groupe TIA, Terminologie et Intelligence Artificielle [Bourigault & Slodzian, 1999]). Les deux approches comportent des inconvénients.

Ceux liés à l'interrogation d'experts comme source exclusive pour l'élicitation

des concepts et l'organisation d'un domaine sont de deux ordres. Tout d'abord, leur temps est précieux, et les tâches d'expertise dans lesquelles ce type de projet les mobilisent sont, somme toute, annexes à leurs fonctions principales : leur sollicitation dans le processus de construction d'ontologie doit être gérée au plus juste. Ensuite, et c'est sans doute le plus important, ils ne sont pas forcément à même d'exprimer leurs connaissances hors propos, de les formaliser en termes de concepts et de relation, et de dégager des primitives sémantiques du domaine *ex nihilo* (pour plus de détail sur ce point, voir notamment [Bourigault & Aussenac-Gilles, 2003] et [Charlet, 2002]). De plus, chaque expert consulté peut avoir un avis propre sur l'organisation du domaine ou la manière de lexicaliser un concept, ce qui ne facilite pas le travail de modélisation consensuelle qu'est la modélisation ontologique. En effet, les concepts et leur organisation doivent être reconnus et partagés par la ou les communauté(s) d'usagers pour que l'ontologie soit exploitable et, plus généralement, exploitée.

Pour dépasser ces limites, l'une des solutions possibles est de se baser dans un premier temps sur un corpus textuel pour acquérir et structurer les concepts du domaine, et de travailler dans un deuxième temps avec un ou plusieurs expert(s) du domaine. En effet, cela allège l'intervention de l'expert, et le met dans la position plus favorable de *validateur*, plutôt que celle d'inventeur d'un système conceptuel. De plus le corpus regroupe des textes de référence, des documents consensuels et reconnus dans le domaine de spécialité considéré, ce qui donne à la fois une certaine objectivité dans l'acquisition des concepts (acquisition *via* un ensemble de référence plutôt que suivant l'expérience d'un spécialiste) et une certaine neutralité dans la lexicalisation des concepts (lexicalisation non biaisée par les partis-pris éventuel de l'expert sollicité).

Les inconvénients concernant la sélection de concepts à partir de corpus textuel sont toutefois également de deux ordres. Il y a tout d'abord la question de la représentativité du corpus construit qui est à valider : il faut que les textes soient représentatifs du point de vue des utilisateurs du système, et qu'ils couvrent tous leurs besoins terminologiques. Et ce dernier mot fait justement le lien avec le deuxième obstacle à la construction d'ontologies à partir de corpus : le fait qu'il n'y a pas de concepts en corpus, mais des unités linguistiques dont le sens est construit contextuellement. Les concepts ont une sémantique référentielle et extensionnelle, les unités linguistiques non. Les concepts sont des constructions *a posteriori* qui symbolisent une notion du domaine. Une étape semble logiquement faire le pas entre un ensemble de textes et une hiérarchie de concepts, il s'agit de la terminologie. En effet, une terminologie a pour vocation d'isoler, de structurer et de définir des unités linguistiques spécifiques : les termes d'un domaine. Parmi les méthodologies de construction d'ontologie à partir de corpus qui répondent à la question du passage d'unités linguistiques à des concepts et relations représentant les primitives sémantiques d'un domaine, la méthodologie développée par Bruno Bachimont [Bachimont, 2000] propose effectivement de construire une terminologie particulière pour établir ce pont : une ontologie différentielle. Nous avons donc choisi de nous placer dans ce cadre méthodologique pour la construction de nos ontologies.

L'ensemble de cette méthodologie comporte trois étapes (ou processus), et fournit un cadre théorique et méthodologique au passage d'une sémantique linguistique

intensionnelle à une sémantique conceptuelle, extensionnelle et formelle. Nous détaillons dans la section suivante ce qu'est une ontologie différentielle, dans le cadre plus général de cette méthodologie de construction d'ontologie, et nous nous attachons à décrire les éléments sémantiques dont il faut disposer pour permettre d'en modéliser à partir d'un corpus textuel.

Les ontologies différentielles

Les ontologies différentielles constituent la première étape de la méthodologie de construction d'ontologies développée par Bruno Bachimont [Bachimont, 2000], qui définit trois processus pour faire le pas entre un corpus textuel et une ontologie opérationnelle (c'est-à-dire prête à être mobilisée dans un système informatique comme un Système à Base de Connaissances). Ces trois processus sont la *normalisation sémantique*, la *formalisation* et l'*opérationnalisation*.

Normalisation sémantique La première étape de la modélisation ontologique concerne l'exploitation d'un corpus textuel, construit en fonction d'une tâche spécifiée et dans un domaine déterminé. Il faut tout d'abord « *Déterminer les notions élémentaires à partir desquelles toutes les connaissances du domaine sont construites* »[Bachimont, 2000, p.308]. Il ne s'agit pas de définir des primitives universelles, mais bien un ensemble de notions propres à un domaine et à une application dans ce domaine. Cependant, même dans un domaine clairement circonscrit, il n'existe pas de primitives sémantiques en tant que telles. Cette opération de recherche des notions fondamentales du domaine ne revient pas à individuer des concepts pré-existants, selon une conception Wüstérienne de la terminologie, mais à trouver des structures en corpus qui marquent la trace linguistique de l'organisation du domaine considéré, et à modéliser et définir « des primitives pour la résolution du problème ». Nous nous plaçons donc ici dans la lignée des travaux en linguistique de corpus.

Il est intéressant de prendre des libellés en langue naturelle pour exprimer des connaissances, parce que « *L'avantage est que le concept reçoit d'emblée une interprétabilité dans le domaine par les spécialistes qui l'utilisent ou le consultent.* »[*op. cit.* p.309] En revanche, à cause de la polysémie inhérente aux libellés décontextualisés, et de la multiplicité des choix possibles pour le libellé d'un terme (par exemple, un médecin pourra utiliser *histoire de la maladie* ou *antécédents du malade* pour exprimer le même concept) ou d'une relation, « *il est nécessaire de contraindre l'interprétation spontanée que fait tout spécialiste de ces libellés, pour que, respectant ces contraintes, tout spécialiste associe les mêmes significations que ses confrères à un libellé.* »[*op. cit.* p.309]. « *En contraignant l'interprétation effectuée par les spécialistes, la même signification est associée quel que soit le contexte, c'est-à-dire indépendamment du contexte. C'est à cette condition que le libellé, pourvu de cette signification, peut fonctionner comme une primitive et être mobilisé pour la représentation formelle des connaissances.* »[*op. cit.* p.309–310].

La sémantique des termes est fixée selon le paradigme intralinguistique de la sémantique différentielle [Rastier *et al.*, 1994] : il définit une unité linguistique dans un

système d'oppositions et de rapprochements avec les autres unités linguistiques qui constituent son environnement sémantique. Pour passer d'une signification linguistique contextuelle dynamique, selon le jeu des sèmes inhérents et afférents [50], à des primitives sémantiques, l'auteur propose d'opérer une normalisation linguistique :

> La normalisation linguistique est le choix d'un contexte de référence, celui de la tâche ou du problème qui motive l'élaboration d'une représentation formelle des connaissances. Le point de vue de la tâche permet au modélisateur de fixer ce que doit être la signification de l'unité linguistique considérée.[*op. cit.* p.311]

Cette normalisation se fait en construisant un réseau d'identités et de différences sémantiques entre les unités linguistiques, dans le contexte du domaine choisi. Le réseau est organisé suivant des principes issus de la sémantique différentielle et les unités normalisées sont placées les unes par rapport aux autres dans une hiérarchie en fonction des éléments de sens qu'elles partagent et des éléments de sens qui leur sont spécifiques. Leur nouveau contexte interprétatif devient donc cette hiérarchie, et les unités linguistiques acquièrent le statut de termes. Pour chaque terme il faut spécifier les éléments sémantiques qui permettent de le relier aux termes les plus proches et ceux qui l'en différencient. Concrètement, la méthodologie prescrit d'instancier pour chaque terme quatre « principes différentiels » :

- Le principe de similarité avec le père : il s'agit de spécifier quel sème (ou élément de sens) un terme partage avec celui qui lui est immédiatement générique dans la hiérarchie : son père terminologique (et futur père ontologique). Ce principe se rapproche de la fonction de « catégorisation » de la définition aristotélicienne, par genre prochain et différence spécifique, et constitue la partie « genre prochain » ;
- Le principe de différence avec le père : il s'agit de spécifier quel sème un terme a en propre par rapport à son père terminologique, ou hyperonyme. Ce principe se rapproche de la différence spécifique de la définition aristotélicienne.
- Le principe de similarité avec le(s) frère(s) : les autres unités terminologiques comparables à un terme sont celles qui se trouvent au même niveau de généricité que lui. Il s'agit alors de définir en quoi ils sont comparables, pour valider un palier ontologique, et justifier de leur présence au même niveau de hiérarchie. Ce principe crée un axe sémantique décrivant la structure horizontale de la hiérachie, alors que les deux principes précédents en définissaient la structure verticale.
- Le principe de différence avec le(s) frère(s) : il s'agit cette fois de déterminer la place précise de l'unité terminologique sur l'axe sémantique défini précédemment : expliciter le sème selon lequel il peut être différencié des autres termes de la fratrie.

[50] Les sèmes inhérents et afférents sont des particules de sens respectivement associées à un terme par défaut, ou par le contexte ; les premiers peuvent être « inhibés » par ce même contexte et les seconds « activés » lorsqu'ils se réalisent. Les autres types de sèmes sont les sèmes génériques : éléments sémantiques communs à tous les membres d'une classe, et les sèmes spécifiques : les éléments qui distinguent un terme des autres termes de la classe à laquelle il appartient.

L'ontologie différentielle propose donc de déduire la structure termino-ontologique des liens sémantiques de rapprochement et de différence entre termes. L'ontologie différentielle est alors comparable à un dictionnaire local, systémique et arborescent : en effet, le but d'un dictionnaire est de définir les différents éléments *comparables* du vocabulaire d'une langue les uns par rapport aux autres : « *Il s'agit, en somme, de fournir une description qui rend compte de tous les emplois observables d'un mot à une période donnée et qui permet de distinguer ce mot de tout autre mot de la même langue et notamment de tout autre mot sémantiquement apparenté* »[Rebeyrolle, 2000, p.11](repris de [Rey, 1990, p.14]). La structure hiérarchique (verticale et horizontale) se rapproche d'un dictionnaire regroupant des définitions de type aristotélicien : la définition de chaque terme est construite par le parcours de l'arbre, et est donnée par l'addition des similarités et différences explicitées depuis la racine de l'arbre jusqu'au terme considéré.

Il existe un autre type d'axe, qui peut également jouer un rôle important dans la modélisation termino-ontologique, et que l'on peut qualifier de « transversal » : il est constitué par les relations de type transversal, telles que nous les avons définies en section 2 du chapitre précédent. En effet, ces relations participent à la structuration de l'ontologie car elles doivent être ancrées au niveau le plus pertinent de la hiérarchie des termes, et peuvent de ce fait induire des choix dans la structuration hiérarchique. Elles peuvent également fournir des critères objectifs pour choisir entre plusieurs modélisations concurrentes. Cet ancrage relationnel contraint également l'interprétation des termes, mais cette fois de manière plus syntagmatique et fonctionnelle qu'essentielle. La structure transversale se rapproche d'un dictionnaire encyclopédique : il montre l'usage du terme en contexte et son rapport au « monde » dans la construction des connaissances.

Catégorisation des axes sémantiques de similarité entre frères termino-ontologiques Nous avons vu que le principe de similarité entre frères constitue une sorte d'axe sémantique selon lequel les sèmes spécifiques des termes de la fratrie pourront s'opposer. A.J. Greimas [Greimas, 1966, p.22–23] répertorie et analyse différents modes d'articulation sémique qui ont été envisagés dans la littérature. Ces modes d'articulation correspondent aux différents types d'axes sémantiques envisageables dans l'ontologie différentielle. Il peuvent être organisés autour de deux pôles, et se déclinent alors suivant ces deux modalités :

- Présence / absence d'un sème, comme dans la comparaison entre les phonèmes [b](*voisé*) et [p](*non voisé*) ;
- Présence d'un sème / présence de sa négation ou de son opposé, comme dans le rapport entre garçon(*masculin*) et fille (*féminin*).

L'axe sémantique peut également fonctionner de manière ternaire, suivant les oppositions :

- Positif / neutre / négatif, comme dans l'exemple : grand vs moyen vs petit, où le sème « moyen » est considéré par Viggo Brondal [Brondal, 1950] comme n'étant « ni grand, ni petit » ;
- Positif / complexe / négatif, dans les cas où le sème intermédiaire « peut

apparaître comme s et non s » : on (*personnel*) vs il (*personnel et impersonnel*) vs cela (*non personnel*) [Greimas, 1966].

Certaines fratries peuvent toutefois contenir plus de deux ou trois termes. Les articulations sémiques se font alors suivant une modalité « scalaire » : un axe sémantique constitue une ensemble de valeurs possibles, et chaque terme de la fratrie correspond à une valeur individuée sur cet axe. Par exemple, sur l'échelle de la chaleur, les termes d'une fratrie peuvent prendre les valeurs de *froid*, *tiède*, *chaud*, *brûlant*, etc. Certains de ces ensembles sont fermés (ils constituent un découpage exhaustif du genre prochain auquel ils appartiennent), d'autres sont ouverts, à l'image de l'ensemble des nombres compris entre 0 et 1. Un exemple concret de fratrie ouverte est celle des matières grasses qui peuvent être employées dans les soins apportés à un enfant dans différentes cultures. L'axe sémantique qui associe toutes les matières grasses (le karité et autres crème de soin pour la peau, celles pour la conservation des aliments, etc.) est l'usage qui est fait de cette matière grasse (usage thérapeutique, cosmétique, alimentaire,...). Il n'est pas nécessaire de prévoir l'ensemble des valeurs possibles de cet axe sémantique, ce qui est fort heureux, mais il suffit de le définir au moyen des principes différentiels pour, en quelque sorte, réserver une place cohérente à d'autres types de matières grasses que l'on voudrait introduire dans l'arbre hiérarchique. En effet, une manière pratique de construire cet arbre est d'adopter une approche incrémentale, par enrichissements successifs.

La sélection des sèmes impliqués dans l'articulation sémique se fait en contexte, lors de la comparaison de deux unités linguistiques en corpus. Il est reconnu que la plupart des unités linguistiques appartiennent à plusieurs paradigmes, comme le mentionne notamment [Rebeyrolle, 2000, p.17], reprenant [Lipka, 1988] et [Geeraerts *et al.*, 1994]. La sélection des éléments de sens qui permettront de les associer à une fratrie dépend donc du paradigme dans lequel on souhaite fixer l'unité linguistique, lors de la normalisation linguistique (c'est-à-dire lors de la sélection du contexte interprétatif de référence). En effet, ce ne sont pas les mêmes éléments de communauté entre frères qui permettent d'associer du *chocolat* à de la *farine* (des ingrédients de cuisine), à des *marrons glacés* (aliments sucrés) et à un *bouquet de fleurs* (l'ensemble des cadeaux destinés à la maîtresse de maison lorsque l'on est invité à dîner[51]. Créer un axe sémantique consiste à « aligner » un ensemble de sèmes communs à deux unités linguistiques. Lors de la construction itérative de la structure terminologique, il arrive que les nouveaux termes remettent en cause l'axe sémantique défini à un stade antérieur. Cela ne remet pas en cause la validité de la structure précédente, mais est dû au fait que ce ne sont pas les mêmes sèmes communs qui sont activés en présence de ce terme supplémentaire. Par exemple, pour reprendre le chocolat, créer une fratrie *sucre / chocolat* revient à considérer le côté calorique de ce dernier ; si l'on rajoute *farine* à cette fratrie, l'axe de similarité sémantique sera orienté vers les *ingrédients de cuisine* que nous avons évoqué plus haut (un des seuls points communs restant entre ces trois termes étant leur propriété commestible). Il existe plusieurs axes sémantiques possibles entre deux termes, et c'est le choix parmi ces

[51] Exemple donné par François Rastier dans le cadre des séminaires *Textes et ontologie* donnés à l'INaLCO en 2003

possibilités qui est remis en cause lors de la construction itérative de l'ontologie.

Implications de la structuration différentielle Puisque la modélisation terminologique se fait suivant des principes sémantiques mutuellement exclusifs, permettant de positionner chaque terme dans la structure en fonction de son rapport aux autres, la structure résultante est strictement arborescente. Les termes sont donc tous reliés à une racine unique. Entre la racine de l'ontologie et les termes du domaine, il y a un niveau de notions d'une grande généricité, qualifié d'ontologie de haut niveau. Ce niveau regroupe des entités mobilisées en tant que principes de classification, par opposition aux termes propres au domaine, mobilisés dans des descriptions. Certains travaux proposent une unification de cette ontologie de haut niveau, dans un souci d'inter-opérabilité, d'échange d'informations ou de standardisation. Ce type de recherche prend sa source dans la recherche des primitives sémantiques universelles, des dix catégories d'Aristote [Catégories 4, 1b25-2a10][52], aux arbres de Porphyre [Porphyre, 1998] pour arriver aux travaux plus récents de John Sowa [Sowa, 1984] ou de Nicola Guarino [Guarino, 1997], en passant, entre autres, par les recherches de Leibnitz sur le calcul mathématique du sens par compositionnalité à partir d'items primitifs. Cependant, nous avons évoqué plus haut le fait que le rattachement des relations transversales pouvait avoir une influence sur la modélisation hiérarchique de l'ontologie : ce rattachement influence notamment l'organisation des termes qui sont hiérarchisés au plus haut niveau. Nous pensons donc (et suivons, entre autres, l'avis de [Kiryakov *et al.*, 2001]) qu'il n'est pas souhaitable d'avoir recours à une ontologie de haut niveau unique, mais plutôt de la modéliser selon une approche ascendante, en partant des concepts du domaine les plus abstraits dont on dispose. En effet, [Troncy & Isaac, 2002] précisent que les différentes modélisations ontologiques réalisées à l'INA regroupent les mêmes concepts de haut niveau, mais qu'ils sont *organisés de manière différente*. Une telle variation est possible grâce à l'approche différentielle de l'ontologie : nous ne cherchons à modéliser que l'existant d'un domaine, les éléments sémantiques nécessaires à la construction de connaissances nécessaires à une application particulière, pas des primitives sémantiques universelles. Autrement dit, contrairement aux approches de Sowa ou de Guarino, nous cherchons à modéliser des données au plus proche des besoins d'un utilisateur et non pas à représenter la Vérité. Cette approche a été adoptée dans l'ontologie MÉNÉLAS, la première ontologie a avoir été modélisée selon des principes différentiels par l'auteur de la méthodologie, lors du projet du même nom [Zweigenbaum *et al.*, 1994]. Ce choix est d'autant plus pertinent que des concepts de type « haut niveau » constituent différents niveaux de profondeur hiérarchique dans l'ontologie différentielle, pouvant aller jusqu'au 13e niveau hiérarchique selon les termes. Il n'est pas possible d'avoir un niveau de description aussi fin avec une ontologie de haut niveau fixe, de type universelle. Une deuxième et plus forte raison qui nous semble jouer en défaveur d'un haut niveau standardisé est le fait que son application implique une modélisation descendante de l'ontologie : si l'on veut s'en servir, il faut partir des catégories abstraites pré-organisées et essayer

[52]Cette référence comme la suivante provient de [Bachimont, 2004, p.133].

de « ranger » les termes et concepts du domaine sous cette conceptualisation. Cette idée va à l'encontre d'une ontologie régionale, modélisée pour être le mieux adaptée à l'application pour laquelle elle va être mobilisée. Nous privilégions donc plutôt une approche de structuration ascendante des termes du domaine, et nous nous appuyons des travaux concernant les catégories de haut niveau pour organiser les termes les plus génériques du domaine entre eux, afin de les rattacher à une racine unique.

Formalisation Après cette étape de normalisation qui conduit à la structuration terminologique sémantiquement motivée de l'ontologie différentielle, le processus de construction d'ontologie se poursuit par une *formalisation*, qui a pour but d'associer aux termes hiérarchisés une sémantique formelle et extensionnelle, les faisant accéder au statut de libellés concepts.

A partir de la structure de termes interdéfinis, liés par des relations hiérarchiques et transversales typées, et organisés selon une catégorisation de haut niveau, on opère un processus de formalisation. Celui-ci a pour but de rendre cette hiérarchie exploitable par un système formel, et consiste, entre autres, à associer aux termes de l'étape précédente une définition formelle. Ils pourront alors servir de primitives sémantiques dans un Système à Base de Connaissance, et acquièrent le statut de concepts (primitifs).

A ce stade de la modélisation, la sémantique des concepts est extensionnelle et référentielle : un concept représente un ensemble d'individus du domaine sélectionné. Il est possible de créer des concepts complexes, appelés *concepts définis*, à partir de ces primitives, au moyen des opérations possibles sur les ensembles (réunion, intersection ou complémentaire). Il est également possible d'avoir de l'héritage multiple, un individu pouvant appartenir à plusieurs classes différentes, des concepts « mixtes » peuvent apparaitre. Par exemple, si les concepts de *scarification* et de *pratiques thérapeutiques* sont mutuellement exclusifs dans le sens où l'une a pour fonction de marquer physiquement le statut d'une personne et l'autre est partiquée dans le but de soigner, une *scarification thérapeutique* héritera de ces deux concepts. Les axiomes (lois liées au domaine), les règles de raisonnement, les contraintes liées aux relations (arité, contraintes de sélection, etc.) et le rattachement explicite des instances aux concepts sont également implémentés dans cette phase (mais de manière indépendante de l'ontologie, à la manière de pointeurs externes).

Opérationnalisation Enfin, la dernière partie de la méthodologie de construction d'ontologie concerne le passage d'un paradigme formel à un système opératoire dans un langage de programmation. Il s'agit de la phase d'*opérationnalisation*.

L'opérationnalisation consiste en une traduction en un langage informatique de l'ontologie, pour qu'elle soit concrètement manipulable dans un système opératoire. Il existe plusieurs types de langages pour représenter des ontologies « opérationnelles ». Nous ne citerons ici que les plus connus, et notamment ceux qui correspondent à des recommandations du World Wide Web Consortium (le W3C[53]). Il

[53]Ce consortium centralise la réflexion à propos de standardisation autour du Web Sémantique,

s'agit des lanagages, RDFS, OIL et OWL.

RDFS (Resource Description Framework Schema) permet de définir la sémantique de données et de métadonnées utilisées dans des graphes RDF, afin d'en permettre l'échange sur Internet. En effet, une fois la sémantique des composant du graphe RDF posée, celui-ci est utilisable dans différentes applications, et peut être échangé, modifié, etc. selon les besoins d'autres personnes que ses concepteurs propres.

OIL et OWL sont des langages associant l'expressivité des modèles hérités des langages dit de Frame avec la puissance de raisonnement liée aux Logiques de Description. Ils permettent de définir des contraintes et des restrictions supplémentaires par rapport à RDFS, définissant avec une finesse plus grande les données modélisées selon leurs constructeurs.

Ces différents langages sont de syntaxe XML, ou existent au moins sous une forme XML (RDF peut par exemple être représenté selon trois paradigmes différents, dont un format XML). Pour plus d'informations à ce sujet, le site du W3C fournit une documentation abondante.

Dans le projet OPALES, les ontologies support des Graphes Conceptuels était écrites en CGXML, également de syntaxe XML, propre au support de raisonnement CoGITaNT.

Support informatique pour la mise en œuvre opérationnelle de la méthodologie Un éditeur d'ontologie, DOE (Differential Ontology Editor[54]) a été développé par Raphaël Troncy et Antoine Isaac [Troncy & Isaac, 2002] dans le cadre du projet OPALES, pour opérationnaliser et automatiser certains points de cette méthodologie. L'éditeur offre, tout d'abord, des outils de création, modification et suppression de concepts et de relations, une représentation graphique de l'arbre ontologique, et des fonctionnalités de recherche et de navigation dans cet arbre. Il propose une interface ergonomique pour associer à chaque terme sa définition encyclopédique et ses principes différentiels (voir la figure 2.5), mais va, bien sûr, au-delà de ces fonctions d'édition.

D'un point de vue théorique, il permet d'effectuer un clivage concret entre les deux premières phases de modélisation ontologique, celle de l'ontologie différentielle et celle de l'ontologie référentielle (celle obtenue après formalisation de la sructure terminologique), et propose d'exporter les ontologies obtenues dans différents formats standard (RDFS, DAML+OIL, OWL et CGXML), afin d'assurer une interopérabilité avec d'autre éditeurs d'ontologie. En effet, certains éditeurs, comme Protégé2000, implémentent des fonctionnalités permettant de concrétiser les différents aspects de la formalisation de l'ontologie. DOE permet de modéliser une structure ontologique avec un fondement linguistique établi ; les concepteurs ne se sont pas focalisés sur le développement de mécanismes de formalisation qui font la force d'outils déjà existants.

et son site (`http://www.w3.org`) détaille la syntaxe et les fondements théoriques des différents langages que nous allons évoquer.

[54]Disponible en ligne à l'URL `http://ina.opales.fr/~public`.

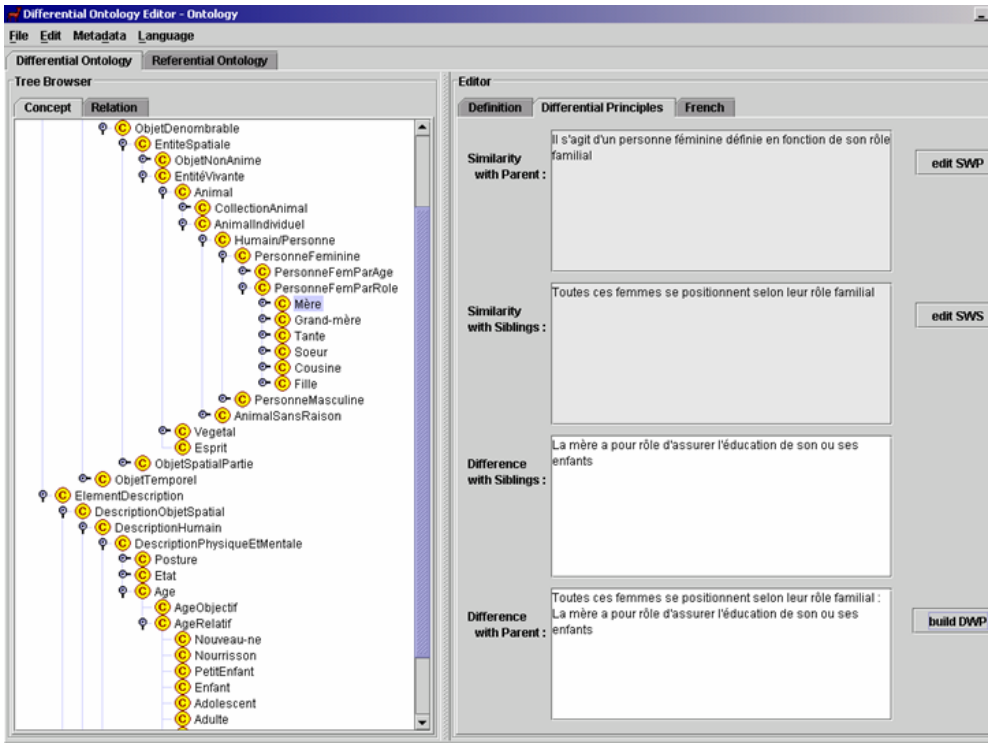


FIG. 2.5 – L'éditeur d'ontologies DOE

Nous déclinons brièvement les apports de l'outils dans les deux étapes de modélisation différentielle et référentielle. Ces deux étapes se réalisent en interaction ; il est donc important d'avoir à la fois accès aux deux volets à tout moment, et de savoir exactement dans quel paradigme on se trouve. Dans la modélisation différentielle, le principal apport de l'outil se situe au niveau des mécanismes liés aux principes différentiels. L'éditeur propage le principe de différence avec le père, une fois qu'il est saisi pour un membre de la fratrie constituant ses hyponymes, à l'ensemble de la fratrie. En effet, il doit être valide pour l'ensemble des co-hyponymes. Un autre principe sémantique partagé par la fratrie est celui de similarité avec les frères, qui est l'axe sémantique selon lequel ils sont comparables. DOE propage donc ce principe à l'ensemble des co-hyponymes. Les fonctionnalités suivantes sont également implémentées :

> Si l'ontologiste veut déplacer une notion dans la taxinomie, les deux principes de similarité de celle-ci sont adaptés à sa nouvelle fratrie. Ensuite, le principe de différence avec le père (DWP) apparait souvent comme la somme de la spécification de l'axe commun aux frères (SWS) et de celle de la différence avec les frères (DWS). [. . .] L'éditeur propose donc de construire lui-même la différence avec le père (DWP) en concaténant simplement les énoncés des principes DWS et SWS. Mais [...] l'utilisateur a tout loisir de changer ce qui ne reste qu'une simple suggestion de l'outil (bouton *build*). [...] On a vu que la modélisation différentielle d'une

> notion permettait de définir celle-ci par l'interprétation de sa position dans la structure hiérarchique globale de l'ontologie. A titre informatif, l'éditeur génère donc, dans l'onglet *Definition*, une *définition différentielle* des notions, en affichant le chemin menant de la racine de l'arbre à la notion considérée et en récapitulant les valeurs des différences avec les pères rencontrées sur ce chemin.[Troncy & Isaac, 2002]

Dans le cadre de la modélisation de l'ontologie référentielle, DOE permet de gérer l'héritage multiple de certains concepts introduits à cette phase.

L'éditeur permet également de définir les signatures des relations organisées sous forme d'arbre dans l'ontologie différentielle. L'arité d'une relation est propagée aux relations qui lui sont plus spécifiques et les domaines et co-domaines sélectionnés au niveau le plus général sont proposés en tant que domaines et co-domaines lors de la création d'une relation plus spécifique.

Récapitulatif

Les ontologies différentielles sont donc la première étape d'un cadre théorique de construction d'ontologie (partiellement mis en œuvre dans DOE) permettant de se baser sur un corpus textuel pour produire des ontologies opérationnelles. Cependant, cette méthodologie ne précise pas la manière de passer d'une collection de textes à une ontologie différentielle : comment choisir et organiser les termes, où trouver les informations sémantiques nécessaires à leur structuration différentielle. C'est précisément sur ces questions que s'est principalement focalisée notre recherche. Des méthodologies et outils de construction d'ontologie à partir de corpus textuel ayant été proposés dans le domaine du Traitement Automatique des Langues (TAL), nous nous sommes intéressée à cette piste et en dressons un rapide état de l'art au chapitre suivant.

Chapitre 3

Etat de l'art sur la construction d'ontologie à partir de corpus textuel et application aux ontologies différentielles

Sommaire

Il existe plusieurs types de méthodes qui sont proposées dans la littérature pour la construction d'ontologies à partir de corpus textuel. Ces méthodes sont basées sur des principes statistiques, des principes linguistiques, ou, le plus souvent, associent les deux. Certains outils implémentent une méthode globale, comme Terminae [Szulman *et al.*, 2002] ou KAON [Voltz *et al.*, 2003]. Nous aborderons la question de ces outils en section 3.3, mais commençons l'état de l'art en matière de construction d'ontologies à partir de corpus textuel par l'ensemble des outils et méthodologies générales susceptibles de remplir l'une ou les deux tâches principales de cette construction. En effet, il est possible de distinguer, suivant la typologie de [Bourigault & Aussenac-Gilles, 2003], deux grandes classes d'outils et de méthodo-

logies, en fonction des deux grandes étapes qu'ils concernent dans le processus global de modélisation ontologique. Les auteurs les déclinent en :

- *Acquisition de termes.* Une première classe regroupe les outils dont la visée est l'extraction à partir du corpus analysé de *candidats-termes*, c'est-à-dire de mots ou groupes de mots susceptibles d'être retenus comme termes par un analyste, et de fournir des étiquettes de concepts. Ces outils diffèrent principalement quant au type de techniques mises en œuvre (syntaxique, statistique, autres).
- *Structuration de termes et regroupement conceptuel.* Les ressources termino-ontologiques se présentent rarement sous la forme d'une liste à plat. Des outils d'aide à la structuration d'ensembles de termes sont donc nécessaires. Dans cette classe, nous évoquerons, d'une part, des outils de classification automatique de termes, et, d'autre part, des outils de repérage de relations. Signalons que beaucoup d'outils d'extraction proposent déjà une structuration des candidats termes extraits.

3.1 Acquisition de termes

Les outils d'acquisition de termes peuvent se diviser schématiquement en deux grands groupes : ceux qui utilisent un étiquetage morpho-syntaxique ou une analyse linguistique complète du corpus, et ceux qui se basent sur des données textuelles brutes.

3.1.1 Acquisition de termes sans analyse linguistique

Les pemiers travaux concernant l'acquisition de termes sur la base de traitement statistique (et sans analyse linguistique) en France datent des recherches autour des segments répétés menés par Ludovic Lebart et André Salem [Lebart & Salem, 1988]. A l'étranger aussi, Choueka [Choueka, 1988] et Hindle et Rooth [Hindle & Rooth, 1990] proposent des systèmes statistiques de repérage de collocations (d'après [Bourigault & Jacquemin, 2000, p.218]). Ces travaux se fondent sur l'idée que des termes polylexicaux d'un domaine auront tendance à aparaître ensemble plus souvent que leurs composantes associées à d'autres mots. Ces travaux exploitent des mesures de similarité pour proposer des associations lexicales récurrentes en tant que termes d'un domaine : « Il existe plusieurs méthodes statistiques appliquées à l'extraction terminologique qui sont pour la plupart fondées sur un principe central, appelé *information mutuelle*. Grosso modo, ce principe veut que l'association récurrente de deux mots ne peut être que le fruit du hasard et est forcément significative »[L'Homme, 2001, p.16]. Un autre type de mesure d'association lexicale communément utilisée est le coefficient de Dice [Velardi *et al.*, 2001]. Ces deux mesures d'association lexicales (Information Mutuelle et coefficient de Dice) correspondent respectivement aux équations suivantes, où E(X) est l'estimation de X, freq(y) est

le nombre d'occurrences d'une expression y en corpus, et w_i et w_j sont des mots :

$$E(M(w_i, w_j)) = log2\frac{Wfreq(w_i \wedge w_j)}{freq(w_i)freq(w_j)}$$

$$Dice(w_i, w_j) = \frac{2freq(w_i \wedge w_j)}{freq(w_i) + freq(w_j)}$$

Ces travaux se focalisent principalement sur les termes polylexicaux. En effet :

> [...]un certain nombre de chercheurs défendent que les unités lexicales complexes sont des suites de mots suffisamment figées pour qu'elles soient identifiées à partir de l'analyse de leur régularités associatives. Ainsi, l'enrichissement linguistique n'est pas considéré comme nécessaire et la seule information générale présente dans les textes doit suffire au repérage des associations lexicales. La communauté scientifique s'est donc défini comme objectif principal de définir des mesures d'association capables de mettre en évidence les attirances entre mots. Suivant ce courant, deux approches ont été privilégiées. D'une part, un certain nombre de recherches proposent la définition de mesures d'association binaires qui font appel aux méthodes d'amorçage pour l'identification de plus de deux mots et d'autre part, un certain nombre d'études proposent la définition de mesures d'association N-aires qui évitent les traitements par amorce.[Dias, 2002, p.26-27].

