ÉCOLE DOCTORALE SPMII
UNIVERSITÉ DE ROUEN LITIS EA4108

THÈSE DE DOCTORAT

Présentée pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ DE ROUEN

Spécialité
Informatique

Suzanne BENTO PEREIRA

Indexation multi-terminologique de concepts en santé

JURY

M	Pierre	ZWEIGENBAUM	Rapporteur
M	Alain	VENOT	Rapporteur
M	Antoine	BUEMI	Examinateur
M	Thierry	LECROQ	Examinateur
Mme	Elisabeth	SERROT	Responsable côté entreprise
M	Michel	JOUBERT	Directeur de thèse
M	Stefan J.	DARMONI	Directeur de thèse

À ma famille,

Remerciements

Je tiens à remercier ici toutes les personnes qui ont rendu possible la réalisation de cette thèse.

Tout d'abord mes encadrants qui ont formé un trio de choc (tel les trois mousquetaires Porthos, Athos et Aramis encadrant le petit d'Artagnan) :

Le professeur Stefan Darmoni incarnant le dynamisme et l'humour qui à base de coups a permis que cette thèse avance,

Le docteur Michel Joubert représentant la sagesse qui a posé les limites et a poussé à la réflexion,

Et le docteur Elisabeth Serrot pour ses analyses méticuleuses.

Puis les différentes équipes pour leur aide, leur soutien et leur amitié :

L'équipe CISMeF (Josette, Gaëtan, Catherine, Benoît, Saoussen, Taieb, Yvan et Badisse),

L'équipe scientifique du Vidal (Mathilde, Josiane, Sophie, Francine, Olivier, Michelle, Blandine, Nicolas, Ghislaine, Gismonde, Jean-Francois),

Ainsi que les rois de la numérisation et accessoirement de la relecture (Laurent, Thierry, Cedric, Remy, Ulrich et Joachim),

Et les personnes externes : Antoine Buemi, Max Silberztein, Philippe Massari, Paul Avillach, Marius Fieschi, Gaëlle Lortal et Lina Soualmia.

Je remercie également les laboratoires LERTIM et LITIS pour m'avoir accueillie,

Ainsi que la société Vidal et son directeur Vincent Bouvier pour son engagement dans ce projet.

Enfin bien sûr ma famille pour son soutient, la relecture de ma soeur Hélène et les sourires du nouveau membre de la famille la petite Liséa.

Résumé

La recherche d'information ainsi que l'aide à la décision nécessitent un accès rapide et efficace aux connaissances contenues dans une collection de documents de santé, ainsi qu'une bonne exploitation des connaissances médicales. L'indexation (description à l'aide de mots clés) permet de rendre ces connaissances accessibles et utilisables. Dans le domaine de la santé, le nombre de ressources électroniques disponibles augmente de manière exponentielle ainsi la nécessité de disposer de solutions automatiques pour faciliter l'accès aux connaissances ainsi que l'indexation est omniprésente. L'objectif de cette thèse a été de développer un outil d'aide à l'indexation automatique multi-terminologique, multi-document et multi-tâche nommé F-MTI (French Multi-terminology Indexer) capable de produire une proposition d'indexation pour les documents de santé. Cet outil a nécessité l'élaboration de méthodes de Traitement Automatique de la Langue Naturelle. Il a été appliqué à l'indexation documentaire dans le catalogue de santé en ligne CISMeF, à l'indexation des données thérapeutiques pour les médicaments et à l'indexation des diagnostics et des actes médicaux pour les dossiers médicaux électroniques.

Mots Clés : Indexation et rédaction du résumé/méthodes ; Stockage et recherche information/méthodes ; Dossiers médicaux ; SNOMED ; Medical Subject Headings ; Healthcare Common Procedure Coding System ; Classification internationale des maladies ; traitement langage naturel ; Vocabulaire contrôlé ; Terminologie ; Algorithme ; Étude évaluation.

Abstract

Information retrieval and decision support systems need fast and accurate access to the content of documents and efficient medical knowledge processing. Indexing (describing using keywords) enables access to knowledge and knowledge processing. In the medical domain, an increasing number of resources are available in electronic format, and there is a growing need for automatic solutions to facilitate knowledge access and indexing. The objectives of my PhD work are the implementation of an automatic multi-terminology multi-document and multi-task indexing help-system namely F-MTI (French Multi-terminology Indexer). It uses Natural Language processing methods to producte an indexing proposition for medical documents. We applied it to resources indexing in a French online health catalogue, namely CISMeF, to therapeutical data indexing for drug medication and to diagnosis and health procedures indexing for patient medical records.

Keywords : Abstracting and Indexing/methods ; Information Storage and Retrieval/methods ; medical records ; Systematised Nomenclature of Medicine ; Medical Subject Headings ; Healthcare Common Procedure Coding System ; International Classification of Diseases ; Natural Language Processing ; vocabulary, controlled ; Terminology ; Algorithms ; Evaluation studies

Table des matières

Introduction générale

0.1 Problématique

Les informations médicales sont nombreuses et très dispersées. Elles sont contenues dans les rapports, articles, livres... sous forme papier ou électronique. Ces informations à l'origine non structurées sont répertoriées, classées et stockées dans des bases de données sous une forme exploitable par un ordinateur (données structurées) dans le but de permettre leur consultation et utilisation.

