Automatisation des tâches documentaires dans un catalogue de santé en ligne.

Aurelie Neveol

Institut National des Sciences Appliquées de Rouen

Thèse

Pour obtenir le grade de

Docteur en informatique de l'INSA de Rouen

Automatisation des tâches documentaires dans un catalogue de santé en ligne

Aurélie Névéol

Soutenue publiquement le 29 Novembre 2005

Jury :

Didier Bourigault (Examinateur)
Stéfan J. Darmoni (Directeur)
Adeline Nazarenko (Rapporteure)
Alexandrina Rogozan (Co-encadrante)
Max Silberztein (Rapporteur)
Pierre Zweigenbaum (Président)

Mon paure houme, o l'ét pâ dan l'Subiet quyi al'ant n'en parler do cancer d'la prostate ! Sonjhe d'allé vouère su' CISMeF, jh'cré beun qu'thiél a'faire o srat espiqué[1].

[1]Mon pauvre homme, ce n'est pas dans le Subiet [*ndt : journal patoisant Saintongeais*] que tu trouveras des informations sur le cancer de la prostate. Consulte plutôt CISMeF, tu auras toutes les explications nécessaires.

Remerciements

Je tiens à remercier chaleureusement :

* mon directeur de thèse, Stéfan Darmoni, pour m'avoir accueillie dans l'équipe CISMeF et ouvert des horizons de recherche variés, pour de nombreuses discussions et débats houleux, pour son bouillonnement d'idées perpétuel et communicatif, et pour savoir créer une ambiance de travail conviviale et motivante.

* ma co-encadrante Alexandrina Rogozan, grâce à qui j'ai beaucoup appris, autant sur le plan scientifique que personnel.

* Pierre Zweigenbaum pour avoir accepté de présider mon jury, pour son expertise scientifique et pratique, ainsi que ses nombreux conseils.

Je remercie également les membres de mon jury pour avoir accepté cette tâche, et pour leurs remarques constructives.

Mes remerciements vont également à Fiametta Namer pour ses encouragements et sa bonne humeur.

Merci à tous les membres de l'équipe CISMeF pour leur accueil chaleureux, leurs explications patientes, leur enthousiasme et leur dynamisme permanents qui m'ont permis d'aller de l'avant.

Je remercie les membres des laboratoires PSI et DYALANG pour de nombreux échanges enrichissants, et en particulier Myriam Bouveret, Nathalie Chaignault, Jean-Philippe Kotowitz et Alain Rakotomamonjy pour leurs précieux conseils à la relecture de mon manuscrit ou lors de la préparation de ma soutenance.

Ce travail doit également beaucoup aux nombreux échanges avec mes collègues (ex-) doctorants.

Une pensée à mes collègues du CHU pour avoir prété une oreille sympathisante à mes soucis quotidiens, et souvent une main secourable - voire une collation réparatrice ! (Ricola, muffin...)

Un clin d'oeil à tous mes copains thésards de Rouen et d'ailleurs... et aux autres aussi, parce que bon, il n'y a pas que les thèses dans la vie !

Merci à ma famille, et en particulier mes parents, pour leur soutien tout au long de ma scolarité, leurs enseignements, leurs encouragements, leur confiance - En résumé : bravo, c'est réussi ! Maintenant que j'ai tous mes brevets, j'espère assister prochaînement à une autre soutenance de thèse.

Une bise à mon frérot, et j'espère que les chemins de la linguistique et de la musique se recroiseront souvent !

Enfin, les mots me manquent pour exprimer toute ma reconnaissance à Sébastien pour son aide innestimable dans ce travail, son soutien de tous les instants et bien plus encore...

Table des matières

Introduction **11**

1 Contexte de travail **15**
1.1 Notion de document et de ressource 15
1.2 Catalogue et Index des Sites Médicaux Francophones : CISMeF 18
1.2.1 Présentation du projet 18
1.2.2 Structure du catalogue 18
1.2.3 Notice CISMeF 19
1.2.4 Accessibilité par l'Homme et par la Machine 22
1.2.5 Présentation de l'équipe et des projets 22
1.2.6 Le corpus CISMeF 24
1.3 Problématique de la thèse 26

2 Terminologie Médicale **29**
2.1 Représentation des connaissances 29
2.1.1 Définitions 29
2.1.2 Eléments de la représentation des connaissances 30
2.1.3 Notions de domaine et de langue de spécialité 31
2.1.4 Relations sémantiques 31
2.1.5 Elaboration de systèmes de représentation des connaissances 32
2.2 Terminologie et représentation des connaissances en Médecine : quelques exemples (CIM-10, SNOMED, MeSH et CISMeF) 34
2.2.1 Classification statistique internationale des maladies et des problèmes de santé connexes : la CIM 34
2.2.2 La Nomenclature systématique des médecines humaine et vétérinaire : la SNOMED 36
2.2.3 Medical Subject Headings : le thésaurus MeSH 37
2.3 La terminologie CISMeF : une terminologie fondée sur le MeSH 41
2.3.1 Utilisation du MeSH par CISMeF 41
2.3.2 Les types de ressource 42
2.3.3 Les métatermes 43
2.4 Traitement automatique de la langue médicale 44
2.4.1 Enjeux du traitement automatique de la langue médicale 44
2.4.2 Ressources linguistiques dans le domaine médical 45
2.4.3 Travaux sur le traitement de la langue médicale 46
2.5 Notre contribution 47

2.5.1 Construction d'un dictionnaire électronique MeSH 47
2.5.2 Traduction automatique de termes médicaux à l'aide de corpus parallèles 55
2.6 Conclusion sur la terminologie médicale 70

3 Veille documentaire **71**
3.1 Définitions 71
3.2 L'activité de veille documentaire 71
3.3 Les techniques et outils pour la Veille Documentaire 73
3.3.1 Collecte de l'information 73
3.3.2 Analyse de l'information 74
3.3.3 Diffusion de l'information 74
3.4 Veille documentaire CISMeF 74
3.4.1 Recherche de ressources 75
3.4.2 Critères de sélection des documents indexés dans CISMeF 76
3.4.3 Formalisation de la procédure de sélection 77
3.4.4 Automatisation de la Veille 78
3.5 Conclusion 81

4 Classification de documents **83**
4.1 Définition et objectifs 83
4.1.1 Définition 83
4.1.2 Classification dans le cadre de CISMeF : Objectifs 84
4.2 Méthodes de classification 84
4.2.1 Vue d'ensemble 84
4.2.2 Classification par Apprentissage 85
4.2.3 Classification avant l'indexation 87
4.2.4 Classification après l'indexation 89
4.3 Conclusion 97

5 Indexation **99**
5.1 Définitions 99
5.2 Les « langages » d'indexation : quelles unités descriptives utiliser pour la représentation d'un document ? 100
5.2.1 Indexation libre et indexation contrôlée 100
5.2.2 Unités descriptives 101
5.2.3 Quelles unités descriptives choisir ? 102
5.3 L'activité d'indexation : un problème complexe 106
5.3.1 L'indexation : une traduction conceptuelle 106
5.3.2 L'indexation : une catégorisation 106
5.3.3 L'indexation : résolution d'un problème mal défini 107
5.4 Critères d'évaluation de l'indexation 107
5.4.1 Objectivité de l'indexation 107
5.4.2 Consistance de l'indexation 108
5.4.3 Qualité de l'indexation 112
5.5 Rôle des titres dans l'indexation 114
5.5.1 Sémantique des Titres 114
5.5.2 Extraction Automatique de titres 115

5.6 Indexation Manuelle, Automatique, Assistée . . . 118
5.6.1 Définitions . . . 118
5.6.2 L'indexation Manuelle . . . 119
5.6.3 L'indexation Automatique . . . 120
5.7 Comparaison des systèmes d'indexation MeSH . . . 123
5.8 MAIF : MeSH Automatic Indexing for French . . . 126
5.8.1 MAIF : approche TAL (Traitement Automatique de la Langue) . . . 126
5.8.2 MAIF : approche k-PPV (dite des « k Plus Proches Voisins ») . . . 132
5.8.3 Fusion des approches . . . 133
5.8.4 Indexation d'une ressource . . . 135
5.9 Evaluation de Systèmes d'Indexation Automatique MeSH . . . 141
5.9.1 Evaluation de MAIF . . . 142
5.9.2 Evaluation des systèmes d'indexation MeSH francophones . . . 150
5.9.3 Evaluation translangue à l'aide d'un corpus parallèle EN/FR . . . 155
5.10 Applications du Système d'Indexation Automatique MAIF . . . 158
5.10.1 Indexation bi-modale Texte-Image . . . 158
5.10.2 Intégration de terminologies (CCAM) . . . 159
5.10.3 Codage des dossiers patients (Indexation CIM-10) . . . 160
5.11 Conclusion sur l'Indexation . . . 162

6 Conclusion et Perspectives **163**
6.1 Réalisations . . . 163
6.2 Perspectives . . . 165

Bibliographie **167**

A Ecran d'accueil de l'outil bibliométrique **183**

B Résultats de l'indexation des Sommaires sur le corpus « diabète » **185**

Introduction

Problématique de la Recherche d'Information

Le but de la Recherche d'Information est de donner accès aux connaissances disponibles dans les fonds documentaires. A l'heure actuelle, l'écrit est un moyen privilégié pour exprimer et véhiculer les connaissances dans tous les domaines. Ainsi, l'objectif de la recherche d'information peut être redéfini comme étant de permettre l'exploitation des informations contenues dans des documents textuels. Depuis quelques années, le nombre de documents électroniques disponibles sur l'immense fond documentaire qu'est l'Internet augmente de manière exponentielle. L'exploitation efficace de ces documents constitue donc un véritable enjeu pour la recherche d'information.

Dans les domaines de la santé et de la bio-médecine, de nombreux travaux ont été entrepris afin de guider les utilisateurs dans leur recherche d'information. Ainsi, la base de données bibliographiques MEDLINE® recense 10,6 millions d'articles scientifiques en langue anglaise indexés à l'aide du MeSH® (Medical Subject Headings), le thésaurus de référence du domaine bio-médical développé et maintenu par la NLM (National Library of Medicine) américaine. En Europe, plusieurs projets (par exemple, HON[2], OMNI[3], DDRT[4] ou CISMeF[5]) ont également vu le jour en se donnant pour objectif de faciliter l'accès de divers publics - les professionnels de santé, les étudiants, les patients, ou encore, le grand public - à une information de santé de qualité dans plusieurs langues européennes, dont le français.

Concrètement, cet engagement se traduit par un travail titanesque de traitement des documents disponibles dans le domaine de la santé, allant du recensement au classement, en passant par la description et l'indexation, sans oublier une analyse de la qualité et de la pertinence des informations. Le coût de ces activités en temps est bien entendu proportionnel au nombre de documents à traiter et les documentalistes à qui ces tâches incombent ne peuvent faire face à la masse de travail que cela implique.

Afin de fournir un service de qualité à l'utilisateur d'information, un catalogue de ressources en ligne doit être suffisamment complet, et mis à jour régulièrement. Pour être viable, le développement d'un catalogue doit impérativement envisager une automatisation au moins partielle des tâches documentaires. Cela implique donc de développer des outils informatiques capables de « comprendre » les documents (Nazarenko, 2004).

En tant qu'utilisateur de tels outils, le documentaliste moderne est alors en mesure de s'acquitter de sa tâche auprès des demandeurs d'information, tout en concentrant ses efforts sur les tâches plus complexes, telles que la sélection des ressources, ou la superindexation.

[2]Health On the Net - `http://www.hon.ch/`
[3]Organising Medical Networked Information - `http://omni.ac.uk/`
[4]Diseases, Disorders and Related Topics - `http://www.mic.ki.se/Diseases/`
[5]Catalogue et Index des Sites Médicaux Francophones - `http://www.cismef.org/`

Dans cette thèse, notre réflexion a porté sur la formalisation des tâches documentaires mises en jeu par la construction et la mise à jour d'un catalogue de santé. Nous avons été amenée à la fois à considérer le problème dans son ensemble, afin de comprendre les relations entretenues par les différentes tâches documentaires et à considérer chaque tâche séparément afin de proposer dans chaque cas une solution automatique susceptible d'alléger le travail des documentalistes, et d'apporter des éléments de réponse en ce qui concerne la « compréhension » des ressources de santé par un système informatique.

Ce travail explore différentes approches pour analyser le contenu des documents, et pour les exploiter. Il s'agit tout d'abord de méthodes d'apprentissage et, principalement, de Traitement de la Langue Naturelle. Dans ce cadre, nous avons également été amenée à utiliser et à développer des outils indispensables pour le traitement de corpus spécialisés, à savoir des terminologies et des dictionnaires du domaine.

La méthode de travail adoptée tout au long de la thèse est une démarche expérimentale habituelle dans le domaine de la santé. Il s'agit d'une démarche séquentielle ou ascendante (« bottom-up ») qui consiste à partir des problématiques concrètes rencontrées pour aller vers la résolution des problèmes scientifiques sous-jacents. Ainsi, pour chaque tâche documentaire, nous avons effectué une analyse du problème, proposé une modélisation sous forme d'automatisation de la tâche, expérimenté la solution ainsi établie. A partir d'une analyse des résultats obtenus, nous avons, lorsque c'était nécessaire, re-itéré sur l'ensemble du processus afin d'ajuster la modélisation proposée initialement et de valider les changements effectués.

Avant d'exposer plus en détail le travail réalisé, nous proposons ci-dessous une définition des points clés de notre problématique.

Tout d'abord, nos travaux on été appliqués a un catalogue de santé en ligne particulier : le Catalogue et Index des Sites Médicaux Francophones (CISMeF). CISMeF a pour vocation de décrire et d'indexer des *ressources* (en premier lieu institutionnelles) de santé disponibles gratuitement sur l'Internet à destination des trois publics suivants : professionnels de santé, étudiants en médecine, patients.

Par ailleurs, le vocable de *ressource* désigne l'objet de notre étude. Il peut prendre de multiples formes (site web ou document numérique) que nous préciserons, tout en soulignant les spécificités dues au format électronique.

Enfin, nous entendons par « tâche documentaire » le traitement des ressources pour la constitution, la maintenance, et la mise à disposition du catalogue. Il s'agit dans notre cas de la *veille stratégique*, de l'*indexation* et de la *catégorisation* de ressources de santé. La *recherche d'information* fait l'objet d'autres travaux au sein de l'équipe CISMeF (réalisés par L. Soualmia (L. Soualmia, Dahamna, & Darmoni, 2005) et A. Loisel (Loisel, Kotovic, Chaignaud, & Darmoni, 2005)).

Organisation du Mémoire

Le premier chapitre présente le contexte de nos travaux, qui s'inscrivent dans le cadre général de la recherche d'information en santé, et plus particulièrement dans le cadre du développement du Catalogue et Index des Sites Médicaux Francophones. Nous revenons sur les notions de *document numérique* et de *ressource*, afin de définir les objets recensés par le catalogue. Nous faisons le bilan des réalisations de CISMeF entre 1995 et 2002, et introduisons

notre problématique, qui s'inscrit dans la continuité du développement de CISMeF.

Le deuxième chapitre aborde les questions de terminologie médicale et de représentation des connaissances en médecine, qui constituent le fondement de l'organisation des informations de santé contenues dans CISMeF. En particulier, la terminologie CISMeF (qui englobe le thésaurus MeSH) est l'outil central permettant d'organiser la sélection et la description des ressources.

Les chapitres suivants présentent les tâches documentaires auxquelles nous nous sommes intéressée et pour lesquelles nous avons apporté notre contribution afin d'automatiser ces processus.

Le troisième chapitre présente la problématique de veille documentaire dans son ensemble, puis dans le cadre particulier de la veille manuelle effectuée dans CISMeF. Nous exposons ensuite notre participation au développement du logiciel CVA, aujourd'hui utilisé dans CISMeF pour la veille documentaire.

Le quatrième chapitre évoque la catégorisation thématique de ressources textuelles. Après une revue des méthodes utilisées dans la littérature, nous présentons les deux approches que nous avons abordées pour traiter la catégorisation de ressources textuelles en spécialités médicales : en amont et en aval de l'indexation MeSH des ressources.

Le cinquième chapitre présente la problématique centrale de nos travaux, l'indexation automatique fine de ressources textuelles à l'aide du thésaurus MeSH. Nous passons en revue les problèmes soulevés par toute tâche d'indexation, à savoir, la caractérisation même de l'activité d'indexation, l'opportunité d'utiliser un vocabulaire contrôlé, les moyens d'évaluer une indexation donnée. En ce qui concerne l'indexation automatique, nous insistons plus particulièrement sur la méthodologie d'analyse des textes, la faisabilité d'une telle tâche, et les outils existants. Nous présentons ensuite l'outil MAIF (accronyme anglais de « Indexation Automatique MeSH pour le Français ») que nous avons développé en relevant le défi de proposer automatiquement une indexation par paires de descripteurs MeSH (mot clé/qualificatif), contrairement à l'approche automatique habituelle qui est fondée sur l'extraction de descripteurs isolés. Les deux approches de Traitement Automatique de la Langue et d'apprentissage (k Plus Proches Voisins) combinées dans MAIF sont détaillées. Finalement, nous résumons les évaluations de MAIF et d'autres logiciels d'indexation MeSH français, suisses et américains que nous avons mises en place et réalisées.

Le dernier chapitre dresse un bilan sur le travail réalisé dans le cadre de cette thèse et rassemble les perspectives de recherche qui s'en dégagent, aussi bien en ce qui concerne l'amélioration pratique des solutions proposées que notre contribution sur l'analyse des documents de santé.

Chapitre 1

Contexte de travail

Dans cette thèse, notre réflexion s'est orientée vers l'automatisation des tâches documentaires dans un catalogue de santé en ligne. Afin de situer cette problématique, il convient de définir au préalable les différentes notions auxquelles elle renvoie. Ainsi, avant de décrire les tâches documentaires dont il est question, nous examinons la définition d'un « document » et précisons la nature des documents avec lesquels nous travaillons. Le contexte spécifique où ils apparaissent, à savoir, un catalogue de santé en ligne, nous amène à prendre en compte les aspects numériques des documents apparaissant sur l'Internet. Par ailleurs, nos travaux ont porté sur un catalogue particulier, le Catalogue et Index des Sites Médicaux Francophones (CISMeF) (Darmoni et al., 2000).

1.1 Notion de document et de ressource

Qu'est ce qu'un document ? Il est communément admis depuis le début du XX^{éme} siècle qu'un document peut prendre de multiples formes, non nécessairement textuelles. Ainsi, Suzanne Briet commence son ouvrage sur la documentation (Briet, 1951) en rappelant le sens classique du document[1] comme « enseignement ou preuve » avant d'illustrer grâce à l'exemple de l'antilope la diversité sous laquelle le document peut se présenter : un spécimen d'Antilope rare capturée en Afrique et exposée au jardin des plantes de Paris est, selon la définition que nous venons de donner, un document - il s'agit bien d'une preuve de l'existence de l'animal. De même, le communiqué de Presse annonçant sa capture, les photographies prises de l'antilope, le rapport des zoologistes l'ayant étudiée etc. sont eux aussi des documents (plus exactement, selon Briet, des documents secondaires). Buckland (Buckland, 1997) reprend la définition du document de S. Briet, ainsi que celles qui ont pu être introduites par la suite et conclut sur l'ambivalence du document en tant que preuve ou information de première main et en tant qu'objet susceptible d'être réapproprié et utilisé par celui qui le consulte.

Accart and Rethy (2003), quant à eux donnent une définition plus technique du document, restreinte à ses formes les plus modernes. Le document serait donc l'« ensemble formé par un support et une information, généralement enregistré de façon permanente et tel qu'il puisse être lu par l'homme ou la machine. » Ici, la notion d'accessibilité au document intervient dans sa définition même.

[1] Cette définition semble toujours en vigueur un demi-siècle plus tard. *Document : Renseignement écrit ou objet servant de preuve ou d'information. (Larousse, 1980)*

Un Site Web est-il un document ? Afin d'examiner cette question, il convient tout d'abord de définir ce que l'on entend par *site web*. A priori, on peut dire qu'un site web est un ensemble de pages HTML[2] regroupées par des hyperliens sous un même domaine, accessible grâce à un navigateur. Cependant, cette définition technique ne préjuge en rien du *contenu* des sites web. En effet, les pages HTML présentent généralement du texte et/ou des images, ce qui permet de les rapprocher d'un document papier traditionnel. Mais qu'en est-il des services web et autres outils souvent disponibles en ligne ? Par ailleurs, certains domaines (sites) hébergent plusieurs centaines de pages, dont le contenu peut être tout à fait hétérogène (textes, images, services web, ...). Il semble donc abusif d'assimiler systématiquement un site web à un document. Afin de contourner ce problème, le W3C (World Wide Web Consortium) évoque la notion de *ressource*[3]. Une ressource est définie comme le plus petit élément constitutif de l'Internet, pouvant être identifié et situé de manière unique grâce à un URI (Unique Resource Identifier) et un URL (Unique Resource Locator). L'existance de sites miroir remet en cause cette unicité, mais nous pouvons néanmoins conclure qu'une ressource peut coincider avec un document, bien que cette notion englobe également d'autres types d'objets (outils). Nous retiendrons d'une part, la diversité du document, - dans le cadre de notre étude, cela signifie que sont susceptibles d'être catalogués dans CISMeF tous les supports électroniques d'information médicale : les cours des facultés de médecine francophones, les articles scientifiques des revues bio-médicales, les recommandations pour la bonne pratique clinique, des dépliants préventifs émanant des autorités sanitaires, des images médicales etc. D'autre part, la perspective de réappropriation du document par l'utilisateur met en évidence l'importance du rôle du documentaliste lors de la sélection des documents dans un catalogue réalisé manuellement (Thirion & Darmoni, 1998), par opposition aux moteurs de recherche dans lesquels la sélection et l'indexation des ressources repose sur des critères statistiques. Dans un catalogue, il s'agit pour le documentaliste de mettre à la disposition du public non seulement le document lui-même, mais également les éléments qui permettent une réflexion sur le contenu et l'usage qu'il est possible d'en faire. On touche ici à la problématique des critères de qualité de l'information, un thème de recherche majeur de l'équipe CISMeF : comment reconnaître un document sérieux d'un document fantaisiste ? Comment évaluer la qualité de l'information ? A ce sujet, la juridiction du Québec met en avant le critère de *fiabilité*. Selon le texte québecois, la fiabilité d'un document repose sur son intégrité, qui doit être maintenue au cours de tout son cycle de vie. La traçabilité des modifications apportées est également à prendre en compte selon d'autres échelles d'évaluation. Les réflexions sur les critères de qualité de l'information dans le domaine de la santé ont donné lieu à l'élaboration de plusieurs grilles d'évaluation : HON[4] , NetScoring[5] (Darmoni, Leroux, Thirion, Santamaria, & Gea, 1999), MedCircle[6] (Darmoni, Mayer, Thomeczek, & Eysenbach, 2002).

L'Internet. Les progrès techniques réalisés depuis l'invention de l'imprimerie font que la fabrication et la publication de documents de toutes sortes sont à la portée de chacun depuis l'avènement de l'Internet. Les conséquences en terme d'accès à l'information se manifestent par des problèmes de recensement de l'information : comment avoir connais-

[2]Hyper Text Mark-up Language
[3]`http://www.w3.org/2001/tag/webarch/errata.html`
[4]`http://www.hon.ch`
[5]`http://www.chu-rouen.fr/netscoring`
[6]`http://www.medcircle.org/`

sance de toutes les nouvelles ressources publiées, de leur localisation sur la Toile, de leur déplacement, de leur disparition... La question du catalogage et de la description des ressources, centrale à nos travaux découle naturellement de l'abondance des informations disponibles. La comparaison d'Internet à une immense « bibliothèque virtuelle » est désormais courante. Dans un article promouvant les bibliothèques traditionnelles, Mark Herring (Herring, 2001) rappelle les limites de cette « bibliothèque virtuelle » :

- la *non-exhaustivité des informations* disponibles malgré leur abondance. La plupart des documents disponibles en ligne sont postérieurs à 1990 et seuls 8% des journaux sont disponibles en ligne. les problèmes liés à la numérisation de documents papiers destinés à être mis en ligne. Cette opération coûteuse est, d'après l'auteur, souvent réalisée au détriment de l'intégrité des documents, suite aux difficultés techniques liées à l'intégration des tableaux ou des notes de bas de page. Le projet récent de « bibliothèque virtuelle » de la société Google souligne l'actualité de ce point.
- le *manque d'outils adaptés à la recherche d'information*, voire leur partialité. En effet, en 2005, seuls quelques outils de recherche d'information sont disponibles et utilisés par la quasi-totalité des internautes. Ainsi, les outils de recherche d'information jouent sur l'internet le rôle des documentalistes dans les bibliothèques traditionnelles : celui de guider l'utilisateur vers l'information.
- l'*absence de contrôle qualité*

Autrement dit, non seulement il faut pouvoir accéder à l'information, mais il faut aussi savoir de quelle information il s'agit, où trouver des informations sur le même thème etc. Cette constatation souligne le rôle important que doivent jouer les professionnels de la documentation pour guider les utilisateurs dans leur recherche d'information, ainsi que dans le développement d'outils automatisés à cet effet. Ainsi, observe-t-on l'émergence de grilles d'évaluation de l'information (évoquées plus haut) mais aussi des services en ligne pour réaliser ce type d'évaluation. Par exemple, la société Temesis[7] propose des formations aux critères de qualité et d'accessibilité des sites Internet et des outils d'évaluation de la conformité des sites à ces critères.

Dimension numérique. Le facteur numérique pose notamment le problème de la *délimitation* de la ressource. En effet, avec l'avènement d'Internet comme moyen de véhiculer l'information et en particulier sur un support multimédia, le document numérique n'est pas une entité autonome que l'on peut tenir entre ses mains. Les possibilités de renvois dynamiques à la fois à l'intérieur et à l'extérieur de la ressource élargissent la navigation à l'Internet entier à partir d'un seul document. Dès lors, comment repérer les limites d'un document numérique ? Quels sont les critères définissant une telle ressource ? Comme le montrent les deux exemples ci-dessous, une simple URL ne peut répondre de manière satisfaisante à ces questions :

1. URL contenant un résumé de la ressource et un lien vers la ressource complète :
   `http://www.ladocfrancaise.gouv.fr/brp/notices/044000497.shtml`
2. URL contenant une (petite) partie de la ressource et des liens vers d'autres parties :
   `http://www-smbh.univ-paris13.fr/smbh/pedago/histologie/histologie_pcem1/conjonctif_histo_p1.html`

(Fowler, Maram, Kouramajian, & Devadhar, 1995) proposent par exemple de considérer comme un document un ensemble de pages web reliées entre elles par des hyperliens

[7] `http://www.temesis.com/` (accédé le 30/08/05)

indépendemment du domaine ou du site éditeur. Ces ensembles de pages web seraient alors identifiés comme des « regroupement naturels » (natural clusters[8]) et définiraient des documents.

Après avoir cerné le support des informations en ligne, nous allons voir à travers un exemple comment les gérer dans le cadre d'un catalogue, afin de faciliter l'accès des utilisateurs aux ressources de santé en français.

1.2 Catalogue et Index des Sites Médicaux Francophones : CISMeF

1.2.1 Présentation du projet

Le nombre de documents disponibles sur l'Internet est considérable et il croît de manière exponentielle : près de 8,1 milliards de pages web recensées en 2005 contre 4,3 en 2004[9] ... Le recensement et la recherche d'information demeurent problématiques : les utilisateurs rencontrent toujours des difficultés à trouver ce qu'ils cherchent malgré les divers outils (catalogues en ligne, moteurs de recherche ou encore méta-moteurs) mis à leur disposition pour les assister dans leur recherche, aussi bien dans le domaine général que dans des domaines spécialisés. L'un des obstacles rencontrés, en particulier dans les domaines scientifiques, est celui de la validité des informations. Afin d'apporter une réponse à cette double préoccupation dans le domaine de la santé (recherche d'information et assurance de qualité de l'information), le projet CISMeF (Darmoni et al., 2000) (ibid. 2001) a été initié en 1995. Initialement mis en place à des fins internes au CHU de Rouen, le catalogue s'est développé pour aider les professionnels de santé, les étudiants en médecine et le grand public dans leur recherche d'information de santé sur l'Internet. A l'heure actuelle, le catalogue est utilisé aussi bien par les internautes en France qu'à l'étranger [10] et sa popularité est en forte progression : le nombre moyen de sessions par jour ouvré était d'environ 5.000 en 2002, contre environ 35.000 en 2005. Les documentalistes CISMeF sélectionnent des ressources en français (n=14.538[11]) en fonction de critères stricts. En moyenne, une cinquantaine de ressources et leur notice descriptive sont ajoutées chaque semaine. Les ressources sont décrites à l'aide d'un ensemble de métadonnées et d'une terminologie structurée du domaine médical. Les notices (dont le contenu est détaillé à la section 1.2.3 ci-dessous) et la terminologie (présentée au chapitre 2) sont stockées dans une base de données relationnelle Oracle®. CISMeF a donc pour objectifs le recensement, la sélection, l'organisation et l'indexation de l'information de qualité en santé disponible sur l'Internet.

1.2.2 Structure du catalogue

CISMeF propose un accès aux ressources de santé du catalogue selon cinq modes de recherche différents :

– Par le biais du *moteur de recherche Doc'CISMeF*. L'utilisateur est invité à formuler une requête en langue naturelle en français ou en anglais (recherche simple), ou bien

[8] Voir (Botafogo, 1993) - cité par (Fowler et al., 1995) pour une méthode d'identification de ces « clusters ».

[9] Chiffres recueillis sur Google (`http://google.fr`) en janvier 2004 et avril 2005

[10] Par exemple, le 22 février 2005, environ 38% des sessions enregistrées avaient pour origine la France, 42% étaient d'origine internationale et 20% d'origine indéterminée.

[11] Le 19 Avril 2005

composée de termes MeSH et/ou CISMeF[12] (recherche avancée), ou encore une requête booléenne utilisant les opérateurs ET, OU, NON (recherche booléenne).
- Par le biais d'un *classement alphabétique des termes MeSH et CISMeF*. Cet index présente la traduction française des termes du thesaurus MeSH et indique également les termes américains originaux. Ce classement donne accès aux mots clés, ainsi qu'aux qualificatifs et aux types de ressources utilisés dans CISMeF. A chaque terme correspond une page HTML descriptive (cf. Fig 1.1), présentant notamment une définition du terme (issue du Vidal™ de la famille, de la traduction des définitions de la NLM, etc.) et des requêtes pré-formatées. Activées par un lien dynamique, ces requêtes utilisent l'outil de recherche pour proposer un accès aux ressources destinées aux professionnels, aux patients ou aux étudiants.
- Par le biais d'un *classement thématique par spécialité médicale*. A chaque spécialité correspond une page HTML présentant tous les termes CISMeF qui lui sont liés sémantiquement. Comme sur les pages consacrées aux termes, des requêtes pré-formatées permettent d'accéder aux ressources relatives soit à la spécialité médicale choisie, soit à l'un des termes qui lui sont sémantiquement liés.
- Par le biais du *moteur de recherche terminologique*. Les requêtes en langage naturel renvoient des informations disponibles dans la terminologie (définition, synonymes, qualificatifs affiliables pour les mots clés, position dans la hiérarchie etc.). Pour chaque terme, des requêtes pré-formatées permettent d'accéder aux ressources correspondantes en français dans le catalogue CISMeF ou en anglais dans la base MEDLINE.
- Par le biais des *types de ressources*.

CISMeF permet également aux utilisateurs d'accéder aux ressources par ces différents moyens en fonction de leur profil (médecin, étudiant, patient).

1.2.3 Notice CISMeF

Contenu d'une Notice CISMeF

A chaque ressource d'information de santé est associée une notice ou fiche descriptive. Ces notices sont similaires à des « annotations » et servent de support à la recherche d'information au sein du catalogue. Il est important de remarquer qu'aucune modification ou annotation ne sont effectuées sur la ressource elle-même. Les notices CISMeF sont élaborées par les documentalistes de l'équipe et contiennent plusieurs types d'information :

- Une présentation contenant des informations générales sur le contenu et la qualité de la ressource : le titre, le nom du ou des auteurs, un résumé succinct, la source, le niveau de preuve, le type de ressource.
- Une classification contenant des informations détaillées sur le contenu de la ressource : la liste des spécialités médicales et des mots clés (ou paires mot clé / qualificatif) MeSH® (Medical Subject Headings).
- Des informations pratiques sur la ressource : l'URL, le format, la langue, le type d'accès (libre, restreint, payant), la date de consultation...

Pour plus de lisibilité, les notices courtes (cf. exemple figure 1.2) présentant une partie des informations sont affichées en réponse à une requête. Une icône cliquable à droite du titre de la ressource permet alors d'accéder à la notice longue (cf. exemple figure 1.3) contenant l'ensemble des informations.

[12] Le thésaurus MeSH et la terminolologie CISMeF sont présentés en détail au chapitre 2.

FIG. 1.1 – Extrait de la page descriptive CISMeF pour le mot clé *diabète gestationnel*

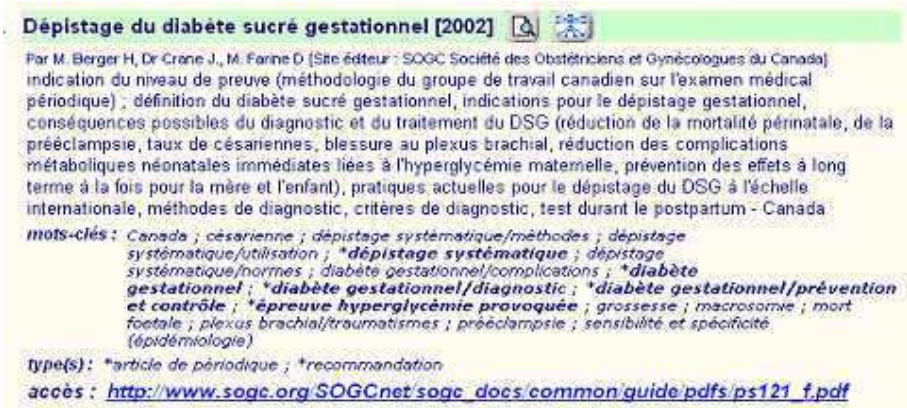

FIG. 1.2 – Extrait de la notice CISMeF n° 9785 (Notice courte)

FIG. 1.3 – Extrait de la notice CISMeF n° 9785 (Notice longue)

Intérêt pour l'utilisateur

Le concept de métadonnée est apparu bien avant l'Internet, mais son intérêt a été décuplé par le nombre croissant de publications électroniques. A l'initiative du W3C, les métadonnées sont utilisées pour décrire le contenu des pages Web. En utilisant un formalisme approprié, les métadonnées permettent donc une caractérisation non ambiguë des ressources et facilitent la recherche d'information dans les pages concernées. Ce formalisme est l'un des outils participant à construction du Web Sémantique. Dans le cadre de l'élaboration du Web Sémantique Médical, les notices CISMeF contiennent des métadonnées standard d'origines diverses (Thirion, Loosli, Douyère, & Darmoni, 2003).

La liste des champs à renseigner dans la notice CISMeF a été établie à l'aide des critères de qualité du Net Scoring[13] et des métadonnées du Dublin Core (DC) (Baker, 2000) pour les champs *auteur, date, description, format, identification, langue, éditeur, type de ressource, droits, sujet* et *titre*. Pour décrire les ressources pédagogiques, onze éléments de la catégorie « Education » du IEEE 1484 LOM (Learning Object Metadata) sont utilisés en plus des autres métadonnées. Par ailleurs, des métadonnées spécifiques à CISMeF, ont été ajoutées pour décrire la qualité ou la localisation de la ressource : *institution, ville, province, pays, type d'accès, partenariat, coût* et *public ciblé*. Deux champs supplémentaires ont été créés pour les ressources destinées aux professionnels de santé : *indication du niveau de preuve* et *la méthode utilisée* pour l'établir (Darmoni, Amsallem, et al., 2003). Les métadonnées HIDDEL[14] ont été introduites dans CISMeF dans le cadre du projet européen MedCircle (Mayer, Darmoni, Fiene, & Al., 2003), qui a pour but d'évaluer la qualité de l'information de santé afin de guider les utilisateurs vers des sources fiables.

L'objet de cette démarche est de fournir des informations sur le contenu et la qualité d'une

[13] `http://www.chu-rouen.fr/netscoring/`

[14] High Information Description Disclosure Evaluation Language - cf. `http://www.medcircle.org/`

ressource de manière synthétique. La santé est l'un des domaines ou la qualité et la fiabilité des informations consultées sont les plus cruciales. Il est donc important de sensibiliser l'utilisateur aux critères qui peuvent attester de la validité des informations données. Les notices très détaillées de CISMeF fournissent les éléments de réponse nécessaires. L'Organisation Mondiale de la Santé (OMS), en date d'un communiqué du 10 mai 2005[15], cite CISMeF parmi les « sites web respectant les bonnes pratiques en matière d'information sur la sécurité des vaccins » (c'est le seul site français mentionné par l'OMS). La saisie de tous ces champs s'effectue sous Oracle, avec un écran de saisie de présentation Access.

1.2.4 Accessibilité par l'Homme et par la Machine

Au début du projet CISMeF le standard HTML 2.0 était utilisé pour que le site soit lisible par la grande majorité des navigateurs. Le standard plus récent HTML 4.0 est maintenant employé. Depuis juin 2001, l'utilisation du standard XML permet une interopérabilité avec d'autres catalogues ou serveurs de ressources (e-learning dans le cadre du projet UMVF[16]). Depuis décembre 2002, CISMeF a adopté le format RDF[17] (into HTML) : les ressources sont maintenant décrites au format RDF d'après les concepts de l'ontologie HIDDEL.

1.2.5 Présentation de l'équipe et des projets

Sous la co-direction du responsable des technologies de l'information et de la communication (Stéfan Darmoni) et du conservateur de la bibliothèque médicale (Benoît Thirion) du CHU de Rouen, l'équipe CISMeF est composée de quatre documentalistes experts du domaine médical, d'un ingénieur de recherche, d'un docteur en informatique et de quatre doctorants. La figure 1.4 ci-dessous illustre les différents projets en cours dans l'équipe, ainsi que le rôle de chacun.

Au centre des activités de l'équipe CISMeF se trouve la terminologie CISMeF (que nous décrivons à la section 2.3. En effet, c'est sur cette terminologie que repose la description des ressources et donc le contenu du catalogue. Il est donc naturel pour l'équipe CISMeF d'être impliquée dans des travaux touchant à la terminologie médicale - essentiellement le MeSH mais aussi d'autres terminologies avec lesquelles des correspondances peuvent s'établir, comme la SNOMED, la CIM-10 ou toute terminologie incluse dans l'UMLS.

Autour de la terminologie CISMeF, se trouvent trois problématiques principales : la recherche d'information (à gauche), l'indexation (à droite) et la définition de critères de qualité de l'information (en haut). Des activités de valorisation (en bas) sont également entreprises, afin de financer le fonctionnement de cette équipe de Recherche et Développement. Nous ne les détaillerons pas ici.

La première étape de la création du catalogue consiste à effectuer une veille documentaire des publications médicales, et à sélectionner des ressources à l'intention des lecteurs ciblés. L'un des aspects importants de la plus-value d'un catalogue réside dans le soin apporté à cette sélection. Ainsi, la problématique de définition de critères de qualité de l'information de santé est primordiale.

[15] cf. `http://www.who.int/mediacentre/news/notes/2005/np09/fr/` pour plus détails sur les critères de qualité de l'information pris en compte par l'OMS.

[16] Université Médicale Virtuelle Francophone - cf. `http://www.umvf.prd.fr`(accédé le 10/01/04)

[17] cf. `http://www.w3.org/RDF/` (accédé le 10/01/04)

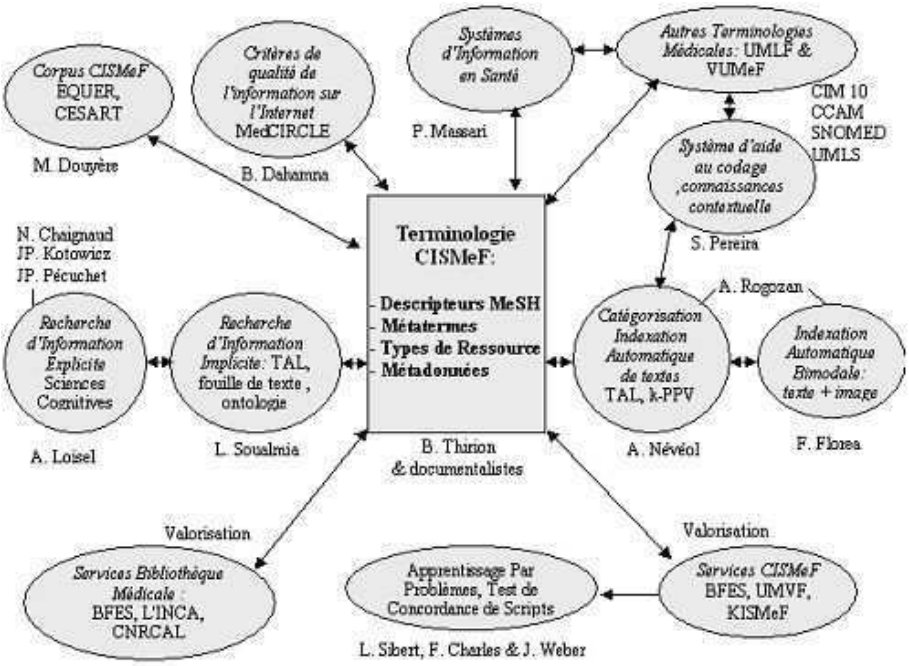

FIG. 1.4 – Organigramme des projets CISMeF

Une fois les documents sélectionnés, l'utilisateur en quête d'information se trouve face à deux stratégies de recherche. L'une, entièrement automatique, s'effectue par le biais d'une requête à l'aide du moteur de recherche Doc'CISMeF. L'autre stratégie sollicite l'intervention d'un expert dans la formulation de la requête automatique. Dans le premier cas (requête entièrement automatique), lors de l'utilisation de la recherche « simple » de Doc' CISMeF (en place depuis 2000), la requête en langage naturel est interprétée à l'aide de la terminologie CISMeF afin d'aboutir à une requête formelle utilisant le vocabulaire de l'indexation. C'est le plus souvent l'échec de cette stratégie, ou l'incompréhension de l'utilisateur face au vocabulaire de l'indexation qui déclenche la deuxième stratégie, l'interrogation du documentaliste/expert. Le rôle de l'expert est alors de « traduire » le besoin d'information exprimé par l'utilisateur dans le langage utilisé pour la description des ressources. Le projet Cogni-CISMeF (Loisel et al., 2005) vise à développer un système de recherche d'information capable de modéliser le dialogue entre utilisateur et expert, afin de réaliser cette traduction.

Pour que ces recherches soient possibles, il est nécessaire que les ressources du catalogue soient indexées. Cette tâche peut être effectuée manuellement par les documentalistes. Mais, en parallèle, deux projets d'indexation automatique portent sur les supports d'information présents dans CISMeF : le texte et l'image.

Nos travaux ont porté sur différents projets (terminologie, veille,...) mais, comme l'indique le schéma, se sont concentrés sur l'indexation automatique de textes médicaux à l'aide de mots clés MeSH.

1.2.6 Le corpus CISMeF

L'ensemble des notices CISMeF décrites ci-dessus constitue un corpus de travail considérable (N∼14.000), permettant d'accéder à des ressources de santé étiquetées notamment à l'aide de descripteurs MeSH. Cependant, ce corpus diffère fondamentalement des corpus classiques en recherche d'information tels que Reuters, Medline ou TREC dans la mesure où ni le texte intégral des ressources ni même un résumé structuré et standardisé n'est directement disponible. Le contenu des ressources est pointé par une (ou plusieurs) URLs. Nous évoquons à la section 3 l'impact de cette caractéristique sur la politique éditoriale et la maintenance du catalogue. Nous évoquons également à la section 5.5 les conséquences techniques sur l'accès au texte intégral des ressources. Dans le cadre de nos travaux en indexation, nous avons été amenée à constituer plusieurs corpus d'évaluation de manière semi-automatique. Ces corpus sont au nombre de trois :

1. **Le corpus monolingue français « diabète » :** constitué de 57 ressources extraites du catalogue CISMeF à l'aide de la requête booléenne « diabète.mc[majeur][18] ». Les ressources choisies sont diversifiées, au niveau du type de ressource (à l'exception des types de ressources tels que « association patient » ou « réseaux coordonnés », etc.) ou de la longueur de la ressource. (740.000 mots). Les titres des paragraphes et/ou le sommaire des ressources ont également été extraits à la main. (16.900 mots).
2. **Le corpus monolingue français « divers » :** constitué de 32 ressources extraites aléatoirement du catalogue CISMeF. (163.000 mots)
3. **Le corpus bilingue anglais-français « ENFR » :** constitué de 51 ressources extraites du catalogue CISMeF à l'aide de la requête booléenne « anglais.la[19] » (295.000 mots - dont 160.000 mots pour la version française seule).

La figure 1.5 présente un extrait de la ressource décrite par la notice CISMeF n°8916. Le texte intégral de cette ressource est inclus dans le corpus « diabète ».

Pour la constitution des corpus, les ressources sélectionnées ont été téléchargées automatiquement puis converties au format texte. Une intervention manuelle a été nécessaire à plusieurs niveaux. Tout d'abord, pour la sélection des ressources, il était nécessaire de vérifier que l'URL correspondait bien au texte intégral de la ressource souhaitée. Dans certains cas, un ajustement manuel a été nécessaire afin de compléter le texte de la ressource, ou d'accéder au texte intégral effectif. Les textes téléchargés ont subi un post-traitement nécessitant une intervention humaine : des informations superflues telles que les menus ou les encarts ne concernant pas la ressource ont été supprimés. Certains caractères spéciaux ont été normalisés. Le coût en temps de l'intervention humaine explique la taille relativement faible de ces corpus par rapport à l'ensemble des ressources décrites dans le catalogue (environ 14 000 ressources). De plus, l'utilisation des corpus dans le cadre de l'évaluation d'outils d'indexation automatique nécessitait une identification des ressources permettant de faire le lien avec l'indexation présente dans les notices CISMeF, considérée lors des évaluations comme l'indexation de référence.

Le corpus bilingue « EN/FR » a également nécessité un traitement particulier à plusieurs niveaux :

[18] Dans cette requête, « .mc » désigne le champ *mot-clé MeSH* et « [majeur] »permet de sélectionner les ressources où ce concept est un thème principal

[19] Dans cette requête, « .la » désigne le champ *langue*

(...)
L'apparition du diabète de type 2 chez l'enfant et ses implications en santé publique

Alors que l'épidémie d'obésité s'étend dans le monde industrialisé, les cliniciens décrivent les premières séries de cas de diabète de type 2 chez l'enfant dans diverses parties du monde. Aux Etats-Unis et au Royaume-Uni, des enquêtes épidémiologiques sont en cours visant à mieux définir l'ampleur et l'évolution du problème et à caractériser les enfants diagnostiqués afin de mieux différencier le diabète de type 2 du type 1. En France où la première série de cas vient d'être publiée, le diabète de type 2 de l'enfant pourrait également être méconnu, mal classé ou non rapporté. Le programme national de prévention en nutrition constitue la première étape de lutte contre ce problème de santé publique potentiel.
(...)

FIG. 1.5 – Extrait du périodique décrit par la notice CISMeF n°8916 - le texte intégral est accessible à partir de la page `http://www.invs.sante.fr/beh/2002/20_21/index.htm`

1. Alignement au niveau des textes. Cet alignement était nécessaire dans le cadre de l'évaluation comparée d'outils d'indexation automatique MeSH anglais et français (détaillée à la section 5.9.3). Il fallait alors disposer du texte d'une ressource en français, de sa traduction en anglais et de l'indexation manuelle MeSH correspondante.
2. Alignement au niveau des paragraphes. Cet alignement était nécessaire dans le cadre de la traduction automatique de termes médicaux (détaillée à la section 2.5.2).

Ainsi, compte tenu de ce double objectif applicatif, certaines ressources, comme les sites des hôpitaux, qui ne présentaient pas d'intérêt pour l'acquisition de traduction de synonymes ont été écartées. D'autres ressources contenaient un résumé anglais d'un article développé en français, ou présentaient les textes sans séparation nette entre les deux langues. Elles ont été également écartées. La majorité des ressources restantes émanaient de sites éditeurs bilingues affiliés au ministère de la santé canadien[20], ce qui est une garantie de la qualité de la traduction disponible. De plus, ces sites observent un classement régulier et organisé des documents dans les différentes langues. Nous étions donc en mesure de déduire l'URL de la version anglaise[21] de la ressource à partir de l'URL de la version française, ou bien dans certains cas, à partir de la ressource elle-même, lorsque celle-ci contient un lien vers la version anglaise. L'alignement au niveau des textes a donc été effectué de cette manière. Nous avons ensuite utilisé une méthode d'alignement au niveau des paragraphes fondée sur le parallélisme entre

[20]La société canadienne de pédiatrie (`http://www.cps.ca`), Santé Canada (`http://www.hc-sc.gc.ca`) et le ministère de la santé et des soins de longue durée de l'Ontario (`http://www.gov.on.ca/health/indexf.html`). Pour l'application de traduction automatique, il n'était pas nécessaire de disposer de l'indexation MeSH des ressources utilisées. Aussi, nous avons complété le corpus de travail bilingue à l'aide de ressources émanant de ces mêmes sites éditeurs, mais non référencées dans CISMeF (N=25). Pour la suite, nous faisons référence à la réunion du corpus ENFR avec ces ressources par corpus « ENFR complété »

[21]Seules les URLs des versions françaises sont disponibles dans CIMeF

la structure d'une ressource et celle de sa traduction. En effet, pour la majorité des ressources, le premier paragraphe de la version française constitue la traduction du premier paragraphe de la version anglaise et ainsi de suite. Nous avons donc procédé à l'alignement au niveau des paragraphes de manière automatique, modulo quelques ajustements réalisés manuellement pour rétablir le parallélisme de structure dans certaines ressources.

Afin de nous assurer de la validité des corpus présentés ci-dessus, nous avons examiné les critères énoncés par (Bommier-Pincemin, 1999) :

Pertinence. Les corpus constitués sont destinés à l'évaluation de systèmes d'indexation MeSH de textes médicaux francophones, dans le cadre d'un catalogue de santé. Les textes choisis doivent donc correspondre au thème de la santé, être annotés à l'aide de descripteurs MeSH et correspondre aux types de documents susceptibles d'être inclus dans un catalogue de santé. Nos corpus étant constitués de ressources référencées dans CISMeF, ils remplissent ces critères.

Représentativité. Les corpus doivent couvrir les phénomènes que l'on cherche à étudier ou à décrire. Dans le cadre de l'indexation MeSH, nous nous sommes particulièrement intéressés à l'utilisation de paires de descripteurs MeSH et de descripteurs obligatoire (cf. section 5.8).

Utilisabilité. Pour que l'étude soit significative, le corpus doit être exhaustif ou statistiquement représentatif.

Le tableau 1.1 ci-dessous présente une description des corpus de travail, ainsi que du corpus global CISMeF (pour ce corpus, nous ne disposons pas d'informations relatives au nombre de mots). Il y a bien cohérence entre la proportion des phénomènes étudiés pour l'indexation (Check tags, paires de descripteurs MeSH) dans chaque corpus et dans le corpus global, bien que la proportion de paires soit plus élevée dans les corpus « ENFR » et « diabète » (48% et 46% respectivement, contre 38% dans le corpus CISMeF). De même, la proportion de check tags est plus élevée dans les corpus « ENFR » et « divers » (13% et 9% respectivement, contre 4% dans le corpus CISMeF). Malgré ces différences, on peut considérer que les corpus de travail sont représentatifs de l'ensemble du corpus CISMeF.

Dans le cadre des projets technolangue[22] EQUER (portant sur l'Evaluation des systèmes de QUEstion-Réponse) et CESART (portant sur l'évaluation dans le domaine des outils d'acquisition de ressources terminologiques à partir de corpus écrits, y compris les outils multilingues) les corpus « diabète » et « divers » pour EQUER et « ENFR » pour CESART ont été repris et complétés automatiquement avec des ressources provenant des mêmes sites éditeurs. Pour les campagnes EQUER et CESART, certaines étapes réalisées manuellement sur les corpus « diabète », « divers » et « ENFR » n'étaient pas nécessaires, car la correspondance entre ressource et notice CISMeF n'était pas exploitée ultérieurement. Par ailleurs, aucune extraction (manuelle ou automatique) des titres de paragraphe et/ou sommaire n'a été réalisée. Cela explique la différence d'échelle entre nos corpus et les corpus des projets EQUER/CESART.

1.3 Problématique de la thèse

La problématique de notre thèse s'inscrit dans la dynamique du catalogue et participe à son développement, sa mise à jour et son adaptation au paysage documentaire en constante

[22] cf. `http://www.technolangue.net/article61.html` pour une description détaillée du projet EQUER et `http://www.technolangue.net/article58.html` pour CESART.

	CISMeF	diabète	divers	ENFR
Nb ressources	14 329	56	32	51
Nb de descripteurs distincts (i.e. mot clé seul ou paire)	24 030	502	345	300
Nb total de descripteurs MeSH	103 376	802	436	430
Nb moyen de descripteurs	7,22	14,32	13,65	8,45
Nb total de paires	39 983	371	154	208
Proportion des paires (%)	38,68	46,26	35,32	48,26
Nb moyen de paires	2,79	6,63	4,81	4,08
Nb total de Check Tags	4371	28	39	57
Proportion des Check Tags (%)	4,23	3,49	8,94	13,26
Nb moyen de Check Tag	0,31	0,50	1,22	1,12
Nb Total de mots	-	740 000	163 000	160 000
Nb moyen de mots	-	12 900	5 100	3 100

Tab. 1.1 – Description des corpus de travail et du corpus CISMeF

évolution qu'est l'Internet. L'ajout d'une nouvelle ressource au catalogue se fait en quatre étapes représentatives des tâches documentaires effectuées au sein du catalogue :

1. Recensement des ressources (Veille documentaire)
2. Validation des ressources (Sélection)
3. Description d'une ressource (Indexation MeSH et création d'une Notice)
4. Mise en ligne de la notice descriptive

La figure 1.6 ci-dessous illustre ce processus, tout en situant le rôle de l'utilisateur qui est à la fois l'initiateur et le bénéficiaire de la chaîne de traitement. En effet, l'utilisateur approche le catalogue avec un besoin d'information à satisfaire. Il peut effectuer sa recherche d'information soit directement à l'aide du moteur de recherche mis à sa disposition, soit par l'intermédiaire d'un documentaliste. Deux cas de figure sont alors possibles :

- le besoin d'information de l'utilisateur ne peut être satisfait- la requête reçoit « zéro réponse » : le processus doit être initié afin de sélectionner des ressources correspondant à la requête, de les décrire et de les intégrer au catalogue.
- le besoin d'information de l'utilisateur ne peut être satisfait- la requête reçoit néanmoins des réponses : contrairement au cas de figure précédent, il n'est pas possible de déterminer si la demande d'information a été réellement satisfaite sans un retour de l'utilisateur lui-même. Afin d'initier un tel dialogue avec l'utilisateur, le projet Cogni-CISMeF (Loisel et al., 2005) a été initié en 2004.
- le besoin d'information de l'utilisateur est satisfait : le catalogue contient une sélection de ressources appropriée à la requête, ainsi que leur description. Il est important de remarquer les deux modes de recherche offerts à l'utilisateur : l'interrogation d'un expert humain et l'utilisation d'un moteur de recherche. Dans les deux cas, l'interface entre utilisateur et ressource finale est le vocabulaire contrôlé utilisé par les documentalistes pour la description des ressources.

Sur la figure 1.6, les tâches documentaires effectuées à la main au début de nos travaux (Septembre 2002) sont représentées en vert et les tâches effectuées automatiquement en rouge :

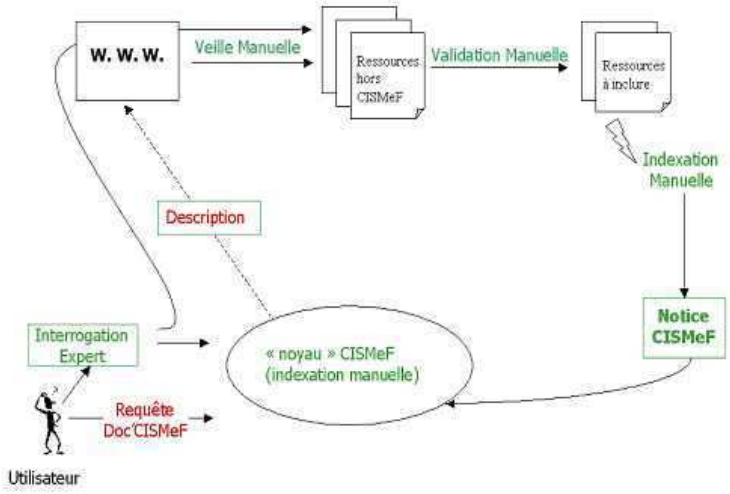

FIG. 1.6 – Fonctionnement de CISMeF avant automatisation des taches documentaires (en 2002)

seule la recherche d'information et l'environnement de mise à jour du catalogue bénéficient d'outils automatiques.

Dans les chapitres suivants, nous présentons les solutions que nous avons pu proposer pour l'automatisation des tâches de veille, de catégorisation et d'indexation. Le bilan de ce travail au chapitre 6 illustre par une nouvelle figure l'impact de nos réalisations sur le fonctionnement de CISMeF.

Chapitre 2

Terminologie Médicale

La problématique abordée dans le cadre du catalogue CISMeF est double : nous nous intéressons à la représentation de ressources de santé et à la mise à disposition de ces ressources auprès des professionnels de la santé, des étudiants en médecine et des patients. Ce chapitre présente les travaux existants les plus proches de notre problématique. Tout d'abord, nous présentons deux approches de la représentation des connaissances, par le biais d'une terminologie ou d'une ontologie. Nous définissons chaque système et discutons des méthodes et enjeux de la construction de telles ressources terminologiques, illustrés par des exemples dans le domaine médical. Par la suite, nous détaillons plus particulièrement le thésaurus MeSH et la terminologie CISMeF, qui sont au coeur de notre thèse, ainsi que la CIM et la SNOMED, deux autres terminologies avec lesquelles nous avons été amenée à travailler. Nous présentons ensuite les directions de recherche actuelles en traitement automatique de la langue médicale, plus particulièrement autour de questions terminologiques. Enfin, nous exposons notre propre contribution à l'enrichissement des ressources terminologiques du domaine.

2.1 Représentation des connaissances

2.1.1 Définitions

Une **terminologie** présente l'ensemble des termes particuliers à une Science, un domaine ou un art (Larousse, 1988), à un groupe de personnes ou à un individu (Office de la langue française, 2000[1]).

« Le terme ontologie, issu de la philosophie de la connaissance, désigne généralement l'ensemble des concepts d'un domaine. » (Zweigenbaum, 1999). Il n'existe pas de consensus permettant de donner une définition plus précise d'une ontologie. Au centre des débats depuis l'avènement du Web Sémantique, la définition, la construction et l'utilisation d'ontologie font l'objet de nombreux travaux et on parle plus volontiers d'ontologies (au pluriel) afin de refléter les multiples facettes que recouvre cette apellation. Plusieurs auteurs, par exemple (Guarino, 1996) ou encore (Dameron, 2003) passent en revue les différentes définitions de la littérature afin d'examiner le type de représentation des connaissances dénoté par le terme ontologie. Pour notre part, nous retiendrons la définition d'une ontologie comme conceptualisation introduite par (Gruber, 1993) puis précisée de manière rigoureuse par (Bachimont, 2000) en ces termes :

[1] d'après le Grand Dictionnaire Terminologique (GDT) - `http://w3.granddictionnaire.com`

« Définir une ontologie pour la représentation des connaissances, c'est définir, pour un domaine et un problème donnés, la signature fonctionnelle et relationnelle d'un langage formel de représentation et la sémantique associée. »

2.1.2 Eléments de la représentation des connaissances

Terminologie et ontologie ont pour objet commun la représentation des connaissances relatives à un domaine - cependant, alors que l'ontologie manipule des concepts, la terminologie - au sens de Wüster (Wüster, 1981)- s'attache à une expression normalisée, figée de ces concepts, les termes[2]. Ainsi, se profilent les trois sommets du triangle aristotélicien (Figure 2.1) : les concepts, les choses, c'est-à-dire les objets du monde décrits de manière abstraite par les concepts et les signes permettant de désigner les concepts, les termes.

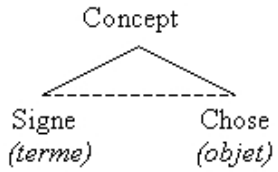

FIG. 2.1 – Le triangle aristotélicien

Cependant, le signe, au sens de Saussure n'est pas nécessairement linguistique et sert plutôt à établir une liaison entre un *signifié* (concept) et un *signifiant*, une image acoustique du signe. Parallèlement, Saussure évoque une fonction *référentielle* qui met le signe en rapport avec les objets de la pensée (par opposition aux choses, qui sont des objets du monde réel). Richards et Ogden illustrent cette conception à l'aide de la triade sémiotique (Figure 2.2) :

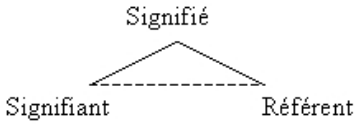

FIG. 2.2 – La triade sémiotique

Notons que les ontologies doivent également manipuler des signifiants puisqu'ils sont l'expression linguistique des concepts et qu'ils constituent, selon (Bachimont, 2000), les données empiriques caractérisant les connaissances à modéliser. Ainsi, la construction d'une ontologie passe par la construction de *primitives sémantiques* représentant les concepts du domaine. Ces primitives sémantiques peuvent être désignées par des « libellés », unités linguistiques semblables aux termes, dont l'interprétation doit être contrainte afin d'en décontextualiser la signification. Comme la Théorie Générale de la Terminologie (TGT), la méthode de construction d'ontologies proposée par Bachimont considère qu'une normalisation sémantique est nécessaire à la représentation des connaissances d'un domaine. Cependant, à l'introspection de l'expert

[2]également appelés « vedettes matières » ou « mots clés » en Science de l'Information. Pour une discussion plus approfondie sur la définition d'un terme, nous renvoyons le lecteur à (Rastier, 1995).

préconisée par Wüster, Bachimont préfère l'analyse de corpus par un expert chargé de décliner le sens des termes selon les contextes dans lesquels ils sont employés et l'intervention d'un ingénieur de la connaissance pour organiser et structurer ces informations.

2.1.3 Notions de domaine et de langue de spécialité

Dans les définitions que nous avons retenues, (cf. section 2.1.1) la représentation des connaissances à l'aide de terminologies ou d'ontologies s'entend dans le cadre d'un domaine particulier de la connaissance et implique un expert familier avec la langue de spécialité utilisée dans le domaine. On peut donc s'interroger sur les notions de domaine et de langue de spécialité. Un domaine est défini comme « l'ensemble des objets qu'embrasse un art, une science » (Larousse, 1988), ce qui pose d'emblée le problème de la délimitation du domaine et par extension des termes entrant dans la langue de spécialité relevant strictement du domaine. (Poibeau, 2005) fait état des discussion de cette question dans la littérature et s'interroge sur l'existence effective de langues de spécialité par opposition à la langue générale. Par une étude statistique distributionnelle sur corpus (Harris, 1991)[3] met en évidence des différences au niveau stylistique et lexical présente dans des corpus issus de différents domaines et la langue générale. Ces observations peuvent s'interpréter comme définissant des sous-langages, ce qui a pour conséquence pratique lors de l'étude d'une langue de spécialité de diminuer les contraintes inhérentes à l'élaboration d'une grammaire de la langue générale. (Poibeau, 2005) rappelle cependant que les langues de spécialité ne se définissent pas tant par une restriction du cadre linguistique général que par une utilisation particulière des structures linguistiques de la langue générale, en privilégiant certains éléments. Ainsi, tout comme il s'avérait difficile de poser les limites du domaine de spécialité, il est également délicat de déterminer ce qui relève de la langue générale et de la langue de spécialité. L'idée d'un *continuum* aussi bien linguistique que cognitif nous semble rendre compte de la situation de manière adéquate. Nous verrons à la section 2.4 que les différentes approches du traitement de la langue médicale reflètent cette situation dans la mesure où certains travaux reposent sur des spécificités de la langue médicale qui ne se retrouvent pas a priori dans d'autres domaines, alors que d'autres travaux sont applicables à (ou même issus de) la langue générale et utilisent simplement un paramètrage lié au domaine médical - par exemple le corpus de travail.