Le système SENTA (Software for the Extraction of N-ary Textual Associations), de Gaël Dias [Dias, 2002], est particulièrement évolué en la matière, car il permet de trouver des associations lexicales N-aires de mots qui ne sont pas forcément contigus. SENTA fonctionne selon un calcul de probabilités conditionnelles et un indice de « cohésion » lexicale entre différents mots d'une fenêtre de mots mobile : cet indice révèle potentiellement un terme du domaine lorsqu'il atteint un pic.

Un autre type d'approche se focalise sur la forme canonique d'un terme pour identifier des termes polylexicaux potentiels sans analyse linguistique préalable d'un corpus. C'est le principe de l'outil ANA [Enguehard & Pantera, 1995]. Les termes sont reconnus au moyen d'égalités approximatives entre mots et d'une observation de répétitions des patrons (extrait de [Bourigault & Aussenac-Gilles, 2003]). ANA n'utilise pas d'analyse linguistique des textes, mais prend comme donnée d'entrée une description linguistique de ce à quoi ressemble un terme dans la langue considérée, c'est-à-dire des informations concernant la composition canonique d'un terme, et un ensemble de mots représentatifs du domaine considéré. Cet outil fonctionne en deux temps. Lors de la première phase, le cogniticien doit définir les données suivantes :

- une liste de mots-vides fréquents (appelée *stop-list*) ;
- un ensemble de termes représentatifs du domaine, listés manuellement ;
- un ensemble de mots suceptibles de faire partie d'un terme polylexical, comme les prépositions et déterminants en français.

Cet ensemble sert de point d'amorce à l'acquisition de termes du domaine, et fonctionne de manière indépendante à la langue employée. L'acquisition à proprement parler se décline selon trois procédures :

- acquisition de termes simples (composés d'un seul mot) récurrents ;
- acquisition de termes simples sur la base de leur fréquente association avec un terme simple, dont ils sont séparés par un mot de la 3e catégorie de la liste précédente (appelés *scheme words*) ;
- acquisition de termes binaires composés d'un terme simple et d'un mot simple fréquemment associés.

Pour calculer des variantes lexicales, l'outil utilise une mesure de distance entre mots inspirée de [Hall & Dowling, 1980] (d'après [Jacquemin & Bourigault, 2003]).

Dans le même ordre d'idée, suivant une approche de type distributionnelle, les travaux de [Vergne, 2003] se fondent sur des grammaires successives à appliquer au corpus pour identifier des termes potentiels. Un premier ensemble de calculs statistiques et de grammaires utilisent des informations contextuelles sur la succession des mots d'un texte pour définir lesquels sont des mots vides (selon la typologie de [Tesnière, 1959] adoptée dans l'article, il s'agit des mots grammaticaux et des auxiliaires) et lesquels des mots pleins [Houben, 2004]. A partir de cette catégorisation des unités graphiques du texte en mots vides/mots pleins, une autre grammaire décrit des alternances mots-vides/mots pleins susceptibles de définir des schémas de termes. Par exemple : un ou plusieurs mots pleins consécutifs (schématisé $P+$), un ou plusieurs mots pleins consécutifs suivis de un ou plusieurs mots vides consécutifs et un ou plusieurs mots pleins consécutifs (correspondant au schéma : $P+v+P+$) et enfin, la forme la plus complexe : $P+v+P+v+P+$ sont potentiellement représentatives d'une unité terminologique. Le système extrait des termes potentiels centrés sur des mots-pleins indépendemment de la langue et sans traitement linguistique. Cette approche a été mise au point dans le but d'indexer des dépêches Internet multilingues.

Ces différentes méthodes s'attachent à trouver des termes polylexicaux, mais un certain nombre de travaux s'intéressent également à la recherche de termes simples. Ces derniers sont plus difficile à repérer sans analyse linguistique, puisque rien dans leur forme ne permet de les distinguer d'autres unités lexicales. Seules des informations concernant leur fréquence peuvent être utilisées à des fins de repérage. Les termes de spécialité peuvent être considérés comme étant les mots simples les plus fréquents dans un corpus centré sur la spécialité en question (les plus fréquents après les mots grammaticaux) ; les outils conçus dans cette optique présentent alors une liste des mots classés par ordre décroissant de fréquence pour aider à repérer ces termes. Une autre méthode consiste à comparer les fréquences d'apparition des mots dans le corpus de spécialité à leur fréquence dans un corpus de référence, considéré comme neutre (du moins par rapport au domaine de spécialité considéré) : les mots qui ont une fréquence significativement plus élevée dans le corpus de spécialité par rapport au corpus de référence sont alors présentés comme des termes simples potentiels de ce domaine.

3.1.2 Acquisition de termes à partir d'une analyse linguistique

Les méthodes d'acquisition de termes à partir d'une analyse linguistique se fondent sur un étiquetage morpho-syntaxique ou un analyse syntaxique complète du corpus. Elles peuvent se diviser en deux approches : celles qui sélectionnent des candidats termes sur la base d'une structure syntaxique canonique d'un terme (par exemple la succession syntaxique d'un *Nom* et d'un autre *Nom*, ou bien la suite *Nom Prep Nom*) et celles qui repèrent des marques de ruptures de termes, c'est-à-dire des configurations qui ne peuvent pas faire partie d'un terme, pour en donner les frontières (ponctuation forte, pronom ou verbe conjugué par exemple). Nomino (anciennement Termino), ACABIT et OntoLearn sont des outils d'acquisition de termes qui appartiennent à la première catégorie, alors que LEXTER, puis SYNTEX appartiennent à la deuxième. SYNTEX propose, en plus des candidats termes construit à partir de syntagmes nominaux, des candidats termes calculés à partir de syntagmes verbaux et adjectivaux. Très peu d'outils prennent en compte ce genre de regroupements dans l'acquisition terminologique.

NOMINO et ACABIT

[Bourigault & Jacquemin, 2000, p.217] font un inventaire des outils d'acquisition terminologique. Le premier outil « affiché par ses auteurs comme spécifiquement dédié à la construction de bases terminologiques » est TERMINO, de l'Université du Québec à Montréal [David & Plante, 1990].

Nomino, anciennement Termino

> Termino [. . .] est un progiciel d'aide au dépouillement terminologique assisté par ordinateur. Il effectue une analyse morpho-syntaxique du texte fourni en entrée, pour repérer des candidats termes, c'est-à-dire des mots ou séquences de mots susceptibles d'être retenus comme termes par le terminologue. Ainsi, ce système précurseur se distingue de façon caractéristique des travaux classiques de l'époque en extraction de collocations, d'une part parce qu'il met en œuvre un traitement syntaxique, et non statistique, du corpus, et d'autre part parce qu'il intègre d'emblée une interface de validation, dite « module de rédaction des fiches ». Celle-ci permet à un utilisateur humain de sélectionner parmi les résultats extraits automatiquement par le logiciel les unités qu'il juge terminologiques et qu'il va décrire dans sa base de données terminologique.

Nomino, décrit par ses auteurs à l'URL `http://www.ling.uqam.ca/nomino/`, effectue une lemmatisation du texte, une analyse morphosyntaxique et une analyse sémantique :

> Le module sémantique de Nomino permet de repérer un concept sous les multiples formes que la langue autorise pour sa réalisation en contexte : « féminin » renvoyant à « femme » ou encore « carcéral » à « prison ».

> Nomino implémente à ce niveau plusieurs procédures propres à détecter les familles dérivationnelles, compositionnelles, étymologiques et proprement conceptuelles. Il assure par ailleurs le traitement des génériques restreints (hyperonymes morphologiques) et plusieurs types d'équivalents sémantiques, détectés sur la base de relations attestées dans plusieurs ouvrages lexicographiques ou encore par la prise en compte des phénomènes de composition et de décomposition des termes savants.

Nomino est intégré à une interface de création de fiches terminologiques (permettant de joindre des documents, comme les textes sources selon lesquels la fiche terminologique a été créée) et d'interrogation de ces dernières.

ACABIT

ACABIT, développé par Béatrice Daille [Daille, 1994] est un outil mixte associant des traitements linguistiques à des filtres statistiques. L'acquisition terminologique dans ACABIT se déroule en deux étapes : (1) analyse linguistique et regroupement de variantes, au cours de laquelle un ensemble de transducteurs analyse le corpus étiqueté pour extraire les séquences nominales et les ramener à des candidats termes binaires ; (2) filtrage statistique, au cours duquel les candidats termes binaires produits à l'étape précédente sont triés au moyen de mesures statistiques (description extraite de [Bourigault & Aussenac-Gilles, 2003]).

OntoLearn

Dans le même ordre d'idée, la détection de termes selon la méthode de construction d'ontologie OntoLearn [Velardi *et al.*, 2001] part d'une analyse linguistique du corpus (étiquetage morpho-syntaxique, segmentation en phrases, analyse des dépendances syntaxiques) et applique des mesures statistiques concernant la cohésion des candidats termes présentés pour minimiser le bruit lié à leur extraction[55]. Plutôt que l'information mutuelle ou le coefficient de Dice, les auteurs définissent une mesure qu'ils appellent « Pertinence par rapport au Domaine » (« Domain Relevance ») et qui joue sur la comparaison des fréquences de termes potentiels dans le corpus de spécialité avec celle des mêmes ensembles de mots dans différents corpus. Comme nous l'avons vu plus haut, cette approche est aussi utilisée pour détecter des termes simples sans analyse linguistique, en corpus de spécialité. Une autre mesure proposée par les auteurs pour calculer ce « Domain Relevance » est la distribution des termes potentiels : les termes les plus importants ou les plus consensuels du domaine seront les plus répétés et *les mieux répartis d'un point de vue distributionnel* sur l'ensemble du corpus. Un poids plus faible est donc accordé aux termes potentiels issus de

[55]Le bruit est la mesure du nombre de réponses proposés qui ne sont pas correctes divisées par l'ensemble des sorties proposées. Il s'agit d'une mesure d'évaluation courante en sciences de l'information, avec le silence, le rappel, la précision et la F-mesure. Le silence est le nombre de réponses non renvoyés par le système. Le rappel mesure le nombre de réponses correctes fournies par rapport à l'ensemble des réponses correctes que le système devrait retrouver et la précision est calculée en divisant le nombre de réponses correctes fournies par le système par le nombre total de réponses proposées. La F-mesure est une mesure d'évaluation globale tenant compte à la fois de la précision et du rappel, avec la possibilité d'associer un poids différent à l'une ou l'autre de ces mesures.

la première analyse qui ont tendance à n'apparaître que dans un sous-ensemble du corpus, voire dans un seul document.

Lexter et Syntex

Lexter [Bourigault, 1994] extrait dans un premier temps des syntagmes nominaux maximaux, en s'arrêtant aux unités linguistiques constituant des frontières des syntagmes ; ces frontières de syntagmes sont constituées par une ponctuation forte, un verbe conjugué, un pronom ou déterminées selon des informations contextuelles complexes. Ensuite, « Le module acquiert par lui-même, à l'aide d'une méthode d'apprentissage endogène sur corpus, les informations de sous-catégorisation des noms et des adjectifs, propres aux corpus, dont il a besoin pour résoudre les cas d'ambiguïté de rattachement prépositionnel. Par exemple, il sera en mesure d'acquérir l'information que, sur tel corpus, le nom "pression" sous-catégorise la préposition "à", et pourra ainsi extraire les syntagmes "pression à l'aspiration" et "pression au refoulement". » [Bourigault & Jacquemin, 2000, p.229]. Et enfin, un dernier module calcule les dépendances syntaxiques au sein des syntagmes nominaux maximaux et construit un réseau de têtes et d'expansions de ces *candidats termes*[56], termes potentiels à valider dans l'interface HTL. Syntex [Bourigault & Fabre, 2000] est le fruit de la recherche poursuivie à l'ERSS par Didier Bourigault et son équipe après Lexter. Syntex propose également des syntagmes verbaux et adjectivaux comme candidats termes et a intégré des traitement linguistiques plus perfectionnés, notamment dans le cadre de la désambiguïsation du rattachement prépositionnel [Bourigault & Frérot, 2005].

FASTR

FASTR, développé par Christian Jacquemin [Jacquemin, 1994], est un outil d'acquisition terminologique qui repère les variantes en corpus d'une même unité lexicale. Ce regroupement permet d'associer différents libellsé à un seul terme. Le système calcule des variantes syntaxiques, morpho-syntaxiques et sémantico-syntaxiques.

3.2 Organisation des termes

L'organisation des termes peut se faire selon différents procédés.
[Nazarenko & Hamon, 2002, p.12] distinguent deux types d'approches : les approches structurelles et les approches contextuelles. « Les approches structurelles reposent avant tout sur la structure interne d'un ensemble de termes pour les mettre en relation », alors que les approches contextuelles reposent, comme leur nom l'indique, sur des informations contenues dans le contexte des termes pour les mettre en relation.

3.2.1 Les approches structurelles de structuration terminologique

Différents niveaux d'information peuvent être exploités pour structurer une terminologie en se basant sur la structure interne des termes : la syntaxe, la morphologie, la structure lexicale ou des connaissances de type sémantique. La structuration peut alors être décrite selon l'angle du type d'information permettant la structuration. La plupart des approches

[56]Respectivement les recteurs et régis des syntagmes nominaux extraits.

se focalisent sur les relations structurantes, telles que nous les avons vues dans la section traitant du thésaurus : relation d'hyperonymie et de meronymie. Certains travaux s'intéressent toutefois à la synonymie, prise parfois au sens large de l'appariement de variantes lexicales d'un même terme : FASTR que nous avons évoqué plus haut et SynoTerm [Hamon, 2000]. D'autres encore accordent une place de choix aux relations transversales [Grabar *et al.*, 2004].

Structuration de terminologie au moyen d'informations syntaxiques

LEXTER et SYNTEX présentent dans leurs interfaces de validation un réseau terminologique structuré par les dépendances syntaxiques entre composantes des candidats termes complexes extraits. Ces composantes sont les Têtes (recteurs) et Expansions (régis) des syntagmes nominaux (pour LEXTER), nominaux, adjectivaux et verbaux (pour SYNTEX) proposés comme termes potentiels du domaine. Ces dépendances peuvent être des indices de relation hiérarchique entre termes : « un terme *t'* construit par modification d'un terme *t* est généralement présenté comme plus spécifique que ce dernier (*coussin de sécurité arrière* est ainsi plus spécifique que *coussin de sécurité*). » (*op.cit.* p.12).

Structuration de terminologie au moyen d'informations morphologiques

Dans le domaine de la médecine, où un grand nombre de termes sont de formation savante, [Zweigenbaum & Grabar, 2000] proposent de structurer un ensemble de termes d'après des relations déduites de leur structure morphologique. Les auteurs regroupent des termes attestés (issus de thésaurus de référence) lorsqu'ils font partie de la même famille morphologique, et partagent donc des éléments de sens. Par exemple, « symbio*se* » et « sybio*tique* » sont rapprochés parce qu'ils partagent une chaîne commune répondant à un certain nombre de contraintes définies par des règles successives. FASTR de [Jacquemin, 1994] a été conçu dans l'objectif de regrouper des synonymes ou variantes lexicales en corpus d'un terme donné à partir de calculs sur la dérivation et des règles morphologiques.

Structuration de terminologie par l'analyse de la structure lexicale des termes

Cette approche, développée notamment dans [Zweigenbaum & Grabar, 2000] à propos de la relation d'hyperonymie, a été également explorée par Fidélia Ibekwe San Juan, notamment dans [Ibekwe-SanJuan, 2005]. Cette dernière expérimentation concerne toutes les relations sémantiques véhiculées par l'inclusion lexicale, et se fonde sur l'idée que l'inclusion de termes dans d'autres termes plus complexes peut dénoter une relation sémantique entre eux. [Ibekwe-SanJuan, 2005] distingue différents types de relations sémantiques régulières entre termes suivant le type de l'inclusion (inclusion à gauche, à droite, à gauche et à droite) et la catégorie morphosyntaxique de la partie incluse (nom ou adjectif). Par exemple, l'auteur a constraté que « l'expansion gauche et l'insertion induisent une relation d'hyponymie. Elle [l'expérimentation] a mis en évidence également le fait que l'expansion droite et gauche-droite induisent une relation d'association ».

Structuration de terminologie au moyen de connaissances sémantiques

Certaines approches utilisent des connaissances sémantiques pour inférer des relations sémantiques entre termes de la ressource à traiter : [Hamon, 2000] se sert de synonymes

extraits d'un ensemble de dictionnaires pour inférer des liens de synonymie en corpus.

3.2.2 Les approches contextuelles de structuration terminologique

Ces approches utilisent le contexte des termes pour en proposer une structuration. Il est, une fois encore, possible de distinguer deux grands courants de méthodes : les approches de type distributionnel et les approches par patrons lexico-syntaxiques.

Structuration terminologique par approche distributionnelle

De manière générale, l'analyse distributionnelle se base sur l'observation en corpus de la redondance des contextes autour de certaines occurrences de ces termes, c'est-à-dire sur leur régularité distributionnelle. La distribution d'une unité linguistique est, en effet, l'ensemble des contextes dans lesquels elle apparait, et les éléments à distribution identique ou comparable forment des classes distributionnelles, qui sont suceptibles de partager des éléments de sens selon le postulat du linguiste américain Leonard Bloomfield. L'analyse distributionnelle est née dans les années 1950, et a été mise en œuvre et popularisée par Zellig S. Harris [Harris, 1968]. Les différents travaux autour de l'analyse distributionelle se positionnent selon les paramètres suivants :

- la définition des contextes pris en compte (dépendances syntaxiques dans LEXICLASS et le module UPERY, fenêtres de mots dans des approches statistiques comme celle abordée dans [Curran & Moens, 2002], par exemple, ou texte entier dans le cas de [Crouch, 1988] et de [Sanderson & Croft, 1999]) ;
- le calcul de la force d'association entre un terme et un contexte ;
- l'algorithme de classification selon lequel les termes considérés comme proches sont regroupés (K plus proches voisins, clustering hiérarchique ascendant, descendant, réseaux de dépendances syntaxiques,...).

Le plupart des approches considèrent les contextes « graphiques » (fenêtre de mots ou document entier) comme contexte, et utilisent des méthodes statistiques pour rapprocher des termes (c'est-à-dire pour calculer la proximité entre deux termes et pour les regrouper selon ce calcul de proximité). Certains outils permettent de sélectionner le paramètre selon lequel les termes doivent être rapprochés (calcul de la matrice de similarité ou de vecteurs de mots), le type de calcul de similarité que l'on souhaite employer, et l'algorithme de classification selon lequel on souhaite que les termes soient organisés (clustering hiérarchique ou non, par exemple). Mo'K [Bisson & Nédellec, 2001], par exemple, est

> un atelier configurable d'aide à la conception d'algorithmes de classification conceptuelle utilisable dans le cadre de l'apprentissage d'ontologies. Il vise à faciliter le travail exploratoire que doit effectuer tout concepteur d'algorithme pour rechercher les méthodes d'apprentissage les plus aptes à accomplir une tâche donnée. Mo'K fournit non seulement des services permettant d'évaluer, de comparer et de caractériser les méthodes de classification, mais il offre également une approche générique permettant une implémentation aisée des mesures de similarité et opérateurs de classification proposés dans la littérature.

D'autres plateformes existent, comme CLUTO (disponible à l'URL `http://www-users.cs.umn.edu/~karypis/cluto/download.html`), Lexico3 (développé par André Salem et

son équipe, disponible pour des fins de recherche à `http://lexico3.no-ip.org/`) ou l'atelier logiciel « Data and Text Mining » (DTM, de Ludovic Lebart, `http://egsh.enst.fr/lebart/`), mais elles nécessitent souvent des connaissances plutôt poussées en statistique, que ce soit pour leur utilisation ou pour l'analyse des résultats. C'est également le cas pour l'atelier Mo'K, et pour les outils qui proposent une visualisation des résultats selon des modalités statistiques (analyse en composantes principales, analyse factorielle, etc.).

Des auteurs comme Didier Bourigault (avec UPERY), Houssem Assadi (avec LEXICLASS) et Benoît Habert (avec Zellig) s'intéressent, eux, aux dépendances syntaxiques récurrentes comme contextes partagés, à la place de l'observation des contextes graphiques simples.

Dans LEXICLASS [Assadi & Bourigault, 2000], s'appuyant sur les données de l'extracteur terminologique LEXTER [Bourigault, 1994], ou dans le module UPERY de la suite SYNTEX-UPERY [Bourigault, 2002], les termes fournis par la première analyse (par LEXTER ou SYNTEX) sont regroupés en fonction des dépendances syntaxiques qu'ils partagent. Ces dépendances peuvent être des « expansions » communes ou des recteurs communs, du fait de la symétrie de cette analyse (elle rapproche les recteurs en fonction de leurs dépendants partagés et également les dépendants en fonction des recteurs qu'ils partagent). L'ensemble des contextes syntaxiques partagés forme un réseau de « voisins en Tête et en Expansion ». Par exemple, dans le cadre d'un corpus de comptes rendus d'hospitalisation[57], le candidat terme *Cure de chimiothérapie* a comme voisins en expansion (c'est-à-dire a comme recteurs partagés) : *Examen*, *Navelbine*, *Cisplatine*, *Doxorubine*, *Taxotere*, *cCarboplatine* et *MIP*. Ces recteurs partagés, à l'exculsion d'examen, forment un groupe sémantiquement homogène : il s'agit de principes actifs thérapeutiques. Les voisins en tête de *Cure de chimiothérapie* sont : *Traitement*, *Bilan*, *Injection*, *Antibiothérapie* et *Radiothérapie*. Cette classe comporte un hyperonyme et deux co-hyponymes de *Chimiothérapie*, respectivement *Traitement* et les deux candidats termes *Antibiothérapie* et *Radiothérapie*. Le logiciel Zellig [Habert, 1998] effectue un regroupement selon un normalisation de ces dépendances et propose une visualisation des résultats sous forme de graphe : les nœuds sont les termes qui partagent des dépendances, et les arcs sont constitués des dépendances et de leur type.

Structuration terminologique par patrons lexico-syntaxiques

L'autre grande méthode de structuration de terminologie fondée sur les contextes des termes est basée sur la définition *a priori* d'une relation sémantique, par exemple l'hyperonymie, 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 basent sur un marqueur, ou pivot (une unité linguistique qui peut être un indice d'une relation lexicale, comme *entre autres* pour la relation d'hyperonymie) et un ensemble de contraintes que le contexte lexical et/ou syntaxique de ce pivot doit remplir pour que les unités lexicales en contexte soient considérées comme potentiellement liées par la relation sémantique voulue. Par exemple, dans le cas de l'hyperonymie et du

[57]Nous présentons ce corpus plus en détail et l'exploitons dans le cadre d'une expérimentation au chapitre 6

marqueur *entre autres*, il faut que la forme sytaxique 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'hyperonymie mettent en relation des couples père-fils potentiels, qui sont intéressants pour la structuration hiérarchique d'une ontologie.

3.3 Outils ou plateformes logicielles intégrant et implémentant une méthodologie de construction d'ontologies à partir de corpus

Comme nous l'avons mentionné dans l'introduction de ce chapitre, il existe à côté des grandes familles de méthodes pour la construction d'ontologies à partir de corpus des « suites logicielles » intégrant une méthodologie globale. Parmi ces suites, nous nous intéressons ici à celles qui partent du traitement d'un corpus textuel, pour arriver à une structure ontologique opérationnelle. Cette condition étant relativement restrictive, nous développerons deux exemples représentatifs de ces « méta-outils » : KAON et Terminae. Le premier a été développé (et est encore en cours de développement) conjointement par l'Institut AIFB de l'Université de Karlsruhe et le FZI Research Center for Information Technologies (Karlsruhe), pour traiter de l'anglais, et le second, dédié au traitement du français, a été mis au point par Sylvie Szulman et Brigite Biébow au LIPN (Laboratoire d'Informatique de Paris-Nord).

3.3.1 KAON

KAON, sigle correspondant à *KArlsruhe ONtology*, est un environnement de construction d'ontologie qui associe différents modules, correspondant à autant de phases identifiées de l'ingénierie ontologique : création, stockage, raffinage, exploitation, maintenance et application d'ontologies. Dans l'ensemble de ce processus, le module qui nous intéresse est celui de « création ontologique » : TextToOnto. Ce module applique des « stratégies de fouille de texte à des corpus textuels pour la création semi-automatique d'ontologies »[58]. KAON est associé à l'éditeur d'ontologie OntoEDIT, dans lequel il est possible de poursuivre la construction de l'ontologie amorcée avec le module TextToOnto. Celui-ci inclut différents traitements : extraction de termes, extraction de relation conceptuelle et algorithme d'« élagage ». Les termes potentiels sont extraits de manière statistique ou au moyen d'expressions régulières définies par l'utilisateur. Les expressions régulières peuvent porter sur un étiquetage morphosyntaxique du corpus réalisé par Qtag. L'extraction de relation se fait sur la base de patrons lexico-syntaxiques, d'un calcul de proximité ou des deux (selon le choix de l'utilisateur). L'élagage propose de retirer des termes de la structure à partir d'un calcul de leur fréquence. Ce module permet également de déduire une ontologie propre à un domaine à partir d'une structure terminologique générique, comme WordNet.

[58]Notre traduction de : TextToOnto supports semi-automatic creation of ontologies by applying text mining algorithms, extrait du manuel d'utilisation de KAON (TextToOnto – A paper for end users), disponible à l'URL `http://www.architexturez.net/FILES/archive/sub.gate.archive/kaon/TextToOntoPaper.pdf`.

3.3.2 Terminae

Terminae est un outil de création et de gestion de terminologie ou d'ontologie, soit plus généralement de ressources termino-ontologiques. Il intègre un environnement d'étude terminologique, un environnement d'aide à la conceptualisation et un système de gestion d'ontologies. L'outil aide le cogniticien à conceptualiser les objets d'un domaine à travers l'étude d'un corpus décrivant le domaine. Cette étude se fait à l'aide de résultats d'outils de Traitement Automatique des Langues visualisables dans Terminae : ceux de LEXTER et/ou SYNTEX, développés par Didier Bourigault ainsi que ceux de SynoTerm développé par Thierry Hamon au Laboratoire d'Informatique de Paris Nord (LIPN). En plus de la possibilité de visualiser ces résultats, Terminae intègre un concordancier et un module de recherche de patrons lexico-syntaxiques dans un corpus étiqueté par Cordial Analyseur[59] : Linguae. Un certain nombre de patrons sont prédéfinis, mais l'interface laisse aussi la possibilité d'en créer de nouveaux. Les différents outils linguistiques d'analyse de corpus accessibles dans Terminae, ou ceux dont les résultats sont visualisables dans l'outil, ont pour but d'aider une personne chargée de modéliser une ontologie à créer des fiches terminologiques à partir des connaissances présentes en corpus. Ces fiches contiennent des champs par défaut (concept auquel se rattachent les termes, synonymes, polysèmes, voir aussi), qu'il est possible de compléter ou personnaliser sous la forme de nouvelles paires attribut/valeur associées au terme en cours de description. L'outil permet de structurer ces termes en réseau et de les formaliser. Le réseau conceptuel résultant peut être visualisé à la fois de manière graphique ou sous forme de hiérarchie dans un éditeur d'ontologie. Le langage de représentation des connaissances utilisé dans Terminae repose sur un formalisme proche des logiques de descriptions, dont les mécanismes d'inférence aident à structurer l'ontologie et à en maintenir la cohérence lors de l'insertion d'un nouveau concept. L'éditeur d'ontologies permet l'importation et l'exportation d'ontologies au format OWL, correspondant à la dernière recommandation du W3C.

Mais au-delà de ces fonctionnalités, Terminae est surtout intéressant du fait qu'il instrumente un processus de construction de ressources termino-ontologiques motivé linguistiquement. Ce processus a été décrit dans [Aussenac-Gilles *et al.*, 2000], et se décline schématiquement en quatre phases, que nous détaillons plus bas. Le modèle des données de Terminae différencie des termes (tels qu'ils apparaissent dans le corpus), des notions (décrites dans des fiches terminologiques), des concepts (décrits à l'aide de fiches de modélisation) et des concepts formels (décrits en logique de description). Ces différentes représentations (figure 3.1) sont utilisées au cours du processus de structuration puis de formalisation des connaissances, chacune rendant compte du résultat de la modélisation à une étape donnée.

A partir de l'observation des occurrences d'un terme en corpus, par exemple au moyen de la visualisation des résultats fournis par LEXTER, l'utilisateur crée une fiche terminologique, associant ce terme à une notion. Chaque notion est décrite à l'aide des informations suivantes : les synonymes et les termes proches, l'ensemble des occurrences associées, une définition et éventuellement des informations lexicales. Cette notion est normalisée pour être représentée dans un réseau conceptuel par un concept (c'est-à-dire, pour les auteurs de l'outil, par un signifié normé) dénoté par le terme et relié à d'autres concepts par des relations sémantiques. Ces relations permettent de structurer le réseau conceptuel ; elles sont définies à partir de relations lexicales trouvées dans le corpus. Une terminologie dans TERMINAE est donc composée de l'ensemble des fiches terminologiques et du réseau

[59] De la société Synapse-développement `http://www.synapse-fr.com/`

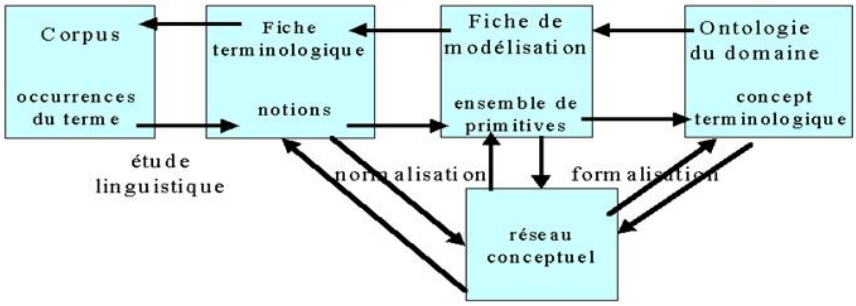


FIG. 3.1 – Les différentes représentations des connaissances dans TERMINAE au cours d'une modélisation

conceptuel. Grâce aux commentaires associés aux notions, il est possible de conserver des particularités d'usage des termes associés. Grâce aux liens qui subsistent des notions vers les termes dans les textes, il est possible

- de rendre compte de phénomènes comme la synonymie (deux termes associés à la même notion) ou la polysémie (un même terme associé à deux notions) ;
- de conserver une trace des raisons qui ont conduit à organiser la terminologie d'une certaine façon, en renvoyant aux extraits du texte qui justifient ces choix.

La dernière étape concerne la formalisation éventuelle de ce réseau. Elle permet de vérifier et de valider la cohérence formelle de la structuration obtenue grâce aux étapes antérieures. Les auteurs se placent dans le cadre d'une démarche peu automatisée : la personne chargée de la construction de la terminologie joue un rôle fondamental pour choisir le contenu de la terminologie et sa structuration à partir de l'interprétation des textes, des résultats des logiciels de TAL et des besoins applicatifs[60].

3.4 Besoins spécifiques liés à la construction d'ontologies différentielles

Nous venons de survoler un ensemble de méthodes et d'outils mis au point dans le domaine du Traitement Automatique des Langues pour la construction de ressources termino-

[60]Cette dernière remarque et une partie du texte ci-dessus est adapté de la présentation donnée par les auteurs de l'outils aux Journées de l'Atala, disponible à l'URL `http://atala.biomath.jussieu.fr/je/010310/Aussenac_Biebow_Szulman.resume.html`.

ontologiques à partir de corpus textuels. Ces méthodes et outils apportent des aides automatiques aux étapes de modélisation suivants :

- recherche de termes du domaine ;
- structuration de ces termes en fonction de relations sémantiques trouvées en corpus ou dans des ressources exogènes ;
- éventuellement définition et formalisation de ces termes.

Pour la construction d'ontologies différentielles, il est indispensable de trouver, en plus des relations sémantiques qui lient les différents termes, les éléments primitifs de sens qui permettent de distinguer un terme de ses voisins : les principes différentiels. Pour chaque terme, on doit donc rechercher en corpus les éléments ou informations suivantes :

- son hyperonyme (son « père ontologique ») ;
- son ou ses co-hyponymes (ses « frères ontologiques ») ;
- le principe sémantique selon lequel on peut le rattacher à son hyperonyme, et celui ou ceux qui l'en différencient ;
- le principe sémantique selon lequel on peut le rapprocher de ses co-hyponymes, et celui selon lequel on peut l'en différencier.

Il nous faut donc tester l'adéquation des méthodes et outils présentés plus haut à la construction d'ontologies différentielles. Les suites d'outils comme KAON, ou celle associée à l'éditeur d'ontologie Protégé 2000 (qui inclut également un plug-in de traitement de texte) sont majoritairement dédiées au traitement de l'anglais. De plus, elles sont élaborées selon une méthodologie qui vise à optimiser le cycle de vie d'une ontologie, mais ne permettent pas de faire une distinction nette entre un terme et un concept. L'étude de [Isaac, 2001] a montré la nécessité de créer un éditeur d'ontologie spécifique pour la création d'ontologies différentielles, les éditeurs existants, même modulaires, n'étant pas adaptés à cette tâche particulière.

Conçu pour traiter du français et pour créer une ontologie selon des principes linguistiques, TERMINAE aurait pu convenir comme plateforme de conception d'ontologie différentielle. Cependant, l'interface ne permet pas de gérer de manière simple les principes différentiels à associer aux différents termes et la « définition systémique » qui en découle. Elle permet de décrire les termes indépendamment les uns des autres, mais les principes différentiels consistent en des rapports sémantiques *entre termes*. Le module de recherche de patrons lexico-syntaxiques s'appuie sur un format de Cordial qui ne correspond pas à nos besoins. Nous avons également utilisé LEXTER pour calculer des candidats termes , mais le format de LEXTER sur lequel nous travaillons ne correspond pas non plus aux requis de Terminae.

Pour ces questions pratiques, mais surtout parce que la gestion semi-automatisée de la construction de principes diférentiels est gérée de manière plus adéquate dans l'éditeur DOE, nous avons choisi de tester les méthodes de construction d'ontologie issues du TAL, mais de ne pas nous servir des suites logicielles disponibles. Nous avons opté pour des programmes indépendants pour créer une maquette du passage d'un corpus étiqueté à un éditeur d'ontologie.