Ces données permettent à un utilisateur (un professionnel de santé ou un patient) d'accéder aux connaissances contenues dans les bases documentaires et de rechercher des informations. Chaque document est décrit dans la base documentaire grâce à des informations sur sa forme et son contenu.

Ces données permettent aussi l'exploitation des connaissances par entre autres des outils d'aide à la décision qui permettent de conseiller les praticiens dans leur pratique quotidienne. Toutes ces connaissances peuvent alors être décrites dans une base de connaissance afin de permettre leur exploitation.

L'indexation permet de traduire des données textuelles non structurées en données structurées. Nous nous intéressons ici à l'indexation contrôlée, c'est-à-dire que la liste de tous les termes formant les données structurées est connue à l'avance et est stockée dans une terminologie.

Cette indexation est le plus souvent effectuée manuellement et prend beaucoup de temps. Des solutions peuvent venir aider l'indexeur dans sa tâche comme des outils facilitant la recherche de termes dans les terminologies d'indexation ou proposant une indexation automatique de documents que l'indexeur n'a plus qu'à vérifier et valider.

Dans notre projet de thèse, nous nous sommes intéressée à ce deuxième type d'outil. Nous nous sommes également intéressée aux moyens de faciliter l'accès aux connaissances contenues dans les bases documentaires.

0.2 En pratique

En pratique, trois applications ont attiré notre attention.

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 [1] recense plus de 18 millions d'articles scientifiques en langue anglaise indexés à l'aide de la terminologie MeSH (Medical Subject Headings). En Europe, plusieurs projets (par exemple : HON [2], Intute [3],...) et notament en France le projet CISMeF [4] ont vu le jour. Ce site répertorie et indexe les documents électroniques d'information institutionnelle de santé en langue francaise afin d'aider les professionnels de santé, les étudiants et les patients à rechercher une information de qualité en santé sur Internet. L'essentiel du travail de l'équipe CISMeF consiste en la maintenance et la mise à jour du catalogue ainsi que son amélioration et son évolution tant en termes de technologies utilisées que de rendement et de facilité d'utilisation pour l'utilisateur. Les indexeurs de l'équipe sont chargés d'indexer toute nouvelle ressource Web sélectionnée, à l'aide de la terminologie MeSH. Internet fournissant une masse de données titanesque en santé (de l'ordre de 7 millions de pages créées par jour tous domaines confondus), il est important de disposer d'outils d'indexation automatique et d'aide à l'indexation afin de faciliter et de rendre plus rapide ce travail.

Dans le domaine du médicament, de nombreux travaux en matière d'aide à la décision permettant de sécuriser les prescriptions existent. C'est le cas des banques de données Thériaque [5], BDSP [6] et notamment de la société Vidal qui diffuse des informations sur le médicament et produit des outils de sécurisation pour les prescriptions. Le travail des indexeurs de l'équipe Vidal consiste à indexer manuellement les Résumés Caractéristiques des Produits (RCP) contenant toutes les informations thérapeutiques pour les médicaments (indications, contre-indications, effets indésirables, etc...) à l'aide des terminologies Vidal dont le TUV (Terminologie Unifiée Vidal). La masse des RCP à traiter provenant de l'AFFSAPS est importante (de l'ordre de 600 à 1200 par mois). Il serait donc nécessaire de disposer d'outils facilitant leur indexation afin de maintenir une base de qualité avec des données à jour.

Dans le domaine de la santé, de nombreux travaux, et notamment ceux du laboratoire LERTIM, s'intéressent à l'élaboration de systèmes d'information hospitaliers performants. Le dossier médical informatisé est l'une des composantes du système d'information en réseaux de l'hôpital. Ce dossier permet de recueillir pour chaque patient toutes les informations qui ont trait à son état de santé et à son parcours de soin. Le recueil des données concernant l'activité de l'hôpital (les pathologies traitées par exemple) et son mode de fonctionnement (exemple : mode de prise en charge) permet de définir les financements nécessaires à l'hôpital. Les données recueillies sont indexées à l'aide des terminologies CIM10 [7] (pour les diagnostics) et CCAM [8] (pour