2.1.4 Relations sémantiques

La TGT et la construction d'ontologies s'intéressent également à l'étude des différentes relations qui peuvent exister entre les concepts du domaine de spécialité traité et donc entre les termes dénotant les concepts de ce domaine. En effet, afin de définir le sens d'une unité linguistique, F. Rastier (Rastier, Cavazza, & Abeillé, 1994) s'appuie notamment sur le paradigme différentiel. Ce principe est fondé sur la définition du sens d'une unité linguistique (ou d'un terme) par rapport à des unités partageant les même contextes, appelées unités voisines. Une unité peut alors être définie en fonction des différences et des ressemblances qu'elle entretient avec ses voisines. Cette approche est nommée « sémantique différentielle » et repose sur les travaux de (Pottier, 1964) (repris par la suite par (Rastier, 1995)) considérant le sens des unités linguistiques selon un ensemble de « sèmes différentiels ».

La théorie Sens-Texte (Mel'čuk, 1997) - notée TST ci-dessous- s'intéresse également au domaine de la sémantique et aux relations entre les unités linguistiques. La TST a pour

[3]Cité par (Poibeau, 2005)

objectif de décrire les mécanismes de correspondance entre le sens (ce que veut exprimer le locuteur et que nous pouvons assimiler à un ou plusieurs concepts éventuellement combinés) et le texte (l'expression du sens que le locuteur souhaite communiquer, que nous pouvons assimiler à un ou plusieurs termes -ou « mots formes », l'unité minimale selon Mel'cuk - généralement combinés en phrases). La TST repose sur la notion de paraphrase. On peut dire qu'elle suit le parcours onomasiologique[4] et postule notamment qu'à un sens peuvent correspondre plusieurs textes.

Ainsi, à un concept (sens élémentaire) peuvent correspondre plusieurs termes (textes) : on voit apparaître l'idée de synonymie. D'autres relations entre termes (telles que « instrument » ou « résultat ») sont encodées par une soixantaine de fonctions lexicales (FL), un formalisme destiné à modéliser les choix lexicaux syntagmatiques et paradigmatiques et à décrire la combinatoire lexicale d'une langue donnée. Les FL sont définies comme des fonctions mathématiques du type $f(x) = y$ où f est la fonction, x est l'argument auquel cette fonction peut être appliquée (un terme, ou une lexie) et y est la valeur, c'est-à-dire l'ensemble des lexies qui peuvent être reliées à x par la relation f. Par exemple, la FL S_{res} permet de définir la relation qui existe entre les termes « radiographie » et « radiographier » : une « radiographie » est le résultat obtenu par l'action de « radiographier » et on note : $S_{res}(radiographier) = radiographie$. Il faut remarquer que dans certains cas, deux mêmes lexies peuvent être liées par des fonctions lexicales différentes. Ainsi, « radiographie » provient également de la nominalisation du verbe « radiographier », modélisée par la FL S_0. Cependant, S_0 et S_{res} définissent des relations différentes sur le plan formel : S_0 apporte une information syntaxique (dérivationnelle) et S_{res} une information sémantique. Cette observation rappelle les liens étroits entretenus par la grammaire et le lexique au sein d'une même langue. Les fonctions lexicales représentent la section phraséologique dans les différents volumes du Dictionnaire Explicatif et Combinatoire du français (Mel'čuk et al., 1984, 1988, 1992, 1999). Le tableau 2.1 présente une illustration des relations prises en compte par les FL. Par ailleurs, l'application des FL à la terminologie et à l'élaboration de dictionnaires spécialisés fait l'objet de divers travaux (Mortchev-Bouveret, 2005), (Claveau & L'Homme, 2005), (Jousse & Bouveret, 2003) ou encore (L'Homme, 2002), (Grimes, 1990), (Frawley, 1988)[5] .

Au niveau de la phrase, la possibilité d'exprimer un sens par plusieurs textes se traduit par l'existence de *paraphrases*, que le locuteur peut utiliser pour exprimer le sens qu'il souhaite véhiculer. Ces paraphrases sont considérées comme quasi-synonymiques. Cependant, la TST prend en compte le fait que le choix du locuteur de l'une ou l'autre formulation est en partie guidé par son *intention communicative*.

Ainsi, bien que la sociolinguistique n'entre pas dans le cadre de la TST, contrairement à la TGT, elle ne considère pas les énoncés et les termes comme totalement neutres. Nous allons voir à la section suivante que la manière dont sont appréhendés les termes a une importance considérable dans le cadre des systèmes de représentation des connaissances.

2.1.5 Elaboration de systèmes de représentation des connaissances

L'émergence de la linguistique de corpus a amorcé une réflexion sur la TGT en replaçant le terme en contexte (Slodzian, 1995). Ainsi, l'une des critiques de (Cabré, 2002) à l'en-

[4]Par opposition à l'approche *sémasiologique* qui part du signe pour aller vers l'idée, l'approche onomasiologique a pour objet la verbalisation d'un concept en langue naturelle, c'est-à-dire la recherche de toutes les expressions permettant de le désigner.

[5]Ces trois derniers auteurs sont cités par (Mortchev-Bouveret, 2005)

Fonction Lexicale	Relation	Exemple
Syn	Synonymie	$Syn(céphalée) = mal\ de\ tête$
$Spec$	Hyponymie	$Spec(fracture\ du\ crâne)$ =fracture orbitaire
$Gener$	Hypéronymie	$Gener(fracture\ du\ crâne) = fracture$
$Anti$	Antonymie	$Anti(inhibiteur) = activateur$
S_0	Nominalisation	$S_0(radiographie) = radiographier$
V_0	Verbalisation	$V_0(soin) = soigner$
S_1	Agent typique	$S_1(document) = documentaliste$
S_{res}	Résultat	$S_{res}(radiographie) = radiographier$
S_{instr}	Instrument	$S_{instr}(endoscopie) = endoscope$

TAB. 2.1 – Exemples de relations modélisées par les Fonctions Lexicales

contre des travaux de Wüster porte sur le caractère, illusoire selon elle, de l'universalité d'une terminologie postulée par la TGT. Elle remarque en effet que « [à] travers ce processus d'uniformisation, la diversité dénominative et conceptuelle de la réalité est passée sous silence. » De son coté, (Holzem, 2000) souligne que la représentation des connaissances proposée par une terminologie est loin d'être universelle ou objective dans la mesure où elle apparaît dans un environnement social et culturel bien particulier[6]. Chaque terminologie reflète le contexte socio-économique et l'état de la connaissance dans le domaine lors de la création ou de la mise à jour. (Holzem, 2000) évoque par exemple le choix de « diversité biologique » comme vedette matière dans le système Rameau de préférence à « biodiversité », un terme plus compact et plus usité pour des raisons de similarité avec le terme anglo-américain « biologic diversity » en vigueur dans version originale (américaine) de la terminologie. On peut également remarquer que le mot clé <*agents anti-vih*> a été introduit dans le MeSH en 1997, car ce type de traitement n'existait pas auparavant. Des mots clés comme <*subvention gouvernementale USA, P.H.S.*>, <*subvention gouvernementale USA, Non-P.H.S.*> et <*subvention gouvernementale Non USA*> rappellent de manière très explicite le caractère américain de cette terminologie.

Un autre facteur déterminant du contenu d'une terminologie est l'usage auquel elle est destinée. Il en va de même pour les ontologies, selon (Bachimont, 2000), qui souligne que « une ontologie est dépendante non seulement du domaine mais aussi de la tâche visée ». En effet, les termes prennent tout leur sens dans une pratique langagière, c'est-à-dire dans le contexte où ils sont employés. Lors de la construction d'une ontologie, l'analyse du corpus de travail permet d'examiner les différents contextes et de choisir une dénomination des concepts adaptée à la résolution du problème posé. Dans le cadre d'une terminologie, la structure paradigmatique qui contient les termes constitue le contexte. En d'autres mots, la signification d'un terme est établie par la place qu'il occupe dans le réseau conceptuel du domaine. Le terme E10.5 de la CIM-10 en est un exemple typique : <diabète sucré insulino-dépendant, avec autre complication précise>. Pour savoir quelles sont les « autres » complications englobées dans ce terme, il faut connaître les complications du diabète insulino-dépendant énumérées dans les termes E10.0 à E10.4. Par ailleurs, en fonction de l'usage anticipé des termes, il est possible

[6]Ces aspects de la construction et de l'utilisation de terminologies sortent de notre domaine de compétence. Il nous semble important d'en évoquer les grandes lignes, mais pour plus de précision, nous invitons le lecteur à se reporter à (Holzem, 2000)

de les présenter de plusieurs manières différentes. Ainsi, le terme <amylose> qui est utilisé pour dénoter une maladie infectieuse peut également désigner la substance responsable de la maladie. La CIM-10, une terminologie orientée vers le recensement des maladies désigne donc la maladie amylose par le terme « amylose »et ne s'intéresse pas à la substance qui en est responsable. En revanche, le MeSH, une terminologie médicale dédiée à la recherche d'information en médecine choisit de désigner la substance responsable de l'amylose par le terme « amylose », alors que la maladie elle même sera dénotée par le terme « amyloïdose ».

Le domaine de la médecine en particulier a été l'un des précurseurs dans la définition des besoins en outils terminologiques, puis dans la construction de tels outils adaptés à ces différents usages (Zweigenbaum, 1999). Nous présentons dans la section suivante les besoins terminologiques en santé tel que l'encodage des dossiers patients, ou la recherche d'information ainsi que certaines terminologies y répondant.

2.2 Terminologie et représentation des connaissances en Médecine : quelques exemples (CIM-10, SNOMED, MeSH et CISMeF)

Il existe dans le domaine bio-médical un grand nombre d'ontologies et de terminologies adaptées aux besoins précis des différents acteurs. Ainsi, GALEN (Rector, Nowlan, & Kay, 1992) est une ontologie médicale généraliste, la GO (Gene Ontology[7]) est dédiée au domaine de la génétique, l'ontologie constituée dans le cadre du projet MENELAS (Zweigenbaum & Consortium-MENELAS, 1995) est ciblée sur les maladies coronariennes. En ce qui concerne les terminologies, la Classification Internationale des Maladies (CIM[8]) sera utilisée pour le codage médico-économique des dossiers patients à des fins statistiques et budgétaires. La Nomenclature Clinique SNOMED[9] est destinée à l'encodage médical plus fin des dossiers électroniques des patients. Le thésaurus MeSH[10] (Medical Subject Headings) a pour objet l'indexation des connaissances médicales et la recherche d'information dans les bases documentaires du domaine de la santé. Nous détaillons ces trois dernières terminologies, qui correspondent aux problématiques de recherche d'information abordées par nos travaux et sont actuellement utilisées au CHU de Rouen.

Les tableaux 2.2 et 2.3 présentent une vue synoptique des besoins et des caractéristiques des différentes terminologies illustrées par nos exemples.

Nous décrivons dans les sections suivantes chaque terminologie indépendamment des autres. Nous mettrons plus particulièrement l'accent sur le MeSH et la terminologie CISMeF, qui sont centrales dans nos travaux.

2.2.1 Classification statistique internationale des maladies et des problèmes de santé connexes : la CIM

La classification statistique internationale des maladies et des problèmes de santé connexes (OMS, 1993) a été publiée pour la première fois en 1933 par l'Organisation Mondiale de la

[7] cf. `http://www.geneontology.org`
[8] cf. `http://www.icd10.ch/index.asp`
[9] cf. `http://www.snomed.org/`
[10] cf. `http://www.nlm.nih.gov/mesh/meshhome.html`

Besoin	Terminologie
Description d'informations : connaissance médicale	MeSH
Caractérisation « orientée » de données : statistiques hôpitaux	CIM10
Caractérisation « ouverte » de données : dossier patient	SNOMED

Tab. 2.2 – Correspondance entre les besoins et les terminologies

Terminologie	Caractéristiques	Exemples
Nomenclature	exhaustivité structuration	SNOMED
Classification	structuration, lien nommés entre les termes	MeSH, CIM10, SNOMED
Thésaurus	normalisation des termes, réduction des ambiguités	MeSH, CIM10, SNOMED
Vocabulaire	définition des termes	MeSH, SNOMED

Tab. 2.3 – Les types de terminologies et leurs caractéristiques

.<liste des categories a trois caracteres>
.04 <maladies endocriniennes, nutritionelles et metaboliques>
.0401E10 <diabete sucre insulino-dependant>
E10.0 <diabete sucre insulino-dependant, avec coma>
E10.1 <diabete sucre insulino-dependant, avec acidocetose>
E10.2 <diabete sucre insulino-dependant, avec complications renales>
E10.3 <diabete sucre insulino-dependant, avec complication oculaire>
E10.4 <diabete sucre insulino-dependant, avec complication neurologique>
E10.5 <diabete sucre insulino-dependant, avec complication vasculaire périphérique>
E10.5 <diabete sucre insulino-dependant, avec autre complication précise>
E10.5 <diabete sucre insulino-dependant, sans complication>
.0401E11 <diabete sucre non insulino-dependant>
.0401E14 <diabete sucre, SAI>
.0401E23 <hyposecretion et autres anomalies de l'hypophyse>
.14 <maladies de l'appareil genito-urinaire>

FIG. 2.3 – Extrait du chapitre 4 de la CIM-10

Santé (OMS), dans le but d'indexer les causes de mortalité et morbidité pour des analyses statistiques et épidémiologiques. Elle permet le recueil de diagnostics à des fins de santé publique ou d'évaluation de l'activité hospitalière. C'est la $10^{\text{ème}}$ version de la CIM (révisée en 1993) qui est actuellement utilisée en France (alors que c'est la $9^{\text{ème}}$ version qui est encore utilisée au Canada et aux Etats-Unis). La CIM-10 est une classification mono-axiale divisée en 21 chapitres permettant de classer les concepts selon le siège anatomique des maladies. La figure 2.3 présente un extrait du chapitre 4 consacré aux maladies endocriniennes, nutritionelles et métaboliques.

Les concepts sont désignés par des termes qui relèvent plus d'un métalangage (Zweigenbaum, 1999) que d'expressions de la langue naturelle. En particulier, l'emploi des parenthèses ou de sigles tels que SAI (Sans Autre Indication) ou NCA (Non Classé Ailleurs) sont caractéristiques du métalangage. La CIM est une hiérarchie à 5 niveaux de profondeur, indiquant une spécialisation des concepts. Elle comporte plus de 18 000 codes et environ 50 000 termes. Par ailleurs, des extensions de codes de la CIM-10 à usage national ont été créées afin de permettre le repérage de certains types de prise en charge et de compléter le codage par des informations à visée purement documentaire, à la demande de certaines sociétés savantes.

2.2.2 La Nomenclature systématique des médecines humaine et vétérinaire : la SNOMED

La Nomenclature systématique des médecines humaine et vétérinaire (Coté & Al., 1997) a été conçue en 1965 pour décrire les lésions anatomopathologiques et radiologiques et a ensuite été étendue à toute la médecine. Introduite en septembre 1993, la SNOMED Internationale ou version III est une nomenclature systématique multiaxiale (elle comporte onze axes orthogonaux), c'est-à-dire qu'elle permet la combinaison de termes provenant d'axes différents. Dans chaque axe, les concepts sont représentés par une série de termes au sein de laquelle on peut distinguer une formulation préférée et des synonymes de diverses natures syntaxiques.

Par ailleurs, chaque axe est hiérarchisé en fonction de la spécialisation des concepts, qui sont reliés par des relations d'hyponymie et de méronymie. Notons qu'il existe aussi des relations transversales plus complexes (entre concepts appartenant à des axes différents). La SNOMED contient 109 023 concepts, désignés par 164 180 termes. La traduction de la SNOMED en français est réalisée par l'équipe du Centre de recherche en diagnostic médical informatisé (CRDMI) et devrait être disponible fin 2005. Autour de cette terminologie, on peut mentionner le projet SNOMED RT qui est un remaniement de la nomenclature SNOMED III visant à épauler ses termes par des descriptions dans un langage de représentation des connaissances. Récemment, la fusion de la SNOMED-RT avec une terminologie britannique, Clinical Terms a donné naissance à la SNOMED Clinical Terms (SNOMED CT), une terminologie des soins de santé cliniques dynamique et validée scientifiquement. Son objectif est de rendre les connaissances de soins de santé plus accessible par toutes les spécialités médicales. La terminologie Core SNOMED CT contient plus de 361 800 concepts de soins de santé et comporte également 975 000 descriptions et près de 1,47 million de relations sémantiques.

2.2.3 Medical Subject Headings : le thésaurus MeSH

Historique, motivation, utilisation

Une première liste officielle de descripteurs a été publiée par la National Library of Medicine (NLM) américaine en 1954 et regroupait des termes issus de plusieurs listes de termes médicaux existantes. Une version revue et corrigée publiée en 1960 constitue le premier thésaurus MeSH. Elle contient alors 4 400 mots clés, contre plus de 22 000 pour la version 2005. Ainsi, le MeSH a été conçu comme un vocabulaire contrôlé dynamique, créé et mis à jour dans le but d'indexer les articles scientifiques des principales revues médicales internationales. La base de données bibliographique MEDLINE[11] regroupe aujourd'hui plus de 10 millions d'articles (en anglais) indexés depuis 1960 à l'aide des descripteurs MeSH par les indexeurs professionnels de la NLM. Le projet CISMeF permet d'étendre ce travail aux ressources francophones disponibles gratuitement sur l'internet à destination des professionnels de la santé, des étudiants en médecine et pharmacie et du grand public. L'INSERM (Institut National de la Santé Et de la Recherche Médicale) a élaboré une version française du MeSH[12], issue de la traduction du MeSH américain. Cependant, il s'agit d'une traduction partielle (l'ensemble des mots clés et des qualificatifs sont traduits, mais certains synonymes ne le sont pas encore). La version française du MeSH comporte 22 995 mots clés, 83 qualificatifs et environ 3 000 synonymes. C'est cette version qui est utilisée par l'équipe CISMeF pour indexer les ressources recensées par le catalogue.

Description « technique »

Le MeSH compte 22 995 mots clés, 83 qualificatifs et environ 57 000 synonymes dans sa version 2005. Les synonymes, ou « entry terms », correspondent à des formulations alternatives des mot clés (par exemple, « cancer sein » est un synonyme de <*tumeur sein*>, « diabète insulinodépendant » est un synonyme de <*diabète de type ii*>).

Certains mots clés, appelés *descripteurs obligatoires*, reflètent la vocation du MeSH à indexer des documents du domaine biomédical. En effet, les descripteurs obligatoires sont des

[11] Accessible grâce au moteur de recherche PubMed sur `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?`

[12] cf. `http://dicdoc.kb.inserm.fr:2010/basismesh/mesh.html`

mots clés qui doivent être obligatoirement être utilisés lors de l'indexation d'un document, si les concepts auxquels ils renvoient apparaissent dans le document. Il s'agit des mots clés correspondant aux classes d'age des patients (pour la médecine humaine) et aux conditions d'expérimentation en laboratoire (pour la médecine vétérinaire).

L'ensemble formé par un mot clé et ses synonymes décrit un concept du domaine biomédical et le mot clé correspond à la « formulation préférée » désignant ce concept. Les qualificatifs sont des termes qui peuvent être associés à un mot clé afin d'en préciser le sens. On utilise le symbole « / » pour séparer mot clé et qualificatif, selon la notation <*mot clé/qualificatif*> et on parle de « paire mot clé qualificatif ». Comme le soulignent (Malaisé, Zweigenbaum, & Bachimont, 2003), cette représentation des concepts par association de deux descripteurs (un mot clé et un qualificatif) est porteuse d'une sémantique plus forte que ce que permet l'utilisation des opérateurs booléens classiques car elle permet de préciser la nature de l'association - la représentation ainsi obtenue se rapproche du niveau de description plus élevé qui caractérise les ontologies.

Un qualificatif ne peut être utilisé seul - cependant, certains termes peuvent être à la fois mot clé et qualificatif comme par exemple, <*thérapeutique*>. Il peut donc être utilisé seul en tant que mot clé, ou associé à un autre mot clé en tant que qualificatif. Par exemple, la paire <*diabète de type ii/thérapeutique*> évoque le traitement du diabète non insulinodépendant, alors que le mot clé <*diabète de type ii*> seul fait référence à la maladie en général. Cependant, toutes les associations mot clé/qualificatif ne sont pas possibles. A chaque mot clé correspond la liste des qualificatifs qui peuvent lui être associés. Ainsi, le qualificatif <*/thérapeutique*> peut être associé au mot clé <*diabète de type ii*>, mais pas au mot clé <*sujet âgé*>. De même, le mot clé <*diabète de type ii*> peut être précisé par le qualificatif <*/thérapeutique*> mais pas par <*/transplantation*> etc.

Les qualificatifs sont organisés hiérarchiquement du plus générique au plus précis[13] (par exemple, <*chimiothérapie*> est un fils de <*thérapeutique*>). Il en est de même pour les mots clés qui sont répartis dans quatorze arborescences thématiques auxquelles correspond un code spécifique : A pour « anatomie » B pour « organismes », C pour « maladie » etc. A chaque mot clé est attribué un code en fonction de l'arborescence dans laquelle il se trouve (par exemple, le mot clé <*main*> a pour code A01.378.800.667 dans l'arborescence « anatomie »). Dans ces arborescences, la relation père-fils correspond à une particularisation des concepts :

1. relation « est une partie de » (méronymie)
   A01.378.800.667 <*main*> → A01.378.800.667.430 <*doigt*>
2. relation « est un type de » (hyponymie)
   A01.378.800.667.430 <*doigt*> → A01.378.800.667.430.705 <*pouce*>
3. relation « est sémantiquement proche de » (aboutness) - révélatrice de l'orientation du thésaurus vers la recherche d'information.
   G03.850.110 <*accidents*> → G03.850.110.060.075 <*sécurité*>

La figure 2.4 présente un extrait de l'arborescence C, avec 8 niveaux de profondeur. Les termes les plus hauts dans les arborescences (par exemple, <*maladies*> ou <*tumeurs*>) correspondent à des concepts généraux et les termes situés au niveau des feuilles correspondent à des concepts spécifiques (par exemple, <*Wolfram, syndrome*>).

Il existe jusqu'à 11 niveaux de profondeur. Les codes attribués aux mots clés reflètent leur position dans les arborescences. Plus un mot clé est situé en profondeur, plus son code

[13] cf. `http://www.chu-rouen.fr/cisme/arboqualificatifs.htm`

C <maladies>
C04 <tumeurs>
C18 <métabolisme et nutrition, maladies>
C18.452 <métabolisme, maladies>
C18.452.090 <amyloïdose>
C18.452.394 <troubles du métabolisme glucidique>
C18.452.394.750 <diabète>
C18.452.394.750.124 <diabète de type 1>
C18.452.394.750.124.960 <Wolfram, syndrome>
C18.452.394.750.149 <diabète de type 2>
C18.452.394.750.774 <état prédiabétique>
C18.452.394.937 <glycosurie>
C18.654 <troubles nutrition>
C23 <troubles liés environnement>

FIG. 2.4 – Extrait de l'arborescence C du MeSH

est long. Il faut remarquer qu'un même terme est susceptible d'appartenir à plusieurs arborescences. Dans ce cas, il comporte un code par arborescence le contenant. Par exemple, le mot clé <*privation de sommeil*> appartient entre autres aux arborescences « maladie » et « troubles mentaux » avec les codes C23.888.592.796.772 et F03.870.400.099 respectivement[14] . Afin d'éviter toute confusion, les mot clés possèdent également un identifiant unique indépendant de leur position dans les diverses arborescences. Par exemple, l'identifiant unique de <*privation de sommeil*> est D012892. Afin d'accéder facilement à toutes ces informations, la NLM a créé le « MeSH browser » pour le MeSH américain, qui permet de naviguer dans les hiérarchies et d'accéder à des pages récapitulatives sur chaque terme (mot clé ou qualificatif) indiquant les hiérarchies auxquelles le terme appartient, la liste des qualificatifs (respectivement, mot clés) auxquels il peut être associé, la liste des synonymes, sa définition et un historique dans la terminologie : date de création, anciennes appellations[15] etc. L'outil de navigation terminologique bilingue de CISMeF fournit le même type d'information pour le MeSH en français (et anglais), ainsi que pour les autres éléments de la terminologie CISMeF décrits à la section 2.3.

Description « linguistique »

D'un point de vue linguistique, les termes MeSH peuvent être classés en trois catégories. On recense :

1. des noms simples
   <*diabète*>, <*main*>, <*grossesse*>
2. des noms composés (syntagmes nominaux libres et noms composés généralisés)
   <*bain de bouche*>, <*acide acétique*>, <*avant bras*>

[14] on peut remarquer que la profondeur du terme peut varier d'une arborescence à l'autre : exemple pour <*privation de sommeil*> : profondeur 4 dans C23 et 3 dans F03.

[15] par exemple, dans le MeSH 2004 la « formulation préférée » du mot clé D003924 était <*diabète non insulinodépendant*>, « diabète de type ii » étant un synonyme. Dans le MeSH 2005, <*diabète de type ii*> est devenu la « formulation préférée » du concept et « diabète non insulinodépendant » son synonyme.

<déchets dangereux>, *<hémodialyse à domicile>*, *<prévention et contrôle>*

3. des expressions typiques du langage documentaire - bien que ces termes ne soient pas des syntagmes nominaux grammaticalement corrects, ils dérivent généralement de noms composés. Nous indiquons les noms composés grammaticaux entre guillemets pour chaque exemple.
   <Alzheimer, maladie>/ « maladie d'Alzheimer »
   <enfant âge pre-scolaire>/ « enfant d'âge pré-scolaire »
   <moule (zoologie)>/ « moule au sens de la zoologie et par opposition à l'objet du même nom »

La notion de nom composé et plus généralement de mot composé est très controversée. Pour plus de précisions, nous invitons le lecteur à se reporter à la discussion de (Savary, 2000) sur ce point. Le problème de la distinction entre nom composé et syntagme nominal libre y est également abordé. Par ailleurs, (Silberztein, 1994) propose quatre critères permettant de distinguer les groupes nominaux libres des noms composés généralisés. Dans le contexte du développement de dictionnaires électroniques du français, l'auteur justifie la lexicalisation[16] de tous les noms composés, qui figurent alors dans les dictionnaires - c'est-à-dire qu'une entrée spécifique sera créée pour inclure un nom composé. Dans le cadre plus particulier de notre application (indexation automatique à l'aide d'un vocabulaire contrôlé), ce choix est d'autant plus pertinent que d'une part, nous cherchons à associer des groupes nominaux à des clés d'index particulières (les termes MeSH), ce qui nécessite un recensement de tous les groupes nominaux concernés et d'autre part, ces groupes nominaux ont justement été choisis comme des mots clés d'un vocabulaire contrôlé parce qu'ils reflètent une institutionnalisation de l'usage dans le domaine médical. Il ne s'agit donc jamais de groupes nominaux libres en tant que tels. Par exemple, le mot clé *<soins longue durée>* en principe exprimé par le syntagme nominal « soins de longue durée » est fréquemment employé tel quel dans les rapports officiels.

Pour bien rendre compte de la diversité des termes à traiter, il faut également remarquer que, bien que la plupart des termes MeSH soient composés de 1 à 3 lemmes, certains peuvent être plus longs (par exemple, *<complexe protéique adaptateur, sous-unité alpha>* comporte 5 lemmes) Nous indiquons dans le tableau 2.4 la répartition des mots clés MeSH en fonction de leur longueur en lemmes :

Ainsi, avec près de 22 000 termes destinés à l'indexation des connaissances en médecine, le MeSH permet une bonne couverture de l'ensemble des domaines de la santé, mais se révèle cependant moins précis que la CIM-10 ou la SNOMED qui ciblent plus particulièrement la description des pathologies et comportent un nombre de termes dans ce domaine supérieur au MeSH. La section suivante présente un cas d'utilisation concrète du MeSH dans un catalogue de santé et les ajustements terminologiques apportés par CISMeF afin d'optimiser l'indexation et l'accès aux connaissances médicales.

[16]Nous entendons ici par « lexicalisation » le *figement lexical* des mots composés - ainsi, les mots composés sont considérés comme des termes, ce qui justifie leur inclusion dans le dictionnaire.

Longueur en lemmes	Nb. de mots clés	Nb. mots clés (%)
1	9 157	40,83
2	8 504	37,92
3	3 129	13,95
4	1 030	4,59
5	344	1,53
6	149	0,66
7	74	0,33
8	24	0,11
9	10	0,04
10	5	0,02
11	2	0,01
Total	**22 430**	**100**

TAB. 2.4 – Répartition des mots clés MeSH en fonction du nombre de lemmes (données MeSH 2004)

2.3 La terminologie CISMeF : une terminologie fondée sur le MeSH

2.3.1 Utilisation du MeSH par CISMeF

Pour l'indexation des documents près de 11 000 mots clés sont utilisés dans les notices CISMeF[17] sur 22 995 (+ 83 qualificatifs, ainsi que l'ensemble des synonymes disponibles en français). Ces 11 000 mots clés ne correspondent pas à un sous ensemble des arborescences MeSH. Ils sont issus de l'ensemble du MeSH, en fonction des besoins de l'indexation - du fait de la politique éditoriale du catalogue, l'accent est mis sur les maladies/pathologies (arborescences C et F03) tandis que les organismes (arborescence B) sont assez peu exploités, comme l'illustre la figure 2.5 ci-dessous. Dans le cadre de VUMeF (Darmoni, Jarousse, et al., 2003), l'équipe CISMeF a créé 2 236 synonymes de termes MeSH et traduit 4 316 définitions de mots clés.

CISMeF propose un accès dynamique aux ressources indexées grâce au moteur de recherche Doc'CISMeF[18] (Thirion et al., 2000).

Cependant, les ressources en ligne susceptibles d'être indexées dans le catalogue sont très hétérogènes, tant au niveau du format, du contenu ou du public visé. Il peut s'avérer difficile de caractériser ce type d'information à l'aide de termes MeSH. Par ailleurs, les termes du thesaurus MeSH sont souvent très spécifiques et il s'avère difficile d'avoir une idée globale de la spécialité médicale concernée par une ressource indexée avec des termes aussi fins. Ces observations ont conduit l'équipe CISMeF a créer des termes plus génériques afin de caractériser les ressources. Deux nouveaux concepts répondant à ce besoin ont été créés, les types de ressource et les métatermes. Nous présentons ci-dessous ces ajouts à la terminologie MeSH, détaillés dans (Douyère et al., 2004).

[17] Le contenu et l'élaboration des notices sont détaillés en 1.2.3

[18] Accessible sur `http://doccismef.chu-rouen.fr`

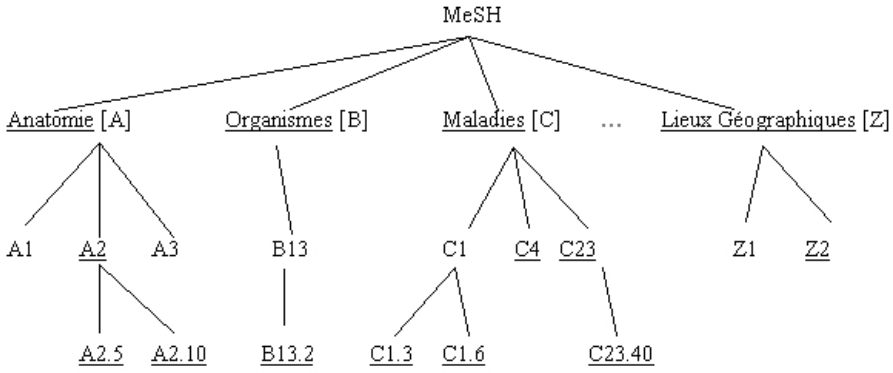

FIG. 2.5 – Le thésaurus MeSH : les termes soulignés sont utilisés par CISMeF

2.3.2 Les types de ressource

Définition

Les types de ressources CISMeF sont une généralisation des « publication types » de Medline - Des types de ressources spécifiques aux ressources en ligne ont été ajoutés, comme par exemple <*association*>, <*réseaux coordonnés*> ou encore <*questions à choix multiple*>. Comme les mots clés et les qualificatifs, les types de ressources sont organisés hiérarchiquement[19]. Par exemple, le type de ressource <*association patients*> est un fils de <*association*>. Les 263 types de ressources CISMeF ainsi que leur définitions sont disponibles sur `http://www.chu-rouen.fr/documed/typeressource.html`. Un type de ressource a pour vocation de décrire la nature de la ressource, tandis que les mot clés (ou paires mot clé/qualificatif) MeSH décrivent son contenu. Par exemple, un cours sur le traitement du cancer du sein sera indexé à l'aide de la paire MeSH <*tumeurs du sein/thérapeutique*> et du type de ressource <*cours*>.

Notion de triplet

Les types de ressources peuvent être utilisés seuls afin de décrire la nature d'une ressource. Cependant, dans certains cas, il est également possible d'affilier un type de ressource à un mot clé ou à une paire mot clé/qualificatif afin de décrire le contenu d'une ressource. On utilise alors le symbole « \ » pour séparer une paire mot clé/qualificatif et un type de ressource, selon la notation <*mot clé/qualificatif\ type de ressource*> et on parle de « triplet ».

Ex : <*pied\radiographie*>, <*tumeurs du sein/diagnostique\question choix multiples*>

L'affiliation de types de ressources aux mots clés (ou paires) MeSH permet une description plus fine des ressources avec un contenu dense et varié telles que les <*cours*> ou les <*recommandations pour la bonne pratique*>. Pour bien comprendre le rôle des triplets, il faut distinguer :

- L'utilisation d'un terme en tant que type de ressource et en tant que qualificatif : certains types de ressources peuvent également être des qualificatifs, comme par exemple <*/\ radiographie*>. Dans ce cas, il faut bien distinguer leur usage en tant que qualificatif et

[19] Remarquons que ce n'est pas le cas des types de publication de Medline.

en tant que type de ressource. Un qualificatif décrira le contenu textuel de la ressource alors qu'un type de ressource décrira le support même de l'information. Par exemple, l'association du mot clé <*pied*> avec le qualificatif <*/radiographie*> désigne un texte ou une explication sur la radiographie du pied. En revanche, l'association du mot clé <*pied*> avec le type de ressource < *radiographie*> indique que la ressource, par exemple un cours d'anatomie, contient (entre autres informations) une radiographie représentant effectivement un pied.

– L'utilisation d'un type de ressource seul ou affilié. Un type de ressource utilisé seul a une portée globale sur la ressource, alors qu'un type de ressource affilié à un mot clé ou à une paire a une portée locale à l'intérieur de la ressource. Il s'utilise lorsqu'une ressource présente des informations sur plusieurs supports distincts. Ainsi, dans le cas où la ressource considérée est une radiographie du pied, on utilisera le type de ressource < *radiographie*> seul pour décrire la nature de la ressource et le mot clé <*pied*> seul pour décrire le contenu de la ressource. Par contre, si la ressource considérée contient une radiographie comme illustration, par exemple dans le cadre d'un cours d'anatomie, on utilisera le type de ressource < *radiographie*> affilié au mot clé <*pied*> ainsi que le mot clé <*anatomie*> pour décrire le contenu de la ressource et on utilisera le type de ressource <*cours*> seul pour décrire la nature de la ressource.

Sans la possibilité d'affilier un type de ressource à un mot clé (ou une paire) MeSH, il n'est pas possible de connaître la portée des types de ressources, ni de savoir à quelle thématique précise se rapportent les types de ressources à portée locale, notamment les images. Par exemple, on ne peut distinguer une ressource décrivant l'anatomie du pied et du poignet et présentant une radiographie du pied et une ressource décrivant l'anatomie du pied et du poignet et présentant une radiographie du poignet. Lors de la recherche d'information, l'utilisation de triplets permet de sélectionner les ressources contenant une radiographie du pied parmi l'ensemble des ressources contenant des radiographies.

2.3.3 Les métatermes

Les métatermes ont été sélectionnés par un expert du domaine bio-médical et correspondent généralement à des spécialités médicales ou à des domaines de la biologie (par exemple, <*cancérologie*> ou <*bactériologie*>). Chacun des 105 métatermes CISMeF est relié à un ou plusieurs termes MeSH (mot clé ou qualificatif) et type de ressource par des liens sémantiques. Par exemple, le métaterme <*cancérologie*> est relié de cette manière aux mots clés MeSH <*cancérologie*> et <*tumeurs*> qui appartiennent à des arborescences MeSH différentes (G et C respectivement). <*cancérologie*> est également relié au type de ressource <*service oncologie hôpital*>. Ainsi, Les métatermes permettent de palier à la finesse de certains descripteurs MeSH en les rassemblant sous une même thématique, ce qui a un impact positif sur la recherche d'information dans le catalogue. Par exemple, une requête sur les « recommandations en cancérologie » obtient seulement 16 réponses si <*cancérologie*> est considéré comme un mot clé MeSH, alors qu'elle obtient 320 réponses si <*cancérologie*> est considéré comme un métaterme[20]. L'utilisation du métaterme permet d'obtenir de meilleurs résultats car, lors de la recherche, au lieu de prendre en compte les seuls mots clés MeSH de l'arborescence G02.403.776.409.515 <*cancérologie*>, la recherche s'effectue également sur les arborescences de tous les termes auxquels le métaterme est sémantiquement relié. Dans le cas

[20]Résultat des recherches booléennes effectuées sur Doc'CISMeF le 24/02/05 : « guidelines.tr ET cancérologie.mc » vs. « guidelines.tr ET cancérologie.mt »

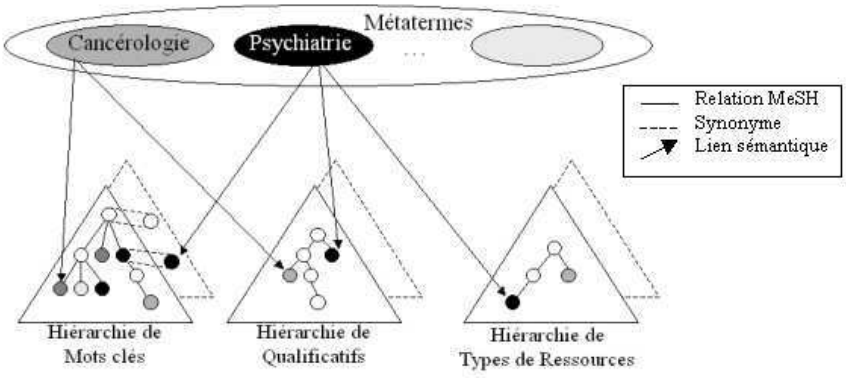

FIG. 2.6 – Structure de la terminologie CISMeF

de <*cancérologie*>, la recherche est par exemple étendue aux arborescences C04 <*tumeurs*> et N02.278.421.556.070 <*services soins cancérologie*>. La liste complète des métatermes est disponible sur `http://www.chu-rouen.fr/ssf/santspe.html`. Il faut remarquer qu'un terme peut être à la fois un mot clé, un qualificatif et un métaterme (par exemple, <*thérapeutique*>). La figure 2.6 inspirée de (L. F. Soualmia, Barry-Greboval, Abdulrab, & Darmoni, 2002) illustre les liens sémantiques qui existent entre les différents éléments de la terminologie CISMeF.

Nous avons rappelé ci-dessus les aspects essentiels de la représentation des connaissances en général et décrit quelques exemples de représentation des connaissances en médecine au travers de quatre terminologies (CIM-10, SNOMED, MeSH et CISMeF). Du fait de l'évolution de la langue médicale d'une part et des besoins terminologiques d'autre part, ces représentations sont l'objet de multiples travaux dans la communauté scientifique. Dans les sections suivantes, nous présentons tout d'abord les grandes lignes des travaux en cours dans le traitement de la langue médicale et nous exposons ensuite notre propre contribution.

2.4 Traitement automatique de la langue médicale

De part la variété et le volume importants des ressources terminologiques développées dans le domaine de la médecine dès la fin du XIX$^{\text{ème}}$ siècle, de nombreux chercheurs ont abordé le traitement automatique de la langue médicale à partir de la deuxième moitié du XX$^{\text{ème}}$ siècle. Nous ne souhaitons pas effectuer ici un état de l'art exhaustif de l'ensemble de ces travaux. Nous présentons simplement les grandes lignes abordées dans le cadre du traitement automatique de la langue médicale et en particulier les travaux touchant à la recherche d'information en médecine.

2.4.1 Enjeux du traitement automatique de la langue médicale

Parmi les applications couramment utilisées dans le domaine de la médecine qui nécessitent une connaissance et une manipulation spécifique de la langue médicale, on dénombre :

- L'indexation
- La classification
- La recherche d'information

Dans ces applications, il s'agit le plus souvent de mettre en correspondance un texte libre (requête utilisateur, contenu d'un document, ...) avec les termes d'un vocabulaire contrôlé (MeSH, CIM-10 etc.). Dans certains cas, la mise en correspondance entre le vocabulaire contrôlé et le texte libre peut également s'avérer utile : par exemple, la re-formulation d'une requête utilisateur en langue naturelle à des fins de désambiguation. Il est également souhaitable que différents systèmes automatiques puissent être inter-opérables. En effet, il peut être intéressant d'effectuer une recherche d'information contextuelle (c'est-à-dire une requête en termes MeSH) à partir d'un dossier patient (codé à l'aide de termes CIM10).

D'une manière générale on peut résumer ces problématiques en observant que le traitement automatique de la langue médicale doit faire face à des problèmes de traduction (qui peuvent d'ailleurs être aussi bien monolingues que plurilingues...) de deux types :

1. Traduction texte libre ↔ vocabulaire contrôlé
2. Traduction vocabulaire contrôlé 1 ↔ vocabulaire contrôlé 2

La disponibilité de systèmes de représentation des connaissances médicales structurées et la mise en correspondance des différentes ressources existantes sont des points clés pour la résolution de ces problèmes.

2.4.2 Ressources linguistiques dans le domaine médical

Le Métathésaurus UMLS (Unified Medical Language System) décrit par (Mc Cray, 1989) apporte des éléments de solution. En effet, il réunit plus de 95 terminologies du domaine médical (dont la CIM-10, la et le MeSH). L'UMLS a pour vocation d'établir des équivalences entre termes issus de chaque représentation du domaine médical. Il s'agit d'un ensemble de « concepts » du domaine médical auxquels sont reliés les termes de toutes les terminologies englobées par l'UMLS. Ce système permet par exemple d'obtenir des équivalences entre des termes issus des diverses terminologies, dans la mesure où ils sont reliés aux mêmes concepts :

(MeSH) <achondroplasie> ↔ (CIM10) <achondroplasie>
(MeSH)<acné> ↔ (CIM10) <acné vulgaire>
(MeSH) <urgence abdominale> ↔ (CIM10) <syndrome abdominal aigu>
Pour ce qui est du traitement de la langue médicale en anglais, l'UMLS constitue donc un outil informatique puissant, permettant d'accéder à de nombreuses informations concernant les variations des termes et les relations qu'ils entretiennent entre eux. Cependant, ce métathésaurus concerne essentiellement des terminologies anglophones, dont l'équivalent dans d'autres langues n'est que partiellement ou pas du tout disponible. Pour le français, plusieurs projets se proposent de fournir de telles ressources aux acteurs de l'informatique médicale francophone :

– S'appuyant sur la dynamique et les acquis de l'UMLS, le projet VUMeF (Darmoni, Jarousse, et al., 2003) a pour objectif d'augmenter la part du français dans l'UMLS, afin de consolider les ressources terminologiques francophones du domaine médical. Ce projet s'intéresse à l'amélioration des traductions existantes pour certaines terminologies - en particulier le MeSH - à la traduction de nouvelles terminologie - en particulier la SNOMED - ainsi qu'à l'intégration de vocabulaires spécifiquement français comme le Catalogue Commun des Actes Médicaux (CCAM). Par ailleurs, le consortium VUMeF a également pour objectif de fournir des outils et des méthodes permettant de mettre en correspondance expression libres et vocabulaires contrôlés, ainsi que l'évaluation de ces

tâches. Par suite, le projet prévoit la mise en oeuvre des ressources ainsi développées pour l'aide au codage dans les dossiers patients et l'indexation automatique de sites Web. Nous reprendrons ces points en détail au chapitre 5.

– Mené en parallèle, le projet UMLF (Zweigenbaum et al., 2003) se donne pour tâche d'effectuer la collecte, la synthèse, la complétion et la validation de ressources lexicales pour le français médical. Par une approche monolingue, il vise à produire un lexique contenant les variantes flexionnelles et dérivationnelles des mots du domaine. Ces informations doivent être encodées dans un format informatique standard afin de favoriser leur intégration dans des systèmes de traitement automatique de la langue médicale.

2.4.3 Travaux sur le traitement de la langue médicale

Dans l'attente de la disponibilité de telles ressources, ou dans le cadre même de leur élaboration, les chercheurs ont pu proposer des solutions pour résoudre les problèmes posés par le traitement de la langue médicale. Plus spécifiquement, quels sont les problèmes rencontrés ?

- Construction de terminologies/ontologies : (Bourrigault, 1994), (Bourigault & Fabre, 2000)
- Traduction d'éléments de terminologies (Claveau & Zweigenbaum, 2005)
- Variation lexicale (Grabar, 2004)
  - caractères (casse, orthographe, accentuation)
    Diabète insulino-dépendant/ diabete insulinodependant
  - ordre des mots
    affectif, symptôme / symptôme affectif
  - absence de mots vides
    accident travail / accident du travail
  - insertion et suppression d'éléments dans les termes (Jacquemin, 1997)
    carence alimentaire en vitamine A/ carence en vitamine A
  - correction orthographique pour améliorer les résultats en Recherche d'Information et en Indexation Automatique (Ruch, 2002)
    gripe / grippe
- Variation morphologique : (Grabar, 2004) et (Namer, 2005)
  massage cardiaque / massage du coeur
- Identification de relations entre termes et expressions (Claveau & L'Homme, 2005)
  - Traitement de la métonymie (Bouaud, Bachimond, & Zweigenbaum, 1996)
    le fond de l'oeil confirme le diagnostic / l'examen du fond d'oeil confirme le diagnostic
  - Traitement des périphrases (Namer, 2005)
    otite/ inflammation de l'oreille

Parmi ces travaux, certains partent de l'étude de la langue générale, sans s'attacher à un domaine de la connaissance particulier et trouvent leur application dans le domaine médical - par exemple (Bourrigault, 1994), (Bachimont, 2000). Cependant, d'autres approches, bien que théoriquement généralisables, exploitent les spécificités (morphologie, régularité des constructions...) du domaine médical - par exemple, (Namer, 2005) (Claveau & Zweigenbaum, 2005). Ainsi, (Sager & Friedman, 1987) présentent également des travaux centrés sur la langue médicale.

2.5 Notre contribution

En regard des grandes lignes du traitement de la langue médicale évoquées ci-dessus, nos travaux se positionnent dans le cadre général de la recherche d'information (et plus précisément de l'indexation). Notre contribution touche plus précisément à la construction de ressources terminologiques (un dictionnaire électronique MeSH) et au traitement de certains types de variations lexicales. Dans le cadre de la construction du dictionnaire, nous avons en effet été confrontée aux variations entre termes et expressions de la langue naturelle et nous avons choisi d'étendre notre contribution à la traduction automatique de synonymes MeSH afin d'enrichir les ressources terminologiques existant pour le français.

2.5.1 Construction d'un dictionnaire électronique MeSH

Dans le cadre du développement du système d'indexation automatique MAIF (décrit à la section 5.8), nous avons été amenée à formaliser et enrichir les ressources terminologiques de santé à notre disposition afin de les rendre exploitables par le système d'indexation. Celui-ci intègre certaines fonctionnalités d'INTEX (Silberztein, 1993), utilisées dans la partie d'analyse de la ressource à indexer. En effet, la première étape du processus d'indexation consiste en une analyse de surface qui permet de localiser des éléments textuels se rapportant à des mots clés ou à des paires de descripteurs MeSH. Ainsi, nous avons pu construire des dictionnaires électroniques MeSH au format DELA, comprenant notamment 4065 « maladies »et une bibliothèque de graphes permettant notamment de reconnaître les descripteurs obligatoires relatifs aux groupes d'âge (Névéol, Douyère, Rogozan, & Darmoni, 2004).

Contenu des dictionnaires

Dictionnaire (Robert, 2004) : Recueil d'unités signifiantes de la langue (mots, termes, éléments) rangées dans un ordre convenu, qui donne des définitions, des informations sur les signes.

Dans le cadre de l'indexation automatique, les dictionnaires « MeSH » doivent tout d'abord nous permettre de repérer les termes MeSH sous les diverses formes qu'il peuvent prendre en langue naturelle. Les mots clés MeSH sont les « unités signifiantes » que nous allons recenser et les liens entre les « formes MeSH » et les mots clés font partie des informations que nous allons inclure dans le dictionnaire. Après ce repérage de « surface », nous souhaitons mettre en oeuvre une analyse plus fine des concepts sous-jacents. Ainsi, les dictionnaires doivent également recenser des informations plus précises sur les termes rencontrés, telles que la nature exacte (est-ce un mot clé ou un qualificatif ?) ou le type de concept désigné (est-ce une maladie, un organe, un vaccin etc.). Nous présentons ci-dessous la manière dont toutes ces informations ont été encodées dans les dictionnaires élaborés.

Un format « à la DELA »

Originellement développés aux Laboratoire d'Analyse et de Description Linguistique (LADL) par B. Courtois et M. Silberztein, les Dictionnaires Electroniques du LADL (DELA) sont destinés à être utilisés par des programmes informatiques de traitement de la langue naturelle. Il existe plusieurs formes de DELA[21] - les DELAS (pour les mots simples et DELAC pour

[21] cf. (Courtois, 1990) pour une description détaillée des différents formats

les mots composés) qui contiennent les formes canoniques des lemmes, ainsi que le code de la table lexique-grammaire s'y rapportant. Ces informations permettent de fléchir le lemme et d'obtenir un dictionnaire des formes fléchies ou DELAF (respectivement DELACF pour les mots composés). Une entrée de dictionnaire au format DELAF se présente sous la forme suivante :

FormeFléchie,FormeCanonique.InformationSyntaxiqueEtLexicographique

Chaque entrée contient trois types d'information : la forme fléchie d'une entrée La forme canonique à laquelle on peut rapporter les formes fléchies, les informations lexicographiques et syntaxiques propres à l'entrée définie par les formes « fléchie » et « canonique ». La forme « fléchie » est séparée de la forme « canonique » par une virgule. La forme canonique est séparée de l'information syntaxique et lexicographique par un point. Pour les besoins de notre problématique (indexation automatique), nous avons utilisé un format légèrement différent du format DELA standard. En effet, après l'analyse du texte pour en extraire les concepts abordés, nous devons « traduire » ces concepts à l'aide des descripteurs MeSH. Nous utilisons donc les termes MeSH comme forme standard (ou forme canonique) à laquelle ramener les lemmes des formes fléchies. Dans certains cas, les entrées s'avèrent strictement identiques au format DELA standard :

diabete,diabete.N+Conc+z1 :ms

Et dans d'autres, pas :

antifongique,antifongiques.N :ms (vs. antifongique,antifongique.N :ms - format DELAF standard)

Par ailleurs, nous avons fait le choix de constituer un dictionnaire désaccentué pour deux raisons. Tout d'abord, le MeSH français n'était pas entièrement accentué dans sa version 2002, disponible au début de nos travaux. Par ailleurs, les ressources sur lesquelles nous travaillons sont issues de l'internet et bien que la majorité d'entre elles proviennent de sources institutionnelles, elles ne sont pas exemptes de coquilles et de fautes d'accentuation diverses. La désaccentuation des ressources et des dictionnaires permet de pallier à certains problèmes de reconnaissance. (Grabar, 2004) montre que ce type de normalisation permet par exemple d'améliorer sensiblement la reconnaissance des requêtes des utilisateurs dans le cadre de la recherche d'information.

Apport lexicographique

Chaque entrée a été complétée avec le code « MeSH » pour les mots clés MeSH, « QMeSH » pour les qualificatifs, « MALADIE » pour les mots clés de l'arborescence C ou F03, « ACTION » pour les mots clés de l'arborescence D27.505, « TECHNIQUE » pour les mots clés de l'arborescence E etc. Nous reprenons ici le principe des classes d'objets introduit par Gaston Gross (Gross, 1992)[22]. Les diverses classes et codes concernés sont rassemblés dans le tableau 2.5. Nous indiquons également la couverture de chaque catégorie dans le dictionnaire.

Ces informations seront par la suite utilisées pour la description de certaines associations mot clé / qualificatif pour l'indexation (évoquée plus loin en 2.5.1).

[22]cité par (Blanco & Bonell, 1998).

Concept	**Code**	**Arborescence MeSH**	**Nb. termes**	**%**
Mot clé MeSH	MeSH	Toutes	19 007	83
Qualificatif MeSH	QMeSH	-	83	100
Type de Ress. CISMeF	TR	-	8	3
Organe	ORGANE	A	1 311	91,6
Maladie	MALADIE	C et F03	4 065	97
Composés chimiques	SUBSTANCE	D sauf D05, D12, D13, D25, D27.505	3 995	82,9
Vaccin	VACCIN	D24.310.894	71	97,3
Action pharmacologique	ACTION	D27.505	254	81,9
Technique thérapeutique	TECHNIQUE	E	1 661	76,4
Spécialité médicale	SPECIALITE	G02.403.776	38	90,5
Personne humaine	Hum	M	123	67,8
Lieu géographique	Top	Z	353	95

Tab. 2.5 – Détail des informations lexicographiques contenues dans le dictionnaire MeSH (données du 01/06/05)

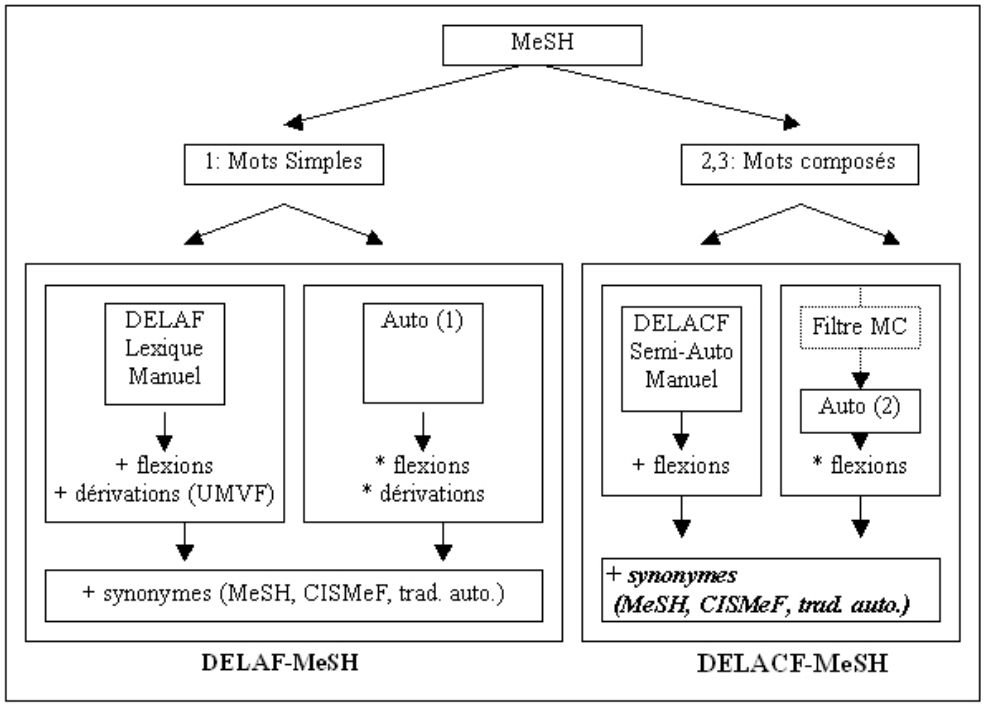

FIG. 2.7 – Méthode de construction des dictionnaires MeSH

Construction des dictionnaires

Les dictionnaires constitués ont été élaborés à partir de trois sources principales :

1. ressources disponibles
   - pour la langue générale : dictionnaires DELA existant, Lexique[23].
   - pour la langue médicale : données UMLF
2. des ajouts (semi)-automatiques
   - bases de synonyme MeSH et CISMeF
   - traduction automatique
   - traitement de certaines expressions typiques récurrentes
3. des ajouts manuels

La figure 2.7 récapitule les différents traitements mis en oeuvre pour élaborer les dictionnaires. Nous détaillons chaque étape dans les paragraphes suivants.

Ressources disponibles pour la langue générale. Certains mots clés MeSH figuraient déjà dans les dictionnaires DELA existants (Courtois, 1990) et nous avons adapté les entrées correspondantes. Elles ont dans un premier temps été converties dans le format DELA-MeSH introduit ci-dessus. Ensuite, chaque entrée a été complétée avec les informations lexicographiques décrites précédemment. (N$\sim$ 12 000)

diabete,diabete.N+MeSH+MALADIE+Conc+z1 :ms

[23]cf. `http://www.lexique.org`

Nous avons également intégré (en partie) des ressources développées par (L. F. Soualmia, Barry, & Darmoni, 2003) pour améliorer la recherche d'information dans le catalogue CISMeF. Tirées du lexique établi grâce à la base FranLex[24], ces ressources rassemblent des variantes orthographiques, des flexions et des dérivations sur les 9 000 mots clés MeSH et 83 qualificatifs utilisés pour indexer les ressources CISMeF au 1er janvier 2003. (N∼ 900)

UMLF. Les travaux effectués dans le cadre du projet UMLF (Zweigenbaum et al., 2003) ont également permis d'enrichir le dictionnaire de mots simples, après validation des résultats obtenus sur le corpus Vidal. (N∼ 3.000)

**spironolactoner,spironolactone.V+MeSH → entrée rejetée*
diagnostiquer,diagnostic.V+MeSH → entrée conservée

Certaines entrées avaient cependant déjà été obtenues grâce aux dictionnaires DELA. Dans les cas de doublons, nous avons privilégié les entrées DELA, qui contenaient plus d'information syntaxiques et sémantiques.

Ajouts manuels. Ces ressources ont été complétées manuellement afin d'étendre la couverture des dictionnaires aux termes relatifs au diabète (Une évaluation préliminaire du système d'indexation portait sur un corpus relatif au diabète) dans un premier temps, puis à toutes les maladies dans un deuxième temps (la politique éditoriale du catalogue CISMeF est ciblée sur les ressources traitant de diverses pathologies. Il est donc prioritaire pour le système d'indexation de pouvoir extraire ce type de mots clés). Ce travail est toujours en cours (N∼ 10 000).

Ajouts semi-automatiques. Nous avons introduit des informations prenant en compte la variabilité sémantique des termes sous forme de synonymes et de formes dérivées. Ainsi, les synonymes INSERM[25] et CISMeF de termes MeSH et leurs flexions ont été ajoutés. (N∼ 4 000)

personne âgee,sujet âge.N+MeSH :fs

Nous avons également utilisé une méthode de traduction automatique des synonymes MeSH américains non traduits pour enrichir la terminologie d'une part et les dictionnaires d'autre part (cf. section 2.5.2). Par exemple, « variola virus » synonyme américain du mot clé <*variole, virus*> traduit automatiquement par « virus variolique » a conduit à créer l'entrée :

virus variolique,variole_ virus.N+MeSH :ms

Par ailleurs, un certain nombre de mot clés du type 3 « langage documentaire » ont pu être traités automatiquement. Bien que ces mots clés ne se trouvent pas tels quels dans un texte en langage naturel, il est possible de déduire du mot clé lui-même quelle est le syntagme nominal lui correspondant en langage naturel. Par exemple, au mot clé <*Wolfram, syndrome*> nous pouvons rattacher le syntagme « syndrome de Wolfram ». De même, au mot clé <*vitamine a, carence*> nous pouvons rattacher le syntagme « carence en vitamine a » etc. Ainsi, il existe plusieurs catégories de mots clés pour lesquels il est possible de générer automatiquement les entrées de dictionnaire associées. Pour la première catégorie, que l'on peut généraliser en <*Nom, syndrome*>, on génèrera l'entrée :

[24] Pour plus d'information sur ce projet cf. `http://perso.limsi.fr/jacquemi/FRANLEX/` (visité le 10/09/05).
[25] Institut National pour la Santé et la Recherche en Médecine

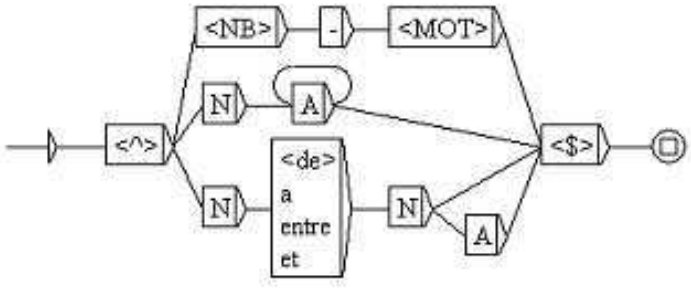

FIG. 2.8 – Automate reconnaissant les mots composés grammaticaux - Noté « filtre MC » sur la figure 2.7

Syndrome d'Nom,Nom_ maladie.N+MeSH+MALADIE :fs

si « Nom » commence par une voyelle ou par un « H » et l'entrée :

Syndrome de Nom,Nom_ maladie.N+MeSH+MALADIE :fs

Dans les autres cas. Pour la deuxième catégorie, <*substance, carence*>, on génèrera les entrées :

carence en substance,substance_ carence.N+MeSH :fs newline carences en substance,substance_ carence.N+MeSH :fp

De cette manière, nous avons pu traiter les mots clés de type maladie, carence, déficit, syndrome, tumeur, infection, acide. Cependant, il reste difficile de traiter automatiquement tous les mots de cette catégorie. Il faut remarquer qu'il est difficile de les distinguer des noms composés grammaticaux. Afin d'en isoler une partie, nous envisageons de les décrire à l'aide d'un graphe syntaxique (cf. figure 2.8) capable de reconnaître les mot clés du type N-ADJ, N de N, les composés chimiques etc.

A ce jour, les 45 000 entrées DELAF et DELACF obtenues représentent une couverture d'environ 83% du MeSH. (soit en moyenne, 2,37 entrées par terme MeSH). Ce travail est toujours en cours afin d'arriver à une couverture proche de 100%.