Deuxième partie

Expérimentations et mise au point de la méthodologie de construction d'ontologie différentielles

Chapitre 4

Expérimentation des méthodologies de construction d'ontologie sur le corpus Petite Enfance d'OPALES

Sommaire

Nous présentons dans cette partie les différentes expérimentations que nous avons menées afin de définir une méthodologie de construction d'ontologies différentielles à partir de corpus textuel, selon des procédés éprouvés dans le domaine du Traitement Automatique des Langues (TAL). Une ontologie différentielle est une terminologie organisée de manière

strictement arborescente, autorisant toutefois des relations transversales entre termes de différentes branches (qui sont tous organisés selon une hiérarchie stricte), et dans laquelle la position de chaque terme est définie selon le rapport sémantique qu'il entretient avec les autres termes. Ce « rapport sémantique » constitue sa définition différentielle, et permet de le classer dans la structure terminologique. Pour construire une telle structure, il faut éliciter, organiser et interdéfinir les différents termes propres au domaine que l'on souhaite modéliser. Nous avons choisi comme source d'information les traces linguistiques des connaissances contenues dans des textes de référence : des corpus spécialisés. Nous présentons le corpus que nous avons compilé pour la construction d'une ontologie dédiée à l'indexation de documents audiovisuels selon le point de vue d'anthropologues s'intéressant à la thématique de la petite enfance, lors de notre participation au projet OPALES (section 4.1), puis les différentes méthodes de TAL que nous avons expérimentées dans ce but (à partir de la section 4.2), et nous dessinons les grandes lignes de la méthodologie que nous avons mise au point (section 4.5). Nous en aborderons les limites et verrons qu'une des solutions possibles pour y répondre est la combinaison de méthodologies.

4.1 Présentation du corpus Petite Enfance

Un corpus est une collection homogène de textes, qui se veulent représentatifs d'un état de langue particulier, et sont liés à un domaine applicatif défini :

> Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extra-linguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue.[Habert *et al.*, 1997, p.11]

Le projet OPALES, comme nous l'avons vu en section 2.2.1, était centré sur la notion de partage et de capitalisation de connaissances spécialisées autour de bases de documents audiovisuels. Ces connaissances sont représentées sous la forme d'annotations, réalisées par des spécialistes du domaine, et portent sur les documents en intégralité ou sur des sections spécifiquement isolées du flux audiovisuel. Ces annotations peuvent être réalisées au moyen de texte libre, de formulaires correspondant à un point de vue spécifié[61] ou de Graphes Conceptuels. Ce dernier type d'annotation demande une « ontologie support », un réseau conceptuel qui fournit l'ensemble des concepts nécessaires aux spécialistes du domaine pour exprimer leurs connaissances. Le corpus concernant la petite enfance, selon un point de vue anthropologique, devait donc permettre de construire ce matériel conceptuel. La constitution de ce corpus était la première étape ce cette recherche, et devait être menée avec soin pour que ce corpus soit satifaisant à la fois au niveau de la couverture du domaine et au niveau de la représentativité terminologique (deux problèmes classiques concernant la constitution de corpus) : il devait englober tous les concepts et toutes les relations dont les spécialistes avaient besoin pour annoter, et permettre de présenter des représentations terminologiques qui leur soient familières.

Pour décider de la couverture du corpus, nous nous sommes tout d'abord interrogée sur la notion de « petite enfance », qui semblait subjective et peu apte à circonscrire un domaine

[61] Les champs du formulaire sont définis de manière générique et sont associés à un *point de vue* ; un utilisateur se logue suivant un point de vue, et a alors accès au formulaire d'annotation qui correpond à son domaine de compétence. Un utilisateur peut écrire des annotations suivant un seul point de vue, mais il peut consulter l'ensemble des annotations en filtrant par type de point de vue, et consulter les annotations faites selon l'ensemble des points de vue.

de connaissance. Ce thème a été défini par un des partenaires du projet, la vidéothèque du CNRS-Bellevue, en fonction de la répartition de son fonds documentaire. En effet, le CNRS-Bellevue était le fournisseur de contenus audiovisuels du projet, et possède un grand nombre de documents concernant les différents types de soins apportés à l'enfant dans diverses cultures, qui ont été réalisés par des anthropologues « sur le terrain » ; ces vidéos constituent le matériel de travail d'un groupe de recherche de ces anthropologues, au Museum National d'Histoire Naturelle (MNHN)[62]. Cette thématique de la petite enfance correspondait donc à un grand nombre de documents audiovisuels, traitant de sujets aussi divers que l'allaitement, l'endormissement, la toilette et le jeu des enfants, mais aussi de la place de la mère dans la vie active (et son influence sur le maternage) et de la participation des différentes personnes à l'éducation ou à la surveillance de l'enfant.

La première définition de ce domaine a été empirique : elle nous a été donnée par l'ensemble des documents de la vidéothèque correspondant à cette thématique, et le premier corpus textuel que nous avons constitué était celui de l'ensemble des notices correspondant à ces documents. Les notices sont rédigées dans une optique de pratique documentaire (qui est le cadre applicatif de la future ontologie) et ont le grand avantage d'avoir été rédigées par les anthropologues auteurs des vidéos : même si elles sont remaniées suivant les pratiques documentaires en vigueur au CNRS-Bellevue, ces remaniements se font en accord avec l'auteur, et nous garantissent par là même une certaine adéquation de la lexicalisation des concepts du domaine avec les conceptions de nos futurs utilisateurs-testeurs du système. Les notices correspondent au point de vue expert des auteurs des documents audiovisuels, normalisé suivant les pratiques documentaires de la vidéothèque.

Cependant, nous nous sommes rendu compte qu'elles étaient d'un niveau de généricité trop élevé pour décrire précisément ce qui intéressait les anthropologues du groupe d'evaluation : les gestes, matériels et types de personnes impliqués dans les différentes actions, ainsi que la succession temporelle de ces actions. En effet, une notice décrit la globalité d'un document en quelques lignes, elle reprend les grandes étapes de sa trame narrative et en souligne les principaux points d'intérêt. Mais le vocabulaire contenu dans ces textes ne permet pas de faire une description détaillée d'une séquence particulière, il ne permet que d'en décrire le cadre général : le pays et l'ethnie concernée, mais pas le fait que l'action soit pratiquée par le père dans un endroit public, et qu'il berce son bébé de bas en haut (et non pas de gauche à droite comme en Europe), par exemple. Pour élargir la couverture terminologique de notre corpus, nous nous sommes intéressée à deux sources différentes : d'une part des principes descriptifs dédiés aux chorégraphies [Laban, 1928] pour avoir l'ensemble des notions de base permettant de qualifier un mouvement, et d'autre part, nous avons utilisé la méthode et les outils mis au point par Natalia Grabar et Sophie Berland [Grabar & Berland, 2001] pour augmenter notre corpus en prenant Internet comme source d'information. Nous allons détailler ces deux points plus avant.

4.1.1 Description du mouvement : caractérisation par le biais de la chorégraphie

La description des différents mouvements selon lesquels les actions étaient pratiquées était l'un des points d'intérêt particulier des anthropologues impliqués dans le projet. En recherchant les paramètres selon lesquels exprimer une description du mouvement et l'intégrer à une structure terminologique selon un paradigme différentiel, nous nous sommes

[62]Laboratoire Soins du corps, Santé, Maladie, Malheurs.

tournée vers la chorégraphie comme source d'information. Il s'agit d'une discipline définie en tant que « Art de composer, diriger, d'ordonner des ballets et des danses ; »[63], qui nécessite donc un ensemble de termes précis pour transmettre des descriptions concernant le mouvement de personnes évoluant sur scène. Un courant de recherche, fondé par Rudolf Laban [Laban, 1928], s'intéresse particulièrement aux systèmes de notation des mouvements. Parmi les différents systèmes élaborés, celui de la Kinétographie (ou Labanotation, du nom de son premier auteur) est particulièrement utilisée dans le milieu de la danse. La Kinétographie utilise des symboles abstraits pour définir les éléments suivants : la direction du mouvement, la partie du corps qui réalise le mouvement, le niveau (intensité) du mouvement et sa durée. Nous en avons déduit un ensemble de propriétés à associer à un mouvement pour le décrire, comme les concepts de DroiteAGAuche, HautEnBas. Ces concepts sont liés à une ACTION par la relation *décrit-par*.

4.1.2 Acquisition semi-automatique d'un complément au corpus de notices

La méthode itérative de construction de corpus mise au point par Natalia Grabar et Sophie Berland est centrée sur un moteur de recherche ciblé, permettant d'aspirer les sites correspondant à une requête complexe[64], et présentant les résultats de la recherche dans un environnement qui en facilite la validation. Nous avons pris comme amorce pour ce corpus étendu les mots et syntagmes significatifs les plus fréquents dans notre corpus de notices[65], espérant obtenir par cet enrichissement des catégories de description plus spécifiques que celles présentes dans les notices. Ce corpus étendu compte 76 700 mots ; il traite de domaines aussi variés que possibles (alimentation de l'enfant, rites de passage, toilette, endormissement, etc.), et couvre un échantillon des zones géographiques traitées par les documents audiovisuels : Afrique, Asie, Amérique du Sud, Europe.

C'est sur ce corpus textuel étendu, appelé corpus « Petite Enfance », que nous avons effectué les premiers tests de méthodologies et outils pour la construction d'ontologies différentielles. Nous les avons testés selon les trois perspectives évoquées plus haut : l'acquisition de termes, la structuration de termes et la recherche de principes différentiels entre termes. Nous présentons les différentes expérimentations concernant l'acquisition des termes suivant une division par « familles » d'outils : outils à base statistique sans analyse de corpus (section 4.2.1), outils fondés sur une approche à fondements linguistiques (section 4.2.2). Les expérimentations concernant la structuration de terminologie sont divisées selon les approches se basant sur la structure interne des termes (section 4.3.1), sur l'analyse distributionnelle (section 4.3.2) et sur les patrons lexico-syntaxiques (section 4.3.3).

[63] Définition du TLFi.

[64] Un fichier contenant l'ensemble des mots-clés est interprété comme une requête complexe, associant l'opérateur booléen *OU* aux différentes lignes, chaque ligne correpondant à une expression où les mots-clés sont associés par le *ET* booléen.

[65] Cette liste d'amorce est la suivante : .

4.2 Acquisition de termes dans le corpus Petite Enfance

Nous avons vu plus haut que la recherche de termes en corpus peut se diviser en deux grandes familles de méthodes : celles s'appuyant sur un étiquetage morphosyntaxique ou une analyse linguistique complète du corpus et celles qui fonctionnent sans analyse de ce type. Dans la première famille d'approches, les premiers travaux français étaient ceux traitant des segments répétés, par Ludovic Lebart et André Salem. Nous avons en conséquence commencé nos tests concernant l'acquisition terminologique à partir de traitement statistique avec le logiciel Lexico3, développé par André Salem et permettant une recherche automatique de segments répétés en corpus.

4.2.1 Acquisition de termes sans analyse linguistique du corpus

Acquisition de termes avec Lexico3

Nous avons testé la recherche de segments répétés sur notre corpus (non analysé) au moyen du logiciel Lexico3[66], développé par André Salem et l'équipe universitaire SYLED-CLA2T (Université Paris 3). Cet outil permet une recherche de « segments répétés » pour la sélection de termes potentiels : des mots dont la fréquence d'apparition en cooccurrence dépasse un seuil fixé. Lexico3 propose également une recherche de la répartition des mots ou termes dans le corpus, pour avoir une idée de leur importance relative (distribution homogène, saillance dans certains passages, marginalité de l'unité linguistique sélectionnée, etc.). Ce logiciel ne prend toutefois pas exactement du texte brut en entrée : il faut découper le corpus en entités documentaires distinctes, associées à un identifiant. Le calcul des segments répétés se fait sur les formes brutes, ce qui inclut les mots grammaticaux (notamment les déterminants) s'ils n'ont pas été retirés du corpus. Notre corpus comptait très peu de segments répétés intéressants. Les unités les plus longues comptent 11 mots (114 séquences pour un nombre d'occurrences fixé à 2, et 9 séquences seulement pour un nombre d'occurrences fixé à 3), mais la première forme qui constitue un terme du domaine en compte 6 (avec seulement deux occurrences en corpus) : *Les rites de la première année*. L'ensemble des segments répétés intéressants sont présentés au tableau 4.2.1, avec dans la colonne de gauche, la longueur du segment, puis le contenu lexical du segment, et enfin sa fréquence, suivant la présentation de l'interface de Lexico3.

Les segments répétés de deux mots sont trop bruités pour être exploités sans filtre complémentaire du fait des nombreux mots grammaticaux qui sont inclus dans leur composition (segments répétés du type *de celui*, *suite de*, *lien avec*, *ils sont*, ...). Une évaluation quantitative sommaire nous a donné l'idée générale qu'ils contenaient très peu de candidats intéressants. Une liste des mots simples triés par ordre de fréquence d'apparition s'avère plus intéressante que le dépouillement manuel de cette dernière catégorie de segments répétés, du fait notamment des déterminants présents dans les termes suggérés, créant une resdondance avec les mots simples (*le maternage*, *maternage*).

Les segments répétés de plus de deux mots qui ont été jugés non pertinents[67] conte-

[66] Librement téléchargeable à des fins de recherche à l'URL `http://lexico3.no-ip.org`

[67] Cette évaluation n'a pas été réalisée par des experts du domaine, elle n'engage que notre propre jugement.

Longueur du segment	Segment Répété	Nb d'occurrences	Longueur du segment	Segment Répété	Nb d'occurrences
6	Les rites de la première année	2	3	l'avant bras	2
5	l'apprentissage de la propreté	9	3	l'assistante sociale	3
5	la venue d'un enfant	2	3	l'arrière plan	2
5	la toilette de l'enfant	2	3	le jeune enfant	4
5	enfants morts en bas âge	2	3	le lait maternel	2
5	les sociétés traditionnelles africaines et	2	3	le nouveau né	9
5	les troubles de la régulation	3	3	le père japonais	2
5	dans le monde des vivants	2	3	groupe d'appartenance	2
4	la bonne santé de	2	3	le post partum	2
4	âge de l'enfant	2	3	la vie quotidienne	7
4	de son dernier né	2	3	le système patrilinéaire	2
4	soins au nouveau-né	2	3	la sage femme	3
4	la divinité des naissances	2	3	les mères japonaises	3
4	la grand-mère maternelle	3	3	les nouveaux nés	3
4	la maîtrise des sphincters	2	3	les sociétés africaines	3
4	un reflux gastro œsophagien	2	3	les sociétés industrielles	3
4	l'imposition du nom	2	3	les sociétés traditionnelles	3
4	la mise au monde	2	3	la prime enfance	2
4	developpement de l'enfant	6	3	la position assise	2
4	la relation avec l'animal	2	3	consultation d'ethnopsychiatrie	5
4	la période de sevrage	2	3	couple mère bébé	3
4	corps de l'enfant	4	3	couple mère enfant	2
4	corps de la mère	2	3	un nom secret	2
4	les pratiques de maternage	2	3	la lignée paternelle	2
4	la relation mère bébé	3	3	la lignée maternelle	3
4	la relation mère enfant	2	3	la jeune fille	3
4	l'alimentation du bébé	2	3	une jeune femme	3
4	le monde des vivants	2	3	rite de passage	3
4	l'eau du bain	2	3	la femme enceinte	é
4	apprentissage de la propreté	13	3	pleurs du bébé	2
4	un rituel de réconciliation	2	3	alimentation du bébé	2
4	professionnels de la périnatalité	2	3	L'allaitement maternel	2
4	Les technoiques de maternage	2	3	la vie familiale	2
4	Les rites de passage	2	3	la sage femme	3
3	donne le sein	3	3	étape du développement	2
3	le campement pygmée	3	3	test de Gesell	2
3	le cordon ombilical	2			

Tab. 4.1 – Liste des segments répétés constituant des termes potentiels du domaine donnée par Lexico3

naient des mots grammaticaux (*dans la vie de la*, *soit parce qu'ils ont*) ou ne présentaient pas d'intérêt dans le domaine de spécialité qui nous intéresse : *et un arbre de décision*, *une stratégie d'évaluation*, *des états dysphoriques*, *de petits cailloux*... Une autre catégorie de segments répétés représentait des séquences complexes, comme *mère procède à la toilette* ou *elle lui donne le sein*, *l'alimentation et le sevrage*, par exemple, qui sont à découper en entités plus simples pour être intégrées dans l'ontologie.

Notre corpus étant petit et faiblement redondant, une recherche de segments répétés n'a donné que peu de résultats intéressants. En effet, dans une perspective d'équilibre entre les sous-thématiques, ce corpus aborde les différents aspects des soins à la petite enfance avec le minimum de répétition possible, pour ne pas privilégier un aspect par rapport aux autres. Or la récurrence lexicale (qu'il ne vérifie pas) est le premier présupposé sur lequel s'appuient les approches à base de statistique. Il n'était pas possible de définir dans Lexico3 une liste de mots-vides à exclure de la composition des segments répétés, et le fait que le logiciel prenne en compte ces mots-vides a entraîné beaucoup de bruit : beaucoup de récurrences lexicales étaient en fait dues à des déterminants ou conjonctions de coordinations, mais pas à des termes complexes « réellement » redondants. Cette première approche ne nous a pas donné de piste terminologique immédiate pour explorer la distribution de certaines unités en corpus : elle n'a pas fait apparaître d'unités saillantes autour desquelles organiser la terminologie. Le regroupement de type clustering n'a pas non plus donné de bons résultats, toujours à cause de la petitesse et de la non-redondance du matériel textuel.

Acquisition de termes avec Hyperbase

Le logiciel Hyperbase, d'Etienne Brunet (Université de Nice - Sophia Antipolis) implémente également cette recherche de segments répétés, ainsi qu'un classement des termes potentiels en fonction de leur spécificité par rapport à un corpus littéraire de référence : les 3000 textes français de la base textuelle FRANTEXT (composée de 80 % d'œuvres littéraires et de 20 % d'ouvrages scientifiques et techniques), réalisée par l'INALF, devenu ATILF (`http://www.atilf.fr`)[68]. Ce logiciel demande également un traitement du corpus spécifique pour pouvoir fonctionner. Aucune de ces deux fonctionnalités (recherche de segments répétés et association d'un score de spécificité aux mots les plus fréquents) n'a donné de résultats immédiatement exploitables. Le première raison en est que les calculs n'étaient pas réalisés sur l'ensemble du corpus, pourtant de petite taille, mais sur chaque passage, ce qui ne permettait pas d'avoir une vue d'ensemble des termes saillants de ce corpus ; la deuxième raison est la faible redondance de notre corpus, évoquée plus haut. Enfin, le présupposé selon lequel la fréquence des termes spécifiques à un domaine sera différente de celle de ces termes dans un corpus d'un genre différent, sur lequel se fondent également les calculs statistiques pour la recherche de termes spécialisés, ne fonctionne pas sur notre cor-

[68] L'unité mixte de recherche ATILF, *Analyse et Traitement Informatique de la Langue Française*, est née au 1er janvier 2001 du rapprochement de l'Institut National de la Langue Française (INALF - CNRS) et de LANDISCO (Langue Discours Cognition - Université Nancy 2), et s'est défini comme axes de recherche l'histoire de la langue française, les approches comparatives entre langues (Langue moderne et contemporaine), et la linguistique informatique. Dans le cadre de ce dernier axe, elle a continué le travail initié par l'INALF autour d'un dictionnaire informatisé de référence que nous avons déjà évoqué, le TLFi, et d'un corpus de textes français informatisés : FRANTEXT. La base de données textuelles FRANTEXT(`http://www.atilf.fr/frantext.htm`) permet la consultation de plus de 3000 textes français du XVI au XXème siècle ; l'interface d'interrogation est conçue en vue de recherches littéraires, linguistiques, lexicographiques et/ou stylistiques.

pus. S'il s'avère que dans la description des soins à la petite enfance, les termes spécialisés existent bien (les termes appartenant au domaine de l'anthropologie), un ensemble terminologique non négligeable concerne cependant la description de l'enfant, des personnes, du mobilier, de l'habitat, des ustensiles employés, etc. La plupart des mots de cet ensemble terminologique sont de type courant : ils figurent dans un dictionnaire de langue courante, et ont de grandes chances de se retrouver dans un corpus à forte dominante littéraire.

Conclusion

Le domaine choisi[69], la taille et la faible redondance de notre corpus le rendent peu propice au traitement statistique brut. Nous avons alors choisi de nous tourner vers les approches basées sur un étiquetage et/ou une analyse linguistique du corpus pour trouver les termes saillants du domaine. Ces termes devront être à la fois représentatifs du domaine choisi, mais aussi appropriés à la tâche de description de documents audiovisuels dans laquelle ils vont être mobilisés.

4.2.2 Acquisition de termes avec analyse linguistique

Nous avons choisi d'étiqueter et d'analyser notre corpus avec un outil d'analyse linguistique. Notre choix s'est porté sur Cordial Analyseur, développé par la société Synapse Développement[70] pour l'analyse linguistique, qui comporte une fontionnalité d'étiquetage morphosyntaxique. Cet outil est largement utilisé dans la communauté du TAL et atteint des performances qui le place dans le groupe de tête des étiqueteurs (avec seulement 2% d'erreurs dans l'étiquetage). De plus, il propose un certain nombre de fonctionnalités qui nous ont paru intéressantes : calcul des dépendances syntaxiques et extraction des mots les plus fréquents (et des segments répétés, mais ici non plus, ils n'étaient pas d'un grand secours), avec leur fréquence.

Une autre raison de notre choix est que l'étiquetage de Cordial est un des formats d'entrée pour l'analyse du corpus par LEXTER. Plutôt que de redéfinir nous-mêmes des patrons de termes, nous nous sommes servis de LEXTER pour produire un ensemble de candidats termes à partir de notre corpus. LEXTER a produit 32 776 candidats termes à partir de notre corpus. Le nombre de termes polylexicaux a été évalué sur un échantillon des 3 500 premier candidats termes, et n'a compté que 370 occurrences (hormis les décompositions d'expressions en anglais considérées par l'outil comme des termes complexes et décomposés mot par mot). Parmi ces derniers, seuls 227 (soit environ 6%) correspondaient effectivement à des candidats termes, les autres étant des regroupements non pertinents, que ce soit d'un point de vue syntaxique (ailes tout, bébé à peine, bonhomme têtard...), ou pour la description de documents audiovisuels concernant la petite enfance (arbres de décision,

[69] Ou plus généralement le fait de s'attaquer à de la description de documents audiovisuels : [Lespinasse, 2002] a également convenu que les approches de type statistique étaient difficilement exploitables dans ses propres travaux, bien qu'ils portent sur un corpus traitant de politique et étant très volumineux. Ce corpus avait été constitué de manière à être très homogène : il est composé de notices documentaires sélectionnées à partir de descripteurs spécifiques (ceux figurant dans les arborescences du thésaurus INA sous les descripteurs de ÉLECTIONS et de POLITIQUE INTÉRIEURE), et porte sur une période historique donnée : les années 1980, la « décennie Mitterrand ».« Le tri par segments répétés fait apparaître peu d'expressions complexes (plus de deux mots) » [Lespinasse *et al.*, 2000].

[70] `http://www.synapse-fr.com`

acide gras essentiel, abus de langage manifeste,...). Notre corpus comporte donc effectivement très peu de termes polylexicaux, ce qui confirme l'inadéquation de la recherche de termes par segments répétés ou techniques analogues.

Le fait que la plupart des termes soient des mots simples peut tenir du fait qu'il s'agit majoritairement d'un vocabulaire issu de la « langue courante » : il est très peu spécialisé au sens où il continent beaucoup de mots ordinaires, mais pris dans un sens spécifique. En effet, les anthropologues s'intéressent à des activités « courantes » comme le bercement, l'endormissement, etc. autour de la petite enfance, mais les considèrent dans une perspective d'analyse, et en font des termes de leur domaine. De plus, les anthropologues auxquels est destiné notre système souhaitaient disposer d'un vocabulaire issus de leurs textes de spécialité, mais permettant de décrire des situations de la vie ordinaire. Il est alors possible que les termes potentiellement pertinents de cette dernière catégorie soient des mots simple du fait de l'économie du langage courant mobilisé dans ces descriptions.

Cette caractéristique joue également un rôle sur les choix de méthode de structuration terminologique : celle basée sur la forme interne des termes polylexicaux semble povoir être laissée de côté, au profit des dépendances syntaxiques données par l'outil pourcomplexes qui nous concernent. En effet, les seules séries de termes polylexicaux que nous avons trouvés étaient en relation *Nom-Adjectif*. Ce rattachement est calculé et présenté de manière adéquate à la fois dans LEXTER et dans SYNTEX.

Nous avons ensuite testé les différences de ces candidats termes avec ceux proposés en sortie de SYNTEX, intégrant l'analyse linguistique à son processus d'extraction. Le fichier de sortie compte cette fois 3401 candidats termes complexes, chaque ligne comprenant, entre autres informations[71], le candidat terme simple central, son dépendant et le type de la dépendance (un extrait de ce fichier est présenté en tableau 4.2, représentant les premiers contextes autour du terme *enfant*, par ordre alphabétique). Le terme le plus productif, c'est-à-dire celui avec lequel le plus de candidats termes sont construits, est *enfant* (productivité de 122), suivi de *bébé* (38), et de *mère* (33). En partant de cette information, et en naviguant entre les dépendances des termes (de tête en expansion, et d'expansion en tête), comme dans l'outil LEXICLASS de Houssem Assadi [Assadi, 1998], il semble possible d'avoir une idée générale assez pertinente de l'organisation du domaine. Nous nous sommes donc basée sur ce format de données et ce treillis de dépendances pour tester une première organisation terminologique suivant les principes de l'analyse distributionnelle.

4.3 Structuration de termes

La deuxième étape classique de construction de terminologie ou d'ontologie est la structuration terminologique. Nous avons vu (en section 3.2) que deux grands types de méthodes pouvaient être mobilisés, lorsque l'on ne dispose pas de ressources extérieures au corpus : les approches se basant sur la structure des termes et celles se basant sur leur contexte.

[71] L'ensemble des colonnes présente, de gauche à droite, la productivité du contexte, la productivité du terme, le numéro affecté au recteur, le type de la relation entre le recteur et le régi, le numéro affecté au terme (68 pour le terme *enfant*), le terme, le numéro dans syntex et enfin la fréquence du candidat terme complet.

prod-contexte	**prod-terme**	**num-recteur**	**rel**	**num-terme**	**recteur**	**terme**	**num-synt**	**freq**
4	122	5176	de	68	développement	enfant	6558	31
13	122	122	SUJ	68	avoir	enfant	954	19
27	122	122	OBJ	68	avoir	enfant	2177	15
18	122	44	SUJ	68	être	enfant	3471	11
4	122	813	de	68	corps	enfant	4758	11
2	122	1726	de	68	santé	enfant	12081	11

TAB. 4.2 – Extrait du fichier de sortie de SYNTEX

4.3.1 Structuration des termes par l'analyse de la structure des termes

Ce type de méthodes se divise à nouveau suivant le type d'information prise en compte dans la structure des termes pour les organiser. Un premier groupe de travaux s'intéresse à la morphologie pour trouver des relations entre termes. Si cette approche est intéresante dans des domaines comme la médecine, où un grand nombre de termes sont de composition savante ([Zweigenbaum & Grabar, 2000], [Namer & Zweigenbaum, 2004]), elle nécessite la mise en œuvre de logiciels (comme FASTR) ou la mise au point de règles complexes, fondées sur des connaissances morphologiques. Dans le contexte de notre corpus Petite Enfance, la plupart des termes qui nous intéressent sont de type « général » (ils sont susceptibles de figurer dans un dictionnaire de langue générale), et les relations sémantiques inscrites dans leur construction morphologique sont limitées. Sur l'ensemble des 3401 candidats termes extraits par SYNTEX, seuls 41 groupes de têtes de syntagmes peuvent êtres associés (manuellement) sur la base de liens morphologiques ou étymologiques. Dans ces 41 groupes, 29 sont des groupes de deux candidats termes (corporel - corps, culture - culturel, éducatif - éducation, ethnie - ethnographie, reconnaissance - reconnaître,...). Certains de ces groupes de deux « têtes de candidats termes » ne sont pas pertinents dans le cadre descriptif qui nous intéresse (document - documentaire, industriel - industrialiser, par exemple). Les groupes de plus de deux candidats termes sont, eux, tous pertinents en termes de description[72] mais entretiennent des relations sémantiques complexes entre eux. Il s'agit des groupes suivants :

- Aliment, Alimentaire, Alimentation ;
- Enfance, Enfant, Infantile ;
- Maternage, Maternel, Mère ;
- Médecin, Médecine, Médical ;
- Mort, Mortalité, Mourir ;
- Naissance, Naître, Néonatal ;
- Paternel, Patrilinéaire, Père ;
- Personnalité, Personne, Personnel ;
- Préscolaire, Scolaire, Scolarité ;
- Protecteur, Protection, Protéger ;

[72] Ils concernent tous des sujets revenant dans plusieurs documents audiovisuels.

Ils ne semblent pas pouvoir être structurés de manière simple au moyen des liens morphologiques qu'ils entretiennent.

Les approches concernant la structuration de termes en fonction de leur composition lexicale servent surtout à structurer des termes plus complexes à partir de la projection en corpus d'un ensemble de termes simples déjà structurés : nous ne disposons pas de ces informations, et souhaitons avoir des critères objectifs pour la modélisation des termes simples au même titre que des termes complexes. Nous avons alors abordé les deux méthodologies principales permettant d'organiser des termes en fonction de leurs contextes : l'approche de l'analyse distributionnelle et celle par patrons lexico-syntaxiques.

4.3.2 Structuration des termes par analyse distributionnelle

Comme nous l'avons vu plus haut, LEXTER et SYNTEX fournissent en sortie un réseau de termes structurés en fonction de leurs dépendances syntaxiques. Ces dépendances syntaxiques sont typées dans SYNTEX, et les catégories principales sont : SUJ, OBJ, adj, de[73]. Nous avons mis en œuvre le principe de construction de classe notionnelle, ou classe distributionnelle développé par Houssem Assadi [Assadi, 1998] ou Benoît Habert (dans le logiciel Zellig), mais au niveau des dépendances verbe-sujet et verbe-objet (les dépendances de ce type sont calculées dans UPERY). L'idée derrière le fonctionnement de LEXICLASS est

> d'inspiration harrissienne [Harris, 1968] : rapprocher les syntagmes nominaux ayant des contextes similaires et constituer ainsi des classes de candidats termes, dont certaines seront interprétables sémantiquement et constitueront l'amorce d'une organisation de l'ontologie en champs conceptuels. Plus précisément, si l'on décrit chaque candidat terme par son contexte terminologique [...], on peut utiliser des méthodes statistiques de classification ascendante hiérarchique (CAH) pour obtenir des classes de candidats termes [Assadi & Bourigault, 2000, p.246].

Ainsi l'outil crée des classes lexicales, appelées *champs notionnels*, qui sont plus ou moins homogènes et aident à conceptualiser le domaine choisi en sous-catégories. Le terme le plus fréquent dans notre corpus étant ENFANT, nous avons cherché à voir quels champs notionnels pouvaient être modélisés autour de lui. Nous avons conçu un programme paramétrable, prenant en entrée un terme ou une liste de termes, éventuellement une relation syntaxique et éventuellement un seuil d'occurrences. Le ou les terme(s) est/sont recherchés dans le tableau de données fourni par SYNTEX ; si aucune relation syntaxique n'est spécifiée, ni aucun seuil d'occurrences, le programme extrait l'ensemble des contextes liés au(x) terme(s) spécifié(s). Si une relation syntaxique est entrée, le système ne recherche que les compléments correspondant à cette relation syntaxique, et si un seuil est fixé, le programme ne renvoie que les valeurs correspondant à la relation syntaxique *et* ayant un nombre d'occurrences en corpus supérieur ou égal au le seuil spécifié. Par exemple, si l'on recherche ENFANT dans tous les contextes où il est sujet, le programme renvoie la liste des verbes dont ENFANT est le sujet. Une deuxième passe du programme va ensuite récupérer l'ensemble des sujets de ces verbes, pour créer un groupe potentiellement homogène de « sujets ». Une

[73]Il s'agit des catégories les plus productives relevées sur le corpus Petite Enfance, l'ensemble des catégories relevées est le suivant : à, ADJ, ADV, après, autour de, avec, chez, comme, contre, dans, de, derrière, dès, en, entre, EPI (épithète), NNPR (nom propre), OBJ, par, pour, selon, sous, SUJ, sur.

Allaitement, allergie, attitude, auteur, balinais, beau corps, bébé, cérémonie, choix, choix de enfant, corps, douleur, enfant, étude, famille, femme, fesse, fille, garçon, idée, initiative, japonais, lait, lait industriel, lait maternel, maman, Maximilien, médecin, mère, mère japonais, naissance, nouveau-né, objectif, pédiatre, père, personne, sage-femme, second, sevrage, vie

TAB. 4.3 – Test de champ notionnel sans nombre minimal d'occurrences

troisième étape permet de généraliser le processus : le programme récupère l'ensemble des verbes dont tous les termes collectés précédemment sont sujets, et en extrait l'ensemble des sujets. Ces opérations peuvent être répétées jusqu'à ce que le champ notionnel ne soit plus augmenté. Dans un développement ultérieur, le nombre d'itérations souhaitées devrait aussi figurer dans les paramètres de la ligne de commande. Nous avons constaté empiriquement que deux itérations étaient suffisantes autour de la notion d'ENFANT. Suivant le seuil (correspondant au nombre d'occurrences) fixé, la classe obtenue est plus ou moins homogène, mais également plus ou moins riche. Si l'on considère l'ensemble des verbes dont ENFANT est sujet, et l'ensemble de leurs sujets fournis par l'analyse de SYNTEX pour lancer l'étape de généralisation du processus, on obtient les données présentées en tableau 4.3.

On peut distinguer les grands « types » sémantiques suivants parmi les candidats termes rapprochés : des personnes (mère, père, bébé, pédiatre,...), des éléments liés à l'alimentation (lait, lait maternel, sevrage), mais aussi des éléments difficiles à relier, constituant du bruit informationnel : vie, initiative, idée,... Le fait de fixer un nombre d'occurrences minimal de 2 réduit considérablement le bruit, et ne donne plus que les deux catégories suivantes :

- Les personnes : bébé, enfant, femme, fille, garçon, maman, mère, mère japonais, père ;
- L'alimentation : allaitement, lait ;

On obtient une classe totalement homogène à partir d'un seuil de 3 occurrences, mais certains candidats termes qui nous semblaient importants disparaissent du groupe : il ne comporte plus que *bébé*, *enfant*, *femme*, *mère*, *père*. Le choix du seuil dépend donc du type de traitement que l'on souhaite associer à ce type de programmes : s'il s'agit d'un traitement entièrement automatique, il vaut mieux choisir un seuil élevé, afin d'avoir des classes homogènes. Comme nous nous plaçons plutôt dans une optique de validation manuelle, un seuil minimal est suffisant.

D'autre part, la méthode expérimentée ici utilise un type de relation syntaxique à la fois (par exemple SUJ). En combinant plusieurs relations syntaxiques (par exemple SUJ et OBJ), c'est-à-dire en tenant compte des mots qui sont le plus souvent à la fois sujet et objet des mêmes verbes que ENFANT, comme c'ets le cas dans Zellig et UPERY, il serait également possible d'avoir des classes plus précises.