1. Accessible via `http://www.ncbi.nlm.nih.gov/pubmed/`
2. Accessible via `http://www.hon.ch/index f.html`
3. Accessible *via* `http://www.intute.ac.uk/healthandlifesciences/medicine/`
4. Catalogue et Index des Sites Médicaux Francophones accessible *via* `http://www.chu-rouen.fr/cismef/`
5. Accessible *via* `http://www.theriaque.org/`
6. Accessible *via* `http://www.bdsp.ehesp.fr/`
7. Classification Internationale statistique des Maladies et des problèmes de santé connexes 10ème révision
8. Classification Commune des Actes Médicaux

les actes). De plus l'utilisation d'une nouvelle terminologie, la SNOMED 3.5 [9], devrait être mise en place prochainement. Cette indexation est fastidieuse pour les médecins et le temps nécessaire n'est dès lors pas consacré à traiter le patient. Une indexation descriptive de l'ensemble des informations du dossier des patients pourrait aussi permettre un meilleur suivi des soins. Les masses d'informations à traiter sont très importantes. Pour exemple, l'hôpital de Rouen répertorie 1 080 384 patients et 182 808 comptes rendus d'hospitalisation en 2005. Il serait donc utile pour les médecins de disposer d'outils d'aide à l'indexation pour l'indexation de leurs dossiers médicaux.

0.3 Objectifs

L'objectif que nous nous sommes fixée est de créer un outil générique destiné à l'indexation automatique de documents. Celui-ci a été developpé afin de permettre l'indexation des dossiers patients en CIM10, CCAM et SNOMED 3.5, des sites médicaux en MeSH et des RCP en TUV.

Ce travail explore différentes approches pour analyser le contenu des documents, et pour les exploiter. Il s'agit principalement de méthodes de Traitement Automatique du Langage Naturel (TALN).

Nous nous sommes également intéressée aux moyens de faciliter l'accès aux connaissances contenues dans les bases documentaires sur Internet et dans les dossiers patients.

0.4 Organisation du mémoire

La rédaction des différents chapitres suit le raisonnement qui a été entrepris dans la réalisation de cette thèse. Nous avons adopté une démarche séquentielle ou ascendante (« bottom-up ») qui consiste à partir de problématiques concrètes à aller vers la résolution des problèmes scientifiques adjacents. Ainsi, pour chaque tâche d'indexation, nous avons effectué une analyse du problème. À partir de ces analyses, nous avons proposé des méthodes qui ont été expérimentées et évaluées. Ces évaluations ont permis de définir les limites rencontrées, de proposer des applications possibles de l'outil et d'aborder les perspectives envisageables.

Dans le premier chapitre, nous exposons le contexte des travaux effectués : contexte administratif et scientifique. Ce chapitre permet de rendre compte des besoins exprimés par les équipes CISMeF, Vidal et LERTIM qui ont mené à l'élaboration du sujet de cette thèse.

Le deuxième chapitre aborde l'analyse de l'état de l'art relatif à notre sujet qui a permis de définir les domaines de recherche abordés : la recherche d'information électronique et notamment sur l'Internet, la construction de bases de connaissances et les systèmes d'aide à la décision. Les différentes tâches d'indexation mises en

9. Nomenclature Systématique de Médecine humaine et vétérinaire version 3.5

évidence dans le premier chapitre sont aussi présentées : la terminologie MeSH et la politique d'indexation des ressources en MeSH au sein de l'équipe CISMeF, le codage médico-économique pour les dossiers patients et les terminologies associées ainsi que l'indexation des RCP à l'aide des terminologies Vidal. Nous présentons aussi les travaux existant en matière d'aide à l'indexation automatique. En fin de chapitre, les axes d'améliorations possibles ainsi que notre contribution dans le domaine sont explicités.

Le troisième chapitre présente le fonctionnement de l'outil F-MTI (French Multi-Terminology Indexer), un outil d'indexation multi-terminologique, multi-document et multi-tâche générique en mesure de reproduire automatiquement les tâches d'indexation décrites réalisées habituellement à la main. Nous présentons aussi les différentes méthodes élaborées.

Dans le chapitre 4, nous présentons les différentes évaluations menées. Ces évaluations portent sur les performances de F-MTI «en situation». L'indexation produite à l'aide de la CIM10, de la CCAM et de la SNOMED pour les comptes rendus d'hospitalisation y est évaluée. Nous présentons aussi les évaluations concernant l'indexation des ressources Web à l'aide du MeSH et des RCP à l'aide du TUV. F-MTI a aussi été comparé à d'autres outils d'indexation automatique.

Un cinquième chapitre permet d'aborder les différentes mises en application envisagées.

Le sixième chapitre résume et permet de discuter les principaux résultats ainsi que d'évoquer les différentes perspectives.

Le chapitre 7 présente notre contribution en matière d'accès aux connaissances pour les professionnels de santé, les patients et les étudiants ayant besoin dans leur quotidien d'informations de santé, que ce soit dans le cadre de l'apprentissage de nouvelles connaissances, d'aide à la décision ou de suivi d'une prise en charge. Des méthodes prenant en compte le contexte et permettant des accès simplifiés à la bonne information, au bon moment et pour la bonne personne sont présentées.

Enfin, 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.

Première partie

Contexte et état de l'art

Chapitre 1

Contexte et besoins

1.1 Introduction

Dans ce chapitre, nous exposons le contexte des travaux effectués. Dans un premier temps, nous décrivons le contexte administratif avec une présentation des différentes équipes impliquées. Nous rendons compte du contexte scientifique par une brève description des travaux de chacune des équipes. Enfin, nous faisons la synthèse des différents besoins exprimés qui ont mené à l'élaboration du sujet de cette thèse.