Bibliothèque de transducteurs

Reconnaissance de mots clés Pour rendre compte de la grande variabilité de certains mot clés (eg. <*adulte âge moyen*>, <*centre rééducation et réadaptation*>), nous avons également constitué une bibliothèque de transducteurs destinée à les repérer dans les ressources. La figure 2.9 présente le transducteur réalisé pour le descripteur obligatoire <*adulte âge moyen*>. Ainsi, les expressions « patients âgés de 50 ans » ou « femme entre 45 et 55 ans » sont reconnues comme équivalentes de <*adulte âge moyen*>. Ce transducteur complète la couverture du mot clé par les entrées de dictionnaire :

adulte d'âge moyen,adulte âge moyen.N+MeSH :ms
adultes d'âge moyen,adulte âge moyen.N+MeSH :mp

Reconnaissance de Paires MeSH Par ailleurs, l'indexation MeSH se fait également à l'aide de paires mot clé/qualificatifs (*vs.* mot clés isolés). Ainsi, nous développons en parallèle des ressources (notamment des transducteurs) permettant de reconnaître directement

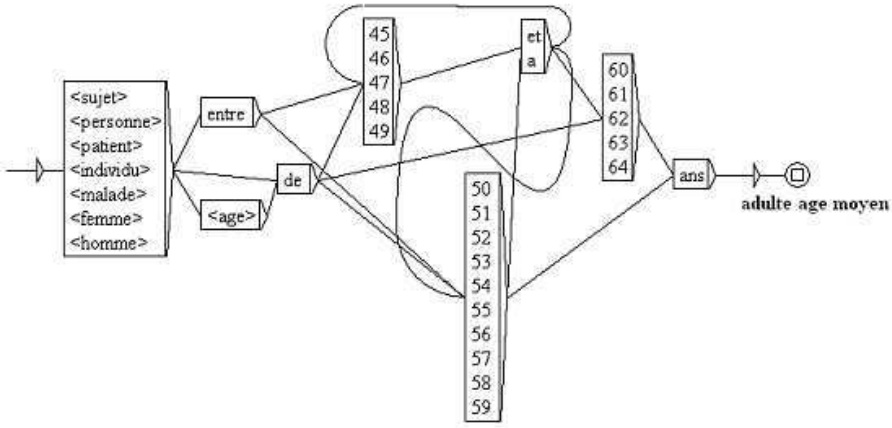

FIG. 2.9 – Transducteur reconnaissant le mot clé <*adulte âge moyen*>

les paires mot clé/qualificatif selon les règles d'indexation définies par un expert du domaine. Cette démarche s'inscrit dans un effort de formalisation des savoir-faire mis en oeuvre par les indexeurs humains. Nous procédons en plusieurs étapes :

- Entretien avec l'expert : dans une discussion informelle autour de ressources déjà indexées, l'ingénieur de la connaissance amène l'expert à expliquer le choix des paires utilisées pour l'indexation à partir d'éléments de la ressource et notamment du titre de la ressource.
- Identification de comportements récurrents : après étude de plusieurs ressources indexées avec des paires similaires (par exemple, comportant le même qualificatif) l'ingénieur de la connaissance essaie d'identifier une méthode récurrente de production des paires étudiées. Il propose à l'expert une reformulation des choix d'indexation sous forme de règle générale.
- Validation : l'expert valide les règles qui lui sont soumises, ou propose des modifications si nécessaire.
- Implémentation : construction d'un transducteur permettant d'identifier une paire mot clé/qualificatif à partir de la règle établie (quand cela est possible)

Par exemple, la règle :

> *Identification de déclencheurs tels que « lutter contre la MALADIE » ou « vaccin anti- MALADIE » → la paire <MALADIE/prévention et contrôle> doit être utilisée pour l'indexation.*

a donné lieu au transducteur illustré par la figure 2.10 :

Dans certains cas, les règles ne peuvent pas être implémentées sous forme de transducteurs :

> *Si le mot clé <biopsie> est sélectionné, ainsi qu'une MALADIE de l'arborescence C04 → la paire <MALADIE/anatomie pathologique> doit être utilisée pour l'indexation.*

Contrairement aux transducteurs (et dictionnaires) qui constituent la première étape de l'indexation (l'analyse de la ressource), ce type de règle doit être pris en compte dans une étape ultérieure de l'indexation, la révision des mots clés et paires candidats. Le processus d'indexa-

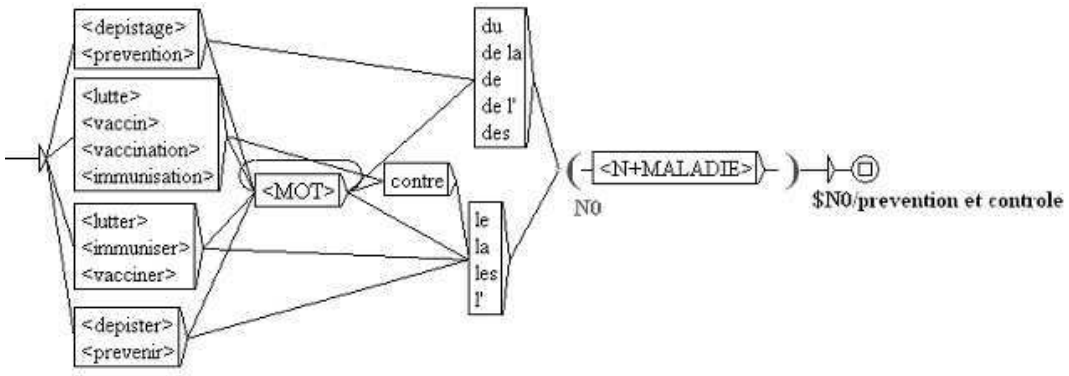

FIG. 2.10 – Transducteur reconnaissant la paire *<MALADIE/prévention et contrôle>*

tion automatique dans son ensemble, ainsi que les étapes qui le composent sont décrits dans la section 5.8.

Perspective : construction d'autres ressources. Les dictionnaires et graphes décrits ci-dessus sont en premier lieu destinés à être utilisés par le module d'analyse de surface d'un système d'indexation automatique MeSH, afin de repérer directement des mots clés MeSH dans les textes à indexer. L'avantage manifeste de cette méthode est d'offrir une grande précision sur les termes extraits : tout terme extrait correspond effectivement à une occurrence du concept MeSH concerné. Par contre, il est possible que toutes les occurrences des concepts MeSH ne soient pas extraites. Ainsi, dans la phrase « Le patient présente une tumeur inquiétante au nez. », le mot clé *<tumeur nez>* ne sera pas repéré car l'entrée

tumeur inquiétante au nez,tumeur nez.N+MALADIE+MeSH :fs

ne figure pas dans le dictionnaire. Cependant, il n'est pas justifié d'ajouter cette entrée dans le dictionnaire, dans la mesure où toutes les occurrences de l'expression « tumeur ADJ au nez » peuvent théoriquement être rencontrées également. Pour prendre en compte ce type d'expression, nous envisageons de compléter le dictionnaire DELAF-MeSH avec une nouvelle catégorie d'entrées (codée LMeSH) correspondant aux lemmes composant les mots clés MeSH. Ainsi, dans notre phrase exemple, les lemmes « tumeur » et « nez » seraient repérés et il serait possible de les combiner pour retrouver le mot clé *<tumeur nez>*.

Les abréviations, qui sont un phénomène fréquent dans les publications scientifiques médicales, doivent faire l'objet d'un traitement particulier. Deux types d'abréviations peuvent être rencontrées.

1. les abréviations « attestées », utilisées par l'ensemble de la communauté médicale mais qui demeurent ambiguës. C'est le cas de IVG qui peut être interprété soit comme "interruption volontaire de grossesse" ou comme "intervention ventricule gauche". Dans ce cas, seul le contexte permet de trancher.
2. les abréviations « locales », utilisées dans le cadre du document et explicitées par l'auteur soit en début de document (« liste des abréviations utilisées ») soit après la première utilisation (par exemple, « Le diabète gestationnel (DG) touche particulièrement... »). Une solution pour prendre en compte ces abréviations serait de construire automatique-

ment un dictionnaire local pour chaque document qui viendrait compléter les ressources disponibles de manière adaptative.

Dans le cadre de l'indexation combinée texte/image (Florea, Rogozan, Bensrhair, & Darmoni, 2005) nous travaillons sur les légendes d'images et les textes dédiés à l'imagerie médicale afin d'étendre notre bibliothèque de transducteurs à la reconnaissance de triplets mot clé / qualificatifs \ type de ressource. (cf. section 5.10.1)

Conclusion. Nous avons détaillé le travail mis en oeuvre pour la construction de ressources terminologiques de santé en français, sous la forme de dictionnaires électroniques et de transducteurs. Les premières évaluations du système d'indexation utilisant les ressources terminologiques décrites montrent que les ressources développées permettent d'extraire les termes MeSH de manière satisfaisante (cf. section 5.8).

Cependant, il est nécessaire d'enrichir ces ressources afin d'augmenter la couverture MeSH et de prendre en compte d'autres aspects de la complexité lexicale. Nous envisageons également d'adapter une partie de ce travail sur l'anglais, notamment en ce qui concerne les apports lexicographiques dans les dictionnaires électroniques.

2.5.2 Traduction automatique de termes médicaux à l'aide de corpus parallèles

Ce travail s'inscrit dans la continuité du développement de ressources médicales partagées par le catalogue CISMeF et le système d'indexation automatique MAIF.

La traduction de termes spécialisés (comme par exemple les synonymes de mots clés MeSH) fait appel à une double compétence, à la fois en traduction pour les paires de langues concernées et dans le domaine de la spécialité traitée. En pratique, il peut s'avérer difficile de trouver ou de former un expert sur ces deux points spécifiques.

Dans ce contexte, nous avons étudié deux méthodes de traduction automatique des synonymes américains du MeSH, afin d'enrichir la terminologie CISMeF tout en limitant l'intervention des experts : il s'agit d'une méthode de traduction statistique et d'une méthode d'appariement sous-phrastique. Nous avons donc isolé les synonymes américains non traduits en français pour les mots clés MeSH utilisés par CISMeF, et utilisé plusieurs corpus parallèles du domaine médical afin d'en extraire la traduction en français des synonymes qui y sont présents. Dans un premier temps, nous nous sommes limitée à la traduction directe de termes complets (Névéol & Ozdowska, 2005). Puis, compte tenu des résultats obtenus, nous avons étendu notre travail à la traduction compositionnelle de certains termes composés(Ozdowska, Névéol, & Thirion, 2005). A titre d'illustration de la tâche que nous nous fixons, le tableau 2.6 présente un échantillon des termes MeSH actuellement disponibles en français et en anglais.

Corpus de travail

Pour ce travail, nous avons utilisé plusieurs corpus parallèles anglais-français du domaine médical : le corpus « ENFR complété » décrit à la section 1.2.6, le corpus « Hansard[26] », composé de débats sur des questions de santé publique à la chambre des communes du parlement canadien (40 000 mots) et le corpus « RCP » constitué dans le cadre du projet PERTOMed[27],

[26]L'ensemble du Hansard peut-être consulté à l'aide de l'outil TransSearch à l'url `http://www.tsrali.com`

[27]Sous la responsabilité scientifique de M-C Jaulent, INSERM ERM 202 (`http://www.spim.jussieu.fr`, rubrique "Projets de Recherche")

Mot clé MeSH américain	Mot clé MeSH français	Synonyme MeSH américain (à traduire)
cardiovascular agents	agents cardiovasculaires	cardiovascular drugs
cell division	division cellulaire	cytokineses
milk, human	lait femme	breast milk
skin diseases, vesiculobullous	dermatoses bulleuses	neddon wilkinson disease
terpenes	terpenes	isoprenoids

Tab. 2.6 – Extrait du MeSH 2004

qui rassemble 94 « Résumés de Caractéristiques Produits » (600 000 mots).

Traduction statistique de termes simples

En collaboration avec Y.C. Chiao, nous avons d'abord appliqué une méthode statistique conçue pour la recherche d'équivalents traductionnels de termes à l'aide de corpus comparables[28]. Appliquée sur un corpus parallèle (le corpus « ENFR complété ») de taille quatre fois moins importante que les corpus comparables utilisés lors de l'évaluation de cette méthode (Chiao & Zweigenbaum, 2002) (340 000 mots contre 1 200 000 mots) nous avons pu obtenir des résultats équivalents pour la traduction de synonymes MeSH[29]. Avec la mesure cosine, une traduction correcte est proposée au rang 1 pour 25% des termes (vs. 23% dans (Chiao & Zweigenbaum, 2002)) et dans les 10 premiers rangs pour 47% des termes (vs. 61% dans (Chiao & Zweigenbaum, 2002)).

Cependant, seuls les termes simples peuvent être traduits avec cette méthode. Par ailleurs, un travail de validation manuel important s'avère nécessaire pour inclure les synonymes traduits dans la terminologie. L'étape de validation semble difficilement automatisable dans la mesure où les candidats traduits sont des termes extraits du corpus, donc relatifs au domaine médical. Pour ces raisons pratiques, nous avons également envisagé une deuxième méthode (alignement sous-phrastique) d'extraction automatique des synonymes MeSH français.

Traduction par alignement sous-phrastique

En collaboration avec S. Ozdowska, nous avons ensuite envisagé une approche par alignement sous phrastique. Ce travail a été réalisé en deux étapes :

1. Travail sur les corpus « ENFR complété » et « Hansard » d'une part[30] et « RCP » d'autre part, avec les données du MeSH 2003 - soit 5 166 synonymes à traduire : traduction directe dans les cas où les termes sont effectivement présents dans les corpus.

[28] Un *corpus comparable* est composé de textes rédigé dans plusieurs langues - par exemple L1 et L2 pour un corpus bilingue. Sans être des traductions exactes des textes en L2, les textes rédigés en L1 présentent des caractéristiques similaires en terme de sujet traité, de style et de genre.

[29] Dans cette étude, les termes à traduire sont les synonymes (termes simples) MeSH français non traduits (MeSH 2003) présents dans le corpus, soit 47 termes simples.

[30] La taille comparativement réduite de ce corpus par rapport aux deux autres nous a incité à regrouper les corpus EN-FR étendu et Hansard. Par la suite, du fait de la différence de registre (information médicale vs. droit de la médecine) nous avons remis en cause cette stratégie et utilisé le corpus EN-FR étendu seul. Cependant, les résultats semble indiquer qu'une étude contrastée de différents registres peut s'avérer intéressante et nous projetons d'approfondir cette perspective très prochainement.

2. Travail sur les corpus « ENFR complété » et « RCP », avec les données du MeSH 2004 - soit 20 048 synonymes à traduire : traduction compositionnelle dans les cas où l'ensemble des composants des termes sont effectivement présents dans les corpus, alors que le terme entier n'est pas présent.

Nous présentons donc dans un premier temps le principe d'appariement sous phrastique mis en oeuvre pour l'ensemble du travail. Puis, nous motivons l'approche de traduction compositionnelle. Nous présentons ensuite les résultats obtenus pour la traduction directe et la traduction compositionnelle. Après une discussion, nous faisons le bilan de ces expériences et synthétisons les perspectives de recherche que nous envisageons d'adopter pour la poursuite de ce travail.

Principe de la procédure d'appariement Pour la recherche des traductions en français des synonymes MeSH américains, nous avons mis en oeuvre une méthode d'appariement de mots et de syntagmes dite « appariement par propagation syntaxique » (Ozdowska, 2004a, 2004b). Il s'agit d'une approche linguistique d'appariement de segments sous-phrastiques basée sur l'analyse syntaxique bilingue de corpus parallèles anglais/français. Son principe est le suivant : à partir de deux mots qui sont en relation de traduction dans des phrases alignées, appelés couple amorce, le lien d'équivalence est propagé vers d'autres mots en suivant les relations syntaxiques préalablement mises en évidence. Plus précisément, en partant du couple amorce (protective, protecteurs), dont chaque élément est en relation syntaxique avec un nom, on peut apparier (clothing, vêtements) comme le montre la figure 2.11[31].

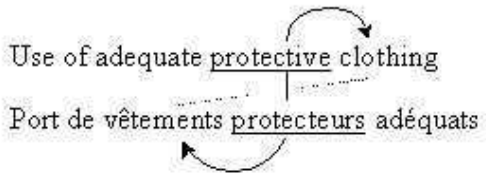

FIG. 2.11 – Principe d'appariement par propagation syntaxique

La technique d'appariement par propagation syntaxique requiert par conséquent que l'on dispose d'un corpus aligné au niveau des phrases, d'outils d'analyse pour les deux langues de travail, le français et l'anglais, ainsi que de couples amorces servant de point de départ à la propagation.

Traitement des corpus Le repérage des relations syntaxiques est pris en charge par les analyseurs Syntex (Bourigault & Fabre, 2000) qui prennent en entrée un corpus étiqueté[32] et identifient, pour chaque phrase du corpus, des relations syntaxiques telles que sujet, objet direct et indirect, modifieur etc. L'appariement s'effectue par conséquent entre des mots lemmes et non des mots formes. Comme la plupart des méthodes travaillant au niveau sous-phrastique, la méthode d'appariement par propagation syntaxique nécessite un corpus préalablement aligné au niveau des phrases. Les corpus de travail dont nous disposons sont alignés de manière fiable uniquement au niveau des paragraphes. Le découpage en phrases étant pris en charge

[31] Le sens des flèches correspond à celui des relations de dépendance syntaxique

[32] L'étiqueteur utilisé pour les deux langues est Treetagger (`http://www.ims.uni-stuttgart.de`).

de manière indépendante dans chacune des deux langues par les étiqueteurs, l'alignement à ce niveau de segmentation est susceptible de présenter des erreurs. Nous avons pris le parti de ne pas corriger les éventuels décalages et avons ignoré, lors du processus de recherche des couples amorces ainsi que de celui de propagation, les phrases non alignées [33].

Identification des couples amorces Les couples amorces permettant d'initialiser le processus de propagation peuvent être fournis au système de différentes manières. Il est possible d'utiliser des ressources lexicales bilingues préexistantes, de construire de telles ressources à partir du corpus ou encore de repérer des cognats, c'est-à-dire des chaînes de caractères identiques ou très proches dans les deux langues. Nous avons, dans un premier temps, choisi de combiner la projection d'une ressource lexicale existante et la recherche de cognats (autres que ceux présents dans la ressource) au niveau des phrases alignées. En effet, nous disposions d'une liste constituée des descripteurs MeSH américains et de leur traduction en français (liste 1), dont nous avons extrait les mots simples[34]. Nous avons ainsi obtenu, à partir d'une liste de 6127 mots, 28139 couples amorces sur un ensemble de 10299 phrases alignées (Tableau 2.15).

Il convient de noter que seuls 556 couples de la liste de départ sont effectivement présents dans le corpus et ont donc pu être utilisés pour la recherche des amorces. Dans un second temps, cette ressource nous est apparue comme insuffisante et ce principalement pour deux raisons.

Premièrement, elle ne contient que des noms, ce qui implique que l'alignement ne peut concerner que des mots qui sont en relation syntaxique avec un nom. Par conséquent, si l'on considère l'exemple de la figure 2.12 il apparaît clairement que l'équivalent français de *nightmare*, qui est l'un des synonymes dont on cherche la traduction, ne pourra être trouvé que si l'on dispose du couple amorce constitué des verbes *continue/durer*, à moins que l'on ne trouve ailleurs dans le corpus une ou plusieurs autres occurrences de *nightmare* et *cauchemar*, toutes deux en relation syntaxique avec des noms amorces.

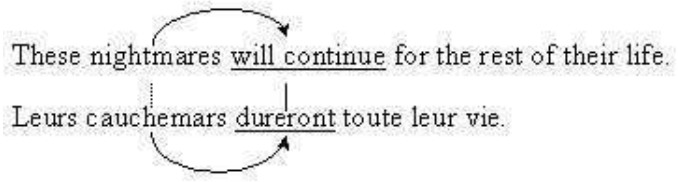

FIG. 2.12 – Propagation syntaxique à partir d'un couple amorce de verbes

Deuxièmement, les noms que cette liste contient relèvent pour la plupart d'un vocabulaire spécialisé relatif au domaine de la santé (Figure 2.13), ceux relevant de la langue générale et susceptibles d'être présents dans le corpus étant peu représentés.

Nous avons donc fait le choix de compléter la liste existante avec des données extraites du corpus (Figure 2.14) afin d'étudier l'influence du nombre et de la diversité des couples amorces sur les alignements obtenus, en termes de catégories grammaticales (restriction aux noms pour la liste 1 versus toutes catégories confondues pour la liste 2) et de type de vocabulaire

[33] D'autres algorithmes d'appariement de phrases en corpus parallèles sont décrits et évalués dans (Véronis, 2000).

[34] Les règles de propagation utilisées actuellement sont fondées uniquement sur les mots simples

bromine	brome
bromizovalum	bromizoval
bromouracil	bromouracile
bronchography	brochographie

FIG. 2.13 – Descripteurs MeSH et leur traduction - extrait de la liste 1

(spécialisé pour la liste 1 versus spécialisé et général pour la liste 2). Pour ce faire, nous avons utilisé une méthode largement répandue dans les travaux sur l'alignement, fondée sur l'hypothèse que les mots qui apparaissent fréquemment dans des segments de texte alignés ont de fortes chances d'être en relation de traduction (Gale & Church, 1991; Ahrenberg, Andersson, & Merkel, 2000). Afin d'isoler des couples de mots en relation de traduction dans nos corpus, nous avons utilisé la mesure d'association Jaccard avec des seuils et des techniques de filtrage de la liste des associations obtenues identiques à ceux décrits dans (Ozdowska, 2004a, 2004b) :

- calcul du Jaccard pour les mots dont la fréquence sur l'ensemble du corpus est égale ou supérieure à 5
- sélection des associations pour lesquelles la valeur du Jaccard est égale ou supérieure à 0,2
- filtrage de la liste des associations par reconnaissance des cognats et par vérification de la réciprocité de l'appariement

die	mourir
monitor	surveiller
next	prochain
often	souvent

FIG. 2.14 – Traductions extraites du corpus par calcul des cooccurrences - extrait de la liste 2

Enfin, nous avons fusionné les deux listes (liste 2) pour obtenir au total un ensemble de 8 866 couples de mots qui ont à leur tour été projetés au niveau des phrases alignées, ce qui a permis d'identifier 39 903 couples amorces (Tableau 2.15). 2 093 des couples de la liste 2 ont pu être pris en compte lors de la phase de repérage des couples amorces.

Corpus	**ENFR complété / Hansard**	**RCP**
nb. phrases alignées	10 299	18 034
nb. couples amorces (liste 1)	28 139	51 600
nb. couples amorces (liste 2)	39 903	72 136

FIG. 2.15 – Repérage des couples amorces : influence du lexique bilingue utilisé

Une fois les couples amorces repérés, la propagation syntaxique des liens d'appariement

repose sur différents patrons de propagation dont on a pour le moment limité la définition aux cas de correspondance directe, c'est-à-dire ceux où la configuration syntaxique est identique dans les deux langues. Comme décrit dans (Ozdowska, 2004a), chaque patron rend compte de la catégorie grammaticale des mots sources et des mots visés par la propagation, de la relation syntaxique qui sert de base à la propagation ainsi que du sens dans lequel cette dernière s'effectue.

Traduction compositionnelle : pourquoi ? Nous avons constaté que sur l'ensemble des termes à traduire, soit 20 048 avec les données MeSH 2004, seule une petite partie est effectivement présente dans les corpus de travail, soit environ 300 pour le corpus « ENFR complété » et 407 pour le corpus « RCP ». Par ailleurs, bien qu'un grand nombre de termes complexes n'apparaissent pas dans ces corpus, il n'en va pas de même pour leurs constituants. Ainsi, nous avons évalué que les constituants de près de 5 000 synonymes se trouvent dans chacun des corpus de travail. Par exemple, bien que le terme *premature birth* ne soit présent ni dans l'un ni dans l'autre des deux corpus, on trouve ses composants, *premature* et *birth*, notamment dans les contextes suivants :

> *They may care for immunocompromised patients (including **premature** infants)...*
> *Ils peuvent s'occcuper de patients immunodéprimés (y compris de bébés **prématurés**)...*
>
> *The infant can be vaccinated at **birth**...*
> *L'enfant pourra être vacciné après sa **naissance**...*

C'est pourquoi, afin d'augmenter la couverture des termes complexes traduits, nous avons envisagé de déduire la traduction des synonymes à partir des traductions respectives des mots qui les composent : c'est la traduction compositionnelle (cf. figure 2.16).

> *They may care for immunocompromised patients (including **premature** infants)...*
> *Ils peuvent s'occcuper de patients immunodéprimés (y compris de bébés **prématurés**)...*
> premature ↔ prématuré
> *The infant can be vaccinated at **birth**...*
> *L'enfant pourra être vacciné après sa **naissance**...*
> birth ↔ naissance
> ---
> premature birth ↔ naissance prématurée[35]

FIG. 2.16 – Traduction compositionnelle

Nous reprenons ainsi le principe de compositionnalité (Frege, 2001; Janssen, 1997) selon lequel le sens d'un terme complexe résulte directement du cumul des sens des unités qui le composent et nous l'appliquons à la traduction des termes complexes. En effet, nous tenons pour acquis que ce principe s'observe dans les deux langues mises en correspondance. Autrement dit, nous faisons l'hypothèse que la traduction des composants des termes complexes anglais pris individuellement permet d'inférer des termes complexes équivalents corrects en français.

[35]Le changement de flexion a fait l'objet d'un ajustement manuel le cas échéant.

Traduction compositionnelle : Application. Le principe de traduction compositionnelle présente bien évidemment des limites comme le soulignent de nombreux travaux en traduction automatique (par exemple, (Rosetta, 2003)). En effet, même si l'on connaît la traduction française de chacun des constituants du terme *breast milk* (*breast* se traduit par sein et *milk* par lait), il est difficile d'en déduire la traduction du terme complet qui est *lait maternel*. A la non-compositionnalité du sens s'ajoute alors la non-correspondance structurelle entre les termes (ici, la structure *N N* du terme en anglais diffère de la structure *N Adj* de sa traduction en français). Néanmoins, comme le montre Daille (1994), il existe des régularités dans la manière de rendre les structures syntaxiques des termes anglais en français. Par exemple, les termes anglais de structure *Adj N* sont régulièrement traduits en français par des termes de structure N Adj. D'autres correspondances, telles que $N1$ *Prep* (*Det*) $N2$ ou $N1$ *de* $N2$, sont beaucoup plus rares. Gaussier (2001) estime la probabilité de correspondance anglais/français *Adj N*/*N Adj* à 0,84. Nous faisons donc l'hypothèse que la correspondance structurelle va de pair, dans la plupart des cas, avec le principe de compositionnalité. Afin de tester le principe de traduction compositonnelle, nous nous sommes limités dans un premier temps aux termes de structure *Adj N*. Le corpus « ENFR complété » contient 489 synonymes de ce type pour lesquels chaque constituant a pu être traduit grâce à la procédure d'appariement de mots. (respectivement 635 synonymes pour le corpus « RCP »[36]). Nous avons calculé la traduction de ces synonymes selon la méthode suivante : $T(Adj\ N) = T(N) + T(Adj)$ où $T(N)$ et $T(Adj)$ correspondent à la traduction de N et Adj. Si on reprend l'exemple de la Figure 2.16 :

$T(premature\ birth) = T(birth) + T(premature)$

$\rightarrow T(premature\ birth) = naissance\ prématurée$

Dans le cas où plusieurs équivalents français sont proposés pour l'un et/ou les deux composants N et Adj du synonyme à traduire, nous avons fait le choix de ne retenir que la première traduction proposée, à savoir la plus fréquente, nommée par la suite « candidat de rang 1 ». Nous discuterons des avantages et inconvénients qui découlent de ce choix.

Résultats : Traduction des termes entiers A l'aide des corpus « ENFR complété + Hansard » et « RCP », une traduction a été proposée pour 217 synonymes au total, soit 4,2% de l'ensemble des synonymes MeSH 2003 à traduire. Après validation par un expert en terminologie médicale, 133 nouveaux termes ont été inclus dans la terminologie (plus des flexions et/ou dérivations de ces termes le cas échéant - soit 190 termes au total). La précision est de 71% sur le corpus « ENFR complété + Hansard » et de 70% sur le corpus « RCP ». Le rappel est respectivement de 54% et 60%. Ainsi, la méthode utilisée offre une précision globale de 70% pour un rappel global de 57%. La figure 2.17 présente les performances obtenues en fonction du nombre d'occurrences des synonymes dans le corpus « ENFR complété + Hansard ». Les points (*/°) représentent les valeurs de précision et de rappel pour un nombre d'occurrences donné. Afin de simplifier la représentation, les courbes en trait plein montrent la précision moyenne pour les termes de basse fréquence (une seule occurrence dans le corpus), de faible fréquence (entre deux et cinq occurrences), de moyenne fréquence (entre six et dix occurrences) et de haute fréquence (plus de 10 occurrences).

Le tableau 2.7 présente le nombre de traductions extraites pour chaque corpus, en fonction du lexique utilisé pour repérer les couples amorces. Les traductions proposées se recoupent parfois : la traduction d'un synonyme peut être extraite des deux corpus ou à l'aide des deux lexiques pour un corpus donné. Nous indiquons également le nombre total de synonymes

[36]Soit au total 915 synonymes distincts, si on tient compte des termes présents dans les deux corpus.

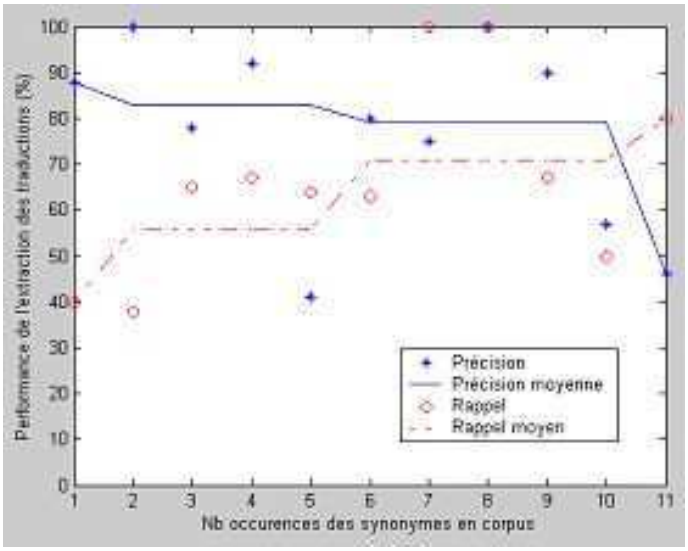

FIG. 2.17 – Performance de la méthode en fonction du nombre d'occurrences des synonymes dans le corpus « ENFR complété + Hansard »

distincts pour lesquels au moins une traduction a été extraite pour chaque corpus, puis pour l'ensemble des deux corpus.

Analyse globale des résultats pour la traduction des termes entiers. Des résultats complémentaires sont proposés si l'on utilise la liste 2 pour le repérage des couples amorces (102 synonymes traduits vs. 115 avec la liste 1 seule pour le corpus « ENFR complété + Hansard »). Il est donc souhaitable de travailler avec des couples amorces appartenant à des registres de langue et à des catégories grammaticales diverses afin d'optimiser les résultats. En observant les performances de notre méthode en fonction du nombre d'occurrences des synonymes traduits dans le corpus, on constate que les résultats sont inégaux. La meilleure précision (88%) est obtenue pour les termes qui n'apparaissent qu'une seule fois dans le corpus. Ce résultat met en évidence un avantage de la méthode utilisée par rapport à des méthodes statistiques, comme par exemple les modèles IBM (Och & Ney, 2003), qui nécessitent un grand nombre d'occurrences des termes pour proposer un alignement fiable. Par contre, pour un grand nombre de ces termes, aucune traduction n'est proposée. La différence de rappel entre les deux corpus peut donc s'expliquer par le fait que le corpus « ENFR complété + Hansard » comporte une proportion plus élevée de synonymes de fréquence 1 (45%) que le corpus « RCP » (35%). En revanche, pour les termes très fréquents, le rappel est optimal (80%). La précision est cependant moins intéressante du fait du nombre important de candidats erronés proposés. Nous avons étudié plus précisément les causes de silence (mesure équivalente à 1 - rappel) et de bruit (mesure équivalente à 1 - précision).

Analyse du silence. L'impossibilité de trouver une traduction s'explique par l'une des raisons suivantes :

– Erreur d'alignement au niveau des phrases : dans près de 30% des cas, les termes non traduits se trouvent dans des phrases qui n'ont pu être correctement alignées.

	Lexique MeSH seul (liste 1)	**Lexique MeSH + cooccurrences** (liste 2)
	corpus « ENFR complété + Hansard »	
Nb traductions extraites	102	115
Nb total traductions	116 distincts	
Précision et Rappel	P = 71% - R = 54%	
	corpus « RCP »	
Nb traductions extraites	139	146
Nb total traductions	148 distincts	
Précision et Rappel	P = 70% - R = 60%	
	corpus « ENFR complété + Hansard » et « RCP »	
Nb total traductions	217 distincts	
Précision et Rappel	P = 70% - R = 57%	

TAB. 2.7 – Nombre de traductions extraites de chaque corpus en fonction du lexique

– Différence structurelle entre les deux langues : les phrases présentent une différence de formulation qui ne permet pas la propagation.

Occluded dialysis access grafts
Occlusion des courts-circuits artério-veineux (dialyse)

– Appariement de type m-n : le nombre d'unités à mettre en correspondance est différent dans les deux langues :

. . .particularly if diarrhea is accompanied by weight loss, ***hematochezia****, . . .*
*. . .en particulier si la diarrhée s'accompagne d'une perte de poids, de l'****émission de selles sanglantes*** *. . .*

Or, pour le moment, la méthode d'appariement employée ne permet de trouver que des correspondances entre des mots simples ou des syntagmes constitués de deux mots pleins. Tout autre type d'appariement, notamment les appariements m-n, n'est pas pris en compte.

– Absence de couple amorce : aucun couple amorce ne permet d'atteindre par propagation les termes que l'on cherche à traduire ou bien le couple amorce trouvé est incorrect :

. . .because of the low ***wages*** *paid to the maker.*
. . .en raison du maigre ***salaire*** *versé au fabricant.*

Ni le couple *paid/versé* ni *low/maigre* n'ont pu être extraits par le calcul du Jaccard à cause d'une fréquence de cooccurrence insuffisante.

– Longueur des termes : les termes que l'on cherche à apparier comportent plus de deux mots pleins. Ils ne sont donc pas pris en charge par la méthode d'appariement, comme nous venons de l'évoquer ci-dessus.

– Erreur d'analyse syntaxique : il y a une erreur d'analyse syntaxique dans l'une ou l'autre des deux langues et on ne dispose pas de la relation syntaxique nécessaire à la propagation. C'est le cas pour la préposition *par* en français qui n'a pu être rattachée à son recteur, *meurtre*, dans l'exemple suivant[37] :

[37]Dans cet exemple et dans les suivants, les couples amorces sont soulignés, les termes que l'on cherche à apparier sont en gras

The issue of euthanasia, better known to many of us as ***mercy*** *killing...*
L'euthanasie, mieux connue de beaucoup sous le vocable de meurtre par ***compassion****...*

– Absence de relation syntaxique : les occurrences de l'un ou des deux termes apparaissent dans des configurations syntaxiques isolées, par exemple entre virgules. Elle ne sont liées aux autres mots de la phrase par aucune relation syntaxique :

...an increase in spontaneous abortion, ***stillbirth****, or prematurity.*
...un risque accru d'avortement spontané, de ***mortinaissance*** *ou de prématurité.*

– Terme non traduit : le synonyme dont on cherche la traduction n'est pas traduit dans la langue cible mais repris tel que en langue source dans la partie cible du corpus.

Analyse du bruit Parmi les traductions erronées proposées, on retrouve plusieurs type d'erreurs :

– traduction incomplète : par exemple, « banque » proposé comme traduction du synonyme « databanks » pour lequel la traduction attendue était « banque de données ». Ceci s'explique par le fait que la méthode ne peut traiter les appariements de type m-n.
– traduction par une forme dérivée du terme : par exemple, « varicelleuse » proposé comme traduction du synonyme « varicella » pour lequel la traduction attendue est « varicelle ». Comme proposé dans (Debili, 1997), deux types d'appariements sont à distinguer. Les premiers mettent en relation des unités qui peuvent être considérées comme traduction l'une de l'autre aussi bien dans le contexte linguistique où elles apparaissent qu'en dehors de ce dernier. On parlera alors d'appariements non contextuels. Les seconds mettent en correspondance des unités qui peuvent être considérées comme traduction l'une de l'autre seulement dans le contexte linguistique où elles apparaissent. Il s'agit d'appariements contextuels. L'appariement *varicella/varicelleuse* relève du second type. Considérons l'un des couples de phrases dont cette correspondance a été extraite :

Recurrences of varicella-like rash have been reported by 4% to 13% of individuals who had previous ***varicella*** *infection.*
Des cas récurrents d'éruption varicelliforme ont été observés chez 4% à 13% des personnes ayant déjà eu une infection ***varicelleuse****.*

Il apparaît clairement que l'appariement *varicella/varicelleuse*, obtenu à partir du couple amorce *infection/infection*, ne résulte pas d'une erreur de l'algorithme de propagation mais qu'il s'agit d'un appariement contextuel « seulement recevable en contexte et non reconnu comme étant une traduction acceptée de manière générale et dès lors répertoriée en principe dans un dictionnaire bilingue » (Debili, 1997).

– traduction par un terme relevant du même champ lexical ou par un hyperonyme : par exemple, *anticoquelucheux* ou *maladie* proposés comme traductions du synonyme *pertussis* pour lequel la traduction attendue était *coqueluche*. Il s'agit, comme ci-dessus, d'appariements contextuels :

... varicella rates of 3% to 4% per year are expected to occur after ***varicella*** *vaccination.*
... des taux annuels de varicelle de 3% à 4% après la vaccination ***antivaricelleuse****.*

Nous avons, dans les deux cas, des différences de formulation dans les deux langues qui peuvent apparaître de manière récurrente dans le corpus, reflétant un usage établi.

Résultats : Traduction compositionnelle. Avant de présenter les résultats, nous exposons la méthode d'évaluation adoptée.

Méthode d'évaluation L'évaluation de notre travail s'effectue en deux étapes. Dans un premier temps, nous avons évalué la traduction des composants (*N* et *Adj*) puis la traduction compositionnelle des synonymes MeSH du type *Adj N* qui en est déduite. L'évaluation des traductions des composants puis des synonymes a été réalisée manuellement sur un échantillon représentatif de taille 100. Chaque traduction extraite est jugée correcte ou erronée par un traducteur (AN) et une vérification en dictionnaire ou en corpus est effectuée. Dans le cas de composants polysémiques, toutes les traductions possibles ont été considérées comme correctes dans la mesure où, hors contexte, il n'est pas possible de préférer l'une à l'autre. Ainsi, les traductions *médicament* et *drogue* du composant *drug* sont considérées comme correctes. Par contre, la traduction *drogue cardiovasculaire* du terme *cardiovascular drug* a été considérée comme erronée car, dans ce contexte, seul le terme *médicament cardiovasculaire* est une traduction correcte. Pour la traduction compositionnelle, nous avons classé les erreurs de traduction rencontrées en deux catégories : erreurs dues à la traduction erronée d'au moins un composant et erreurs dues à la non-compositionnalité.

Dans un second temps, nous avons procédé à une validation manuelle des synonymes obtenus par traduction avec un expert en terminologie médicale (BT), afin de les inclure dans la terminologie CISMeF. L'expert ne prend en compte que le terme et le synonyme proposé et juge si ce dernier est valable ou non. Les deux étapes de la validation sont manuelles et il faut remarquer qu'il existe un biais inévitable constitué par le jugement (dans certains cas subjectif) de l'expert, traducteur ou terminologue.

Résultats. Le Tableau 2.8 présente la précision de la méthode employée pour la traduction des composants simples (noms et adjectifs) utilisés par la suite pour la traduction compositionnelle. La colonne 1 indique la précision obtenue si tous les candidats extraits sont pris en compte. En moyenne, 1,5 candidats ont été proposés pour chaque adjectif du corpus « ENFR complété » (respectivement 2,1 pour le corpus « RCP ») et 2,7 candidats ont été proposés pour chaque nom du corpus « ENFR complété » (respectivement 3,4 pour le corpus « RCP »). La colonne 2 indique la précision obtenue en considérant seulement les candidats de rang 1 (les plus fréquents).

Le Tableau 2.9 présente une évaluation quantitative de cent traductions compositionnelles pour chaque corpus.

Analyse globale des résultats pour la traduction compositionnelle Nous constatons que la précision de la traduction des composants est nettement meilleure si on ne considère que les candidats de rang 1, plutôt que l'ensemble des candidats (pour les adjectifs, 85% au rang 1 vs. 64% en considérant tous les candidats pour le corpus « ENFR complété » et 91% au rang 1 vs. 57% en considérant tous les candidats pour le corpus « RCP »). On peut cependant remarquer une différence de précision notable sur l'ensemble des candidats entre les adjectifs et les noms. Cette différence vient en partie du grand nombre de candidats proposés pour certains noms - par exemple, pour le corpus RCP, 15 candidats pour vaccine (dont trois traductions correctes) ou, pour le corpus CISMeF, 26 candidats pour maladie (dont trois traductions correctes). En terme de précision, la sélection des candidats de rang 1 constitue un avantage certain pour la traduction compositionnelle et semble par conséquent pleinement

	Tous les candidats	**Candidats rang 1**
	corpus « ENFR complété »	
Noms	39%	80%
Adjectifs	64%	85%
	corpus « RCP »	
Noms	31%	89%
Adjectifs	57%	91%
	corpus « ENFR complété + Hansard » et « RCP »	
Nb total traductions	217 distincts	
Précision et Rappel	P = 70% - R = 57%	

TAB. 2.8 – Précision de la traduction des composants par corpus

Corpus	**« ENFR complété »**	**« RCP »**
Traduction correcte	72	75
dont synonymes validés	46	47
Traduction erronée	28	25
dont traduction erronée d'un composant	24	12
dont erreur due à la compositionnalité	4	13

TAB. 2.9 – Évaluation de 100 traductions compositionnelles par corpus

justifiée. Par contre, elle peut représenter un inconvénient en terme de rappel. En effet, elle ne permet de proposer qu'une traduction et une seule pour chaque synonyme américain alors que, dans certains cas, plusieurs traductions sont disponibles et pourraient être retenues par le terminologue. Ainsi, pour le terme *stable population*, la traduction compositionnelle proposée suite à la sélection des candidats de rang 1 est *population stable*. Or, si l'on prend en compte les candidats de rang supérieur à 1, on obtient une autre traduction correcte, à savoir *population constante*. Il en est de même pour le synonyme *vaginal injuries*, correctement traduit par *lésions vaginales*, pour lequel une autre traduction pourrait être obtenue en tenant compte des autres candidats pour la traduction de *injuries* : *atteintes vaginales*. L'impact de la sélection au rang 1 sur le rappel, mais aussi sur la précision, mérite donc d'être étudié de manière plus approfondie, notamment à la lumière des travaux tels que ceux menés en recherche d'information par Hull and Grefensette (1996), par exemple, qui tiennent compte de tous les candidats et opèrent un filtrage en corpus pour ne conserver que les termes pertinents pour la requête. Statistiquement, on pouvait attendre une précision de $0,84 * 0,85 * 0,80$[38], soit environ 57% pour le corpus « ENFR complété » (respectivement 68% pour « RCP »).

[38]La probabilité qu'un terme de la forme *Adj N* soit correctement traduit par compositionnalité selon notre méthode dépend de la probabilité qu'il y ait une correspondance entre ce terme et sa traduction en *N Adj* (0,84 d'après (Gaussier, 2001)), de la probabilité que le premier terme soit correctement traduit (estimée à 0,85 pour le corpus « ENFR complété » d'après les résultats du tableau 2.8) et de la probabilité que le second terme soit correctement traduit (estimée à 0,80).

Les résultats obtenus (72% et 75% respectivement) sont de cet ordre et même supérieurs. La correspondance anticipée entre *Adj N* et *N Adj* est effectivement observée : au moins 72 cas sur 100 pour le corpus « ENFR complété » (respectivement 75 pour « RCP »). Pour les cas où la traduction compositionnelle proposée était incorrecte, nous avons étudié plus précisément les sources d'erreur. Celles-ci sont liées soit à la traduction des composants, soit aux limites de la compositionnalité.

Erreur de traduction des composants Nous avons identifié trois principaux types d'erreur pour ce qui est de la traduction des composants :

1. erreur de traduction de l'un/des deux composants. Par exemple, dans *agricultural crops*, la traduction proposée pour *crops* est *poussées*. Il s'agit en effet de l'une des traductions possibles pour ce mot dans les corpus ; on la trouve notamment dans le contexte
   (...) appear in successive ***crops****/ (...) apparaissent en* ***poussées*** *successives*
   Par contre, ce n'est pas la traduction attendue dans *agricultural crops*. L'équivalent français résultant de la traduction compositionnelle, *poussées agricoles*, est donc erroné.
2. erreur de traduction de l'un/des deux composants ayant pour origine la sélection au rang 1. La traduction proposée pour *alcoholic cirrhosis* est *cirrhose alcoolisée* alors que la traduction correcte est *cirrhose alcoolique*. Le bon équivalent pour *alcoholic*, *alcoolique*, apparaît en seconde position dans la liste des traductions possibles. La sélection se faisant au rang 1, seule la traduction la plus fréquente, *alcoolisée*, est prise en compte.
3. erreur de traduction de l'un/des deux composants ayant pour origine sa/leur polysémie. Le terme *drug*, par exemple, se traduit en français soit par *drogue* soit par *médicament*. Dans les corpus de travail, c'est son emploi au sens de *drogue* qui est le plus fréquent, le sens *médicament* étant moins représenté. Ainsi, le terme *cardiovascular drug* a été traduit par *drogue cardiovasculaire*, *médicament cardiovasculaire* étant la traduction correcte attendue.

Même si la précision est globalement meilleure lorsque l'on sélectionne la traduction des composants au rang 1, il apparaît que, tout comme pour le rappel, ce critère est dans certains cas trop contraignant.

Limites de la compositionnalité Les erreurs de traduction rencontrées confirment deux limites du principe de compositionnalité :

- échec de la correspondance structurelle *Adj N*/*N Adj*. Par exemple, le synonyme *bacterial count* a été traduit par *nombre bactérien*, la traduction attendue étant *nombre de bactéries*. Nous avons donc affaire à une correspondance structurelle de type *Adj N*/*N de N*, dont la probabilité est estimée à seulement 0,13 dans (Gaussier, 2001). On voit ici que la correspondance des structures limite le nombre de traductions qui peuvent être obtenues avec notre méthode. Il peut arriver que la « formulation préférée » ne soit pas trouvée, au profit d'une forme moins usitée. Par exemple, le synonyme *medical school* a été traduit par *école médicale* qui est un terme peu employé par rapport à la traduction la plus usitée, *école de médecine*, qui a une structure différente de *N Adj*.
- échec de l'équivalence des sens. Le synonyme *cold sore* a été traduit par *lésion froide* vs. *feu sauvage* qui est l'une des traductions correctes. Il s'agit ici d'une collocation du même type que *breast milk* : la traduction du tout ne peut être déduite de la traduction des composants. L'autre traduction correcte, *bouton de fièvre*, présente à la fois le problème de l'équivalence structurelle et sémantique.

Enrichissement de la terminologie. La dernière étape de notre travail a consisté à valider les synonymes obtenus par extraction des traductions avec un expert en terminologie médicale, afin de les inclure dans la terminologie CISMeF. Lors de cette phase de validation, nous avons rencontré plusieurs cas où il n'a pas été possible d'inclure la traduction proposée dans la terminologie bien que cette dernière ait été correcte. Lors de la phase de validation des synonymes traduits, nous avons rencontré plusieurs cas où il n'a pas été possible d'inclure la traduction proposée dans la terminologie bien que cette dernière ait été correcte. Ces cas relevaient soit de la différence de complexité lexicale entre les deux langues, soit d'une différence d'usage.

Différence d'usage. Tout d'abord, certaines traductions se sont révélées ambiguës en français. Par exemple, le synonyme <*cirrhosis*> du mot clé américain <*fibrosis*> (en français, <*fibrose*>) a été traduit par « cirrhose » grâce à notre méthode. Le terme « cirrhosis » a donc été correctement traduit par « cirrhose » mais, en français, « cirrhose » a une connotation restreinte qui se limite à la cirrhose du foie - d'ailleurs, le mot clé MeSH américain <*liver cirrhosis*> est traduit en français par <*cirrhose*>. Il n'est donc pas possible d'utiliser « cirrhose » comme synonyme de « fibrose ».

Bien que certains termes soient des équivalents traductionnels en français et en anglais, l'usage des termes dans chaque langue met au jour un glissement de sens et la traduction, bien que correcte, peut finalement se révéler ambiguë en français. Par exemple, le synonyme <*intravenous infusion*> du mot clé américain <*intravenous perfusion*> (en français, <*perfusion intraveineuse*>) a été traduit par *injection intraveineuse* grâce à notre méthode. Cette traduction est acceptable dans la mesure où, en anglais, les termes *perfusion/injection/infusion* semblent être plus facilement utilisés les uns pour les autres, en dépit de la différence conceptuelle entre *perfusion/infusion* (administration continue) et *injection* (administration instantanée) qui, en français, est beaucoup plus marquée. La traduction *injection intraveineuse* reflète l'imprécision dont font preuve certains locuteurs anglophones dans l'usage de ces termes.

Par ailleurs, il existe bien un mot clé MeSH différent pour désigner le concept d'administration instantanée : <*intravenous injection*> en anglais et <*injection intraveineuse*> en français. Il n'est donc pas possible d'utiliser <*injection intraveineuse*> comme synonyme de <*perfusion intraveineuse*> sans introduire une ambiguïté dans la terminologie. Dans une optique de validation semi-automatique des résultats obtenus, il semble alors pertinent de filtrer les cas où le synonyme traduit sera automatiquement rejeté. Pour cela, il suffit de vérifier d'une part que le synonyme traduit est différent du mot clé auquel il se rapporte et d'autre part que le synonyme traduit ne correspond à aucun autre mot clé de la terminologie.

Différence de complexité lexicale Certains concepts peuvent être désignés par deux termes synonymes en anglais pour lesquels il n'existe qu'une seule et même désignation en français. Par exemple, le synonyme <*microbiological phenomena*> du mot clé américain <*microbiologic phenomena*> (en français, <*phénomène microbiologique*>) a été traduit par *phénomène microbiologique* grâce à notre méthode. Cette traduction est correcte. Cependant en français, il n'existe qu'un seul terme pour désigner le concept *phénomène microbiologique* contrairement à l'anglais où il en existe deux : le terme formé sur l'adjectif *microbiological* et le terme formé sur l'adjectif *microbiologic*. Ainsi, en français, le terme et le synonyme américains ne peuvent être différenciés. Ce phénomène peut également être observé pour les mots simples

(par exemple, <*scar*> et <*cicatrix*> ont pour seul équivalent le terme <*cicatrice*>).

Inversement, pour certains synonymes, plusieurs des traductions proposées ont pu être incluses dans la terminologie. Par exemple, le synonyme <*cannabis smoking*> du mot clé américain <*marijuana smoking*> (en français, <*consommation de marijuana*>) a été traduit grâce à notre méthode par <*consommation de cannabis*> et <*inhalation de cannabis*>, qui ont tous les deux été inclus dans la terminologie.

Conclusion sur la traduction automatique de termes MeSH. Afin d'enrichir une terminologie médicale francophone, nous avons proposé et mis en oeuvre une méthode de traduction automatique de termes MeSH américains à l'aide de corpus parallèles du domaine. Nous avons pu ajouter dans un premier temps 133 synonymes à la terminologie CISMeF. La méthode d'extraction des traductions par propagation des liens d'équivalence à l'aide des relations syntaxiques offre une précision globale de 70% pour un rappel de 57%. Une analyse plus fine des résultats montre que ces performances sont optimales pour la traduction de termes de fréquence moyenne (autour de 5 occurrences). Pour les termes de basse fréquence, le rappel est faible et inversement, pour les termes très fréquents, la précision chute.

Afin d'exploiter pleinement nos corpus de travail et d'élargir la couverture de notre méthode, nous avons étendu notre approche à la traduction compositionnelle des termes absents des corpus mais dont les composants se trouvaient dans les corpus. Nous avons pu ajouter 91 synonymes supplémentaires à la terminologie CISMeF, issus de l'échantillon de 200 traductions validées. La méthode de traduction compositionnelle offre une précision globale de 73% sur cet échantillon et ce, moyennant une sélection des traductions des composants au rang 1. L'application du principe de compositionnalité et de correspondance structurelle entre les termes s'avère par conséquent pertinente pour la traduction en français des synonymes MeSH américains de structure *Adj N*. Une analyse plus fine des résultats montre que si, de manière globale, elle améliore considérablement la précision, l'application du critère de sélection au rang 1 s'avère parfois trop contraignante. En effet, dans certains cas elle ne permet pas de proposer la bonne traduction alors que celle-ci peut être trouvée en tenant compte des équivalents de rang inférieur ; dans d'autres, elle ne permet de retenir qu'une seule traduction là où plusieurs sont possibles et seraient susceptibles d'être retenues dans la terminologie. L'utilisation des tous les candidats pour la traduction compositionnelle impliquerait cependant une multiplication des termes à valider. Pour cette raison, nous envisageons un filtrage automatique des candidats à l'aide du corpus constitué par l'Internet[39]. En effet, les termes pour lesquels aucune ocurrence n'est trouvée sur l'Internet sont probablement erronés ou peu usités. Il est donc envisageable de les rejeter automatiquement. Par contre, il semble peu prudent de valider automatiquement les termes pour lesquels il est possible de trouver un minimum d'occurrences sur l'Internet.

Par ailleurs, aussi bien pour la traduction directe que compositionnelle, il apparaît également que l'enrichissement de la terminologie ne relève pas d'une simple traduction des synonymes. Il est impératif de tenir compte de la complexité lexicale et de l'usage des termes dans les deux langues afin de valider l'ajout de synonymes correctement traduits.

Pour la poursuite de ces travaux, nous envisageons d'affiner les critères de sélection des traductions des composants mais aussi d'étendre la traduction compositionnelle aux termes de structure autre que *Adj N*, qu'il s'agisse de termes constitués de deux mots pleins ou plus. Enfin, nous souhaitons mettre nos techniques d'extraction automatique des traductions

[39]Par exemple, à l'aide de l'API Google ou Yahoo !

à l'épreuve d'un nouveau corpus : le corpus parallèle CESART.

2.6 Conclusion sur la terminologie médicale

Nous avons présenté les deux systèmes de représentation des connaissances les plus courants dans le domaine de la médecine : les terminologies et les ontologies. Alors que les ontologies sont centrées sur *les concepts*, les terminologies ont pour objet *les termes*, qui en sont la verbalisation. En dépit de cette opposition apparente, les deux systèmes s'intéressent néanmoins également aux relations entre les termes et les concepts. De plus il apparait que pour être utilisés efficacement, terminologies et ontologies doivent être construites pour répondre à un besoin précis des acteurs du domaine de spécialité concerné. Ainsi, nous avons présenté plusieurs terminologies du domaine médical, destinées au codage médico-économique, à l'indexation des dossiers patients et à la recherche d'information en santé. A travers une revue succinte des travaux en traitement de la langue médicale nous avons illustré les utilisations de ces terminologies, et les améliorations souhaitables. Finalement, nous avons présenté notre propre contribution, principalement destinée à l'indexation de ressources de santé à l'aide du MeSH, présentée au chapitre 5. Les dictionnaires MeSH et la bibliothèque de transducteurs obtenus permettent une couverture du MeSH de 83%. Ils sont également enrichis par la traduction automatique de synonymes américains du MeSH, pour laquelle nous avons pu proposer plusieurs solutions.

Chapitre 3

Veille documentaire

La veille documentaire constitue le premier maillon de la chaîne de traitement de l'information de santé francophone mise en place au sein du catalogue CISMeF. Après avoir défini et situé cette activité documentaire dans un contexte global, nous décrirons plus précisément le processus de veille mis en place dans CISMeF et son automatisation partielle.

3.1 Définitions

La Veille documentaire [1] consiste en une collecte de documents issus de diverses sources (bases de données ou revues spécialisées etc.), permettant de rassembler des informations sur un domaine précis. Elle s'appuie sur des techniques de recherche documentaire appelées « cueillette active » (pull, en anglais « tirer ») et « cueillette passive » (push, en anglais « pousser ») afin d'assurer la mise à jour des informations récoltées.

La cueillette active, ou pull consiste en une surveillance régulière des sources d'informations connues pour le domaine concerné (par exemple, des revues spécialisées), afin d'en « tirer » les éléments les plus récents et enrichir ainsi le fond documentaire déjà disponible.

La cueillette passive, ou push consiste à recueillir des données qui ont été « poussées » vers le veilleur par un tiers en fonction de critères spécifiés au préalable (par exemple, l'annonce de publications de la part d'un spécialiste). On parle aussi de *diffusion sélective de l'information*.

Ainsi, il apparaît que la veille va en fait plus loin que la simple collecte d'information : il s'agit d'une collecte orientée par un domaine d'application et par le profil des utilisateurs avérés ou potentiels. Elle doit donc s'accompagner d'une sélection des documents collectés et d'une re-direction des flux d'information sélectionnés.

3.2 L'activité de veille documentaire

La veille documentaire a longtemps été le fait des bibliothécaires. La loi de Bradford, qui fut publiée pour la première fois en 1934 (Bradord, 1934), met en évidence la préoccupation ancienne des professionnels de l'information de réaliser une sélection de documents - donc

[1] Ce paragraphe a été rédigé à l'aide des notes de cours de l'université de Montréal (Michaud & Waller, 2005)

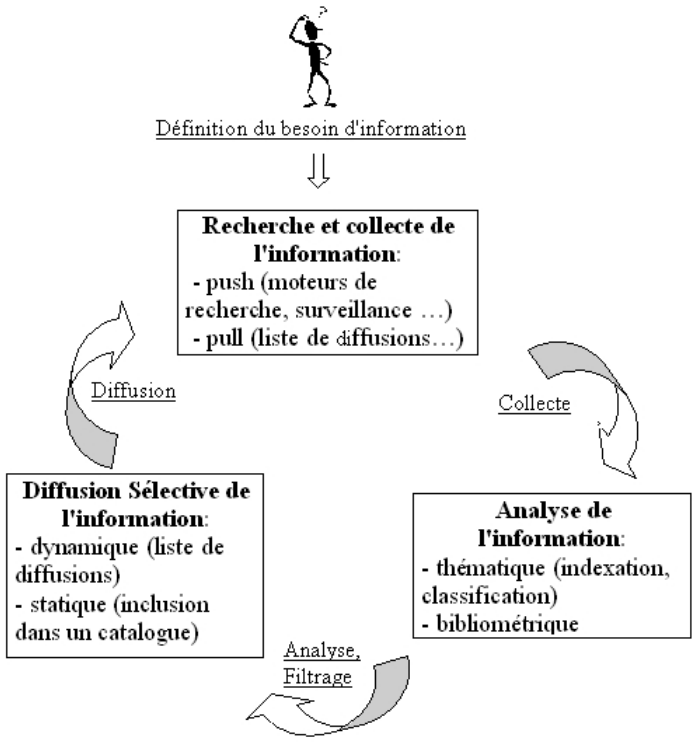

FIG. 3.1 – Le Cycle de la Veille

une veille - efficace. La loi de Bradford permet d'identifier des « revues de référence » pour chaque spécialité en analysant la répartition des articles d'un domaine donné dans les journaux scientifiques. Les documentalistes peuvent ainsi cibler les sources d'informations les plus pertinentes et les plus complètes pour chaque domaine.

L'activité de veille documentaire implique une collecte active ou passive, puis une analyse de l'information, qui sera ensuite diffusée sélectivement[2] aux utilisateurs des services de documentation (Thomas, 2004; François, 2003). Toutes ces étapes (illustrées par la fig. 3.1) apportent une valeur ajoutée à l'information, qui peut ensuite être pleinement exploitée par l'utilisateur. Ainsi, récemment, le développement de plus en plus rapide des nouvelles technologies et la recherche de compétitivité face à la concurrence font de l'information un facteur stratégique décisif dans tous les secteurs d'activité de la recherche ou de l'industrie. On assiste à un nouvel essor de la veille documentaire sous la forme de « veille stratégique » ou « veille concurrentielle » qui sont alors liées au contexte concurrentiel de l'entreprise[3].

[2]La sélection peut se faire à deux niveaux : au niveau de la collection (il s'agit alors d'un filtrage des informations collectées avant leur inclusion dans la collection) ou au niveau des utilisateurs (il s'agit alors de la prise en compte des préférences ou des souhaits exprimés par les utilisateurs quant aux informations qui les intéressent.

[3]Pour plus de détails sur ces types de veille, nous invitons le lecteur à consulter le document de La Maison des Sciences de l'Homme-Alpes (UMS 1799 du CNRS) `http://www.msh-alpes.prd.fr/veille/defveille.htm`, visité le 28/02/05

3.3 Les techniques et outils pour la Veille Documentaire

A travers une étude de cas fictive, Thomas (2004) présente un panel d'outils pouvant être utilisés pour la veille documentaire. L'auteur distingue notamment les outils commerciaux de veille intégrée (par exemple, Arisem de la société Thales[4], ou AMI « Market Intelligence » de la société Albert[5]) ou des outils non nécessairement spécifiques à la veille, mais pouvant être utilisés dans les diverses étapes du cycle de veille et intégrés pour constituer un outil dédié. Il faut en effet souligner que les étapes du cycle de veille (cf. figure 3.1) peuvent s'apparenter à d'autres activités du traitement de l'information et de la langue naturelle. La définition du besoin d'information est une formalisation conceptuelle qui peut être mise en oeuvre par la formulation de requêtes. La collecte d'information relève de la *fouille de données* ou de la *recherche d'information*, soit sur l'Internet soit dans d'autres bases documentaires. La sélection ou le filtrage d'information relève de la *catégorisation* ou de la *classification de documents*. L'analyse des documents collectés peut consister en une *indexation* ou une *annotation* des ressources à l'aide de mots clés... Ainsi, la « veille documentaire » peut englober un large spectre de tâches documentaires. Nous en faisons une revue rapide ci-dessous.

3.3.1 Collecte de l'information

La collecte de l'information relève à la fois de la recherche d'information et de la surveillance (active ou passive) de sources d'information connues. Ainsi, des outils de recherches classiques tels que les moteurs de recherche d'information et leurs interfaces dédiées[6] peuvent être utilisées. Les progrès récents dans le domaine de l'Intelligence Artificielle Distribuée, notamment avec les Systèmes Multi-Agents (SMA) ont permis d'avancer dans la voie d'une veille automatisée. Un SMA est composé d'agents autonomes capables de communiquer entre eux, d'interagir avec leur environnement pour résoudre la tâche qui leur est assignée et de s'adapter aux changements dudit environnement. Les agents sont également en mesure d'apprendre à partir de leur propre expérience, c'est-à-dire de leurs interactions passées avec l'environnement et les autres agents du système.

D'après (Sycara, 1998)[7] les principales caractéristiques d'un agent sont le caractère situé (interaction entre environnement et agent), l'autonomie, l'adaptabilité et la sociabilité. Ferber (1995), quant à lui donne une définition détaillée des SMA et des agents qui les composent, sans exclure les agents humains. Retenons qu'il existe deux types d'agents : les agents cognitifs et les agents réactifs. Les agents cognitifs sont des agents dits « intelligents ». Ils ont une représentation de leur environnement, ils sont capables de planifier leur comportement, de communiquer par l'intermédiaire de messages, de négocier etc. Les agents réactifs sont moins complexes. Contrairement aux agents cognitifs, ils n'ont aucune représentation de leur environnement, ils fonctionnent sur le principe stimuli/action et ne possèdent pas d'autre forme de communication. Cependant, ils sont plus simples à mettre en oeuvre, et se révèlent adaptés à certaines tâches de la veille documentaire sur Internet, comme la diffusion d'information

[4]La page d'information de la société Thales sur Arisem (parfois appelé « DigOut4U ») est disponible à l'URL : `http://www.arisem.com/fr/arisem/` (accédé le 02/07/05).

[5]Une documentation sur la version 3.0 du logiciel AMI Market Intelligence est disponible sur le site de la société Albert à l'URL : `http://www.albert.com/documentation.html` (accédé le 30/08/05).

[6]Par exemple, les API Google (`http://www.google.com/apis/` - accédé le 02/04/05) et Yahoo ! (`http://developer.yahoo.net/search/` - accédé le 02/08/05) permettent une utilisation automatique de ces outils, qui peuvent par exemple être intégrés à un système de veille.

[7]Cité par (Pauchet, 2005)

aux abonnés d'une liste, ou l'alerte d'un utilisateur lors de mouvement sur une page web surveillée.

3.3.2 Analyse de l'information

L'analyse des documents collectés peut porter sur de nombreux aspects aussi bien relatifs au contenu qu'à des éléments connexes à l'information véhiculée par le document. En ce qui concerne le contenu, l'analyse de document vise à en extraire les concepts caractéristiques, soit sous forme d'indexation par mot clés ou texte libre (cf. section 5), de résumé automatique de texte (Minel, 2004), de catégorisation de document (cf. section 4) ou de clustering (Lelu & Ferhan, 1998), éventuellement illustré par une cartographie des thèmes traités (Lelu, Halleb, & Delprat, 1998; Roy, 2005). Outre la loi de Bradford (Bradord, 1934) évoquée plus haut, les techniques de bibliométrie permettent une analyse orientée sur les acteurs (étude des auteurs (Lotka, 1960) et de leur affiliation) sur les supports de communication (étude des types de document) et sur l'évolution de ces éléments dans le temps (étude des dates de publication). La bibliométrie s'intéresse également à la répartition des concepts dans les documents, sous forme de mots (Zipf, 1935). Afin d'étendre la portée de la veille, d'autres fonctionnalités comme la traduction des résultats de l'analyse (proposée par exemple par Arisem) peuvent être envisagées.