Nous nous sommes intéressée aux classes distributionnelles créées autour des relations SUJ et OBJ, mais également à des dépendances sytaxiques simples, notamment aux catégorisations en complément circonstanciel de lieu de temps fournies par Cordial sur notre corpus. La distinction entre ces deux dernières catégorisation était imparfaite, mais l'ensemble des deux catégories fournissait un ensemble lexical intéressant pour compléter la description de nos documents audiovisuels suivant le principe des « 5 W's and an H » : who, what, why, where, when, and how ? Ce principe cherche à décrire une scène en fonction de ses

participants, de l'action impliquée, de sa localisation spatio-temporelle et de sa motivation. Ces catégories ont été formalisées par Audrey Tam et Clement Leung [Tam & Leung, 2001], qui proposent une description d'une image selon une annotation structurée comprenant au moins un descripteur parmi les quatre types suivants :

- Un agent, un *enfant* par exemple (qui peut avoir différentes caractéristiques, comme l'âge, l'appartenance à une ethnie, etc.) ;
- Une action, comme *manger* ;
- Un objet (un *fruit*) ;
- Un cadre (une *place de village*, un *moment de la journée*), ce cadre peut aussi être décrit au moyen d'attributs.

Ce cadre descriptif s'applique tout à fait au cas qui nous intéresse : il permet de représenter l'ensemble des descriptions qui nous intéressent, au moyen de certains ou de l'ensemble des types de descripteurs. Il reprend notamment les catégories définies dans les guides d'usage établis à l'intention des documentalistes traitant de documents audiovisuels. Nous avons vu que ces catégories conceptuelles ont l'avantage de pouvoir être partiellement renseignées au niveau lexico-terminologique par les groupes créés par analyse distributionnelle autour des realtions SUJ et OBJ et par l'analyse de la phrase telle qu'elle est calculée par Cordial (complément de temps, complément de lieu). La question de la motivation de l'action correspond à une analyse sémantique qui n'intéressait pas les anthropologues : ils préféraient une étude comparative limitée au plan descriptif. Ce dernier point terminologique n'était donc pas à intégrer à notre ontologie.

Cependant, nous nous sommes aperçue que si l'analyse distributionnelle et l'analyse syntaxique de la phrase donnaient des catégories intéressantes d'un point de vue descriptif, les classes obtenues n'étaient pas structurées de manière interne (ou très peu, comme dans le cas de l'hyperonymie liant certains termes : lait/lait industriel, lait/lait maternel), ni les unes par rapport aux autres. Les relations sémantiques devaient être déduites d'une interprétation humaine des classes et des rapports entre celles-ci. Le champ notionnel des personnes humaines : bébé, enfant, femme, mère, père, ne contient pas de critères objectifs pour en classifier hiérarchiquement les différents constituants.

Il s'agit d'un écueil bien connu en TAL : l'analyse distributionnelle ne permet pas d'avoir des relations typées entre (candidats) termes, les regroupements sémantiques sont parfois hétérogènes et leur interprétation demande une expertise manuelle. Un autre problème régulièrement soulevé est le nombre important de candidats termes produits par l'outil : notre corpus étant faiblement redondant, nous n'avions pas de critères numériques pour faire des sélections parmi l'ensemble des propositions. Il s'avère que ce type d'approche est surtout intéressante pour des corpus de grande taille, homogènes et redondants au niveau lexical. Nous gardons donc à l'esprit le modèle du domaine que cette approche permet de lier à une analyse syntaxique du corpus, même sur un corpus comme le nôtre, et testons l'adéquation de l'extraction par patrons lexico-syntaxiques pour la construction d'une ontologie différentielle à partir de notre corpus Petite Enfance. En effet, les patrons lexico-syntaxiques sont modélisés pour extraire des couples de termes vérifiant une relation sémantique explicitement sélectionnée, la question du typage de cette relation est donc donnée *a priori*.

4.3.3 Structuration des termes par l'application de patrons lexico-syntaxiques

Emmanuel Morin propose une méthode en 7 étapes pour acquérir en corpus des patrons lexico-syntaxiques correspondant à la relation sémantique d'hyperonymie [Morin, 1998]. Il s'intéresse en particulier à cette relation car elle permet de construire l'ossature verticale d'une terminologie. Cette méthode, destinée à un terminologue ou à un ingénieur de la connaissance, se décline de la façon suivante :

1. Choisir la relation sémantique pour laquelle on souhaite créer des patrons lexico-syntaxiques (e.g. l'hyperonymie).
2. Fournir une amorce constituée de couples de termes qui respectent la relation précédemment spécifiée. Cette liste peut être obtenue à partir d'un thésaurus, d'une base de connaissances, ou bien constituée manuellement. Par exemple, un thésaurus d'agronomie fournira le couple en relation d'hyperonymie suivant : *calcium* EST-UN *cation*.
3. Extraire du corpus l'ensemble des phrases qui contiennent les précédents couples. Ainsi le couple (« *cation* », « *calcium* ») sélectionne dans le corpus [AGRO] la phrase « *Des cations tels que le sodium, le potassium, le calcium et le magnésium peuvent être dosés par une méthode de routine* ».
4. Trouver un environnement commun qui généralise les phrases (ou certaines phrases) extraites à l'étape 3. Cet environnement, décrit sous la forme d'une expression lexico-syntaxique, révèle un candidat patron lexico-syntaxique.
5. Retenir les candidats patrons les plus pertinents.
6. Utiliser les nouveaux patrons pour extraire de nouveaux candidats couples de termes.
7. Retenir les candidats couples de termes les plus pertinents. Ces nouveaux couples sont ajoutés à ceux de la liste initiale, puis le processus est réitéré à partir de l'étape d'extraction (étape 3).

Nous n'avions pas de couples de termes en relation d'hyperonymie pour amorcer le processus, mais Patrick Séguéla [Séguéla, 2001] a répertorié les patrons lexico-syntaxique d'hyperonymie et de meronymie qu'il a mis au point au cours de sa recherche. Nous avons repris ses patrons et les avons appliqués à notre corpus. Notre corpus étant étiqueté et lemmatisé, nous avons généralisé certains des patrons au moyen de la forme lemmatisée du marqueur (l'élément lexical central du patron), et celui-ci étant au format XML[74] nous avons testé les possibilités du langage XSLT pour exprimer les patrons lexico-syntaxiques. En effet, XSLT est un langage spécifiquement créé pour transformer des documents XML et il a des possibilités intéressantes en termes d'expressivité. Il permet, en effet, de faire des recherches sur le contenu des éléments balisés, sur les types et valeurs d'attributs associés, de faire des recherches en contexte (dans les N mots suivants ou les N mots précédents, par

[74]eXtensible Markup Language, un langage de représentation sémantique des données, où les étiquettes des balises (les tags) ne représentent pas des éléments de présentation, comme en HTML. Dans le cadre de ce mémoire, nous nous en servons pour marquer la syntaxe de nos corpus. Pour plus de précisions, voir notamment les recommandations du W3C sur `http://www.w3.org/XML/`. Nos balises délimitent les phrases et les mots du corpus, suivant l'analyse de Cordial -qui reconnait quelques locutions complexes- et associent à chaque forme graphique un identifiant unique, son lemme, sa catégorie morphosyntaxique, sa fonction syntaxique dans la phrase -sujet, par exemple- et le verbe recteur dont elle dépend, sous forme de couples attribut/valeur.

exemple), contexte qui est structuré sous forme d'arbre. Pour exprimer un patron lexico-syntaxique, il faut pouvoir spécifier une recherche positive dans le contexte gauche et droit du marqueur (marqueur précédé et/ou suivi d'un élément lexico-syntaxique), une recherche négative dans ce même contexte droit et gauche (marqueur non précédé et/ou non suivi d'un élément lexico-syntaxique), une recherche floue : marqueur précédé et/ou suivi (ou non) d'un élément lexico-syntaxique à une position comprise entre 0 et N mots, ou à une position non spécifiée. XSLT permet d'exprimer toutes ces nuances, et nous avons donc implémenté nos patrons dans ce langage.

Le tableau 4.4 présente un aperçu des résultats de l'extraction d'hyperonymes au moyen des patrons lexico-syntaxiques définis ou repris dans la littérature par Patrick Séguéla : les colonnes présentent, de gauche à droite, le marqueur central du patron lexico-syntaxique à l'origine de l'extraction de l'énoncé, deux ensembles lexicaux supposés contenir les termes en relation d'hyperonymie dans l'énoncé extrait, et enfin l'énoncé extrait.

Marqueur	**Terme 1**	**Terme 2**	**Enoncé**
est	, qui	signe d' alarme	[. . .]la douleur , qui est un signe d' alarme , ne doit pas être maintenue inutilement au-delà des coliques du nourrisson ,. . .
est	main à la bouche	moyen d' analyse orale pour le bébé	mettre sa main à la bouche est un moyen d' analyse orale pour le bébé ,. . .
est	c'	objet qui porte d' odeur de la mère	Le doudou , par définition , c' est un objet qui porte d' odeur de la mère ,. . .
parmi	acheter mille articles ,	peluches	Il ne s' agit pas non plus d' acheter mille articles , parmi lesquels peluches , poupons mous , couvertures , mouchoirs etc. . .
comme	substance	de l' alcool	vous n' avez pas absorbé de substances , comme par exemple de l' alcool , une drogue douce , une drogue dure , ou cocktail , certains médicaments , bref , tout ce qui fait que votre vigilance est modifiée ,. . .

Tab. 4.4 – Exemples de l'extraction d'hyperonyme au moyen de patrons lexico-syntaxiques définis ou repris par Patrick Séguéla

La recherche d'hyperonymes au moyen de ces patrons lexico-syntaxiques a donné des résultats intéressants (142 énoncés sur un total de 546 extractions), mais pose également un certain nombre de problèmes :

– Le bruit : les patrons renvoient un grand nombre de phrases qui ne correspondent pas à la relation sémantique attendue. 234 énoncés contenaient des unités linguistiques qui ne nous intéressaient pas (par exemple, le fait d'avoir une relation d'hyperonymie

entre *coq de combat* et *animal*), 14 extractions étaient liées à des erreurs de segmentation dans le corpus, et enfin 156 énoncés contenaient des termes liés par une autre relation sémantique (84 caractérisation, 48 synonymie, 9 définition, 1 méronymie et 14 « autres ») ;

– La question du point de vue : un certain nombre d'énoncés présentent une relation d'hyperonymie qui est « à validité locale », comme *L'allergie est un de nos fléaux actuels* ou *la carotte, c'est un moyen de faire ses gencives*. Ces formes ne reflètent pas un point de vue pertinent pour hiérarchiser notre terminologie différentielle, mais plutôt des aspects accessoires liés aux termes potentiellement intéressants : ils en montrent la fonction plutôt que l'essence. Certains des termes potentiels extraits ne sont pas non plus forcément intéressants pour la description des documents audiovisuels traitant de la petite enfance, comme la notion d'*allergie* de l'exemple précédent.

La première de ces deux remarques rejoint les recherches d'Anne Condamines sur la nécessaire adaptation des patrons lexico-syntaxiques aux corpus auxquels ils sont appliqués [Condamines, 2003] : en effet, les patrons de Patrick Séguéla ont été mis au point sur des corpus de textes techniques et scientifiques. Il faut les modifier pour qu'il donnent de meilleurs résultats sur notre corpus, en termes de précision, et il faut acquérir de nouvelles formes de l'expression de l'hyperonymie spécifiques à notre matériel de travail pour améliorer le rappel.

De plus, les deux remarques soulèvent un autre point largement débattu dans la communauté du TAL, à savoir qu'un outil d'extraction de ce type doit posséder une interface de validation, afin de traiter de manière efficace les données présentées, et de pouvoir continuer une chaîne de traitement vers une terminologie structurée uniquement avec les données validées. Se pose alors la question de la forme que doit prendre cette interface de validation, que nous aborderons plus loin (voir le chapitre 5), et celle de définir le profil de la personne en charge de la validation : à quel moment l'expert du domaine doit-il valider les données afin d'optimiser son intervention ? Les deux remarques nous montrent qu'il y a en fait deux niveaux d'évaluation : la distinction entre bruit et énoncés pertinents peut être faite dans une certaine mesure par un linguiste ou ingénieur de la connaissance en charge de la modélisation ontologique, en revanche la question de la pertinence des éléments « bien formés » extraits par rapport à l'application doit être discutée avec les experts du domaines (représentatifs des) futurs utilisateurs de l'application.

L'extraction d'éléments lexicaux liés par une relation sémantique typée présente donc un intérêt concret pour la construction d'ontologie à partir de matériel textuel non redondant et relativement peu abondant. Nous allons à présent voir dans quelle mesure cette méthode et l'analyse distributionnelle permettent de répondre à la recherche de principes différentiels, afin de définir les fondements méthodologiques les plus adaptés pour un outil d'aide à la construction d'ontologie *différentielle*.

4.4 Recherche de principes différentiels

4.4.1 Qualification des principes différentiels par l'analyse distributionnelle

Les outils d'acquisition terminologique fonctionnant selon les principes de l'analyse distributionnelle que nous avons détaillé (Lexter et Syntex) présentent en sortie un réseau de termes, qui sont liés par leurs dépendances syntaxiques. Les unités linguistiques diffé-

renciant un terme complexe d'un terme plus simple obtenu par sa décomposition peuvent être pertinentes pour qualifier l'axe de différence avec le père et l'unité linguistique représentant la chaîne commune entre ce terme complexe et le terme plus simple peut servir à qualifier l'axe de similarité avec le père. Par exemple, dans le cas d'une expansion adjectivale (comme dans l'exemple de la section 3.2), l'adjectif constitue la différence spécifique entre le terme complexe *coussin de sécurité arrière* et son hyperonyme *coussin de sécurité*. La chaîne commune peut également faire figure de principe de communauté avec le père : les deux termes sont des *coussins de sécurité*.

Cependant, une différenciation de type adjectivale ne correspond pas toujours à une relation de type inclusive dans une ontologie : le concept de *coussin de sécurité arrière* peut également être considéré comme une unité complexe, construite de *coussin de sécurité* associé à la relation de localisation, dont le codomaine (ou concept-cible) peut prendre les valeurs d'*avant* ou d'*arrière*. Ainsi, *coussin de sécurité arrière* ne figurera peut-être pas en tant que tel dans l'ontologie, mais sera « créé à la demande » lorsqu'un utilisateur de l'ontologie aura besoin de ce concept. Ce point est indépendant du fait qu'il y ait bien une relation d'hyperonymie entre les deux termes concernés, mais concerne des choix de modélisation ontologiques qui peut être lié, lui, à la catégorie morphosyntaxique du régi.

Concernant les autres dépendances, sur les 1945 premiers candidats termes fournis par Lexter [75], seuls des regroupements autour de 9 à 11 candidats termes centraux étaient intéressants. Il s'agit, par exemple, des ensembles terminologiques autour de :

- absoption : absorption de comprimé, absorption de eau ;
- activités : activités de habillage, de le enfant, de le mère, de soin,... ;
- anthropologie : anthropologie cognitif, culturel, social, structural, de le petite enfance, de le maladie, de le santé,...

Le nombre total de structures pertinentes obtenues est très faible, cette fois encore.

Nous avons vu, par ailleurs, que notre corpus comportait relativement peu de candidats termes complexes, et donc une recherche de principes différentiels basée sur la composition des termes est une solution intéressante, mais ne concerne pas la majorité des futurs concepts de l'ontologie. L'exploration de corpus au moyen de patrons lexico-syntaxiques permet-elle de trouver d'autres types de principes différentiels ?

4.4.2 Qualification des principes différentiels par recherche de patrons lexico-syntaxiques

Si les énoncés correspondant à une mise en correspondance d'un terme avec son hyperonyme sont utiles à la structuration verticale de la future ontologie, ils ne donnent pas toujours d'indices pour élaborer les principes différentiels entre ces termes. Les meilleurs candidats pour ce type d'information sont les énoncés que nous avons extraits au moyen de patrons modélisés à partir de marqueurs de type exemplificatoire[76] et définitoire[77], selon

[75] Ceux correspondant à la première lettre A quand ils sont classés par ordre alphabétique.

[76] Il s'agit de la liste suivante : comme, comme par exemple(s), du type, de type, tel que, pareil à.

[77] Il s'agit des verbes et participes suivants : appeller, assimiler, baptiser, coder, confondre, correspondre à, désigner, identifier, marquer, nommer, noter, représenter, qualifier, signer, signifier, susdénommé(e)(s), susnommé(e)(s), susdit(e)(s) symboliser. Patrick Séguéla a plus précisément décrit une liste de formes pertinentes issues de ces verbes dans sa thèse et les a associés à des prépositions pour réduire la polysémie de ces marqueurs.

les dénominations de Patrick Séguéla.

Nous référant à un article de [Veronis & Ide, 1990], nous pouvons envisager une exploitation d'énoncés à intérêt définitoire extraits de corpus dans le but de découvrir des principes de similarité sémantique entre termes, en adaptant une partie de leurs résultats concernant les définitions de dictionnaire. Les auteurs s'intéressent à la désambiguïsation sémantique des mots d'un corpus en utilisant des informations contextuelles et un réseau de neurones construit à partir du parcours lexical des définitions d'un dictionnaire. Chaque entrée du dictionnaire (ou *definiendum*) est associé au contenu lexical des mots pleins composant ses définitions (ou *definiens*), plus précisément à un parcours lexical par définition. Ce parcours lexical consiste en un ensemble de nœuds, représentant les mots « pleins » de la définition[78], liés selon leur ordre syntagmatique. Le but des auteurs est de comparer ces contenus lexicaux afin de déterminer le sens d'un mot dans une position particulière en corpus. Par exemple, dans le cas de *pen* (stylo ou enclos, en anglais), s'il est en contexte avec *book*, (le livre), le parcours lexical de leurs définitions respectives permettra de choisir le sens de *stylo*, puisqu'il partagera des mots comme *write* (écrire) avec une des définitions de *book*. Si *pen* est en contexte avec *goat* (la chèvre), la comparaison de leurs deux ensembles de parcours lexicaux permettra de choisir le sens d'enclos. Le problème auquel les auteurs se heurtent est celui du recouvrement lexical : le seul ensemble des mots pleins composants les définitions ne permet d'en rapprocher qu'un nombre restreint. L'approche donne de meilleurs résultats lorsque la définition initiale est en quelque sorte complétée des définitions des différents mots la composant, créant un parcours lexical plus diversifié.

Notre problématique n'est pas la désambiguïsation sémantique, mais cette expérience permet également de mettre en lumière que le mot de *write* (écrire) est le point commun entre *pen* (le stylo) et *book* (le livre) : *write* peut servir de base à l'élaboration d'un axe de similarité sémantique entre *pen* et *book*, l'un étant l'instrument et l'autre le produit résultant de l'action. L'extraction d'énoncés définitoires en corpus pourrait permettre, selon cette approche, de rapprocher des termes proches et la comparaison de ces énoncés définitoires pourrait permettre d'aider un modéliseur à définir un axe de similarité entre termes, futurs concepts d'une ontologie. De plus, en nous appuyant sur une comparaison de définitions de dictionnaire, nous pouvons penser que les mots différents entre deux définitions de termes proches peuvent également permettre de construire le principe de différence avec le frère. Par exemple, entre les définitions de *mère* et de *père* extraites du Petit Robert[79] :

- Mère : Femme qui a mis au monde *un ou plusieurs enfants* ;
- Père : Homme qui a engendré, qui a donné naissance à *un ou plusieurs enfants*.

les mots communs de *un ou plusieurs enfants* peuvent servir de base à la modélisation de leur axe de similarité, les mots les différenciant, à savoir *Femme qui a mis au monde* et *Homme qui a engendré, qui a donné naissance à* sont aussi suceptibles de constituer le principe de différence selon lequel ils s'opposent l'un l'autre. Il est possible d'en distinguer deux axes : l'opposition femme/homme, et celle entre *qui a mis au monde*/*qui a engendré, qui a donné naissance à*, la deuxième formulation posant d'ailleurs problème dans cette optique d'opposition.

Nous avons alors cherché à voir dans quelle mesure la recherche d'énoncés définitoires pouvait être à la base d'une méthodologie complète de construction d'ontologie différentielle à partir de corpus : permettent-ils également de faire de l'acquisition et de la structuration

[78]Par opposition aux mots « vides », ou « mots outils », correspondant notamment aux déterminants et mots grammaticaux, qui ne sont pas pris en compte ici.

[79]Edition de 1993.

terminologiques ?

4.5 Sélection d'une technique pour l'élaboration d'une méthode globale d'aide à la construction d'ontologies différentielles : l'extraction d'énoncés définitoires par patrons lexico-syntaxiques

[Chukwu & Thoiron, 1989] utilisent la recherche d'énoncés définitoires en corpus pour repérer des termes spécialisés, ce type d'énoncés permet donc de faire de l'acquisition terminologique.

Les différents travaux consacrés à la définition ou aux énoncés définitoires , avec entre autres ceux de [Martin, 1990], de [Pearson, 1998], de [Rebeyrolle, 2000], ou encore de [Sager, 2001], s'accordent pour dire qu'elles et ils sont réalisés suivant différentes relations sémantiques. Ces relations sémantiques comprennent l'hyperonymie (notamment dans le cas de la classique définition aristotélicienne par genre prochain et différence spécifique) : ces énoncés sont alors susceptibles de contenir les informations nécessaires à la structuration verticale de la terminologie différentielle.

Comme nous venons de le voir, ils peuvent également renseigner un modéliseur d'ontologie sur les principes différentiels selon lesquels qualifier les axes sémantiques reliant des termes co-hyponymes. Nous avons alors choisi de nous intéresser plus précisément à l'extraction d'énoncés définitoires en corpus, et à leur exploitation suivant ces différents aspects pour l'aide à la modélisation d'ontologies différentielles. Les exemples et expérimentations que nous présentons au chapitre 5 sont extraits et réalisés sur notre corpus de test, le corpus Petite Enfance, et un corpus de validation dont nous nous sommes servie pour valider nos hypothèses et tester la réutilisabilité de notre méthode. Ce deuxième corpus traite de diététique.

Chapitre 5

Description de la méthode globale : SODA (Structuration d'Ontologie Différentielle Assistée)

Nous cherchons à exploiter des énoncés définitoires dans une triple perspective, pour l'aide à la construction d'ontologies différentielles : l'acquisition terminologique, la structuration terminologique et la recherche de principes différentiels entre termes. Nous définissons plus en détail ce que nous entendons par *énoncés définitoires* et présentons les différentes méthodologies qui ont été proposées pour les repérer en corpus (section 5.1). Nous avons mis au point un outil d'exploration de corpus ciblé sur la recherche d'énoncés définitoires et permettant une validation intermédiaire des résultats par un ingénieur de la connaissance ou un linguiste en charge de la modélisation ontologique. Nous en détaillons le fonctionnement et l'implémentation en section 5.3.

Nous décrivons ensuite ses différents modules, et présentons une évaluation de leur application à deux corpus : le corpus de test traitant de la Petite Enfance que nous avons détaillé plus haut et un corpus d'évaluation traitant de diététique (section 5.5). Nous discutons des problèmes rencontrés, et exposons en section 5.8 les limites de cette approche.

5.1 Enoncés à intérêt définitoire : définition et méthodologies d'extraction

5.1.1 Qu'est-ce qu'un énoncé à intérêt définitoire ?

Josette Rebeyrolle décrit les énoncés définitoires comme « des structures qui permettent de réaliser un acte de définition en discours » [Rebeyrolle & Tanguy, 2000]. Concrètement, il s'agit de formulations naturelles (au sens où elles ne sont pas contraintes par un cadre de production lexicographique) qui utilisent la fonction métalinguistique de la langue [Jackobson, 1970]. D'un point de vue pratique, nous suivons la caractérisation pragmatique adoptée par Ingrid Meyer [Meyer, 2001], à savoir qu'un « contexte définitoire »[80] est une phrase qui peut servir de définition dictionairique, ou qui donne au moins un élément

[80]Il s'agit, selon cet auteur, de l'un des types d'énoncés riches en connaissances : « knowledge rich (defining) context »

sémantique propre à en construire une. La définition dictionnairique prototypique est celle issue des Topiques d'Aristote (chapitre I, section 5), qui est de ce fait qualifiée d'aristotélicienne. Elle propose de définir un objet au moyen de son genre prochain (l'élément sémantique qui lui est le plus proche et qui possède un sens plus générique que lui) et de ses différences spécifiques : ce qui permet de distinguer l'objet en question de son genre prochain. Ce point de vue peut se schématiser ainsi : *Species = Genus + Differentia*.

Il s'agit d'une définition de type classificatoire, qui implique une relation d'hyperonymie entre le definiendum, noté *Species* plus haut, et son genre prochain, le *Genus*. Mais il est possible de trouver d'autres types de relations sémantiques dans une définition de dictionnaire, et à plus forte raison dans un énoncé définitoire en corpus. Il n'est alors plus possible de parler de « genre prochain », c'est pourquoi nous parlons de relation sémantique entre les deux « termes principaux » de l'énoncé : le definiendum et le terme qui correspond à la position du Genus. Au nombre des relations sémantiques que l'on peut trouver entre ces deux termes, il y a la meronymie (*le doigt est la partie de la main qui...*), la fonction (*un médicament agit sur certains types de symptômes*, ou *traite certains types d'affections*), ou la caractérisation (*le lait maternel est riche en agents anti-infectieux*).

Pour évaluer la différence entre des énoncés définitoires en corpus et des définitions du dictionnaire, nous avons annoté manuellement l'ensemble des énoncés définitoires d'une version non étiquetée du corpus Petite Enfance, et en avons extrait (manuellement) les definienda. Le tableau 5.1 présente les différents types d'énoncés définitoires répertoriés sur le corpus Petite Enfance, avec leur fréquence en corpus.

Nous avons ensuite cherché si ces definienda (au nombre de 262[82]) étaient définis dans un dictionnaire de langue générale de référence, le TLFi. Lorsque les termes étaient définis, nous avons comparé la définition du TLFi avec l'énoncé définitoire pour voir s'ils recouvraient les mêmes notions ou s'il s'agissait de points de vues différents sur le terme (en effet, l'hypothèse sous-jacente est que l'énoncé définitoire représentera plutôt un point de vue « spécialisé » et la définition, rédigée selon une perspective et des règles terminographiques précises, représentera plutôt un point de vue correspondant à l'acception courante du terme). Pour 74 de ces 262 termes (soit 28% des énoncés), la définition du TLFi représentait un point de vue différent de celui de l'énoncé définitoire, et il n'y avait que 32 termes pour lesquels la définition correspondait à l'énoncé définitoire (ce qui représente 12% des énoncés). Elle était alors plus complète et plus longue que l'énoncé extrait du corpus. Nous avons également typé les definienda qui n'étaient pas définis dans le TLFi (voir le tableau 5.2).

Certains termes sont décrits par plusieurs énoncés définitoires en corpus, parmi lesquels certains peuvent correspondre au point de vue adopté dans le TLFi : c'est pour cette raison

[82]Certains definienda ont plusieurs énoncés définitoires les caractérisant en corpus.

Type de l'énoncé	**Nombres d'énoncés**
Formel (aristotélicien)	76
Semi-formel (caractérisation)	43
Informel[81]	235
Total	354

Tab. 5.1 – Répartition des énoncés définitoires dans le corpus Petite Enfance

que la somme des termes définis et non définis dans le TLFi dépasse 262. Nous pouvons voir que dans 236 cas (162 termes non définis et 74 termes définis d'une autre manière que celle envisagée dans le corpus) il est nécessaire de s'appuyer sur des énoncés définitoires plutôt que sur des définitions de dictionnaire pour avoir un point de vue « local » sur un terme (ou pour avoir un point de vue tout court). Cette observation nous conforte dans l'idée de faire de l'acquisition et de la structuration terminologique à partir d'énoncés définitoires extraits en corpus plutôt qu'au moyen de définitions, même dans le cas où elles existent et sont disponibles.

Typologie unifiée des énoncés définitoires

Il existe différentes typologies de définitions ou d'énoncés définitoires, établies en fonction de différents critères. Elles peuvent se diviser en trois grandes catégories :

- Les typologies fondées sur le but de la définition : définition linguistique *vs* définition encyclopédique [Picoche, 1977], qui sont respectivement des définitions s'intéressant à la description de l'usage d'un mot dans une langue donnée ou à la description de la chose à laquelle le mot réfère dans le monde ;
- Les typologies fondées sur le type de la paraphrase de reformulation du sens (le definiens) : définition formelle, semi-formelle, ou informelle [Trimble, 1985], [Flowerdew, 1992, p.202-221]. La définition formelle se conforme au schéma aristotélicien que nous avons vu plus haut : *Species = Genus + Differentia*. Une définition semi-formelle associe le definiendum à ses caractéristiques spécifiques ou à ses attributs [Meyer, 2001, p.279-302]. Une définition non formelle a pour but de « définir d'une manière générale, afin que le lecteur puisse reconnaître un élément familier dans la description d'un nouveau mot » [Trimble, 1985]. Ce type de définition peut consister en une association d'un mot avec un synonyme, une paraphrase informelle ou un phénomène de dérivation (comme dans *Jardinet : petit jardin*, où la définition conerne en fait le morphème *-et*) ;
- Les typologies fondées sur la relation sémantique qui associe le definiendum au « premier mot » du definiens. Martin [Martin, 1990, p.86-95] considère ainsi quatre catégories de definitions paraphrastiques (opposées aux définitions métalinguistiques) : dérivationnelle (*justification : action de se justifier*), approximative (*quiche : une sorte de tarte...*), métonymique (*manche : partie d'un habit*), hyperonymique, qui se décline encore en :

Type de l'énoncé	Nombres d'énoncés
Anecdotique	47
Nom propre	33
Terme polylexical spécialisé	39
Terme en langue étrangère	18
Terme polylexical de la langue courante	15
Terme spécialisé simple	10
Total	162

Tab. 5.2 – Type des termes ayant un énoncé définitoire en corpus et n'étant pas définis dans le TLFI

Type de l'énoncé	**Forme du marqueur**
Copulatif	*un X est un Y qui*
Equivalence	*equivalent à*
Caractérisation	*attribut de*, *qualité de*,...
Analyse	*composé de*, *« équippé de »*, *fait de*,...
Fonction	*avoir la fonction de*, *le rôle de*, *utiliser X pour Y*,...
Causalité	*causer X par Y*, *obtenir X par*,...

TAB. 5.3 – Enoncés définitoires comprenant un marqueur lexical de type « général »

- Positive : *aguicher : provoquer par...* ,
- Négative : *céder : ne plus résister à la pression*,
- Conjonctionnelle : *voler : se soutenir et se déplacer dans les airs.*

Dans cette dernière catégorie, et compilant les résultats de [Martin, 1983], [Martin, 1992], [Chukwu & Thoiron, 1989], [Condamines, 1993] et de [Loffler-Laurian, 1983], Alain Auger [Auger, 1997] propose une typologie unifiée des énoncés définitoires qui reprend tous les types que nous avons évoqués au paragraphe précédent. Sa typologie décline les énoncés définitoires en fonction des marqueurs linguistiques auxquels ils sont associés ; ces marqueurs peuvent nous servir de base pour la création de patrons lexico-syntaxiques afin d'extraire ces énoncés de nos corpus. Alain Auger distingue trois catégories d'énoncés définitoires, en fonction du type de niveau linguistique mis en œuvre pour les exprimer :

- Enoncés définitoires exprimés au moyen de marqueurs linguistiques de « bas niveau » : indices de ponctuation, comme la parenthèse, le tiret d'incise ou les guillemets ;
- Enoncés définitoires exprimés au moyen de marqueurs lexicaux : marqueurs linguistiques ou métalinguistiques ;
- Enoncés définitoires exprimés au moyen de marqueurs linguistiques de « haut niveau » : des tournures syntaxiques comme l'anaphore ou l'apposition.

Les énoncés définitoires indentifiables au moyen de marqueurs lexicaux peuvent se diviser en deux catégories : celles à marqueurs lexicaux « génériques » et celles à marqueurs explicitement métalinguistiques (que [Rebeyrolle & Tanguy, 2000] qualifient respectivement d'énoncés définitoires indirects et directs). Ces types d'énoncés et leurs marqueurs correspondants sont détaillés dans les tableaux 5.3 et 5.4.

Ces différents indices ont été exploités (en totalité ou en partie) dans des travaux visant au repérage et à l'extraction automatique d'énoncés définitoires en corpus. Nous présentons un survol de ces travaux dans la section suivante (5.1.2).

5.1.2 Méthodologies pour le repérage d'énoncés définitoires en corpus

Différents travaux s'intéressent au repérage d'énoncés définitoires en corpus. Ils se déclinent en trois grandes familles méthodologiques très proches : elles s'appuient sur des indices lexicaux (ou lexico-syntaxiques) susceptibles d'être des marqueurs de ces énoncés

et des informations contextuelles permettant de les différencier d'autres types d'occurrences (de ces marqueurs) en corpus. Certains de ces indices ont été listés dans la section précédente, sous le nom de marqueurs. Nous illustrerons ces trois courants de recherche par les travaux d'Emmanuel Cartier [Cartier, 1997], de Jennifer Pearson [Pearson, 1999] et Josette Rebeyrolle [Rebeyrolle, 2000], et enfin, pour la troisième famille de méthode, par ceux de Smaranda Muresan et Judith Klavans [Muresan & Klavans, 2002].

Repérage d'énoncés définitoires par exploration contextuelle [Cartier, 1997] suit la méthodologie de l'exploration contextuelle initiée par Jean-Pierre Desclés [Desclés, 1997]. Cette méthodologie consiste à éliciter un certain nombre d'éléments lexicaux susceptibles d'être des indices d'un énoncé définitoire. Le linguiste doit ensuite définir un certain nombre de règles visant à caractériser plus précisément les contextes dans lesquels ces indices sont véritablement des marqueurs d'énoncés définitoires. Quand une phrase continent un des indices listés, et qu'elle répond aux contraintes définies dans les règles contextuelles, elle peut être considérée comme un énoncé définitoire, et être traitée automatiquement.

Repérage d'énoncés définitoires par patrons lexico-syntaxiques Jennifer Pearson [Pearson, 1999] et Josette Rebeyrolle [Rebeyrolle, 2000] se sont servis de patrons lexico-syntaxiques. Comme nous l'avons déjà vu (voir la section 3.2.2), cette méthodologie a tout d'abord été décrite dans [Hearst, 1992], et son principe consiste à décrire le contexte lexical et syntaxique d'une occurrence en corpus d'une paire de termes connus pour être liés par une relation sémantique spécifique. Les premiers travaux en la matière se sont intéressés à l'hyperonymie, mais Pearson et Rebeyrolle se sont servis de ce principe de modélisation lexico-syntaxique pour repérer et extraire des énoncés définitoires en corpus. Rebeyrolle a évalué le rappel et la précision de ses patrons sur différents corpus, en les subdivisant en fonction de types de patrons. Elle obtient une précision allant de 18 à 79 %, et un rappel de 95 à 100 %, suivant les patrons impliqués dans les extractions. Les patrons à marqueurs métalinguistiques ont donné de très bons résultats, et ceux avec des marqueurs linguistiques plus génériques, de moins bons.