1.2 Contexte administratif

Les travaux présentés dans ce mémoire sont le résultat de ma thèse d'informatique débutée en mars 2006 [1]. Cette thèse est réalisée dans le cadre d'une bourse CIFRE [2]. Cette thèse a été conduite par trois partenaires : la société Vidal, le laboratoire LERTIM et le laboratoire LITIS.

Le LITIS [3] est le Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes. Il est l'unité de recherche dans le domaine des Sciences et Technologies de l'Information et de la Communication (STIC) de Haute-Normandie. Depuis mars 2006, le LITIS est reconnu en tant qu'Équipe d'Accueil EA4108. Le LITIS est pluridisciplinaire associant praticiens et théoriciens à la jonction de l'informatique, de la reconnaissance des formes, du traitement du signal et des images, de la médecine et des mathématiques.

La société Vidal [4] est une filiale de CMP Medica (Group United Business Me-

1. Les travaux de thèse ont démarré officieusement en septembre 2005, après six mois de stage de master 2 en Informatique médicale au sein de l'équipe CISMeF

2. Les conventions CIFRE (Conventions Industrielles de Formation par la Recherche) associent, autour d'un projet de recherche, trois partenaires : une entreprise, un jeune diplômé et un laboratoire. L'Association nationale de la recherche technique (ANRT) est responsable de la gestion et de l'animation des conventions CIFRE (`http://www.anrt.asso.fr/fr/espace cifre/accueil.jsp?index=2`).

3. Site Internet du laboratoire : `http://www.litislab.eu/`

4. Site Internet de la société : `http://www.vidal.fr/index.htm`

dia [5]), leader international de l'information professionnelle aux entreprises dans les secteurs, entre autres, de la santé, de la technologie et des média. Elle diffuse des informations sur le médicament aux professionnels de santé, aux industries pharmaceutiques et au grand public.

Enfin, le LERTIM [6], le Laboratoire d'Enseignement et de Recherche sur le Traitement de l'Information Médicale, a été labellisé Équipe d'Accueil EA3283. La recherche autour de l'informatique médicale se développe dans ce laboratoire depuis 1990.

Les travaux de thèse ont été effectués au sein des équipes suivantes :

- l'équipe CISMeF dirigée par le professeur Stéfan Darmoni et le conservateur de la bibliothèque médicale Benoît Thirion au Centre Hospitalo-Universitaire de Rouen [7]. L'équipe CISMeF appartient à l'axe «Traitement de l'Information en Biologie et Santé» (TIBS) du laboratoire LITIS. L'équipe est constituée d'un professeur, d'un conservateur, de trois documentalistes spécialistes de l'indexation médicale (les indexeurs), de deux ingénieurs de recherche et de trois doctorants (voir la figure 6.1 pour avoir un apercu du rôle de chacun).
- l'équipe scientifique du Vidal dirigée par Jean-Francois Forget. Les travaux ont été co-encadrés par Elisabeth Serrot responsable de l'équipe chargée de la création et de la maintenance des données thérapeutiques, ainsi que de l'indexation des RCP (Résumé des Caractéristiques du Produit [8]). Elle est constituée de six pharmaciens spécialistes de l'indexation (les indexeurs), d'une pharmacienne chargée des thesaurus et d'une linguiste spécialisée dans le Traitement Automatique du Langage.
- l'équipe du LERTIM dirigée par le Professeur Marius Fieschi au Centre Hospitalo-Universitaire de La Timone à Marseille. Les travaux ont été co-encadrés par Michel Joubert. L'équipe est constituée de trois professeurs, neuf maîtres de conférences, cinq praticiens et assistants hospitaliers, deux intervenants du service de santé des armées en Épidémiologie et Santé Publique, deux ingénieurs et techniciens et sept doctorants.

1.3 Contexte scientifique

1.3.1 Travaux de l'équipe CISMeF

1.3.1.1 Domaine de Recherche de l'équipe CISMeF

De nombreux travaux ont été entrepris par l'équipe CISMeF dans le domaine de la recherche d'information en santé et plus particulièrement dans la recherche documentaire afin de guider les utilisateurs dans leur quête d'informations médicales.

5. Site Internet du groupe : `http://www.cmpmedica.com/`
6. Site Internet du laboratoire : `http://cybertim.timone.univ-mrs.fr`
7. Le site Internet du CHU de Rouen : `http://www.chu-rouen.fr/`
8. Les RCP comportent les données cliniques des spécialités pharmaceutiques ayant fait l'objet d'une AMM (Autorisation de Mise sur le Marché) attribuée par l'Afssaps (Agence francaise de sécurité sanitaire des produits de santé).