Ces différentes analyses possibles d'un document peuvent servir de fondement à la mise en place de grilles de qualité. Les résultats de l'analyse de l'information permettent d'envisager sa diffusion sélective : ils peuvent être utilisés pour filtrer les informations collectées aussi bien en fonction de critères qualitatifs établis par ces référentiels, que de critères personnels choisis par les utilisateurs.

3.3.3 Diffusion de l'information

La diffusion de l'information peut-être rapprochée de la collecte d'information, car elle représente un flux d'information similaire, entre auteur ou documentaliste vers documentaliste ou utilisateur. Ainsi, les outils évoqués plus haut tels que les agents conviennent également à cette étape, dans le cas d'une diffusion « dynamique » (push) où l'information doit être fournie directement à un utilisateur connu. Une méthode de diffusion « statique » consiste à recenser les documents collectés et sélectionnés, en indiquant les résultats des analyses qui en ont été faites. Il s'agit de la constitution d'un fond documentaire semblable à un catalogue (Darmoni et al., 2000).

Afin de faire le point sur les outils de veille existant pour chacune des étapes du cycle comme pour l'ensemble de l'activité (outils intégrés), l'Institut National de l'Information Scientifique et Technique (INIST) a initié un projet national de recensement et d'évaluation des outils de veille[8]. Malgré l'émergence de nombreuses solutions technologiques pour la veille, Thomas (2004) insiste sur le rôle central de l'expert du domaine dans le processus de veille, qu'il est à l'heure actuelle impossible d'automatiser entièrement.

3.4 Veille documentaire CISMeF

Dans le cadre de CISMeF, nous entendons par *veille* la recherche sur l'Internet et la sélection de ressources électroniques susceptibles d'être intégrées au catalogue. Nous expose-

[8]Site officiel du projet : `http://veille-srv.inist.fr/~OutilsVeille/Public/` (accédé le 02/08/05).

rons la méthode manuelle de recherche des ressources utilisée entre 1995 et 2002, puis nous détaillerons les critères de sélection des ressources et enfin nous présenterons les solutions informatiques mises en oeuvre pour automatiser une partie de ces étapes.

3.4.1 Recherche de ressources

La recherche de ressources à intégrer dans le catalogue met en oeuvre les techniques de cueillette active et passive décrites ci-dessus en 3.1. CISMeF recense des documents électroniques en ligne. L'outil de prédilection pour effectuer la veille documentaire est donc l'Internet, mais il n'est pas utilisé de façon exclusive. L'analyse de certains journaux papiers est également effectuée systématiquement, comme par exemple le numéro « Spécial Informatique » du *Quotidien du Médecin*.

La cueillette passive : listes de diffusion, appel à ressources sur la page d'accueil CISMeF.

La documentaliste de l'équipe CISMeF chargée de la veille stratégique est inscrite à une dizaine de listes de diffusion signalant par courrier électronique aux abonnés la publication de nouvelles ressources sur certains sites. Les publications signalées par ces listes portent soit sur des thèmes variés comme le droit, la santé, l'éducation, la culture, (par exemple, les listes de la documentation française, du sénat ou du service public) soit sont ciblées sur la santé (par exemple, « infectioliste »). Les courriels envoyés par les listes de diffusion contiennent généralement une phrase d'introduction (ex : « Dernières publications sur le site de l'Assemblée Nationale ») ainsi qu'une liste plus ou moins structurée des documents en question, comprenant les informations suivantes :

- Titre du document
- URL et lien vers le document
- Brève description du document et/ou de son contenu

D'autres ressources sont soumises pour ajout au catalogue par leurs auteurs ou par des utilisateurs réguliers de CISMeF. Dans ce cas, l'informateur est invité à remplir un formulaire (accessible depuis la page d'accueil[9]) contenant des champs fondés sur les métadonnées du DublinCore similaires à ceux renseignés lors de la constitution d'une notice : titre de la ressource, auteur, public ciblé, URL, spécialité médicale concernée, indexation MeSH etc. Ces informations peuvent guider les documentalistes lors de la sélection de la ressource et dans l'éventualité d'une sélection, les éléments renseignés pourront être repris (et validés) pour la constitution de la notice.

La cueillette active : visite régulière de sites. Une liste de sites publiant ou référençant régulièrement des documents relatifs à la santé a été construite, enrichie et mise à jour par l'équipe CISMeF depuis la création du catalogue en 1995. Ces sites sont classés en fonction de la fréquence à laquelle ils sont mis à jour (quotidienne, hebdomadaire, mensuelle, annuelle) et en conséquence visités un fois par jour, semaine, mois ou année. Les sites visités sont soit des catalogues généralistes (ex : Nomade, Yahoo) soit des sites spécifiques au domaine de la santé (ex : Santé Canada, ANAES). Quel que soit leur domaine d'action, ils ont en commun la particularité d'annoncer sur une page « nouveauté » spécifique les sites ou documents nouvellement publiés ou référencés, avec les liens correspondants. Une fois sur la page "nouveauté", la démarche est sensiblement la même pour chaque site : les documents non consultés précédemment sont identifiés

[9] Accessible sur `http://cismef.chu-rouen.fr/general/formulaire.html`- accédé le 20/11/03

grâce à une mémoire du site lors du précédent passage (sous forme d'impression datée de la page dans le cadre de la veille manuelle). Des pages légèrement différentes, comme la rubrique Internet de F. Eveillard[10] sont également surveillées.
Les ressources récoltées par le biais de ces différentes méthodes pouvant être redondantes, il convient de vérifier qu'un document n'a pas déjà été sélectionné ou encore qu'il n'est pas déjà répertorié dans la base CISMeF. Se pose ensuite le problème de la sélection des documents relatifs à la santé (pour les sites ou listes généralistes) et parmi eux, les documents pertinents pour CISMeF.

3.4.2 Critères de sélection des documents indexés dans CISMeF

Les cueillettes active et passive fournissent un grand nombre de ressources d'intérêt inégal et il convient de les trier afin de sélectionner celles qui ont réellement leur place dans le catalogue. Cette sélection constitue une direction de recherche majeure des équipes CISMeF et GCSIS. Tout d'abord, les ressources sélectionnées doivent répondre à l'objectif global du projet CISMeF et traiter de questions à l'usage des patients, des étudiants en médecine ou des professionnels de la santé. Il s'agira par exemple, d'une question d'internat pour les étudiants (les ressources liées à l'Examen National Classant sont recensées depuis 2004), d'une recommandation pour la bonne pratique clinique pour les médecins, d'un document de vulgarisation sur le dépistage et le traitement d'une maladie pour les patients.

Pour les ressources destinées à ces trois types de public, les critères de sélection s'appuient principalement sur la **source** et la **qualité** de la ressource. Ainsi, tout site ou document à caractère commercial (site de société pharmaceutique ou autre organisme de vente, site contenant de la publicité...) est rejeté, quel que soit son contenu.

Sensible à l'indication de la qualité des documents disponibles sur l'Internet, l'équipe CISMeF a participé à la mise au point d'une grille d'évaluation à cet usage (Darmoni et al., 1999). Au total, 49 critères entrent en ligne de compte dans le Net Scoring[11]. Chaque critère comporte une pondération propre pour un score maximum de 312 points. Pour des raisons pratiques, l'évaluation d'une ressource avec l'ensemble de ces critères est impossible dans le contexte de CISMeF : une demi-journée de travail est nécessaire pour établir le score d'une ressource. Les critères de qualité du Net Scoring entrent cependant en compte dans la sélection des ressources. Une attention particulière est portée à la mention explicite du nom des auteurs, des éditeurs, ainsi que les dates de publication et de mise à jour des ressources. Ces critères montrent que, comme d'autres grilles de score, le Net Scoring évalue le *contenant* plus que le *contenu*. Afin de compléter cette approche pour les documents « sensibles » fondés sur la médecine factuelle, c'est-à-dire traitant de diagnostique ou de thérapeutique, l'équipe CISMeF a retenu un critère majeur dénotant la qualité du contenu. Il s'agit de l'indication du niveau de preuve[12] selon la définition de la FNCLCC (Fédération Nationale des Centres de Lutte Contre le Cancer). De plus, comme il existe plus de vingt méthodes dans le monde pour déterminer le niveau de preuve (Darmoni, Amsallem, et al., 2003), la méthode utilisée dans chaque ressource est précisée.

[10] La Revue du Praticien, `http://www.33doc.com` accédé le 06/07/05

[11] Le référentiel complet des critères de qualité de l'information de santé sur l'Internet (Net Scoring) est disponible sur `http://www.chu-rouen.fr/netscoring`

[12] Pour consulter la définition du niveau de preuve de la FNCLCC, nous invitons le lecteur à consulter la page « Evidence-based Medicine » de CISMeF à l'URL : `http://www.chu-rouen.fr/ssf/profes/evidencebasedmedicine.html` (visité le 01/03/05).

Les documents retenus présentent de préférence un caractère institutionnel et de manière générale, les documents émanant des sites gouvernementaux (ministère de la santé, ministère de la justice, sénat etc.) des universités de médecine, des hôpitaux, des agences nationales reconnues dans le domaine médical (ANAES, InVS etc.), des sociétés savantes en médecine (généralement, une société est retenue par spécialité médicale - par exemple la Société Française de Radiologie (SFR) pour l'imagerie médicale) sont considérés comme fiables. Cependant, les documents éphémères comme les projets de lois ne sont pas retenus. En effet, si la loi n'est pas votée, ce document n'est pas réellement d'actualité et si elle l'est, le texte de loi correspondant sera indexé.

Par ailleurs, des sites d'associations et quelques sites personnels peuvent être retenus. Il s'agit souvent de sites mis en ligne par des patients qui peuvent faire bénéficier d'autres patients de leur expérience. Ils donnent des informations sur la vie au quotidien avec la maladie, proposent des forums de discussion sur ce sujet, signalent l'existence de structures de soutien pour les patients ou leur entourage etc. Avant toute sélection, il convient de vérifier que les ressources ne comportent pas de publicité et qu'elles ne sont pas affiliées à des organismes commerciaux.

La dimension « en ligne » des documents référencés soulève la question de la maintenance des liens dans la durée. Ce problème a été abordé récemment pour la références aux publications en ligne par Ho (2005). L'étude des liens contenus dans les articles de trois revues scientifiques en ligne montre qu'un nombre important de liens sont brisés et que les articles les plus anciens contiennent d'autant plus de liens brisés. Aucune modification ou stockage des documents originaux n'est effectuée dans CISMeF, afin de permettre aux utilisateurs de visualiser des ressources à jour. Une fiche descriptive contenant un lien électronique est créée pour chaque ressource. Il est donc essentiel pour la maintenance du catalogue que les ressources ainsi pointées puissent être trouvées à une URL stable sur la Toile. Une variabilité trop importante des URLs ou des contenus peut être un critère discriminant, même pour des ressources dont le contenu vérifie l'ensemble des critères de qualité précédemment énoncés.

3.4.3 Formalisation de la procédure de sélection

Dans l'optique d'une automatisation de la veille (recherche puis sélection de ressources) il convient de formaliser la procédure décrite ci-dessus. Pour résumer l'étape de recherche, en ce qui concerne la cueillette active, il s'agit de surveiller les flux d'informations de sites présélectionnés à intervalle régulier. Les *nouvelles ressources*[13] ainsi détectées viennent s'ajouter aux ressources fournies par la cueillette passive.

La sélection des ressources s'effectue selon trois types de critères :

1. Critère suffisant : toute ressource satisfaisant à un tel critère est sélectionnée.
   - Documents publiés par l'ANAES, l'OMS ou un organisme partenaire de CISMeF
   - Directives révisées et énoncés de principe
   - Rapports et travaux de l'ordre des médecins
   - Avis du comité consultatif d'éthique
2. Critère incompatible : toute ressource satisfaisant à un tel critère est rejetée.
   - Ressource traitant de sujets autres que santé

[13]Par « nouvelles ressources » nous entendons ici les ressources qui ne figurent pas dans le catalogue. En effet, une ressource rédigée en 2000 (par exemple un cours) peut être mise en ligne en 2004 et être intéressante pour le catalogue bien qu'elle ne soit pas *récente*.

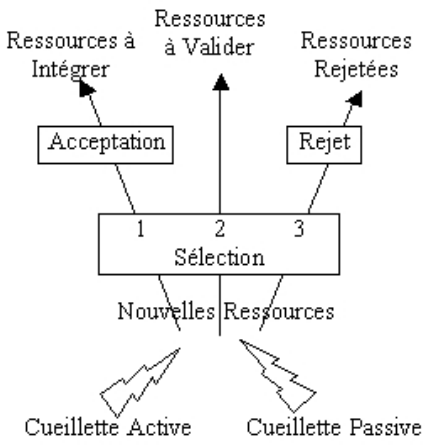

FIG. 3.2 – Sélection des ressources à intégrer dans CISMeF

- Ressource en langue autre que français
- Ressource contenant des bandeaux publicitaires
- L'un des types de ressource suivants : Projet de loi, Communiqué de presse, Discours, Campagne du ministère de la santé.

3. Critère compatible : toute ressource satisfaisant à tel critère est susceptible d'être sélectionnée.
   - Site personnel ou association de patients
   - Cours d'une école/faculté de médecine
   - Ressource traitant d'une question de santé

Dans le cadre d'une sélection automatique, l'identification des deux premiers types de critères (suffisant et incompatible) permet de traiter complètement une ressource. Si seul le critère compatible est satisfait, une validation manuelle devra compléter l'analyse automatique (cf. Figure 3.2).

Il faut remarquer que les noms de domaines ne sont pas des critères significatifs de sélection. En effet, si les sites ".com" ont souvent un caractère commercial, certaines associations de patients sont hébergées sous cette extension. De même, les sites « .org » - en principe institutionnels - peuvent cacher des sites commerciaux.

3.4.4 Automatisation de la Veille

Comme l'indiquent (Foulonneau, 1999)[14] et (Thomas, 2004) il existe de nombreux outils commerciaux pour la veille documentaire et stratégique. Cependant, leur coût élevé et la particularité de notre procédure nous ont conduit à développer un nouvel outil, le Veilleur Automatique CISMeF (CVA). CVA a été initialement développé par B. Dahamna, l'ingénieur de l'équipe CISMeF, puis amélioré par trois étudiants de l'INSA de Rouen (département ASI) encadrés par A. Rogozan. Mon rôle a été de définir les spécifications de cet outil après une analyse de la méthode manuelle de veille mise en oeuvre au sein de CISMeF.

[14]Rapport de l'enssib sur les logiciels de veille disponible sur `http://www.enssib.fr/autres-sites/dessid/dessid99/gedfoulo.pdf` (visité le 01/03/05)

CVA peut-être considéré comme un *agent réactif* dans la mesure où, une fois activé c'est-à-dire, après réception d'un **stimulus** :

- De manière **indépendante**, il explore récursivement les liens contenus dans les pages listées dans la base de donnée
- De manière **autonome**, il traite les nouveaux liens rencontrés qui sont soit signalés à l'utilisateur (veilleur humain), soit ignorés
- en fonction des réponses de l'utilisateur lors de la découverte de liens, la base de donnée reliée à CVA est mise à jour et CVA **s'adapte** à ces mises à jour pour continuer son exploration.

Les principales fonctions de l'outil CVA ont été développées en Java (JSDK 1.4) par B. Dahamna. Il s'agit de la création et de la modification de trois tables d'URLs (à explorer/à éviter/à valider) en environnement Oracle, et de l'écriture du spider : parcours de la liste d'URLs à explorer et extraction des liens contenus dans les pages pointées jusqu'à une profondeur fixée, filtrage des résultats à l'aide de la liste d'URLs à éviter, présentation de la liste résultante des URLs à valider, avec possibilité de les marquer comme « à éviter ». Des améliorations supplémentaires ont été apportées par les étudiants de l'INSA, principalement au niveau de l'interface du logiciel développée à l'aide des librairies Java Swing et AWT. Ainsi, la création d'onglets permet d'accéder séparément aux URLs concernant la veille quotidienne, hebdomadaire et mensuelle. Des fonctionnalités ergonomiques permettent de suivre les étapes des traitements lancés (traitement en cours ou terminé) ou de faire fonctionner CVA en tâche de fond. Les autres améliorations ont porté sur la détection de certains liens précédemment ignorés (liens shtml et liens affichés par des servlets), ainsi que sur la possibilité de définir une profondeur d'exploration spécifique à chaque URL, au lieu d'une profondeur globale pour l'ensemble des sites. Nous reprenons ci-dessous plus en détail le fonctionnement de l'outil.

Utilisé par la documentaliste de l'équipe CISMeF chargée de la Veille, CVA comporte une interface conviviale à base d'onglets (cf. Figures 3.3-3.4). L'ajout d'une URL dans le système (Figure 3.3) s'accompagne d'un classement en fonction de la fréquence à laquelle la veille doit y être accomplie (hebdomadaire, quotidienne, mensuelle). CVA permet de définir pour chaque URL à surveiller la profondeur à laquelle l'exploration doit se poursuivre (profondeur de 0 - nouvelle URL située dans la page courante à 3 - nouvelle URL située dans une page atteinte en suivant jusqu'à trois liens).

La base de donnée avec laquelle CVA interagit contient trois listes d'URLs : les URLs à explorer (listées dans les onglets « quotidienne », « hebdomadaire »etc.), les URLs à éviter (qui peuvent contenir des liens nouveaux, mais qu'il n'est pas nécessaire d'explorer car les ressources correspondantes satisfont à un critère *incompatible* et ne pourront être retenues), les URLs connues (qu'il n'est pas nécessaire d'explorer). Ces deux derniers types d'URLs apparaissent dans l'onglet « Sites à éviter ou connus », où elles peuvent être consultées et modifiées.

Lors du *lancement* de la veille (fonctionnalité présente dans l'onglet « lancement », illustré par la figure 3.4) CVA recherche les liens vers de « nouvelles »URLs qui sont présents dans les pages correspondant aux URLs sélectionnées préalablement dans les onglets « quotidienne », « hebdomadaire »etc. Ainsi il est possible d'effectuer aussi bien une veille ciblée sur une seule URL qu'une veille globale sur l'ensemble des URLs à surveiller. L'affichage des résultats permet de vérifier quelles sont les URLs traitées : dans le cadre « liens trouvés », CVA affiche les caractéristiques des URLs traitées. Par exemple, `http://www.sante.gouv.fr/drees/etude-resultat/doc.html` est une URL surveillée avec une profondeur 0. La ligne suivante correspond à une autre URL surveillée, ce qui signifie qu'aucun nouveau lien n'a été trouvé

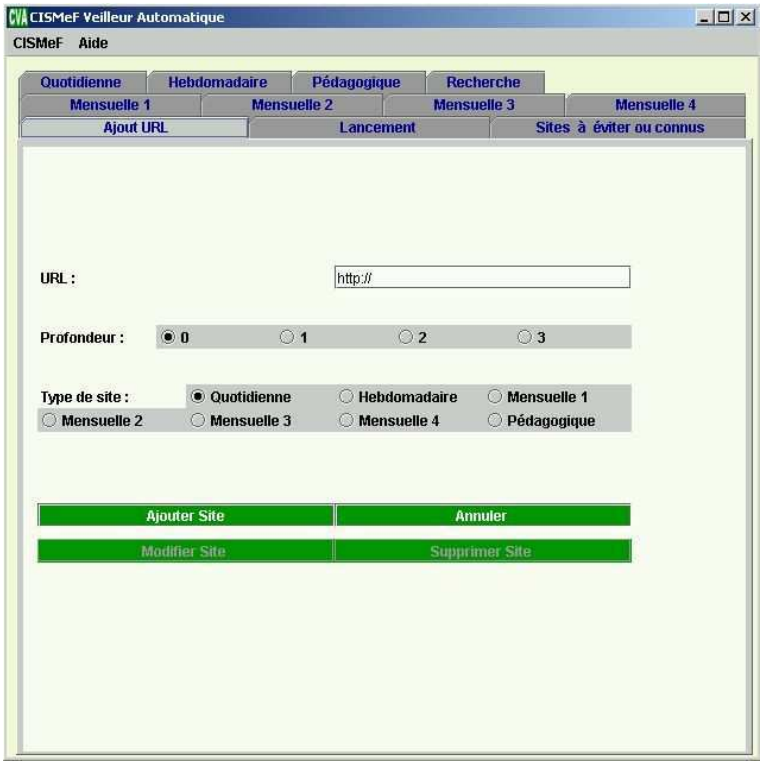

FIG. 3.3 – Copie écran de CVA : ajout d'une URL

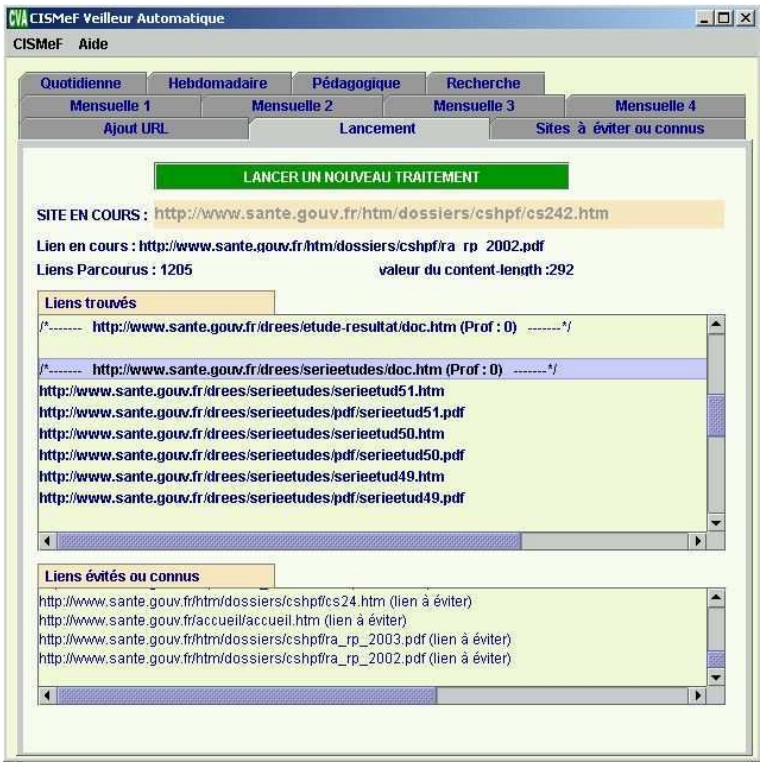

FIG. 3.4 – Copie écran de CVA : veille hebdomadaire du 07/07/05.

sur la page `http://www.sante.gouv.fr/drees/etude-resultat/doc.htm`. Par contre, on peut observer qu'une série de nouveaux liens correspondant à des documents htm et pdf a été extraite de la page `http://www.sante.gouv.fr/drees/serieetudes/doc.htm`. Le cadre suivant, « Liens évités ou connus » signale les URLs à éviter et les URLs déjà contenues dans la base qui ont été rencontrées lors du traitement.

Les liens vers de nouvelles URLs ainsi repérés par l'intermédiaire de CVA sont ensuite validés manuellement par les documentalistes de l'équipe CISMeF et ajoutés (manuellement) dans la liste des URLs « en attente »d'être indexées.

Par ailleurs, dans le cadre d'une collaboration entre CISMeF et l'INCa (Institut National du Cancer), la veille effectuée grâce à CVA est complétée par l'outil commercial « Ami-Veille » pour les ressources de cancérologie. « Ami-Veille » est configuré spécifiquement par les documentalistes de l'INCa pour signaler les ressources électroniques traitant de thèmes liés au cancer. Ces ressources sont ensuite indexées comme les autres ressources référencées dans CISMeF et rassemblées dans un portail dédié au Cancer : KISMeF[15].

3.5 Conclusion

Ainsi, CVA permet de répondre de manière satisfaisante au problème de la recherche de nouvelles ressources à inclure dans le catalogue et dans une certaine mesure procède au filtrage de certaines ressources non conformes aux critères de qualité CISMeF (grâce à la liste des

[15] cf. `http://doccismef.chu-rouen.fr/servlets/KISMeF` (accédé le 30/08/05)

URLs « à éviter »). Pour les cas où ce filtrage n'est pas possible, il faut dans un premier temps opérer une sélection des ressources traitant du domaine de la santé. Ce problème s'apparente à la classification thématique, dont nous traitons au chapitre 4.

CVA est utilisé quotidiennement par la documentaliste de l'équipe CISMeF chargée de la veille.

Chapitre 4

Classification de documents

Dans le processus de traitement de l'information de santé CISMeF, la classification de documents peut intervenir à deux niveaux. D'une part, comme nous venons de le voir dans le cadre de la veille documentaire, une méthode de classification serait utile pour reconnaître automatiquement les ressources traitant de santé des ressources traitant d'autres thèmes. Cependant, la classification intervient principalement dans l'étape qui suit la veille documentaire, à savoir la constitution de notices. Après avoir été sélectionnées, les ressources de santé sont décrites afin de faciliter la recherche d'information au sein du catalogue. La classification permet une description générale[1] du contenu de la ressource. Nous définissons ci-dessous l'activité de classification telle que mise en oeuvre pour la description des ressources au sein des notices longues CISMeF. Nous présentons ensuite les deux approches utilisées et nous les comparons à celles de la littérature.

4.1 Définition et objectifs

4.1.1 Définition

> La **classification**[2] ou *catégorisation* consiste à répartir les objets d'un univers entre (au moins) deux *classes* ou *catégories*.

Cette organisation de l'univers en groupes d'objets similaires peut être fondée sur des catégories pré-définies. On parle alors de catégorisation supervisée, ou de classification. Dans le cas où les catégories dans lesquelles les objets à classer doivent être répartis ne sont pas connues *a priori*, on parle de catégorisation non supervisée, ou encore de clustering (Lelu & Ferhan, 1998)

Nous nous intéressons ici à la catégorisation supervisée des ressources du catalogue CISMeF, soit des ressources essentiellement textuelles. Un des objectifs de la catégorisation textuelle est de classer les ressources en fonction de leur thème (d'autres classes possibles sont : la langue dans laquelle elles sont rédigées, le genre littéraire, ou bien l'identité de l'auteur, ...). Les catégories utilisées par le catalogue Yahoo![3] sont un exemple représentatif de découpage thématique supervisé : « finance », « santé », « loisirs »etc. Comme nous l'avons exposé dans le chapitre précédent (3), les ressources CISMeF traitent de questions relatives à la

[1] c'est-à-dire à un niveau de granularité moins fin que l'indexation MeSH.
[2] D'après (Manning & Shütze, 2000)
[3] `http://www.yahoo.fr`

santé. Nous devons donc utiliser une catégorisation plus fine afin de déterminer les spécialités médicales auxquelles se rapportent les ressources. Un tel découpage du domaine médical en sous-domaines de spécialité a été effectué avec l'élaboration de la liste des métatermes décrits au chapitre 2, soit 83 spécialités en 2003, au début de ce travail. En 2005, le nombre de métatermes est 126. La tâche de catégorisation que nous nous fixons dans le cadre du catalogue CISMeF utilise une centaine de classes représentant des spécialités médicales bien définies.

Il s'agit donc de catégorisation supervisée.

4.1.2 Classification dans le cadre de CISMeF : Objectifs

La classification figure dans la description des ressources présentée par les notices longues CISMeF (décrites en 1.2.3), au même titre que l'indexation MeSH. Classification et indexation offrent une représentation complémentaire du contenu des ressources dans la mesure où l'indexation a pour objet une représentation fine des concepts abordés, et la classification a pour objet une représentation globale des thèmes traités. Il semble alors évident que ces deux représentations ne sont pas indépendantes et que l'une peut servir de base à l'autre :

- Si la classification intervient avant l'indexation, elle pourra être utilisée comme connaissance contextuelle pour désambigüer certaines expressions : par exemple, en *cardiologie* l'acronyme IVG sera interprété comme « Intervention Ventricule Gauche » alors qu'en *chirurgie* il s'agira d'une « Interruption Volontaire de Grossesse ».
- Si la classification intervient après l'indexation, elle pourra dans une certaine mesure être déduite des mots clés ou paires attribués à la ressource : par exemple, on pourra inférer qu'une ressource indexée à l'aide du mot clé <*diarrhée nourrisson*> traite de *pédiatrie*.

Nous avons envisagé ces deux approches et proposé différentes méthodes permettant d'effectuer la classification *avant* (Rogozan, Névéol, & Darmoni, 2003) ou *après* (Névéol, Soualmia, et al., 2004) l'indexation.

4.2 Méthodes de classification

4.2.1 Vue d'ensemble

Dans le domaine de la classification textuelle, on se propose de résoudre le problème suivant :

> Etant donné un texte T, un ensemble de classes C_k (k fixé) quelle(s) est(sont) la(les) classe(s) qu'il est pertinent d'associer à T ?

Dans la définition du problème, les paramètres importants à prendre en compte sont :

- La nature et taille de la **base d'apprentissage**[4] disponible
- Le **nombre de classes k** - si k=2, on parle de classification binaire. - si k>2, on parle de problème *multi-classe*.
- Le **nombre maximum de classes** qui peuvent être attribuées à un texte T.

[4]Nous entendons par base d'apprentissage un jeu de textes pour lesquels les catégories sont connues. Cet étiquetage est généralement effectué manuellement, ce qui explique qu'il soit parfois difficile de disposer d'une base importante. Dans le cas de CISMeF, les ressources disponibles en 2002 (environ 10 000) ont été étiquetées à la main.

– Autres données disponibles

Le choix d'une méthode de classification dépend de ces paramètres : par exemple, si aucune base d'apprentissage n'est disponible, on pourra s'interroger sur l'opportunité d'en construire une afin d'utiliser une méthode d'apprentissage, ou bien choisir une méthode sans apprentissage.

Un premier type de méthode de classification consiste à analyser T directement, afin de découvrir des éléments qui permettent d'établir sa pertinence relativement à une ou plusieurs des classes de l'ensemble C_k. La mise en oeuvre de cette stratégie repose sur la mise en correspondance des éléments caractéristiques de T avec les classes C_k. Pour ce faire, il est possible d'utiliser les données de la base d'apprentissage (⇒ arbres de décision, réseaux...). Il est également possible de se fonder sur d'autres données disponibles (par exemple, une ontologie ⇒ Catégorisation Hiérarchique Ascendante) ou sur l'expertise d'un spécialiste du domaine concerné (⇒ système heuristique, ou à base de règles).

Une seconde méthode consiste à étudier un ensemble T_n de textes préalablement étiquetés à l'aide des classes C_k. L'analyse de T a alors pour but d'établir des ressemblances entre T et les textes étiquetés issus de la base d'apprentissage T_n afin d'en déduire quelle(s) classe(s) attribuer à T. On parle alors de méthode d'apprentissage(⇒ Machines à Vecteur Support (SVM), k-PPV)

Les limitations des méthodes de classifications sont liées au *temps* de mise en place du système (en particulier pour les méthodes requérant l'intervention d'un expert, ou la construction de ressources terminologiques), à la *taille* de la base d'apprentissage (dans certains cas, elle conditionne directement les performances) ou à la *complexité* des calculs mis en jeu pour les problèmes multi-classes (par exemple, pour les SVM). Ainsi, de nombreux travaux en classification visent à minimiser la taille de la base d'apprentissage[5] en introduisant des méthodes de classification dites semi-supervisées. On utilise alors une base d'apprentissage de petite taille qui peut être complétée par des éléments classés automatiquement, pour lesquels l'exactitude de la classification inférée peut-être estimée à l'aide de probabilités (Nigam, McCallum, Thrun, & Mitchell, 2000). D'autres travaux s'intéressent à la réduction de la complexité calculatoire, par l'intermédiaire d'une sélection de paramètres appropriée (par exemple, (Joachim, 1998) et (Dumais, Platt, Heckerman, & Saham, 1998) utilisent l'information mutuelle ; (Wiener, Perdersen, & Weigend, 1995) la méthode du chi-2).

Il apparaît que la plupart des méthodes de classification actuelles sont au moins en partie fondées sur un apprentissage à partir d'une base d'exemples étiquetés. Nous présentons les méthodes les plus courantes dans la section suivante, qui ont également été utilisées dans le cadre de l'indexation, tâche documentaire au centre de nos travaux[6].

4.2.2 Classification par Apprentissage

Une méthode d'apprentissage consiste à construire une fonction qui relie un vecteur de paramètres à son appartenance à une classe ou, le plus souvent, à sa probabilité d'appartenance à une classe. Ainsi, toutes méthodes d'apprentissage ont en commun une étape préliminaire à la classification proprement dite, la construction d'un vecteur de paramètres destiné à représenter les objets à classer[7] .

[5]Tout en maintenant de bonnes performances

[6]D'autres méthodes de catégorisation thématiques sont par exemple décrites dans (Manning & Shütze, 2000) ou (Han & Kamber, 2001)

[7]Cependant, certains objets peuvent être classifiés tels quels, sans utiliser de représentation intermédiaire

Sélection de paramètres

Dans le cas de la classification de textes, les vecteurs de paramètres représentant les textes sont le plus souvent des sacs de mots, mais il peut également s'agir d'autres unités d'indexation[8]. Les valeurs de chaque paramètre peuvent être binaires (par exemple, présence ou absence d'un mot dans le texte) ou non (par exemple, la valeur peut être calculée à partir de la fréquence du mot dans le texte et/ou dans la collection, comme dans le modèle proposé par (Salton & McGill, 1983)). Le vocabulaire d'une langue comptant un nombre considérable de mots (par exemple, le Petit Robert recense 60 000 mots français en 2004) il paraît indispensable de trouver une méthode de sélection des paramètres afin de réduire la dimensionalité du problème traité. La comparaison des résultats obtenus par (Joachim, 1998) et (Dumais et al., 1998) sur ce point semble indiquer que la sélection de paramètre peut également améliorer les performances. En effet, avec 300 paramètres, (Dumais et al., 1998) obtiennent une précision « break-even » de 87% en utilisant un SVM linéaire contre 84,2% pour le SVM linéaire de (Joachim, 1998) avec au moins 500 paramètres.

Classification

Rocchio

L'algorithme de Rocchio (Rocchio, 1971) est l'un des pionniers en catégorisation de documents. Développé dans le cadre de la recherche d'information, il s'appuie sur le « relevance feedback » des utilisateurs pour améliorer un paramétrage initial. Les caractéristiques des retours positifs sont additionnées au vecteur initial, alors que celles des retours négatifs sont retranchées. Le vecteur ainsi amélioré contient finalement les caractéristiques « moyennes » des documents recherchés. Appliqué à la catégorisation, on cherche à construire un vecteur caractéristique de chaque classe à partir de la base d'apprentissage. Pour chaque classe C_j.(j=1, ..., k), soient N_{C_j} le nombre de documents de la base d'apprentissage étiquetés à l'aide de la classe C_j et N le nombre total de documents de la base. On note $\overrightarrow{x} = (x_1, \ldots, x_n)$ un vecteur représentatif d'un document de la base d'apprentissage. On calcule alors le vecteur représentatif $\overrightarrow{c_j}$ de la classe C_j selon l'équation 4.1 :

$$\overrightarrow{c_j} = \alpha \cdot \overrightarrow{c_{init}} + \beta \cdot \frac{1}{N_{C_j}} \sum_{i \in C_j} \overrightarrow{x_i} - \gamma \cdot \frac{1}{N - N_{C_j}} \sum_{i \notin C_j} \overrightarrow{x_i} \tag{4.1}$$

Comme aucun vecteur initial n'est a priori disponible, on choisit $\alpha = 0$. Les valeurs de β et γ étant fixées, le classement d'un nouveau document s'opère en calculant la distance entre son vecteur représentatif $\overrightarrow{d}$ et les vecteurs représentatifs des classes $\overrightarrow{c_j}$ (j=1, ..., k). La ou les classes les plus proches sont attribuées au nouveau document. Les variantes de cette méthode portent sur le paramétrage des constantes β et γ (cependant, les valeurs choisies sont souvent 1 et 0, respectivement - par exemple dans (Dumais et al., 1998)) et le choix de la distance (par exemple, Dumais et al. (1998) utilisent la mesure Jacquart, Vinot, Grabar, and Valette (2003) utilisent la mesure cosinus). Les avantages de cette méthode sont, outre la simplicité de mise en oeuvre, la rapidité de l'apprentissage et la faible complexité calculatoire.

***K* Plus Proches Voisins (*k*-PPV)**

k-PPV est un algorithme issu de la reconnaissance des formes qui a été adapté à de nombreux autres domaines, y compris la classification de documents où il s'est révélé efficace

[8]Pour une description détaillée de ces unités d'indexation cf. section 5.2

(Yang & Chute, 1994; Lam, Ruiz, & Srinivasan, 1999). Contrairement à l'algorithme de Rocchio où l'apprentissage précède le classement, avec les k-PPV, il se fait dynamiquement. Le classement d'un nouveau document s'opère en calculant la distance euclidienne entre son vecteur représentatif et les vecteurs représentatifs des documents de la base d'apprentissage. Les k plus proches sont sélectionnés et la ou les classes majoritaires sont attribuées au nouveau document. Une variante consiste à pondérer le vote des voisins en fonction de la distance avec le document à classer. La complexité calculatoire dépend de la taille de la base d'apprentissage et non pas du nombre de classes, comme c'était le cas pour l'algorithme de Rocchio.

Machines à Vecteur Support (SVM)

Proposés par Vapnik en 1979, les SVM (Vapnik, 1998) sont maintenant fréquemment appliqués à la classification textuelle. Plusieurs études comparatives concluent à leur supériorité dans ce domaine (Dumais et al., 1998), (Joachim, 1998). Le principe des SVMs est de définir un hyperplan séparant l'espace où se trouvent les exemples positifs de l'espace où se trouvent les exemples négatifs. Cette séparation des données est en fait un problème d'optimisation consistant à maximiser la distance entre la frontière des classes et les vecteurs d'exemples les plus proches : ce sont les vecteurs support. Une variante de cette méthode permet de traiter les problèmes non-linéairement séparables en attribuant une pénalité aux vecteurs situés du « mauvais » côté de la frontière. Le classement d'un nouveau document est fondé sur la situation de ce document par rapport à l'hyperplan ainsi défini. Les SVMs s'appliquent au départ à des problèmes binaires, mais peuvent également être adaptés à des problèmes multiclasses. Dans ce cas, il est nécessaire de construire un SVM par classe, ce qui permettra de déterminer pour chaque classe si elle est pertinente pour un nouveau document.

4.2.3 Classification avant l'indexation

Nombre de méthodes sont fondées sur une analyse de la ressource, afin d'en extraire des vecteurs de données caractéristiques (généralement des sacs des mots) qui peuvent ensuite servir de base à une comparaison avec les vecteurs caractérisant les éléments d'une base d'apprentissage, ou bien à l'inférence de la catégorisation appropriée. La description vectorielle des ressources comporte notamment une étape de segmentation du texte avec une prise en compte de la variabilité morphologique des termes. Dans notre cas, ce type de travail pourrait s'avérer redondant avec l'extraction de descripteurs correspondant à la tâche d'indexation proprement dite. Des travaux (Teahan & Harper, 2001) montrent que les modèles statistiques de compression PPM (Prediction by Partial Match) introduits par (Bell, Cleary, & Witten, 1990) ont des performances intéressantes (comparables à celles des SVMs) pour la classification thématique, tout en abordant le problème d'une manière globale. Nous avons donc choisi de les adapter à notre problématique de classification de ressources de santé.

Principe de la classification à l'aide de modèles statistiques de compression

Chaque classe (spécialité médicale) est modélisée par un modèle de compression, qui est optimisé pour la compression de textes appartenant à la classe en question. Ainsi, un texte donné sera compressé de manière optimale par le modèle correspondant à la classe à laquelle il appartient.

Pour chaque modèle M_k(i=1, ..., N), on obtient un taux de compression (A_i) supérieur pour les textes traitant de la même spécialité que pour d'autres textes.

Afin de déterminer la classe d'un texte inconnu, celui-ci est compressé avec les différents modèles. Puis, la comparaison des taux de compression permet de classer les spécialités médicales par ordre de pertinence.

Compression d'un même texte par chaque modèle : il existe M_k *tel que* $A_k \geq A_i$ *(i=1, ..., N) et* i $\neq$ k. *Alors,* k *est la spécialité recherchée.*

Apprentissage des Modèles de compression

Pour chaque spécialité, nous avons constitué un corpus positif composé de ressources traitant de la spécialité concernée[9] et un corpus négatif composé de ressources traitant d'autres spécialités. Ces corpus sont utilisés pour entraîner les modèles de compression et déterminer l'ordre optimal (taille du contexte) pour chaque modèle. Pour ce faire, plusieurs tailles de contexte sont utilisées successivement (n=1, ..., 5). L'ordre retenu est celui pour lequel la différence entre les taux de compression positif et négatif est la plus grande. Pour nos expérimentations, nous avons utilisé l'implémentation en C de l'algorithme de compression PPM réalisée par F. Bellard[10] .

Résultats

Dans un premier temps, des modèles de compression ont été construits pour les quatre spécialités médicales les plus représentés dans le catalogue CISMeF[11] : Cardiologie, Chirurgie, Pédiatrie et Psychologie. Les corpus d'entraînement et de validation sont composés de 10 ressources pour chaque spécialité. Dans chaque cas, la différence de taux de compression entre les corpus positifs et négatifs est maximale pour des modèles PPM d'ordre 4. Sur un corpus de test de 50 ressources, la précision obtenue a été de 60%. La même expérimentation réitérée pour 10 contextes offre une précision de seulement 15%.

Discussion

Performances de la méthode Les résultats relativement encourageants obtenus sur quatre contextes ne sont pas reproduits à plus grande échelle, avec dix contextes par exemple. Le succès de la méthode utilisée repose en grande partie sur la qualité des données d'entraînement et on pourrait notamment mettre en cause la taille des corpus utilisés. Cette objection soulève un premier problème. Il serait en effet possible de travailler avec des corpus plus importants pour les spécialités les plus représentées dans le catalogue, mais cela n'est pas envisageable pour l'ensemble des 126 spécialités. D'autre part, pour éviter tout biais, les corpus *positifs* doivent impérativement représenter des contextes disjoints - c'est-à-dire qu'ils doivent être constitués de textes traitant de la spécialité à modéliser à l'exclusion des autres spécialités. Or, les ressources du catalogue CISMeF se limitent rarement à un seul contexte. Dans nos expériences, nous avons sélectionné des textes traitant *en majeur* de la spécialité recherchée et n'ayant aucune autre spécialité en majeur (c'est-à-dire, des textes ayant cette spécialité comme thème central, mais pouvant également développer d'autres thèmes secondaires - indiqués par une spécialité en *mineur*). Ce critère de sélection réduit

[9] A l'aide de la requête booléénne « *spécialité*.mt[majeur] », où « .mt » désigne le champ *métaterme* et « [majeur] »permet de sélectionner les ressources où ce concept est un thème principal

[10] `http://fabrice.bellard.free.fr/stat/rapport/rapport.html`

[11] En 2003.

considérablement le nombre de ressources disponibles pour constituer les corpus de travail[12] et ne semble néanmoins pas suffisant pour obtenir des résultats de classification satisfaisants...

En pratique, il apparaît donc que la méthode de catégorisation à l'aide de modèles de compression n'est pas applicable à une catégorisation aussi fine (une centaine de classes à l'intérieur d'un sous-domaine). Cependant, il peut être pertinent de l'utiliser dans le cadre d'une catégorisation exclusion, par exemple pour distinguer les ressources relevant du domaine de la santé dans lors de la veille. En effet, nous avons vu au chapitre 3 que l'un des progrès qui restent à accomplir dans le domaine de l'automatisation de la veille documentaire concerne la distinction entre les ressources traitant de santé et celles traitant d'un autre sujet. Cette tache de classification binaire devrait être prise en charge par la méthode que nous avons présentée, dans la mesure où la constitution de corpus de travail modélisant les contextes serait moins problématique.

Comparaison avec d'autres travaux

A titre indicatif, on peut constater que (Teahan & Harper, 2001) obtiennent une précision de 82,1% sur 1à catégories de Newsgroups de Usenet.

Une autre méthode de classification textuelle utilisant une approche globale de la représentation des ressources est présentée dans (Poulos, Papavlasopoulos, & Chrissikopoulos, 2004). Après avoir converti les textes à catégoriser en une expression symbolique, les auteurs utilisent un algorithme issu de la géométrie computationelle pour réduire la dimensionalité du problème et procéder à la catégorisation. Les résultats obtenus sont excellents quand la méthode est appliquée à la catégorisation de 15 textes courts (environ 500 mots) en trois catégories très distinctes (géométrie, biologie, finance). Cependant, il ressort de la conclusion de ces travaux que la complexité algorithmique de cette méthode la rend difficilement applicable sur des textes longs (qui doivent alors être segmentés en sacs de 100 mots). Par ailleurs, aucune expérimentation de la méthode n'a été réalisée sur une catégorisation plus fine comme les spécialités médicales.

Globalement, il semble que des méthodes plus classiques soient plus adaptées pour la catégorisation de documents. Nous rejoignons sur ce point les conclusions de (Franck, Chui, & Witten, 2000) qui estimaient que l'inconvénient majeur des modèles de compression pour la catégorisation de textes réside dans l'absence de souplesse au niveau de la sélection de paramètres, contrairement aux autres méthodes d'apprentissages comme celles présentées à la section 4.2.2. Dans des travaux plus récents, (Teahan & Harper, 2003) les auteurs proposent une méthode permettant d'introduire une sélection de paramètres dans les modèles de catégorisation par compression. Cependant, les résultats indiquent que cette méthode semble plus appropriée pour une catégorisation en fonction de la langue dans laquelle sont rédigés les textes ou en fonction des auteurs que pour une catégorisation thématique.

4.2.4 Classification après l'indexation

L'algorithme de catégorisation proposé ici est fondé sur une représentation formelle des concepts du domaine médical contenue dans la terminologie CISMeF. Il exploite l'indexation manuelle des ressources disponibles dans la base de données CISMeF. Cet algorithme repose sur un principe semblable à celui de la Classification Hiérarchique Ascendante (Bouroche &

[12]Le problème de la quantité de données étiquetées disponibles pour l'apprentissage de tâches de classification est par exemple abordé par (Nigam et al., 2000) qui proposent une méthode utilisant des données non étiquetées pour compléter les données étiquetées déjà disponibles. La méthode est appliquée à la catégorisation textuelle sur une vingtaine de classes.

Saporta, 1983) dans la mesure où il utilise les liens sémantiques existant entre les métatermes et les mots clés MeSH, les qualificatifs et les types de ressource afin d'extraire une liste de métatermes (des *pères* conceptuels, ou super-concepts) qui sera associée à une ressource donnée (dans la Classification Hiérarchique Ascendante, les liens utilisés sont créés dynamiquement à partir d'une méthode statistique). Cette liste de métatermes est alors ordonnée en fonction du poids cumulé des termes (mots clés, qualificatifs, types de ressource) ayant permis de l'inférer. Rappelons que les métatermes ont été introduit dans la terminologie CISMeF en 1997 afin d'optimiser la recherche d'information. Ici, nous utilisons les métatermes avec un tout autre objectif, la catégorisation de ressources.

Comme nous l'avons vu précédemment, chaque ressource recensée dans CISMeF est indexée par une liste de mots clés MeSH, associés ou non à des qualificatifs et par une liste de types de ressource. Par l'intermédiaire des liens sémantiques de la terminologie CISMeF (Figure 2.6 au chapitre 2), l'algorithme associe chaque élément de ces listes à un ou plusieurs métatermes. Ainsi, si un terme (mot clé, qualificatif ou type de ressource) est lié à plusieurs métatermes, chacun de ces métatermes sera retenu pour la catégorisation. Par exemple, le mot clé <*infections à VIH*> nous conduira à retenir les métatermes *virologie* et *infectiologie* tandis que le mot clé <*facteur socioéconomique*> conduira à retenir le seul métaterme *économie*. Le tableau 4.1 illustre les liens sémantiques mis en jeu pour la classification d'une ressource entière. Dans la colonne de droite « métatermes inférés » le signe « + » indique que plusieurs métatermes ont été inférés à partir du terme d'indexation correspondant, figurant dans la colonne centrale.

Par ailleurs, pour obtenir la catégorisation finale, l'algorithme calcule deux scores pour chaque métaterme retenu : un score « mineur »et un score « majeur ». Le score « mineur » correspond au nombre de mots clés et de couples (mot clé/qualificatif) mineurs à partir desquels le métaterme considéré a été retenu. De même, le score « majeur » correspond au nombre de types de ressource, de mots clés majeurs et de couples (mot clé/qualificatif) majeurs à partir desquels le métaterme considéré a été retenu. Ainsi, pour obtenir la catégorisation recherchée, les métatermes peuvent être classés par ordre de scores « majeurs » décroissants, les scores « mineurs » permettant de départager les éventuels *ex-aequo*. La figure 4.1 présente la méthode de calcul des scores.

Soit une ressource décrite par :
- *n mots clés MeSH* M_1, M_2*, ..., M_n *(une étoile signale les mots clés majeurs)*
- *m qualificatifs* Q_1, Q_2*, ..., Q_m *(les qualificatifs majeurs proviennent des paires mot clé/qualificatif majeures)*
- *p types de ressource* T_1, T_2, ..., T_p

Les liens sémantiques de la terminologie CISMeF permettent d'inférer de ces trois ensembles (mot clés, qualificatifs et types de ressource) une liste de k métatermes MT_1, MT_2*, ..., MT_k*. Pour chacun de ces k métatermes, un score mineur et un score majeur sont calculés :*
- *majeur(*MT_i*)= Card{M*, Q*, R, liés à* MT_i*}, le nombre de termes d'indexation majeurs et de types de ressource sémantiquement liés à* MT_i
- *mineur(*MT_i*)= Card{M, Q, liés à* MT_i*} , le nombre de termes d'indexation mineurs sémantiquement liés à* MT_i

FIG. 4.1 – Méthode de calcul des scores attribués à chaque métaterme pour la classification

	Terme	Métatermes inférés
Indexation MeSH	distribution selon sexe	-
	facteur socioéconomique	économie
	France	-
	*groupe soutien social	-
	*infection à VIH	virologie + infectiologie
	*infection à VIH/*épidémiologie*	virologie + infectiologie + *épidémiologie*+ *statistiques*
	*infection à VIH/*psychologie*	virologie + infectiologie + *psychologie*
	logement	environnement et santé publique
	Paris	-
	rôle médecin	-
	*soutien social	-
	*syndrome d'immunodéficience acquise	virologie + allergie et immunologie
	*syndrome d'immunodéficience acquise/*épidémiologie*	virologie + allergie et immunologie + *épidémiologie*+ *statistiques*
	*syndrome d'immunodéficience acquise/*psychologie*	virologie + *psychologie* + allergie et immunologie
	troubles liés à une substance	addictologie + allergologie
Type de Ressource	*article de périodique	-

TAB. 4.1 – Liens sémantiques utilisés pour la classification de la ressource n°9982

Les métatermes ayant un score majeur non nul sont dit « majeurs » et ils sont représentés avec un nombre d'étoiles correspondant à ce score. Pour reprendre l'exemple du tableau 4.1, le métaterme *virologie* est retenu à partir des deux mots clés majeurs <*infection à VIH*> et <*syndrome d'immunodéficience acquise*>, des deux paires majeures <*infections à VIH / épidémiologie*> et <*syndrome d'immunodéficience acquise / épidémiologie*> et des deux paires mineures <*infections à VIH / psychologie*> et <*syndrome d'immunodéficience acquise / psychologie*>. Dans cet exemple, *virologie* aura donc un score majeur égal à 4 et un score mineur égal à 2. Quatre étoiles lui seront attribuées (cf. figure 4.2 à la section résultats 4.2.4.)

Evaluation

Afin d'évaluer la pertinence des résultats obtenus, les catégorisations proposées par l'algorithme ont été comparées avec les catégorisations établies par un documentaliste CISMeF sur les mêmes ressources. La catégorisation manuelle étant relativement coûteuse en temps (environ 20 minutes par ressource), la taille du corpus d'évaluation a été volontairement limitée. Ainsi, l'algorithme de catégorisation a été évalué sur un échantillon de 123 ressources choisies aléatoirement dans le catalogue CISMeF. Le tableau 4.2 représente le nombre de spécialités à extraire pour chaque ressource, selon la documentaliste.

Nombre de Ressources	**Nombre de Spécialités**	**%**
Au plus 1 spécialité	31	25,20%
2 spécialités	32	26,02%
3 spécialités	32	26,02%
4 spécialités et plus	28	22,76%
Total	123	100%

Tab. 4.2 – Nombre de spécialités à extraire par ressource

La catégorisation manuelle, considérée comme notre référence (c'est-à-dire, fournissant la liste des spécialités pertinentes correctement ordonnées) a été réalisée après avoir pris connaissance des résultats de l'algorithme et présente l'avantage de proposer une validation de la catégorisation automatique.

Une évaluation quantitative a été réalisée à l'aide des mesures de précision et de rappel, usuelles en Sciences de l'Information. La précision représente le nombre de spécialités médicales correctement extraites par le système divisé par le nombre total de spécialités extraites. Le rappel représente le nombre de spécialités médicales correctement extraites par le système divisé par le nombre de spécialités attendues d'après la documentaliste.

Grâce à une évaluation qualitative, nous avons souhaité mettre en évidence la pertinence de la catégorisation proposée par l'algorithme. Trois niveaux ont été distingués : la catégorisation « sans faute » pour laquelle l'algorithme fournit exactement la même catégorisation que le documentaliste, la catégorisation « pertinente » pour laquelle l'algorithme fournit une catégorisation qui présente beaucoup de points communs avec celle du documentaliste, mais n'est cependant pas exacte et la catégorisation « non pertinente » qui n'a rien en commun avec la catégorisation du documentaliste ou présente un silence (catégorie omise). Plus exactement, une catégorisation « pertinente » est qualifiée de « bonne » si elle présente au plus deux erreurs mineures touchant moins de 50% des catégories attribuées manuellement. Elle est qualifiée de « moyenne » si elle présente trois ou quatre erreurs touchant 50% des catégories attribuées

manuellement. Enfin, elle est qualifiée de « mauvaise » si elle présente plus de quatre erreurs touchant plus de 50% des catégories attribuées manuellement.

Les écarts observés entre les deux catégorisations (manuelle et automatique) sont dus à quatre types d'erreurs : proposition par l'algorithme d'une spécialité non pertinente (bruit), oubli d'une spécialité pertinente (silence), erreur d'ordre dans le classement des spécialités et erreur de pondération majeur/mineur correspondant à ces spécialités.

Résultats

La figure 4.2 est un exemple de catégorisation obtenue par notre algorithme pour une ressource répertoriée dans le catalogue CISMeF. Sur le corpus d'évaluation, la catégorisation CISMeF offre une précision de 80,75% (au total 298 spécialités correctes sur les 369 proposées), soit un bruit de 19,25%. Le rappel est de 93,41% (au total 298 spécialités correctes sur les 319 attendues), ce qui correspond à un silence global de 6,59%. Ces résultats sont très satisfaisants et montrent que la catégorisation automatique des ressources est presque exhaustive.

FIG. 4.2 – Catégorisation de la ressource n° 9982 « Aides apportées aux personnes atteintes par l'infection à VIH-SIDA » (`http://www.sante.gouv.fr/drees/etude-resultat/er-pdf/er203.pdf`)(accédé le 08/09/03)

Le tableau 4.3 présente la pertinence de la catégorisation proposée par l'algorithme sur l'ensemble des ressources et le tableau 4.4 présente la répartition des types d'erreurs décrits au paragraphe précédent.

Pertinence	Nombre de Ressources	%
Catégorisation sans faute	45	36,58 %
Catégorisation pertinente (bonne)	32	26,01 %
Catégorisation pertinente (moyenne)	20	16,26 %
Catégorisation pertinente (mauvaise)	5	4,06 %
Catégorisation non pertinente	21	17,07 %
Total	123	100%

TAB. 4.3 – Pertinence de la catégorisation

Types d'erreurs	Nombre de Ressources
Spécialité inappropriée (bruit)	36
Spécialité omise (silence)	21
Erreur d'ordre	39
Erreur de pondération	
Majeur au lieu de mineur	42
Mineur au lieu de majeur	7

Tab. 4.4 – Typologie des erreurs de catégorisation observées

Discussion

Les principaux résultats de ce travail (précision de 80,75% et rappel de 93,41%) sont satisfisants. De plus, on constate que l'algorithme propose la catégorisation exacte (c'est-à-dire préconisée par le documentaliste) dans plus d'un tiers des cas. Les cas de catégorisation « non pertinente » sont généralement obtenus sur les ressources devant être indexées par une ou deux spécialités pour lesquelles un silence, parfois acompagné de bruit, intervient. Dans beaucoup des cas que nous qualifions de « pertinents », les erreurs observées sont très légères. Quelques cas seulement cumulent plusieurs types d'erreurs (4% de « mauvaise » catégorisation).

Notre méthode d'évaluation, qui a le désavantage de ne pas être réalisée en aveugle, permet de juger plus précisément la pertinence de chaque élément de la catégorisation automatique, afin de mettre en évidence le bruit et le silence de l'algorithme. Le silence mis en évidence par les résultats porte sur certains domaines spécifiques de la terminologie. Ainsi, une ressource indexée avec les mots clés <*voyage*> et <*médecine tropicale*> et avec le type de ressource *information patient et grand public* ne permet de retenir aucun métaterme car ces mots clés ne sont liés à aucun métaterme[13]. Ainsi, une analyse des résultats permet de dégager une liste de termes de la terminologie CISMeF pour lesquels il est nécessaire d'instaurer des liens vers des métatermes existants, ou encore de créer des métatermes adaptés à cet effet. Dans notre exemple, <*médecine tropicale*> est une spécialité médicale et il faut donc créer un métaterme *médecine tropicale* qui sera lié au mot clé <*médecine tropicale*>[14]. La création d'un tel métaterme enrichira la terminologie CISMeF et permettra d'améliorer les performances de l'algorithme de catégorisation, mais n'aura pas d'incidence sur la recherche d'information dans CISMeF. Malgré ces manques dans la terminologie, qui ont été comblés depuis, le silence de la catégorisation est très faible, sans doute car les lacunes touchaient principalement des spécialités peu abordées dans CISMeF (par exemple, *biologie cellulaire*, *biochimie* ou *chirurgie esthétique*).

En revanche, il est impossible de réduire le bruit observé sans affecter les performances de la recherche d'information. En effet, pour réduire le bruit, il faudrait identifier les liens provoquant une catégorisation « bruitée »et les supprimer. Il faut également remarquer que les métatermes proposés par l'algorithme de catégorisation et jugés superflus par la documentaliste ne constituent pas une mauvaise description du document en soi, mais ils nuisent à l'aspect synthétique de la catégorisation.

[13] Au début de notre étude.

[14] Un lien entre le métaterme existant *patient* et le mot clé *information patient et grand public* a également été rajouté.

Pour ce qui est de l'identification majeur/mineur des métatermes de la catégorisation, on observe six fois plus d'erreurs pour les mots clés « mineurs » identifiés comme « majeurs » (42 ressources sur 123) que l'inverse (7 ressources sur 123). Ceci résulte en fait de l'importance accordée au type de ressource qui n'est pas pondéré à l'origine, mais que nous considérons comme un terme « majeur » dans la constitution de la catégorisation. Suite à cette étude, l'équipe CISMeF a décidé d'instaurer une pondération majeur / mineur pour les types de ressource, comme pour les termes MeSH. En conséquence, le calcul des scores majeur et mineur de chaque métaterme a été ajusté à cette modification : en fonction de sa pondération, un type de ressource intervient désormais soit dans le calcul du score majeur s'il est majeur, soit dans le calcul du score mineur s'il est mineur.

Comparaison avec d'autres travaux

(Teahan & Harper, 2001) montrent que les modèles statistiques de compression PPM ont des performances intéressantes, tout en relevant d'une approche globale ne nécessitant aucune extraction de mots clés *a priori*. Nous avons montré au paragraphe 4.2.3 que cette approche ne convenait pas à la classification de ressources à l'intérieur d'un même domaine. L'application des SVMs sur le corpus Reuters (Dumais, 1998) montre que cette méthode peut s'appliquer sur un nombre de classes dépassant la centaine. Pour notre étude, nous ne disposions que de données limitées sur les textes du corpus de travail (cf. section 1.2.6) et il était notamment impossible d'utiliser le texte intégral des ressources. La représentation des ressources à l'aide de descripteurs MeSH et CISMeF que nous avons utilisée pour notre algorithme de classification hiérarchique résultait en un problème de très grande dimension pour lequel la sélection de paramètres représentatifs était difficile, en raison du faible nombre d'occurrence de chaque descripteur sur l'ensemble du corpus. Les résultats montrent clairement que les performances de notre algorithme sont directement affectées par les manques qui peuvent exister dans notre représentation des connaissances. Comme le souligne Bodenreider dans des travaux similaires (Bodenreider, 2000), l'implémentation de notre méthode de catégorisation nous permet d'optimiser notre terminologie. En effet, alors que les relations sémantiques de l'UMLS (Unified Medical Language System) constituaient une barrière pour améliorer les performances (Bodenreider, 2000), nous avons la possibilité d'étendre la couverture de notre terminologie. Suite à l'évaluation de l'algorithme que nous décrivons, nous avons introduit 18 métatermes (ainsi qu'une série de liens sémantiques associés) dans la terminologie CISMeF en décembre 2002. Une nouvelle étude informelle a permis de compléter l'ensemble des métatermes (et des liens sémantiques associés) de sorte que la terminologie compte aujourd'hui 126 métatermes. Les résultats obtenus par (Larkey & Croft, 1996) à l'aide d'une combinaison de classifieurs pour une tâche de classification similaire sont inférieurs et conduisent les auteurs à recommander l'utilisation d'une telle méthode de classification comme un outil interactif : la combinaison de classifieurs serait à privilégier pour aider les documentalistes dans leur tâche, mais il n'est pas souhaitable de l'utiliser sans validation humaine. Malgré le bruit et le silence constatés, l'algorithme CISMeF offre une catégorisation pertinente satisfaisant aux critères de classement et de pondération dans presque deux tiers des cas.

Ré-utilisabilité de la catégorisation

Application en bibliométrie. L'algorithme présenté ci-dessus a servi de base à la création d'un outil bibliométrique (Darmoni et al., 2005) disponible sur l'intranet du CHU de Rouen

depuis mi-2004. Adapté pour traiter les fichiers d'indexation de MEDLINE (au format xml), l'algorithme de catégorisation bibliométrique permet de connaître les spécialités médicales dont traite un ensemble de ressources indexées sur PubMed à l'aide de descripteurs MeSH. Ce nouvel algorithme diffère de celui que nous venons de présenter à la fois sur le plan de la méthode et sur le plan de la couverture :

1. *Méthode* : Les types de ressources CISMeF n'ont pas réellement d'équivalent dans MEDLINE. Seuls deux types de publication MEDLINE (par exemple, meta-analyse - relié au métaterme biostatistique) auraient pu être pris en compte, ce qui aurait pu causer un biais important. Ainsi, seuls les mots clés et les qualificatifs MeSH sont pris en compte pour le calcul des spécialités médicales.
2. *Couverture* : L'algorithme de catégorisation initial était destiné à la catégorisation de ressources isolées afin de compléter la description des ressources au sein du catalogue. Bien que la catégorisation de ressources isolées reste possible avec l'outil bibliométrique, celui-ci est destiné à la catégorisation de collections de ressources. Dans ce cas, il est appliqué récursivement sur l'ensemble des ressources d'une collection, afin de produire une catégorisation en spécialités de la collection.

Ainsi, cet outil bibliométrique permet par exemple de connaître les spécialités médicales traitées par un journal scientifique, ou par un auteur. Il peut se révéler un outil efficace pour la constitution de corpus spécialisés dans le domaine de la médecine.

Application aux dossiers patients La catégorisation proposée peut également s'appliquer aux dossiers électroniques patients. La catégorisation peut directement s'appliquer aux dossiers indexés (automatiquement ou manuellement) avec le MeSH. Cependant, en pratique, la CIM-10 est la terminologie couramment employée pour le codage médico-économique des dossiers patients. L'idée est alors d'adapter notre algorithme de catégorisation à la CIM-10 en créant des liens sémantiques entre les métatermes existant et les codes CIM-10 pertinents. Les équivalences MeSH/CIM-10 contenues dans l'UMLS peuvent permettre d'amorcer ce travail automatiquement à partir des liens existants entre métatermes et termes MeSH. Cependant, l'intervention d'un expert sera indispensable pour valider les liens obtenus, et compléter le réseau sémantique si nécessaire.