Repérage d'énoncés définitoires par règles linguistiques Smaranda Muresan et Judith Klavans [Muresan & Klavans, 2002] ont construit un logiciel d'extraction d'énoncés définitoires spécifiquement dédié au domaine médical : DEFINDER. Son principe de fonctionnement est basé sur des règles linguistiques (centrées elles aussi sur les marqueurs de

Type de l'énoncé	**Forme du marqueur**
Type de l'énoncé	Forme du marqueur
Désignation	*« désigner »*, *« vouloir dire »*,...
Dénomination	*« nommer »*
Systémique[83]	*« écrire »*, *« épeler »*, *« le nom »*,...

TAB. 5.4 – Enoncés définitoires comprenant un marqueur lexical de type métalinguistique

définition), et leur rappel est augmenté par une phase d'apprentissage semi-automatique sur corpus. Les auteurs ont évalué leur système en fonction de l'utilité des énoncés définitoires produits pour la compréhension de notions médicales (pour des spécialistes ou des non spécialistes du domaine) par rapport à des définitions issues de dictionnaires médicaux. Leur précision atteint 87 % et le rappel 75 % suivant leurs critères d'évaluation.

Motivation de notre choix Nous avons choisi la seconde méthodologie pour notre propre expérimentation : celle des patrons lexico-syntaxiques. Cela nous permet, d'une part, de nous appuyer sur une analyse préalable concernant la langue française réalisée par [Rebeyrolle, 2000], et d'autre part, la forme des patrons lexico-syntaxique même est particulièrement adéquate pour l'extraction des différents éléments qui nous intéressent *dans* l'énoncé définitoire : les deux « termes principaux » et la qualification de la relation sémantique qui les lie.

Nous avons adapté les différents patrons que d'autres chercheurs ont développé antérieurement (notamment Josette Rebeyrolle et Jennifer Pearson), et recherché dans notre corpus de test (corpus Petite Enfance) et dans la littérature s'intéressant à la définition, la paraphrase [Fuchs, 1994] ou la reformulation, d'autres marqueurs propres à repérer des énoncés définitoires. Nous nous sommes notamment intéressée à un marqueur de bas niveau très productif : la parenthèse, et avons modélisé quatre patrons autour de cet indice particulier. Nous nous sommes également intéressée à qualifier automatiquement la relation sémantique selon laquelle la définition était exprimée, ce qui n'était pas abordé dans les travaux antérieurs.

5.2 Corpus de validation : le corpus Diététique

Le corpus de validation, autrement dit, celui sur lequel nous avons testé la généricité et la validité inter-corpus de nos programmes informatiques, concerne la diététique. Cette thématique a été abordée dans la perspective du maintien des personnes en bonne santé, car cette optique correspond à un ensemble d'émissions télévisuelles (comme *Santé à la Une*) et à une problématique intéressant le domaine médical. Il a été constitué automatiquement à partir des articles francophones référencés par le portail CISMeF (`http://www.chu-rouen.fr/cismef/`) sous les arborescences « diététique » et « nutrition », convertis au format texte. Il compte 480 Kmots.

5.3 Principes de fonctionnement et enchaînement des différents modules

5.3.1 Extraction d'énoncés définitoires

La première question qui se pose est celle du repérage des énoncés définitoires. Les patrons lexico-syntaxiques que nous employons pour cela s'articulent autour d'un ensemble de marqueurs, c'est-à-dire des mots ou expressions qui sont souvent révélateurs d'un énoncé définitoire : « défini comme », « c'est-à-dire », « emploie le terme de ». La finalité d'un patron lexico-syntaxique est de préciser les contextes lexicaux et syntaxiques dans lesquels un marqueur introduit bien l'une des relations sémantiques recherchées. Par exemple, autour du marqueur « comme », nous avons défini les deux patrons suivants :

– définir *MOTS** comme *SN* ;
– *SN* comme *DET{1,3} SN* ;

Ces patrons peuvent utiliser des informations sur les formes (comme), lemmes (définir), catégories morphosyntaxiques (DET pour déterminant) et fonctions syntaxiques calculées précédemment. L'étoile signifie la présence facultative de l'élément et les accolades la possibilité de répéter un nombre déterminé de fois l'élément la précédant. Nous limitons généralement notre extraction d'énoncés définitoires au contexte de la phrase parce que c'était le plus souvent suffisant. Dans la plupart des exceptions en la matière, les phrases commençaient par *il s'agit de* ou par un pronom personnel. Nous avons alors développé des algorithmes spécifiquement dédiés à ces cas particuliers, qui extraient la phrase courante et sa précédente.

5.3.2 Sélection des « termes principaux » dans les énoncés à intérêt définitoire

La deuxième question à traiter est celle du repérage des unités lexicales en relation. Nous avons défini pour cela deux modalités :

– si le marqueur est un verbe, nous extrayons son sujet et son objet direct dans l'énoncé, s'il en contient, et sinon, nous extrayons respectivement :
  – le groupe syntaxique ayant la même fonction que le nom précédant le marqueur ;
  – le groupe syntaxique ayant la fonction du premier mot plein suivant le marqueur ;
– si le marqueur n'est pas un verbe, nous extrayons les groupes syntaxiques précédant et suivant le marqueur de la manière décrite ci-dessus.

Les dépendances syntaxiques et la délimitation des groupes syntaxiques sont calculés par Cordial.

Dans les cas où deux marqueurs doivent être présents dans la phrase (*définir* associé à *comme*,...), nous ne spécifions pas la position relative des deux marqueurs dans la phrase, et extrayons les sujets et objets ou les contextes droits et gauches du verbe. Ce procédé rudimentaire donne toutefois des résultats de l'ordre de 55 % de précision [Malaisé *et al.*, 2004b] et permet de factoriser les patrons. La qualité de cette extraction dépend en grande partie de la qualité de la segmentation initiale des phrases et de leur analyse. L'interface de notre outil d'extraction laissant la possibilité de corriger les unités lexicales proposées, nous avons essayé de proposer une séquence qui couvre au minimum l'unité lexicale intéressante, même si elle est entourée de mots parasites, à une extraction plus ciblée, mais donnant moins de résultats.

5.3.3 Organisation hiérarchique des termes par extraction d'énoncés à intérêt définitoire : recherche d'hyperonymie

La dernière question concerne la détermination des relations sémantiques. Chaque patron sert à détecter une relation sémantique entre les unités lexicales extraites (hyperonymie, paradigme, etc.). Cette relation est proposée à l'utilisateur pour validation ou correction éventuelle dans l'interface d'extraction. Par exemple, les patrons autour de la parenthèse permettent d'extraire des unités en relation de paradigme (des éléments lexicaux n'étant ni synonymes ni antonymes, mais co-hyponymes potentiels, comme une *mère* et un *père*), d'hyperonymie (le *moïse* (ou *berceau*)) ou de définition « fonctionnelle » : « [...

] l'*anthropologie* (qui *s'occupe surtout des contextes et des significations culturelles*), [...] ». Cet ensemble de relations est proposé pour correction lors de l'extraction de phrases correspondant à un patron lexico-syntaxique lié à la parenthèse (voir figure 5.1), avec en tête de liste la relation qui correspond spécifiquement à ce patron.

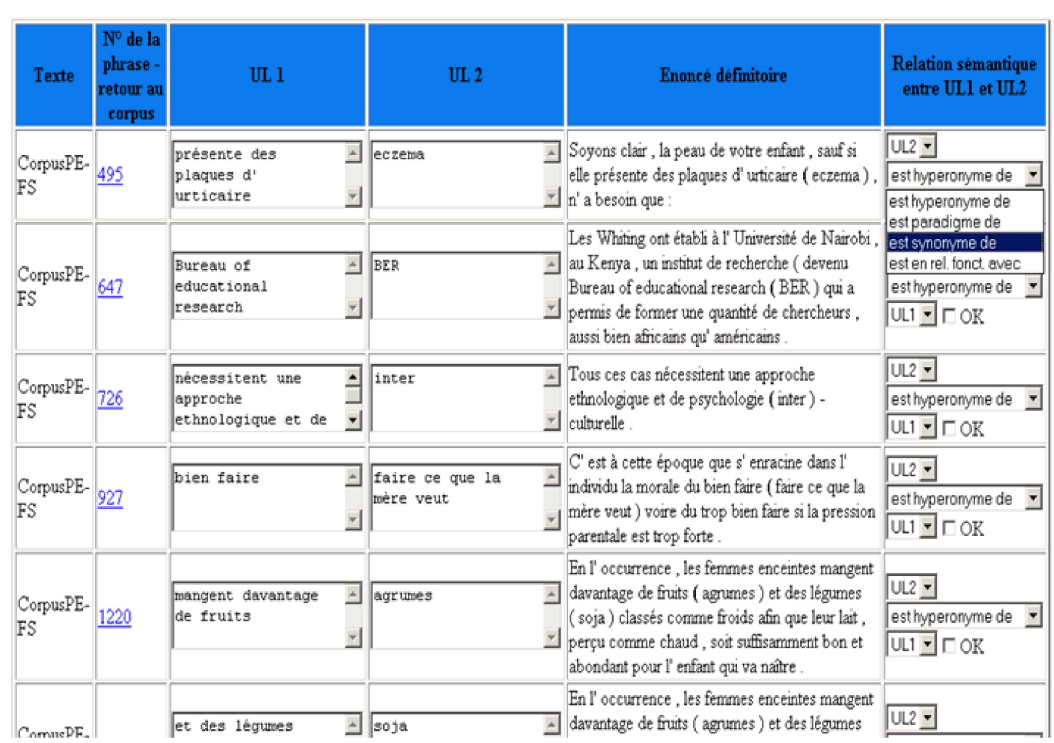
5.4 Implémentation de la méthode

L'ensemble de la méthode est implémentée en XSLT, à l'aide du processeur XLST Xalan (`http://xml.apache.org/#xalan`), par souci de cohérence avec des choix antérieurs concernant le format du corpus (XML) et nos outils de gestion d'ontologie (éditeur DOE, [Troncy & Isaac, 2002]). Les patrons sont directement exprimés en XSLT, ce langage permettant d'exprimer toutes les composantes des patrons don nous avions besoin : spécification positive et négative des contextes droits et gauches autour de marqueurs, définition des types de contextes à prendre en compte et de la nature des éléments du contexte à prendre en compte (forme graphique, lemme, catégorie morphosyntaxique, groupe syntaxique, etc.). Le résultat du traitement est un formulaire HTML, comportant des champs modifiables, qui sert d'interface de validation à l'utilisateur. Cet accès permet à un non spécialiste du domaine d'effectuer une première phase de validation (sur la base de son intuition de linguiste ou d'ingénieur de la connaissance), avant de présenter les résultats à un expert du domaine. Nous avions vu qu'un accès intermédiaire aux résultats était important, et cette interface est une réponse à cette nécessité de travailler en (au moins) deux étapes. La figure 5.1 montre une copie d'écran de cette interface. Chaque rangée du tableau correspond à l'appariement d'un patron avec une phrase du corpus ; elle présente les deux unités lexicales extraites, la phrase concernée et la relation sémantique proposée entre les deux unités lexicales ; un lien permet de retourner au corpus pour y examiner la phrase dans son contexte d'origine, et la dernière colonne permet de valider l'association.

Le formulaire de validation est associé à une base de donnée MySQL. Nous avons créé des programmes pour exporter des données qui y sont stockées dans un format XML compatible avec l'éditeur DOE, et dans un format texte dont nous verrons l'intérêt en section 5.7.1.

Mise au point des patrons lexico-syntaxiques Nous avons constitué une liste de marqueurs pour le repérage des énoncés définitoires à partir, notamment, des travaux de [Auger, 1997], [Rebeyrolle, 2000], et de [Fuchs, 1994]. C'est en précisant leurs contextes d'usage que nous avons construit nos patrons lexico-syntaxiques, suivant la méthodologie de [Séguéla & Aussenac-Gilles, 1999]. La nécessité de leur adaptation en patrons liés au corpus tient à la polysémie des marqueurs lexicaux suivant le domaine décrit (par exemple « baptiser » peut être un marqueur de définition fiable, sauf dans le cas d'un corpus centré sur la petite enfance, où il désignera plutôt le baptême), et à une variabilité des formes syntaxiques qui sont plus ou moins complexes à décrire suivant le genre de documents pris en compte dans le corpus. Nous distinguons quatre types de marqueurs, définissant quatre groupes de patrons :

- Les marqueurs métalinguistiques à utiliser indépendamment (au nombre de 9) : appeler, baptiser, définir comme, dénommer, dénoter, désigner, nommer, signifier, vouloir dire ;
- Les marqueurs métalinguistiques nominaux (11) : appellation, acception, concept, dénomination, désignation, expression, mot, nom, notion, terme, vocable, *à associer*



Texte	N° de la phrase - retour au corpus	UL 1	UL 2	Enoncé définitoire	Relation sémantique entre UL1 et UL2
CorpusPE-FS	495	présente des plaques d' urticaire	eczema	Soyons clair , la peau de votre enfant , sauf si elle présente des plaques d' urticaire (eczema) , n' a besoin que :	UL2 est hyperonyme de (est hyperonyme de / est paradigme de / est synonyme de / est en rel. fonct. avec)
CorpusPE-FS	647	Bureau of educational research	BER	Les Whiting ont établi à l' Université de Nairobi , au Kenya , un institut de recherche (devenu Bureau of educational research (BER) qui a permis de former une quantité de chercheurs , aussi bien africains qu' américains .	est hyperonyme de UL1 OK
CorpusPE-FS	726	nécessitent une approche ethnologique et de	inter	Tous ces cas nécessitent une approche ethnologique et de psychologie (inter) - culturelle .	UL2 est hyperonyme de UL1 OK
CorpusPE-FS	927	bien faire	faire ce que la mère veut	C' est à cette époque que s' enracine dans l' individu la morale du bien faire (faire ce que la mère veut) voire du trop bien faire si la pression parentale est trop forte .	UL2 est hyperonyme de UL1 OK
CorpusPE-FS	1220	mangent davantage de fruits	agrumes	En l' occurrence , les femmes enceintes mangent davantage de fruits (agrumes) et des légumes (soja) classés comme froids afin que leur lait , perçu comme chaud , soit suffisamment bon et abondant pour l' enfant qui va naître .	UL2 est hyperonyme de UL1 OK
CorpusPE-		et des légumes	soja	En l' occurrence , les femmes enceintes mangent davantage de fruits (agrumes) et des légumes	UL2

FIG. 5.1 – Interface de visualisation et de validation des extractions

à un verbe support parmi : appliquer, donner, employer, prendre, porter, recevoir, référer, renvoyer, réserver, utiliser ;
- Les marqueurs lexicaux n'étant pas explicitement métalinguistiques, ou ceux de reformulation (21) : c'est-à-dire, en d'autres termes, soit, à savoir, en quelques sortes, une sorte de, enfin, il s'agit de, entendre par, vouloir dire, indiquer, comme, dit, par exemple, autrement dit, même chose que, équivaloir à, employer pour, marque, expliquer, préciser ;
- Les ponctuations : parenthèses, guillemets et tirets d'incise sont également mentionnés dans la littérature. Nous nous sommes intéressée aux contextes définitoires autour de la parenthèse, et l'observation du premier corpus nous a permis de mettre au point quatre patrons synthétisant des contextes « intéressants » autour de cet indice de bas niveau. Nous nous sommes aperçu que, outre la définition, nous pouvions extraire des paradigmes (aide à la modélisation horizontale de l'ontologie) et des hyperonymes. Nous avons alors décidé d'inclure ces quatre schémas à l'évaluation de la méthode, même s'ils n'étaient pas exclusivement ciblés sur l'extraction de définitions.

Nous avons implémenté un ensemble de 74 patrons pour la recherche d'énoncés définitoires exprimés selon des relations sémantiques d'hyperonymie ou de synonymie. Les patrons sont écrits en XSLT, la figure 5.4 donne un apperçu du fichier regroupant les patrons pour la recherche d'énoncés définitoires « hyperonymiques » autour de la parenthèse. Le détail des patrons lexico-syntaxiques modélisés figure en annexe.

5.5 Recherche d'énoncés définitoires, acquisition et structuration terminologiques avec SODA

Ces évaluations sont compilées des résultats publiés dans [Malaisé et al.*, 2004b] et [Malaisé* et al.*, 2004a].* Dans une première expérimentation, nous sommes intéressée aux énoncés définitoires réalisés au moyen des relations sémantiques d'hyperonymie et de synonymie ainsi qu'aux énoncés susceptible de permettre l'extraction de co-hyponymes potentiels, comme le couple père/mère, par exemple. Nous avons appliqué l'ensemble des patrons correspondants à nos deux corpus (le corpus de test Petite Enfance et le corpus d'évaluation baptisé « Diététique »), et avons procédé à différentes évaluations des extractions. Les mesures classiques pour évaluer ce type d'extractions sont la précision (la proportion d'extractions correctes parmi les résultats du système) et le rappel (la proportion d'extractions du système parmi les résultats attendus selon un étalon de référence). Nous avons relevé à la main l'ensemble des énoncés définitoires du corpus Petite Enfance, et les avons également relevés sur un échantillon aléatoire du corpus d'évaluation (sur 13 textes extraits des 132 du corpus traitant de la diététique) pour constituer cet étalon de référence. Cependant, comme nous avons également extrait des énoncés contenant des éléments de type « co-hyponymes », nous n'avons pas pu calculer le rappel de nos extractions. Nous en présentons donc la précision : la précision de l'extraction des énoncés définitoires et la précision des termes principaux issus de ces énoncés.