Internet connaît depuis le début des années 90 un grand essor mondial avec une croissance soutenue de l'ordre de 7 millions de pages par jour et l'ensemble dépasse les 10 milliards. Pour les utilisateurs en quête d'information médicale, il est devenu très difficile de rechercher des informations sur le Web, compte tenu de la quantité énorme de sites et de documents médicaux disponibles. Chacun peut publier des informations médicales sur le Web, aussi il est devenu difficile de retrouver de l'information de qualité et correctement recensée.

1.3.1.2 Les travaux de l'équipe CISMeF

L'équipe CISMeF a développé le site du Catalogue et Index des Sites Médicaux Francophone [9] (CISMeF) en février 1995 (voir figure 1.1). Il répertorie et indexe les documents électroniques d'information institutionnelle de santé en langue francaise afin d'aider les professionnels de santé, les étudiants et les patients à rechercher une information de qualité en santé sur Internet. Quatre raisons ont motivé l'élaboration du catalogue : la profusion des informations toujours grandissante en santé sur le Web, la nécessité d'accéder à des informations fiables et de qualité en médecine, l'inexistence de moteur de recherche spécialisés et efficaces, et la difficulté de distinguer les informations destinées aux professionnels de celles destinées aux patients.

FIGURE 1.1 – Le site CISMeF

Le site CISMeF est un site assez populaire puisque le nombre d'utilisateurs uniques se connectant à CISMeF est d'environ 27 000 par jour ouvré (dont 37,8% en France et 38,4% en Algérie).

Le catalogue CISMeF est aussi un important fond documentaire qui contient plus de 47 000 ressources [10] avec une grande diversité de formes (recommandations, cours, sites d'associations de patients, forums etc...) et de formats (documents PDF, sites Web, documents PowerPoint etc...).

L'essentiel du travail de l'équipe consiste en la maintenance et la mise à jour du

9. L'accès au catalogue se fait *via* les URL suivantes : `http://www.chu-rouen.fr/cismef.fr` ou `http://www.cismef.org`.
10. données de mai 2008.

catalogue ainsi que son amélioration et son évolution tant en termes de technologies utilisées que de recensement de nouvelles ressources et de facilité d'utilisation pour l'utilisateur.

L'ajout d'une nouvelle ressource [11] au catalogue s'effectue en quatre étapes :

1. Recensement des ressources potentielles par une veille stratégique quotidienne (via des annuaires multidisciplinaires francophones, des sites majeurs et bien d'autres)
2. Sélection des ressources selon des critères de qualité fondés sur le NetScoring [12] (critères de qualité de l'information de santé sur Internet [Darmoni98, Darmoni03a]). Cette sélection est faite de manière rigoureuse par des professionnels de l'information appuyés par des réseaux d'experts
3. Chaque ressource est décrite dans une notice (voir un exemple de notice courte [13] figure 1.2) afin d'être plus facilement retrouvée par le moteur de recherche CISMeF. Un ensemble de métadonnées est associé à la ressource par les in-

FIGURE 1.2 – Exemple d'une notice courte

dexeurs [14] :

- Caractéristiques externes de la ressource : le titre, les auteurs, le type de ressource, la cible, la langue, la date, la source (pays, site éditeur), des informations sur la qualité du document [Darmoni98], l'URL, le format, le type d'accès et la date de consultation.
- Informations sur le contenu du document : un résumé succinct élaboré par les indexeurs, et des mots clés décrivant les notions principales abordées dans le document (mots clés généraux et substances issus de la terminologie CISMeF [15], voir section 2.4.1.2 pour une description de la terminologie CISMeF

11. Les sites web ou documents numériques sont des documents particuliers que nous appellerons ressources.
12. Voir `http://www.churouen.fr/netscoring`
13. Seules les principales données pour chaque ressource sont présentées, il existe aussi dans CISMeF une notice longue avec toutes les caractéristiques disponibles.
14. L'indexeur pratique la description et l'indexation de ressources.
15. La terminologie CISMeF contient l'ensemble des mots-clés pouvant être assignés à une ressource.

et des méthodes d'indexation).
L'activité qui consiste à assigner au document des mots clés s'appelle l'«indexation». Il existe différents niveaux d'indexation. Le choix de la méthode d'indexation est opéré par l'indexeur à l'étape 2 lors de la sélection des ressources. Le premier niveau est une indexation purement manuelle (faite par des humains à la main) pour les ressources de priorité haute comme les recommandations qui ont besoin d'être indexées rapidement pour être diffusées rapidement auprès des médecins. L'indexation de niveau 2 est une indexation supervisée qui consiste en une indexation automatique effectuée par un programme informatique sur le titre de la ressource. Les indexeurs sont ensuite chargés de valider et modifier à la main si nécessaire cette indexation. Elle est destinée aux ressources de qualité mais moins urgentes que celles du premier niveau. Enfin, l'indexation de niveau 3 est une indexation purement automatique (sans validation humaine *a posteriori*) sur le titre pour les ressources de priorité faible dont la qualité et l'utilité ne nécessitent pas une indexation précise ou dont le thème est déjà abondamment traité dans CISMeF. Le catalogue contient 18 807 ressources indexées manuellement, 7 317 ressources supervisées et 14 752 resources indexées automatiquement.