La catégorisation des dossiers permettrait d'une part de rendre le contenu du dossier plus abordable pour les patients (quand ils accèderont à leur dossier électronique) et serait particulièrement utile aux professionnels de santé dans le cas de patients présentant des pathologies complexes affectant de multiples organes et nécessitant un traitement par différents spécialistes.

Application multilingue Dans un contexte plus global, l'algorithme de catégorisation CISMeF peut être utilisé pour la catégorisation de ressources de santé indexées à l'aide de termes MeSH, que ce soit en anglais ou en français, puisque la terminologie CISMeF (présentée au chapitre 2) est bilingue français/anglais. Ainsi, la méthode présentée ci-dessus pourrait permettre de classer en spécialités médicales des ressources répertoriées par des portails de santé reconnus au niveau international tel que OMNI (URL : `http://omni.ac.uk/`), Cliniweb (URL : `http://www.ohsu.edu/cliniweb/`) ou HealthInSite (URL : `http://www.healthinsite.gov.au/`). Une extension à d'autres langues pour lesquelles le MeSH a été traduit (par exemple, l'espagnol ou l'allemand) est également envisageable avec un coût minimum correspondant à la traduction des métatermes et à la transposition des liens sémantiques.

4.3 Conclusion

Nous avons présenté deux méthodes destinées à établir une catégorisation synoptique de spécialités médicales reflétant les thèmes abordés dans les ressources de santé avant ou après indexation de ces ressources. Il apparaît que la catégorisation fondée sur l'indexation des ressources par des couples (mot clé/qualificatif) MeSH et des types de ressources CISMeF soit la méthode la plus efficace. Cette méthode exploite également les liens sémantiques existant entre les éléments de la terminologie CISMeF. Une évaluation sur 123 ressources choisies au hasard fournit des résultats très satisfaisants, qui ont permis d'une part d'enrichir la terminologie CISMeF et d'autre part de montrer qu'il était tout à fait pertinent de mettre l'algorithme introduit en production dans le catalogue CISMeF. Récemment, cette méthode a également été utilisée pour réaliser un outil bibliométrique (à titre d'illustration, nous présentons l'écran d'accueil de cet outil en annexe). La méthode de catégorisation par compression s'est révélée inadaptée à une catégorisation aussi fine à l'intérieur d'un même domaine. Cependant, elle pourrait être appliquée au filtrage de ressources de santé dans le cadre de la veille documentaire.

Chapitre 5

Indexation

L'indexation est avant tout une des étapes sur lesquelles se fonde la recherche d'information à l'intérieur d'une collection de documents. Le cadre plus large de l'indexation est donc celui de la recherche documentaire : étant donné un utilisateur avec un besoin d'information, il s'agit de lui proposer un fond documentaire contenant l'information cherchée (collection de documents adéquate) ainsi que les moyens d'extraire rapidement un document contenant l'information voulue de la collection. Pour effectuer cette dernière étape (extraction d'un document de la collection), on se propose de représenter à la fois les documents de la collection et l'information cherchée, appelée « requête ». Etant donné ces représentations, on définit alors une mesure de comparaison et on considère que l'information cherchée se trouve dans les documents les plus similaires à la requête d'après la comparaison effectuée. Après avoir défini précisément la tache d'indexation, nous détaillerons les différents aspects de la représentation des documents mise en jeu par l'indexation. Nous exposerons ensuite les méthodes de représentation existant pour tous types de documents et en particulier les ressources de santé. Nous aborderons alors la problématique de l'indexation automatique MeSH à travers une présentation contrastée des systèmes existant. Nous exposerons ensuite notre contribution en détaillant le développement du système MAIF. A travers une série d'évaluations, nous présentons les performances de MAIF, comparées à celles d'autres systèmes d'indexation MeSH français, suisses et américains. Finalement nous évoquons les applications possibles de MAIF en dehors du cadre strict de l'indexation MeSH.

5.1 Définitions

L'indexation est définie par la norme ISO 5963 comme une méthode pour l'examen de documents, l'analyse des sujets qui y sont traités et la sélection de termes d'indexation. Ainsi, l'indexation apparaît comme une tâche centrale en documentation. L'examen de quelques définitions données de cette activité font ressortir les aspects qui la caractérisent, tout en laissant la place aux diverses formes qu'elle peut prendre :

1. Indexation (TLF[1]). Le fait de dresser un répertoire ou une liste, généralement alphabétique, des sujets traités, des noms cités dans un ouvrage, suivis des références aux pages, aux paragraphes.

[1]Trésor de la Langue Française informatisé, disponible librement sur `http://atilf.atilf.fr/tlf.fr`

2. Indexation (*NF Z 47-102*, 1978). Activité consistant en l'analyse de documents et la retranscription en langage documentaire des concepts contenus dans ces documents.

La première définition renvoie aux index de fin de livre et la deuxième, plus générale, aux index figurant par exemple dans les catalogues des bibliothèques. Cependant, ces deux définitions se rejoignent sur l'aspect d'analyse des documents et sur l'extraction du contenu sémantique de ceux-ci, un point de vue partagé par (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1992). La différence réside dans la représentation du contenu sémantique : soit de manière hiérarchisée et en y insérant des références (Aït El Mekki & Nazarenko, 2004), soit comme une simple liste de concepts. En considérant l'indexation dans une perspective de recherche d'information, (Anderson & Perez-Carballo, 2001) observent un consensus sur la manière de procéder des documentalistes et évoquent une étape d'analyse du document suivie d'une étape de traduction des notions identifiées dans le langage d'indexation.

– l'indexation se voit redéfinie comme une activité de traduction : est-ce entièrement exact ? De quel type de traduction s'agit-il ?
– Il apparaît qu'il existe un langage d'indexation, ou langage documentaire : quel est-il ? Quelle est la nature des unités descriptives appropriées pour l'indexation ?

5.2 Les « langages » d'indexation : quelles unités descriptives utiliser pour la représentation d'un document ?

Le rôle des unités descriptives attribuées à un document lors de la phase d'indexation est double (Salton & McGill, 1983). Il doit à la fois être descriptif (c'est-à-dire représentatif du contenu du document) et discriminant (c'est-à-dire qu'il doit mettre en évidence ce qui distingue le document à l'intérieur de la collection)[2].

5.2.1 Indexation libre et indexation contrôlée

De nombreuses unités descriptives peuvent être utilisées pour la représentation d'un document. Avant de faire une description plus détaillée de chaque unité, remarquons d'abord qu'il existe deux types d'indexation, utilisant un langage d'indexation différent :

1. **L'indexation libre** utilise librement tous les mots d'une langue naturelle donnée, voire même des groupes de caractères appelés n-grammes (Halleb & Lelu, 1997). Ce type d'indexation est notamment utilisé par les moteurs de recherche procédant à une indexation entièrement automatique comme Google. Dans ce cas, l'indexation d'une ressource consiste en une liste de tous les mots (définis ici par des suites caractères séparés par un blanc ou un signe de ponctuation) contenus dans la ressource, auquel un filtrage ou une certaine normalisation pourront être appliqués (Salton & McGill, 1983). Cependant, l'indexation libre peut également être utilisée dans un cadre manuel ou semi-automatique comme IndDoc (Aït El Mekki & Nazarenko, 2004). Dans l'indexation libre, l'ensemble des unités descriptives qui peuvent être utilisés n'est pas connu *a priori*.

[2]Cette idée est également reprise par (Soergel, 1994) qui distingue deux approches de l'indexation, l'une recentrée sur le document (« entity-oriented ») qui cherche à décrire précisément le document et l'autre centrée sur la recherche d'information (« request-oriented »), qui cherche à utiliser le vocabulaire des utilisateurs du système de recherche d'information pour l'indexation, afin de discriminer le document à l'intérieur de la collection. On voit que ces approches ne sont pas incompatibles, mais qu'elle peuvent jouer sur le choix des unités descriptives utilisées.

2. **L'indexation contrôlée** utilise de manière contrainte les entités répertoriées dans une liste pré-définie. Le nombre de termes d'indexation susceptibles d'être utilisés est connu, il s'agit des termes contenus dans la liste de référence (terminologie) choisie. Cette terminologie définit également la forme des termes d'indexation utilisés. Il peut s'agir de termes ou d'expressions de la langue naturelle, de termes d'un méta-langage dit « langage documentaire[3] », employé pour souligner le caractère normatif attribué au descripteur, ou bien de symboles choisis pour représenter un concept de manière normative et unique.

On peut dire qu'indexation libre et indexation contrôlée se distinguent par la connaissance *a priori* ou non des unités descriptives à utiliser. Cependant, nous avons vu au chapitre 2 que les termes contrôlés ne sont pas nécessairement observables directement dans les textes à indexer. L'utilisation d'unités descriptives relevant de l'indexation libre peut s'avérer une étape intermédiaire nécessaire/utile.

5.2.2 Unités descriptives

Après une description d'une sélection d'unités descriptives[4], nous discutons des conséquences du choix d'un type particulier d'unité sur la description des documents et la recherche d'information.

- N-grammes : des groupes de n caractères permettant une description statistique de la langue par apprentissage des probabilités d'apparition de chaque groupe dans un corpus de la langue étudiée ((Manning & Shütze, 2000; Bell et al., 1990)). Plus la valeur de n est élevée, plus le modèle est précis. En revanche, la taille du corpus requis et des données d'apprentissage à stocker pour constituer le modèle croit également. Ainsi, en pratique, les modèles bi-grammes ($n = 2$) ou tri-grammes ($n = 3$) sont utilisés car ils permettent le meilleur compromis[5]. Afin d'utiliser des modèles plus précis ($n = 4$, 5 ou 6) (Lelu et al., 1998) opèrent une sélection des n-grammes retenus pour la représentation en fonction de leur fréquence dans le corpus d'apprentissage. Sur le même principe, il est également possible d'envisager une représentation à l'aide de groupes de n mots.
- Mots : suites de caractères séparés par un blanc ou un signe de ponctuation - c'est l'unité descriptive la plus simple, proposée par (Salton & McGill, 1983).
- Mots-formes : ici, le découpage du texte en mots formes tient compte tient compte du sens des unités dans la langue. Nous renvoyons ici le lecteur à la discussion sur la définition d'un mot composé évoquée au chapitre 2.
- Racine (ou radical) : Forme primitive d'où dérivent les mots d'une même famille (Larousse). On parlera de racine « libre » quand la racine correspond à un mot du lexique et de racine « liée » quand la racine ne constitue pas un mot à elle seule.
- Lemmes : Forme standard (i.e. masculin singulier pour les noms et adjectifs, infinitif pour les verbes) à laquelle sont ramenées les formes fléchies des mots.
- Codes grammaticaux : Ce sont les parties du discours qui peuvent être associées aux mots formes du texte. Il est possible d'utiliser différents jeux de codes grammaticaux.

[3](Jacquemin, 1997) définit les langages documentaires comme une liste contrôlée de termes d'indexation ayant fait l'objet d'une validation humaine.

[4]La liste n'est pas exhaustive - d'autres unités descriptives sont également employées dans la littérature, par exemple les « segments répétés » (Salem, 1984). Nous avons essayé ici de présenter une sélection représentative des différents niveaux de représentation existants.

[5]Voir par exemple (Rosenfeld, 2000) pour une description des modèles statistiques de la langue utilisés en recherche d'information, reconnaissance de la parole, traduction automatique et les améliorations proposées.

– Termes : expressions normalisées des concepts d'un domaine de spécialité. (cf. chapitre 2.)
– Requête : expression complexe combinant des instances des concepts d'un domaine de spécialité (Berrios, Cucina, & Fagan, 2002).

Afin d'illustrer les différentes représentations possibles, le tableau 5.1 ci-dessous donne la représentation de la phrase

(A) Le diabète de type 1 représente 20% des cas de diabète sucré.

à l'aide de chacune des unités descriptives que nous venons de définir.

Unité descriptive	**Représentation**
Tri-grammes	Le ;dia ;bèt ;e d ;e t ;ype ; 1 ;rep ;rés ;ent ;e 2 ;0% ; des ; ca ;s d ;e d ;iab ;ète ; su ;cré ;
Mots	Le ;diabète ;de ;type ;1 ;représente ;20% ;des ;cas ;sucré
Mots formes	Le ;diabète de type 1 ;représente ;20 ;% ;des ;cas ; de ;diabète sucré
Racines	Le ;diabèt ;de ;typ ;1 ;représent ;20 ;% ;de ;cas ;de ; diabèt ;sucr
Lemmes	Le ;diabète_de_type_1 ;représenter ;20 ;% ;de ;le ;cas ; de ;le ;diabète_sucré
Codes Grammaticaux[6]	Dms ;NCms ;VP3s ;Dmp ;NCmp ;P+Dmp ; NCms ;DPms ;NCms
Termes MeSH 2005	Diabète de type i ; diabète
Requête	-

Tab. 5.1 – Représentation d'une phrase à l'aide de différentes unités descriptives

5.2.3 Quelles unités descriptives choisir ?

Si l'on observe la phrase suivante :

(B) L'incidence de ce type de maladie est de 1 pour 100 000 habitants contre 1,6 chez les patients atteints de diabète insipide.

On voit d'emblée que la représentation choisie conditionne le nombre d'unité descriptives partagées par les énoncés A et B : avec une représentation par tri-grammes ou par codes grammaticaux, le nombre d'unités communes est assez élevé. Avec une représentation par mots ou par racines, on observe quelques unités communes. En revanche, avec une représentation par mots formes, seul le mot-outil « de » est commun aux deux énoncés, qui ne partagent aucun termes. D'une manière générale, il semble souhaitable pour l'indexation de rassembler les termes similaires sous une forme commune. Cependant, un niveau de description trop abstrait peut entraîner un silence total comme c'est la cas avec l'indexation par requête pour les enoncés A et B. (Lelu et al., 1998) justifient l'utilisation des n-grammes par la simplicité de mise en oeuvre de la méthode, qui a en outre l'avantage de prendre en compte dans une certaine mesure les graphies différentes d'un même mot, voire les fautes d'orthographe qui

[6]Les catégories grammaticales utilisées ici sont celles définies dans (Abeillé & Clément, 2003).

pourraient être récurrentes dans le corpus d'apprentissage. Bien que de nombreux travaux se fondent avec un certain succès sur l'utilisation de mots (Dumais, 1998), Bécue-Bertaut (2003) souligne l'intérêt du repérage des unités polylexicales (mots formes), indûment segmentées par un découpage en formes graphiques (mots).

Les opérations de racinisation et de lemmatisation, qui consistent respectivement à extraire d'un mot la racine ou le lemme correspondant sont des procédés souvent utilisés dans une optique de normalisation des termes dans le cadre de l'indexation de documents, et plus largement de la recherche d'information. Ainsi, dans la procédure d'indexation automatique qu'il décrit, Salton (Salton & McGill, 1983) préconise la désuffixation afin de ramener les mots de l'index à leur racine. Les méthodes de racinisation les plus utilisées sont celles proposées par Lovins (Lovins, 1968) fondée sur un désuffixage et recodage séparés et Porter (Porter, 1980) fondée sur un désuffixage et recodage simultanés.

D. Hull et G. Grefenstette (Hull & Grefensette, 1996)[7] ont étudié l'efficacité de ces deux techniques pour la recherche d'information dans une base documentaire indexée à l'aide de racines ou de lemmes des textes, respectivement. Leur hypothèse était que les deux méthodes sont *a priori* équivalentes, à condition de formuler les requêtes de manière adaptées à l'approche choisie. Ils ont cependant constaté qu'en anglais, lors de la racinisation, un certain nombre d'erreurs (rapprochement de termes qui n'appartiennent pas à la même famille, ou au contraire, omission de rapprocher des termes qui devaient l'être) ont un effet négatif sur la précision et le rappel. Ce phénomène est nettement amplifié sur le français, qui est une langue morphologiquement plus complexe que l'anglais. Les travaux de (Kraaij & Pohlmann, 1993) sur l'anglais reconnaissent également la supériorité d'une approche morphologique (lemmes) sur la désuffixation (racines). Ces observations ont donné lieu à plusieurs projets importants sur l'analyse morphologique et flexionnelle du francais, tel que le développement de Flemm (Namer, 2000), ou les travaux de N. Grabar (Grabar & Zweigenbaum, 2000). Par ailleurs, dans un rapport du projet FRANLEX[8], G. Dal souligne que « les travaux sur des langues dites à morphologie riche comme le slovène (Popovic & Willett, 1992) ou le néerlandais ((Kraaij & Pohlmann, 1996) donnent un avantage plus grand à la morphologie. Il en va de même pour le français où les techniques de désuffixage ne fournissent pas des informations aussi complètes et aussi riches qu'une base de données lexicales conjoignant des informations structurelles et sémantiques sur les mots construits. »

Ainsi, il apparaît que la racinisation et la lemmatisation sont des méthodes quasi-équivalentes (certains travaux considèrent la lemmatisation comme plus efficace) pour les langues à morphologie simple comme l'anglais, mais que la lemmatisation est significativement plus efficace pour les langues à morphologie complexe tel que le français. On peut également remarquer que pour certaines langues qui ont une morphologie très productive, la lemmatisation est une méthode difficile à appliquer. Pour l'allemand, des travaux récents (Marko, Daumke, Schultz, & Hahn, 2003) se portent en faveur de la racinisation et pour l'arabe, une étude réalisée par (Al-Kharashi & Evens, 1994) montre que l'utilisation de racines pour les termes de l'index est une meilleure méthode que l'utilisation de lemmes ou de mots. Un numéro récent de la revue Corpus alimente le débat sur le choix des unités descriptives et conduit (Pincemin, 2004) au constat qu'il n'existe pas a priori de meilleures unités descriptives, mais que chaque type d'unité oriente la description dans une direction plutôt thématique pour les

[7] cité par (Namer, 2000).

[8] Rapport en ligne disponible sur `http://perso.limsi.fr/jacquemi/FRANLEX/franlexv2.html`, accédé le 12/11/03.

lemmes ou les mots formes et plutôt stylistique pour les codes grammaticaux.

Il faut également remarquer que la plupart des travaux en recherche d'information mentionnés ci-dessus se fondent sur une indexation des documents à l'aide d'unités descriptives extraits du texte libre à l'aide de différentes méthodes. Nous allons maintenant nous intéresser à l'indexation à l'aide d'un vocabulaire contrôlé, par opposition aux unités descriptives libres.

Enjeu du langage d'indexation

Le langage d'indexation défini l'espace de représentation du document indexé. Ce choix stratégique orientera la recherche d'information (Soergel, 1994). En fonction du contexte de Recherche d'Information anticipé, l'une ou l'autre solution sera choisie (cf. Tableau 5.2).

Contexte de Recherche d'Information	**Langage d'Indexation**
Moteur de Recherche (eg. Google)	Libre
Livre	Libre/ Contrôlée
Catalogue (eg. Yahoo, CISMeF)	Contrôlée

Tab. 5.2 – Langage d'indexation utilisé en fonction du contexte de Recherche d'Information

Faut-il utiliser un vocabulaire contrôlé pour l'indexation ?

Dans le cadre de l'élaboration d'un index de fin de livre, la société américaine des indexeurs (ASI[9]) estime par exemple que le langage d'indexation utilisé (libre ou contrôlé) importe peu, dans la mesure où l'index rend bien compte du contenu de l'ouvrage et de l'analyse qui en a été faite lors de l'indexation. Par ailleurs, dans le cas d'une indexation contrôlée, une connaissance approfondie du vocabulaire contrôlé est nécessaire pour être en mesure d'effectuer aussi bien une indexation de qualité qu'une recherche efficace. On peut donc se demander si le choix d'une telle indexation est réellement pertinent. Plusieurs études de Salton (Salton 1979-91) à ce sujet concluent que l'utilisation d'un vocabulaire contrôlé donne des résultats équivalents ou légèrement supérieurs pour la recherche d'information à l'utilisation de mots libres comme termes d'indexation. Dans ce sens, les études de Leonard (1977) et Markey (1984) montrent que la consistance de l'indexation augmente en moyenne de 15% avec l'utilisation d'un vocabulaire contrôlé. Dans le cadre d'un système de recherche d'information, une certaine régularité dans l'indexation des documents est souhaitable afin de permettre aux utilisateurs de développer des stratégies de recherche efficaces. Pour Soergel (1994), l'utilisation d'un vocabulaire contrôlé permet une meilleure adéquation entre le langage de l'indexation et celui des requêtes. Leininger (2000) estime que le choix d'un vocabulaire contrôlé pour l'indexation des ressources d'une base documentaire permet de favoriser

[9] American Society of Indexers - Site `http://www.asindexing.org` consulté le 12/06/05.

la précision lors des recherches d'information, alors qu'une indexation libre favoriserait le rappel. Il observe également que l'utilisation d'un vocabulaire contrôlé est conditionnée par l'existence d'un thésaurus adapté à la base documentaire considérée. L'analyse de (Sparck-Jones, 1995) sur les résultats des campagnes d'évaluation TREC abonde également dans ce sens. Dans la communauté TREC, l'accent est mis sur un rappel élevé (« high recall is the normal requirement »). De plus, le large spectre (à la fois au niveau des thémes, genres, auteurs, sources etc.) des documents fait qu'aucun vocabulaire contrôlé n'est réellement adapté à la collection. Si on prend également en compte le fait que la collection de travail change à chaque campagne et donc que les thématiques traitées sont susceptibles d'être également différentes d'une campagne à l'autre, on ne peut que conclure avec l'auteur que l'utilisation d'un vocabulaire contrôlé dans ces conditions est totallement inaptée. (« it is foolish to rely primarily on any kind of relatively fixed, controlled, indexing language »).

Le grand nombre de terminologies médicales disponibles fait de la médecine un domaine particulièrement propice à l'indexation contrôlée. Une étude récente de la NLM (Wilbur & Kim, 2003) sur le domaine médical met en évidence la supériorité des termes MeSH sur les mots libres grâce à deux critères : la transparence et la prévisibilité. Chaque terme ayant une signification bien précise dans le thésaurus, les ambiguïtés sont réduites (transparence), ainsi que le choix de termes à utiliser pour exprimer un concept donné (prévisibilité).

Mise en correspondance de Racines et de Lemmes avec des termes contrôlés

La racine ***card*** peut être associée à 17 mots MeSH (*cardiaque*, *cardiovasculaire*...) si on ne considère que ceux qui contiennent la racine en début de mot et à 34 mots si on prend également en compte ceux qui contiennent card en milieu de chaîne (*myocarde*, *tachycardie*...). Ces ensembles de mots entrent respectivement dans la composition de 65 et 102 mots clefs MeSH différents. Cet exemple permet d'entrevoir les difficultés liées à la mise en correspondance des racines avec les mots clés MeSH. Dans la littérature, les systèmes automatique d'indexation contrôlée s'appuient largement sur la représentation par lemmes : le système FASTR (Jacquemin & Royauté, 1994) utilise un formalisme à trois niveaux fondé sur la reconnaissance de lemmes et l'exploitation d'informations morphologiques et sémantiques pour le français et l'anglais. D'autres systèmes d'extraction de mots clés du domaine médical (MeSH ou CIM-10) travaillent également à partir de lemmes (Pouliquen, 2002; Lovis, 1996).

Conclusion sur le choix d'un langage d'indexation

Nous ne pouvons que généraliser la conclusion de (Pincemin, 2004) et conclure qu'il n'existe pas a priori de langage d'indexation supérieur à tous les autres, qu'il soit contrôlé ou libre. Le choix d'un langage d'indexation doit se faire en considérant la tâche d'indexation dans son contexte spécifique et repose principalement sur les points suivants :

- L'indexation est-elle principalement descriptive ou discriminante ?
- La recherche d'information doit elle être principalement pertinente ou exhaustive ?
- Le domaine des documents traités est-il connu, délimité ? Si oui, est-il convenablement couvert par des terminologies spécialisées ?

Le choix de CISMeF

L'indexation des documents du catalogue CISMeF à l'aide du MeSH semble pleinement justifiée par les résultats de (Wilbur & Kim, 2003). Indépendemment des éléments issus de

la littérature scientifique, le choix d'un vocabulaire contrôlé pour CISMeF résulte également (1) d'une volonté de démarcation par rapport à un moteur de recherche généraliste tel que Google et (2) d'une volonté de reconnaissance à l'intérieur de la communauté médicale. Le choix du MeSH correspond à ces critères. De plus, ce thésaurus est développé par la NLM depuis plus de quarante ans, ce qui garantit sa pérenité. Par ailleurs, son utilisation pour la base bibliographique la plus utilisée dans le domaine de la santé (MEDLINE) en a fait la terminologie de référence pour la gestion de connaissances en santé.

5.3 L'activité d'indexation : un problème complexe

5.3.1 L'indexation : une traduction conceptuelle

Le rôle du documentaliste est celui de médiateur entre les documents et les utilisateurs à la recherche d'informations - parallèle avec le traducteur. Cependant, comme le souligne (Holzem, 2000), il ne s'agit pas d'une traduction entre deux langues, mais d'une « réduction en quelques mots clés des principaux concepts contenus dans un écrit ». En ce sens, l'indexation se rapproche du résumé de texte, qui consiste en une autre forme (*complémentaire* d'après (Lancaster, 1991)) de réduction d'un document faisant ressortir les concepts clés et la trame d'un document. Cependant, dans l'indexation, la langue du document importe moins que l'information véhiculée. C'est pourquoi l'indexation a pour but d'exprimer cette information de manière universelle, à l'aide de termes sélectionnés spécifiquement : on retrouve ici l'idéal de simplification et de désambiguïsation de la communication visé par Wüster. C'est donc tout naturellement que les terminologies sont le véhicule privilégié du contenu sémantique des ressources indexées. Cependant, comme nous l'avons vu au chapitre 2, les terminologies ne sont pas réellement universelles et la traduction conceptuelle ne se fait pas sans heurt. Pour reprendre l'exemple de (Holzem, 2000) sur les vedettes matières Ramaux biniou/musette, l'indexeur est parfois confronté au paradoxe de la traduction évoqué par Paul Ricoeur (Ricoeur, 2004) : bonheur de transmettre le message de l'auteur au lecteur *vs.* deuil de ne pouvoir y parvenir que de manière imparfaite. Dans l'indexation, cette imperfection résulte du fait que les terminologies n'offrent qu'une possibilité de description limitée aux concepts recouverts par les mots clés. En ce sens, l'indexation s'apparente à une catégorisation des documents en fonctions des mots clés de la terminologie. Selon la terminologie utilisée, la catégorisaion sera plus ou moins fine- rappelons par exemple que le MeSH comporte environ 23 000 mots clés, contre environ 150 000 pour la SNOMED.

5.3.2 L'indexation : une catégorisation

Les langages documentaires, les terminologies utilisées pour l'indexation sont des formes appauvries de la langue, et indexer revient aussi à classer les documents avec des étiquettes correspondant aux termes. Ainsi, Kleber (1990) assimile concepts et catégories. Rosch and Mervis (1975) remarquent que « les catégories sont généralement désignées par des noms » alors même que Rastier (1995) estime que la nominalisation et l'une des étapes de l'accession d'un mot au statut de terme. Bertrand (1993) affirme également que l'indexation est bien une activité de catégorisation. Dans l'indexation, il s'agit de déterminer de quoi traite le document à indexer. Transcrire la réponse à cette question à l'aide de termes d'indexation revient à une alternative binaire : le document aborde-t-il, oui ou non le concept représenté par le terme x ? Dans l'affirmative, la sélection de x comme terme d'indexation doit être motivée par les

deux autres aspects que l'indexeur doit prendre en compte, selon Lancaster (1991) : d'une part, la place que le document doit occuper dans la collection où il s'inscrit et d'autre part, les centres d'intérêt des lecteurs potentiels. Ces deux critères font sans aucun doute appel au jugement de l'indexeur et conduisent à se poser la question de l'objectivité de l'indexation (évoquée en 5.4.1).

5.3.3 L'indexation : résolution d'un problème mal défini

Considérée sous l'angle de la psychologie cognitive, l'indexation est un problème « ouvert », dans la mesure il n'existe pas de solution unique optimale, mais plutôt un éventail de solutions possibles et acceptables (Lancaster, 1991). L'indexation fait également partie des problèmes dits « mal définis », par opposition aux problèmes « bien définis » (Simon, 1973)[10]. Un problème est considéré comme « mal défini » si toutes les informations nécessaires à sa résolution ne sont pas présentes dans l'énoncé du problème. Plus précisément, face à leur tâche, les indexeurs ne disposent que d'une représentation mentale incomplète et imprécise de certaines composantes élémentaires du problème[11]. Un flou subsiste quant au but précis à atteindre : Combien de descripteurs faut-il choisir ? Quels sont les concepts à privilégier ? L'ensemble des contraintes liées au problème peut également faire l'objet d'un certain flou : par exemple, certains vocabulaires contrôlés (le MeSH, le Thesaurus of Psychological Index Terms, ...) comportent une série de descripteurs « obligatoires » (les Check Tags) qui doivent être systématiquement utilisés si le concept auquel ils renvoient est abordé dans le document à indexer ; cependant, faut-il sélectionner un Check Tag si le concept est simplement mentionné dans une partie secondaire du document ? Pour être en mesure de résoudre le problème (c'est-à-dire de produire l'indexation d'un document), l'indexeur doit dans un premier temps résoudre les interrogations soulevées par le flou sur les composantes élémentaires. L'indexeur doit donc préciser la représentation mentale du problème dont il dispose en y ajoutant ses propres contraintes afin de transformer un problème « mal défini » en problème mieux défini, qu'il sera possible de résoudre. Cette étape de redéfinition du problème fait nécessairement appel au jugement de l'indexeur et nous renvoie à la question de l'objectivité de l'indexation, qui était également soulevée par les activités de traduction et de catégorisation desquelles nous avions rapproché l'indexation. Finalement, on peut dire que la traduction et la catégorisation de documents sont également des problèmes mal définis, ce qui explique que l'indexation puisse être traitée comme l'un ou l'autre.

5.4 Critères d'évaluation de l'indexation

5.4.1 Objectivité de l'indexation

Est-il possible- voire souhaitable- que l'indexation soit objective ? Bertrand (1993) observe que malgré une forte volonté de normaliser et donc d'objectiver l'indexation, ce but est rarement atteint en pratique. A ce sujet, on peut par exemple remarquer que la subjectivité même des terminologies utilisées pour l'indexation rend toute objectivité utopique. La mesure d'autres aspects cognitifs tels que la connaissance que l'indexeur a de la terminologie et du domaine de travail, ou la psychologie générale de l'indexeur viennent conforter ce constat.

[10] Cité par (David, Giroux, Bertrand-Gastaldy, & Lanteigne, 1995).

[11] Les composantes élémentaires d'un problème sont : but, état initial, opérations possibles et contraintes - selon (David et al., 1995).

Cependant, les critères avancés par Lancaster (1991) dépassent ce constat sur l'impossibilité d'objectivité et l'auteur affirme même qu'il est nécessaire que l'indexation soit partiale, car elle est régie par une politique éditoriale subjective. Lancaster (1991) et Soergel (1994) s'accordent pour dire que l'indexation est un rouage dans le fonctionnement d'un système de Recherche d'Information centré sur l'utilisateur. Ainsi, l'indexation doit tenir compte des besoins d'information des utilisateurs et assurer la cohésion entre les documents présents dans le système. Ces principes semblent naturellement impliquer qu'une fois défini le cadre précis dans lequel l'indexation est réalisée, ces contraintes ajoutées à la normalisation du langage de l'indexation devraient assurer une indexation régulière, sinon objective. Cette « régularité » attendue de l'indexation ne va pourtant pas de soi. Pour la définir et l'étudier, on parle de « consistance de l'indexation ».

5.4.2 Consistance de l'indexation

Définition

La consistance de l'indexation est une notion qui vise à apprécier la concordance entre des indexations proposées pour un même document par deux indexeurs ou deux méthodes d'indexations différentes. Idéalement, si les règles d'indexation sont bien définies, deux indexeurs différents devraient produire la même indexation pour un même document : c'est la consistance inter-indexeur. De même, un même indexeur devrait produire la même indexation pour un même document à deux moments donnés : c'est la consistance intra-indexeur.

Facteurs de consistance / variabilité

Dans les faits, on observe des écarts entre les indexations réalisées dans ces deux situations (deux indexeurs différents à un même moment, un même indexeur à deux moments différents). La consistance inter-indexeur semble meilleure dans le cas d'une indexation contrôlée, par opposition à une indexation libre : l'étude de (Berrios et al., 2002) sur l'indexation libre de 3 chapitres de livre rapporte une consistance moyenne de 35% contre 50% [12] pour les études de consistance réalisée par (Funk, Reid, & McGoogan, 1983) et (Leininger, 2000) sur l'indexation d'articles scientifiques à l'aide de deux vocabulaires contrôlés, le MeSH et le *Thésaurus of Psychological Index Terms*. De plus les études de (Markey, 1984) et (Leonard, 1977) arrivent aux mêmes conclusions.

Par ailleurs, (Landes & Spidal, 2003) montrent que les différences d'indexation pour un même ouvrage entre deux indexeurs ayant reçu les mêmes instructions peuvent s'expliquer par une différence de formation (technique vs. généraliste) et d'expérience (indexation d'ouvrages techniques vs. généralistes). (Leininger, 2000) évoque également une série d'autres facteurs susceptibles d'influer sur la variabilité de l'indexation, tels que les outils (logiciels, manuels...) à disposition des indexeurs ou l'environnement dans lequel l'indexation à lieu. Par exemple, le dernier document indexé peut avoir une influence sur l'indexation en cours car l'indexeur pourra avoir tendance à mieux repérer les similitudes ou différences entre les deux documents. Pour résumer les facteurs de variabilité de l'indexation, on peut dire qu'ils sont de deux types :

1. Des facteurs internes : il s'agit des connaissances propres à l'indexeur, acquises au cours de la formation ou de l'expérience, ainsi que de son jugement propre ou ses préférences personnelles.

[12] Les trois évaluations ont été réalisées à l'aide de la mesure de Hooper, définie à la section suivante.

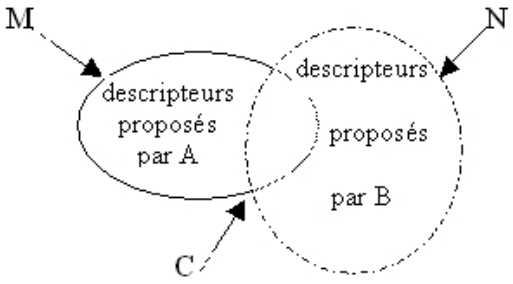

FIG. 5.1 – Répartition des descripteurs attribués par A et B pour un même document

2. Des facteurs externes : il s'agit des règles d'indexation, du vocabulaire contrôlé (le cas échéant), des contraintes temporelles imposées pour l'indexation (le cas échéant), des outils disponibles, de l'environnement de travail...

Mesures de consistance

Les nombreuses études réalisées sur la consistance [13] - que ce soit pour l'indexation ou pour d'autres problèmes cognitifs tels que la formulation de requêtes (Saracevic & Kantor, 1988) ou la construction de documents hypertextes (Furner, Ellis, & Willett, 1999), utilisent un spectre de mesures de la consistance assez large. Furner et al. (1999) choisissent par exemple d'adapter une mesure de similarité (mesure cosine). D'autres travaux utilisent des mesures dédiées. Ainsi, un état d'art de la consistence inter-indexeur datant de la fin des années soixante-dix recense sept mesures différentes. Les mesures les plus usitées à ce jour semblent être la mesure de Hooper (Hooper, 1965), la mesure de Rolling (Rolling, 1981) et le taux de recouvrement (Saracevic & Kantor, 1988). Nous les détaillons ci-dessous.

Soient A et B deux indexeurs, proposant chacun une liste de descripteurs à attribuer à un document. Soient M le nombre de descripteurs proposés par A exclusivement (i.e. non proposés par B) et N le nombre de descripteurs proposés par B exclusivement (i.e. non proposés par A[14]) - a priori, $N \neq M$. Enfin, soit C le nombre de termes proposés à la fois par A et par B. La figure 5.1 illustre cette répartition.

On peut alors définir :

– Mesure de Hooper (Hooper, 1965) : Cette mesure évalue simplement la proportion de descripteurs proposés par les deux indexeurs à la fois, sur l'ensemble des descripteurs proposés par l'un ou l'autre des indexeurs :

$$CH = \frac{100 * C}{M + N + C} \tag{5.1}$$

– Mesure de Rolling (Rolling, 1981) : Cette mesure accorde un poids supplémentaire aux descripteurs témoignant d'un consensus entre les deux indexeurs (les descripteurs

[13] Remarquons cependant qu'aucune étude récente ne porte sur la consistance de l'indexation dans le domaine de la santé, et en particulier l'indexation MeSH.

[14] En principe, ces ensembles tiennent compte des descripteurs strictement proposés par les deux indexeurs, sans considérer la proximité sémantique éventuelle des descripteurs. Ainsi, si l'indexeur A propose le descripteur « céphalée » et l'indexeur B « mal de tête », on considèrera que des descripteurs différents ont été proposés, bien que les concepts dénotés soient similaires.

proposés par les deux indexeurs à la fois), par rapport aux descripteurs témoignant d'une divergence d'appréciation (les descripteurs proposés par l'un des indexeurs seulement).

$$CR = \frac{100 * 2C}{M + N + 2C} \tag{5.2}$$

– Taux de recouvrement (Saracevic & Kantor, 1988) : Cette mesure diffère des précédentes, car elle ne place pas les deux listes de descripteurs au même niveau. En effet, elle permet d'évaluer le taux de recouvrement d'une liste par rapport à une autre. Ainsi, si on considère la liste proposée par l'indexeur A comme référence, on utilisera la mesure SA et si on considère la liste proposée par l'indexeur B comme référence, on utilisera la mesure SB. On a :

$$SA = \frac{100 * C}{M + C} \qquad et \qquad SB = \frac{100 * C}{N + C} \tag{5.3}$$

Afin d'illustrer l'utilisation de chacune de ces mesures, nous allons évaluer la consistance des listes de descripteurs du tableau 5.3 :

Liste A[15]	**Liste B**
diabète de type ii	grossesse
diabète de type ii/chimiothérapie	hypoglycémiants/ administration et posologie
grossesse	sujet âgé
hypoglycémiants	diabète de type ii/thérapeutique
hypoglycémiants/ administration et posologie	diabète de type ii/complications
hypoglycémiants/classification	hypoglycémiants/effets indésirables
hypoglycémiants/effets indésirables	hypoglycémiants
interactions médicamenteuses	diabète de type ii
suivi soins patient	
sujet âgé	

Tab. 5.3 – Deux listes de descripteurs MeSH attribués à un même document

Quatre descripteurs sont proposés seulement par A : <*diabète de type ii*>, <*hypoglycémiants / classification*>, <*interactions médicamenteuses*> et <*suivi soins patient*>. Trois descripteurs sont proposés seulement par B : <*diabète de type ii/thérapeutique*>, <*diabète de type ii/complications*> et <*hypoglycémiants/administration et posologie*>. Cinq descripteurs sont proposés à la fois par A et B : <*grossesse*>, <*sujet âgé*>, <*hypoglycémiants/effets indésirables*>, <*hypoglycémiants*>, <*diabète de type ii*>. Selon les notations établies ci-dessus, nous avons donc $M = 4$, $N = 3$ et $C = 5$. On peut donc calculer :

$$CH = \frac{100 * 5}{4 + 3 + 5} = \frac{500}{12} = 41,7\%$$

[15]Les listes comportent des descripteurs attribués au document « Le diabète de type 2 ». La liste A correspond au contenu de la notice CISMeF numéro 115 ; la liste B correspond aux huit premiers descripteurs proposés par le système d'indexation automatique MAIF lors d'une évaluation réalisée en 2004.

$$CR = \frac{100 * 10}{4 + 3 + 10} = \frac{1000}{17} = 58,8\%$$

$$SA = \frac{100 * 5}{4 + 5} = \frac{500}{9} = 55,6\% \qquad et \qquad SB = \frac{100 * 5}{3 + 5} = \frac{500}{8} = 62,5\%$$

On constate que, selon la mesure utilisée, la consistance entre les listes A et B varie considérablement : de 41,7% avec la mesure de Hooper à 62,5% avec le recouvrement considérant la liste B comme référence.

Il est également intéressant de noter qu'une étude portant sur la recherche d'information (Saracevic & Kantor, 1988) a montré que les termes utilisés par différentes personnes pour rechercher la même information ne concordent que rarement (taux de recouvrement moyen de 27%). Il apparaît que le problème de consistance n'est pas propre à l'indexation et qu'il se rencontre en fait dans d'autres problèmes faisant appel à une traduction conceptuelle dans un langage contrôlé, par exemple, la formulation de requêtes dans un moteur de recherche.

Consistance de l'indexation dans CISMeF

Dans la littérature, les études sur la consistance de l'indexation sont réalisées soit :

- A partir de doublons, c'est à dire des documents indexés deux fois par inadvertance : c'est le cas par exemple pour les études réalisées par (Funk et al., 1983) ou (Leininger, 2000).
- A partir de documents indexés par plusieurs indexeurs différents dans le cadre précis de l'évaluation : c'est, par exemple, le cas pour l'étude réalisée par (Berrios et al., 2002) ou (Landes & Spidal, 2003).

Chacune des méthodes présente des avantages et des inconvénients. Les deux facteurs principaux sont le nombre de documents disponibles pour l'étude et l'attitude des indexeurs pendant leur tâche. Si la double indexation est réalisée spécifiquement pour l'étude, les indexeurs savent donc que la consistance de l'indexation va être évaluée et il est possible que cela affecte leur travail. De plus, l'indexation manuelle étant coûteuse en temps, il est souvent difficile de travailler avec un nombre élevé de documents. En revanche, si l'étude est réalisée à partir de doublons, un plus grand nombre de documents peuvent être disponibles et on peut être assuré que les indexeurs n'ont pas eu conscience que la consistance de l'indexation serait évaluée lors de la constitution des index. Cependant, l'inconvénient de cette méthode est qu'il n'est généralement pas possible de connaître l'identité des deux indexeurs. Par conséquent, il n'est pas exclu qu'un document ait été indexé deux fois par le même indexeur. De même, il n'est pas possible de maîtriser certains paramètres comme l'expérience ou la formation des deux indexeurs participant à l'étude, sans parler de l'impact des variations dues à l'évolution de la terminologie CISMeF (le nombre de mots clés MeSH utilisés par CISMeF à partir de 2000 est nettement plus important que sur la periode 1995-2000 du fait de l'automatisation de certains aspects de la gestion de CISMeF), ou des règles d'indexation CISMeF.

Pour des raisons pratiques, nous avons étudié la consistance de l'indexation dans CISMeF à partir de 10 ressources indexées en double entre 1998 et 2005. La mesure utilisée est celle de Hooper (que nous notons CH). La consistance a été évaluée pour l'ensemble des descripteurs (mots clés ou paires mot clé/qualifcatif MeSH), pour les mots clés majeurs et pour les types de ressources. La consistance globale est donc de 30% et de 40% pour les mots clés majeurs. Ces résultats concordent avec l'étude de (Funk et al., 1983) qui observaient également une consistance plus élevée pour les mots clés majeurs (61%, également avec la mesure de Hooper) par rapport à la consistance globale (CH=33%). La consistance sur les types de ressources

est plus élevée que pour les mots clés : 55% sur notre échantillon. En raison du faible nombre de ressources concernées par l'indexation des Check Tags (3 ressources sur 10 dans notre échantillon) ou des descripteurs géographiques (2 ressources sur 10) nous avons estimé que les chiffres obtenus seraient peu représentatifs de la consistance sur ces types de descripteurs.

5.4.3 Qualité de l'indexation

Nous avons vu en 5.2.3 qu'il est souhaitable pour l'efficacité de la recherche d'information que l'indexation à l'intérieur d'une collection documentaire soit consistante. La consistance peut donc être considérée comme un critère de qualité de l'indexation. Cependant, il ne s'agit pas d'un critère suffisant : l'indexation peut être invariablement mauvaise, comme invariablement bonne (Cooper, 1969)[16]. Il est donc nécessaire d'envisager également d'autres mesures de qualité. Le problème principal de l'évaluation de l'indexation, souligné par (Lancaster, 1991)est qu'il n'existe pas d'indexation « de référence » à laquelle confronter l'indexation (humaine ou automatique) à évaluer. Ainsi, les méthodes utilisées dans la littérature pour évaluer l'indexation en tant que telle sont au nombre de deux :

1. Méthode a priori : comparaison de l'indexation à un « gold standard », une indexation prise comme référence, élaborée par un indexeur expert. Ce type d'évaluation correspond finalement à la mesure de la consistance entre l'indexation étudiée et l'indexation prise comme référence, si on considère l'indexation comme un tout indissociable. Cependant, dans le cas de l'indexation automatique, la plupart des systèmes proposent une liste ordonnée de descripteurs susceptibles d'être utilisés pour l'indexation. Dans ces listes, le premier descripteur proposé est considéré par le système comme plus pertinent que le second et ainsi de suite. On voit ici que l'ordre d'apparition des descripteurs dans la liste est important. En conséquence, les mesures de qualité utilisées doivent également prendre en compte le rang des descripteurs extraits automatiquement.
2. Méthode *a posteriori* : validation de l'indexation par un indexeur expert.

Une autre solution pour évaluer l'indexation consiste à replacer cette tache dans le cadre plus général de la recherche d'information. On peut alors faire l'hypothèse que la qualité de l'indexation aura un impact direct sur la recherche d'information. On peut alors observer les performances d'un système de recherche d'information utilisant l'indexation que l'on souhaite évaluer (Kim, Aronson, & Wilbur, 2001). Cependant, ces méthodes impliquent le jugement subjectif d'un expert, pour apprécier la pertinence des mots clés sélectionnés pour l'indexation ou des documents retournés pour la recherche d'information.

Le tableau 5.4 illustre la répartition des termes d'indexation qui peuvent être attribués à une ressource donnée. Selon les notations du Tableau 5.4, A+B+C+D représente l'ensemble

	$\in$ Indexation Manuelle	**$\notin$ Indexation Manuelle**
$\in$ Indexation Automatique	A	B
$\notin$ Indexation Automatique	C	D

TAB. 5.4 – Distribution des termes d'indexation pour une ressource donnée

[16]Cité par (Funk et al., 1983)

des termes d'indexation disponibles (soit, pour le MeSH N= 507.845[17]) Ainsi, soient P la précision, R le rappel, et F la F-measure.

$$P = \frac{A}{A+B}$$

$$R = \frac{A}{A+C}$$

$$F = \frac{1}{\alpha.\frac{1}{P} + (1-\alpha).\frac{1}{R}} \quad avec \quad 0 \leq \alpha \leq 1$$

α représente le poids attribué à la précision. Si précision et rappel sont considérés comme étant d'importance égale, on fixe $\alpha = 0,5$.

La précision représente les termes correctement extraits par le système. On peut aussi évoquer le silence ($S = 1 - R$) pour évaluer la proportion de termes n'ayant pas été extrait. Le rappel représente la couverture du système. On peut aussi évoquer le bruit ($N = 1 - P$) pour évaluer la proportion de termes erronés extraits par le système (faux positifs), ou la pureté (Pureté=$\frac{D}{B+D}$) pour évaluer la proportion d'erreurs d'indexation (extraction d'un terme erroné) évitées par le système (Soergel, 1994). Il faut remarquer que cette dernière mesure n'est utilisable que dans le cas de l'indexation contrôlée. En effet, pour être en mesure d'évaluer D, il faut connaître le nombre total de descripteurs qui peuvent être attribués à un document lors de l'indexation, qu'elle soit manuelle ou automatique.

En Recherche d'Information et notamment dans les campagnes d'évaluation TREC, une importance particulière est accordée au rappel des systèmes évalués (Sparck-Jones, 1995). Ainsi, il est donc d'usage de représenter les performances des systèmes à l'aide de la « 11-point average precision ». La précision est calculée aux onze points de rappel fixe 0%, 10%, ...et 100% et la moyenne des valeurs obtenues est calculée. Pour une représentation graphique des résultats, il est également d'usage de tracer une courbe précision / rappel faisant figurer l'ensemble de ces valeurs. La Recherche d'Information accorde également un intérêt tout particulier au point de rupture (en anglais, « break-even point ») c'est-à-dire, le rang où la précision est idéalement égale au rappel. Ainsi, si n termes sont attendus pour un document donné, la précision au point de rupture correspond à la précision au rang n.

Cependant, toutes ces mesures sont effectuées avec comme point focal un document précis. Dans le cadre de l'évaluation d'un système automatique d'indexation ou de catégorisation, on peut également envisager comme (Lam et al., 1999) d'effectuer des mesures centrées sur *les termes d'indexation ou catégories*, ou bien centrées sur les *décisions* prises par le système. Ainsi, la mesure centrée sur les catégories utilisées par exemple par (Soergel, 1994) ou (Lam et al., 1999) consiste à calculer précision et rappel pour un terme d'indexation donné. La précision correspond alors au nombre de documents auxquels le terme d'indexation a été correctement[18] attribué divisé par le nombre total de documents auxquels le terme d'indexation a été attribué par le système automatique. De même, la mesure centrée sur les décisions prend en compte l'ensemble des documents de l'évaluation pour calculer précision et rappel. La précision calculée par (Lam et al., 1999) correspond alors au nombre total de cas où les termes d'indexation ont été correctement attribués divisé par le nombre total d'attributions effectuées par le système automatique.

[17]Soit 484.985 paires MC/Q possibles et 22.860 mots clés isolés.

[18]Par « correctement attribué », nous entendons que le terme a été attribué conjointement par le système et par l'indexation manuelle de référence.

Cette grande diversité des mesures de qualité de l'indexation a pour conséquence principale la difficulté de comparaison entre différents travaux. En effet, il semble difficile de tirer des conclusions sur des systèmes évalués sur des corpus différents et/ou avec des métriques différentes. Par ailleurs, cette situation donne lieu à une réflexion renouvelée sur la définition de la (des) meilleure(s) mesure(s)[19].

5.5 Rôle des titres dans l'indexation

5.5.1 Sémantique des Titres

Les titres de documents et de paragraphes revêtent une importance toute particulière dans le cadre de l'indexation. Les instructions de la norme (*NF Z 47-102*, 1978) concernant l'analyse du document (la première étape de l'indexation) stipulent que le document ne doit pas être lu dans son intégralité, mais qu'une attention particulière doit être portée aux titres et sous titres (ainsi qu'aux résumés, paragraphes d'introduction et de conclusion, ainsi qu'aux légendes des illustrations).

Afin de refléter ces instructions, qui sont rigoureusement observées par les documentalistes dans leur pratique quotidienne, certains systèmes d'indexation automatique attribuent effectivement un poids supplémentaire aux mots clés qui sont extraits du titre du document (MTI (Aronson, Mork, Gay, Humphrey, & Rogers, 2004)) ou des titres de paragraphes (NOMINDEX (Pouliquen, 2002)). Nous avons nous même mené une expérience comparant les indexations issues de documents entiers ou des titres de paragraphes de ceux-ci (Névéol, 2004), qui montrait l'intérêt de ce type d'approche pour la précision de l'indexation. En revanche, le rappel était peu satisfaisant.

Cependant, en observant les titres de quelques documents « Le diabète et les maladies rénales », « Film : techniques d'immobilisation plâtrées », « Publication n° 03-011-02 » et de quelques paragraphes « Comment le diabète affecte-t-il les reins ? », « Partie 1 », « Conclusions et recommandations », il apparaît que tous les titres n'ont pas le même pouvoir informatif et n'ont pas le même degré d'adéquation avec le contenu sémantique du paragraphe ou du document auquel ils se rapportent. On distingue dans ces quelques exemples des instances des deux premières définitions du titre données par le Robert ; d'une part le nom donné au document et d'autre part la désignation du sujet traité. Genette (1987) considère ces derniers comme des titres « thématiques », par opposition aux titres « rhématiques » qui dénotent la *fonction* du document, plutôt que son *contenu*[20]. Virbel (2002) modifie quelque peu cette classification afin de rendre compte des titres qui relèvent des deux catégories et donnent des informations à la fois sur la fonction et sur le contenu du document. Une étude sur la sémiotique du titre d'une nouvelle d'Edgar Poe (Barthes, 1985) met en avant le rôle d'ouverture au texte joué par le titre. Ainsi, un titre thématique permet à l'auteur d'exprimer ce qui, selon lui, constitue l'aspect le plus important de son oeuvre. Nous pouvons transposer cette remarque à tous les types de documents, considérer effectivement le titre comme contenant les concepts essentiels traités dans le document.

[19] A ce sujet voir par exemple (Nakache & Métais, 2005)

[20] Pour la constitution des notices CISMeF, il faut remarquer que l'on s'intéresse à la fois au contenu (titre, mots clés, catégorisation...) et à la fonction du document (type de ressource).

5.5.2 Extraction Automatique de titres

En pratique, l'extraction automatique de titres de ressources (c'est-à-dire, l'extraction du titre d'un document - ou des titres de paragraphes- à partir d'une URL) est un problème complexe qui relève de l'ingénierie. En effet, il faut prendre en compte la multiplicité des formats de ressources existants (html, pdf, doc, ppt etc.) ainsi que les différents modes d'encodage possibles des informations relatives aux titres (balises <titre>, <H1>, métadonnées...) quand cet encodage existe, ce qui est loin d'être systématique. Malgré les efforts de standardisation des formats de documents électroniques prônés par le W3C dans le cadre du développement du Web Sémantique, le format des ressources de la Toile est encore très hétérogène.

Nous avons réalisé une première expérience d'extraction automatique de titres sur les deux formats de ressource les plus fréquemment traités par CISMeF - les formats html (ou assimilés - htm, asp, php, etc.) et pdf. Pour les documents html, le titre extrait est celui contenu entre les balises « titre » ou « title » du document html.

<title> Titre HTML </title>

Dans le cas d'un balisage multiple, les différents titres candidats sont concaténés. De même, pour les documents pdf, le titre extrait est celui contenu dans la balise dédiée.

/Title(Titre PDF)

A l'aide de cette méthode, nous avons procédé à l'extraction automatique du titre de 554 URLs de ressources « en attente » d'être intégrées au catalogue CISMeF. Le tableau 5.5 présente les résultats obtenus.

Nb ressources	%	**Extraction Titre**	**Exemple de Titre**	**Cause**
69	20	Pas d'extraction	-	Format non traité : .ppt, .doc, .zip, .rtf
57	17	Problème d'extraction	-	ressource sans titre, format non conforme
43	13	Extraction non significative	« Document sans titre », « Acrobat PDFWriter 4.05 pour Windows »	L'auteur n'a pas spécifié de titre significatif
38	11	Titres rhématiques	« Index », « Généralités »	L'auteur a choisi un titre ne décrivant pas le contenu
132	39	Extraction significative	« Trisomie 21 », « Fondements de l'ergothérapie »	L'auteur a choisi un titre significatif

Tab. 5.5 – Extraction automatique de titre pour 339 ressources (URLs accédées le 21/11/2004)

On constate que la méthode proposée offre une précision d'au moins 39% sur les ressources auxquelles il a été possible d'accéder (soit 339 ressources, 151 ressources renvoyaient un code

d'erreur 404 le 21/11/04 et 64 URLs supplémentaires renvoyaient également un code 404 le 23/06/05 lors de la suite de nos expériences).

L'extraction du titre à partir du texte de la ressource même pourrait améliorer ces performances sur des formats tels que .doc et .rtf, ainsi que pour les documents pdf ne contenant aucun titre dans la balise prévue (silence observé dans 37% des cas). En revanche, il semble difficile de distinguer les titres non significatifs (24% des cas dans notre expérience, soit 48% des titres effectivement extraits) des titres significatifs (c'est dire, les titres « thématiques » constituant réellement une description du contenu sémantique de la ressource).

Malgré un silence relativement important (37% des URLs accédées), les résultats de cette première expérience sont encourageants dans la mesure où 52% des titres effectivement extraits sont pertinents. Afin d'améliorer la méthode d'extraction, nous avons décidé de prendre en compte d'autres formats de ressources (.doc, .rtf et .ppt). Dans ce cas, le titre est extrait du document même par application de l'heuristique suivante (H) :

> *« conversion au format texte et extraction de la première ligne du document texte obtenu ». (H)*

Par ailleurs, nous effectuons une reconnaissance minimale (par comparaison brute de chaînes de caractères) des types de titres non significatifs récurrents, c'est-à-dire ceux contenant les mots « Microsoft » et « Adobe », ou comprenant moins de 15 caractères. Dans ce cas, nous appliquons également l'heuristique (H). La figure 5.2 présente l'algorithme complet d'extraction de titre à partir d'une URL.

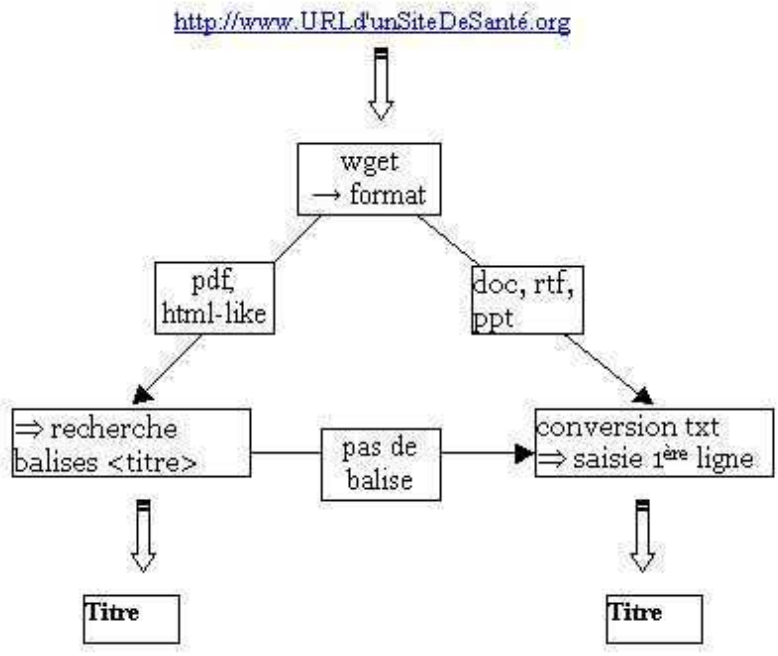

FIG. 5.2 – Algorithme d'extraction du titre d'une ressource

Afin d'évaluer l'impact de ces modifications sur l'extraction effective des titres, nous avons appliqué le nouvel algorithme sur les 339 URLs précédemment traitées. Le tableau 5.6 présente les résultats obtenus lors de cette seconde expérience.

Nb ressources	%	**Extraction Titre**	**Exemple de Titre**	**Cause**
32	9	Pas d'extraction	-	Format non traité : .zip, PDF protégé
57	17	Problème technique d'extraction	-	PDF « image », diapositives PPT imbriquées dans html...
21	6	Extraction non significative	« BMbioch.fm », « Corel Office Document »	L'auteur n'a pas spécifié de titre significatif
57	17	Titres rhématiques, site éditeurs, ...	« Publications , 04-707-02 » « Faculté de Médecine de Lyon Sud : cours »	Le titre extrait ne décrit pas le contenu
172	51	Extraction significative	« Trisomie 21 », « Fondements de l'ergothérapie »	L'auteur a choisi un titre significatif

Tab. 5.6 – Extraction automatique de titre pour 339 ressources (URLs accédées le 23/06/2005)

Les améliorations apportées semblent pertinentes, dans la mesure où le silence est réduit à 26% et le nombre de titres correctement extraits augmente sensiblement (51%).

Par ailleurs, la part des titres rhématiques et autres informations liées à la ressource augmente également. Ce type d'extraction est erronée en ce qui concerne le titre de la ressource, mais peut s'avérer utile dans le cadre d'une utilisation supervisée : l'information pourra être ré-utilisée lors de la validation humaine.

Parmi les ressources pour lesquelles aucune extraction n'est possible (N=32), la majorité sont des fichiers pdf protégés par un mot de passe. Il n'est donc pas possible d'accéder au contenu automatiquement. Cependant dans certains cas, le nom du fichier semble significatif et peut tenir lieu de titre - ex : « retrovirus.pdf », « SyndromeAlcoolisationFoetale.pdf ».

Cette deuxième expérience permet de mieux cerner les limites de l'extraction automatique de titres. Il n'est pas possible à notre connaissance d'extraire certains titres pour des raisons techniques (pdf image, pdf protégé par mot de passe, diapositives ppt imbriquées dans html). Du fait de l'hétérogénéité des formats et de la qualité des ressources, notre méthode est mise en échec par certains fichiers (principalement html et pdf avec balisage non standard).

L'accès au « texte » de la ressource pour pallier l'absence de titre dans les balises habituelles (html et pdf) s'avère utile dans certains cas, mais le bilan reste mitigé. En effet, il arrive que la première ligne du fichier ne corresponde pas au titre de la ressource mais :

- A un numéro de page (« 1 », « -78- », ...)
- Au nom de l'auteur ou du site éditeur(« Maîtrise STAPS - Année 2002-2003 - Université Paris 13 - UFR SMBH Didier Chapelot »)
- Au texte lui-même (« Un interrogatoire plus approfondi du patient vous informe qu'il a été traité l'année précédente par Terbinafine pendant 9 mois pour une onychomycose

et qu'il a des antécédents d'hépatite virale dans l'enfance. »)

Les problèmes rencontrés se rapportent donc à la fois à la structuration du document par l'auteur et à la lisibilité du document par la machine. Il est possible d'apporter une solution à certains de ces obstacles, mais il reste difficile d'attaquer le problème dans sa globalité - afin de procéder à une extraction de titre efficace, il serait souhaitable de pouvoir reconnaître le type de difficulté présentée par un document en particulier, ce qui impliquerait notamment de disposer d'une sorte de grammaire des titres, permettant de reconnaître un titre possible d'un titre erroné. Ce travail sort du cadre de notre approche « baseline », et nous nous contentons de la suggérer sans la mettre en oeuvre.

Néanmoins, le système d'extraction de titres présenté ci-dessus, qui a été développé au cours de cette thèse, sera utilisé prochainement par l'équipe CISMeF pour extraire automatiquement les titres des ressources à indexer dans le catalogue (URLs en attente). Une correction manuelle permettra de rectifier les titres erronés et de réutiliser les informations de type « auteur » ou « type de ressource » qui auraient pu être extraites à la place du titre. Par ailleurs, ce système est également intégré au système d'indexation automatique MAIF (cf. section 5.8) qui sera utilisé pour indexer automatiquement certaines des ressources en attente (en particulier les ressources pédagogiques).

5.6 Indexation Manuelle, Automatique, Assistée

L'indexation peut être caractérisée par le type d'unités descriptives utilisées pour la réaliser, mais également par les moyens mis en oeuvre pour extraire ces unités descriptives : l'homme et la machine sont *a priori* capables de réaliser un tel travail, avec un résultat toutefois souvent différent. Après avoir défini brièvement les notions d'« indexation manuelle », « indexation automatique et « indexation assistée », nous tentons de les caractériser.

5.6.1 Définitions

L'indexation manuelle dénote l'attribution de descripteurs (issus d'un vocabulaire contrôlé ou libres) à un document (texte, image, livre etc.) par une personne humaine, sans utilisation d'un système automatique d'indexation.

L'indexation automatique dénote l'attribution de descripteurs (issus d'un vocabulaire contrôlé ou libres) à un document (texte, image, livre etc.) par un logiciel ou un ensemble de programmes informatiques, sans qu'il y ait d'intervention humaine (autre que la conception et la mise en marche du système)

L'indexation semi-automatique, parfois appelée indexation assistée, est un compromis entre les deux méthodes précédentes. Le plus souvent, une indexation automatique sert de support à l'indexeur humain, qui corrige et enrichit la liste de descripteurs extraits automatiquement. De manière plus générale, l'indexation semi-automatique dénote un système informatique nécessitant une intervention humaine à une ou plusieurs étapes de son fonctionnement afin d'attribuer une liste de descripteurs au document indexé.

L'indexation manuelle présente l'avantage d'être précise, mais se révèle très coûteuse en temps et requiert une formation solide pour les indexeurs. L'indexation automatique présente l'avantage d'être très rapide, mais peut s'avérer souvent erronée. L'indexation semi-automatique a pour but de combiner les avantages de ces deux méthodes et de permettre une indexation rapide et précise.