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<!-- cette feuille de style lance les traitements : construction du tableau de visualisation, recherche d'énoncés
définitoires -->
<!-- début de la feuille de transformation -->
- <xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
  <xsl:output method="html" encoding="ISO-8859-1" indent="yes" />
  <xsl:include href="PresentationFormValidation.xsl" />
  <xsl:include href="cadreFormValidation.xsl" />
  <!-- template qui évite une mauvaise gestion des noeuds textuels -->
  <xsl:template match="text()" />
  <!-- MAIN : ce template gère les contextes autour de la parenthèse où l'on observe régulièrement une relation de
  type hyperonymique -->
- <xsl:template match="w[@lemme='(']">
  - <xsl:choose>
      <!-- cas de la parenthèse précédée d'un nom commun -->
    - <xsl:when test="preceding-sibling::w[1][starts-with(@catMS,'NC')]">
      - <xsl:if test="following-sibling::w[1][starts-with(@catMS,'NC')]">
        - <xsl:if test="string-length(following-sibling::w[1]/@lemme)!= '1'">
            <!-- si ce n'est pas une lettre seule -->
          - <xsl:if test="string-length(following-sibling::w[1]/@lemme)!= '2'">
            - <xsl:if test="following-sibling::w[1][not(starts-with(@lemme,'1'))]">
              - <xsl:if test="following-sibling::w[2][@lemme=')']">
                - <xsl:call-template name="contexteGD">
                    <xsl:with-param name="pivot" select="'('" />
                    <xsl:with-param name="stop" select="following-sibling::w[1]/@fonction" />
                    <xsl:with-param name="positionApres" select="1" />
                    <xsl:with-param name="fs" select="preceding-sibling::w[1]/@fonction" />
                    <xsl:with-param name="positionAvant" select="1" />
                  </xsl:call-template>
                </xsl:if>
              </xsl:if>
            </xsl:if>
          </xsl:if>
        </xsl:if>
      </xsl:when>
```

FIG. 5.2 – Exemple de patron lexico-syntaxique XSLT

5.5.1 Méthodologie d'évaluation des extractions des énoncés et des unités linguistiques

La validation concernant l'énoncé global s'articule en deux points : est-il conforme à la relation sémantique annoncée (hyperonymie, synonymie, « co-hyponymie ») ; et, si ce n'est pas le cas, cet énoncé nous intéresse-t-il ? Certains patrons peuvent être sous-spécifiés par rapport au corpus, et extraire des phrases ne correspondant pas aux énoncés attendus. Mais si l'extrait s'avère être intéressant parce qu'il présente des relations structurantes et des unités lexicales propres à être intégrées dans l'ontologie, nous ne le rejetons pas, mais essaierons dans un deuxième temps de créer deux patrons distincts et de gérer la nouvelle relation observée. Par exemple, le marqueur « signifier » permet d'extraire des énoncés définitoires et des phrases exprimant des implications liées au domaine, respectivement « [. . .] le signe hiéroglyphique sa qui signifie la protection. » et « Les investissements dans le développement de la petite enfance pourraient signifier de meilleurs services [. . .] ». Ce deuxième type d'énoncé est « mis en réserve » pour créer deux patrons spécifiques à partir du même marqueur. Nous avons donc deux catégories dans notre évaluation : énoncé conforme à la relation sémantique annoncée, ou énoncé intéressant à un autre titre.

La validation des termes principaux (des unités lexicales, représentées par le sigle UL pour plus de concision) extraites de la phrase s'articule également en deux points : l'UL proposée est-elle pertinente, et son extraction est-elle correcte ? C'est-à-dire qu'un premier niveau de jugement consiste à vérifier que les UL extraites sont bien celles que l'on cherche à avoir à partir de l'énoncé définitoire. Par exemple, dans la phrase : « Le concept "éducation" est souvent défini de façon étroite, parfois même comme uniquement la scolarisation. », il s'agit de trouver au minimum éducation et scolarisation. Un deuxième niveau concerne la délimitation de l'extrait proposé par rapport à l'UL intéressante. Trois cas de figure peuvent se produire : ou l'extrait correspond exactement à l'UL, ou celui-ci englobe l'UL (nécessitant un nettoyage manuel de la séquence), ou celui-ci ne contient pas toute l'UL (et il faut rajouter manuellement les parties manquantes). Nous considérons que l'extrait correspond exactement à l'UL même s'il compte l'article précédant le nom ou le groupe nominal constituant l'UL.

5.5.2 Application aux deux corpus

Nous avons appliqué notre méthode et les patrons au corpus Petite Enfance (tableaux 5.5 et 5.6) et au corpus Diététique (tableaux 5.7 et 5.8). Les tableaux 5.5 et 5.7 présentent dans les premières colonnes le nombre de patrons lexico-syntaxiques correspondant à chacun des groupes définis plus haut (voir le paragraphe 5.4) : les patrons centrés sur des verbes métalinguistiques (ligne Méta1), sur des éléments lexicaux métalinguistiques combinés : un terme et un verbe métalinguistique (ligne Méta2), sur des éléments lexicaux non spécifiquement métalinguistiques (ligne Ling.), et enfin sur des marqueurs de ponctuation (ligne Ponct.). Les colonnes suivantes présentent le nombre d'extractions de phrases correspondant à la relation sémantique prédite par le patron (définition, hyperonymie, taxème), le nombre de phrases ne correspondant pas à la relation prédite mais jugées intéressantes (et servant de base pour raffiner le patron actuel suivant les différents cas). Les tableaux 5.6 et 5.8 concernent l'évaluation des unités lexicales : sont-elles correctes ou fausses, et, si correctes, sont-elles exactement extraites par nos programmes, incluses dans l'extrait proposé ou partiellement extraites ? Les données sont d'abord présentées sous la forme du nombre d'extraits validés, puis suivant leurs taux de précision.

Type	**Nb patrons**	**Nb extraits**	**Phrases suivant la rel. sem.**	**Taux de précision**	**Phrases intéressantes**	**Total**	**Précision des deux types d'énoncés**
Méta1	7	32	16	50 %	11	27	84 %
Méta2	20	14	7	50 %	4	11	78 %
Ling	24	97	31+7[2]	39 %	34	72	74 %
Ponct	4	79	22	27 %	7+45[1]	74	93 %
Total	55	222	83	37 %	101	184	82 %

TAB. 5.5 – Evaluation des extractions d'énoncés définitoires sur le corpus Petite Enfance

Type	**Nb patrons**	**Nb UL extraites**	**UL exacte**	**Taux de précision**	**UL incluse**	**Précision des deux types**	**UL incomplète**	**UL fausse**
Méta1	7	54	14	26 %	14	52 %	19	7
Méta2	20	22	2	9 %	8	45 %	3	9
Ling	24	144	6	4 %	48	37 %	6	84
Ponct	4	148	92	62 %	29	81 %	22	5
Total	55	368	114	31 %	99	57 %	50	105

TAB. 5.6 – Evaluation des extractions d'unités lexicales sur le corpus Petite Enfance

Au total, sur le premier corpus, la précision des énoncés correspondant à la relation sémantique (précision « stricte ») est de 37% et celle des unités lexicales est de 31%. Si l'on prend en compte également les phrases « intéressantes », cette précision s'élève à 82% : par exemple, 45 (correspondant au [1] du tableau) des phrases repérées par l'un des patrons de ponctuation apportaient une relation de traduction entre les deux unités lexicales. Cette relation n'était une hyperonymie, mais une reformulation dans une autre langue que l'on peut également considérer comme intéressante.

Les patrons des deux groupes métalinguistiques obtiennent la meilleure précision stricte (50%), mais un nombre absolu d'énoncés plus faible (23 au total contre 32 pour le groupe Ling et 22 pour le groupe Ponct). On retrouve une opposition classique entre rendement et précision. Cette observation peut aussi être liée au fait que les textes ne relevant pas de pratiques terminographiques auraient plus souvent recours à des gloses de reformulation pour introduire des énoncés définitoires, et ces gloses sont repérées par nos patrons des groupes Ling et Ponct. Par exemple, les énoncés définitoires sont plutôt introduits par des tournures comme « c'est-à-dire » (« l'embonpoint — *c'est-à-dire* la répartition harmonieuse du poids sur l'ensemble du corps — ») que par des expressions plus spécifiquement métalinguistiques (du type « Les berceuses japonaises sont *appelées* komori-uta, *terme qui désigne* proprement les "chansons de la garde d'enfant" »). En fait, il s'avère que les expressions métalinguistiques sont majoritairement employées dans ce corpus pour définir des termes en langue étrangère, afin d'introduire des conceptions liées à la petite enfance dans la culture japonaise, par exemple. Dans le cas des autres notions, comme pour « embonpoint », il s'agit de notions qui ont une acception courante en francais, et aui n'ont pas besoin d'être définis avec la même précision que des mots inconnus. Le lecteur est supposé avoir une intuition concernant la notion, et l'auteur n'a donc besoin que d'en préciser les éléments spécifiques à sa conception propre ou à celle qu'il décrit dans une culture particulière. Les mêmes modalités ne sont donc pas employées suivant le degré de spécialisation du vocabulaire. De plus, les définitions « bien formées » sont plutôt lourdes stylistiquement (elles créent une coupure dans la narration qui est plus importante qu'une simple glose de reformulation), et trouvent rarement leur place ailleurs que dans des ouvrages à vocation explicitement didactique. Hors des livres de cours ou autres ouvrages de ce type, les patrons d'extraction de glose de reformulation auront des chances de donner de meilleurs résultats. Mais comme ces patrons peuvent engendrer beaucoup de bruit informationnel, il peut être judicieux de les combiner avec des ressources sémantiques extérieures suceptibles de valider des relations proposées entre termes (utiliser un dictionnaire de spécialité sur l'anthropologie pour avoir des connaissances additionnelles sur les relations sémantiques potentielles entre termes, par exemple).

Concernant l'évaluation du système d'extraction, nous pouvons noter que certaines des unités lexicales incomplètes ne pouvaient être complétées dans le corpus par notre stratégie d'extraction parce que les données manquantes étaient implicites ou suivaient un principe de rattachement (anaphorique) complexe. Sinon, dans le cas de 7 énoncés ([2] du tableau), la compréhension globale du sens de la définition nécessitait également la lecture de la phrase précédente.

Une première remarque doit précéder l'examen des tableaux 5.7 et 5.8 : la conversion en format texte de certains des documents HTML du corpus a causé une segmentation excessive de certaines phrases. Des énoncés pertinents ont ainsi été coupés, empêchant les patrons de trouver les unités lexicales correctes et retournant des énoncés trop incomplets pour permettre une validation. Ces énoncés ont été considérés comme faux, faisant chuter

Type	**Nb patrons**	**Nb extraits**	**Phrases suivant la rel. sem.**	**Taux de précision**	**Phrases intéressantes**	**Total**	**Précision des deux types d'énoncés**
Méta1	7	29	20	69 %	2	22	75 %
Méta2	20	10	6	60 %	1	7	70 %
Ling	24	307	55	18 %	15	70	22 %
Ponct	4	365	38	10 %	30+39[3]	107	29 %
Total	55	711	119	16 %	87	206	29 %

TAB. 5.7 – Evaluation des extractions d'énoncés définitoires sur le corpus Diététique

Type	**Nb patrons**	**Nb UL extraites**	**UL exacte**	**Taux de précision**	**UL incluse**	**Précision des deux types**	**UL incomplète**	**UL fausse**
Méta1	7	44	14	31 %	4	41 %	12	14
Méta2	20	14	4	28 %	0	28 %	6	4
Ling	24	140	16	11 %	32	34 %	16	76
Ponct	4	214	100	46 %	56	73 %	13	45
Total	55	412	134	32 %	92	54 %	47	139

TAB. 5.8 – Evaluation des extractions d'unités lexicales sur le corpus Diététique

de manière drastique le taux de précision, qui n'atteint que 16% de précision « stricte ». Un choix de l'étiquetage a également induit un bruit non négligeable dans l'extraction des phrases autour de la parenthèse. En effet, certains documents contiennent des références de type *(A)* ; *A*, dans ce cas, est étiqueté comme nom commun par Cordial et la forme *SN (NomCommun)* étant un des patrons, des extraction incorrectes ont été générées.

Nous avons à nouveau été confrontée au problème lié à la prédiction de la relation sémantique d'hyperonymie (^{3} du tableau) : les extraits du corpus correspondaient cette fois à une expansion d'acronymes, qui peut être considérée comme une sorte de définition, mais de type synonymique. Nous pouvons par ailleurs remarquer que la qualité de l'extraction des UL dépend de la complexité syntaxique du patron. Plus le patron est de bas niveau plus les informations contextuelles sont efficaces pour l'extraction (elles donnent alors de bons résultats), plus il implique une description linguistique complexe, plus des informations de dépendance fonctionnelle entrent en jeu, compliquant la tâche. En effet, pour avoir une meilleure précision, il faudrait alors lister l'ensemble des formes syntaxiques que l'énoncé est susceptible d'avoir, multipliant par là même le nombre de nos patrons. Nous avons voulu rester dans une optique de factorisation, qui, si elle ne donne pas une précision optimale en termes d'exraction des UL, reste toutefois relativement indépendante des choix lexicaux et syntaxiques propres à un corpus particulier.

5.5.3 Conclusion de ces évaluations

Nous avons présenté une méthode ciblant en corpus des énoncés définitoires ou intéressants dans une perspective de construction ontologique. L'évaluation de cette méthode nous a permis de soulever certains points, notamment la difficulté d'avoir des taux de précision élevés lorsque l'on s'intéresse à des marqueurs linguistiques de reformulation plutôt que des unités lexicales métalinguistiques, remarque également formulée dans [Rebeyrolle, 2000].

5.5.4 Evaluation de la relation sémantique entre les deux UL extraites

Nous avons évalué plus spécifiquement la précision et une forme de rappel de la relation sémantique proposée par le système entre les deux UL extraites, lors d'une deuxième expérimentation. Cette fois-ci, nous cherchions à voir si les relations sémantiques associées aux différents patrons lexico-syntaxiques lors de l'analyse du corpus Petite Enfance restaient valides quand les patrons lexico-syntaxiques étaient appliqués à un autre corpus. Nous avons donc évalué la précision des relations sémantiques trouvées par rapport à la valeur de la relation attendue pour les énoncés définitoires extraits au moyen de nos patrons sur le corpus Diététique. Pour cette expérimentation, nous n'avons plus considéré que les relations d'hyperonymie et de synonymie (relations qui sont en général considérées comme centrales dans la modélisation terminologique).

Dans les 13 textes extraits du corpus Diététique, nous avons trouvé 90 énoncés définitoires de type hyperonymique et 22 de type synonymique. Les premiers énoncés représentent près de 45% des énoncés définitoires de l'ensemble considéré, et les seconds en représentent environ 11%. Comme certains énoncés définitoires comportaient plusieurs relations sémantiques, l'hyperonymie représentaient près de 40% des relations observées dans les énoncés, et la synonymie près de 10%.

	Hyperonymie	**Synonymie**
Nombre d'énoncés extraits	270	585
Précision(def)	61%	66%
Précision(rel)	26%	15%
Rappel(rel)	4%	36%

TAB. 5.9 – Evaluation de la relation sémantique véhiculée par l'énoncé définitoire

L'évaluation des extractions de SODA sur le corpus Diététique est présenté au tableau 5.9. La précision est divisée en deux mesures : Précision(def) correspond à la précision de l'énoncé définitoire (à savoir le nombre d'extractions correspondant effectivement à des énoncés définitoires) et Précision(rel) correspond à la mesure de la précision concernant la relation sémantique. Le rappel ne concerne que la relation sémantique, et correspond au nombre d'énoncés définitoires exprimés selon l'hyperonymie ou la synonymie divisé par le nombre de ces énoncés attendus.

Nous avons ensuite détaillé le comportement des différents patrons, en fonction de leurs marqueurs (voir **??**), et avons remarqué que certains patrons étaient plus polysémiques que d'autres. Cette constatation semble être une évidence, mais la surprise vient du fait que certains marqueurs de type métalinguistiques sont moins fiables que des marqueurs de type linguistiques pour l'extraction de la relation, et de l'énoncé définitoire. Il semblerait également que chaque patron véhicule une relation sémantique privilégiée sur un corpus donnée, ce qui implique qu'une fois que cette association est fixée (par l'analyse d'un sous-corpus, par exemple), elle pourrait permettre de prédire une relation sémantique entre termes de manière relativement stable.

5.6 Gestion des données validées

Comme nous l'avons vu plus haut, nous avons associé nos formulaires de validation à une base de données MySQL, où les données peuvent être modifiées, complétées et/ou exportées. Les deux formats d'export possible sont un format XML et un format texte. Le format XML correspond à une recommandation du W3C (format OWL), choisi pour pouvoir ouvrir les hiérarchies de termes et les manipuler dans des éditeurs d'ontologie acceptant ce format (DOE, Protégé2000, OilED, WebODE par exemple)[84]. Suivant la relation sémantique validée entre les unités lexicales extraites, le programme d'export XML structure les données hiérarchiquement (dans le cas de l'hyperonymie) ou associe les deux libellés à un même concept (dans le cas de la synonymie). La figure 5.3 présente un exemple de structure arborescente créée automatiquement à partir de la validation du formulaire de la figure 5.1.

L'export au format texte associe chaque terme validé à son énoncé définitoire, constituant une sorte de dictionnaire local. C'est sur des données sous ce format que nous avons réalisé l'expérimentation suivante, mais les données proviennent d'enoncés définitoires balisés manuellement sur le corpus Petite Enfance (et non pas d'extractions automatiques).

[84]Le format OWL correspond à un format d'ontologie formelle, dans l'idée de ses concepteurs. Nous en reprenons simplement l'architecture pour nos hiérarchies de termes non formelles, dans un souci de compatibilité avec les principaux éditeurs d'ontologie.

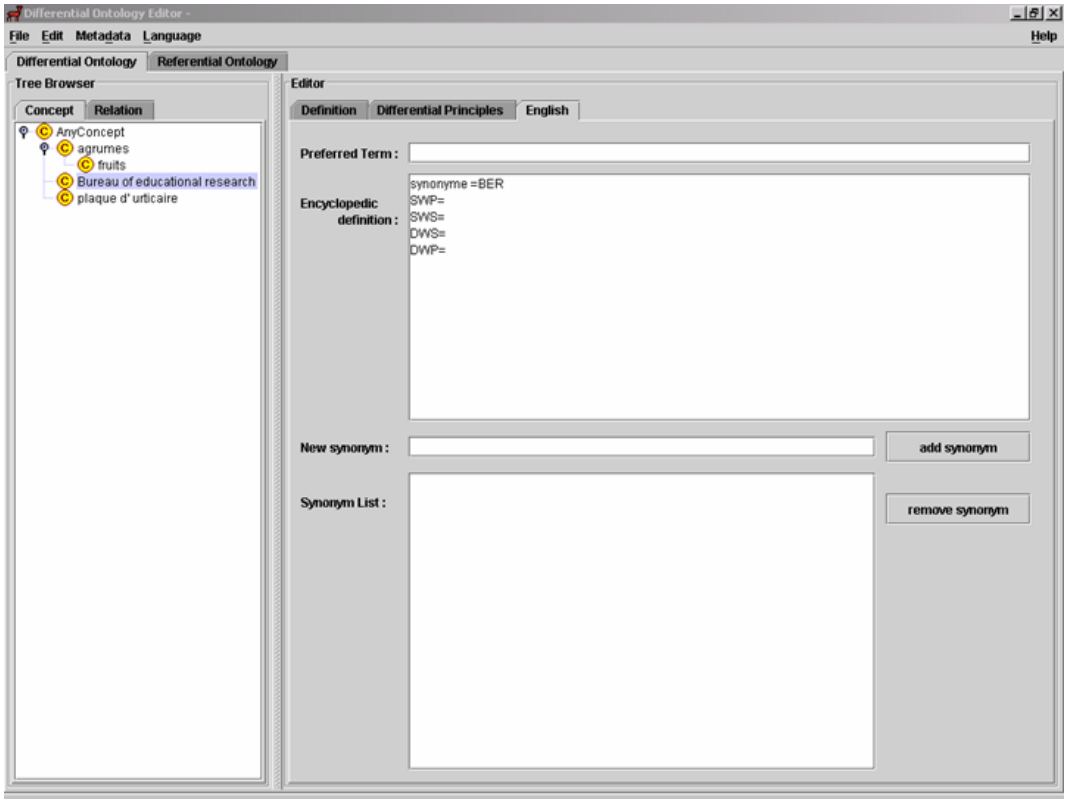


FIG. 5.3 – Hiérarchie terminologique construite automatiquement à partir de la validation d'un formulaire d'extraction

Cette expérimentation vise à rapprocher des co-hyponymes potentiels et à donner des éléments lexicaux pour la qualification de leur axe de similarité sémantique.

5.7 Structuration horizontale et recherche des axes différentiels à partir des énoncés à intérêt définitoires

L'un des axes les plus intéressants à exploiter pour la construction d'ontologies différentielles est l'utilisation du contenu lexical des énoncés définitoires pour aider à définir des axes de similarité sémantique entre frères. Nous faisons pour cela l'hypothèse que des mots « signifiants » partagés par deux énoncés définitoires peuvent servir de base à la construction du principe de similarité sémantique entre les termes définis par ces énoncés définitoires, que nous supposons alors pouvoir être comparables. La recherche de mots communs à des énoncés définitoires, si nos expériences vérifient notre hypothèse, devrait donc nous permettre de regrouper des termes en tant que frères ontologiques (les termes définis partageant des éléments de définition), et de donner des éléments lexicaux pertinents pour qualifier leur axe de communauté sémantique. La recherche de frères ontologiques potentiels aide le modéliseur à structurer l'axe horizontal de l'ontologie, et la recherche de principes différentiels est la partie la plus épineuse de la modélisation d'une ontologie différentielle.

Pour tester cette hypothèse, nous nous sommes fondée sur les formes « brutes » des énoncés définitoires annotés manuellement du corpus Petite Enfance, sans lemmatisation ni racinisation, mais avec un filtrage des mots outils par liste de mots vides. Cette liste contient des mots grammaticaux (déterminants, conjonction de coordination, etc.), des lettres isolées (présentes dans l'énumération de certains exemples) et des pronoms. Pour éviter l'écueil soulevé dans [Veronis & Ide, 1990], à savoir que le recouvrement lexical entre différentes définitions d'un dictionnaire est très pauvre (voir la section 4.4.1), nous avons suivi la solution proposée par ces auteurs et enrichi les énoncés définitoires initiaux des énoncés définitoires des mots composant ces énoncés. Par exemple, si l'on s'intéresse à la définition en dictionnaire de *Mère*, à savoir *Femme qui a mis au monde un ou plusieurs enfants* selon le Petit Robert, on voit qu'elle comporte le mot *enfants*. *Enfant* est associé par ailleurs, dans le même dictionnaire, à l'énoncé définitoire suivant : *Etre humain dans l'âge de l'enfance*. Le principe est alors d'augmenter le contenu lexical de la définition « originelle » de *Mère* par l'ensemble des mots de la définition de *Enfant* (voir le tableau 5.10), puis de procéder de même pour tous les autres mots de la définition « originelle » définis par ailleurs dans le dictionnaire.

Nous voyons que *Mère* est toujours associé à *Femme* et *enfants*, mais qu'à présent ce mot est en plus lié à *Etre humain* et *enfance*. Nous avons appliqué ce principe à nos énoncés définitoires, qui contenaient également des termes associés à des énoncés définitoires en corpus. Nous avons comparé ces énoncés définitoires « étendus » de manière automatique[85] et créé en sortie un tableau décrivant des proximités sémantiques potentielles entre termes associés à des énoncés définitoires. Le tableau de « proximité sémantique potentielle » est construit autour des termes partageant des mots signifiants dans leurs énoncés définitoires « étendus », et présente les termes en question, le nombre de mots que leurs énoncés

[85]Par un programme Perl.

Phase 1	Mère	Femme qui a mis au monde un ou plusieurs enfants
Phase 1	Enfant	Etre humain dans l'âge de l'enfance
Phase 2	Mère	Femme qui a mis au monde un ou plusieurs enfants Etre humain dans l'âge de l'enfance

TAB. 5.10 – Principe de construction de l'« énoncé définitoire étendu »

Terme 1	**Terme 2**	**Nb mots partagés**	**Mots communs**
activité paternelle	jeu	2	jouer bébé

TAB. 5.11 – Une ligne de la table de similarité

définitoires ont en commun et les mots qu'ils partagent. Le tableau 5.11 représente, par exemple, la ligne concernant les termes *activité paternelle* et *jeu*. Ils partagent deux mots dans leurs énoncés définitoires : jouer et bébé.

5.7.1 Evaluation de la recherche de frères ontologiques et des principes différentiels

Le tableau 5.12 montre le nombre de termes rapprochés par la méthode décrite plus haut, et par la programme informatique correspondant à son implémentation. La première colonne donne le nombre de termes partageant au moins un mot signifiant de son énoncé définitoire avec celui d'un autre terme ; la deuxième colonne donne l'intervalle de mots partagés par des énoncés définitoires dans ce corpus, et les troisième, quatrième et cinquième colonnes s'intéressent à l'évaluation de la qualité des termes rapprochés : nombre de termes qui sont des co-hyponymes corrects (3e colonne), nombre de termes partageant une autre relation sémantique (colonne 4) et nombre de termes ne partageant aucune relation sémantique, ou une relation très éloignée (colonne 5).

Deux questions peuvent être abordées à partir de ces données :

- d'un côté, y a-t-il un critère permettant de distinguer les couples de termes validés en tant que co-hyponymes des autres ? Nous allons examiner si les seuls critères présents dans ce tableau, à savoir le nombre de mots de leurs énoncés définitoires qu'ils partagent et la fréquence de ce ou ces mots partagé(s) peuvent être utilisés en tant que critères de distinction.
- et d'un autre côté, nous pouvons nous interroger sur la représentativité des mots que des énoncés définitoires de termes co-hyponymes partagent : sont-ils des candidats intéressants pour la construction du principe de similarité entre frères ontologiques ?

Les exemples présentés plus bas (tableaux 5.13 à 5.21) sont extraits du tableau de « proximité sémantique potentielle » (voir le tableau 5.11). L'analyse de ce tableau a montré que le rapprochement de deux termes pouvait être significatif même s'ils ne partagent qu'un seul mot dans leurs énoncés définitoires « augmentés » (tableau 5.13), et même si ce seul mot commun est très fréquent dans l'ensemble des énoncés définitoires, comme dans

Nb de termes partageant un ou plusieurs mots	**Nb de mots partagés**	**Nb de co-hyponymes corrects**	**Nb de termes partageant une autre relation sémantique**	**Nb de termes non reliés**
1400	17 à 1	351	390 (incluant 22 hyperonymes et 46 meronymes)	659

TAB. 5.12 – Evaluation du degré de similarité entre des termes partageant des mots dans leurs énoncés définitoires

activité paternelle	pratique de maternage	1	bain

TAB. 5.13 – activité paternelle et pratique de maternage sont proposés comme frères ontologiques, et leur point comun est le *bain* de l'enfant

l'exemple du tableau 5.14. 222 paires de termes ne partageant qu'un seul mot ont été jugés pertinentes, elles représentent 63.3% des co-hyponymes validés.

Cependant, l'association de deux termes sur la base d'un seul mot commun, qui plus est si ce mot est fréquent, peut également être une erreur (tableau 5.15). 601 paires de co-hyponymes invalides ne partageaient qu'un seul mot, ce qui représente 91.3% des paires de termes incorrectes.

Nous avons alors associé à chaque mot un indice de fiabilité, suivant leur fréquence d'apparition dans les énoncés définitoires : plus un mot est présent (et plus sa distribution est homogène dans le corpus), moins il est pertinent en tant que candidat pour rapprocher des co-hyponymes. Nous avons utilisé une formule dérivée de la mesure classique du tf.idf pour calculer cet indice. Cette formule est présentée ci-dessous, et représente le poids (W) associé à un mot « i » : *tf i,j* est la fréquence du mot « i » dans l'énoncé définitoire « j », *dfi* est la fréquence documentaire, c'est à dire le nombre d'énoncés définitoires différents contenant le mot « i » et N le nombre total d'énoncés définitoires.

$$W(_{i,j}) = 1 + log(tf_{i,j}) \times log\frac{N}{df_i}$$

Nous avons ensuite réalisé une classification ascendante hiérarchique à partir des données du tableau de contingence obtenu par ce calcul, tableau pondéré au moyen de l'indice de fiabilité que nous avons détaillé plus haut. Un extrait en est présenté tableau 5.16. Chaque ligne du tableau représente un terme associé à un énoncé définitoire dans le corpus

dormir	allaitement	1	bébé

TAB. 5.14 – Dormir et *allaitement* sont des co-hyponymes corrects, malgré le fait qu'ils ne partagent qu'un seul mot, *bébé*, qui est très fréquent en corpus

allaitement	activité paternelle	1	bébé

TAB. 5.15 – L'*allaitement* et l'*activité paternelle* sont proposés en tant que frères ontologiques

	Bain	Bébé	...
Activité paternelle	4,95	2,93	...
Allaitement	0	2,93	...

TAB. 5.16 – Extrait du tableau de contingence

et parageant au moins un mot dans sa définition avec un autre terme du corpus (afin d'avoir le minimum de valeurs nulles dans la matrice), et chaque colonne représente l'ensemble des mots partagés par au moins deux énoncés définitoires. L'intersection de chaque ligne avec une colonne symbolise la présence ou l'absence du mot dans l'énoncé définitoire correspondant au terme en question. La valeur de 0 dénote l'absence de ce mot dans l'énoncé définitoire, et une valeur numérique autre représente l'indice de fiabilité associé au mot présent dans l'énoncé.

Nous avons obtenu des groupes intéressants, regroupant par exemple les professions liées à l'enfance et mentionnées dans le corpus (deux clusters homogènes), ou encore les rituels et cérémonies. Cependant, l'ensemble des groupes ne sont pas homogènes. Pour obtenir un ensemble de classes cohérentes, il faut partitionner les termes en une centaine de classes, pour environ 300 termes, ce qui n'est pas sensiblement différent de rapprocher les termes deux à deux. Les deux groupes homogènes que nous avons mentionnés apparaissent déjà, quant à eux, avec un partitionnement en 10 classes, et se trouvent ensuite subdivisés (pas forcément de manière heureuse). Une interprétation possible de ce phénomène est que ces deux classes représentent les éléments centraux du corpus, à modéliser dans l'ontologie.

Par ailleurs, l'analyse de la table montre également que lorsque des termes sont liés par plus d'un mot, ces mots peuvent être pertinents pour construire l'axe de similarité entre les termes concernés (exemple du tableau 5.17). Cette pertinence décroit lorsque le nombre de mots partagés passe à un seul (tableau 5.18). Les termes associés par erreur étaient souvent ceux partageant un seul mot très fréquent, ou un mot polysémique comme dans l'exemple 5.19. Des techniques de désambiguïsation sémantique pourraient permettre de résoudre ce dernier problème.

Nous avons également constaté que, même en corpus, certains termes définis faisaient partie de plusieurs paradigmes. Par exemple, « alimentation » partage le sème « rituel » avec d'autres pratiques rituelles, et partage le sème « quotidien » avec d'autres actions

alimentation	traitement du placenta	4	rituel naissance pratique post-partum
centre de santé	poste de santé	2	structure de soins

TAB. 5.17 – « alimentation » et le « traitement du placenta » sont deux « rituel de naissance » et des « pratique post-partum » ; « centre de santé » et « poste de santé » sont deux « structure de soins »

jaune d'œuf	colostrum	1	corps

TAB. 5.18 – Le « jaune d'oeuf » est comestible, ainsi que le « colostrum », le premier lait sécrété par la mère après la naissance ; cependant l'élément lexical qui les rapproche est « corps »

axonge	possession	1	corps

TAB. 5.19 – « Corps » réfère tour à tour à un « corps gras », l'axonge, et au corps humain

de la vie ordinaire. En effet, l'« alimentation » peut être considérée comme une activité rituelle ou comme une activité quotidienne.

Comme nous l'avons vu au tableau 5.12, les paires de termes proposées peuvent être des co-hyponymes, des termes sans relation, mais également des termes partageant diverses relations sémantiques, comme l'hyperonymie (exemple tableau 5.20), ou la méronymie (exemple tableau 5.21).

Ce fait peut être considéré comme une des limites de cette approche, qui vise à associer automatiquement des co-hyponymes et à trouver la valeur sémantique qui justifie leur rapprochement. Mais le fait de trouver d'autres relations sémantiques est relativement courant dans les résultats d'extraction automatique (SynoTerm n'extrait pas *que* des synonymes, l'approche de [Zweigenbaum & Grabar, 2000] ne relie pas *que* des hyperonymes, par exemple). Le fait de trouver des termes en relation d'hyperonymie peu aider à renforcer la structuration verticale déjà amorcée lors de la validation des énoncés définitoires (au moyen des termes et relations sémantiques extraits du corpus), ou à proposer de nouvelles relations sémantiques.

La véritable limite est liée à la méthode elle-même, qui propose de comparer des unités lexicales. Les trois phrases suivantes, extraites du corpus d'évaluation, expriment le même genre de contenu sémantique (le definiendum est à chaque fois une maladie fréquente), mais ils ne partagent presque aucun élément lexical :

- *L'anorexie mentale* (AM) est une pathologie qui semble actuellement de plus en plus fréquente ;
- *Le diabète* est une maladie qui touche de plus en plus de citoyens gaspésiens, québécois et canadiens ;
- *L'hypertension artérielle* est une affection très fréquente, touchant plus de quinze pour cent de la population adulte.

Ces definienda pourront difficilement être rapprochés sur la base d'une comparaison lexicale. Mais, malgré la simplicité de la méthode testée, cette expérience nous a permis de lier 450 paires de véritables co-hyponymes, ce qui représente déjà le volume d'une ontologie de taille moyenne. Et de plus, cette approche nous donne la possibilité de trouver

alimentation de l'enfant	alimentation	1	pratique

TAB. 5.20 – « alimentation de l'enfant » est un hyponyme de « alimentation »

allaitement	pratique de maternage	2	joue joue

TAB. 5.21 – « allaitement » fait partie du « maternage »

une amorce aux principes différentiels nécessaires à la validation et à la structuration de l'ontologie différentielle. Elle peut également être une piste pour trouver les principes de différence entre co-hyponymes : pour deux termes validés en tant que frères, les éléments lexicaux différant d'un énoncé définitoire à l'autre peuvent peut-être aider à caractériser l'axe de différence entre frères ontologiques.

5.8 Evaluation des limites de cette approche

Nous avons déjà abordé une des limites de cette approche : le fait qu'elle est basée sur une comparaison de niveau lexical la rend intéressante comme première approche, mais demande de mettre en œuvre des moyens linguistiques plus conséquents pour être généralisée. Un des avantages de l'extraction d'énoncés définitoires est de centrer les extractions sur le point de vue des auteurs des documents : ce type d'énoncé traduit un apparté, une manière de souligner un élément important de la part d'un auteur. Un de ses inconvénients est le peu dénoncés retournés, même dans une annotation manuelle de corpus (c'est-à-dire indépendamment d'un biais constitué par faible rappel éventuel de nos programmes) : nos corpus ne sont pas explicitement didactiques et les énoncés définitoires concernent des notions bien spécifiques, mais peu nombreuses. Une des possibilités pour répondre à cette pénurie est le fait de combiner différentes approches : l'approche par patrons lexico-syntaxiques et énoncés définitoires donne un cœur intéressant et ciblé à la future ontologie, qui peut éventuellement être complété par d'autres techniques. Une des pistes que nous allons aborder concerne la compatibilité et la complémentarité de l'approche par analyse distributionnelle *à la Harris*, puisqu'elle fournit, elle, un grand nombre de candidats termes mais peu de pistes pour les structurer. L'idée serait alors, si possible, de compléter au moyen de ces candidats termes une structure minimale déjà hiérarchisée. Le choix de l'analyse distributionnelle comme complément au système testé dans SODA découle également du retour d'utilisation du système OPALES. Nous avons pu observer à quel point les deux points de vue (strictement hiérarchique et de type distributionnel) étaient complémentaires dans l'utilisation d'un tel système. Nous développons ce point au chapitre suivant.

5.9 Retour de l'évaluation du système OPALES par les utilisateurs

Les expérimentations préliminaires visant à élaborer une méthode TAL pour l'aide à la construction d'ontologies différentielles ont eu lieu lors de notre participation au projet OPALES, et si la méthode présentée ci-dessus (SODA) a été développée ultérieurement, c'est sur la base de cette première expérience qu'elle a été réalisée. Non seulement nous l'avons mise au point en testant nos diverses hypothèses sur le corpus qui a été constitué dans le cadre d'OPALES (corpus Petite Enfance), mais nous avons également pu conserver à l'esprit les enseignements liés au retour d'usage de l'ontologie Petite Enfance (réalisée

manuellement à l'époque du projet) par les utilisateurs d'OPALES pour l'annotation des documents audiovisuels. En effet, une ontologie différentielle est modélisée dans un but applicatif précis, il est donc important de l'évaluer en usage, et de vérifier son adéquation par rapport à l'objectif fixé lors de son intégration dans un système opérationnel. L'interface du sytème d'annotation d'OPALES a été développée par l'équipe du LIRMM de Montpellier, en collaboration avec Antoine Isaac et Raphaël Troncy (INA) pour ce qui est de la gestion de l'ontologie (visualisation, navigation, interrogation) ; et un des buts du projet était de tester l'intérêt des Graphes Conceptuels dans la gestion documentaire. Le retour des utilisateurs nous a renseigné sur deux aspects : les réactions des utilisateurs face à l'annotation au moyen de l'ontologie et des Graphes Conceptuel et l'apport du système en termes de précision lors d'une recherche documentaire.

5.9.1 Prise en main des Graphes Conceptuels et de l'ontologie par les utilisateurs

Si le côté graphique des Graphes Conceptuel a séduit les utilisateurs d'OPALES lors des séances de présentation du système, la rédaction des premières annotations n'est pas allée sans difficultés. L'inconvénient principal était la sélection des concepts et des relations dans l'ontologie pour réaliser les annotations. En effet, la première interface donnait accès à l'ontologie à partir de sa racine pour rechercher les concepts et relations propres à exprimer les connaissances désirées. Le chemin de la racine aux concepts effectivement manipulables pour décrire un document est relativement long (notre ontologie possède un niveau abstrait — ou ontologie de haut niveau, voir le pragraphe à ce propos à la section 2.2.2 du chapitre 2 — plutôt conséquent) et ardu à parcourir : les principes différentiels permettant de choisir de dérouler une branche plutôt qu'une autre sont très théoriques au niveau de la racine. Une liste des sous-concepts, des fils directs d'un concept, permettrait peut-être à un non spécialiste des ontologies d'avoir une vue d'ensemble donnant plus de corps à la définition différentielle. Par exemple, la définition différentielle permettant de distinguer des OBJETS TEMPORELS d'OBJETS SPATIAUX est la suivante : *Type de repérage privilégié de l'objet : Objet repéré temporellement.* Si l'on rajoute à cette définition les libellés des concepts fils d'OBJETSTEMPORELS, à savoir ACTION et PRATIQUE, cette définition devient plus intelligible.

Lors de la recherche de relations, les utilisateurs n'en vérifiaient pas la signature pour être sûrs de créer une association licite, mais se fiaient au libellé de la relation. Un des moyens de répondre à cette source d'erreur potentielle serait de présenter chaque concept sélectionné avec l'ensemble des relations sémantiques qui sont susceptibles d'en partir (les relations dont il est le domaine) ou d'y arriver (les relations dont il est le co-domaine).

Un autre accès possible aux concepts et relations de l'ontologie est la recherche par ordre alphabétique. Cependant, comme la synonymie n'était pas gérée dans cette interface de recherche, le fait de trouver un concept (ou une relation) impliquait d'en connaître l'existence, et de parcourir la liste alphabétique pour en trouver le libellé. Par exemple, le concept de PERSONNE peut également être libellé ÊTRE HUMAIN ou HOMME, ce qui pose un problème lors de la recherche alphabétique. Cependant, si le concept (ou la relation) est trouvé(e) dans cette liste, le fait de le(la) sélectionner dans l'interface de visualisation alphabétique permet de déplier automatiquement l'arbre ontologique correspondant jusqu'à son emplacement dans la hiérarchie. Une fois le concept(ou la relation) repéré(e) dans la hiérarchie ontologique, l'utilisateur accède aux différentes informations véhiculées