Ces métadonnées proviennent de plusieurs référentiels dont 11 champs (parmi les 15) du Dublin Core [Dekkers03, Thirion04] et certains champs du IEEE 1484 LOM (Learning Object Metadata avec sa version francaise LOM-FR [16]). Les métadonnées HIDDEL [17] ont aussi été introduites dans le cadre du projet européen MedCircle [Mayer03].

4. L'ajout définitif au catalogue par la mise en ligne de la notice de la ressource

En moyenne, une cinquantaine de ressources par semaine sont indexées manuellement et ajoutées au catalogue.

Depuis l'année 2000, Doc'CISMeF, l'outil de recherche intégré au site CISMeF, donne un accès précis et rapide aux ressources. Son interface permet à l'aide de requêtes saisies par l'utilisateur d'obtenir une série de documents susceptibles de contenir l'information recherchée par celui-ci (c'est ce qu'on appelle la recherche documentaire). L'utilisateur n'a plus qu'à sélectionner la ressource qu'il désire et rechercher lui-même l'information qui l'intéresse à l'intérieur. Ces ressources sont présentées par ordre chronologique, les ressources indexées manuellement sont présentées en premier. Puis les ressources supervisées et enfin celles indexées automatiquement les suivent.

Différents modes de recherche d'information (accessibles depuis la page d'accueil de CISMeF voir figure 1.1) sont possibles :

- La recherche simple permet à l'utilisateur peu expérimenté de saisir une requête sous forme d'expressions libres en francais ou en anglais. Le système est alors chargé d'exprimer cette requête sous forme de mots clés (voir figure 1.3) afin

16. Pour plus d'informations sur les métadonnées LOM voir `http://www.lomfr.org`
17. Pour plus d'informations sur les métadonnées HIDDEL voir `http://www.medcircle.org`

de retourner les ressources qui ont été indexées à l'aide de ces mots-clés.

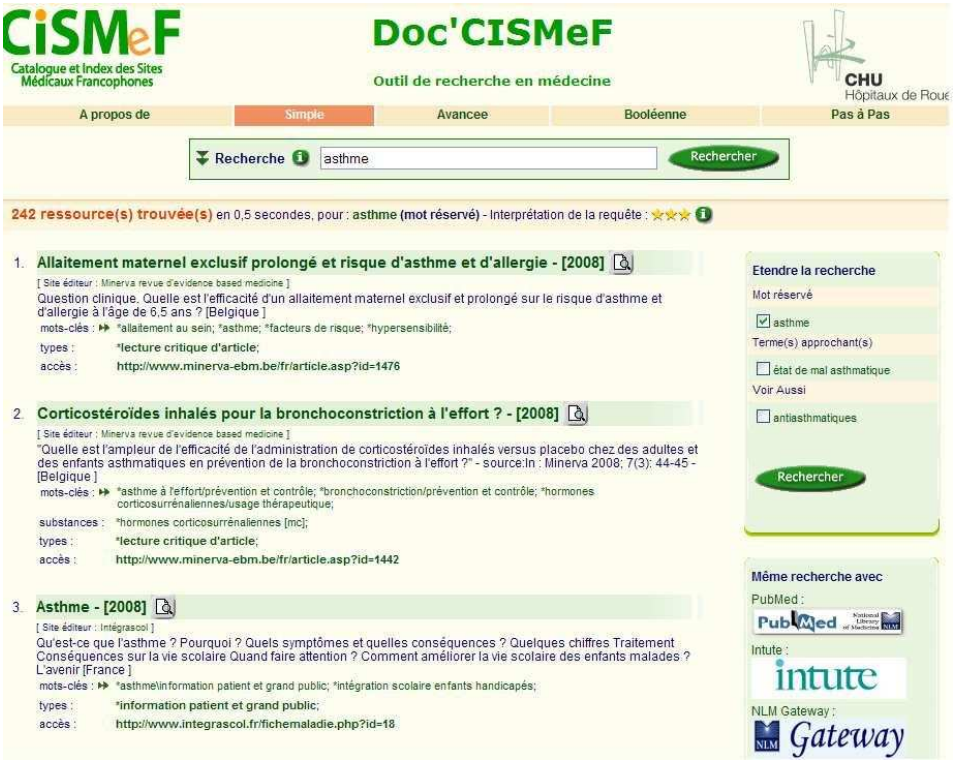

FIGURE 1.3 – Exemple de recherche simple avec Doc'CISMeF

- La recherche avancée permet des recherches plus poussées facilitées par l'utilisation d'un formulaire contenant des listes déroulantes et permettant de combiner plusieurs champs (mots clés, thème, type de ressources, année, etc...) avec des opérateurs booléens (ET, OU, SAUF).
- La recherche booléenne pour les utilisateurs expérimentés s'effectue à l'aide d'un langage de requêtes particulier utilisant des opérateurs booléens et des caractères spéciaux.
- Une recherche *via* le serveur de terminologie [18] permet de rechercher des informations à partir d'un mot clé. La recherche sur le mot clé peut être affinée (grâce à l'association de qualificatifs) avant d'être lancée sur CISMeF pour retrouver des documents en francais ou sur MEDLINE [19] pour retrouver des documents en anglais [20] [Thirion07].