5.6.2 L'indexation Manuelle

L'indexation manuelle est une tâche complexe, nécessitant une connaissance approfondie du domaine concerné par les documents à indexer, du thésaurus utilisé s'il s'agit d'une indexation contrôlée, des règles d'indexation et des besoins des utilisateurs susceptibles de consulter les documents indexés. L'indexeur doit faire preuve de capacités d'analyse afin d'extraire les concepts importants des documents et de synthèse afin de résumer les notions abordées en quelques mots clés. Bien que de nombreux systèmes d'indexation automatique existent à l'heure actuelle, l'indexation manuelle reste privilégiée pour une indexation thématique. Son inconvénient principal est le coût humain de la tache. Par exemple, dans le cadre de CISMeF, une cinquantaine de ressources sont indexées par l'équipe de documentalistes chaque semaine, soit un coût moyen d'environ 2 heures par ressource.

Plusieurs solutions sont envisagées pour réduire les délais d'indexation. Par exemple, des catalogues comme Yahoo ! ou CISMeF demandent aux utilisateurs qui soumettent des sites à répertorier de les accompagner de mots clés. Ainsi, la simple vérification de l'indexation proposée permet de gagner du temps. D'autres projets prennent le parti d'utiliser in extenso l'indexation soumise par les utilisateurs pour les documents de la base. Par exemple, le projet « gimp-savy » a pour but de mettre à la disposition du public une base d'images libre de droits entièrement indexée par des utilisateurs volontaires, qui peuvent indexer des images auxquelles aucun mot clé n'a été attribué[21] ou ajouter des mots clés qui aurait été omis pour les images déjà indexées. L'inconvénient de cette méthode est qu'elle ne permet pas de corriger les erreurs d'indexation constatées et qu'elle laisse la base à la merci de plaisantins ou d'internautes mal intentionnés qui peuvent bruiter l'indexation[22]. Cependant, dans cet esprit, l'approche communautaire proposée par les wikis[23] permettrait de contourner cette difficulté et d'homogénéiser les indexations. Ainsi, dans le cadre du projet UMVF, l'équipe CISMeF a mis en place une interopérabilité avec la faculté de Grenoble permettant l'intégration automatique des notices réalisées par les enseignants de cette université afin de référencer leurs ressources pédagogiques dans CISMeF. Cependant, il s'avère que la qualité des métadonnées fournies dans ce cadre se démarque notablement de celles habituellement fournies par les documentalistes de l'équipe. En conséquence, de nombreuses corrections doivent être apportées par les documentalistes CISMeF sur ces notices. Cette expérience récente montre que l'indexation distribuée est une stratégie qui nécessite une excellente coordination ainsi qu'un professionalisme de la part des acteurs pour être vraiment efficace.

D'un point de vue méthodologique, il existe un consensus (Anderson & Perez-Carballo, 2001) selon lequel l'indexation manuelle s'effectue en deux étapes :

1. Analyse du texte
2. Traduction dans le vocabulaire contrôlé (le cas échéant)
   Auxquelles peut s'ajouter une troisième :
3. Relecture, révision, application de règles d'indexation

[21] La liste des images qui restent à indexer est disponible à l'URL `http://gimp-savy.com/cgi-bin/unindexed.cgi`. Une autre page du site (`http://gimp-savy.com/PHOTO-ARCHIVE/tips_on_indexing.html`) explique la démarche à suivre pour indexer les images et donne quelques exemple.

[22] Par exemple, les mots clés « Ivan » et « Cécile » attribués à une image représentant des cétacés

[23] « Un wiki est un site Web dynamique permettant à tout visiteur de modifier les pages à volonté. Il permet non seulement de communiquer et diffuser des informations rapidement (...), mais de structurer cette information pour permettre d'y naviguer commodément ». Pour plus de détails, sur les wikis, nous invitons le lecteur à consulter l'article complet de l'encyclopédie « wikipédia » sur `http://fr.wikipedia.org/wiki/Wiki` (accédé le 03/08/05)

En fait, c'est cette dernière étape qui fait la spécificité de l'indexation manuelle car c'est à ce moment que se fait le choix des descripteurs les plus représentatifs du texte en fonction du profil des utilisateurs et des règles éditoriales propres à la collection documentaire. C'est à cette étape que le jugement de l'indexeur entre en jeu. (David et al., 1995) l'estiment cruciale :

> « *le moment critique dans le processus [de l'indexation] est celui de la décision de conserver ou de rejeter un descripteur. En d'autres termes, bien que les indexeurs suivent tous les mêmes procédures (...) pour analyser les documents, leurs critères d'appréciation de ce qui constitue une bonne indexation semblent varier.* »[24]

5.6.3 L'indexation Automatique

Méthodes d'indexation

Nous présentons ci-dessous les deux grandes catégories de modèle d'indexation et de recherche d'information : le modèle vectoriel et le modèle probabiliste. Nous détaillons également le principe de l'indexation sémantique latente, qui représente l'un des modèles intermédiaires les plus courants. Finalement, nous discutons l'adéquation de ces modèles à un contexte de l'indexation contrôlée et nous présentons les particularités de ce type d'indexation.

Le Modèle Vectoriel. Le modèle vectoriel (en anglais, « vector space system ») a été introduit par Salton à la fin des années 1970 et amélioré pendant les années qui ont suivi. Il est par exemple décrit dans (Salton & McGill, 1983) ou (Salton, 1989) et implémenté dans le système SMART. Il met en oeuvre les trois étapes présentées ci-dessus. L'analyse du texte consiste en une segmentation du texte en mots (séquences de caractères séparés par un espace). La traduction dans le langage d'indexation consiste en une racinisation des termes La sélection des candidats termes utilise un anti-dictionnaire (ou stoplist) pour supprimer de la liste de candidats les mots les plus fréquents, susceptibles de fausser la représentation du contenu sémantique du texte. Il également possible de sélectionner les termes à l'aide de leur coefficient de discrimination (cf. figure 5.3). Le coefficient de discrimination $C\_disc(k)$ d'un terme k est calculé par comparaison du taux de similarité moyen des documents de la collection avec et sans le terme k : $C\_disc(k) = Sim(k) - Sim$

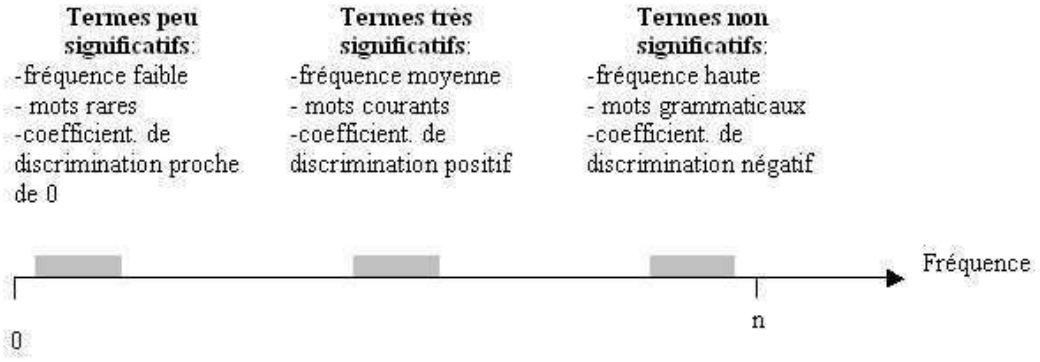

FIG. 5.3 – Sélection des termes en fonction de leur fréquence- d'après (Salton & McGill, 1983)

[24] Notre traduction.

Par suite, un score est attribué à chacun des termes retenus. Ce score peut-être fondé sur la fréquence du terme dans le document (*term frequency*, tf) et l'inverse de la fréquence du terme dans les documents de la collection (*inverse document frequency*, idf). Le calcul du poids reflète l'intuition que l'importance d'un terme doit être (1) proportionnelle à sa fréquence le document (plus un terme est fréquent dans le document, plus il est représentatif) et (2) inversement proportionnelle au nombre total de documents dans lequel il apparaît (plus un terme apparaît dans un grand nombre de documents, moins il est discriminant). Une normalisation de ce poids peut également être ajoutée, afin de prendre en compte la longueur du document. Cette méthode de calcul de poids est appelée *tf*idf* et accorde une importance élevée aux termes fréquents dans peu de documents d'une collection.

Finalement, on peut dire que le poids du terme t dans le document d dépend des trois facteurs suivants :

1. la pondération locale de t dans d (nombres d'occurrences de t dans d)
2. la pondération globale de t dans la collection (nombre de documents où t apparaît)
3. la normalisation appliquée

En fonction de la méthode de calcul de chacun de ces facteurs (naturelle, notée n, logarithmique notée l ou cosine notée c) on obtient une série de schémas de pondération tf*idf désignés par trois lettres correspondant à chacun des facteurs. Le tableau 5.7 résume les combinaisons possibles.

Occurrences du terme dans le document	**Fréquence du terme dans les documents de la collection**	**Normalisation**
n (naturel) $tf_{t,d}$	n (naturel) df_t	n (pas de normalisation)
l (logarithmique) $1 + log(tf_{t,d})$	$t \qquad log(\frac{N}{df_t})$	c (cosinus) $\frac{1}{\sqrt{w_1^2+w_2^2+...+w_n^2}}$
a (augmenté) $0,5 + \frac{0,5*tf_{t,d}}{max_t(tf_{t,d})}$	-	-

TAB. 5.7 – Les différents schémas tf*idf - d'après (Manning & Shütze, 2000). $tf_{t,d}$ représente le nombre d'occurrences du terme t dans le document d, df_t représente le nombre de documents où t apparaît, N représente le nombre total de documents dans la collection et w_t représente le poids du terme t.

L'Indexation Sémantique Latente. Le modèle d'Indexation Sémantique Latente (en anglais, « Latent Semantic Indexing » ou LSI) est fondé sur le modèle vectoriel, dont il propose d'élargir le champ. Introduit par Deerwester, Dumais, Furnas, Landauer, and Harshman (1990), il réfute l'hypothèse sous-jacente du modèle vectoriel selon laquelle tous les termes utilisés pour représenter un document sont indépendants. Le modèle cherche alors à découvrir les dépendances latentes entre les termes d'indexation, afin de les exploiter dans la représentation des documents. En pratique, ce modèle cherche à pallier le problème posé par les synonymes (source de silence en Recherche d'Information) et les homographes (source de bruit). Ainsi, le modèle LSI propose de transformer la représentation vectorielle classique afin d'utiliser une

nouvelle représentation permettant de rapprocher les documents sémantiquement proches, plutôt que ceux qui partageraient simplement le plus de termes. Intuitivement, il repose sur l'idée qu'un document indexé avec des termes tels que « insuline » ou « glycémie » peut être pertinent pour une requête sur « diabète » même s'il n'est pas indexé avec « diabète » dans la mesure où les trois termes « insuline », « glycémie » et « diabète » sont des co-occurrents fréquents. En pratique, la nouvelle représentation de la collection est obtenue à partir de la matrice composée des représentations vectorielles de chaque document. Cette matrice est décomposée selon le procédé de « Décomposition en Valeurs Singulières » grâce auquel les valeurs singulières - non significatives - sont écartées et une nouvelle représentation de taille inférieure peut être obtenue. Deerwester et al. (1990) montrent que ce type de représentation dans le cadre d'un système de recherche d'information offre des performances équivalentes ou supérieures à une représentation vectorielle, telle qu'implémentée dans SMART. Cependant, l'un des inconvénients de la LSI est la difficulté d'interprétation de la représentation finale - ce qui peut rendre ce modèle difficilement adaptable à un problème d'indexation contrôlée.

Une autre méthode d'indexation prenant en compte la non-indépendance entre les termes de la collection sélectionnés pour la représentation des documents consiste à utiliser une matrice de co-occurrence pour modifier la représentation vectorielle classique (cf. par exemple les travaux de (Besançon, Chappelier, Rajman, & Rozenknop, 2001)). Dans ce cas, la matrice des représentations vectorielles de la collection est multipliée par la matrice de co-occurrence des termes. On obtient alors une représentation dans le même espace, donc plus facilement interprétable que dans le cas de la LSI.

Le Modèle Probabiliste. Le modèle probabiliste, introduit dans le cadre de la Recherche d'Information (Sparck-Jones, Walker, & Robertson, 2000), cherche a estimer la probabilité de pertinence d'un document de la collection par rapport à une requête formulée par un utilisateur. Le poids attribué à chaque unité descriptive d'un document reflète donc la probabilité que ce terme soit un descripteur pertinent pour le document en question. Ce modèle, implémenté dans le système OKAPI[25] est décrit précisément dans (Sparck-Jones et al., 2000).

Cas de l'indexation contrôlée. Ces trois systèmes (vectoriel, LSI et probabiliste) utilisent au départ des mots simples pour l'indexation. Ils ont tous les trois été introduits dans le cadre de la Recherche d'Information et plus particulièrement celui des conférences TREC. Ainsi, les travaux dont ils sont issus ont écarté d'emblée la possibilité d'utiliser un vocabulaire contrôlé pour l'indexation et se sont concentrés sur l'aspect exhaustif de la description, plus que sur l'aspect descriptif. Ainsi, les modèles OKAPI et LSI s'avèrent difficilement transposables dans le contexte d'une indexation contrôlée, ou même fondée sur des unités descriptives linguistiquement complexes assimilables à des mots clés. Par contre, le modèle vectoriel, plus simple, a servi de point de départ pour beaucoup de travaux. Ainsi, certains systèmes, bien qu'ils n'utilisent pas un vocabulaire contrôlé particulier, opèrent un regroupement des variantes des termes afin de réduire l'espace d'indexation et de sélectionner des unités significatives. On compte parmi ces systèmes FASTR (Jacquemin, 1994), INTDoc (Aït El Mekki & Nazarenko, 2004), ou le système d'indexation développé par (Lahtinen, 2000). D'autres approches utilisent des ressources comme WordNet pour effectuer des regroupements dans des tâches d'indexation ou de clustering (Hunh, Wermter, & Smith, 2004).

[25]Les fonctionalités d'OKAPI pour l'indexation et la recherche d'information (entre autres tâches) sont disponibles pour l'anglais, l'arabe et le chinois sur`http://www.lemurproject.org/lemur/overview.html`.

Dans le cadre d'une indexation contrôlée, il est fréquent d'utiliser des automates-dictionnaires pour repérer les unités d'indexation (Gaudinat, Boyer, Baujard, & Ruch, 2002), (Pouliquen, 2002) ou (Lovis, 1996). Une autre stratégie consiste à traiter l'indexation contrôlée comme un problème de catégorisation supervisée. Ainsi, les méthodes de classifications usuelles telles que les k Plus Proches Voisins (Lam et al., 1999; Yang & Chute, 1994), les SVM (Cai & Hofmann, 2004), ou la méthode de Rocchio (Rocchio, 1971) peuvent être envisagées. En parallèle avec ces méthodes d'indexation, qui permettent d'obtenir une liste de descripteurs candidats, certains travaux portent sur la sélection des descripteurs candidats. Ce filtrage peut être opéré de plusieurs manières :

- Soit en réduisant judicieusement le texte à indexer, afin de ne conserver que la partie où les descripteurs pertinents sont susceptibles de se trouver. Ainsi, certains systèmes d'indexation utilisent le résumé des textes à indexer et non le texte intégral (MeSHMap, MTI). Des travaux récents étudient également la possibilité de sélectionner efficacement certaines portions du résumé (Ruch et al., 2005). D'autres travaux sur les textes entiers montrent que la sélection des propositions principales est une méthode de réduction avantageuse du bruit de l'indexation (Corston-Olivier & Dolan, 1999).
- Soit en révisant la liste de candidats extraits. La méthode la plus simple consiste à sélectionner les N premiers candidats extraits, supposés être les plus pertinents. La valeur de N peut être fixée arbitrairement, ou en fonction du « break-even point » moyen observé sur un jeu de test (Ruch, Baud, & Geissbühler, 2003; Gaudinat et al., 2002). D'autres systèmes, comme MTI, appliquent un certain nombre de règles d'indexation qui permettent de compléter la liste de candidats et une méthode de filtrage permettant de supprimer d'autres candidats, non nécessairement en fin de classement. Des travaux récents (Joubert, Peretti, Gouvernet, & Fieschi, 2005) étudient l'attribution d'une pondération majeur/mineur aux candidats, ce qui permet par exemple de sélectionner les descripteurs majeurs.

5.7 Comparaison des systèmes d'indexation MeSH

Comme nous l'avons exposé au paragraphe 5.6.3, il existe de nombreuses approches pour l'indexation automatique, y compris pour une tâche d'indexation contrôlée comme l'indexation MeSH. Plusieurs méthodes ont été implémentées dans les systèmes existants. Le tableau 5.8 a pour objet de résumer leurs principales caractéristiques, afin d'offrir une comparaison des systèmes MTI (Aronson et al., 2004), MeSHMap (Ruch et al., 2003), HON-MeSHMapper (Gaudinat et al., 2002), NOMINDEX (Pouliquen, 2002) et MAIF (cf. section 5.8) Deux évaluations comparatives de ces systèmes ont été réalisées au cours de cette thèse. Les résultats sont présentés dans les sections 5.9.2 et 5.9.3.

NOMINDEX (Pouliquen, 2002). L'objectif de NOMINDEX est de reconnaître les concepts médicaux contenus dans une phrase en langue naturelle et de les utiliser pour créer une base de données de documents consultable facilement. Nomindex utilise un lexique tiré de l'ADM (Aide au Diagnostic Médical) (Lenoir, Michel, Frangeul, & Chales, 1981), qui contient 130 000 termes, synonymes, mots composés, préfixes et suffixes. Dans un premier temps, les mots du document à indexer sont mis en correspondance avec les termes ADM. Par exemple,

[26] Pour une description des méthodes statistiques employées dans MTI, nous invitons le lecteur à se reporter à la description du système disponible à l'URL `http://ii.nlm.nih.gov/mti.shtml` (accédé le 01/07/05).

Syst. / Caract.	MeSHMap	HON-MeSHMapper	NOMINDEX	MAIF	MTI
Lexique et synonymes	MeSH	UMLS	MeSH et ADM	MeSH et CISMeF	UMLS
Unités d' indexation	Termes isolés	Termes isolés	Termes isolés	Termes isolés ou paires	Termes isolés
Extraction des concepts	Termes entiers Termes partiels (mots composés)	Termes entiers Termes partiels	Termes entiers	Termes entiers	Termes entiers Termes partiels tri-grammes
Gestion de la variation terminologique	Autorise l'insertion d' un mot ou d'une lettre dans les termes	Fenêtre de recherche des mots composés	Informations flexionelles	Informations flexionelles et dérivationelles	Fenêtre de recherche des mots composés analyse des variantes terminologiques
Règles d' indexation	-	-	-	MeSH + CISMeF	NLM
Hiérarchie	Non	Oui	Non	Oui	Oui
Structure du texte	Titre	-	Titres et sous-titres	Titre	Titre
Méthode	Vector Space (VS) System	TAL	TAL	TAL + k-PPV	TAL/VS+ n-grammes+ related citations
Statistiques	tf*idf sur le thésaurus MeSH	tf*idf	tf*idf sur la collection considérée	tf*idf sur les notices CISMeF « probabilités » de sélection	Utilisation avancée[26]
Langue	EN+FR	EN+FR	FR	FR	EN
Taille de l'index	fixe	fixe	fixe	variable	fixe

TAB. 5.8 – Comparaison des systèmes d'indexation MeSH

« céphalée »est rattaché à « mal de tête ». Ensuite, les termes de l'ADM sont rattachés à leurs équivalents MeSH, ainsi qu'à leur Identifiant Unique (CUI) dans l'UMLS. Un score fondé sur tf*idf est calculé pour chaque concept dentifié. Cette analyse est exploitée dans plusieurs applications : l'indexation, la recherche de documents similaires et la synthèse de documents.

HONMeSHMapper (Gaudinat, 2002). Le système HONMeSHMapper a été développé en 1997 en parallèle avec MARVIN (Multi-Agent Retrieval Vagabond on Information Networks) dans le but d'indexer et de rechercher des documents de santé en ligne. HONMeSHMapper fait partie d'un extracteur terminologique plus générique, fondé sur les ressources de l'UMLS. Adapté à l'extraction MeSH, il a été utilisé dans le cadre du projet européen

WRAPIN (Gaudinat et al., 2004) et fait partie des outils dont disposent les évaluateurs lors de la procédure d'accréditation des sites de santé de HON. HONMeSHMapper est fondé sur une séries d'expressions régulières reconnaissant les mots MeSH simples ; les termes MeSH composés de plus d'un mot sont alors identifiés à l'intérieur d'une fenêtre de cinq mots. Une approche sac-de-mots permet de prendre en compte la distribution des composants des termes MeSH composés identifiés dans tout le document. Finalement, un premier score inversement proportionnel à leur fréquence répertoriée par MedHunt est attribué aux termes extraits. Ce score est ensuite pondéré par la place du terme dans la hiérarchie MeSH.

MeSHMap (Ruch, 2003). MeSHMap aborde l'indexation MeSH comme un problème de recherche d'information : le document à indexer est considéré comme une requête en langage naturel, à laquelle il faut associer des documents issus du thésaurus MeSH, les mots clés. Cette approche repose sur l'indexation de résumés d'articles scientifiques, par opposition à des textes intégraux. MeSHMap a un fonctionnement similaire a celui de HONMeSHMapper, dans la mesure où il combine deux types de classifieurs :

1. l'un, fondé sur des expressions régulières, qui extrait les composants MeSH (en autorisant certaines variations, comme l'insertion d'un caractère - par exemple « insulino-dépendant » sera reconnu pour le composant « insulinodépendant ») et les termes MeSH formés de ces composants (à ce niveau, la variation terminologique est prise en compte en autorisant l'insertion d'un mot à l'intérieur des termes - par exemple, « maladies très rares » sera reconnu pour le terme « maladies rares »).
2. l'autre, fondé sur les sacs de mots, qui extrait les racines des composants sur l'ensemble du texte et utilise la méthode « Vector Space » présentée plus haut pour attribuer un score à chacun des termes du thésaurus MeSH.

La première approche privilégie en principe la précision puisque seuls les termes MeSH effectivement présents dans le texte peuvent être extraits. La seconde approche privilégie au contraire le rappel, puisque tous les termes MeSH contenant au moins l'une des racines identifiées sont considérés comme candidats à l'indexation. Ruch et al. (2003) ont montré que la combinaison de ces approches donne des résultats supérieurs à ceux obtenus avec chacune des méthodes séparément.

Medical Text Indexer (MTI) (Aronson, 2004). MTI est issu de la combinaison de deux approches d'indexation MeSH, à savoir une approche fondée sur le Traitement Automatique de la Langue Naturelle implémentée dans le système MetaMap (MMI) et une approche statistique appelée « PubMed Related Citations » (PRC). MTI intègre les resultats obtenus grâce à une méthode de post-traitement originale.

1. MetaMap (Aronson, 2001) permet d'analyser un texte et d'en extraire des concepts de l'UMLS auxquels un score est attribué selon leur fréquence et leur pertinence. Les concepts UMLS sont ensuite restreints aux termes MeSH correspondants (Bodenreider, Nelson, Hole, & Chang, 1998).
2. L'algorithme PRC (Kim et al., 2001) extrait une liste ordonnée de termes MeSH à partir d'un titre et d'un résumé d'article en recherchant les articles plus proches dans la base MEDLINE. Cette recherche s'effectue sur la base des mots en commun, en tenant compte de la longueur relative des résumés.

Les deux listes de termes obtenues grâce à ces approches sont combinées de manière pondérée (poids de 7 pour MMI vs. 2 pour PRC) et en tenant compte des co-ocurrences et de la présence des termes dans le titre. Cette liste finale de candidats est ensuite révisée à l'aide d'une série de règles d'indexation destinée à filtrer les termes MeSH non pertinents. Trois niveaux de filtrage sont possibles, afin de privilégier soit la précision, soit le rappel.

1. Le filtrage « strict » ne conserve que les candidats extraits par les deux approches : la liste obtenue comporte peu de termes (précision élevée, rappel faible)
2. Le filtrage « medium » conserve les meilleurs candidats extraits par chaque méthode et élimine les termes trop généraux : la liste obtenue comporte un bon nombre de termes (précision et rappels moyens)
3. Le filtrage « minimum » est appliqué avant en toute circonstance et avant les filtrages « medium » ou « strict » qui sont optionnels. Il consiste en l'application des règles d'indexation NLM, ainsi que d'autres règles définies par l'expérience de l'utilisation du système. La liste de candidats obtenue comporte un mélange acceptable de termes corrects et incorrects (précision faible, rappel élevé).

Les règles appliquées dans le filtrage minimum conduisent à l'addition ou à la suppression d'un terme MeSH, à la modification du rang d'un terme de la liste, fondée sur la présence d'autres termes.

5.8 MAIF : MeSH Automatic Indexing for French

Nous présentons plus particulièrement dans cette section le système MAIF (MeSH Automatic Indexing for French) développé dans le cadre de notre thèse. La Figure 5.4 présente une vue d'ensemble du fonctionnement de MAIF, destiné à produire une indexation à l'aide de descripteurs MeSH (mots clés ou paires mot clé/qualificatif) à partir de l'URL d'une ressource.

Ainsi, le système comporte une étape préliminaire (« pré-traitement ») au cours de laquelle le support nécessaire à l'indexation proprement dite est obtenu à partir de l'URL : il s'agit d'un *texte* pour l'approche TAL et d'un *titre* pour l'approche k-PPV (cf. figure 5.4). Cette étape est détaillée à la section 5.5 ci-dessus. A partir de ces supports, deux approches sont mises en oeuvre pour produire deux listes de descripteurs MeSH candidats : l'approche TAL est détaillée dans la section 5.8.1 et l'approche k-PPV est présentée dans la section 5.8.2.

5.8.1 MAIF : approche TAL (Traitement Automatique de la Langue)

Cette approche aborde le problème de l'indexation comme une *traduction conceptuelle* : à un texte rédigé en langue naturelle (le français), on se propose de faire correspondre une liste de concepts, représentés par des mots clés (ou paires mot clé/qualificatifs) du MeSH. Pour ce faire, on va procéder à une analyse du texte en langue naturelle afin d'en extraire les descripteurs MeSH pertinents.

Fonctionnement du système

L'approche TAL suit les étapes de l'indexation manuelle : analyse du texte pour en tirer les concepts abordés, traduction des concepts dans le langage de l'indexation, à savoir le MeSH, puis révision de la liste de candidats obtenus.

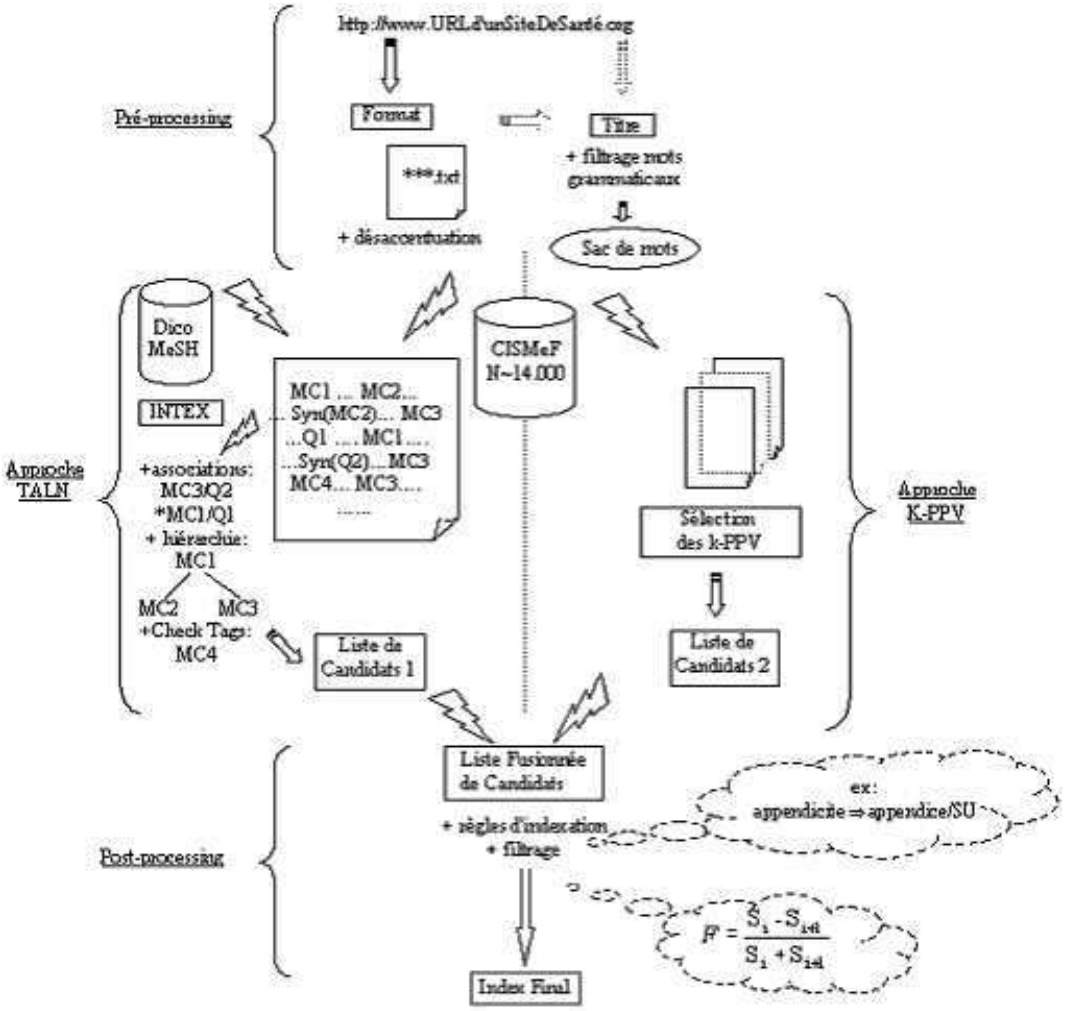

FIG. 5.4 – Processus d'Indexation Automatique implémenté dans MAIF

Extraction de concepts médicaux. De nombreuses solutions sont envisageables pour procéder à l'extraction de concepts médicaux. Nous présentons la démarche que nous avons adoptée et détaillons les choix qui ont été faits.

Rapprochement avec la Recherche de documents. Pour commencer, il est intéressant de rappeler le parallèle qui existe entre la *Recherche de documents* et l'*Indexation des documents*. En effet, lorsqu'une requête est soumise au moteur de recherche « simple » Doc'CISMeF, la première opération effectuée est la traduction de la requête en termes MeSH. C'est précisément le but que nous nous fixons en indexation : traduire un texte en un ensemble de descripteurs MeSH appropriés. Il semble alors pertinent de se demander s'il est possible d'adapter l'algorithme d'extraction MeSH de la recherche d'information pour l'indexation. Cependant, en approfondissant la question, les différences entre les deux tâches surgissent. La principale est le support même de l'extraction MeSH : une requête en langue naturelle soumise à un moteur de recherche comporte au plus une dizaine de mots, peut être mal orthographiée, rédigée en dépit des règles de grammaire et ciblée sur un, deux, voire trois mots clés. A l'inverse, un texte comporte plusieurs centaines de mots, il est construit de façon structurée et peut aborder un nombre important de concepts. La deuxième différence importante réside dans l'objectif derrière l'analyse réalisée : dans le cas de la requête, on privilégiera plutôt le *rappel* afin de répondre de manière exhaustive au besoin d'information de l'utilisateur. A l'inverse, dans le cas du texte, on cherchera à privilégier la *précision* afin de garantir une bonne description des concepts abordés dans le document. Finalement, l'aspect technique de traitement d'une requête constitue une contradiction supplémentaire : chaque requête est segmentée en mots (suites de caractères séparés par un espace ou une ponctuation) ; après filtrage des mots vides, un sac de mot est constitué afin d'interroger la base de données terminologique. Une recherche est ensuite lancée sur le terme CISMeF le plus proche (ou dans certains cas, les deux termes les plus proches). Transposée à l'analyse d'un texte, cette technique s'avère difficilement applicable. Tout d'abord, il nous semble que la représentation d'un texte entier par un sac de mots unique (implantée dans MeSHMap (Ruch et al., 2003)) a des limites non négligeables. Cette représentation implique la perte de toute la structure du texte. En particulier, elle ne permet pas de conserver d'information sur la séquencialité des unités (mots), ni sur leur proximité spaciale. Ces informations peuvent pourtant se révéler déterminantes pour une reconnaissance précise des termes complexes, composés de plusieurs lemmes. Une solution intermédiaire consisterait à conserver le découpage en phrases et à considérer des sacs de mots issus de chaque phrase du texte. Cependant, se pose alors le problème de la taille variable des sacs de mots issus des phrases, et du nombre également variable de termes à reconnaître dans chaque phrase. Ces réflexions nous ont conduit à la conclusion qu'il n'est théoriquement pas pertinent, et techniquement difficilement réalisable de traiter ces deux problèmes de la même manière. En résumé, le traitement d'une requête dans le cadre de la recherche de documents et d'un texte dans le cadre de l'indexation sont deux activités pour lesquelles une séquence textuelle doit être « traduite »dans le langage documentaire utilisé dans le système de Recherche d'Information. Cependant, leur mise en oeuvre doit faire l'objet d'un traitement spécifique pour plusieurs raisons :

- longueur de la séquence à analyser (quelques mots *vs.* quelques centaines de mots)
- finalité de l'extraction de concepts (exhaustivité vs. précision)
- nombre de concepts à extraire (connu : un ou deux *vs.* inconnu et variable : de zéro à une dizaine)

Définition des contraintes. Les tests pratiques effectués au début de cette thèse au sein de l'équipe CISMeF à l'aide d'un logiciel d'indexation MeSH (NOMINDEX) nous ont permis

d'avancer dans la définition des contraintes liées à l'indexation automatique. La nécessité de mettre l'accent sur la *précision* fait l'unanimité parmi les documentalistes. En effet, pour l'indexation MeSH, leur expérience les amène à dire qu'« il est plus rapide d'indexer sans aucune aide, plutôt que de corriger une indexation trop bruitée ». Par ailleurs, l'extraction de termes isolés est aussi à déplorer car elle ne répond pas au besoin réel de l'indexation et garantit la nécessité de corrections supplémentaires. De même, les descripteurs obligatoires devraient être sélectionnés systématiquement. Le dernier point soulevé concerne le nombre de descripteurs candidats proposés par le système automatique : plusieurs dizaines. Il semble souhaitable de réduire la taille de cette liste de manière pertinente, car plus le nombre de termes à consulter est élevé, plus le temps passé à les prendre en compte est long et les bénéfices de l'outil automatique diminuent en conséquence. Notre cahier des charges est donc axé sur les points suivants :

- privilégier la précision
- utiliser des paires mots clé/ qualificatif
- utiliser les descripteurs obligatoires
- proposer un index de taille appropriée

Solutions envisagées. Etant donné notre contrainte principale, la précision, nous avons d'emblée choisi d'écarter toute approche fondée sur la reconnaissance des racines des termes médicaux. En effet, d'une part, (comme nous l'avons mentionné plus haut) il semble difficile de rattacher une racine au terme approprié dans le cas du MeSH (par exemple, *card* correspond à au moins 17 termes MeSH). D'autre part, les racines sont utilisées dans d'autres logiciels afin de privilégier le rappel (Ruch et al., 2003; Gaudinat et al., 2002). Nous nous sommes donc orientée vers la reconnaissance des termes médicaux tels qu'ils apparaissent en corpus - sous forme de lemmes. Dès lors, nous avons envisagé l'utilisation de deux logiciels de traitement automatique de la langue : un logiciel dédié à l'extraction contrôlée de termes, FASTR, et un logiciel plus général d'analyse de corpus, INTEX. Ces deux outils offrent la possibilité d'effectuer des traitements intégrés à un logiciel d'indexation. FASTR (qui a par ailleurs été utilisé avec succès pour l'extraction de termes MeSH sur des corpus anglais (Hamon, 2005)) présente l'avantage de traiter de nombreux cas de variation terminologique, tels que l'inversion des constituants d'un terme (« rare maladie »pour <*maladie rare*>), ou l'insertion de mots à l'intérieur d'un terme (« maladie très rare »pour <*maladie rare*>). Le traitement de ces cas est également possible avec INTEX, mais nécessite une mise en oeuvre spécifique, impliquant la production de ressources (dictionnaires et/ou grammaires locales) à cet usage. Ainsi, pour l'extraction stricte de termes, les deux logiciels semblent équivalents, avec une souplesse supplémentaire pour FASTR. Cependant, l'extraction de termes *isolés* ne correspond qu'à une partie de la tâche que nous nous fixons. En effet, nous souhaitons également être en mesure d'extraire des paires mots clés/qualificatfs MeSH (par ex. <*diabète/prévention et contrôle*>), qui dans la plupart des cas n'apparaissent pas en corpus sous une forme qui peut être apparentée à une variation du terme constitué par la paire. A notre connaissance, FASTR n'inclus pas de solution pour la reconnaissance de tels termes. En revanche, INTEX offre la possibilité de construire des grammaires locales afin de décrire certaines expressions, et donc d'extraire directement des paires mot clé/qualificatif. Finalement, notre choix s'est donc porté sur ce dernier logiciel.

Mise en oeuvre. Nous avons construit les dictionnaires MeSH et la bibliothèque de transducteurs décrits à la section 2.5.1, afin de les appliquer à l'aide d'INTEX. Ainsi, l'extraction des concepts médicaux est réalisée grâce à une analyse de surface consistant à repérer les entrées de nos dictionnaires et transducteurs dans le texte à indexer. Cette méthode a l'avan-

tage d'être précise, dans la mesure où seuls les éléments référencés dans les dictionnaires seront reconnus. Par contre, certains concepts qui ne sont pas exprimés à l'aide de termes ou d'expressions spécifiques ne pourront être extraits. Par exemple, une ressource faisant état d'une étude clinique réalisée en France et en Allemagne devrait être indexée à l'aide du mot clé <*étude comparative*> Le concept n'est explicitement mentionné à aucun endroit de la ressource et seule une analyse globale pourrait permettre de l'extraire. Notre méthode ne traite pas ce type de cas. L'expérience des indexeurs de l'équipe CISMeF révèle que ces cas sont beaucoup plus rares que les cas où les concepts sont effectivement dénotés par des termes ou des expressions précises. Nous avons donc choisi de concentrer nos efforts sur l'extraction de concepts *explicites*.

Traduction en descripteurs MeSH. L'extraction des concepts médicaux à l'aide de transducteurs nous fournit deux types d'informations :

1. le terme ou l'expression tel qu'il a été rencontré dans le texte
2. le mot clé, le qualificatif ou la paire MeSH qui leur a été associée dans le dictionnaire ou la bibliothèque.

Ainsi, une première traduction en descripteurs MeSH est effectuée en même temps que l'extraction des concepts médicaux (grâce aux fonctionnalités pour automates et transducteurs d'INTEX).

Cependant, la liste de descripteurs obtenue à ce stade n'est que préliminaire. Deux types de révisions sont effectuées :

1. Extraction multiple (un même terme fait partie d'une entrée du dictionnaire de termes simples, du dictionnaire de termes composés et d'une expression contenue dans la bibliothèque de transducteurs.) : seule l'occurrence la plus longue est prise en compte.
2. Qualificatif isolé : un appariement « local » avec un mot clé de la même phrase est recherché :
   - En cas de conflit d'appariement (le qualificatif isolé peut être apparié avec plusieurs mots clés), les données statistiques issues du catalogue CISMeF (fréquence des paires en jeu) sont utilisées pour départager les paires possibles.
   - En cas de succès, l'occurrence du qualificatif isolé est remplacée par la paire formée.
   - En cas d'échec, le qualificatif est mis de côté pour un appariement « global » avec l'un des deux mots clés les plus fréquents du texte.

Calcul de score. Un score S_i est calculé pour chaque mot clé (ou paire) i. Conformément au consensus entre les indexeurs de l'équipe CISMeF, nous avons établi une règle d'indexation pour les descripteurs obligatoires. Ils doivent être sélectionnés presque systématiquement : leur score S_i prend une valeur maximum s'ils apparaissent plus d'une fois. Pour les autres mots clés (ou paires) extraits, le score S_i est fondé sur le nombre d'occurences du concept dans la ressource. Dans le cas où plusieurs mots clés d'une même hiérarchie sont extraits (cf. figure 5.5), une réallocation équitable du score du père entre les différents enfants est effectuée, afin de ne retenir que les termes les plus précis.

Par exemple, si nous avons 10 occurrences pour <*diabète*>, 15 pour <*diabète de type 1*>, et 4 pour <*diabète de type 2*> les relations hiérarchiques entre <*diabète de type 1*> et <*diabète*> d'une part, et entre <*diabète de type 2*> et <*diabète*> d'autre part entraineront une modification des scores. <*diabète de type 1*> aura un score de $15 + (10/2) = 20$ et

<maladies endocrines>
<**diabète**>
<**diabète de type 1**>
<diabète expérimental>
<diabète gestationnel>
<**diabète de type 2**>

FIG. 5.5 – Exemple de relations hiérarchiques

<*diabète de type 2*> un score de $4 + (10/2) = 9$, alors que <*diabète*> sera éliminé de la liste des candidats.

Conformément aux indications de Salton (Salton & McGill, 1983) nous calculons un poids $tf * idf$ à partir du nombre d'occurrences, afin de favoriser les termes très fréquents dans la ressource qui sont également suffisamment représentatifs au sein de la collection (le catalogue CISMeF).

Coefficient de sélection. Afin d'améliorer le classement des termes, nous avons envisagé d'adapter une méthode utilisée par (Lahtinen, 2000) dans son travail sur l'indexation libre. Il s'agit de la pondération des scores obtenus avec ce que nous appellerons un *coefficient de sélection*. Ce coefficient, calculé sur l'ensemble des ressources contenues dans le catalogue CISMeF, est défini pour un terme t donné comme indiqué par la formule 5.4. R représente une ressource du catalogue et I_R l'indexation manuelle MeSH pour cette ressource.

$$cs(t) = \frac{Card\{(t \in R) \bigcap (t \in I_R)\}}{Card\{t \in R\}} \tag{5.4}$$

Le coefficient de sélection représente le nombre de ressources dans lesquelles le terme t apparait divisé par le nombre de ressources pour lesquelles t a effectivement été sélectionné comme terme d'indexation. Ainsi, un coefficient de sélection proche de 1 indique que dès que le terme apparait dans le texte d'une ressource, il doit être sélectionné (par exemple, un check tag - le coefficient calculé pour <*nourrisson*> est de 0,63). En revanche, un coefficient plus proche de zéro indique que le terme apparait très souvent dans les ressources, mais n'est que rarement sélectionné. C'est par exemple le cas de termes très généraux comme <*sang*> - le coefficient calculé est de 0,01.

En pratique, nous avons pu calculer un coefficient de sélection pour 8 153 mots clés, à l'aide de requêtes soumises au moteur de recherche Doc'CISMeF[27] :

1. la requête « motclé.pt OU synonyme_motclé.pt [28] » a été utilisée pour évaluer le nombre de ressources dans lesquelles le concept dénoté par le mot clé MeSH <*motclé*> apparaissait en plein texte.
2. la requête « motclé.mc » a été utilisée pour évaluer le nombre de ressources indexées avec le mot clé MeSH <*motclé*>.

[27] En mai 2005.

[28] Dans cette requête, « .pt » désigne une occurrence dans le texte intégral de la ressource. « motclé » désigne le mot clé MeSH traité et « synonyme_motclé » désigne l(es) éventuel(s) synonyme(s) du mot clé.

Nous n'avons pas évalué de coefficient de sélection pour les paires mot clé/qualificatif car il nous semblait difficile d'estimer correctement le nombre d'occurrences des paires en plein texte à l'aide d'une requête.

Avec cette méthode de calcul, les coefficients doivent être compris entre 0 et 1. Cependant, il faut remarquer que certains mots clés sont sélectionnés plus souvent qu'ils n'apparaissent en plein texte par exemple, <*distribution selon sexe*> (score 2,12) : il s'agit de mots clé représentant un concept dénoté plutôt par l'ensemble de la ressource, que par des termes ou expressions précis. Dans certains cas, le mot clé n'apparait même jamais en plein texte, ce qui pose un problème pour le calcul du coefficient (division par 0). Afin de normaliser les coefficients obtenus, nous avons fixé le coefficient des ces deux derniers types de mot clés à 1. Aux mots clés pour lesquels il n'a pas été possible de calculer un coefficient de sélection, nous avons attribué la valeur moyenne calculée sur les autres coefficients, soit 0,16.

5.8.2 MAIF : approche k-PPV (dite des « k Plus Proches Voisins »)

Cette approche aborde le problème de l'indexation comme une *catégorisation* : on se propose de déterminer à quelles catégories MeSH prédéfinies (l'ensemble des mots clés ou paires mot clé/qualificatif du MeSH) s'apparente le texte étudié. Pour ce faire, on va construire une représentation du texte et rechercher dans une base de connaissances les textes pré-étiquetés les plus proches. A partir de cet ensemble de « proches voisins » (les textes étiquetés extraits de la base), on pourra inférer une liste de catégories à associer au nouveau texte considéré.

Représentation du texte

Pour cette approche, nous avons choisi d'utiliser une représentation du texte fondée sur le titre. Ce choix relève de facteurs à la fois théoriques et pratiques :

- Facteurs théoriques : Nous avons souligné précédemment l'importance du titre en indexation (cf. section 5.5). Afin de tenir compte de ce facteur et de diversifier notre approche du problème de l'indexation (vs. travail sur texte intégral présenté en 5.8.1) une représentation fondée sur le titre nous a semblé pertinente. Par ailleurs, les systèmes d'indexation MeSH de la littérature (MTI, Nomindex) utilisent les titres des textes indexés avec un certain succès.
- Facteur pratique : La base de connaissances avec laquelle nous travaillons ne nous permet pas d'accéder au texte intégral des ressources étiquetées. En revanche, la base contient les titres des textes.

Les textes sont donc représentés ici par un « sac de mots » issus du titre après filtrage des mots grammaticaux de notre « anti-dictionnaire ». Par exemple, le titre « Le diabète et les maladies rénales » sera représenté par le sac de mots (diabète, maladies, rénales), les mots grammaticaux « Le », « et » et « les » ayant été filtrés.

Recherche des « k-Plus Proches Voisins »

La recherche des « k-Plus Proches Voisins » potentiels se fait en deux temps :

1. Extraction des titres contenant au moins l'un des mots du sac de mots
2. Pour chaque titre extrait, calcul d'un score de similarité et classement afin de ne retenir que les k premiers.

Pour le calcul du score de similarité, deux méthodes ont été étudiées. La première est fondée sur le décompte des mots non-grammaticaux en commun. Par exemple, « Le diabète et les maladies rénales » et « Diabète et grossesse » ont un mot non-grammatical en commun : « diabète ». Le mot grammatical « et », présent dans les deux titres, n'intervient pas dans le décompte. Le score de similarité entre les deux titres sera donc de 1. La deuxième méthode est fondée sur la distance de Lewenshtein, ou distance d'édition (Levenshtein, 1966) entre les titres. Selon Lewenshtein, la distance entre deux chaînes de caractères peut être caractérisée par le nombre d'opérations d'édition à réaliser sur une chaîne pour obtenir la deuxième. Ces opérations d'édition, ou transformations élémentaires sont de trois types : la suppression (ex : diabète → diabte), l'insertion (ex : diabète → diabèrte) et la substitution (ex : diabète → diabyte). La distance d'édition entre deux chaînes correspond alors au nombre minimum de transformations élémentaires à réaliser pour passer d'une chaîne à l'autre. La distance d'édition est une mesure appliquée dans des domaines aussi variés que la correction orthographique (Dameron, 2003), la reconnaissance vocale (Sankoff & Kruskal, 1983) ou encore l'alignement de textes multilingues (Simard et al., 2005). En fonction de l'application, on peut envisager d'attribuer un coût différent à chaque type d'opération, comme le propose (Damereau, 1964), voire même à chaque opération (Parmentier, 1998), afin de refléter sa probabilité d'occurrence dans l'application ciblée. Pour notre part, nous n'avons pas différencié les opérations et nous avons considéré le taux de similarité entre deux titres comme étant l'inverse de la distance d'édition les séparant. Par exemple, la distance d'édition entre « Le diabète et les maladies rénales » et « Diabète et grossesse » est de 20. Le score de similarité correspondant est donc 0,050. L'avantage de cette méthode par rapport au simple décompte des mots grammaticaux est de ne pas segmenter les mots composés ou les collocations. Par ailleurs, les éventuelles fautes d'orthographe ou de frappe affectent peu le score. Ainsi, avec la distance d'édition, le score de similarité entre « Le diabèrte et les maladies rénales » et « Diabète et grossesse » sera de 0,048 alors qu'il sera de 0 avec la méthode de décompte des mots communs.

Obtention d'une indexation

Soit R la ressource à indexer et soient $V_1, ..., V_k$ ses k plus proches voisins. Soit P un mot clé (ou paire mot clé/qualificatif) candidat pour l'indexation de R ainsi que $Score(P)$ le score qui lui est attribué. La liste des mots clés (ou paires) candidats à l'indexation de R est déduite de l'indexation de ses k voisins selon l'algorithme de la figure 5.6 :

Il résulte de cette méthode que le nombre de mots clés (ou paires) attribués à R correspond au nombre de mots clés (ou paires) distincts attribués à l'ensemble des k voisins. De même, chaque mot clé (ou paire) se voit attribuer un score correspondant au nombre de mots clés auxquels ce même mot clé (ou paire) est également attribué, soit entre 1 et k.

5.8.3 Fusion des approches

Ajustement des scores Nous avons envisagé deux procédés pour combiner les indexations obtenues avec les approches TAL et k-PPV :

Le premier prend en compte le rang des mot clés (ou paires) résultant de l'indexation TAL et k-NN. Ainsi, nous attribuons à chaque mot clé (ou paire) un nouveau score égal à la somme de ses rangs dans les deux méthodes. Si un mot clé n'a été extrait que par une seule méthode, son rang pour l'autre méthode est considéré comme nul. Les mots clés (ou paires)

Pour $i = 1$ *à* k*, faire :*
    Tant que P, mot clé ou paire attribué à V_i *, faire :*
        Si P candidat,
            $Score(P) \leftarrow Score(P) + 1$
        Sinon,
            $Score(P) \leftarrow 1$
        Fin Si
    Fin Tant que
Fin Pour

FIG. 5.6 – Approche k-NN : Algorithme permettant d'obtenir les candidats à l'indexation et leur score

sont ensuite classés par score croissant.

Le second prend en compte le score relatif mot clés (ou paires) résultant de l'indexation TAL et k-NN. Le score relatif d'un mot clé (ou paire) correspond à son score initial divisé par la somme des scores attribués par la méthode correspondante. Comme pour la méthode du rang, nous attribuons à chaque mot clé (ou paire) un nouveau score égal à la somme de ses scores relatifs dans les deux méthodes. Si un mot clé n'a été extrait que par une seule méthode, son score relatif pour la deuxième méthode est considéré comme nul. Les mots clés (ou paires) sont ensuite classés par score décroissant.

Dans les deux cas, nous avons placé en tête du classement les mots clé (ou paires) extraits conjointement par les deux méthodes.

Dans la suite de nos travaux, nous envisageons l'étude d'autres méthodes pour combiner les résultats. Nous évaluerons notamment la pondération des apports de chaque méthode (par exemple, à l'aide de coefficients fondés sur l'expérience comme ceux utilisés dans MTI). Nous testerons également des méthodes statistiques (régression logistique).

Application des règles d'indexation Les règles d'indexation sont de deux types :

1. Des règles d'indexation issues de la NLM préconisant l'utilisation d'un mot clé MeSH de préférence à une paire mot clé / qualificatif pour représenter un même concept. Ainsi, la règle « *<MC1/Q1>* → *<MC2>* » indique qu'il convient de remplacer la paire *<MC1/Q1>* par le seul mot clé *<MC2>*. Par exemple :

   <coeur/transplantation> → *<transplantation cardiaque>*

2. Des règles d'indexation CISMeF[29] préconisant l'introduction d'un mot clé ou d'une paire complémentaire dans l'indexation d'une ressource. Ainsi, la règle « *<MC1/Q1*[30]*>* + *<MC2/Q2>* » indique qu'il convient d'ajouter la paire *<MC2/Q2>* à l'indexation d'une ressource déjà indexée avec la paire *<MC1/Q1>*. Par exemple,

[29] Certaines de ces règles ont été extraites de manière automatique, puis validées manuellement (L. Soualmia et al., 2005). Les autres règles ont été proposées par les documentalistes à partir de leur expertise dans le domaine médical.

[30] On considère ici que les qualificatifs peuvent être « vides », c'est-à-dire que les règles CISMeF peuvent statuer sur des mot clés seuls ou associés à un qualificatif.

<appendicectomie> + <appendicite/chirurgie>

En pratique, lorsqu'une règle CISMeF est appliquée, si le mot clé (ou la paire) à ajouter ne figure pas dans la liste des candidats, le score du mot clé ou de la paire ayant entrainé l'application de la règle (*<MC1/Q1>* d'après nos notations) lui est attribué. Si le mot clé (ou la paire) à ajouter figure déjà dans le liste des candidats, son score est augmenté du score attribué à *<MC1/Q1>*.

Sélection des descripteurs à attribuer à la ressource Soit N le nombre de mot clés (ou paires) candidats à l'indexation extraits à l'aide de l'une des méthodes ci-dessus. Soit S_i le score attribué au $i^{\text{ème}}$ candidat. On suppose que les candidats sont classés par ordre de scores décroissants, de sorte que $S_1 > \cdots > S_i > \cdots > S_N$. Pour $i = 1, \cdots, N-1$, on calcule :

$$F = \frac{S_i - S_{i+1}}{S_i + S_{i+1}}$$

Le seuil retenu sera i tel que F soit maximum. Habituellement utilisée en traitement du signal (Abdallah, 1998), cette fonction permet de détecter une rupture dans la continuité des scores et donc dans la pertinence des candidats proposés. Ce seuil adaptatif est une alternative à la sélection d'un nombre constant et arbitraire de mots clés (ou paires) pour chaque ressource. Adapter la taille de l'index à chaque ressource permet à la fois de refléter la pratique des documentalistes et de prendre en compte le fait que les performances d'un système automatique peuvent varier d'une ressource à l'autre.

5.8.4 Indexation d'une ressource

Afin d'illustrer le fonctionnement du système d'indexation implémenté dans MAIF, nous allons détailler et commenter l'indexation d'un exemple réel. Il s'agit d'un extrait de la ressource décrite par la notice CISMeF n°8916. Le texte intégral de cette ressource est inclus dans notre corpus « diabète ».

Approche TAL

La première étape de l'indexation est l'analyse de surface du texte désaccentué, effectuée avec nos dictionnaires et grammaires locales, décrits à la section 2.5.1. La figure 5.7 présente le résultat de cette analyse : les éléments du dictionnaire de mots simples sont soulignés (mot simple), les éléments du dictionnaire de mots composés sont surlignés (mot composé), les phrases sont numérotées et mises entre crochets ($\{\text{phrase}\}_N$). Aucun des patrons recensés par les grammaires n'a été identifié.

On peut tout d'abord observer un recoupement entre les éléments recensés par les deux dictionnaires (mot simples et composés). Par ailleurs, on peut également observer que les formes singulier et pluriel de « enfant » sont repérées. A première vue, l'ensemble des éléments exploitables pour l'indexation MeSH ont bien été identifiés, à part, en toute rigueur, « diabète de type 1 » qui aurait du être identifié dans l'expression « différencier le diabete de type 2 du type 1 » de la phrase 3. Ce type d'élision stylistique n'est pas pris en charge par notre analyse de surface[31]. De plus, certains des éléments retenus confirment ce désavantage lié à

[31]Le modèle vectoriel (Salton, 1989), par contre permet de pallier ce problème. Cependant, il occasionne également du bruit non désiré dans d'autres cas.

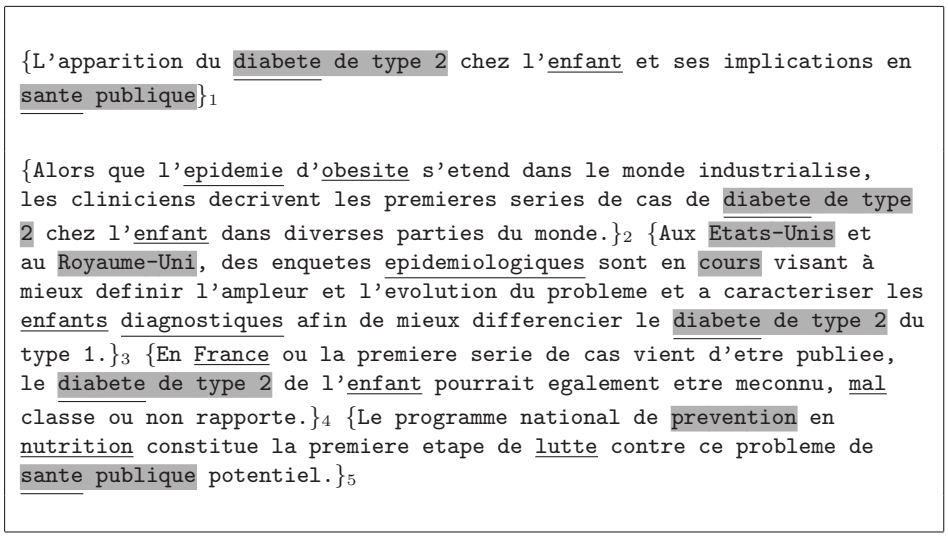

FIG. 5.7 – Résultat de l'analyse de surface effectuée sur le texte après désaccentuation (les mots simples sont soulignés, les mots composés sont surlignés, les phrases sont numérotées).

l'analyse de surface : il s'agit des termes polysémiques « cours », « mal »et « lutte ». Une analyse syntaxique du texte est à envisager pour résoudre ces ambiguités. Reconnu comme un adjectif, « mal » ne pourrait être traduit par <*douleur*>. Le tableau 5.9 présente le résultat de la traduction des éléments textuels en termes MeSH, ainsi que la fréquence de chaque descripteur dans le texte. Les qualificatifs non appariés sont représentés en italique dans la 2ème colonne ; ils seront l'objet d'un traitement particulier dans l'étape suivante.

Pour les séquences identifiées par les deux dictionnaires, seules les plus longues sont retenues. Ainsi, « diabete » est abandonné au profit de « diabete de type 2 »et « sante » est abandonné au profit de « sante publique ». On observe par contre que « cours » est traduit par le qualificatif <*/enseignement et éducation*>, « mal » par le mot clé <*douleur*> et « lutte » par le mot clé éponyme. Cette dernière erreur est en fait due à la tournure anaphorique utilisée par l'auteur dans la phrase 5, qui fait référence au « diabète de type 2 » par « ce probleme de sante publique potentiel ». L'analyse nécessaire pour résoudre cette ambiguité automatiquement est très complexe et sort du cadre de notre étude. Cependant, afin d'illustrer l'utilisation des grammaires locales par notre système, nous allons également considérer une formulation alternative de la phrase 5 dans laquelle l'anaphore a été résolue. L'analyse de cette phrase (notée 5-ALT) est illustrée par la figure 5.8. Les notations utilisées sont les mêmes que pour la figure 5.7. Le segment analysé par notre grammaire locale est représenté en italique.

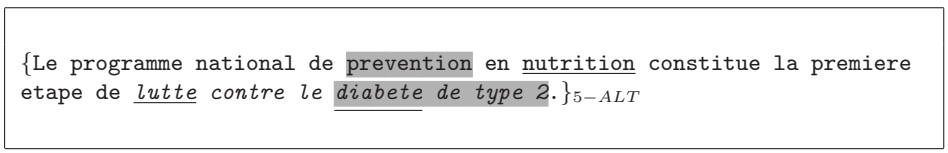

FIG. 5.8 – Formulation alternative de la phrase 5 et résultat de l'analyse de surface (les mots simples sont soulignés, les mots composés sont surlignés, les expressions sont en italique).

Eléments rencontrés	Descripteurs MeSH correspondants	Fréquence
diabete de type 2	<diabète de type 2>	4
	~~<diabète>~~	0
enfant	<enfant>	3
enfants	<enfant>	1
sante publique	<santé publique>	2
	~~<santé>~~	0
epidemie	<épidémie>	1
obesite	<obésité>	1
Etats-Unis	<Etats-Unis d'Amérique>	1
Royaume-Uni	<Grande Bretagne>	1
epidemiologiques	*</épidémiologie>*	1
cours	*</enseignement et éducation>*	1
diagnostiques	*</diagnostic>*	1
France	<France>	1
mal	<douleur>	1
prvention	*</prévention et contrôle>*	1
nutrition	<nutrition>	1
lutte	<lutte>	1

TAB. 5.9 – Traduction de l'analyse en terme MeSH. (Les qualificatifs non appariés sont représentés en italique dans la colonne centrale)

On observe cette fois un recouvrement entre les éléments identifiés par les dictionnaires et la grammaire locale. Le Tableau 5.10 présente les modifications de traduction et de décompte des occurrences occasionnées par le remplacement de la phrase 5 par la phrase 5-ALT dans notre texte. Par soucis de clarté, les éléments inchangés ne figurent pas dans le tableau.

Eléments rencontrés	Descripteurs MeSH correspondants	Fréquence
diabete de type 2	<diabète de type 2>	3
	<santé publique>	1
lutte contre le diabete de type 2	<diabète de type 2 /prévention et contrôle>	1
	~~<lutte>~~	0
	~~<diabete>~~	0
	~~<diabete de type 2>~~	0

TAB. 5.10 – Modifications occasionnées par la formulation alternative de la phrase 5

Comme précédemment, dans les cas de recouvrement, seules les séquences les plus longues sont retenues. Ainsi, les trois éléments « lutte » « diabete » et « diabete de type 2 » sont abandonnés au profit de « lutte contre le diabete de type 2 ». L'expression est alors traduite par la paire *<diabète de type 2/prévention et contrôle>*.

Phrase	Mots clés	Qualificatifs	Appariements
3	<Etats-Unis> <Royaume-Uni> <enfant> <diabète de type 2>	</épidémiologie> </enseignement et éducation> </diagnostic>	<diabète de type 2 /épidémiologie> <diabète de type 2 /enseignement et éducation> <diabète de type 2 /diagnostic>
5	<nutrition> <lutte> <santé publique>	</prévention et contrôle>	-
5-ALT	<nutrition>	</prévention et contrôle>	-

Tab. 5.11 – Appariement des qualificatifs isolés dans les phrases 3 et 5.

Mots clés les + fréquents	Qualificatifs	Appariements
<diabète de type 2> <enfant>	</prévention et contrôle>	<diabète de type 2 /prévention et contrôle>

Tab. 5.12 – Appariement du qualificatif isolé *prévention et contrôle*.

La phase suivante de l'indexation consiste à apparier les qualificatifs isolés, dans un premier temps à l'intérieur de la phrase où ils figurent. Trois qualificatifs isolés ont été identifiés dans la phrase 3 et un dans la phrase 5 (idem dans le cas de la phrase 5-ALT). La stratégie d'appariement consiste à rechercher un mot clé candidat à l'appariement à l'intérieur de la même phrase et de départager les paires potentielles à l'aide de l'historique CISMeF en cas de conflit. Le tableau 5.11 présente le résultat des appariements effectués.

Les qualificatifs de la phrase 3 ne peuvent être affiliés qu'au seul mot clé *<diabète de type 2>*. Trois affiliations sont donc générées. Pour les phrases 5 et 5-ALT, le qualificatif *</prévention et contrôle>* ne peut être affilié à aucun des mots clés candidats. Ainsi, à l'issue de la stratégie d'appariement phrastique, le qualificatif *</prévention et contrôle>* reste isolé. La seconde stratégie d'appariement est donc appliquée. Elle consiste à considérer les deux mots clés les plus fréquents comme candidats, les paires potentielles étant départagées à l'aide de l'historique CISMeF en cas de conflit. Le tableau 5.12 présente le résultat de cette stratégie d'appariement.

Le qualificatif *</prévention et contrôle>* ne peut être affilié au descripteur obligatoire *<enfant>*, car cette association n'est pas autorisée par le MeSH. Il est donc affilié au mot clé *<diabète de type 2>*. Le tableau 5.13 présente la liste des descripteurs retenus avec leur fréquence à l'issue de ces étapes, en considérant la phrase 5 (à gauche, colonne 2) ou la phrase 5-ALT (à droite, colonne 5).

La phase suivante de l'indexation consiste à exploiter les relations hiérarchiques entre les mots clés MeSH. Dans notre cas, tous les mots clés identifiés appartiennent à des arborescences

[32]Pour des raisons de place, nous représentons les qualificatifs sous forme abrégée : EP pour épidémiologie, DG pour diagnostic, etc.