par le contexte du nœud choisi : père et frères ontologiques, principes différentiels pour les concepts, père, frères ontologiques et signature pour les relations.

La solution mise en œuvre dans le projet a été de passer par des Graphes Patrons. En effet, OPALES permet de s'approprier des annotations existantes[86], soit pour reprendre le point de vue d'un anotateur précédent et ajouter des précision ou amendements à sa vue, soit pour modifier une annotation propre réalisée antérieurement[87]. Lorsque l'on modifie un graphe, le fait de cliquer sur un concept ou une relation mène l'utilisateur à la place de cet élément dans l'ontologie. En utilisant cette propriété, nous avons défini un Graphe Patron en concertation avec les utilisateurs (pour plus de détails voir la section 2.2.1), qui est copiable pour créer de nouvelles annotations. Une fois ce Graphe copié dans l'espace de travail (la zone cerclée de rouge dans la figure 5.4), le fait de cliquer sur ses différentes composantes permet d'accéder à l'ontologie à un niveau de profondeur hiérarchique plus simple à manipuler et comprendre que le haut niveau : l'utilisateur accède directement au niveau des concepts propres au domaine. Il reste bien sûr toujours les possibilités de naviguer dans l'ontologie pour en sélectionner des concepts ou relations à partir de la racine et de rechercher un concept ou une relation dans la liste alphabétique.

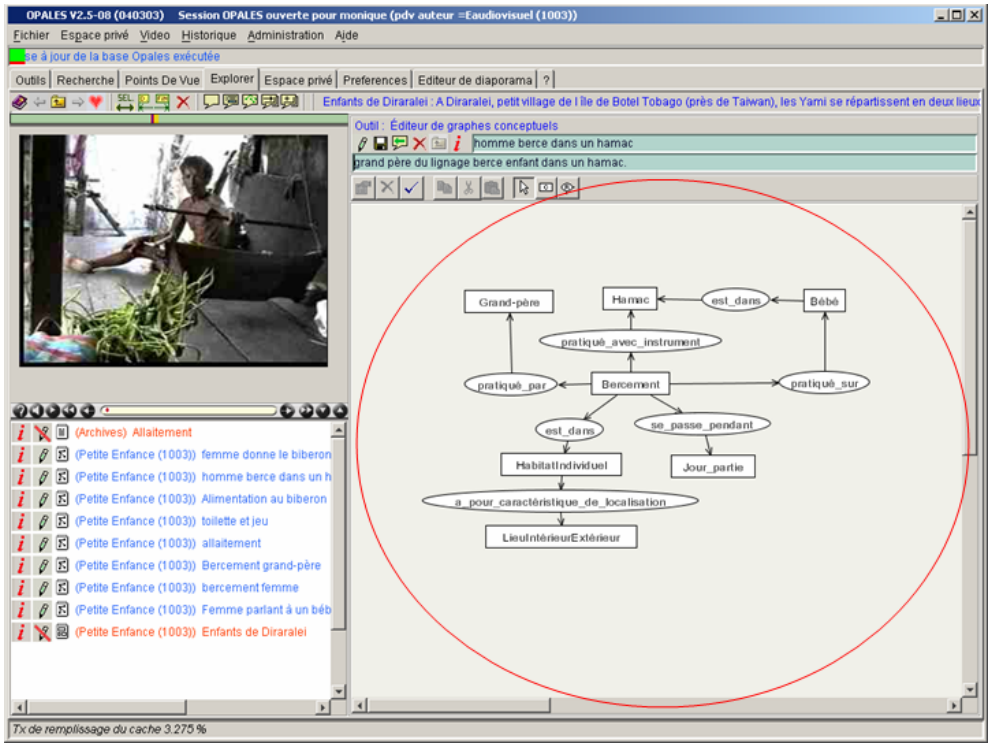


FIG. 5.4 – Interface d'annotation par Graphes Conceptuels d'OPALES

L'utilisation humaine d'une ontologie pour l'annotation de documents nécessite donc une association du modèle hiérarchique des concepts et relations avec une organisation du

[86] L'interface permet en fait de dupliquer une annotation réalisée par un auteur différent et d'en modifier la copie.

[87] Dans ce cas, l'annotation antérieure peut directement être modifiée (sans copie intermédiaire), puisqu'il s'agit d'une intervention du même auteur.

domaine plus proche des connaissances des spécialistes du domaine. Dans notre travail, il s'agit d'une organisation du domaine déductible de notre corpus selon les principes de l'analyse distributionnelle.

5.9.2 Précision lors de la recherche documentaire

Composer une requête au moyen de concepts simples, sans les lier par des relations explicites sélectionnées dans l'ontologie, revient à faire une recherche par mots-clés issus d'un thésaurus hiérarchisé (associant aux mots-clés leurs sous-concepts par une fonctionnalité d'« explosion »). Le rajout de relations permet de préciser la sémantique souhaitée dans l'association entre concepts, de faire le tri dans les réponses renvoyées et éventuellement de raffiner le sens d'une requête antérieure afin d'en améliorer la précision. Cette latitude dans l'interface de recherche permet de ne pas perdre l'acquis lié aux système documentaires plus classiques (la facilité de la requête simple), tout en offrant de meilleurs résultats en termes de précision si l'utilisateur souhaite opérer des requêtes plus poussées. Le rajout de relations à la requête n'est donc pas une contrainte, mais bien un potentiel supplémentaire.[Lespinasse, 2002] avait souligné qu'il n'était pas possible de faire la distinction entre des émissions *de* politique et des émissions *sur la* politique, lors de la constitution de son corpus de travail, avec les outils documentaires de l'INA. Ce type de distinction est possible dans OPALES, ainsi que la spécification de relations temporelles (rechercher des ENDORMISSEMENT après la TOILETTE au lieu de toutes les séquences traitant à la fois de ces deux types d'actions) ou spatiales (endormissement dans les bras du père, par exemple). Pour que les raisonnements documentaires adéquats fonctionnent à partir de ces relations, il faut une structure conceptuelle homogène, définie à partir d'une arborescence stricte dans notre cas.

L'utilisation de ce type de système pour des non spécialistes des sciences cognitives entraîne donc deux implications qui peuvent être contradictoires. Une organisation de type strictement hiérarchique est nécessaire pour la gestion informatique des données, et une mobilisation humaine de l'ontologie demande un accès aux données proche de l'organisation du domaine dont ils sont familiers, gardant la trace des relations transversales et hiérarchiques de l'ontologie, mais en évitant un parcours de l'arborescence depuis la racine. Nous avons vu dans nos premières expérimentations (chapitre 4) que l'*analyse distributionnelle* était propre à fournir ce type d'organisation du domaine. L'*extraction d'énoncés définitoires* réalisés selon un point de vue hyperonymique permet de structurer hiérarchiquement un ensemble de termes, ce qui aboutit à une structure propre au traitement informatique des données ontologiques. Nous avons donc voulu tester la compatibilité de ces deux méthodes et leur apport mutuel. Nous avons pu réaliser ce test, déjà entrepris par d'autres (comme [Caraballo, 1999] et [Cimiano *et al.*, 2004]), lors de notre collaboration au projet PERTOMed, que nous développons au chapitre suivant.

Chapitre 6

Mise en œuvre de SODA dans un projet médical et comparaison des résultats avec ceux donnés par l'Analyse Distributionnelle

Sommaire

Certaines données présentées dans ce chapitres sont reprises de la publication [Baneyx et al.*, 2005].*

Etant donné les intérêts respectifs des approches par analyse distributionnelle et par patrons lexico-syntaxiques pour la construction de ressources termino-ontologiques, ou plus précisément pour nous, pour la constuction d'ontologies différentielles, nous avons cherché à voir dans quelle mesure ces deux approches étaient compatibles. Certains travaux présentent une combinaison de méthodes statistiques ou issues du TAL pour la construction d'ontologies à partir de corpus (voir [Bourigault & Lame, 2002], [Caraballo, 1999] ou [Cimiano *et al.*, 2004], par exemple). Leurs résultats plaident également en faveur d'une approche mixte : ils obtiennent de meilleurs résultats en capitalisant les sorties résultant de différents traitements, ou en les associant dans une chaîne de traitement.

Nous avons pu tester l'intérêt et la compatibilité des résultats obtenus par analyse distributionnelle avec ceux obtenus par patrons lexico-syntaxiques lors de notre collaboration

au projet PERTOMed. Ce projet s'intéresse, entre autres, à la modélisation à la demande de ressources termino-ontologiques dans le domaine médical.

Il existe différentes terminologies dans le domaine médical. Nous avons évoqué le MeSH, utilisé pour l'indexation de bases documentaires ; d'autres structures terminologiques sont des classifications pour le codage de diagnostics (la Classification Internationale des Maladies, CIM10, en est un exemple). Certaines sont post-coordonnées et permettent la création de nouvelles notions par la combinaison de notions élémentaires (comme dans SNOMED), et certaines contiennent des relations typées pour lier les notions primitives en notions complexes, notamment la SNOMED-CT. Différents efforts visent à construire des structures terminologiques de plus en plus formelles et contenant des relations typées, comme la SNOMeD-CT. Et enfin, différents projets de recherche se basent sur de véritables ontologies pour la gestion du dossier patient ou l'aide au codage en cardiologie (projet Ménélas [Zweigenbaum *et al.*, 1994]), ou encore pour l'échange d'informations médicales non ambigües dans le projet GALEN [Rodrigues *et al.*, 1999]. Le projet PERTOMed se place dans la lignée de ces recherches, et vise à sélectionner un ensemble d'outils et de méthodes pour la construction et l'utilisation de ressources terminologiques et ontologiques dans le domaine médical. Nous avons collaboré, pour cette expérimentation, avec Audrey Baneyx et Jean Charlet, du laboratoire INSERM U729.

6.1 Le projet PERTOMed

Le projet de recherche PERTOMed (Production et évaluation de ressources terminologiques et ontologiques dans le domaine de la médecine, `http://www.spim.jussieu.fr/Pertomed`) est financé par le CNRS dans le cadre du programme TCAN[88]. 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. Face à la multiplication des sources d'informations potentiellement accessibles et face à l'augmentation de la production textuelle, les limites actuelles des outils de traitement de l'information ne se situent pas du côté de leurs performances pour stocker et traiter rapidement de gros volumes, mais de leur capacité à prendre en compte les spécificités des vocabulaires métier. 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, le travail auquel nous avons collaboré consiste à construire une ontologie différentielle dans le domaine de la pneumologie. Cette ontologie sera intégrée à 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, qui sera proposé à des pneumologues. Nous nous plaçons une fois encore dans le cadre de la construction d'ontologie à partir de corpus. Afin de sélectionner les éléments terminologiques pertinents au codage des actes médicaux en pneumologie, nous disposons d'un corpus de comptes rendus d'hospitalisation en pneumologie, rédigés par un ensemble de spécialistes dans différents hôpitaux, collectés et traités par Audrey Baneyx. Nous disposons également de la version éléctronique d'un livre de cours de pneumologie, deuxième

[88] Projet en cours.

corpus également fourni par Audrey Baneyx. Nous présentons ces deux corpus plus avant dans la section suivante.

6.2 Les corpus

Dans le but de couvrir autant que possible la terminologie en usage dans la description d'actes médicaux (en pneumologie), la future ontologie devant être exploitée dans une perspective d'aide au codage de ces actes médicaux, Audrey Baneyx a collecté des comptes rendus d'hospitalisation (corpus intitulé [CRH]) dans plusieurs hôpitaux de l'Assistance Publique-Hôpitaux de Paris, répartis comme suit : Créteil : 326 CRH, Hôtel-Dieu : 97 CRH, Kremlin-Bicêtre : 125 CRH, Pitié-Salpêtrière : 57 CRH, Saint Antoine : 372 CRH, Tenon : 61 CRH. Au total nous disposons de 1 038 CRH, ce qui est un volume raisonnable au vu de l'étude de [Le Moigno *et al.*, 2002], qui a établi que 600 comptes rendus (environ 350 000 mots) semblait être un minimum 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 correspond environ à 823 000 mots. Ces corpus sont sous des formats inexploitables par les outils d'analyse du langage. Ils ont donc été traités par ma collègue et moi-même : les fichiers ont été convertis au format texte, « nettoyés », anonymisés, segmentés, associés à des identifiants de section et de phrase, étiquetés, analysés morphosyntaxiquement par Cordial Analyseur de la société Synapse, 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 (Extensible Markup Language). Nous disposons de ces deux corpus sous deux formats : un format pour le traitement des données par SYNTEX, réalisé par Audrey Baneyx et un format XML pour le traitement des données par nos programmes d'extraction, dont nous nous sommes chargée.

6.3 L'expérimentation

Nous avons testé indépendamment l'analyse distributionnelle et l'application des patrons lexico-syntaxiques développés dans SODA sur les deux corpus (section 6.3.1). Nous verrons que leur genre et leurs caractéristiques les rendent chacun plus adaptés à un type de traitement. Nous avons alors voulu analyser la compatibilité des résulats de deux analyses différentes, réalisées sur deux corpus différents (section 6.3.2). Nous avons pour cela mesuré (manuellement) le recouvrement terminologique des deux structures obtenues à partir des corpus et la compatibilité des deux structures hiérarchiques.

6.3.1 Application des deux types de traitement aux deux corpus

Application de l'analyse distributionnelle aux deux corpus

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

1. Les termes extraits par SYNTEX-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[89] : 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.
3. 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], Audrey Baneyx s'est intéressée, dans un premier temps, aux 679 syntagmes nominaux ayant plus de 12 occurrences en corpus. L'analyse syntaxique permet de dégager trois grands axes particulièrement pertinents : les pathologies, les signes et les traitements/examens. L'interface de visualisation et de validation des résultats comporte un indice de « pertinence » qu'il est possible d'affecter à chaque candidat terme. Cet indice est compris entre 1 et 6. Ma collègue s'en est servi pour faire une classification préliminaire des candidats termes : l'indice de 1 est associé aux candidats termes jugés non pertinents, et les indices compris entre 4 et 6 sont affectés aux syntagmes sémantiquement proches de l'un des trois axes identifiés. Ce regroupement permet de commencer une première phase de travail sur le rapprochement par contexte et laisse 292 syntagmes nominaux sur lesquels élaborer le cœur de l'ontologie.

D'après les résultats de l'analyse distributionnelle, le candidat terme *cure de chimiothérapie* a le plus grand nombre de voisins en Expansion (52 voisins), et sa fréquence d'apparition est également la plus haute, avec 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, *Traitement* est le père ontologique de *Chimiothérapie*, le terme central du syntagme le plus fréquent *cure de chimiothérapie*, et *Bilan*, *Antibiothérapie*, *Injection* et *Radiothérapie* sont des frères ontologiques potentiels de ce concept : il s'agit d'un ensemble de traitements. Les voisins en Tête regroupent des principes médicamenteux : *Navelbine*, *Cisplatine*, *Doxorubicine*, *Taxotere*, *Carboplatine*, *MIP*. Une relation peut d'ores et déjà être envisagée entre les *Traitements* regroupés grâce aux voisins en Expansion et ces principes.

Cette méthode de regroupement donne de bons résultats et simplifie la tâche de modélisation d'un ingénieur de la connaissance. L'analyse distributionnelle sur le corpus [CRH] donne ainsi des candidats termes pertinents pour la construction de l'ontologie qui est saisie dans l'éditeur d'ontologie DOE [Troncy & Isaac, 2002]. Cette première phase de construction ontologique a donné lieu à la modélisation de 292 concepts primitifs.

[89] Il s'agit des candidats termes qui partagent respectivement des recteurs et des régis avec le candidat terme concerné. Pour plus de détail sur ces notions, voir le chapitre 3.

Application des patrons lexico-syntaxiques de SODA aux deux corpus

Nous appliquons sur le corpus [CRH] les patrons lexico-syntaxiques de recherche d'énoncés définitoires développés antérieurement (voir le chapitre précédent). 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 compétence, 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[90], sur un total de 199 extractions. Nous avons par exemple extrait la phrase suivante : « Perfusion de LAROXYL dans le cadre de *troubles du sommeil* (*insomnie*). ». 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 NC(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 : Traitement par PRINCIPE ACTIF(r), 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*, ... ;
- Et 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 première phrase (soit la phrase suivant le titre) par *Il s'agit de...* . Nous avons développé un patron reprenant la phrase précédente dans le cas où un énoncé potentiellement définitoire commence par *Il s'agit de*, mais nous nous heurtons ici à un problème de rattachement sémantique : la mention *Il s'agit de* ne se rapporte 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 *Il s'agit de*. Enfin, le troisième type d'erreurs rencontré rejoint le comportement que nous avions déjà observé sur le corpus Diététique. Il semblerait que le marqueur *indiquer* ne soit pas pertinent ou demande des contraintes explicites dans le domaine médical.

L'analyse des résultats montre 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, fait déjà soulevé antérieurement et développé dans [Condamines, 2003], que le fonctionnement de certains patrons est fortement lié à des différences de genres textuels.

Nous appliquons ensuite ces patrons au corpus [LIVRE]. Il relève d'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[91].

[90] 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 ».

[91] Ce qui représente une précision de 14.89 %, soit près de 15 % des extractions.

Type	Nb	Exemple
Termes identiques, à la normalisation terminologique près	3	*Asthme*[CRH] ***vs*** *asthme*[LIVRE], *Emphyseme*[CRH] ***vs*** *emphysème*[LIVRE]
Variantes lexicales comparables	3	*BronchoPneumopathie*[CRH]/*bronchopathie chronique obstructive*[LIVRE], *SaturationEnAir*[CRH] ***vs*** *saturation en oxygène*[LIVRE]
Niveaux de granularité différents	18	*Adenopathie*[CRH] ***vs*** *Adénopathies médiastinales*[LIVRE], *MaladieRespiratoire*[CRH] ***vs*** *maladies pulmonaires*[LIVRE]

TAB. 6.1 – Comparaison des termes des deux hiérarchies terminologiques

Nous avons suivi la méthode décrite au chapitre 5 pour exploiter ces énoncés définitoires. Pour rappel, les groupes extraits sont présentés au valideur dans une interface HTML : un formulaire où il est possible de les modifier, de valider les relations sémantiques[92] 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 World Wide Web consortium (W3C). Les hiérarchies créées sont visualisables dans un éditeur d'ontologie comme DOE.

6.3.2 Comparaison des résultats issus de l'application des deux méthodes aux corpus

Procédure de comparaison

Nous comparons les terminologies structurées construites par les deux méthodes précédentes (analyse distributionnelle des [CRH] et patrons sur le [LIVRE]) pour voir dans quelle mesure elles sont compatibles et complémentaires. Pour cela, nous comparons tout d'abord manuellement les 292 termes de la future ontologie (arborescence [CRH]) avec la structuration à un niveau issue de la validation des extractions d'énoncés définitoires (arborescence [LIVRE]), comprenant 119 candidats termes ou groupes syntaxiques plus larges. Nous trouvons 24 ensembles de termes comparables[93] à différents titres[94]. Ces variantes sont détaillées au tableau 6.1.

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

[92] Il s'agit d'un pluriel car nous ne pensons pas que la seule relation sémantique pertinente pour la construction d'ontologie soit l'hyperonymie.

[93] 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].

[94] Le fait que les termes soient effectivement comparable n'a pas été validé par un médecin, mais a été évalué en fonction du classement des différents termes dans le thésaurus MeSH.

définitoire ne met pas en relation deux « synonymes » au sens large), nous observons que les deux structures peuvent être identiques, complémentaires ou divergentes (voir tableau 6.2).

Les derniers exemples de ce tableau 6.2, 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] a cause d'un problème de cible : les termes plus simples peuvent correspondre à des notions supposées acquises au niveau d'étude auquel s'adresse ce manuel. Elle peut également être due au mode d'extraction de ces termes et marquer une des spécificités de l'extraction par patrons lexico-syntaxiques.

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

Comparaison des termes du corpus [Livre]

Nous cherchons à comprendre pourquoi les termes qui n'apparaissent que dans la structure terminologique construite à partir du corpus [Livre] ne se retrouvent pas dans l'ontologie [CRH]. Nous étudions, pour cela, s'ils ont une ou plusieurs occurrences dans le corpus [CRH], auquel cas l'analyse distributionnelle aurait dû les isoler. Il y a 83[95] 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 6.3.

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 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 Syntex. 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].

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

Nous nous intéressons ensuite aux 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 6.4.

[95] 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.

68 termes sont associés 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 6.5.

Nous détaillons ensuite l'analyse des 47 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 du type résolution d'anaphore), afin de savoir s'il faut augmenter notre système de nouveaux patrons, en adapter certains ou relâcher des contraintes (tableau 6.6). Une partie des 47 énoncés analysés sont extraits par notre système (23), mais n'ont pas été validés lors de l'analyse des réponses. En effet, ils donnent une définition partielle ou suivant un hyperonyme inattendu jugé non pertinent lors de la validation des formulaires.

6.3.3 Conclusion et perspectives à cette expérimentation

Nous avons présenté une expérimentation associant deux types de traitements TAL pour la construction de hiérarchies terminologiques. Chaque méthodologie est adaptée à un type et genre de corpus : l'analyse distributionnelle donne de bons résultats sur un corpus redondant et riche en termes spécialisés, et l'extraction par patrons lexico-syntaxiques est efficace sur un corpus didactique à la structure régulière. Nous avons montré qu'il existe une relative compatibilité entre les deux ensembles terminologiques extraits, mais surtout une complémentarité intéressante des structures arborescentes développées par ces deux méthodes, bien que les traitements aient porté sur des corpus différents. La divergence même des structures est un point intéressant, car elle dénote la possibilité d'organisations conceptuelles différentes au sein du domaine considéré, connaissance précieuse pour un ingénieur des connaissances. Nous avons également 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 sophistiquées d'appariement (des techniques mettant en jeu des connaissances de type dérivationnelles ou de résolution d'anaphores, par exemple), et, d'autre part, d'améliorer la précision (et peut-être le rappel) des patrons lexico-syntaxiques, ce qui implique leur adaptation à la spécificité du domaine médical. Toutefois, ces mêmes patrons permettent déjà de repérer d'autres relations propres au domaine médical, qu'il est alors intéressant d'isoler et de spécifier 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.

6.4 Mise en correspondance des résultats des deux traitments sur deux corpus différents

Comme nous l'avons vu en introduction de ce chapitre, certains travaux proposent également une approche associant plusieurs types de traitements. Guiraude Lame [Bourigault & Lame, 2002] cherche à typer au moyen de patrons lexico-syntaxiques certaines relations sémantiques entre termes extraits par Syntex, selon les principes de l'analyse distributionnelle. Sharon Caraballo [Caraballo, 1999] regroupe des termes issus

d'un corpus journalistique par l'application de patrons[96] et cherche des relations d'hyperonymie entre les clusters obtenus au moyen des patrons lexico-syntaxiques définis par [Hearst, 1992]. [Cimiano *et al.*, 2004] comparent, eux, les structurations issues de quatre approches différentes :

- l'application des patrons lexico-syntaxiques de [Hearst, 1992] ;
- une heuristique de classification proche de la structuration par analyse de l'inclusion lexicale, stipulant que si un terme T1 est inclus dans un terme T2, et que ce dernier est modifié par un terme ou un adjectif, il est possible de déduire la relation sémantique T1 *est-hyperonyme-de* T2, comme dans l'exemple : Conférence / Conférence internationale implique Conférence *est-hyperonyme-de* Conférence internationale ;
- une recherche généralisée sur Internet : les auteurs utilisent une des API proposées par Google pour dénombrer le nombre de fois où un certain nombre de patrons apparaissent sur Internet. Ces patrons sont de la forme *<T1>s such as <T2>*, où T1 et T2 sont extraits d'une ontologie de référence réalisée à la main par un cogniticien expert ;
- l'utilisation de la relation de subsomption présente dans WordNet.

Le but de ces auteurs n'est toutefois pas de fusionner différentes modélisations, ni de mettre à jour différentes modélisations possibles d'un domaine, mais de reproduire de manière aussi automatisée que possible l'ontologie de référence modélisée à la main. Ils ne valident donc pas la pertinence des différentes hiérarchisations obtenues, et leur problématique est, somme toute, assez éloignée de la nôtre. Cependant, elle montre toutefois que l'utilisation de ressources hétérogènes pour la génération de hiérarchies terminologiques donne de meilleurs résultats que l'utilisation exclusive d'une seule méthode. En effet, les auteurs concluent que les méthodes à base de patrons ont une précision intéressante mais un faible rappel, du fait du petit nombre d'occurrences de ces patrons, alors que les méthodes impliquant des ressources comme le Web ou WordNet ont un rappel conséquent, mais une précision très faible. Cette conclusion concernant le rapport entre la précision et le rappel est un grand classique en recherche d'information, et peut être rapprochée de notre propre constatation concernant l'extraction d'énoncés définitoires et l'utilisation de l'analyse distributionnelle pour la construction d'ontologies différentielles.

Ces différentes expérimentations nous confortent dans l'idée de la complémentarité des deux approches que nous avons comparées. Mais elles ne nous donnent pas de piste concrète quant à la manière de les mettre en rapport et à leurs intérêts respectifs. Nous allons reprendre ces deux points ci-dessous, avant d'entamer une discussion concernant les perspectives possibles à ces développements et, d'une manière plus générale, à ce travail (chapitre 7).

6.4.1 Pistes pour la mise en correspondance de ces deux méthodes de structuration terminologique de manière plus automatique

Dans la littérature, l'idée est de partir des résultats donnés par l'analyse distributionnelle et de typer les relations entre termes proches. Nous suggérons plutôt de construire

[96] Elle regroupe les termes partageant un maximum de « conjonctifs » et d'« appositives » (des termes en relation du type *A et B* ; *A ou B* ; *A, anciennement B*) dans son corpus. Nous verrons (dans le chapitre 7) que nous avons également modélisé un certain nombre de ces patrons dans l'espoir de trouver des co-hyponymes ontologiques.

une première structure avec les termes les plus saillants du domaine, sélectionnés au moyen d'énoncés définitoires. En effet, ces énoncés mettent l'accent sur

- les notions explicitement définies dans les corpus de spécialité, qui devraient correspondre à des termes du domaine ;
- les notions ayant une acception différentes en corpus de spécialité (ce qui met également en évidence un terme du domaine), au moyen des énoncés moins formels de reformulation, de paraphrase et d'exemplification.

Cette première structure terminologique peut ensuite être complétée par le jeu de navigation entre Tête et Expansion de syntagme, pour les groupes réguliers et productifs, ou entre fonctions syntaxiques dans les dépendances données par les outils d'analyse distributionnelle. Pour pouvoir passer d'un ensemble de termes à l'autre, une association basée sur la morphologie (avec des regroupements de variantes morphologiques comme ceux proposés par FASTR) ou la structure interne doit être envisagée. Une méthode complète mettrait donc en œuvre l'ensemble des trois méthodes de structuration de terminologie émergeant en TAL.

6.4.2 Intérêts respectifs de ces deux méthodes

Si une structure arborescente est conforme aux principes théoriques sous-tendant la construction d'ontologies différentielles, nous avons vu (en section 5.9 du chapitre précédent) que ce n'est pas la forme idéale de navigation pour la sélection humaine des données. Il est beaucoup plus naturel pour les utilisateurs de naviguer dans les données issues de l'analyse distributionnelle, car elles permettent de refléter une organisation du domaine de type syntagmatique, de par leur construction : si les classes issues de cette méthode regroupent des termes liés de manière paradigmatique, le lien entre les différentes classes est issu de leur association syntagmatique. Il est par exemple possible d'associer des classes de personnes à des classes d'actions par le fait que les premières sont régulièrement le sujet des secondes. Il pourrait donc être intéressant d'associer une structure transversale navigable à l'arborescence des concepts de l'ontologie, afin de permettre à un utilisateur non seulement de comprendre la sémantique d'un concept en accédant à sa hiérarchie, mais également par l'accès direct aux relations dont ce concept est le domaine ou le co-domaine. L'interface d'annotation devrait pouvoir proposer l'ensemble des relations partant d'un concept donné, ou y arrivant, pour permettre de construire des connaissances rapidement, plutôt que de à naviguer tour à tour dans les arborescences respectives des concepts et des relations lorsque l'on cherche à les lier dans un graphe. De même, à la sélection d'une relation, les principales branches des concepts (du domaine) source et cible devraient être sélectionnables, d'après sa signature.

Il faut également offrir à la personne qui se sert d'une ontologie, dans un système donné, un accès au niveau des connaissances du domaine. Le haut niveau de l'ontologie doit pouvoir être consultable, pour comprendre la sémantique des concepts plus bas dans l'arborescence, mais il n'est pas utilisé dans des descriptions concrètes. Ces concepts de haut niveau n'étant en principe pas capturés par l'analyse distributionnelle, ce type d'analyse pourrait donner des critères objectifs pour définir les connaissances du domaine, et permettre de fournir un accès au niveau adéquat de spécialisation des concepts d'une ontologie pour une application donnée.

Pour permettre un accès de type alphabétique efficace aux concepts d'une ontologie, il faudrait également mettre en place une gestion efficace des synonymes, pour associer autant de libellés que possible au terme sélectionné pour représenter un concept dans l'ontologie.

L'analyse distributionnelle pourrait permettre de créer les réseaux de concepts et de relations que nous venons d'évoquer, ainsi que les associations de type synonymiques entre différents termes. Une approche par patrons permettrait, elle, de créer la version « informatique » des données, et de donner des pistes pour la lexicalisation des principes de similarité et de différence sémantique entre concepts différentiels. A partir du moment où l'on pose que les stuctures terminologiques fournies par ces deux approches sont complémentaires, ce type d'association est raisonnablement envisageable.

Nous allons à présent aborder les perspectives que l'on peut envisager au système d'extraction et d'exploitation d'énoncés définitoires présenté, et nous concluerons sur les perspectives globales offertes par la recherche présentée dans ce mémoire.

Type	**Exemple**	**Commentaire**
Identique ou comparable	*Broncho pneumopathie/Asthme* [CRH] ***vs*** *Bronchopathie/Asthme à dyspnée continue* [Livre]	Pour comparer ces deux hiérarchies, nous avons regardé comment ces trois notions étaient organisées dans le MeSH. Les deux premières sont classifiées sous *Poumon, maladie*, alors qu'*Asthme* 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	*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 de *Insuffisance ventriculaire droite*.
Divergente	*EtatPathologique/ MaladieRespiratoire / Bronchite* ET *Signe / Toux* [CRH] ***vs*** *Toux avec expectoration / Bronchite chronique* [Livre]	Dans le MeSH, la toux est classifiée à la fois comme *signe-symptôme* et comme *pathologie* : les deux sources textuelles illustrent chacune un de ces aspects. Nous avons choisi de privilégier le point de vue abordé dans le corpus [CRH].
	EtatMorphologique / AnomalieMorphologique / Lesion / Atelectasie – Adenopathie [CRH] ***vs*** *Opacité médiastinale / Adénopathies médiastinales* ET *Opacité dense arrondie / Atéléctasie par enroulement...* [Livre]	Dans les deux cas, *Atélectasie* et *Adénopathie* sont co-hyponymes, mais leurs hyéronymes sont contradictoires : dans l'arborescence [CRH], les *Opacités* sont classifiées sous *Signes*. Il s'agit d'une erreur d'interprétation de la relation sémantique dans les extractions à partir du corpus [Livre]. Une *Opacité* est bien un élément d'une image médicale qui est interprétée comme le *signe* d'une *adénopathie*. L'extrait du corpus était toutefois ambigu : *Adénopathies médiastinales. Il s'agit des opacités médiastinales les plus fréquentes...*

Tab. 6.2 – Comparaison des deux structures terminologiques

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] (*spondylarthrite* [Livre] *vs spondylarthropathie* [CRH]), ou présence de la tête du syntagme complexe (*mésothéliome malin diffus* [Livre] *vs mésothéliome ou mésothéliome pleural* [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. 6.3 – Comparaison des termes propres à la terminologie construite à partir du corpus [Livre] avec le corpus [CRH]

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

Tab. 6.4 – Comparaison des termes propres à l'ontologie basée sur les corpus [CRH] et [Livre]

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 retrouvées 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. 6.5 – 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 *repose sur* la cortico-thérapie...

TAB. 6.6 – Évaluation des énoncés à intérêt définitoire pouvant être renvoyés par des patrons lexico-syntaxiques

Chapitre 7

Conclusions et Perspectives

Sommaire

Nous avons voulu présenter dans ce mémoire l'intérêt des ontologies par rapport aux systèmes à base de mots-clés pour la gestion documentaire, mais surtout faire une proposition pour l'aide à la modélisation d'ontologies différentielles à partir de corpus textuels.

Cette modélisation s'appuie sur une méthodologie éprouvée dans le domaine du Traitement Automatique des Langues : l'utilisation de patrons lexico-syntaxiques. Dans notre cas, nous nous sommes plus particulièrement intéressée à l'extraction d'énoncés définitoires par ce moyen. En effet, les énoncés définitoires permettent de repérer des termes spécifiques à un domaine, des relations sémantiques entre termes propres à aider à la structuration d'une terminologie (structuration assez rigoureuse pour que la terminologie en question puisse servir de base à une ontologie) et de donner des pistes pour la sélection de principes différentiels entre termes. Ce dernier point est une des difficultés particulières de la construction d'ontologies différentielles.

Cette méthode présente des limites, notamment le faible nombre d'énoncés renvoyés à partir d'un corpus, mais peut être couplée à d'autres méthodologies issues du TAL. L'analyse distributionnelle est un candidat particulièrement intéressant dans cette optique parce que, d'une part elle propose un grand nombre de candidats termes, structurés en réseau de dépendances, et d'autre part parce qu'elle permet de reconstruire une vision « syntagmatique » du domaine proche des conceptions des spécialistes. Comme l'ontologie s'adresse à ces spécialistes (ce sont eux qui se serviront de la structure conceptuelle dans une application concrète), l'approche distributionnelle est une clé d'entrée plus adaptée que le parcours de l'arborescence des concepts pour la manipulation de l'ontologie.

Les premières perspectives liées à ce travail peuvent se concevoir à deux niveaux : au niveau des développements de la chaîne de traitements créée (SODA) et au niveau de la mise en correspondance des deux types de modèles à associer (c'est-à-dire l'arbre hiérar-

chique et le réseau des dépendances syntaxiques). Nous avons également envisagé le cahier des charges d'une maquette pour une manipulation humaine intuitive d'une ontologie différentielle, et nous terminons ce mémoire par une brève présentation d'autres applications qui pourraient se servir des programmes d'extraction d'énoncés définitoires que nous avons développés.

7.1 Perspectives en termes de développement

Nous avons évoqué plusieurs fois le fait que les énoncés définitoires étaient réalisés selon différentes relations sémantiques (chapitre 5 section 5.1), et que l'ensemble de ces relations sémantiques étaient intéressantes à intégrer à l'ontologie. Les relations autres que l'hyperonymie ou la synonymie représentent des moyens d'exprimer des connaissances du domaine et peuvent, dans certains cas, donner des arguments objectifs pour la modélisation des hiérarchies. Par exemple, les deux modélisations présentées figure 7.1 sont toutes les deux valides d'un point de vue sémantique : elles représentent deux manières différentes de structurer les concepts mis en relation. Cependant, si une relation transversale, *se-déplacer-vers* par exemple, doit être rattachée à une de ces deux arborescence, vu qu'elle ne s'applique pas au concept de PLANTE, la modélisation de gauche est plus adaptée, et permet le rattachement présenté en figure 7.2. La relation transversale et la logique de son rattachement conceptuel permet de trancher entre ces deux modélisations potentielles concurrentes.

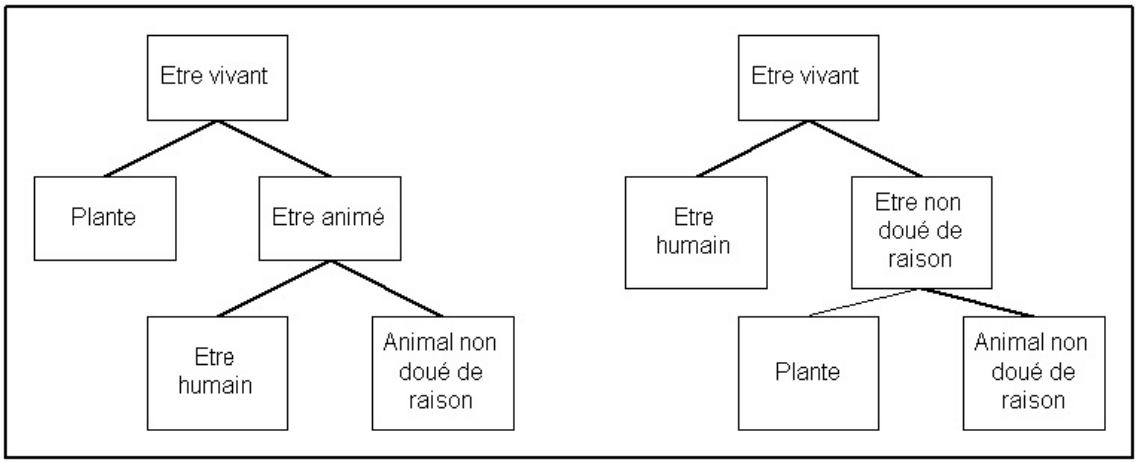


FIG. 7.1 – Deux modélisations arborescentes différentes des mêmes concepts

Une extension du modèle présenté devrait donc aussi prendre en compte les relations transversales, et pour cela, un nouveau groupe de patrons devrait être modélisé autour des marqueurs définitoires de méronymie, de caractérisation, de « fonction »,...

D'autre part [Caraballo, 1999] et [Bourigault & Lame, 2002] s'intéressent aux conjonctions de coordination *et* et *ou* pour trouver des termes entretenant des relations sémantiques. Cette relation s'apparente majoritairement, d'après nos propres expérimentations, à de la co-hyponymie ou à de la synonymie. Nous avons modélisé un ensemble de patrons autour de *ou* et de *soit X, soit Y*, qui donnent lieu à la construction d'un axe de co-hyponymie

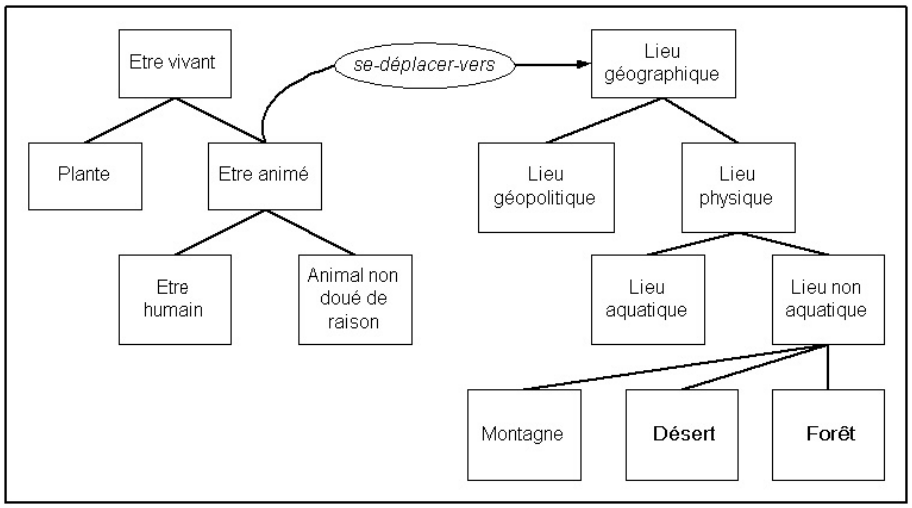


FIG. 7.2 – Choix d'une modélisation grâce à l'ancrage de la relation *se-déplacer-vers*

s'ils sont validés. La création de patrons autour de *et* est plus complexe, et celle autour d'environnements de listes (de la forme *des X, Y et Z*, par exemple) a été envisagée mais non implémentée. D'après le libellé des termes rapprochés en tant que co-hyponymes, s'ils ont un hyperonyme validé par ailleurs, l'éditeur DOE les classe dans la hiérarchie correspondante. Ce classement automatique sur la base du libellé du concept peut être source de bruit, du fait de la polysémie possible de ce libellé, et devrait être pris en charge par un algorithme de vérification particulier, que nous n'avons pas encore mis au point. Cet algorithme devrait idéalement permettre d'associer un concept à son hyperonyme même si leurs libellés sont différents. Cette fonctionnalité pourrait être gérée automatiquement par un classifieur sur la base des définitions formelles des différents concepts. La question à résoudre alors est l'éventualité (et les modalités) du passage de la définition terminologique associée aux libellés de concepts à cette définition formelle.