Par ailleurs, CISMeF donne accès à d'autres sites spécialisés dans la recherche de documents dans le domaine de la santé. L'accès à ces sites est donné de manière contextuelle dans CISMeF (voir l'onglet «même recherche avec» figure 1.3). Par exemple,

18. Le serveur de terminologie est accessible *via* l'url : `http://www.churouen.fr/terminologiecismef/`.

19. Base de données bibliographique en anglais accessible *via* `http://www.ncbi.nlm.nih.gov/pubmed/`.

20. CISMeF est conforme aux standards W3C (`http://www.w3c.org`) (XML qui permet une interopérabilité avec d'autres moteurs de recherche, OWL, RDF).

si l'utilisateur recherche des recommandations (le système a détecté le concept «recommandations» dans la requête tapée par l'utilisateur) alors, lui est proposé à côté des ressources CISMeF, un accès à d'autres sites de référence pour les recommandations afin d'étendre sa recherche (NGC, OMNI, etc...). Le même principe est utilisé pour l'accès aux sites dédiés aux étudiants, aux patients ou aux moteurs de recherche généralistes. Plus de 70 sites en anglais et en francais connus dans le domaine pour leur fiabilité ont été choisis et référencés et les requêtes correspondantes élaborées. En effet, chaque site a des modalités d'interrogation différentes (mode de recherche, langage de requête particulier) que l'équipe CISMeF a exploité au maximum afin de reformuler automatiquement, dans le moteur de recherche ciblé, la requête de départ de l'utilisateur dans CISMeF. Parmi ces sites se trouve notamment le moteur de recherche Google. Compte-tenu de la difficulté de retrouver des documents de qualité sur ce site, l'établissement d'un partenariat Google/CISMeF, a permis de restreindre l'accès de Google à une liste de sites de qualité sélectionnés par l'équipe CISMeF pour le domaine médical [21] et pour les médicaments [22].

1.3.1.3 Les différents projets

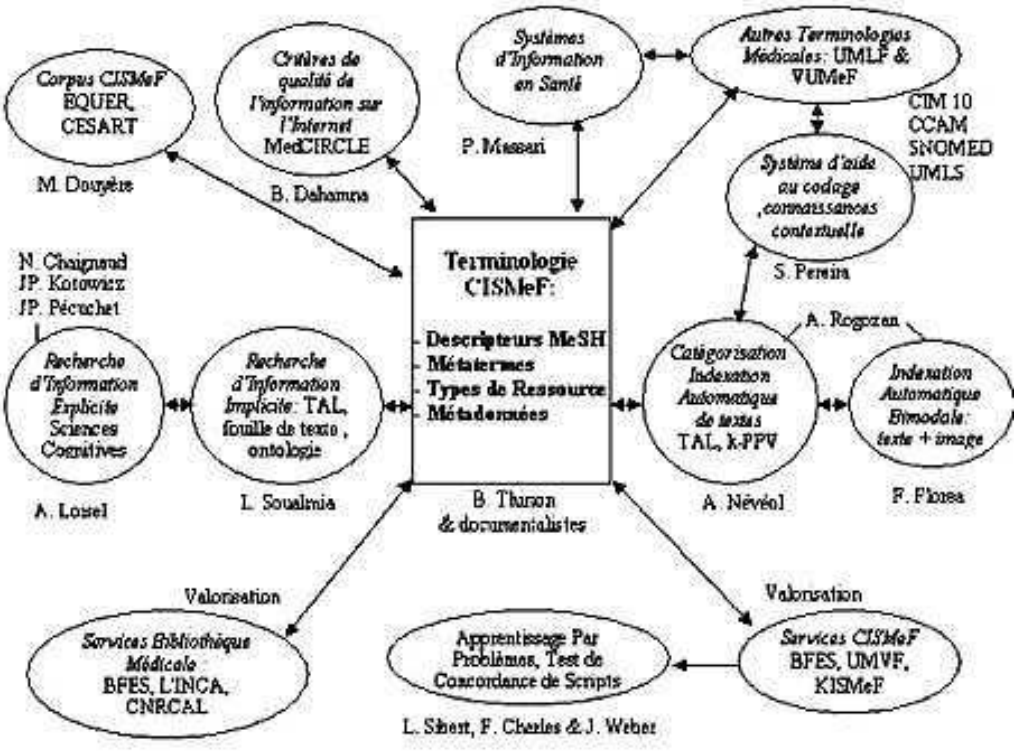

FIGURE 1.4 – Les différents projets de l'équipe CISMeF

De nombreux travaux [23] visant à améliorer la recherche d'information dans CIS-

21. http://www.google.com/custom?hl=fr&lr=lang fr&client=google-coop-np&cof=AH.
22. http://www.google.com/coop/cse?cx=015430007758165987576\%3Ab3cmgan4uas&hl=fr.
23. Les publications engendrées par ces travaux sont disponibles *via* http://www.chu-rouen.