Descripteurs[32]	F.	Score	Descripteurs	F.	Score
<diabète de type 2>	4	21,64	<diabète de type 2>	3	16,23
<enfant>	4	9,04	<enfant>	4	9,04
<santé publique>	2	8,93	<santé publique>	1	4,47
<épidémie>	1	6,88	<épidémie>	1	6,88
<obésité>	1	5,09	<obésité>	1	5,09
<Etats-Unis d'Amérique>	1	4,94	<Etats-Unis d'Amérique>	1	4,94
<Grande Bretagne>	1	5,61	<Grande Bretagne>	1	5,61
<diabète de type 2/EP>	1	7,32	<diabète de type 2/EP>	1	7,32
<diabète de type 2/ED	1	7,09	<diabète de type 2/ED	1	7,09
<diabète de type 2/DG>	1	7,12	<diabète de type 2/DG>	1	7,12
<France>	1	2,15	<France>	1	2,15
<douleur>	1	9,52	<douleur>	1	9,52
<diabète de type 2/PC>	1	7,22	<diabète de type 2/PC>	2	14,43
<nutrition>	1	4,79	<nutrition>	1	4,79
<lutte>	1	1			

TAB. 5.13 – Résumé final des descripteurs retenus : fréquences et scores (à droite avec la formulation alternative)

différentes. La hiérarchie n'a donc aucune influence sur l'indexation de ce texte. Ensuite, on calcule le score attribué à chacun des descripteurs retenus. Les colonnes 3 et 6 du tableau 5.13 présentent les scores obtenus par chaque descripteur. On peut constater que le classement varie selon que l'on considère les fréquences ou les scores. En effet, un terme de fréquence moindre peut obtenir un score plus élevé. Par exemple, <*santé publique*> est deux fois plus fréquent que <*douleur*>, pourtant, son score est de 8,93 contre 9,52 pour <*douleur*>.

La liste comporte un descripteur obligatoire de fréquence supérieure ou égale à 2, <*enfant*>. Le score maximal (21,64 ou 16,23 avec la formulation 5-ALT) lui est attribué.

Approche k-PPV

La première étape consiste à filtrer les mots grammaticaux, afin d'obtenir un sac de mots représentant la ressource :

{apparition, chez, diabète, enfant, implications, publique, santé, type, 2}

Les plus proches voisins sont ensuite recherchés dans la base CISMeF. Le tableau 5.14 présente les trois plus proches voisins extraits, à l'aide du décompte des mots non grammaticaux en commun. Les mots communs ayant conduit à la sélection sont représentés en gras.

On constate que la troisième ressource sélectionnée (n°10053) semble globalement moins pertinente - on peut notamment s'interroger sur l'opportunité d'inclure « chez » dans l'anti-dictionnaire (liste des mots à filtrer).

L'indexation de ces voisins conduit à retenir un total de 27 mots clés (ou paires MeSH) comme candidats pour l'indexation de la ressource n°9816. Pour la clarté de notre exemple, nous n'avons pris en compte que les 11 descripteurs majeurs. Le tableau 5.15 présente l'indexation de chaque voisin, et la liste de candidats qui en résulte, avec leurs scores.

N° CISMeF	Titre
16526	Réduction du risque de **diabète** de **type 2** **chez** les **enfants** autochtones du Canada
9811	**Diabète** de **type 2** ou **diabète** non insulino-dépendant (DNID)
10053	Allergies et hypersensibilités de **type** 1 **chez** l'**enfant** et **chez** l'adulte : aspects épidémiologiques, diagnostiques et principes du traitement

Tab. 5.14 – Trois plus proches voisins sélectionnés pour la ressource n° 9816 (les mots communs sont en gras)

N° CISMeF	Indexation MeSH	
16526	dépistage systématique dépistage systématique/normes dépistage systématique/utilisation diabète de type 2 diabète de type 2/diagnostic diabète de type 2/prévention et contrôle prévention primaire	
9811	diabète de type 2	
10053	allergènes hypersensibilité hypersensibilité/diagnostic hypersensibilité/étiologie hypersensibilité/physiopathologie hypersensibilité/thérapeutique	
9816	**Candidats**	**Score**
	diabète de type 2	2
	allergènes	1
	dépistage systématique	1
	dépistage systématique/normes	1
	dépistage systématique/utilisation	1
	diabète de type 2/diagnostic	1
	diabète de type 2/prévention et contrôle	1
	hypersensibilité	1
	hypersensibilité/diagnostic	1
	hypersensibilité/étiologie	1
	hypersensibilité/physiopathologie	1
	hypersensibilité/thérapeutique	1
	prévention primaire	1

Tab. 5.15 – Obtention d'une liste de candidats pour la ressource n° 9816

Fusion des résultats

Nous allons maintenant fusionner les listes de descripteurs MeSH candidats obtenus avec les approches TAL et k-PPV en utilisant la méthode pondérée, qui est fondée sur les scores

relatifs obtenus par les candidats avec chaque méthode. Pour la méthode TAL, la somme des scores pour les descripteurs obtenus avec la formulation originale de la phrase 5 est de 120,74. Ainsi, le score de <*diabète de type 2*> qui était de 21,64 correspond à 17,92% du total. Ainsi, pour la fusion des listes, le score de <*diabète de type 2*> est donc 17,92. De même, pour la méthode k-PPV la somme des scores étant 14, le score de <*diabète de type 2*> passe de 2 à 14,29. Par ailleurs, <*diabète de type 2*> étant présent dans les deux listes, les deux scores obtenus sont additionnés dans la liste fusionnée, et son score final est donc $17,92 + 14,29 = 32,21$. La liste fusionnée de candidats, ainsi que leurs scores sont présentés dans le tableau 5.16.

L'avant-dernière étape du traitement est l'application des règles d'indexation. Dans notre cas, aucun des descripteurs extraits n'entre en jeu dans une régle d'indexation. La liste des candidats reste donc inchangée[33].

Finalement, on sélectionne les candidats qui représentent le mieux le texte à l'aide de la fonction de rupture : le tableau 5.16 présente les valeurs sucessives de la fonction de rupture pour notre exemple (avec la formulation originale pour la phrase 5).

La valeur maximum de la fonction de rupture (0,42) est surlignée. Dans ce cas, le seuil se trouve au rang 2, et les descripteurs sélectionnés pour l'indexation sont <enfant> et <diabète de type 2>.

A titre informatif, les descripteurs proposés par un indexeur professionnel pour indexer le texte étudié sont présentés par la figure 5.9.

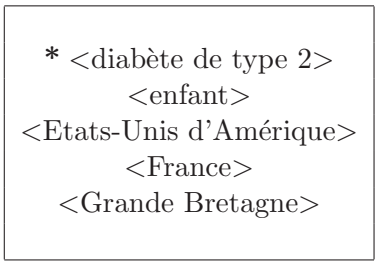

FIG. 5.9 – Liste des descripteurs MeSH attribués manuellement pour l'extrait de la ressource 8916

L'ensemble de ces descripteurs étaient extraits par notre méthode. Le descripteur majeur <*diabète de type 2*> figure dans la liste des descripteurs sélectionnés, ainsi que le descripteur obligatoire <*enfant*>.

5.9 Evaluation de Systèmes d'Indexation Automatique MeSH

Nous présentons ici les différentes évaluations de systèmes d'indexation automatique MeSH que nous avons réalisées dans le cadre de cette thèse. Elles portent tout d'abord sur les deux approches (TAL et k-PPV) mises en oeuvre dans le système MAIF, ainsi que sur leur combinaison. Ces évaluations portent sur l'extraction de *paires mot clé/qualificatif*.

Dans un deuxième temps, nous avons comparé les résultats obtenus par MAIF à ceux d'autres logiciels d'indexation MeSH francophones (HONMeSHMapper, MeSHMap et NO-

[33]Par exemple, si le mot clé <*appendicectomie*> avait fait partie des candidats extraits avec un score de 5,25 la paire <*appendicite/chirurgie*> aurait été rajoutée à la liste avec un score de 5,25.

Descripteurs	Score S	$\frac{S_i - S_{i+1}}{S_i + S_{i+1}}$
<enfant>	32,21	0
<diabète de type 2>	32,21	0,42
<diabète de type 2/PC>	13,12	0,01
<diabète de type 2/DG>	12,98	0,00
<douleur>	7,88	0,24
<santé publique>	7,40	0,03
<diabète de type 2/EP>	7,32	0,01
<allergènes>	7,14	0
<dépistage systématique>	7,14	0
<dépistage systématique/NO>	7,14	0
<dépistage systématique/UT>	7,14	0
<hypersensibilité>	7,14	0
<hypersensibilité/DG>	7,14	0
<hypersensibilité/ET>	7,14	0
<hypersensibilité/PP>	7,14	0
<hypersensibilité/TH>	7,14	0
<prévention primaire>	7,14	0,10
<diabète de type 2/ED>	5,87	0,01
<épidémie>	5,70	0,10
<Grande Bretagne>	4,65	0,05
<obésité>	4,22	0,03
<nutrition>	3,97	0,01
<Etats-Unis d'Amérique>	3,93	0,37
<France>	1,78	0,36
<lutte>	0,83	-

TAB. 5.16 – Application de la fonction de rupture et calcul du seuil (la valeur maximum de la fonction de rupture est surlignée)

MINDEX) et américains (MTI). Ces évaluations portant sur l'extraction de *termes MeSH isolés* ont été présentées récemment lors des conférences AIME (Artificial Intelligence in Medicine Europe) (Névéol, Mary, et al., 2005) et AMIA (American Medical Informatics Association) (Névéol, Mork, Aronson, & Darmoni, 2005).

5.9.1 Evaluation de MAIF

Evaluation de MAIF-TAL

Evaluations préliminaires. Une évaluation préliminaire de l'approche TAL de MAIF a été réalisée sur un sous-ensemble de 10 ressources du corpus « diabète »(Névéol, Rogozan, & Darmoni, 2004). Le dictionnaire utilisé ne comportait alors que des entrées liées aux descripteurs obligatoires, qualificatifs et mots clés du domaine du diabète (soit une couverture MeSH d'environ 1%). Une seconde évaluation réalisée sur les même ressources avec un dictionnaire couvrant cette fois 33% du MeSH a permis d'établir que l'augmentation significative de la taille du dictionnaire n'avait pas eu de répercution trop importante au niveau du bruit

(Névéol, 2004). Ces expériences avaient pour but de valider globalement l'approche et en particulier la méthodologie de construction des dictionnaires utilisés pour l'extraction des concepts médicaux.

Les résultats (précision proche de 50% au rang 4) ont montré que la stratégie d'indexation semblait pertinente. Nous avons également pu observer que les descripteurs obligatoires n'était pas systématiquement extraits par les indexeurs humains, alors qu'il l'étaient par le système. Par ailleurs, certains mots clés très fréquents appartenant au champ lexical du diabète (par exemple, <*insuline*> ou <*sang*> étaient souvent sélectionnés à tort par le système. De même, du fait de leur faible fréquence par rapport aux mots clés isolés dans le catalogue CISMeF, les paires mot clé/qualificatif semblent obtenir un score élevé après application de la normalisation $tf * idf$.

Evaluations plus approfondies : méthodes Les évaluations suivantes ont été réalisées sur l'ensemble du corpus « diabète », à l'aide des dictionnaires ayant une couverture de 33% du MeSH (Névéol, Rogozan, & Darmoni, 2005a). En plus de l'indexation du texte intégral de la ressource, nous avons également extrait automatiquement une indexation pour le sommaire de chaque ressource, disponible dans notre corpus comme nous l'avons indiqué à la section 1.2.6.

Comme précédemment, l'indexation automatique obtenue a été comparée à l'indexation manuelle CISMeF, prise comme référence. Les mots clés (ou paires) extraits par le système automatique étant classés par ordre des scores décroissants, nous avons calculé la précision, le rappel et la F-measure à chaque rang. Par exemple, au rang i, la précision est de $\frac{n}{i}$ et le rappel de $\frac{n}{N}$ avec n le nombre de paires correctement extraites (on a donc $n \leq i$), et N le nombre total de paires à extraire. Remarquons qu'avec cette méthode d'évaluation, le rappel maximum qu'il est possible d'obtenir est inférieur à 100% pour tous les rangs $i < N$.

A partir des observations faites lors des évaluations préliminaires, nous avons plus particulièrement étudié les points suivants :

1. Evaluation du silence de l'indexation manuelle sur les descripteurs obligatoires : pour ce faire, nous avons comparé l'indexation automatique proposée par le système à deux références. D'une part, l'indexation CISMeF telle qu'elle figurait dans le catalogue et d'autre part l'indexation CISMeF complétée par les descripteurs obligatoires de fréquence supérieure ou égale à deux extraits par le système automatique. Dans ce deuxième cas, si l'indexation manuelle préconisait l'extraction de 10 mots clés (ou paires) et que deux descripteurs obligatoires (non compris dans l'indexation manuelle) étaient sélectionnés par le système, nous avons considéré que douze mots clés (ou paires) étaient pertinents pour l'indexation de la ressource : les dix originellement sélectionnés par les indexeurs CISMeF, plus les deux descripteurs obligatoires extraits par MAIF.
2. Correction de la sélection eronnée de certains termes très fréquents : nous avons modifié le score des mots clés (ou paires) extraits à l'aide des coefficients de sélection calculés comme indiqué par la formule 5.4.
3. Evaluation de l'apport de la normalisation $tf*idf$: nous avons comparé les performances obtenues en utilisant la normalisation $tf * idf$ dans le calcul des scores, ou en utilisant simplement les fréquences des concepts.

Résultats Afin d'illustrer les résultats, le tableau 5.17 présente l'indexation manuelle (colonne de gauche) et automatique (colonne de droite - seuls les descripteurs au dessus du Seuil

sont présentés) obtenues pour une ressource du corpus.

Indexation Manuelle	**Indexation Automatique**
∗ diabète de type ii	grossesse
∗ diabète de type ii/chimiothérapie	hypoglycémiants/
	administration et posologie
grossesse	sujet âgé
∗ hypoglycémiants	diabète de type ii/thérapeutique
hypoglycémiants/	diabète de type ii/complications
administration et posologie	
hypoglycémiants/classification	hypoglycémiants/effets indésirables
hypoglycémiants/effets indésirables	hypoglycémiants
interactions médicamenteuses	diabète de type ii
suivi soins patient	insuline
sujet âgé	

Tab. 5.17 – Indexation Manuelle et Automatique pour la ressource n°115.

Les tableaux 5.18 à 5.21 présentent les performances obtenues sur l'ensemble du corpus, en variant les différents paramètres de l'évaluation (descripteurs obligatoires, coefficients de sélection, normalisation $tf * idf$).

Le tableau 5.22 offre une comparaison des résultats obtenus sur l'indexation à l'aide de mots clés isolés par opposition aux paires mot clé/qualificatifs, évaluées dans tous les autres tableaux.

La dernière ligne de chaque tableau (en gras) présente la précision et le rappel moyens obtenus au seuil adaptatif. Le seuil moyen est indiqué entre parenthèses (S=∗seuil_moyen∗).

Rang	$tf * idf$	**Fréquence**
	Précision - Rappel	Précision - Rappel
1	19,30 - 4,93	28,07 - 5,98
4	23,68 - 15,75	27,19 - 18,74
10	17,54 - 27,40	17,54 - 28,00
20	12,19 - 33,28	12,28 - 34,18
50	7,35 - 42,12	7,15 - 41,40
Seuil	**21,61 - 19,40**	**25,72- 19,93**
	(S=32)	**(S=28)**

Tab. 5.18 – Comparaison des scores fondés sur la normalisation $tf * idf$ et la fréquence. (Indexation à l'aide de paires mot clé/qualificatif)

D'après un test du signe, il n'y a pas de différence significative entre les scores fondés sur la fréquence ou la normalisation $tf * idf$ au delà du rang 10. Cependant, la fréquence est une meilleure méthode de classement si nous considérons seulement les premiers rangs. En effet, la différence de précision est statistiquement significative au rang 1 selon un test de Mc

Rang	$tf * idf + cs(t)$	**Fréquence** $+cs(t)$
	Précision - Rappel	Précision - Rappel
1	19,30 - 5,21	24,56 - 7,98
4	24,56 - 15,70	27,63 - 20,54
10	16,32 - 26,58	17,02 - 28,02
20	11,23 - 34,54	11,05 - 33,40
50	7,15 - 42,67	6,96 - 41,21
Seuil	**27,02- 19,98**	**25,12- 15,98**
	(S=5)	**(S=4)**

TAB. 5.19 – Comparaison des scores fondés sur la normalisation $tf * idf$ et la fréquence, en utilisant les coefficients de sélection. (Indexation à l'aide de paires mot clé/qualificatif)

Rang	$tf * idf$+ **DO**	**Fréquence + DO**
	Précision - Rappel	Précision - Rappel
1	49,12 - 8,51	59,65 - 10,04
4	48,68 - 24,11	52,20 - 26,51
10	28,42 - 34,32	28,42 - 34,54
20	17,63 - 39,33	17,72 - 39,95
50	9,81 - 48,46	9,58 - 47,62
Seuil	**43,68- 8,25**	**48,47 -28,49**
	(S=32)	**(S=28)**

TAB. 5.20 – Comparaison des scores fondés sur la normalisation $tf * idf$ et la fréquence, en considérant que les descripteurs obligatoires (DO) sont correctement extraits par le système. (Indexation à l'aide de paires mot clé/qualificatif)

Rang	$tf * idf$+ **DO** $+cs(t)$	**Fréquence + DO** $+cs(t)$
	Précision - Rappel	Précision - Rappel
1	49,12 - 8,95	63,16 - 12,63
4	49,56 - 24,26	53,07 - 28,46
10	27,19 - 33,12	27,89 - 34,51
20	16,67 - 40,28	9,35 - 47,17
50	9,69 - 48,94	9,58 - 47,62
Seuil	**48,65- 28,16**	**49,40 - 24,96**
	(S=5)	**(S=4)**

TAB. 5.21 – Comparaison des scores fondés sur la normalisation $tf * idf$ et la fréquence, en utilisant les coefficients de sélection et en considérant que les descripteurs obligatoires (DO) sont correctement extraits par le système. (Indexation à l'aide de paires mot clé/qualificatif)

Nemar (p=0,025), et au rang 4 selon un test du signe (p=0,008). Les différences observées pour le rappel ne sont pas statistiquement significatives, mais semblent également indiquer une légère supériorité de la fréquence.

A titre indicatif, les résultats obtenus sur l'indexation des « sommaires » sont présentés en

Rang	$tf * idf$	**Fréquence**
	Précision - Rappel	Précision - Rappel
1	42,86 - 11,14	28,57 - 7,21
4	21,43 - 15,96	24,11 - 17,87
10	15,54 - 25,34	15,18 - 23,93
20	10,18 - 29,71	10,36 - 30,89
50	6,24 - 35,33	6,12 - 34,76
Seuil	**24,16- 7,43 (S=47)**	**31,52- 3,19 (S=31)**

TAB. 5.22 – Comparaison des scores fondés sur la normalisation $tf * idf$ et la fréquence. (Indexation à l'aide de termes isolés)

annexe. La principale différence observée est que la fonction de seuil semble inefficace dans le cas des sommaires. Il semble également que la précision soit un peu meilleure dans le cas des sommaires et le rappel un peu moins élevé. Cependant, nous avons résolu de ne pas pousser plus loin la piste de l'indexation des sommaires, car leur extraction automatique sort du cadre de notre étude.

Performances du système. Les résultats de cette évaluation montrent qu'il y a effectivement un silence important de l'indexation manuelle sur les descripteurs obligatoires. Les tableaux 5.18 et 5.20 montrent que le rappel est plus élevé si on considère les descripteurs obligatoires extraits par le système comme étant pertinents pour l'indexation. De plus, la précision des premiers rangs est également plus élevée dans ce cas, car notre système attribue un score maximum aux descripteurs obligatoires - ce qui revient à les classer en tête de la liste de candidats à l'indexation. La sélection des descripteurs obligatoires par le système a été validée par le super-indexeur de l'équipe CISMeF et doit être répercutée sur l'indexation des ressources du corpus dans le catalogue.

L'utilisation des coefficients de sélection pour le calcul des scores n'a que peu d'impact sur la précision et le rappel aux rangs fixes. Cela peut venir du fait que, pour cette étude réalisée en 2004, nous avions calculé les coefficients de sélection à partir du corpus EQUER et disposions des coefficients pour seulement 483 mots clés. Suite à la mise en place de la recherche plein texte dans Doc'CISMeF, nous avons été en mesure de calculer ces coefficients pour plus de 8 000 mots clés comme nous l'indiquions à la section 5.8.1. Nous envisageons donc de nouveaux tests avec ces données plus complètes. Dans notre étude, l'impact des coefficients de sélection est donc limité au seuil moyen, qui est plus bas. Le seuil moyen obtenu (cf. tableaux 5.19 et 5.19) est alors proche du nombre de mots clés (ou paires) attribués manuellement à une ressource du catalogue (4.46 in 2002).

Bien que les performances du systèmes soient meilleures au seuil adaptatif qu'au rang fixe équivalent, la F-measure peut parfois s'avérer plus élevée à un autre rang fixe.

Dans cette étude, les meilleurs résultats pour l'indexation à l'aide de paires mot clé/qualificatif sont obtenus sur l'indexation plein texte utilisant les fréquences et les coefficients de sélection pour le calcul du score, en considérant que les descripteurs obligatoires extraits sont pertinents. Dans ce cas, la précision obtenue au rang 4 est de 53%, ce qui est équivalent aux performances d'autres systèmes de la littérature sur l'extraction de mots clés isolés (Gaudinat et al., 2002).

Ainsi, nous avons considéré que le développement du système était suffisamment avancé pour réaliser une évaluation comparative de tous les systèmes d'indexation MeSH francophone. Nous détaillons cette évaluation à la section 5.9.2.

Contrairement aux autres systèmes existants, notre approche consiste à extraire des paires mot clé/qualificatif. Ainsi, lors de l'évaluation, nous considérons qu'il est incorrect d'extraire un mot clé isolé là où l'indexation manuelle préconise l'utilisation d'une paire. Par exemple, si notre système extrait le mot clé <*diabète*> alors que la paire <*diabète/thérapeutique*> est attendue, nous considérons que le terme extrait n'est pas pertinent. Cette méthode d'évaluation a des conséquences statistiquement significatives sur les résultats (cf. tableaux 5.18 et 5.22). Par exemple, sur un test de Mc Nemar au rang 1 nous avons observé une p-value de 0,005). La précision globale sur l'ensemble des ressources pour l'extraction de paires mot clé/qualificatif était de 10,05% pour toutes les paires extraites, et de 18,31% pour les paires retenues au dessus du seuil. Afin de pondérer ces chiffres, il est intéressant de rappeler que dans les études de consistance de l'indexation (Funk et al., 1983; Leininger, 2000), les descripteurs obligatoires sont les termes sur lesquels la consistance est la plus faible. Les extraire automatiquement est donc un exercice d'autant plus difficile.

Dans l'exemple du tableau 5.17, deux des quatre paires attendues sont extraites et deux des trois mots clés majeurs sont également extraits par le système automatique.

Evaluation de MAIF k-PPV

Nous présentons ici une évaluation de l'indexation MeSH à l'aide de la méthode de recherche des k plus proches voisins présentée à la section 5.8.2. Nous avons étudié plus particulièrement le choix d'une valeur de k et d'une fonction de similarité permettant de choisir les k voisins à utiliser.

Corpus et méthode d'évaluation. Les mesures habituelles de précision, rappel et F-measure sont utilisées. L'indexation obtenue grâce à la méthode automatique est comparée à l'indexation manuelle CISMeF, considérée comme la référence. Pour ces évaluations, nous avons utilisé les corpus « diabète », « misc » et « ENFR ». Nous avons étudié l'influence de la valeur de k, de la mesure de similarité utilisée pour sélectionner les voisins et du corpus d'apprentissage.

Valeur de k. Le tableau 5.23 présente les résultats de l'expérience conduite le 29/09/04 afin de déterminer le nombre de voisins optimal.

- On constate que pour $k > 5$, il n'est pas possible de trouver k voisins pour certaines ressources. Pour éviter un silence complet de la méthode dans ce cas, nous avons choisi d'assouplir le système et d'utiliser l'ensemble des voisins disponibles, même s'il y en a moins de k dans la version opérationnelle de MAIF.
- On observe également que les performances sont meilleures pour $k = 10$

Mesure de Similarité. Le tableau 5.24 présente les résultats de l'expérience conduite le 11/07/05 afin de comparer les performances du calcul de similarité fondé sur le décompte des mots non-grammaticaux en commun dans les titres (colonne de gauche) et sur la distance d'édition (colonne de droite). On constate que les performances sont relativement similaires pour la précision, le rappel et la F-measure, avec un léger avantage en faveur du décompte des

Rg.	k=3	k=5	k=10	k=15
	P - R	P - R	P - R	P - R
1	34,96 - 4,72	51,22 - 6,83	57,02 - 8,65	42,62 - 5,49
3	23,83 - 8,93	33,81 - 12,94	37,15 - 14,53	30,83 - 11,69
5	19,67 - 12,21	27,80 - 17,72	30,08 - 18,77	23,93 - 14,47
7	17,03 - 14,85	22,89 - 19,78	26,10 - 22,96	20,25 - 16,52
10	14,56 - 16,54	18,43 - 22,79	20,83 - 25,29	16,31 - 18,83
50	1,82 - 12,18	6,40 - 39,22	7,55 - 43,07	6,12 - 33,54

TAB. 5.23 – Comparaison des valeurs de k pour la recherche des k-plus proches voisins. (k = 3, 5, 10, 15)

mots non-grammaticaux. Par ailleurs, cette dernière méthode présente également un avantage au niveau du temps de calcul. C'est donc celle que nous retiendrons dans la version opérationnelle de MAIF.

Rg.	Décompte des mots non-grammaticaux	Distance d'édition
	P - R - F	P - R - F
1	54,92 - 7,11 - 12,58	54,92 - 6,98 - 12,38
3	39,10 - 14,04 - 20,66	38,49 - 13,43 - 19,91
5	30,09 - 16,69 - 21,47	30,26 - 17,32 - 22,03
7	25,55 - 19,73 - 22,27	25,71 - 20,72 - 22,95
10	21,67 - 23,40 - 22,50	20,26 - 23,51 - 21,77
50	7,67 - 37,65 - 12,74	6,91 - 36,41 - 11,62
Seuil	**34,69 - 18,00 - 23,70 (S=5,04)**	**33,21 - 16,26 - 21,83 (S=4,41)**

TAB. 5.24 – Comparaison des mesures de similarité pour la recherche des k-plus proches voisins. (k = 10)

Influence du corpus d'apprentissage. Le tableau 5.25 présente les résultats des expériences conduites le 22/09/04 (colonne de gauche) et le 11/07/05 (colonne de droite) afin d'évaluer l'influence du corpus d'apprentissage pour la recherche des k-plus proches voisins. En effet, le corpus d'apprentissage constitué par le catalogue CISMeF est en évolution constante (environ 50 nouvelles ressources par semaine et parfois des suppressions dues aux liens brisés). En fonction du contenu du catalogue au moment de l'indexation, on peut supposer que les k voisins sélectionnés pour une ressource donnée peuvent être différents. Excepté au rang 1 (F-measure de 15,03 le 22/09/04 contre 12,58 le 11/07/05), on peut dire que les performances varient relativement peu.

Rg.	Le 22/09/04	Le 11/07/05
	P - R - F	P - R - F
1	57,02 - 8,65 - 15,03	54,92 - 7,11 - 12,58
3	37,15 - 14,53 - 20,89	39,10 - 14,04 - 20,66
5	30,08 - 18,77 - 23,12	30,09 - 16,69 - 21,47
7	26,10 - 22,96 - 24,43	25,55 - 19,73 - 22,27
10	20,83 - 25,29 - 22,84	21,67 - 23,40 - 22,50
50	7,55 - 43,07 - 12,85	7,67 - 37,65 - 12,74

TAB. 5.25 – Influence du corpus d'apprentissage pour la recherche des k-plus proches voisins. ($k = 10$)

Evaluation des approches fusionnées.

Corpus et méthode d'évaluation

Pour cette évaluation, nous avons utilisé le corpus « misc » et les ressources en français du corpus « ENFR »(Névéol, Rogozan, & Darmoni, 2005b). Les performances des systèmes ont été mesurées à l'aide de la précision et du rappel. Les paires de descripteurs MeSH extraites par MAIF ont été comparées l'indexation manuelle CISMeF.

Nous avons principalement comparé les deux méthodes de fusion présentée à la section 5.8.3 : en tenant compte soit du rang, soit du score relatif attribué aux candidats par chaque approche.

Résultats.

Le tableau 5.26 présente la précision et le rappel obtenus pour chaque méthode séparément (colonnes 1 et 2) et pour la combinaison des méthodes fondée sur le rang (colonne 3) et sur les scores (colonne 4). Conformément aux résultats de la section 5.9.1, nous avons choisi $k = 10$ pour l'approche des plus proches voisins et utilisé la mesure de similarité fondée sur le décompte des mots grammaticaux. Dans le cas où il n'a pas été possible de trouver 10 voisins (10 ressources), nous avons utilisé l'indexation TAL. Nous avons également appliqué la fonction de rupture décrite à la section 5.8.3. Ainsi, la dernière ligne du tableau 5.26 présente la précision et le rappel moyens au seuil de rupture. Le seuil moyen est indiqué entre parenthèses.

Rk	TAL	10-PPV	MAIF-Rang	MAIF-Score
	P - R	P - R	P - R	P - R
1	36 - 5	57 - 6	51 - 6	49 - 6
4	32 - 16	34 - 15	38 - 17	36 - 18
10	22 - 27	22 - 23	26 - 31	27 - 33
50	8 - 40	8 - 40	10 - 53	10 - 53
T	**27 - 21**	**36 - 18**	**34 - 24**	**32 - 25**
	(T=12)	**(T=5)**	**(T=9)**	**(T=9)**

TAB. 5.26 – Précision et rappel de MAIF aux rangs fixes 1, 4, 10, 50 et au seuil adaptatif

La Figure 5.10 présente la F-measure en fonction du rang pour chacune des méthodes.

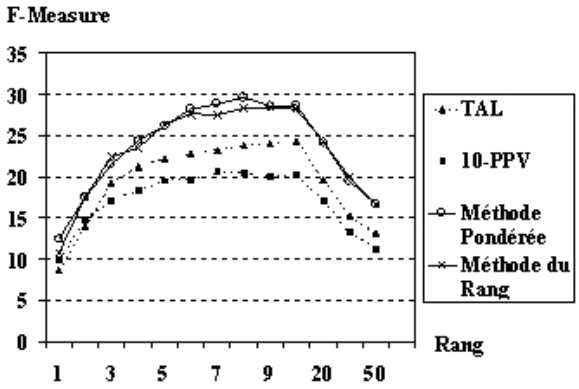

FIG. 5.10 – Courbes F-measure en fonction du rang pour MAIF

Discussion.

D'après le Tableau 5.26, on peut remarquer que la combinaison des deux approches d'indexation présentées offre une précision supérieure ou égale à chacune des approches, ainsi qu'un meilleur rappel. On observe également sur la Figure 5.10 que la F-measure de la combinaison des méthodes, est globalement plus élevée que pour chacune des méthodes séparément. Cela rejoint les conclusions de Aronson et al. (2004) qui observent une amélioration des performances de MTI lorsqu'un module statistique vient compléter le module de traitement linguistique.

Cependant, sur le seul corpus « ENFR », il s'avère que l'approche TAL offre de meilleures performances seule que combinée avec l'approche 10-PPV (cf. tableau 5.30 à la section 5.9.3). Ainsi, il pourrait être pertinent d'attribuer une pondération supérieure aux candidats proposés par l'approche TAL, qui doit en principe privilégier la précision. De fait, dans MTI, la pondération utilisée est de 7 pour l'approche linguistique contre seulement 2 pour l'approche statistique.

5.9.2 Evaluation des systèmes d'indexation MeSH francophones

Corpus et méthode d'évaluation

Pour cette évaluation, nous avons utilisé le corpus « misc » et les ressources en français du corpus « ENFR ». Les performances des systèmes ont été mesurées à l'aide de la précision, du rappel et de la F-mesure. Les descripteurs MeSH extraits par les systèmes ont été comparés aux descripteurs attribués à la ressource par les documentalistes de l'équipe CISMeF. Cette indexation manuelle, que nous avons prise comme référence, comportait des mots clés MeSH et des paires mot clé/ qualificatif. Cependant, trois systèmes sur les quatre évalués (NOMINDEX, MeSHMap et HONMeSHMapper) proposaient une indexation par mot clés isolés et non par paires. Ainsi, nous avons évalué dans cette étude l'extraction de mots clés isolés. Nous avons donc considéré qu'il était correct d'extraire un mot clé isolé dans les cas où l'indexation manuelle préconisait ce même mot clé associé à un qualificatif. Par exemple,

si le mot clé <*diabète*> était extrait alors que la paire <*diabète/chimiothérapie*> était attendue, nous avons considéré que le terme avait été correctement extrait. De même, si les paires <*diabète/chimiothérapie*> et <*diabète/prévention et contrôle*> étaient attendues, nous avons considéré que seul le mot clé <*diabète*> devait être extrait par les systèmes.

Nous avons également étudié de manière qualitative l'indexation proposée par chaque système pour une ressource donnée. Un indexeur expert (BT) a classé les 15 premiers termes extraits par les systèmes en « non pertinent » (NP), « trop général » (TG), « trop précis » (TP), ou « pertinent » (PE).

Résultats

Le tableau 5.27 présente la précision et le rappel obtenus pour chaque système. Nous avons également appliqué la fonction de rupture décrite à la section 5.8.3. Ainsi, la dernière ligne du tableau 5.27 présente la précision et le rappel moyens au seuil de rupture. Le seuil moyen est indiqué entre parenthèses.

Rk	**NOMINDEX**	**HON MeSHMapper**	**MAIF- TAL**	**MeSHMapp**
	P - R	P - R	P - R	P - R
1	13,25 - 2,37	45,78 - 8,63	45,78 - 7,42	13,41 - 1,77
4	12,65 - 9,20	31,93 - 26,41	30,72 - 22,05	15,24 - 10,57
10	12,53 - 22,55	20,61 - 36,96	21,23 - 37,26	11,83 - 18,20
50	6,20 - 51,44	7,76 - 57,81	7,04 - 48,50	5,56 - 39,39
T	**9,70 - 11 (T=6,6)**	**42,23 - 19,80 (T=4,6)**	**29,93 - 29,11 (T=12)**	**12,22 - 5,13 (T=3,09)**

Tab. 5.27 – Précision et rappel des systèmes francophones aux rangs fixes 1, 4, 7, 10 et au seuil adaptatif

La figure 5.11 présente une comparaison des quatre systèmes grâce à la F-measure. Celle-ci augmente de manière stable jusqu'au rang 10 pour NOMINDEX et MeSHMap. Pour HONMeSHMapper et MAIF-TAL, la F-measure augmente fortement jusqu'au rang 3, puis reste stable jusqu'au rang 10.

Le tableau 5.28 présente les quinze premiers mots clés extraits par chacun des systèmes pour une ressource du corpus.

A titre de comparaison, le tableau 5.29 présente les quinze premiers descripteurs (mots clés ou paires) extraits par MAIF pour la même ressource.

Discussion

Performances globales des systèmes. Le tableau 5.27 indique que les meilleurs systèmes évalués obtiennent une précision de 45% au rang 1. (HONMeSHMapper, MAIF-TAL). HONMeSHMapper et MAIF semblent équivalents en terme de précision, mais le rappel est un peu supérieur pour HONMeSHMapper. Ainsi, on observe sur la figure 5.11 que HONMeSHMapper offre globalement une meilleure F-measure.

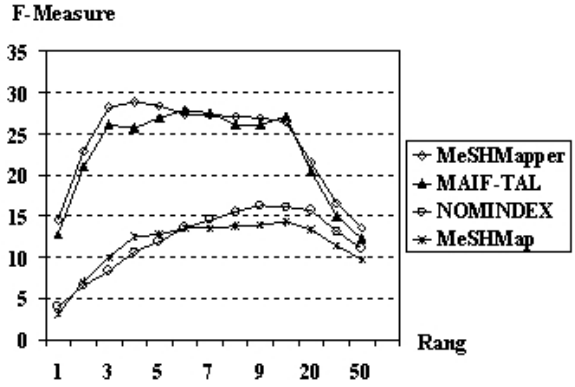

FIG. 5.11 – Courbe présentant la F-measure en fonction du rang pour les systèmes d'indexation MeSH francophones.

Pertinence du Seuil adaptatif. La fonction de seuil semble efficace pour maximiser la précision de HONMeSHMapper et MAIF. En effet, la précision aux seuils est plus élevée que la précision obtenue à un rang fixe équivalent (par exemple, elle est de 42% au seuil 4,6 contre 20% au rang 5 for HONMeSHMapper). Par contre, pour NOMINDEX et MeSHMap, la fonction de seuil est inefficace (pour MeSHMap, la précision et le rappel au seuil 3,09 sont inférieurs à ceux obtenus au rang fixe 3 - 12% et 7% contre 14% et 7%).

Pour MAIF, la F-measure au seuil est supérieure à la F-measure obtenue pour tous les rangs fixes. (29,51 au seuil, contre 27,89 - maximum observé au rang 6). Pour HONMeSHMapper, ce n'est pas le cas (la F-measure au seuil est de 26,9 contre 28,91 - maximum observé au rang 4). De plus, pour HONMeSHMapper, on constate que la F-measure est stable entre les rangs 3 et 10. Ainsi on peut dire que la fonction de seuil n'a pas sélectionné l'endroit idéal pour la coupure.

Analyse qualitative. Le tableau 5.28 présente les quinze premiers termes extraits par chaque système pour une ressource de notre corpus d'évaluation. Pour cette ressource, les termes extraits par NOMINDEX on été jugés « trop généraux » ou « non pertinents », à part <*signes et symptômes digestifs*> qui était « pertinent ». Il faut remarquer que parmi les termes erronés, NOMINDEX extrait plusieurs mots clés appartenant aux mêmes arborescences, comme par exemple, <*signes et symptômes, états pathologiques*>, <*signes et symptômes digestifs*>, <*signes et symptômes*> et <*diarrhée*>. Les termes extraits par HONMeSHMapper étaient également répartis entre les catégories « trop général », « trop précis » ou « non pertinent », à l'exception de <*médecine tropicale*> qui était « pertinent ». Les termes extraits par MeSHMap étaient principalement « trop précis », à part <*médecine tropicale*> et <*diagnostic*>qui étaient « pertinents ». Les mots clés extraits par MAIF étaient soit « trop généraux », « trop précis ». En ce qui concerne les paires (tableau 5.29), la majorité se sont averrées être « non pertinentes », à l'exception de <*diarrhée/diagnostic*> et <*médecine tropicale*> qui étaient « pertinentes ». Cependant, il est intéressant de remarquer que certains des descripteurs « non pertinents » comme par exemple <*diarrhée du nourrisson/diagnostic*> ou <*diarrhée du nourrisson*> sont très proches des termes pertinents <*diarrhée/diagnostic*> et <*diarrhée*>. Dans le cadre d'une indexation semi-automatique, ce type d'erreur serait facile

NOMINDEX	HONMeSHMapper
signes et symptômes, états pathologiques (TG)	**syndrome du côlon irritable** (TP)
estomac et intestin, maladies (TG)	**diarrhée**
diarrhée	**maladies inflammatoires intestinales** (TP)
signes et symptômes digestifs(PE)	**maladie aigüe** (NP)
processus anatomopathologique (NP)	**intestin, maladies** (TG)
maladie de l'appareil digestif (TG)	intestin grêle (TP)
intestin, maladies (TG)	médecine tropicale (PE)
signes et symptômes (TG)	infection (TG)
maladie (TG)	récepteurs activés par la protéinase (NP)
infections bactériennes et mycoses (NP)	infections bactériennes (TG)
lactose (TP)	voyage
infection (TG)	sprue tropicale (TP)
voyage	cyclosporose (TP)
syndrome du côlon irritable (TP)	côlon, maladies (TG)
syndromes de malabsorption (TP)	intolérance au lactose (TP)

MAIF- TAL	MeSHMapp
diarrhée	**schistosomiase intestinale** (TP)
voyage	**examen physique** (TP)
syndrome (TG)	**sprue tropicale** (TP)
signes et symptômes (TG)	**maladies transmissibles** (TG)
colon (TG)	**essai clinique** (TP)
lactose (TP)	**maladie** (TG)
santé (TG)	**perte poids** (TP)
entérite (TP)	cyclospora (TP)
fibre alimentaire (TP)	cryptosporidium TP
colonoscopie (TP)	infections (TG)
suivi soin patient (NP)	intestin (TG)
intestin grêle (TP)	schistosomiase (TP)
Canada (NP)	diagnostic (PE)
amibiase (TP)	campylobacter (TP)
perte poids (TP)	medecine tropicale (PE)

TAB. 5.28 – Exemple d'indexation automatique par *mots clés* proposée par chaque système (ressource CISMeF n° 4485, `http://www.phac-aspc.gc.ca/publicat/ccdr-rmtc/98vol24/24sup/dcc1.html` - accédé le 01/02/05). Les mots clés au dessus du seuil sont en gras, les mot clés également sélectionnés par les documentalistes CISMeF sont soulignés.

à corriger pour un indexeur humain.

Après cette étude des indexations automatiques de la ressource n°4485, les trois termes jugés pertinents (<*diarrhée/diagnostic*>, <*médecine tropicale*> et <*signes et symptômes digestifs*>) ont été ajoutés à l'indexation figurant sur la notice CISMeF de la ressource. Cette situation est intéressante à plusieurs titres :

MAIF
voyage
médecine tropicale (PE)
diarrhée/étiologie
diarrhée/diagnostic (PE)
diarrhée
grossesse (NP)
nourrisson (NP)
syndrome (TG)
colon (TP)
diarrhée du nourrisson/thérapeutique (NP)
purification eau (TP)
coopération internationale (NP)
France (NP)
diarrhée du nourrisson (NP)
pédiatrie/enseignement et éducation (NP)

TAB. 5.29 – Exemple d'indexation automatique par *paires* proposée par MAIF (ressource CISMeF n° 4485, `http://www.phac-aspc.gc.ca/publicat/ccdr-rmtc/98vol24/24sup/dcc1.html` - accédé le 01/02/05). Les mots clés au dessus du seuil sont en gras, les mot clés également sélectionnés par les documentalistes CISMeF sont soulignés.

- d'une part, elle met en évidence la variabilité de l'indexation : bien qu'effectuée manuellement, l'indexation de cette ressource était perfectible.
- d'autre part, elle montre l'intérêt des systèmes automatiques pour la réduction du silence de l'indexation manuelle : grâce aux termes extraits par les systèmes automatiques, l'indexation a pu être complétée et donc améliorée.

Par ailleurs, cette étude montre que le « bruit » généré par les sytèmes automatiques ne provient pas de l'extraction de termes non pertinents. La plupart des descripteurs extraits automatiquement qui ne sont pas également sélctionnés par les indexeurs humains n'ont pas été retenus car ils sont soit trop généraux (dans ce cas, le contenu de la ressource est décrit de manière trop vague pour être utile à l'utilisateur) soit trop précis (dans ce cas, le concept dénoté n'est pas suffisamment développé dans la ressource et l'utilisateur formulant une requête sur ce concept aurait besoin de plus d'information). Ainsi, pour les quatres logiciels d'indexation MeSH étudiés, il serait souhaitable d'intégrer un système de décision capable d'évaluer l'adequation du degré de spécificité des termes extraits pour une ressource.

Perspectives. les ressources terminologiques exploitées par les quatre systèmes sont différentes (CISMeF, ADM, UMLS et WRAPIN), et peuvent être complémentaires. Elles pourraient être intégrées au lexique médical français en cours de développement dans le cadre du projet UMLF (Zweigenbaum et al., 2003) et partagées par tous les systèmes, afin d'améliorer leurs performances. Une évaluation de Nomindex (Mary et al., 2002) a montré que les performances du système pouvaient être améliorées par un enrichissement du lexique. De fait, en utilisant un dictionnaire enrichi pour MAIF sur l'indexation du corpus d'évaluation, on observe une différence significative. La figure 5.12 présente la F-measure obtenue pour MAIF

(approche TAL) avec un dictionnaire couvrant 60% du MeSH (les résultats précédents avaient été obtenus en utilisant un dictionnaire couvrant 33% du MeSH).

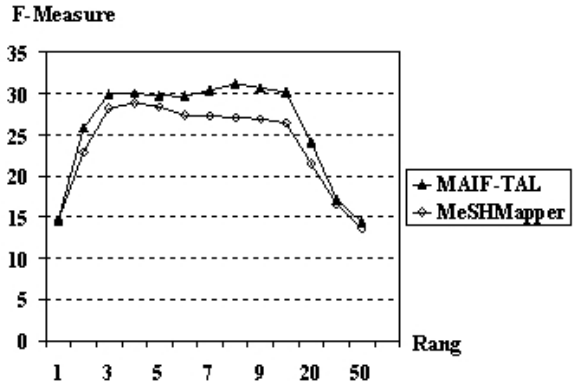

FIG. 5.12 – Courbe présentant la F-measure en fonction du rang pour MAIF-TAL et MeSH-Mapper.

On constate qu'avec ces nouvelles ressoures, MAIF offre de meilleures performances que HONMeSHMapper.

5.9.3 Evaluation translangue à l'aide d'un corpus parallèle EN/FR

Corpus et méthode d'évaluation

Pour cette évaluation, nous avons utilisé le corpus parallèle « ENFR ». La version anglaise des ressources a été soumise à MTI pour en extraire des mot clés MeSH et la version française a été soumise à MAIF afin d'en extraire également des mots clés MeSH. Les performances des systèmes ont été comparées à l'indexation manuelle CISMeF selon la même méthode que pour les extracteurs francophones (cf. section 5.9.2).

Résultats

Le tableau 5.30 présente la précision, le rappel et la F-measure obtenus pour chaque système. Nous avons également appliqué la fonction de rupture décrite à la section 5.8.3. Ainsi, la dernière ligne du tableau 5.30 présente les performances moyennes au seuil de rupture. Le seuil moyen est indiqué entre parenthèses.

La figure 5.13 illustre les performances des deux systèmes dans la configuration où la précision est la meilleure.

Pour MAIF, la F-measure augmente jusqu'au rang 3 et reste stable jusqu'au rang 10, alors que pour MTI, elle augmente jusqu'au rang 2 et décroit lentement jusqu'au rang 10.

Discussion

Ces résultats sont globalement cohérents avec les performances des deux systèmes lors d'autres évaluations (par exemple, (Aronson et al., 2004) pour MTI).

Rk	MTI strict	MTI medium	MTI default
	P - R - F	P - R - F	P - R - F
1	78,43 - 21,24 - 33,42	74,51 - 20,31 - 31,92	74,51 - 20,31 - 31,92
3	54,90 - 40,24 - 46,44	49,65 - 36,00 - 41,74	50,29 - 36,20 - 42,10
5	40,78 - 45,67 - 43,09	39,61 - 45,41 - 42,31	39,22 - 45,02 - 41,92
7	33,08 - 50,78 - 40,06	31,37 - 49,69 - 38,46	31,10 - 48,84 - 38,00
10	26,74 - 55,60 - 36,12	25,49 - 54,59 - 34,75	25,10 - 53,22 - 34,11
T	**38,84 - 53,28 - 44,90 (T=11,18)**	**36,88 - 59,40- 45,50 (T=15,20)**	**33,64 - 61,82 - 43,57 (T=20,90)**
Rk	**MAIF-TAL**	**MAIF-10 PPV**	**MAIF**
	P - R - F	P - R - F	P - R - F
1	54,90 - 9,71 - 16,50	21,95 - 3,49 - 6,02	37,25 - 6,14 - 10,54
3	40,47 - 28,67 - 33,56	14,56 - 8,05 - 10,37	34,61 - 21,71 - 26,68
5	30,59 - 32,63 - 31,57	13,17 - 12,02 - 12,57	26,27 - 26,75 - 26,51
7	26,08 - 39,71 - 31,48	13,95 - 16,61 - 15,16	21,82 - 32,14 - 25,99
10	22,60 - 49,42 - 31,02	12,44 - 20,29 - 15,42	18,24 - 39,29 - 24,91
T	**38,56 - 35,00 - 36,69 (T=5,24)**	**12,35 - 23,18 - 16,11 (T=11,55)**	**27,20 - 36,06 - 31,01 (T=7,46)**

Tab. 5.30 – Performance des systèmes MAIF et MTI à rang fixes et seuil adaptatif

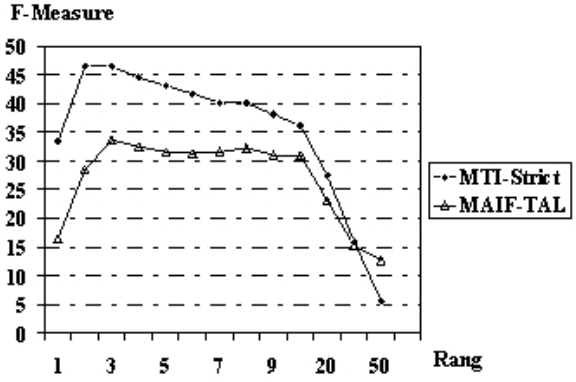

Fig. 5.13 – Courbe présentant la F-measure en fonction du rang pour MTI strict et MAIF-TAL.

Performances globales des systèmes MTI offre de meilleures performances aussi bien à rang fixe qu'au seuil adaptatif. Comme le montre le tableau 5.30 aussi bien la précision que le rappel obtenus par MTI sont systématiquement supérieurs à ceux obtenus par MAIF. La figure 5.13 illustre bien cette constatation : la courbe de F-measure de MTI est au dessus de celle de MAIF (sauf au rang 50). Les différences de performance observées peuvent résulter de plusieurs facteurs :

– MAIF est un logiciel en cours de développement, alors que MTI est un outil qui a été implémenté au cours du projet « Indexing Initiative » de la NLM, et amélioré au cours

des dix dernières années.

– Les ressources linguistiques utilisées sont différentes et en tout état de cause, les ressources bio-médicales disponibles sont beaucoup plus riches en anglais qu'en français.

  1. Couverture du MeSH par MAIF : au moment de l'évaluation, les dictionnaires MeSH utilisés par MAIF couvraient seulement 60% du MeSH et comme nous l'avons montré précédemment, l'exhaustivité des dictionnaires utilisés est un facteur important pour les performances du système. Par ailleurs, bien que cela n'ait pas eu d'impact sur cette étude, il faut également remarquer que l'approche k-PPV a également une couverture du MeSH limitée à celle de la base d'apprentissage sur laquelle elle est fondée, c'est-à-dire le catalogue CISMeF qui a une couverture MeSH de 50%. En revanche, MTI a une couverture MeSH de 100%.
  2. Des ressources linguistiques différentes pour le français et l'anglais : MTI utilise également les ressources du métathésaurus de l'UMLS, qui rassemble plus de 70 terminologies et contient également des liens entre les diférentes formulations d'un même concept bio-médical. Malgré les travaux en cours pour augmenter la part du français dans l'UMLS (Darmoni, Jarousse, et al., 2003), un nombre significatif de termes (y compris 50 000 synonymes MeSH) doivent encore être traduits en français. A terme, nous souhaitons que MAIF s'appuie sur les quatre terminologies les plus utilisées en santé (MeSH, CIM-10, SNOMED et CCAM).
  3. combinaison des méthodes statistiques et linguistiques : La méthode de filtrage de la liste de candidats appliquée par MTI s'avère très efficace : par exemple, au rang 3, la précision obtenue sans filtrage (default) est de 50% alors que celle obtenue avec le filtrage « strict » est de 54%. Les règles d'indexation appliquées par MAIF ont une influence moins significative sur les résultats.

– La combinaison des approches TAL et statistiques n'est pas effectuée avec la même approche.

Cette dernière observation confirme que l'étape de post-traitement des candidats extraits par les systèmes automatiques est très importante. En effet, l'évaluation des systèmes francophones présentée à la section 5.9.2 avait révélé que les candidats proposés par les systèmes automatiques étaient souvent trop généraux ou trop précis.

Parmi les mots clés extraits par MAIF mais non sélectionnés par les indexeurs humains, 52% sont des descripteurs obligatoires (au rang 1). Cette observation reflète le silence de l'indexation manuelle sur les descripteurs obligatoires que nous avions mis en évidence précédemment.

Sur ce corpus, l'approche TAL de MAIF offre de meilleures performances seule que combinée avec l'approche k-PPV. Ceci met en évidence les limites de l'approche k-PPV dont les performances dépendent de la base d'apprentisage utilisée. Dans le cas de MTI, la base utilisée comporte 15 millions de ressources indexées, contre 14.000 pour MAIF. Pour dix ressources du corpus « ENFR » il n'a pas été possible de trouver 10 voisins. Pour d'autres ressources, il s'avère que les « plus proches » voisins traitent en fait de thèmes très différents de ceux abordés dans la ressource à indexer.

Valeur ajoutée de la fonction de Seuil. La fonction de seuil permet de maximiser la précision des deux systèmes. Ainsi, la précision au seuil est plus élevée que la précision observée au rang fixe équivalent. (par exemple elle est de 39% au seuil 11 contre 27% au rang 11 pour MTI « Strict »). Pour MAIF, MTI « Default » et « Medium », la F-measure au seuil

est même supérieure à celle obtenue pour tous les rangs fixes (Par exemple, elle est de 31 au rang 7 contre 27 - qui est le maximum obtenu au rang 4 pour MAIF). Pour MTI « Strict » cependant, la F-measure au seuil est élevée, mais elle atteint son maximum au rang 2 (44,9 au seuil contre 46,4 au rang 2).

Conclusion et Perspectives. Cette étude montre que les sytèmes sont complémentaires.

- MTI offre d'excellentes performances pour l'extraction de mots clés MeSH isolés et met en oeuvre une méthode efficace de filtrage des candidats extraits qui permet d'améliorer encore les résultats.
- MAIF peut extraire des paires mot clé / qualificatif presque aussi efficacement que des mots clés MeSH isolés et utilise une fonction de rupture permettant de sélectionner un nombre optimal de descripteurs pour chaque ressource.

Les ressources linguistiques utilisées par l'approche TAL de MAIF doivent être complétées afin d'atteindre une couverture de 100% du MeSH. De plus, MAIF pourrait mettre en place une méthode de post traitement similaire à celle de MTI pour améliorer les performances.

De même, la méthode d'extraction de paires implémentée dans MAIF pourrait être adaptée à l'anglais pour une implantation dans MTI. La fonction de rupture pourrait également être utilisée par MTI pour sélectionner les descripteurs dans le cadre d'une utilisation entièrement automatique.

Nous avons presenté une évaluation comparative de deux systèmes d'indexation MeSH pour le français et l'anglais à l'aide d'un corpus parallèle. Une indexation par mots clés MeSH isolés a été proposée par MAIF pour la version française des ressources et par MTI pour la version anglaise. Les indexations automatiques ont été comparées à l'indexation manuelle CISMeF, prise comme référence. La meilleure précision (78%) est obtenue par MTI au rang 1, avec un filtrage « strict ». Ce résultat peut s'expliquer par une couverture très complète du MeSH et par l'utilisation d'une méthode de filtrage efficace des candidats. Pour la poursuite de ce travail, nous envisageons d'exploiter la complémentarité des systèmes pour les améliorer. Ainsi, nous pourrions utiliser les données statistiques de MTI, ainsi que la méthode de filtrage pour améliorer les performances de MAIF. De même, il semble intéressant d'adapter les aractéristiques de MAIF (extraction de paires mot clé/qualificatif, fonction de seuil) pour MTI.

5.10 Applications du Système d'Indexation Automatique MAIF

5.10.1 Indexation bi-modale Texte-Image

L'indexation bi-modale Texte-Image est une problématique récente de l'équipe CISMeF (Florea et al., 2005).

L'introduction des types de ressource liés à l'imagerie médicale et l'indexation des ressources à l'aide de *triplets* introduite dans CISMeF fin 2003 (cf. section 2.3.2) nous ont conduit à étendre notre réflexion sur l'extraction de paires mot clé/qualificatif à l'extraction de triplets <*mot clé / qualificatif \ type de ressource*>. Rappelons par exemple que le triplet <*pied\radiographie*> doit être utilisé pour désigner une image médicale de type radiographie représentant un pied, alors que la paire <*pied/radiographie*> doit être utilisée pour désigner la technique d'imagerie médicale de type radiographie appliquée à un pied. Cependant, à l'intérieur d'une ressource, l'expression textuelle « radiographie du pied » peut, selon le contexte, correspondre à l'un ou l'autre de ces concepts. Afin de les différencier, il convient

par exemple d'identifier les légendes associées aux images médicales. Tout comme l'extraction des titres et sous titres d'une ressource évoquée en 5.5, l'extraction des légendes d'image constitue un problème à part entière que nous n'avons pas abordé.

L'intérêt d'une combinaison de l'analyse textuelle des légendes d'image et de l'analyse de texture des images réside dans la complémentarité des résultats, d'autant que seuls 6 types d'images sont pour l'instant identifiés à l'aide de la texture (Florea et al., 2005).

Nous avons choisi d'aborder l'extraction de triplet à l'aide la méthode utilisée pour l'extraction des paires : l'analyse de corpus en compagnie d'un indexeur expert. Afin d'illustrer ce travail, nous présentons ci-dessous l'analyse d'une expression liée à des types de ressource d'imagerie : « ArthroScanner de l'épaule ». Un *arthroscanner* est une technique diagnostique résultant de l'analyse d'une *arthrographie* (c'est-à-dire, la radiographie d'une articulation) et d'un scanner (une < *tomodensitométrie*> au sens de la terminologie CISMeF). Ainsi, l'expression « arthroScanner de l'épaule » dénote tout d'abord les concepts de <*épaule**tomodensitométrie*> et <*épaule**arthrographie*>. Par ailleurs, pour être complète, l'indexation doit mentionner les pathologies qui peuvent être identifiées grâce à un arthroScanner de l'épaule, soit la fracture de l'épaule etc., ce qui entraine que les concepts de <épaule_ fracture\tomodensitométrie> et <épaule_ fracture\arthrographie> etc. sont également liés à cette expression. On peut avoir un raisonnement similaire avec d'autres expressions, telles que « ArthroIRM de l'épaule ». Il apparait donc que les expressions liées aux types de ressource image peuvent être très productives pour l'indexation. L'exemple que nous venons d'analyser peut être formalisé par la règle suivante :

> *«* TR image *de l'*organe O *»* → <*organe O**TR image*> *ET* <*pathologie liée à O**TR image*>

En toute rigueur, seuls les mots clés liés aux pathologies effectivement abordées dans la ressource ont leur place dans l'indexation. En conséquence, des règles de post processing doivent permettre de réviser la liste finale de candidats en ce sens.

L'indexation automatique à l'aide de triplets CISMeF est une tâche fort complexe, mais qui peut être traitée grâce aux outils que nous avons mis en place dans MAIF pour l'indexation à l'aide paires MeSH. Au delà de ce constat, à ce jour, nous n'avons fait qu'ébaucher ce chantier, qui constitue une vaste perspective pour nos travaux.

5.10.2 Intégration de terminologies (CCAM)

Chaque terminologie médicale est destinée à un usage précis et il peut s'avérer difficile voire coûteux de l'adapter à une utilisation non anticipée lors de sa création (Bachimont, 2000). Il est alors intéressant d'établir des correspondances entre les différentes terminologies afin de diversifier les usages possibles. Ainsi, l'extracteur MAIF a été utilisé pour établir des correspondances entre les termes de la CCAM et les termes MeSH. La correspondance MeSH proposée par MAIF pour chaque terme CCAM a ensuite été validée de manière corrective par un expert MeSH. Cette expérience a été conduite sur un échantillon de 195 termes CCAM. Selon l'expert, 287 termes MeSH étaient attendus au total. Sur 234 termes proposés par MAIF, 221 étaient corrects. Ainsi, pour cette tache de correspondance, MAIF offre une précision de 94% et un rappel de 77%. Par ailleurs, (une vingtaine) de synonymes ont été ajoutés dans la terminologie CISMeF et dans nos dictionnaires MeSH à partir de l'analyse du silence de la correspondance automatique. Le tableau 5.31 présente un échantillon des résultats obtenus.

Termes CCAM	Equivalents MeSH
Équilibre	*Équilibre locomoteur*
Oreille, sans précision	*Oreille*
Glandes salivaires	*Glande salivaire*
Système respiratoire, sans précision	~~Respiration~~ appareil respiratoire

TAB. 5.31 – Correspondance CCAM/MeSH ; Les termes MeSH correctement proposés par MAIF sont en italique, les termes erronés sont barrés et les rectifications de l'expert sont surlignées.

5.10.3 Codage des dossiers patients (Indexation CIM-10)

Dans le cadre d'un Mastère (Pereira, 2005) que j'ai partiellement encadré avec S. Darmoni, une étude de faisabilité concernant le codage médico-économique des dossiers patient à l'aide de la Classification Internationale des Maladies (version 10) a été réalisée au sein de l'équipe CISMeF. Dans cette étude, une indexation automatique CIM-10 est obtenue à partir de l'indexation automatique MeSH réalisée avec la méthode TAL (hiérarchie MeSH et associations mot clé / qualificatifs non prises en compte) suivie d'un transcodage MeSH ↔ CIM-10 effectué grâce aux données UMLS. Ce protocole et son évaluation sont illustrés par la figure 5.14.

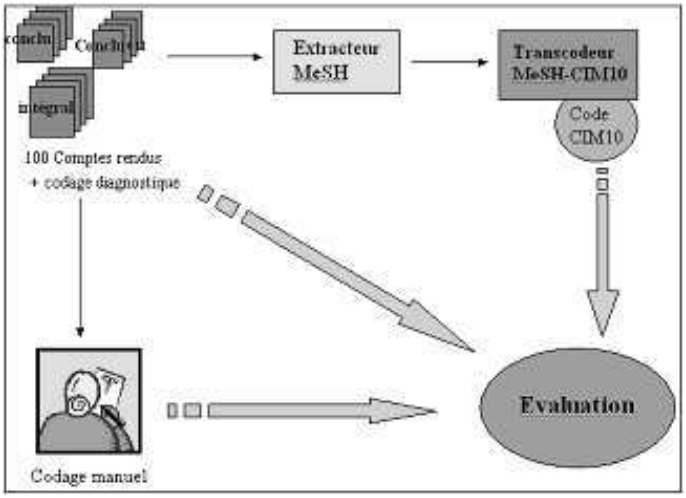

FIG. 5.14 – Processus d'Indexation Automatique CIM-10 d'après (Pereira, 2005)

Cette méthode est considérée comme une méthode minimale (méthode « baseline » au sens de (Manning & Shütze, 2000)) pour le codage CIM-10 pour deux raisons : Tout d'abord, le thésaurus MeSH étant une terminologie destinée à l'indexation de connaissances en médecine, et non à la description des données contenues dans le dossier patient (telles que : symptômes, diagnostic, description des soins administrés, ordonnance...), il semble peu probable qu'il soit possible d'obtenir un codage adapté à un usage différent. Par ailleurs, nous anticipons également une perte d'information lors du transcodage (traduction des mots clés MeSH

en termes CIM10), dans la mesure où tous les termes MeSH n'ont pas d'équivalent CIM-10 (environ 10% dans cette première étude). Cependant, les résultats obtenus à l'aide de cette méthode nous permettront de juger de l'adéquation du développement d'un extracteur spécifique CIM-10 sur le modèle de l'extracteur MeSH. L'évaluation de la méthode est réalisée sur un échantillon représentatif de dossiers patients issus des services de cancérologie et de pneumologie du CHU de Rouen. L'indexation automatique obtenue est confrontée à deux références manuelles. L'une établie avant même l'indexation automatique (codage médico-économique réalisé par le personnel hospitalier sur l'ensemble des dossiers) et l'autre établie après l'indexation automatique (codage médical réalisé par un expert-pneumologue sur les dossiers du service de pneumologie). Une première indexation donne des résultats mitigés.