Et enfin, un système d'apprentissage pourrait être associé à l'extraction des énoncés définitoires pour en sélectionner les plus pertinents de manière (semi-)automatique, lors de leur application à un nouveau corpus. Un système de classement des réponses renvoyées serait également appréciable. Il peut se baser sur le type des marqueurs du patron (méta-linguistique, linguistique ou de bas niveau, par ordre décroissant de fiabilité), le nombre des marqueurs trouvés dans l'énoncé (de un à trois, par ordre croissant de fiabilité), et/ou de nouveaux critères (ou une combinaison de critères) élaborés par apprentissage. Ces critères peuvent être d'un autre niveau linguistique que les critères lexicaux évoqués dans cette étude (critères de « bas niveau », utilisant des informations de type ponctuation ou des statistiques données par Cordial sur les textes analysés, ou critères de « haut niveau »). Et enfin, des techniques d'apprentissage, ou l'utilisation d'outils comme Caméléon et Prométhée, qui calcule des rapprochements entre patrons potentiels ou permet d'en « apprendre » sur corpus, devraient permettre d'adapter à moindre coût les patrons que nous avons développés dans le contexte de nouveaux corpus, ou permettre d'en modéliser de nouveaux à partir des marqueurs de base, sur un nouveau corpus.

7.2 Amorce d'un cahier des charges pour la comparaison des résultats de patrons lexico-syntaxiques et de l'analyse distributionnelle

La première des modifications à apporter au système actuel pour permettre une meilleure interopérabilité entre lui et les résultats de l'analyse distributionnelle serait d'intégrer une extraction à deux niveaux : une extraction des formes du texte pour la validation humaine des données, et une extraction de la forme lemmatisée des textes et unités lexicales sélectionnées, qui seront, elles, stockées dans la base de données. Comme l'analyse distributionnelle réalisée dans SYNTEX et UPERY stocke également les résultats dans une base de données, la comparaison des formes lemmatisées des deux types de traitements serait déjà un premier pas vers leur utilisation complémentaire.

Ensuite, plusieurs types de traitements peuvent être envisagés, notamment la recherche de synonymes, une analyse de type morphologique et la décomposition des termes complexes.

La recherche de synonymes permettrait de regrouper des termes avec des libellés différents. Elle pourrait être mise en œuvre par exemple en utilisant une analyse distributionnelle de « deuxième ordre », c'est-à-dire en créant des classes distributionnelles de cooccurrents *de cooccurrents* (voir l'expérimentation de [Bertels, 2005] pour un appreçu de la méthode et de ses résultats), ou au moyen de synonymes donnés par des ressources extérieures au corpus, comme dans les recherches de [Hamon, 2000].

Une analyse morphologique comme celle réalisée dans FASTR permet également de regrouper des variantes terminologiques autour d'une forme canonique : les adjectifs dérivés d'un nom, par exemple. L'utilisation d'une telle ressource, ou d'une analyse de ce type serait certainement profitable au regroupement des deux types d'analyse qui nous intéressent, comme tendent à le montrer les travaux de [Bourigault & Jacquemin, 1999] : les deux auteurs montrent la complémentarité des traitements de FASTR par rapport aux extractions de LEXTER.

Et enfin, l'analyse distributionnelle se base sur un réseau de termes décomposés en fonction de leurs dépendances syntaxiques, créant des groupes à partir de candidats termes simples. Pour comparer ces résultats avec notre propre système d'extraction, et les enrichir au moyen des informations sémantiques contenues dans les dépendances calculées par SYNTEX, il nous faut en décomposer les termes complexes extraits. Pour cela, les travaux autour de la relation sémantique entre un terme inclus dans un terme complexe, réalisés notamment par [Zweigenbaum & Grabar, 2000] et [Ibekwe-SanJuan, 2005], seront une source d'information précieuse.

Ces propositions visent à permettre un rapprochement entre le contenu lexical de deux structures différentes, en se focalisant sur l'appariement de termes entretenant des relations sémantiques de (para)synonymie ou des relations syntaxiques d'inclusion lexicales. Au niveau de la complémentarité des approches, il serait appréciable d'inscrire les termes extraits par patrons dans leur réseau de dépendances syntaxiques pour construire une vue du domaine intuitive aux utilisateurs. Inversement, les relations sémantiques validées lors de l'extraction par patrons pourraient servir à typer certaines des relations sémantiques proposées par SYNTEX, non typées par défaut. Sur le plan de l'appariement des structures mêmes, les travaux de la communauté du Web Sémantique sur la fusion (ou « mapping ») d'ontologies peut également être une source d'inspiration.

7.3 Amorce d'un cahier des charges pour un outil d'annotation fonctionnant sur la base d'une ontologie différentielle

D'un point de vue ergonomique, nous avons également abordé le fait que, si une structure strictement hiérarchique (une structure verticale constituée exclusivement de la relation d'inclusion) est le fondement d'une ontologie différentielle, cette structure n'est pas idéalement adaptée à un parcours humain des données.

Une interface d'accès aux données à un niveau médian (le niveau des termes du domaine) est préférable pour l'utilisation humaine de ce type d'ontologie. Cette interface, pour être proche de la conception que les utilisateurs ont du domaine, peut se fonder sur les résultats de l'analyse distributionnelle pour présenter les catégories principales et l'organisation sytagmatique des termes telle qu'elle est représentée dans les documents de référence. Cette interface correspondrait aux Graphes Patrons que nous avons créés pour la manipulation des données dans le projet OPALES. Elle devrait être liée à l'arbre ontologique, et en permettre le parcours. Pour une meilleure compréhension des libellés des concepts et une manipulation plus simple des données, la navigation dans l'arbre devrait être proposée de manière hiérarchique mais aussi de manière transversale : il devrait être possible de naviguer de concept en concept non seulement selon la relation père-fils, mais aussi selon les *autres* relations présentes dans l'ontologie. Pour l'instant, DOE ne permet qu'une navigation parallèle dans l'arbre des concepts et l'arbre des relations. Un lien « physique » entre les deux hiérarchies serait profitable à une utilisation humaine de l'ontologie.

7.4 Utilisation des programmes dans un autre cadre applicatif

Les programmes développés au cours de cette thèse peuvent également servir dans d'autres cadres que celui pour lequel ils ont été initialement envisagés. Ces programmes ont par exemple été réemployés dans un système de questions-réponses. Ils avaient alors vocation à proposer des réponses pour des questions de type définitoires, ciblées sur le domaine médical (voir [Malaisé *et al.*, 2005]).

Ils peuvent également servir à mesurer le degré de spécificité de certains termes en corpus : les termes associés à des énoncés définitoires peuvent être recherchés dans un dictionnaire de langue générale ou de spécialité, et ses définitions (éventuelles) comparées à l'énoncé définitoire. Ils peuvent également servir à augmenter un ensemble de définitions, par exemple dans un contexte de traduction, en extrayant les énoncés définitoires relatifs à des néologismes.

Annexe A

Exemple de notice issue du système documentaire de l'INA

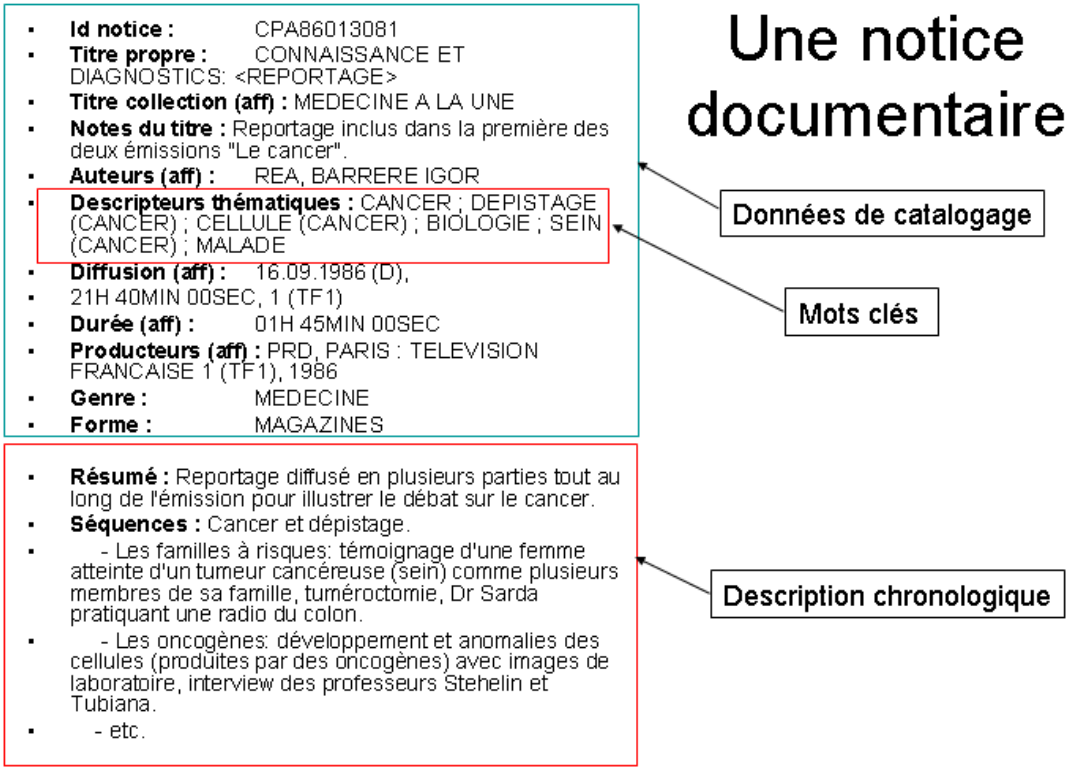


FIG. A.1 – Notice documentaire

Annexe B

Liste des patrons lexico-syntaxiques

Les trois tableaux suivants sont respectivement une légende des abréviations et du système de notation utilisés pour transcrire les patrons lexico-syntaxiques la liste des patrons lexico-syntaxiques implémentés pour la recherche d'énoncés définitoires (tableau en deux parties) et la liste des patrons lexico-syntaxiques modélisés pour la recherche de taxèmes.

A P X	A précédé de X
A NP X	A non précédé de X
A NP[nb] X	A non précédé de X à la position nb (X n'est pas le nbième mot/lemme/catégorie morpho-syntaxique précédant A)
A NP[nb1-nb2] X	A non précédé de X entre la position nb1 et nb2
A S X	A suivi de X
A NS X	A non suivi de X
A NS[nb] X	A non suivi de X à la position nb (X n'est pas le nbième mot/lemme/catégorie morphosyntaxique suivant A)
A NS[nb1-nb2] X	A non suivi de X entre la position nb1 et nb2
CatMS	catégorie morphosyntaxiques ()
DET	déterminant
NC	nom commun
PCT	ponctuation
VB	verbe
VINF	verbe à l'infinitif
VIND	verbe à l'indicatif
VPARP	verbe au participe passé
MOT	un mot quelconque
CHIFFRE	chiffre entre 0 et 9
ET	suite du patron lexico-syntaxique
(A|B)	A ou B
A1,3	1 à 3 répétitions de A
A[CatMS :(VIND|VPARP)]	A ayant comme catégorie morphosyntaxique « verbe à l'indicatif » ou « verbe au participe passé »
sorte	lemme de sorte
« sorte »	forme graphique de sorte

Tab. B.1 – Légende des abréviations et des notations

Marqueur	patron
Une sorte de	sorte P DET
Par exemple	exemple P[1] par ET P[2] MOT ET S[2]MOT
Etre un	être NS[1] pas ET P[1] « ce » ET P[2] MOT ET S[1-3] (« des »|« les »|« un »|« une »|« le »|« la »)
A savoir	savoir P[1] « à » ET NP[2] CatMS[PCT]
(	(P[1] CatMS[NC] ET S[1] CatMS[NC] ET NS[1] CHIFFRE ET S[2])
C'est-à-dire	dire P[1] à ET P[2] être
C'est-à-dire	(cad|c.a.d.|c'est-à-dire)
C'est-à-dire	d P[1]a ET P[2]d
Au sens de	sens P[1] « au »
En d'autres termes	terme P[1] autre
Soit	soit P[1] CatMS[PCT] ET NP soit
Enfin	enfin P[1] CatMS[PCT]
Il s'agit de	agir P[1] se ET NS[3] CatMS[VB]
Entendre par	entendre S[1-6] par
Par ... entendre	par S[1-6] entendre
Vouloir dire	vouloir S[1] dire
Indiquer	indiquer NS[1] (pas|par|sur)
Comme	comme P définir
Comme	comme S définir
Comme	comme P[1] CatMS[NC] ET S[1] CatMS[DET]1,3
Dire à l'indicatif ou au participe passé	dire[CatMS :(VIND|VPARP)]
Ou	ou P[1] ,
Autrement dit	dire P[1] autrement ET P[2] MOT
Même chose que	même S[1] chose ET S[2] que
De même que	même P[1] de ET S[1] que
Equivaloir à	équivaloir

TAB. B.2 – Patrons lexico-syntaxiques de recherche d'énoncés définitoires 1 / 2

Employer pour	employer S[1] pour
Action de	action S[1] de ET S[2] CatMS[VINF]
Préciser le sens	préciser S sens
(appeler | nommer | référer | dénommer | désigner | dénoter | signifier | définir)	(appeler | nommer | référer | dénommer | désigner | dénoter | signifier | définir) NP[1] se ET NS[1-4] « à » ET NS[1] .
(nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception)	(nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception) NP[1] « au » ET P[1] CatMS[DET] ET S porter NS[1-6] sur
(nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception)	(nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception) P(prendre | recevoir | appliquer | employer | réserver | utiliser | donner | renvoyer | référer | définir)
(nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception)	(nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception) S (prendre | recevoir | appliquer | employer | réserver | utiliser | donner | renvoyer | référer | définir)
Etre (le | ce) (nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception)	être S (le | ce) ET S[2-3] (nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception)
Sous le (nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception)	sous S[2] (nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception)

TAB. B.3 – Patrons lexico-syntaxiques de recherche d'énoncés définitoires 2 / 2

Soit	soit P[1] CatMS[PCT] ET P soit
Ou	ou NP[1] ,

TAB. B.4 – Patrons lexico-syntaxiques liés à la recherche de taxèmes

Bibliographie

[Amar, 2000] Amar M. (2000). *Les fondements théoriques de l'indexation.* Parissko : Association des professionnels de l'information et de la documentation (ADBS).

[Assadi, 1998] Assadi H. (1998). *Construction d'ontologies à partir de textes techniques – application aux systèmes documentaires.* Thèse de doctorat, Université Paris 6.

[Assadi & Bourigault, 2000] 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.

[Auger, 1997] Auger A. (1997). *Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles.* Thèse de doctorat, Université de Neuchâtel.

[Aussenac-Gilles *et al.*, 2000] Aussenac-Gilles N., Biebow B. & Sylvie S. (2000). Modélisation du domaine par une méthode fondée sur l'analyse de corpus. In *9e journées francophones d'Ingénierie des Connaissances IC'2000*, Toulouse.

[Baader, 2003] Baader F. (2003). *The Description Logic Handbook – Theory, Implementation and Applications.* Cambridge University Press.

[Bachimont, 2000] 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. 305–324. Paris : Eyrolles.

[Bachimont, 2004] Bachimont B. (2004). *Arts et sciences du numérique : ingénierie des connaissances et critique de la raison computationnelle.* Habilitation à diriger des recherches, Université de Technologie de Compiègne.

[Baneyx *et al.*, 2005] 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 Conférence TIA-2005*, Rouen. *À paraître*.

[Berners-Lee *et al.*, 2001] Berners-Lee T., Hendler J. & Lassila O. (2001). The semantic web a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific american.* version en ligne seulement.

[Bertels, 2005] Bertels A. (2005). A la découverte de la polysémie des spécificités du français technique. In *12e édition de la conférence sur le Traitement Automatique des Langues (TALN 2005).*

[Bisson & Nédellec, 2001] Bisson G. & Nédellec C. (2001). Aide à la conception de méthodes de classification pour la construction d'ontologies : l'atelier Mo'K. In H. Brian, Ed., *Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 2001)* : Hermès.

[Bourigault, 1994] Bourigault D. (1994). *LEXTER, un logiciel d'EXtraction de TERminologie. Application à l'acquisition des connaissances à partir de textes.* PhD thesis, Ecole des Hautes Etudes en Sciences Sociales, Paris.

[Bourigault, 2002] Bourigault D. (2002). Analyse distributionnelle étendue. In *9e Conférence Annuelle sur le Traitement Automatique des Langues (TALN 2002)*, p. 75–84.

[Bourigault & Aussenac-Gilles, 2003] Bourigault D. & Aussenac-Gilles N. (2003). Construction d'ontologies à partir de textes. In *10e Conférence Annuelle sur le Traitement Automatique des Langues (TALN 2003)*, volume 2, p. 27–50. Batz-sur-Mer.

[Bourigault *et al.*, 2004] Bourigault D., Aussenac-Gilles N. & Charlet J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intellignece Artificielle (RIA)*, **18**(1/2004), 87–110.

[Bourigault & Fabre, 2000] Bourigault D. & Fabre C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de grammaire*, **25**, 131–151.

[Bourigault & Frérot, 2005] Bourigault D. & Frérot C. (2005). Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique. In *Actes des 12e journées sur le Traitement Automatique des Langues*, Dourdan.

[Bourigault & Jacquemin, 1999] Bourigault D. & Jacquemin C. (1999). Term extraction + term clustering : An integrated platform for computer-aided terminology. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL 99).*

[Bourigault & Jacquemin, 2000] Bourigault D. & Jacquemin C. (2000). Construction de ressources terminologiques. In J.-M. Pierrel, Ed., *Ingénierie des langues*, p. 215–233. Paris : Hermès.

[Bourigault & Lame, 2002] 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), 129–150. Adeline Nazarenko and Thierry Hamon (resp.).

[Bourigault & Slodzian, 1999] Bourigault D. & Slodzian M. (1999). Pour une terminologie textuelle. In *Terminomogie et Intelligence Artificielle, actes du colloque de Nantes.* 10-11 mai 1999, Nantes. Tutoriel à TIA 1999, paru dans Terminologies nouvelles n 19.

[Brondal, 1950] Brondal V. (1950). *Théorie des propositions. Introduction à une sémantique rationnelle.* Copenhague : E. Munksgaard.

[Cabré, 1999] Cabré M. T. (1999). *Terminology and Lexicography, Research and Practice*, chapter Terminology. Theory, methods and applications. John Benjamins : Amsterdam/Philadelphia.

[Caraballo, 1999] 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.

[Cartier, 1997] Cartier E. (1997). La définition dans les textes scientifiques et techniques : présentation d'un outil d'extraction automatique de relations définitoires. In *2e Rencontres Terminologie et Intelligence Artificielle (TIA'97)*, p. 127–140, Toulouse.

[Charlet, 2002] Charlet J. (2002). *L'ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales.* Habilitation à diriger des recherches, CHU Pitié-Salpêtrière. 10 décembre 2002.

[Choueka, 1988] Choueka Y. (1988). Looking for needles in a haystack. In *User-Oriented Context Based Text And Image Handling (RIAO'88)*, p. 609–623, Cambridge.

[Chukwu & Thoiron, 1989] Chukwu U. & Thoiron P. (1989). Reformulation et repérage des termes. *La Banque des Mots*, **Numéro spécial CTN - INaLF - CNRS**, 23–53.

[Cimiano *et al.*, 2004] 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, 1993] Condamines A. (1993). Un exemple d'utilisation de connaissances de sémantique lexicale : acquisition semi-automatique d'un vocabulaire de spécialité. *Cahiers de lexicologie*, **LXII**, 25–65.

[Condamines, 2003] Condamines A. (2003). *Sémantique et corpus spécialisé : constitution de bases de connaissances terminologiques.* Habilitation à diriger des recherches, Université de Toulouse Le Mirail.

[Crouch, 1988] Crouch C. J. (1988). Construction of a dynamic thesaurus and its use for associated information retrieval. In *eleventh international conference on Research and Development in Information Retrieval*, p. 309–320, Grenoble, France. 13–15 June.

[Curran & Moens, 2002] Curran J. R. & Moens M. (2002). Scaling context space. In *Proc 38th ACL*, p. 231–238, Philadelphie : ACL.

[Dailland *et al.*, 2004] Dailland F., Leuthereau A. & Vallée H. (2004). *AIDE MEMOIRE D'INDEXATION - MeSH (Medical Subject Heading) et FMeSH (version française) pour le catalogage.* Bibliothèque Universitaire de Médecine de Paris XI (Paris-Sud), INSERM DISC (Département de l'Information Scientifique et de la Communication), Le Kremlin Bicêtre, Hôpital Paul Brousse - Villejuif.

[Daille, 1994] Daille B. (1994). *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques.* Thèse de Doctorat, Université de Paris 7.

[David & Plante, 1990] David S. & Plante P. (1990). De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes. *Intelligence Artificielle et Sciences Cognitives au Québec*, **3**(3), 140–154.

[Desclés, 1997] Desclés J.-P. (1997). Systèmes d'exploration contextuelle. In C. Guimier, Ed., *Co-texte et calcul du sens*, p. 215–232. Caen : Presses Universitaires de Caen.

[Dias, 2002] Dias G. (2002). *Extraction automatique d'associations lexicales à partir de corpora.* Thèse de Doctorat d'Université, Université d'Orléans.

[Enguehard & Pantera, 1995] Enguehard C. & Pantera L. (1995). Automatic natural acquisition of a terminology. *Journal of Quantitative Linguistics*, **2**(1), 27–32.

[Flowerdew, 1992] Flowerdew J. (1992). Definitions in science lectures. *Linguistics*, **13**(2), 202–221.

[Fuchs, 1994] Fuchs C. (1994). *Paraphrase et énonciation.* Paris : Ophrys.

[Geeraerts *et al.*, 1994] Geeraerts D., Grondelaers S. & Bakema P. (1994). *The structure of lexical variation : meaning, naming and context.* Berlin–New York : Mouton de Gruyter.

[Genest, 2000] Genest D. (2000). *Extension du modèle des graphes conceptules pour la recherche d'information*. Thèse de doctorat, Université Montpellier II, Montpellier. Laboratoire d'Informatique, de Robotique et de Mico-électronique de Montpellier (LIRMM) spécialité Informatique 18 décembre 2000.

[Genest & Salvat, 1998] Genest D. & Salvat É. (1998). A platform allowing typed nested graphs : How cogito became cogitant. In *6th International Conference on Conceptual Structures (ICCS 98) , session Conceptual Graph Tools*, volume 1453, p. 154–161 : Springer. Lecture Notes in Artificial Intelligence.

[Gomez-Perez *et al.*, 2004] Gomez-Perez A., Fernandez-Lopez M. & Corcho O. (2004). *Ontological Engineering with examples from the areas of Knowledge management, e-Commerce and the Semantic Web*. Springer.

[Gouadec, 1990] Gouadec D. (1990). *Terminologie : Constitution des données*.

[Grabar & Berland, 2001] Grabar N. & Berland S. (2001). Construire un corpus Web pour l'acquisition terminologique. In *Terminologie et intelligence artificielle*, p. 44–54, Nancy.

[Grabar & Hamon, 2004] Grabar N. & Hamon T. (2004). Les relations dans les terminologies structurées : de la théorie à la pratique. *Revue d'Intelligence Artificielle (RIA)*, **18**(1), 57–85.

[Grabar *et al.*, 2004] Grabar N., Malaisé V., Marcus A. & Krul A. (2004). Repérage de relations terminologiques transversales en corpus. In *11e journées sur le Traitement Automatique des Langues*, Fes, Maroc.

[Greimas, 1966] Greimas A. J. (1966). *Sémantique structurale*. Paris : Larousse.

[Guarino, 1997] Guarino N. (1997). Some organizing principles for a unified top-level ontology. In *Proceedings of the AIII Spring Symposium on Ontological Engineering*.

[Habert, 1998] Habert B. (1998). *Des mots complexes possibles aux mots complexes existants : l'apport des corpus*. Habilitation à diriger des recherches en linguistique, Universite Lille III - Charles de Gaulle.

[Habert *et al.*, 1997] Habert B., Nazarenko A. & Salem A. (1997). *Les linguistiques de corpus*. Paris : Armand Colin–Masson.

[Hall & Dowling, 1980] Hall P. A. & Dowling G. R. (1980). Approximate string matching. *Computing survey*, **12(4)**, 381–402.

[Hamon, 2000] Hamon T. (2000). *Variation sémantique en corpus specialisé : Acquisition de relations de synonymie à partir de ressources lexicales*. PhD thesis, Université Paris 13 - Villetaneuse. Jury Henry Boccon-Gibod (Examinateur) Béatrice Daille (Examinatrice) Christophe Fouqueré (Directeur) Benoît Habert (Rapporteur) Adeline Nazarenko (Co-directrice) Jean Véronis (Rapporteur) Université : Université Paris 13 - Villetaneuse Discipline : Informatique Date de soutenance : 19/12/2000 Lieu de soutenance : Salle L322 - Institut Galilée - Université Paris Nord.

[Harris, 1968] Harris Z. (1968). *Mathematical Structures of Language*. New-York : John Wiley and Sons.

[Hearst, 1992] Hearst M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In A. Zampolli, Ed., *Computational Linguistics (CoLing'1992)*, p. 539–545, Nantes.

[Hindle & Rooth, 1990] Hindle D. & Rooth M. (1990). Structural ambiguity and lexical relations. In *DARPA Speech and Natural Language Workshop*, Hidden Valley.

[Houben, 2004] Houben F. (2004). Mot vide, mot plein ? comment trancher localement. In *8e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2004)*, p.-, Fès, Maroc.

[Ibekwe-SanJuan, 2005] Ibekwe-SanJuan F. (2005). Inclusion lexicale et proximité sémantique entre termes. In *6e rencontres Terminologie et Intelligence Artificielle (TIA 2005)*, p. 45–58, Rouen. 4 et 5 avril 2005.

[Isaac, 2001] Isaac A. (2001). Vers la mise en œuvre informatique d'une méthode de conception d'ontologies. Rapport de dea, Laboratoire LaLICC, Université Paris IV Sorbonne.

[Isaac *et al.*, 2004] Isaac A., Courounet P., Genest D., Malaisé V., Nanard J. & Nanard M. (2004). Un système d'annotation muli-forme et communautaire de documents audiovisuels : Opales. In *Journée du Document Numérique (SDN 2004), atelier "Modèles documentaires de l'Audiovisuel"*, La Rochelle. 22 juin.

[Jackobson, 1970] Jackobson R. (1970). *Essais de linguistique générale.* collection "Points". Paris.

[Jacquemin, 1994] Jacquemin C. (1994). Fastr : A unification-based front-end to automatic indexing. In *Intelligent Multimedia Information Retrival Systems and Management (RIAO'94)*, p. 34–37, New-York : CID, Paris.

[Jacquemin & Bourigault, 2003] Jacquemin C. & Bourigault D. (2003). *The Oxford Handbook of Computational Linguistics*, chapter Term Extraction and Automatic Indexing, p. 599–615. Oxford Univesity Press.

[Kiryakov *et al.*, 2001] Kiryakov A., Simov K. & Dimitrov M. (2001). Ontomap : Portal for upper-level ontologies. In *FOIS 2001 Conference*, p. 47–58, Ogunquit, Maine, USA.

[Laban, 1928] Laban R. (1928). *Schrifttanz.* Wien.

[Le Moigno *et al.*, 2002] 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 *13e journées francophones d'ingénierie des Connaissances (IC 2002)*, Rouen.

[Lebart & Salem, 1988] Lebart L. & Salem A. (1988). *Analyse statistique des données textuelles.* Paris : Dunod, Bordas.

[Lespinasse, 2002] Lespinasse K. (2002). *Acquisition sémantique en langue générale : l'heure de vérité de la paradocumentation textuelle pour l'indexation des documents audiovisuels sur la politique.* Thèse de doctorat en sciences du langage, option linguistiques de corpus, Université de Paris 3, Sorbonne nouvelle. 3 mai 2002.

[Lespinasse *et al.*, 2000] Lespinasse K., Habert B. & Bachimont B. (2000). Le péritexte, un sésame pour les données audiovisuelles ? l'analyse exploratoire d'un corpus hétérogène de notices documentaires interprétant des documents audiovisuels. In *Journées d'Analyse statistique de Données Textuelles (JADT 2000).*

[L'Homme, 2001] L'Homme M.-C. (2001). Nouvelles technologies et recherche terminologique. techniques d'extraction des données terminologiques et leur impact sur le travail du terminographe. In *L'impact des nouvelles technologies sur la gestion terminologique*, Toronto : Université de York. 18 août.

[Lipka, 1988] Lipka L. (1988). *Understanding the lexicon : meaning, sense and word knowledge in lexical semantics*, chapter A rose is a rose : on simple and dual categorization in natural language, p. 355–366. Niemeyer : Tubingen.

[Loffler-Laurian, 1983] Loffler-Laurian A. (1983). Typologie des discours scientifiques : deux approches. *Études de linguistique appliquée*, **51**, 8–20.

[Malaisé *et al.*, 2005] Malaisé V., Delbecque T. & Zweigenbaum P. (2005). Recherche en corpus de réponses à des questions définitoires. In *12e édition de la conférence sur le Traitement Automatique des Langues (TALN 2005)*.

[Malaisé *et al.*, 2004a] Malaisé V., Zweigenbaum P. & Bachimont B. (2004a). Detecting semantic relations between terms in definitions. In S. Ananadiou & P. Zweigenbaum, Eds., *COLING 2004 CompuTerm 2004 : 3rd International Workshop on Computational Terminology*, p. 55–62, Geneva, Switzerland : COLING.

[Malaisé *et al.*, 2004b] Malaisé V., Zweigenbaum P. & Bachimont B. (2004b). 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.

[Martin, 1992] Martin E. (1992). *Éléments pour un système de reconnaissance des contextes thématiques dans l'exploration d'un corpus textuel*. Thèse de doctorat, Université de Paris-Sorbonne.

[Martin, 1983] Martin R. (1983). *Pour une logique du sens*. Paris : Presses Universitaires de France.

[Martin, 1990] Martin R. (1990). La définition "naturelle". In J. Chaurand & F. Mazière, Eds., *La définition*, p. 86–95. Paris : Larousse.

[Meyer, 2001] Meyer I. (2001). Extracting knowledge-rich contexts for terminography. In D. Bourigault, M.-C. L'Homme & C. Jacquemin, Eds., *Recent advances in Computational Terminology*, p. 279–302. Amsterdam/Philadelphia, PA : John Benjamins Publishing Company.

[Morin, 1998] Morin E. (1998). Prométhée : un outil d'aide à l'acquisition de relation sémantiques entre termes. In *5e conférence annuelle sur le Traitement Automatique des Langues Naturelles*, p. 172–181, Paris.

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

[Mugnier & Chein, 1998] Mugnier M.-L. & Chein M. (1998). Conceptual structures : Theory, tools and applications. In *Lecture Notes in Artificial Intelligence*, volume 699. Berlin : Springer-Verlag.

[Muresan & Klavans, 2002] Muresan S. & Klavans J. L. (2002). A method for automatically building and evaluating dictionary resources. In *the Language Resources and Evaluation Conference (LREC 2002)*, p. 231–234, Las Palmas, Spain.

[Namer & Zweigenbaum, 2004] 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 th World Congress on Medical Informatics*, p. 535–539, San Francisco, Ca.

[Nazarenko & Hamon, 2002] Nazarenko A. & Hamon T. (2002). Structuration de terminologie : quels outils pour quelles pratiques ? *Traitement automatique des langues*, **43**(1), 7–18.

[Newell, 1982] Newell A. (1982). The knowledge level. *Artificial Intelligence*, **18 (1)**, 87–127.

[Pearson, 1998] Pearson J. (1998). *Terms in context*. Amsterdam/Philadelphia : John Benjamins Publishing Company.

[Pearson, 1999] Pearson J. (1999). Comment accéder aux éléments définitoires dans les textes spécialisés ? In *Terminologie et intelligence artificielle (TIA'1999)*, p. 21–38, Nantes, France.

[Picoche, 1977] Picoche J. (1977). *Précis de lexicologie française*. Paris.

[Porphyre, 1998] Porphyre (1998). *Isagoge. Sic et Non*. Vrin.

[Rastier *et al.*, 1994] Rastier F., M. C. & Abeillé A. (1994). *Sémantique pour l'analyse*. Paris : Masson.

[Rebeyrolle, 2000] Rebeyrolle J. (2000). *Forme et fonction de la définition en discours*. Thèse de doctorat, Université de Toulouse II - Le Mirail.

[Rebeyrolle & Tanguy, 2000] Rebeyrolle J. & Tanguy L. (2000). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire*, **25**, 153–174.

[Rey, 1990] Rey A. (1990). Polysémie du terme définition. In *La définition*. Librairie Larousse.

[Rodrigues *et al.*, 1999] Rodrigues J. M., Trombert-Paviota B., Rectorb A., Baude R., Clavela L., Abriala V., Idira H. & Verya J. M. (1999). Galen, il existe quelque chose derrière les mots : leur signification et au-delà le savoir médical. In *Innovation Stratégique en Information de Santé (ISIS)*.

[Ryle, 1949] Ryle G. (1949). *The Concept of Mind*. Harmondsworth : Penguin.

[Sager, 2001] Sager J. C. (2001). *Essays on Definition*. Amsterdam : John Benjamins.

[Sanderson & Croft, 1999] Sanderson M. & Croft B. (1999). Deriving concept hierarchies from text. In *22nd annual international ACM SIGIR conference on Research and Development in Information Retrieval*, p. 206–213, Berkeley, CA USA. 1519 August.

[Schreiber *et al.*, 2001] Schreiber G., Dubbeldam B., Wielemaker J. & Wielinga B. (2001). Ontology-based photo annotation. *IEEE Intelligent Systems*, **May/June**.

[Séguéla & Aussenac-Gilles, 1999] Séguéla P. & Aussenac-Gilles N. (1999). Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In *9e journées francophones d'Ingénierie des Connaissances (IC 1999)*, p. 79–88, Palaiseau.

[Sowa, 1984] Sowa J. F. (1984). *Conceptual Structures : Information Processing in Mind and Machine*. London : Addison-Wesley.

[Sowa, 1994] Sowa J. F. (1994). Knowledge representation : Logical, philosophical and computational foundations. In *Preliminary Edition, 2th International Conference on Conceptual Structures (ICCS'94)*.

[Spark Jones & Willet, 1997] Spark Jones K. & Willet P. (1997). Thesaurus. In *Readings in Information Retrieval*. Elsevier Science.

[Srinivasan, 1992] Srinivasan P. (1992). *Information Retrieval : Data Structures and Algorithms*, chapter 9. Thesaurus construction. Prentice hall, New Jersey.

[Szulman *et al.*, 2002] Szulman S., Biébow B. & Aussenac-Gilles N. (2002). Structuration de terminologies à l'aide d'outils de TAL avec TERMINAE. *Traitement automatique des langues*, **43**(1), 103–128. Adeline Nazarenko and Thierry Hamon (resp.).

[Séguéla, 2001] Séguéla P. (2001). *Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques*. Thèse de doctorat, Université Toulouse 3.

[Tam & Leung, 2001] Tam A. M. & Leung C. H. C. (2001). Structured natural-language description for semantic content retrieval. *Journal of the American Society for Information Science*.

[Tesnière, 1959] Tesnière L. (1959). *Éléments de syntaxe structurale*. Paris : Klincksieck.

[Trimble, 1985] Trimble L. (1985). *English for Science and Technology : A Discourse Approach*. Cambridge, MA.

[Troncy, 2004a] Troncy R. (2004a). *Formalisation des connaissances documentaires et des connaissances conceptuelles à l'aide d'ontologies : application à la description de documents audiovisuels traitant du cyclisme*. Thèse de doctorat, Joseph Fourier - Grenoble 1.

[Troncy, 2004b] Troncy R. (2004b). Le raisonnement dans les descriptions documentaires : l'apport de la représentation des connaissances. In *14e journées francophones d'Ingénierie des Connaissances (IC 2004)*.

[Troncy & Isaac, 2002] Troncy R. & Isaac A. (2002). DOE : une mise en œuvre d'une méthode de structuration différentielle pour les ontologies. In *13e journées francophones d'Ingénierie des Connaissances (IC 2002)*, p. 63–74, Rouen.

[Tudhope *et al.*, 2002] Tudhope D., Binding C., Blocks D. & Cunliffe D. (2002). Representation and retrieval in faceted systems. In M. Lopez-Huertas, Ed., *Proceedings of the International Conference on Information Systems and Knowledge Organisation*.

[Turner *et al.*, 2000] Turner J. M., Hudon M. & Devin Y. (2000). Text as a tool for organizing moving image collections. In *Actes du 28e congrès de l'Association canadienne des Sciences de l'information*, Edmonton. 2000 05 28 - http ://www.slis.ualberta.ca/cais2000/turner.htm.

[Velardi *et al.*, 2001] Velardi P., Missikof M. & Fabriani P. (2001). Using text processing techniques to automatically enrich a domain ontology. In *Proceeding of ACM-FOIS*.

[Vergne, 2003] Vergne J. (2003). Un outil d'extraction terminologique endogène et multilingue. In *10e Conférence Annuelle sur le Traitement Automatique des Langues (TALN 2003)*, volume 2, p. 139–148, Batz-sur-Mer. 11-14 juin 2003.

[Veronis & Ide, 1990] Veronis J. & Ide N. (1990). Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *13th conference on Computational Linguistics (CoLing'1991)*, p. 389 – 394, Helsinki, Finland.

[Voltz *et al.*, 2003] Voltz R., Oberle D., Staab S. & Motik B. (2003). Kaon server - a semantic web management system. In *Alternate Track Proceedings of the Twelfth International World Wide Web Conference*, p. 139–148, Budapest, Hungary : ACM. 20-24 May 2003.

[Wüster, 1981] Wüster E. (1981). L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses. In *Textes choisis de Terminologie*, volume Vol 1 Fondements théoriques de la terminologie, p. 55–114. Québec : GISTERM, Université de Laval.

[Zweigenbaum, 1999] Zweigenbaum P. (1999). Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances. *Innovation Stratégique en Information de Santé*, **2–3**, 27–47.

[Zweigenbaum *et al.*, 1994] Zweigenbaum P., Bachimont B., Bouaud J., Charlet J., Ben Saïd M., Boisvieux J.-F., Beux P. L., Delamarre D., Burgun A., Willems J. L., Spyns P., Guillotin T., Fargues J., Landau M.-C., Moens M., Whittemore G., Grover C. & Mikheev A. (1994). *MENELAS System and User Documentation.* Deliverable report AIM-Menelas 14, DIAM-SIM/INSERM U.194.

[Zweigenbaum *et al.*, 2002] Zweigenbaum P., Darmoni S. J., Grabar N., Douyère M. & Benichou J. (2002). An assessment of the visibility of MeSH-indexed medical web catalogs through search engines. *Journal of the American Medical Informatics Association*, **8**(suppl), 954–958.

[Zweigenbaum & Grabar, 2000] Zweigenbaum P. & Grabar N. (2000). Liens morphologiques et structuration de terminologie. In *IC 2000 : Ingénierie des connaissances*, p. 325–334.

Résumé

Des ressources telles que les terminologies ou les ontologies sont utilisées dans différentes applications, notamment dans la description documentaire et la recherche d'information. Différentes méthodologies ont été proposées pour construire ce type de ressources, que ce soit à partir d'entrevues avec des experts du domaine ou à partir de corpus textuels.

Nous nous intéressons dans ce mémoire à l'utilisation de méthodologies existantes dans le domaine du Traitement Automatique des Langues, destinées à la construction d'ontologies à partir de corpus textuels, pour la construction d'un type de ressource particulier : des ontologies différentielles. Ces ontologies sont structurées selon un système d'identité et de différence sémantique entre leurs constituants : les termes du domaine et des catégories dites *de haut niveau*.

Nous présentons différentes expérimentations qui ont été menées pour éliciter, structurer, définir et interdéfinir les éléments terminologiques pertinents à la réalisation d'une tâche particulière. Notre premier contexte applicatif a été le projet OPALES, et nous devions fournir à des anthropologue le vocabulaire conceptuel destiné à annoter des documents audiovisuels traitant de la petite enfance. Nous nous sommes servie du corpus constitué à cette occasion pour tester les méthodologies et outils linguistiques proposés pour l'aide à la construction d'ontologie, et avons défini notre propre chaîne de traitement. Celle-ci, appellée SODA, est basée sur l'extraction et l'exploitation d'énoncés définitoires en corpus pour repérer des éléments terminologiques, les structurer et donner des éléments de communauté sémantique permettant de les comparer.

Mots-clés: Ontologie différentielle, Traitement Automatique des Langues, Enoncé définitoire.

Abstract

Resources like terminologies or ontologies are used in a number of applications, including documentary description and information retrieval. Different methodologies have been proposed to build such resources, on the basis of experts' interviews or of textual corpora.

This thesis focuses on the use of existing Natural Language Processing methodologies, meant to help the building of ontologies from textual corpora, to build a particular type of resource : differential ontologies. These ontologies are structured according to a system of semantic identities and differences between their constituents: terms of the domain and categorisation items called "top level categories".

We present different experiments that we have done to elicit, structure, define and "interdefine" the terminological items relevant for a given task. Our first use case was the OPALES project, in which we had to provide a group of anthropologists with the conceptual vocabulary that they needed to annotate audiovisual documents about childhood. We have used the textual corpus that we have built in this project to test linguistic tools and methodologies for building ontologies from textual data, and we have defined our own

programs. The suite of resulting programs is called SODA, and they focus on the extraction and use of defining contexts in corpora to spot terminological items, to structure them and to provide semantic similarity information that enables to compare them.

Keywords: Differential Ontologies, Natural Language Processing, Defining Context.