Cependant, une analyse de l'indexation automatique obtenue permet d'identifier deux sources importantes de silence. Tout d'abord, le dictionnaire MeSH utilisé ne contient pas tous les termes qu'il serait souhaitable d'identifier dans les dossiers (couverture de 33% du MeSH au début de ce travail). D'autre part, l'ensemble des liens UMLS permettant le transcodage MeSH $\leftrightarrow$ CIM-10 est également incomplet. Des liens entre les synonymes MeSH et les termes CIM-10 ont été rajouté dans la table de transcodage, afin de la compléter. Cependant, l'extracteur MeSH fait déjà le lien entre les synonymes MeSH et les termes MeSH. Cette modification de la table de trancodage est utile sur le plan de la description des liens entre les termes MeSH et CIM-10, mais n'aura pas d'impact sur le codage automatique par notre méthode. Une deuxième indexation des dossiers avec un dictionnaire MeSH complété (couverture de 83%)permet d'évaluer l'impact de ce facteur.

On observe une diminution du silence (il était de 75% lors de la 1ère expérience, contre 71% lors de la deuxième, pour les compte-rendus de cardiologie) - certains termes ne sont toujours pas extraits par la méthode du fait de la variabilité de certaines expressions non prises en compte par le dictionnaire. Par ailleurs, la table de transcodage reste incomplète, notamment du fait de la restriction du transcodage à des équivalences unaires (un terme MeSH correspond à un terme CIM-10). Il semble en effet pertinent dans certains cas de pouvoir créer des équivalences entre n mots clés MeSH et m termes CIM-10. Par exemple pour (n=2, m=1) :

$$<\textit{diabète de type I}>_{MeSH} + <\textit{coma diabétique}>_{MeSH} \leftrightarrow <\textit{diabète sucre insulino-dépendant, avec coma }>_{CIM-10}$$

Ce type d'équivalence pourrait améliorer les résultats du codage automatique CIM-10.

On observe également que certains codes sont proposés par les professionnels de santé alors que les concepts auxquels ils se rapportent ne figurent pas dans le dossier. Ainsi, ces codes sont issus d'une connaissance du médecin extérieure au dossier du patient et il n'est pas possible de les extraire automatiquement par la seule analyse du dossier. Afin d'enrichir le codage automatique, deux sources de connaissances extérieures au dossier patient peuvent être utilisées :

1. Les pathologies associées aux médicaments prescrits dans le dossier
2. La base de connaissance constituée par les autres dossiers patient du service

L'exploitation du système de reconnaissance des médicaments et pathologies associées du VIDAL permet de mettre en oeuvre la première de ces approches, et d'extraire l'ensemble des pathologies recherchées (rappel proche de 70%). Cependant, la précision reste faible.

5.11 Conclusion sur l'Indexation

Dans ce chapitre, nous avons établi que l'activité d'indexation est un problème complexe, qui nécessite une redéfinition par les acteurs afin d'être résolu. Dans le cadre de l'indexation automatique, on peut ainsi l'aborder de plusieurs manières. Les approches implémentées dans le système MAIF envisagent tour à tour l'indexation comme une traduction conceptuelle (approche TAL) ou comme une catégorisation (approche k-PPV). Les deux approches de MAIF ont été développées avec l'objectif principal d'extraire des *paires* des descripteurs MeSH, et de fournir un index de taille adaptée à une ressource donnée et aux performances du système sur cette ressource.

Les évaluations effectuées ont permis de valider les méthodes d'extraction de paires mot clé / qualificatifs, et la fonction de rupture. Par ailleurs, MAIF a également été comparé favorablement aux autres systèmes d'indexation MeSH francophones sur l'extraction de descripteurs MeSH isolés. La confrontation avec plusieurs systèmes d'indexation (les systèmes francophones et MTI) a permis de mettre en lumière la complémentarité des ressources et des approches utilisées, ainsi que les efforts à faire pour les systèmes francophones par rapport à la référence internationale qu'est MTI. Ces constatations ouvrent des perspectives de recherche pour l'amélioration des divers systèmes, à travers la mise en commun des ressources linguistiques pour le français, et des méthodes pour MTI.

Finalement, plusieurs projets dans le domaine de la terminologie Médicale (CCAM) ou de l'indexation automatique (indexation bimodale texte/image ou codage CIM-10) ont montré que MAIF peut être mis à profit dans des activités connexes à l'indexation de ressources de santé. Ainsi, il entrera dans l'architecture d'un extracteur SNOMED et CIM-10 dont le développement est amorcé avec la thèse de S. Pereira. Il pourra également être adapté pour l'indexation à l'aide de triplets mot clé / qualificatif \ type de ressource, un travail qui s'inscrit dans le cadre de la thèse de F. Florea.

Chapitre 6

Conclusion et Perspectives

Pour conclure sur ce travail, nous revenons sur les différentes réalisations en les replaçant dans le cadre global du projet CISMeF. Puis, nous reprenons les perspectives de recherche envisagées.

6.1 Réalisations

Dans cette thèse, nous nous sommes principalement intéressée à l'analyse de ressources textuelles médicales, et en particulier à l'indexation automatique de ressources de santé à l'aide de descripteurs MeSH.

Nous avons contribué à l'enrichissement des ressources terminologiques disponibles dans le domaine médical en proposant des méthodes de traduction automatique des synonmyes américains du MeSH pour compléter les dictionnaires MeSH et la bibliothèque de transducteurs que nous avons développés. Ces ressources terminologiques ont permis d'enrichir la terminologie CISMeF, et peuvent être exploitées au sein du catalogue, en particulier dans le cadre de l'indexation automatique.

En nous démarquant de l'approche traditionnelle, qui consiste à extraire des termes MeSH isolés, nous avons proposé deux méthodes pour extraire des paires mot clé /qualificatif MeSH. Ces deux approches (TAL et k-PPV) ont été implémentées et combinées dans le système MAIF. Au travers de plusieurs évaluations, nous avons montré que MAIF offre des performances satisfaisantes pour l'extraction de paires MeSH et des performances au moins équivalentes à celles des autres systèmes d'indexation MeSH francophones pour l'extraction de descripteurs isolés. Par ailleurs, nous avons également montré que les méthodes d'extraction de paires mot clé/qualificatif mises en place par MAIF, ainsi que le calcul d'un seuil adaptatif pour la sélection des descripteurs pouvaient être mises à profit pour améliorer les performances du système d'indexation américain MTI.

Afin de caractériser les documents de manière plus générale, nous avons également étudié deux méthodes de catégorisation de ressources de santé pour le catalogue CISMeF, l'une intervenant avant l'indexation MeSH et l'autre après. Cette dernière est actuellement utilisée quotidiennement dans le catalogue, et a donné lieu à l'élaboration d'un outil bibliométrique.

La figure 6.1 présente une vue d'ensemble du fruit de notre travail de thèse au sein de l'équipe CISMeF.

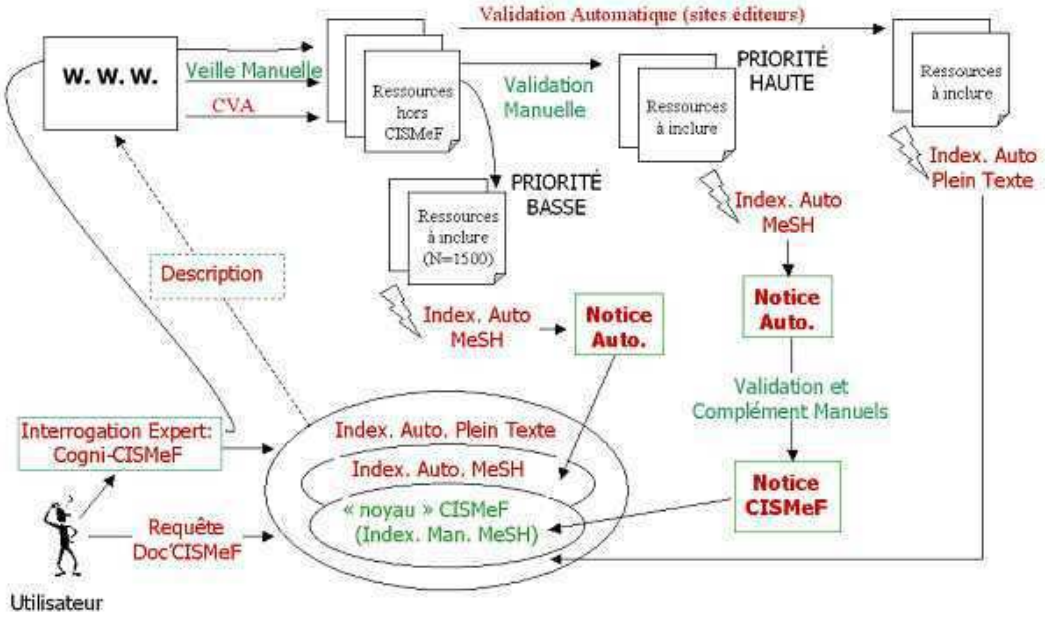

FIG. 6.1 – Fonctionnement de CISMeF après automatisation des taches documentaires (perspective 2006)- en vert les tâches effectuées manuellement, en rouge les tâches automatisées

Plusieurs progrès sont visibles par rapport au fonctionnement du catalogue au début de nos travaux. La sélection des ressources à ajouter au catalogue est effectuée semi-automatiquement à l'aide de l'outil de veille automatique CVA présenté à la section 3. L'attribution d'une priorité au ressources retenues résulte d'une réflexion au sein de l'équipe sur l'utilisation effective d'un outil d'indexation automatique MeSH. Compte tenu des performances de MAIF, il a été décidé d'utiliser cet outil de deux manières : en mode automatique pour les ressources de priorité « basse » et semi-automatique pour les ressources de priorité « haute »). En plus de l'indexation MeSH ainsi obtenue, les ressources bénéficient également d'une indexation plein texte de « bas niveau[1] »qui est utilisée par le moteur de recherche quand les stratégies de recherche utilisant le vocabulaire contrôlé sont en échec. Par suite, la catégorisation en spécialités médicales est déduite de l'indexation MeSH à l'aide de l'algorithme présenté à la section 4.2.4. A partir de ces éléments, une notice décrivant la ressource peut être ajoutée dans le catalogue.

En parallèle à la thèse que j'ai menée, l'équipe CISMeF a amélioré de façon drastique l'algorithme de la recherche simple (L. Soualmia et al., 2005) et contribué au développement de nouvelles ressources terminologiques dans le domaine de la santé (Douyère et al., 2004).

Bien que cela ne figure pas explicitement, l'utilisateur intervient indirectement à chacune des étapes du processus CISMeF. En effet, le catalogue étant destiné à un certain type d'utilisateur (étudiant, médecin, patient) la politique éditoriale de sélection des ressources doit refléter les besoins d'information en santé ce public. De même, la définition des priorités est fondée sur les demandes effectives. Par exemple, l'analyse des requêtes des utilisateurs

[1]Cette indexation consiste en une segmentation brute des textes, avec stockage des mots ainsi recensés dans un index. Elle est effectuée à l'aide d'un outil d'indexation Oracle.

permet de classer en priorité « haute »les ressources correspondant à des demandes répétées qui n'ont pu être satisfaites (requêtes ayant abouti à « zéro réponse »pour une recherche sur la terminologie CISMeF). Pour ce qui est de l'indexation, elle doit décrire la ressource de façon à faire ressortir les aspects qui sont susceptibles d'intéresser les utilisateurs. Ainsi, il apparaît évident que « le processus de recherche d'information est centré sur l'utilisateur. Sans une bonne compréhension des besoins de l'utilisateur, quelle que soit la qualité des autres éléments du système - y compris l'indexation - l'utilisateur ne pourra être satisfait. » (Lancaster, 1991)[2]

Les éléments d'automatisation des tâches documentaires mis en place dans le catalogue CISMeF tiennent compte de cette réalité. L'intervention des documentalistes aux points clés (validation des ressources issues de la veille, validation de l'indexation MeSH des ressources de priorité « haute ») du traitement de l'information médicale dans CISMeF en témoigne.

6.2 Perspectives

Terminologie Médicale. Dans le cadre des projets VUMeF et UMLF auxquels nous avons participé, nous entendons poursuivre (en collaboration avec Sylwia Ozdowska) l'enrichissement des ressources médicales francophones grâce à la traduction automatique de synonymes MeSH. L'application des méthodes exposées à la section 2.5.2 à la traduction des synonymes du MeSH 2005 à l'aide du corpus CESART est prévue pour fin 2005. Par ailleurs, nous souhaitons également approfondir l'étude des corpus utilisables pour ce travail, en comparant les résultats obtenus sur un corpus médical tel que CESART et sur un corpus juridique tel que le Hansart.

Les résultats obtenus (nouveaux synonymes traduits) seront intégrés à la terminologie CISMeF, ainsi qu'à nos dictionnaires MeSH. Ces derniers seront mis à la disposition de la communauté scientifique par l'intermédiaire de NOOJ et enrichis dans le cadre de la thèse de Suzanne Pereira avec le développement d'un extracteur SNOMED.

Indexation. L'intégration de MAIF dans le processus CISMeF doit être finalisée dans les semaines qui suivront la soutenance de cette thèse et intégrer l'utilisation de NOOJ, qui prendra le relai d'INTEX.

Une amélioration du système réside dans l'utilisation de données statistiques issues de CISMeF et/ou de MedLine et plus précisément les coefficients de sélection et les coocurrences entre termes.

Un travail important reste également à accomplir pour généraliser l'extraction des paires mot clé / qualificatif et l'application des règles d'indexation destinées à réviser la liste des candidats MeSH proposés par MAIF. Nous envisageons d'accomplir ce travail également pour l'anglais afin d'implémenter l'extraction de paires mot clé / qualificatif dans MTI. Ce point précis pourrait faire l'objet d'une collaboration avec des collègues de la NLM dans le cadre d'un post-doc en 2006. Nos travaux pourront également s'orienter sur l'adaptation d'une partie de l'algorithme de MTI pour le français. Des travaux récents (non publiés) de V. Claveau et P. Zweigenbaum ont mis en évidence la difficulté d'une telle tâche.

D'autres perspectives sont également ouvertes pour des travaux autour de MAIF dans l'équipe CISMeF : l'indexation automatique à l'aide de triplets (collaboration avec Filip Flo-

[2]Notre traduction.

rea), l'intégration de MAIF ou de son architecture dans des systèmes d'indexation CIM-10 et SNOMED (collaboration avec Suzanne Pereira, dans le cadre de sa thèse).

Evaluation. La mise en oeuvre des différents points mentionnés ci-dessus dans MAIF et MTI devra donner lieu à de nouvelles évaluations, afin d'observer l'impact sur les performances des logiciels. Notre expérience lors des précédentes évaluations sur la construction et la manipulation des corpus pourra faciliter ces futures évaluations.

Par ailleurs, l'utilisation effective de MAIF par les documentalistes devra également être évaluée, avec une attention particulière portée aux questions suivante : MAIF permet-il un gain de temps lors de l'indexation ? MAIF permet-il de réduire le silence de l'indexation manuelle ? Il pourra également être opportun d'envisager plusieurs modes de présentation des candidats MeSH extraits par MAIF (sous forme de page web, sous forme de fiche access pré-remplie...), afin d'évaluer celui qui parait le plus utile et le plus pratique pour les documentalistes.

Recherche d'Information. Enfin, le rapprochement qui peut être fait entre l'approche d'indexation k-PPV de MAIF et l'algorithme PRC implémenté dans MTI devrait prochainement donner lieu à un nouvel outil de recherche dans CISMeF (en collaboration avec Badisse Dahamna et Suzanne Pereira), proposant aux utilisateurs de visualiser les ressources les plus proches d'une ressource donnée. Pour cela, nous nous appuierons sur les travaux de Kim et al. (2001) sur l'algorithme « Pubmed Related Citations ».

Par ailleurs, une collaboration avec Thibault Roy (GREYC, Caen) est en cours sur le thème de la cartographie de corpus de santé. Pour cela, nous nous fondons sur l'algorithme de catégorisation après indexation MeSH décrit à la section 4.2.4, sur l'outil bibliographique qui en a découlé ainsi que sur les travaux de Roy (2005). Afin de faciliter l'exploration des corpus catégorisés, nous souhaitons proposer aux utilisateurs des cartes regroupant les documents d'un corpus parmi les spécialités médicales les plus fréquentes et indiquant visuellement les proximités thématiques entre documents.

Bibliographie

Abdallah, I. (1998). *Segmentation et codage de signaux de parole par critères entropiques*. Unpublished doctoral dissertation, Université du Maine.

Abeillé, A., & Clément, L. (2003). *Annotation morpho-syntaxique.* Document en ligne disponible sur `http://www.llf.cnrs.fr/fr/Abeille/guide-morpho-synt.02.pdf` (consulté le 12/06/05).

Accart, J. P., & Rethy, M. P. (2003). *Le métier de documentaliste* (2ème ed.). Paris : Editions du Cercle de la librairie.

Ahrenberg, L., Andersson, M., & Merkel, M. (2000). A knowledge-lite approach to word alignment. Kluwer.

Al-Kharashi, I. A., & Evens, M. W. (1994). Comparing words, stems and roots as index terms in an arabic information retrieval system. *JASIS*, *45*(8), 548–560.

Anderson, J. D., & Perez-Carballo, J. (2001). The nature of indexing : how humans and machines analyze messages and texts for retrieval. part i : Research, and the nature of human indexing. *Information Processing and Management*, *2*(37), 231–254.

Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus : the metamap program. In *Proc. AMIA Symp. 2001* (pp. 17–21).

Aronson, A. R., Mork, J. G., Gay, C. W., Humphrey, S. M., & Rogers, W. J. (2004). The nlm indexing initiative's medical text indexer. In *Proc. Medinfo 2004* (pp. 268–272).

Aït El Mekki, T., & Nazarenko, A. (2004). L'index de fin de livre, une forme de résumé indicatif ? *Revue Traitement Automatique des Langues*, *1*(45), 121–150.

Bachimont, B. (2000). Engagement sémantique et engagement ontologique : Conception et réalisation d'ontologies et ingénierie des connaissances. (Ingénierie des Connaissances, Evolutions Récentes et Nouveaux Défis)

Baker, T. (2000). A grammar of dublin core. *Digital-Library Magazine*, *6*(10).

Barthes, R. (1985). *Analyse textuelle d'un conte d'edgar poe, l'aventure sémiologique*. Paris : Seuil.

Bécue-Bertaut, M. (2003). Comparaison des structure induites sur un ensemble de réponses ouvertes par le choix de l'unité statistique. *Corpus*(2). (La distance intertextuelle)

Bell, T., Cleary, J. G., & Witten, I. H. (1990). *Text compression.* NJ : Prentice Hall.

Berrios, D. C., Cucina, R. J., & Fagan, L. M. (2002). Methods for semi-automated indexing for high precision retrieval. *Journal of American Medical Informatics Association*, *9*(6), 637–651.

Bertrand, A. (1993). *Compréhension et catégorisation dans une activité complexe. l'indexation de documents scientifiques*. Unpublished doctoral dissertation, Université de Toulouse le Mirail.

Besançon, R., Chappelier, J. C., Rajman, M., & Rozenknop, A. (2001). Improving text representations through probabilistic integration of synonymy relations. In *Proc. of the*

Xth International Symposium on Applied Stochastic Models and Data Analysis (ASMDA'2001) (pp. 200–205).

Blanco, X., & Bonell, C. (1998). Vers une structuration syntactico-sémantique de la terminologie médicale. *Cahiers de Grammaire - Université Toulouse le Mirail*(23).

Bodenreider, O. (2000). Using UMLS semantics for classification purposes. In *Proc. AMIA Symp. 2000* (pp. 86–90).

Bodenreider, O., Nelson, S. J., Hole, W. T., & Chang, H. F. (1998). Beyond synonymy : exploiting the UMLS semantics in mapping vocabularies. In *Proc. AMIA Symp. 1998* (pp. 815–819).

Bommier-Pincemin, B. (1999). *Diffusion ciblée automatique d'informations :conception et mise en oeuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*. Unpublished doctoral dissertation, Université Paris IV Sorbonne. (ch. VII pp. 415–427)

Botafogo, R. A. (1993). Cluster analysis for hypertext systems. In *Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 116–125).

Bouaud, J., Bachimond, B., & Zweigenbaum, P. (1996). Traitement de la métonymie basé sur un modèle du domaine et sur une recherche heuristique de chemin dans des graphes. In *Actes de TALN 1996).*

Bourigault, D., & Fabre, C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire - Université Toulouse le Mirail*(25), 131–151.

Bouroche, J. M., & Saporta, G. (1983). *L'analyse des données* (2$^{\text{ème}}$ ed.). PUF. (Collection Que Sais-je ?)

Bourrigault, D. (1994). *Lexter, un logiciel d'extraction de terminologie. application à l'acquisition des connaissances à partir de textes*. Unpublished doctoral dissertation, Ecole des Hautes Etudes en Sciences Sociales, Paris.

Bradord, S. C. (1934, sep). Sources of information on specific subjects. *Engineering*, *137*, 85–86.

Briet, S. (1951). *Qu'est-ce que la documentation ?* Paris : Editions documentaires et industrielles.

Buckland, M. K. (1997, sep). What is a document ? *Journal of the American Society of Information Science*, *48*(9), 804–809.

Cabré, M. T. (2002). Terminología y lingüística : La teoría de las puertas. *Estudios de Lingüística Española (ELIES)*, *16*. (Disponible sur `http://elies.rediris.es/elies16/Cabre.html` (accédé le 10/04/05))

Cai, L., & Hofmann, T. (2004). Hierarchical document categorization with support vector machines. In *Proc. CIKM 2004* (pp. 396–402).

Chiao, Y. C., & Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In *Proc. COLING 2002)* (pp. 1208–1212).

Claveau, V., & L'Homme, M. C. (2005). Apprentissage par analogie pour la structuration de terminologie- utilisation comparée de ressources endogènes et exogènes. In *Actes de TIA 2005)* (pp. 59–70).

Claveau, V., & Zweigenbaum, P. (2005). Traduction de termes biomédicaux par inférence de transducteurs. In *Actes de TALN 2005)* (pp. 253–362).

Cooper, W. S. (1969). Is inter-indexer consistency a hobgoblin ? *Am Doc*(21).

Corston-Olivier, S. H., & Dolan, W. B. (1999). Less is more : Eliminating indexing terms rom subordinate clauses. In *Proc. of the 37th ACL meeting)* (p. 349-356).

Coté, R. A., & Al. (1997). *The systematised nomenclature of human and veterinary medicine : Snomed international.*

Courtois, M., B et Silberztein. (1990). *Dictionnaires électroniques du français.* Paris : Larousse.

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques.* Unpublished doctoral dissertation, Université Paris 7.

Damereau, F. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, *7*, 659–664.

Dameron, O. (2003). *Modélisation, représentation et partage de connaissances anatomiques sur le cortex cérébral.* Unpublished doctoral dissertation, Université Rennes 1.

Darmoni, S. J., Amsallem, E., Haugh, M. C., Lukacs, B., Chalhoub, C., & Leroy, J.-P. (2003). Level of evidence as a future gold standard for the content quality of health resources on the internet. *Methods of Information in Medicine*, *42*(3), 200–225.

Darmoni, S. J., Jarousse, E., Zweigenbaum, P., Le Beux, P., Namer, F., Baud, R., et al. (2003). VUMeF : Extending the french involvement in the UMLS metathesaurus. In *Proc. AMIA Symp. 2003* (p. 824).

Darmoni, S. J., Leroux, V., Thirion, B., Santamaria, P., & Gea, M. (1999). Netscoring : critères de qualité de l'information de santé sur internet. *Les enjeux des industries du savoir*, 29–44.

Darmoni, S. J., Leroy, J.-P., Baudic, F., Douyère, M., Piot, J., & Thirion, B. (2000). CISMeF : a structured health resource guide. *Methods of Information in Medicine*, *39*, 30–35.

Darmoni, S. J., Mayer, M. A., Thomeczek, C., & Eysenbach, G. (2002). MedCIRCLE : un projet européen pour évaluer la qualité de l'information de santé. In *Internet et pédagogie médicale.*

Darmoni, S. J., Névéol, A., Renard, J. M., Gehanno, J. F., Soualmia, L. F., Dahamna, B., et al. (2005). A MEDLINE categorization algorithm. *BMC*, in press.

David, C., Giroux, L., Bertrand-Gastaldy, S., & Lanteigne, D. (1995). Indexing as problem solving- a cognitive approach to consistency. In *Proc. ACSI/CAIS.*

Debili, F. (1997). L'appariement : quels problèmes ? In *Actes des 1ères JST FRANCIL de l'AUPELF UREF* (pp. 199–206).

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of American Society for Information Science*, *6*(41), 391–407.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1992). Le traitement du langage naturel dans la recherche d'information. *Interface intelligente dans l'information scientifique et technique*, 103–130.

Douyère, M., Soualmia, L., Névéol, A., Rogozan, A., Dahamna, B., Leroy, J.-P., et al. (2004, dec). Enhancing the MeSH thesaurus to retrieve French online health resources in aquality-controlled gateway. *Health Info Libr J.*, *21*(4), 253–61.

Dumais, S. (1998). Using svms for text categorization. *IEEE Intelligent Systems Magazine, Trends and Controversies*, *13*(4).

Dumais, S., Platt, J., Heckerman, D., & Saham, M. (1998). Inductive learning algorithms and representations for text categorization. *CIKM*, 148–155.

Ferber, J. (1995). *Les systèmes multi-agents : vers une intelligence collective.* Inter-Editions.

Florea, F. I., Rogozan, A., Bensrhair, A., & Darmoni, S. J. (2005). Medical image retrieval by content and keyword in a on-line health-catalogue context. In *Proc. Mirage 2005* (p. 229-236).

Fowler, J., Maram, S., Kouramajian, V., & Devadhar, V. (1995). Automated MeSH Indexing of the World Wide Web. In *AMIA Symp.* (pp. 893–897).

Franck, E., Chui, C., & Witten, I. H. (2000). Text categorization using compression models. In *Proc. Conference on Data Compression* (p. 55).

François, C. (2003). *Outils de veille : Typologie.* Présentation aux rencontres des professionnels de l'IST. (Disponible sur `http://www.inist.fr/rencontresIST/docpdf/4veille.PDF` (consulté le 01/08/05))

Frawley, W. (1988). New forms of specialised dictionaries. *International Journal of Lexicography, 1*(3), 89–213.

Frege, G. (2001). On sense and reference. Blackwell.

Funk, M. E., Reid, C. A., & McGoogan, L. S. (1983). Indexing consistency in MEDLINE. *Bull. Med. Libr. Assoc., 2*(71), 176–183.

Furner, J., Ellis, D., & Willett, P. (1999). Inter-linker consistency in the manual construction of hypertext documents. *ACM Computing Surveys, 4es*(31). (On-line supplement : article no. 18)

Gale, W. A., & Church, K. W. (1991). Identifying word correspondences in parallel text. In *Proc. of the DARPA Workshop on Speech and Natural Language.*

Gaudinat, A., Boyer, C., Baujard, V., & Ruch, P. (2002). Evaluation de l'extraction de termes mesh pour les systèmes de recherche d'information dans le domaine médicale. In *Actes des 9ièmes Journées Francophones d'Informatique Médicale.*

Gaudinat, A., Joubert, M., Aymard, S., Falco, L., Boyer, C., & Fieschi, M. (2004). Wrapin : New generation health search engine using umls knowledge sources for mesh term extraction from health documentation. In *Proc. Medinfo 2004* (pp. 356–360).

Gaussier, E. (2001). General considerations on bilingual terminology extraction. John Benjamins Publishing Company.

Genette, G. (1987). *Seuils* (Seuil, Ed.). Paris.

Grabar, N. (2004). *Terminologie médicale et morphologie. acquisition de ressources morphologiques et leur utilisation pour le traitement de la variation terminologique.* Unpublished doctoral dissertation, Université Paris 6.

Grabar, N., & Zweigenbaum, P. (2000). Automatic acquisition of domain-specific morphological resources from thesauri. In *Proc. RIAO 2000 : Content-Based Multimedia Information Access* (p. 765-784).

Grimes, J. (1990). Inverse lexical functions. Ottawa University Press.

Gross, G. (1992). Forme d'un dictionnaire électronique. Presse de l'Université du Québec.

Gruber, T. (1993). A translation approach to portable ontology. Philosophia Verlag.

Guarino, N. (1996). *Understanding, building and using ontologies.* Document en ligne sur `http://ksi.cpsc.ucalgary.ca/KAW/KAW96/guarino/guarino.html`. ((consultté le 15/06/05))

Halleb, M., & Lelu, A. (1997). Hypertextualisation automatique multilingue à partir des fréquences des n-grammes. *Hypertextes et hypermédias, 1*(2-3-4), 275–287.

Hamon, T. (2005). Indexing specialized documents : are terminological resources sufficient ? In *Actes de TIA 2005.*

Han, J., & Kamber, M. (2001). Mining text databases. In *Data mining - concepts and techniques* (pp. 428–44). Morgan Kaufman Series in Data Management Systems.

Harris, Z. (1991). *A theory of language and information. a mathematical approach.* Oxford : Oxford University Press.

Herring, M. Y. (2001, avr). Ten reasons why the internet is no substitute for a library. *American Libraries*, 76–78.

Ho, J. (2005). Hyperlink obsolescence in scholarly online journals. *Journal of Computer-Mediated Communication*, *10*(3). (article 15. `http://jcmc.indiana.edu/vol10/issue3/ho.html` (consulté le 01/07/05))

Holzem, M. (2000). Termes d'indexation et construction des connaissances. 43–52. (Sémantique des termes spécialisés - Collection Dyalang)

Hooper, R. S. (1965). *Indexer consistency tests : origin, measurement, results and utilization* (Tech. Rep.). IBM Corporation. (Bethesda, MD)

Hull, D. A., & Grefensette, G. (1996). Experiments in multilingual information retrieval. In *Proc. of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Hunh, C., Wermter, S., & Smith, P. (2004). Predictive top-down knowledge improves neural exploratory bottom-up clustering. In *Proc. ECIR* (pp. 154–166).

Jacquemin, C. (1994). Fastr : a unification-based front-end to automatic indexing. In *Proc. RIAO'94* (pp. 34–47).

Jacquemin, C. (1997). *Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus.* Mémoire d'habilitation à diriger des recherches. (Université de Nantes.)

Jacquemin, C., & Royauté, J. (1994). Retrieving terms and their variants in a lexicalised unification-based framework. In *Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 132–141).

Janssen, T. M. V. (1997). Compositionality. Elsevier.

Joachim, T. (1998). Text categorization with support vector machines : Learning with many relevant features. In Springer-Verlag (Ed.), *Proc. of the Tenth European Conference on Machine Learning (ECML'98)* (pp. 137–142).

Joubert, M., Peretti, A. L., Gouvernet, J. F., & Fieschi, M. (2005). Refinement of an automatic method or indexing medical litterature : a preliminary study. In *Actes de MIE 2005.*

Jousse, A. L., & Bouveret, M. (2003). Lexical functions to represent derivational relations in specialised dictionaries. *Terminology*, *9*(1), 71–98.

Kim, W., Aronson, A. R., & Wilbur, W. J. (2001). Automatic mesh term assignment and quality assessment. In *Proc. AMIA Symp. 2001* (pp. 319–323).

Kleber, G. (1990). *La sémantique du prototype. catégorie et sens lexical* (PUF, Ed.). (Coll. « Linguistique Nouvelle »)

Kraaij, W., & Pohlmann, R. (1993). Viewing morphology as an inference process. In *Proc. of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 191–203).

Kraaij, W., & Pohlmann, R. (1996). Viewing stemming as recall enhancement. In *Proc. of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 40–48).

Lahtinen, T. (2000). *Automatic indexing : an approach using an index term corpus and combining linguistic and statistical methods.* Unpublished doctoral dissertation, University of Helsinki.

Lam, W., Ruiz, M., & Srinivasan, P. (1999). Automatic text categorisation and its application to text retrieval. *IEEE Transactions on Knowledge and Data Engineering*, *11*(6), 865–879.

Lancaster, F. W. (1991). *Indexing and abstracting in theory and practice.* University of Illinois : Champaign, IL.

Landes, D., & Spidal, D. (2003). An index comparison project : The effects of two indexers' diverse backgrounds on creating an index from a software manual. In *Proc. ASI-IASC/SCAD.*

Larkey, L. S., & Croft, W. B. (1996). Combining classifiers in text categorization. In *Proc. of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 289–297).

Leininger, K. (2000, mar). Interindexer consistency in psycINFO. *Journal of Librarianship and Information Science*, *1*(32).

Lelu, A., & Ferhan, S. (1998). Clustering a textual dataflow by incremental density-modes seeking. In *Proc. IFCS'98.*

Lelu, A., Halleb, M., & Delprat, B. (1998). Recherche d'information et cartographie dans des corpus textuels à partir des fréquences de n-grammes. In *Actes des JADT* (pp. 391–400).

Lenoir, P., Michel, J. R., Frangeul, C., & Chales, G. (1981). Réalisation, développement et maintenance de la base de données A.D.M. *Médecine informatique*(6), 51–56.

Leonard, L. E. (1977). Inter-indexer consistency studies, 1954-1975 : a review of the literature and summary of study results. *University of Illinois Graduate School of Library Science Occasional Papers*(131).

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.*(6), 707–710.

L'Homme, M. C. (2002). Fonctions lexicales pour modéliser les relations sémantiques entre termes. *Revue TAL*(2).

Loisel, A., Kotovic, J. P., Chaignaud, N., & Darmoni, S. J. (2005). Un système de dialogue homme-machine pour un moteur de recherche de documents médicaux. In *Poster pour IHM 2005.*

Lotka, A. J. (1960). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*(16), 317–323.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical translation and computational linguistics*(11), 22–31.

Lovis, C. (1996). *Codage médico-économique des diagnostics et procédures.* Thèse de Mèdecine. (Université de Genève)

Malaisé, V., Zweigenbaum, P., & Bachimont, B. (2003). Vers une combinaison de méthodologies pour la structuration de termes en corpus. In *Actes de ISKO 2003.*

Manning, C. D., & Shütze, H. (2000). *Foundations of statistical natural language processing.* Cambridge, Massashusetts : MIT Press.

Markey, K. (1984). Interindexer consistency tests : a literature review and report of a test of consstency in indexing visual materials. *Library and Information Science Research*, *2*(6), 155–177.

Marko, K., Daumke, P., Schultz, S., & Hahn, U. (2003). Cross-language mesh indexing using morpho-semantic normalization. In *Proc. AMIA Symp. 2003* (pp. 425–429).

Mary, V., Pouliquen, B., Le Duff, F., Darmoni, S. J., Segui, A., & Le Beux, P. (2002). Automatic conceptual indexing of french pharmaceutical theses. *Stud Health Technol Inform*(90), 388–392.

Mayer, M. A., Darmoni, S. J., Fiene, M., & Al. (2003). MedCIRCLE - modeling a collaboration

for internet rating, certification, labeling and evaluation of health information on the semantic world-wide-web. In *Medical Informatics Europe* (pp. 667–672).

Mc Cray, A. T. (1989). The UMLS semantic network. *SCAMC - Washington D.C.*, 503–507.

Mel'čuk, I. (1997). *Vers une linguistique sens-texte.* Paris- Collège de France. (Leçon inaugurale.)

Mel'čuk, I., Arbatchewsky-Jumarie, N., Elnitsky, L., Iordanskaja, L., Lessart, A., Dagenais, L., et al. (1984, 1988, 1992, 1999). *Dictionnaire explicatif et combinatoire du français contemporain. recherches lexico-sémantiques, volumes I-IV.* Montréal : Les Presses de l'Université de Montréal.

Michaud, S., & Waller, M. (2005). *La veille documentaire : Demeurez au courant des nouvelles publications.* CEFES–BLSH – Université de Montréal. (Notes de cours)

Minel, J. L. (2004). Le résumé automatique de textes : solutions et perspectives. *Traitement Automatique de la Langue*, *45*(1). (Numéro Spécial)

Mortchev-Bouveret, M. (2005). Modélisation de relations sémantiques dans un dictionnaire spécialisé au moyen de fonctions lexicales. In L'Homme & Vandaele (Eds.), *Lexicographie et terminologie : compatibilité des modèles et des méthodes, actes du colloque ACFAS.*

Nakache, D., & Métais, E. (2005). Evaluation : nouvelle approche avec juges. In *Inforsid* (p. 555-570).

Namer, F. (2000). Flemm : Un analyseur flexionnel du français à base de règles. *Revue Traitement Automatique des Langues*, *41*(2).

Namer, F. (2005). Morphosémantique pour l'appariement de termes dans le domaine médical : approche multilingue. In *Actes de TALN 2005* (pp. 63–72).

Nazarenko, A. (2004). *Donner accès au contenu des documents textuels : acquisition de connaissances et analyse de corpus spécialisés.* Mémoire d'Habilitation à Diriger des Recherches en Informatique. (Université Paris 13)

NF Z 47-102. (1978). (Principes généraux pour l'indexation des documents)

Nigam, M. K., McCallum, A., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine Learning*, *39*(2/3), 103–134.

Névéol, A. (2004). Indexation automatique de ressources de santé à l'aide d'un vocabulaire contrôlé. In *Actes de RECITAL 2004* (pp. 105–114).

Névéol, A., Douyère, M., Rogozan, A., & Darmoni, S. J. (2004). Construction de ressources terminologiques en santé pour un système d'indexation automatique. In *Actes des journées INTEX/NOOJ 2004.*

Névéol, A., Mary, V., Gaudinat, A., Boyer, C., Rogozan, A., & Darmoni, S. J. (2005). A Benchmark Evaluation of the French MeSH Indexing Systems. In *Actes de AIME 2005* (pp. 251–255).

Névéol, A., Mork, J. G., Aronson, A. R., & Darmoni, S. J. (2005). Evaluation of French and English MeSH Indexing Systems with a parallel corpus. In *Actes de AMIA 2005* (p. à paraître).

Névéol, A., & Ozdowska, S. (2005). Extraction bilingue de termes médicaux dans un corpus parallèle. In *Actes des 5ème journées Extraction et Gestion des Connaissances* (pp. 655–666). Toulouse : Cépadués.

Névéol, A., Rogozan, A., & Darmoni, S. J. (2004). Automatic indexing of health resources in French with a controlled vocabulary for the CISMeF catalogue : a preliminary study. In *Actes de MEDINFO 2004* (p. 1772).

Névéol, A., Rogozan, A., & Darmoni, S. J. (2005a). Automatic indexing of online health

resources for a french quality controlled gateway. *Information Processing and Management*, à paraître.

Névéol, A., Rogozan, A., & Darmoni, S. J. (2005b). Indexation automatique de ressources de santé à l'aide de paires de descripteurs MeSH. In *Actes de TALN 2005.*

Névéol, A., Soualmia, L., Douyère, M., Rogozan, A., Thirion, B., & Darmoni, S. J. (2004). Using CISMeF MeSH « Encapsulated » Terminology and a Categorization Algorithm for Health Resources. *International Journal of Medical Informatics*, *1*(73), 57–64.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19–51.

OMS, O. (1993). *Classification statistique internationale des maladies et des problèmes de santé connexes.* (Dixième révision)

Ozdowska, S. (2004a). Appariement bilingue de mots par propagation syntaxique à partir de corpus français/anglais alignés. In *Actes de RECITAL 2004* (pp. 125–134).

Ozdowska, S. (2004b). Identifying correspondences between words : an approach based on a bilingual syntactic analysis of french/english parallel corpus. In *Proc. of the Workshop on Multilingual Linguistic Resources COLING'04.*

Ozdowska, S., Névéol, A., & Thirion, B. (2005). Traduction compositionnelle automatique de bitermes dans des corpus anglais/français alignés. In *Actes de TIA 2005* (pp. 83–94).

Parmentier, F. (1998). *Spécification d'une architecture émergente fondée sur le raisonnement par analogie : Application aux références bibliographiques.* Unpublished doctoral dissertation, Université Henri-Poincaré - Nancy 1.

Pauchet, A. (2005). *Modélisation cognitive d'interactions humaines dans un cadre de planification.* Unpublished doctoral dissertation, INSA de Rouen.

Pereira, S. (2005). *Etude de faisabilité de différentes méthodes d'optimisation du codage médico-économique.* Master d'Informatique Médicale. (Université Paris V)

Pincemin, B. (2004). *Compte rendu du n°2 de la revue* Corpus *sur « la distance intertextuelle ».* Texto. (Disponible sur : `http://www.revue-texto.net/Parutions/CR/Pincemin_CR.html`. (Consulté le 11 /06/05))

Poibeau, T. (2005). Parcours interprétatifs et terminologie. In *Actes de TIA 2005.*

Popovic, M., & Willett, P. (1992). The effectiveness of stemming for natural-language acess to solvene textual data. *Journal of American Society for Information Science*, *5*(43), 384–390.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *3*(14), 130–137.

Pottier, B. (1964). Vers une sémantique moderne. *Travaux de linguistique et de littérature*(1), 107–138.

Pouliquen, B. (2002). *Indexation de textes médicaux par indexation de concepts, et ses utilisations.* Unpublished doctoral dissertation, Université Rennes 1.

Poulos, M., Papavlasopoulos, S., & Chrissikopoulos, V. (2004). A text categorization technique based on a numerical conversion of a symbolic expression and an onion layers algorithm. *Journal of Digital Information*, *6.1.*

Rastier, F. (1995). Le terme : entre ontologie et linguistique. *La banque des mots*(7), 35–65. (Actes des 1ères journées Terminologies et Intelligence Artificielle - Numéro Spécial)

Rastier, F., Cavazza, M., & Abeillé, A. (1994). *Sémantique pour l'analyse.* Paris : Masson.

Rector, A. L., Nowlan, W. A., & Kay, S. (1992). Conceptual knowledge : the core of medical information systems. In *Proc. MEDINFO 1992* (pp. 1420–1426).

Ricoeur, P. (2004). *Sur la traduction* (Bayard, Ed.). Paris. (ISBN n° 2-227-47367-3)

Rocchio, J. J. (1971). *The smart retrieval system : Experiments in automatic document*

processing. New–Jersey : Prentice-Hall. (Chapter 14, Relevance Feedback in Information Retrieval pp. 313–323)

Rogozan, A., Névéol, A., & Darmoni, S. J. (2003). Using compression models for health resources categorisation prior to indexing. In M. Dojat, E. Keravnou, & P. Barahona (Eds.), *Proc. AIME 2003 - LNAI 2780* (pp. 81–85). Springer.

Rolling, L. (1981). Indexing consistency, quality and efficiency. *Information Processing and Management*, *2*(17), 69–76.

Rosch, E., & Mervis, C. B. (1975). Family ressemblances :study in the internal structure of categories. *Cognitive Psychology*(7), 573–605.

Rosenfeld, R. (2000). Two decades of statistical language modeling : Where do we go from here. In *Proc. of the IEEE 88(8).*

Rosetta, M. T. (2003). Compositional translation by rosetta : a book review. *Computational Linguistics*, *21*(4).

Roy, T. (2005). Une plate-forme logicielle dédiée à la cartographie thématique de corpus. In *Actes de RECITAL 2005* (pp. 545–555).

Ruch, P. (2002). Information retrieval and spelling errors : Improving effectiveness by lexical disambiguation. In *Proc. ACM-SAC, Information Access and Retrieval Track.*

Ruch, P., Baud, R., & Geissbühler, A. (2003). Learning-free text categorization. In M. Dojat, E. Keravnou, & P. Barahona (Eds.), *Proc. AIME 2003 - LNAI 2780* (pp. 199–204). Springer.

Ruch, P., Marty, J., Baud, R., Geissbühler, A., Tbahriti, I., & Veuthey, A.-L. (2005). Latent argumentative pruning for compact MEDLINE indexing. In *Proc. AIME 2005 - lnai 2780.* Springer.

Sager, R. N., & Friedman, C. (Eds.). (1987). *Medical language processing : Computer management of narrative data*. Addison-Wesley.

Salem, A. (1984). La typologie des segments répétés dans un corpus, fondée sur l'analyse d'un tableau croisant mots et textes. *Les cahiers de l'analyse de données*, *4*(9), 489–500.

Salton, G. (1989). *Automatic text processing : The transformation, analysis, and retrieval of information by computer*. Reading, MA : Addison-Wesley.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York : McGraw-Hill.

Sankoff, D., & Kruskal, J. B. (1983). *Time warps, string edits, and macromolecules : The theory and practice of sequence comparison*. Reading, Massachusetts : Addison-Wesley Publishing Company, Inc. (ISBN n° 0-201-07809-0)

Saracevic, T., & Kantor, P. (1988). Study of information seeking and retrieving : Part iii. searcher, searches and overlap. *Journal of American Society for Information Science*, *3*(39), 197–216.

Savary, A. (2000). *Recensement et description des mots composés- méthodes et applications*. Unpublished doctoral dissertation, Université Paris 7 et Université de Marne-la-Vallée.

Silberztein, M. (1993). *Dictionnaires électroniques et analyse automatique de textes : le système intex* (Masson, Ed.). Paris.

Silberztein, M. (1994). Groupes nominaux libres et noms composés lexicalisés. *Linguisticae Investigationes*, *XVII*(2), 405–425.

Simard, M., Cancedda, N., Cavestro, B., Dymetman, M., Gaussier, E., Goutte, C., et al. (2005). Une approche à la traduction automatique statistique par segments discontinus. In *Actes de TALN 2005* (Vol. 1, pp. 233–242).

Slodzian, M. (1995). Comment revisiter la doctrine terminologique aujourd'hui ? *La banque*

des mots(7), 11–18. (Actes des 1ères journées Terminologies et Intelligence Artificielle - Numéro Spécial)

Soergel, D. (1994). Indexing and retrieval performance : the logical evidence. *Journal of American Society for Information Science*.

Soualmia, L., Dahamna, B., & Darmoni, S. J. (2005). *Some strategies for health information searching* (Tech. Rep.). CISMeF.

Soualmia, L. F., Barry, C., & Darmoni, S. J. (2003). Knowledge-based query expansion over a medical terminology oriented ontology on the web. In M. Dojat, E. Keravnou, & P. Barahona (Eds.), *LNAI 2780 - Proc. AIME 2003.* Springer.

Soualmia, L. F., Barry-Greboval, C., Abdulrab, H., & Darmoni, S. J. (2002). Modélisation et représentation des connaissances dans un catalogue de santé. In *Actes de IC 2002* (pp. 139–149).

Sparck-Jones, K. (1995). Reflections on trec. *Information Processing and Management*, *31*(3), 291–314.

Sparck-Jones, K., Walker, S., & Robertson, S. E. (2000). A probalistic model o inormation retrieval : development and comparative experiments (part 1). *Information Processing and Management*, *36*(3), 779–808.

Sycara, K. (1998). Multiagent systems. *AI Magazine*(19).

Teahan, W. J., & Harper, D. J. (2001). Using compression based language models for text categorization. In J. Callan, B. Croft, & J. Lafferty (Eds.), *Workshop on Language Modelling and Information Retrieval* (pp. 83–88).

Teahan, W. J., & Harper, D. J. (2003). *Using compression based language models for text categorization*. Kluwer Accademic Publishers.

Thirion, B., & Darmoni, S. J. (1998). Les sites médicaux francophones sur internet : le devoir d'ingérence des bibliothèques. *Bulletin des Bibliothèques de France*(3), 42–45.

Thirion, B., Lacoste, B., Videau, S., Douyère, M., Leroy, J. P., Goupy, G., et al. (2000). Doc'cismef : un outil de recherche utilisant le thesaurus mesh. *La Revue du Praticien - Médecine Générale, 2000*, *506*(14), 1427–8.

Thirion, B., Loosli, G., Douyère, M., & Darmoni, S. J. (2003). Metadata element set in a quality-controlled subject gateway : a step to a health semantic web. In *Medical Informatics Europe.*

Thomas, A. (2004). Les outils de veille en sept étapes. *Veille magazine*(74), 36–40.

Vapnik, V. N. (1998). *The statistical learning theory*. Springer.

Vinot, R., Grabar, N., & Valette, M. (2003). Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'internet. In *Actes de TALN 2003.*

Virbel, J. (2002). Eléments d'analyse du titre. *Inscription Spatiale du Langage : structures et processus*, 123–132.

Véronis, J. (Ed.). (2000). *Parallel text processing : Alignment and use of translation corpora*. Dordrecht : Kluwer Academic Publishers.

Wiener, J. O., Perdersen, A. S., & Weigend, A. (1995). Neural network approach for topic spotting. In *Proc. of the 4^{th} Symposium on Document Analysis and Information Retrieval (SDAIR'95)* (pp. 317–332).

Wilbur, W. J., & Kim, W. (2003). The dimensions of indexing. In *Proc. AMIA Symp. 2003* (pp. 714–719).

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. textes choi-*

sis de terminologie 1, fondements théoriques de la terminologie. Presses de l'université de Laval.

Yang, Y., & Chute, C. D. (1994). An example-based mapping method for text categorisation and retrieval. *ACM Trans. Information Systems*, *12*(3), 252–277.

Zipf, G. K. (1935). The behaviour of words. *The psycho-biology of language*, 20–48.

Zweigenbaum, P. (1999). Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances. *ISIS*.

Zweigenbaum, P., Baud, R., Burgun, A., Namer, F., Jarousse, E., Grabar, N., et al. (2003). UMLF : construction d'un lexique médical francophone unifié. In *Actes des JFIM 2003.*

Zweigenbaum, P., & Consortium-MENELAS. (1995). MENELAS : coding and information retrieval from natural language patient discharge summaries. IOS Press.

Table des figures

1.1 Extrait de la page descriptive CISMeF pour le mot clé *diabète gestationnel* . 20
1.2 Extrait de la notice CISMeF n° 9785 (Notice courte) 20
1.3 Extrait de la notice CISMeF n° 9785 (Notice longue) 21
1.4 Organigramme des projets CISMeF 23
1.5 Extrait du périodique décrit par la notice CISMeF n°8916 25
1.6 Fonctionnement de CISMeF avant automatisation des taches documentaires (en 2002) 28

2.1 Le triangle aristotélicien 30
2.2 La triade sémiotique 30
2.3 Extrait du chapitre 4 de la CIM-10 36
2.4 Extrait de l'arborescence C du MeSH 39
2.5 Utilisation du MeSH par CISMeF 42
2.6 Structure de la terminologie CISMeF 44
2.7 Méthode de construction des dictionnaires MeSH 50
2.8 Automate reconnaissant les mots composés grammaticaux 52
2.9 Transducteur reconnaissant le mot clé <*adulte âge moyen*> 53
2.10 Transducteur reconnaissant la paire <*MALADIE/prévention et contrôle*> . . 54
2.11 Principe d'appariement par propagation syntaxique 57
2.12 Propagation syntaxique à partir d'un couple amorce de verbes 58
2.13 Descripteurs MeSH et leur traduction 59
2.14 Traductions extraites du corpus par calcul des cooccurrences 59
2.15 Repérage des couples amorces : influence du lexique bilingue utilisé 59
2.16 Traduction compositionnelle 60
2.17 Performance de la traduction directe en fonction du nombre d'occurrences des synonymes 62

3.1 Le Cycle de la Veille 72
3.2 Sélection des ressources à intégrer dans CISMeF 78
3.3 Copie écran de CVA : ajout d'une URL 80
3.4 Copie écran de CVA : veille hebdomadaire du 07/07/05. 81

4.1 Méthode de calcul des scores attribués à chaque métaterme pour la classification 90
4.2 Catégorisation de la ressource n° 9982 93

5.1 Répartition des descripteurs attribués par A et B pour un même document . 109
5.2 Algorithme d'extraction du titre d'une ressource 116
5.3 Sélection des termes en fonction de leur fréquence 120

5.4 Processus d'Indexation Automatique implémenté dans MAIF 127
5.5 Exemple de relations hiérarchiques 131
5.6 Approche k-NN : Algorithme permettant d'obtenir les candidats à l'indexation et leur score 134
5.7 Résultat de l'analyse de surface effectuée par MAIF-TAL 136
5.8 Formulation alternative de la phrase 5 et résultat de l'analyse de surface . . . 136
5.9 Liste des descripteurs MeSH attribués manuellement pour l'extrait de la ressource 8916 141
5.10 Courbes F-measure en fonction du rang pour MAIF 150
5.11 Courbe présentant la F-measure en fonction du rang pour les systèmes d'indexation MeSH francophones. 152
5.12 Courbe présentant la F-measure en fonction du rang pour MAIF-TAL et MeSH-Mapper. 155
5.13 Courbe présentant la F-measure en fonction du rang pour MTI strict et MAIF-TAL. 156
5.14 Processus d'Indexation Automatique CIM-10 160

6.1 Fonctionnement de CISMeF après automatisation des taches documentaires . 164

A.1 Ecran d'accueil de l'outil bibliométrique de catégorisation 183

Liste des tableaux

1.1 Description des corpus de travail et du corpus CISMeF 27

2.1 Exemples de relations modélisées par les Fonctions Lexicales 33
2.2 Correspondance entre les besoins et les terminologies 35
2.3 Les types de terminologies et leurs caractéristiques 35
2.4 Répartition des mots clés MeSH en fonction du nombre de lemmes 41
2.5 Informations lexicographiques contenues dans le dictionnaire MeSH 49
2.6 Extrait du MeSH 2004 56
2.7 Nombre de traductions directes extraites 63
2.8 Précision de la traduction des composants par corpus 66
2.9 Évaluation de 100 traductions compositionnelles par corpus 66

4.1 Liens sémantiques utilisés pour la classification de la ressource n°9982 91
4.2 Nombre de spécialités à extraire par ressource 92
4.3 Pertinence de la catégorisation 93
4.4 Typologie des erreurs de catégorisation observées 94

5.1 Représentation d'une phrase à l'aide de différentes unités descriptives 102
5.2 Langage d'indexation utilisé en fonction du contexte de Recherche d'Information 104
5.3 Deux listes de descripteurs MeSH attribués à un même document 110
5.4 Distribution des termes d'indexation pour une ressource donnée 112
5.5 Extraction automatique de titre pour 339 ressources (URLs accédées le 21/11/2004) 115
5.6 Extraction automatique de titre pour 339 ressources (URLs accédées le 23/06/2005) 117
5.7 Les différents schémas tf*idf 121
5.8 Comparaison des systèmes d'indexation MeSH 124
5.9 Traduction de l'analyse en terme MeSH. 137
5.10 Modifications occasionnées par la formulation alternative de la phrase 5 . . . 137
5.11 Appariement des qualificatifs isolés dans les phrases 3 et 5. 138
5.12 Appariement du qualificatif isolé *prévention et contrôle*. 138
5.13 Résumé final des descripteurs retenus : fréquences et scores 139
5.14 Trois plus proches voisins sélectionnés pour la ressource n° 9816 140
5.15 Obtention d'une liste de candidats pour la ressource n° 9816 140
5.16 Calcul du seuil adaptatif 142
5.17 Indexation Manuelle et Automatique pour la ressource n°115. 144
5.18 Comparaison des scores fondés sur la normalisation $tf * idf$ et la fréquence . 144
5.19 Comparaison des scores fondés sur la normalisation $tf * idf$ et la fréquence (avec coefficients de sélection) 145

5.20 Comparaison des scores fondés sur la normalisation $tf * idf$ et la fréquence (avec descripteurs obligatoires) 145
5.21 Comparaison des scores fondés sur la normalisation $tf * idf$ et la fréquence (avec descripteurs obligatoires et coefficients de sélection) 145
5.22 Comparaison des scores fondés sur la normalisation $tf * idf$ et la fréquence . 146
5.23 Comparaison des valeurs de k pour la recherche des k-plus proches voisins. . 148
5.24 Comparaison des mesures de similarité pour la recherche des k-plus proches voisins. 148
5.25 Influence du corpus d'apprentissage pour la recherche des k-plus proches voisins. 149
5.26 Précision et rappel de MAIF 149
5.27 Précision et rappel des systèmes francophones 151
5.28 Exemple d'indexation automatique par *mots clés* proposée par chaque système 153
5.29 Exemple d'indexation automatique par *paires* proposée par MAIF 154
5.30 Performance des systèmes MAIF et MTI à rang fixes et seuil adaptatif . . . 156
5.31 Correspondance CCAM/MeSH 160

B.1 Comparaison des scores fondés sur la normalisation $tf * idf$ et la fréquence . 185
B.2 Comparaison des scores fondés sur la normalisation $tf * idf$ et la fréquence (avec coefficients de sélection) 185
B.3 Comparaison des scores fondés sur la normalisation $tf * idf$ et la fréquence (avec descripteurs obligatoires) 186
B.4 Comparaison des scores fondés sur la normalisation $tf * idf$ et la fréquence (avec descripteurs obligatoires et coefficients de sélection) 186

Annexe A

Ecran d'accueil de l'outil bibliométrique

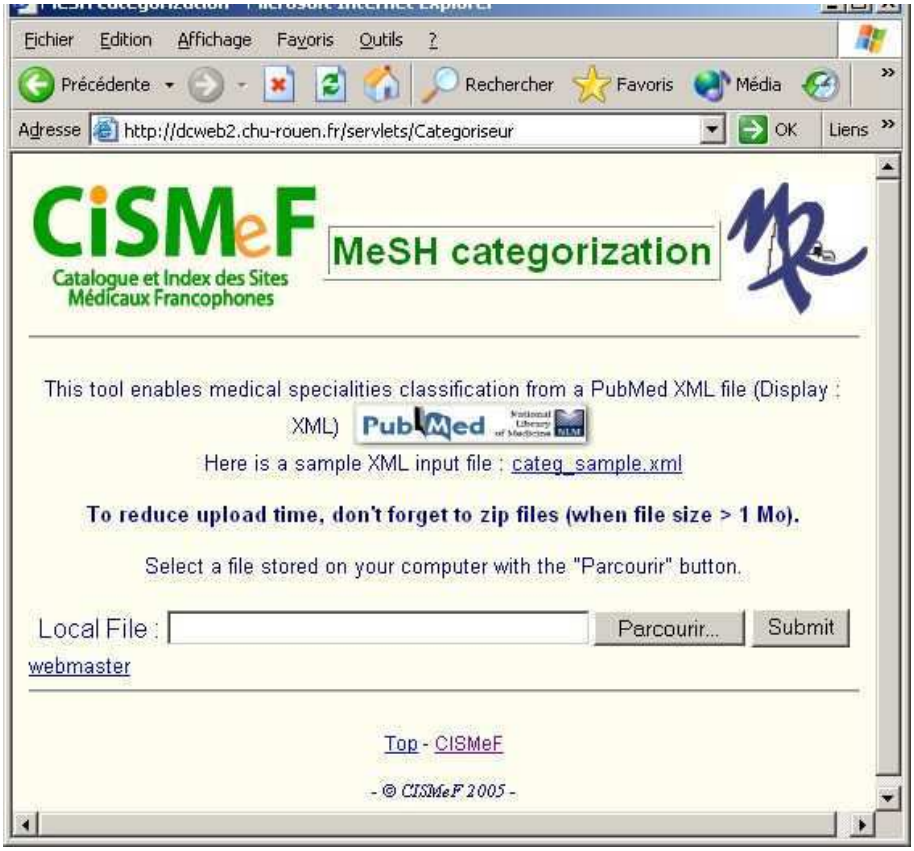

FIG. A.1 – Ecran d'accueil de l'outil bibliométrique de catégorisation- disponible sur l'intranet du CHU de Rouen

Annexe B

Résultats de l'indexation des Sommaires sur le corpus « diabète »

Rang	$tf * idf$	**Fréquence**
	Précision - Rappel	Précision - Rappel
1	45,45 - 10,49	50,91 - 12,05
4	33,72 - 20,28	31,98 - 19,88
10	26,77 - 23,84	24,19 - 21,97
20	29,23 - 24,15	28,85 - 22,69
50	14,50 - 24,00	18,00 - 26,00
Seuil	**38,07 - 13,11**	**29,91- 6,44**
	(S=4)	**(S=2)**

Tab. B.1 – Comparaison des scores fondés sur la normalisation $tf * idf$ et la fréquence. (Indexation à l'aide de paires mot clé/qualificatif)

Rang	$tf * idf + cs(t)$	**Fréquence** $+cs(t)$
	Précision - Rappel	Précision - Rappel
1	49,09 - 12,07	54,55 - 13,78
4	36,05 - 22,51	35,47 - 21,56
10	26,45 - 23,87	25,81 - 22,90
20	26,62 - 23,92	28,08 - 22,85
50	15,00 - 22,25	28,08 - 22,85
Seuil	**47,61- 15,11**	**12,26- 1,56**
	(S=2)	**(S=1)**

Tab. B.2 – Comparaison des scores fondés sur la normalisation $tf * idf$ et la fréquence, en utilisant les coefficients de sélection. (Indexation à l'aide de paires mot clé/qualificatif)

Rang	$tf * idf$+ **DO**	**Fréquence + DO**
	Précision - Rappel	Précision - Rappel
1	49,09 - 10,65	54,55 - 12,22
4	35,47 - 28,58	33,72 - 20,19
10	27,74 - 24,23	25,16 - 22,39
20	30,38 - 25,05	30,00 - 23,62
50	16,00 - 26,25	19,50 - 28,00
Seuil	**39,49- 13,33**	**31,33 -6,67**
	(S=4)	**(S=2)**

TAB. B.3 – Comparaison des scores fondés sur la normalisation $tf * idf$ et la fréquence, en considérant que les descripteurs obligatoires (DO) sont correctement extraits par le système. (Indexation à l'aide de paires mot clé/qualificatif)

Rang	$tf * idf$+ **DO** +$cs(t)$	**Fréquence + DO** +$cs(t)$
	Précision - Rappel	Précision - Rappel
1	52,73 - 12,24	58,18 - 13,95
4	37,79 - 22,81	37,21 - 21,86
10	27,74 - 24,39	27,10 - 23,42
20	30,77 - 24,77	29,23 - 23,77
50	16,50 - 24,75	21,00 - 31,00
Seuil	**49,04- 15,33**	**13,70 - 1,79**
	(S=2)	**(S=1)**

TAB. B.4 – Comparaison des scores fondés sur la normalisation $tf * idf$ et la fréquence, en utilisant les coefficients de sélection et en considérant que les descripteurs obligatoires (DO) sont correctement extraits par le système. (Indexation à l'aide de paires mot clé/qualificatif)

Automatisation des tâches documentaires dans un catalogue de santé en ligne

A. Névéol - Thèse d'Informatique de l'INSA de Rouen

Résumé :

La Recherche d'Information a pour objectif de permettre aux utilisateurs d'accéder rapidement et efficacement au contenu d'une collection de document. Dans le domaine de la santé, le nombre de ressources électroniques disponibles augmente de manière exponentielle, et la nécessité de disposer de solutions automatiques se fait sentir à plusieurs étapes de la chaîne d'information. Les documents, en particulier les textes, doivent être sélectionnés selon des critères de qualité pour être inclus dans des catalogues ; ils doivent également être décrits à l'aide de mots clés et catégorisés en spécialités médicales afin de faciliter les recherches effectuées dans les catalogues. Ces taches constituent un défi pour le Traitement Automatique de la Langue Naturelle car elles impliquent une « compréhension » du contenu des documents par un système automatique. Ce travail de thèse engage une réflexion sur la répartition des tâches documentaires entre l'homme et la machine dans le cadre particulier du Catalogue et Index des Sites Médicaux Francophones (CISMeF). A ce titre, il aborde l'automatisation des tâches documentaires dans le catalogue de santé en ligne CISMeF. Cette thèse apporte une contribution au développement de ressources linguistiques en français pour le domaine de la santé, et présente des systèmes de veille documentaire et de description automatiques de ressources de santé. Sur ce dernier point, l'accent a été mis sur l'indexation à l'aide de paires de descripteurs issues du thésaurus MeSH.

Mots clés : Indexation automatique , terminologie médicale, traitement automatique de corpus

Abstract :

Information Retrieval aims at enabling users to access the content of documents quickly and efficiently. In the medical domain, an increasing number of resources are available in electronic format, and there is a growing need for automatic solutions at several levels. Documents, and in particular texts, need to be selected and quality assessed to appear in catalogues ; they need to be described with keywords and categorized within medical specialties to allow searches within the catalogues. These tasks are a challenge for Natural Language Processing, as they imply an « understanding » of the documents content by an automatic system. My PhD work has addressed this issue, and applied it to the automatisation of documentary tasks in a French online health catalogue, CISMeF. More specifically, this work has involved contributing to the enrichment of linguistic resources available in French for the medical domain, and developping systems for document watch and resource description. In this particular area, my focus was on MeSH automatic indexing with Main Heading/Sub Heading pairs.

Keywords : Automatic indexing, medical terminology, corpus analysis