MeF ont été menés afin de permettre une recherche d'information plus intelligente et plus efficace (voir figure 1.4 pour une revue des projets).

Au centre des activités de l'équipe CISMeF se trouve la terminologie CISMeF (que nous décrivons à la section 2.4.1.2). C'est en effet, sur cette terminologie que reposent les principaux travaux pour la recherche d'information dans le moteur CISMeF. Elle permet de décrire les ressources (indexation des ressources à l'aide de mots-clés appartenant à la terminologie), la requête d'un utilisateur (traduction de la requête à l'aide de mots-clés appartenant à la terminologie) et ainsi, de faire correspondre une requête à un ensemble de documents du catalogue (cette notion est davantage décrite dans 2.2.1.0.2). L'équipe CISMeF est donc très impliquée dans des travaux touchant à l'enrichissement des terminologies médicales francaises notamment pour le MeSH (terminologie médicale généraliste qui constitue les bases de la terminologie CISMeF) et pour d'autres terminologies telles que la CIM10, la SNOMED et la CCAM. L'équipe CISMeF fait évoluer jour après jour sa terminologie [Douyère04], elle participe aussi avec d'autres équipes à des projets collaboratifs. Ainsi, l'équipe CISMeF a participé, en collaboration notamment avec l'équipe du LERTIM et la société Vidal, aux projets UMLF [Zweigenbaum03] et VUMeF [Darmoni03b] de 2003 à 2007, dont le but était d'enrichir les terminologies médicales francaises dans l'UMLS [24]. CISMeF a aussi collaboré avec la société Memodata (PME spécialiste des dictionnaires) en vue d'enrichir le catalogue de nombreuses définitions et traductions en plusieurs langues. D'autres travaux ont aussi été menés pour mieux comprendre le langage médical courant utilisé par les usagers non spécialistes du domaine dans l'élaboration de leurs requêtes [Darmoni02].

Des travaux pour faciliter la recherche des utilisateurs ont aussi été effectués : un dialogue homme-machine actuellement à l'étude avec le projet Cogni'CISMeF [Loisel07] et une recherche d'information implicite avec le système KnowQuE (Knowledge-based Query Expansion) [Soualmia03, Soualmia04].

L'indexation d'une ressource à l'aide de mots clés est à la base de la recherche d'information dans le catalogue. C'est l'une des tâches les plus importantes et malheureusement la plus coûteuse en temps lors de l'ajout d'une nouvelle ressource au catalogue. En effet, elle est réalisée à la main et demande une fine analyse du document et de la terminologie ainsi que de bonnes connaissances métier. Étant donné le nombre croissant de ressources médicales de qualité sur Internet, l'équipe CISMeF a cherché à augmenter sa productivité en disposant d'outils automatiques pour l'indexation. Ainsi en 2005, les travaux de thèse d'A. Névéol [Névéol05b, Névéol05a] ont mené à l'élaboration du système MAIF (MeSH Automatic Indexing in French) un système d'indexation automatique pour le MeSH. D'autres travaux ont porté sur l'indexation automatique et la recherche bimodale (combinée) texte et image (travaux de Philippe Florea [Florea07b, Florea07a]).

CISMeF a, enfin, su valoriser ses travaux de recherche avec l'aboutissement de nombreux projets industriels. Le catalogue CISMeF a ainsi donné naissance à d'autres

`fr/l@stics/publis.html`.

24. L'Unified Medical Langage System contient plus de 100 terminologies médicales en différentes langues, celui-ci est décrit dans la section 2.3.2.3.3.

portails d'information grâce à des partenariats avec des industriels, ces portails venant directement interroger le moteur de recherche CISMeF sur un type de document précis. Le site CISMeF-Bonnes pratiques [25] permet de ne diffuser que les recommandations de bonnes pratiques pour les médecins. Le portail PIH (Portail Institutionnel du Handicap [26] RNTS 2005) créé en collaboration notamment avec la société TEmis (PME spécialiste du text mining), permet de rechercher des informations sur le handicap. Le portail KISMeF est né d'une collaboration avec l'Institut National du cancer (INCA), pour rechercher des informations autour de la spécialité Cancérologie pour les patients (2005-2007). Un portail pour l'industrie pharmaceutique a aussi été réalisé avec le laboratoire Lilly. Dans le même cadre, on peut citer la création du moteur de recherche Doc'UMVF [Cuggia07] (2002-2005) avec l'UMVF (Université Médicale Virtuelle Francophone [27]). L'extension de la recherche dans CISMeF vers d'autres moteurs de recherche en santé francais et anglophones a aussi débouché sur un partenariat avec la société Vidal pour l'extension de recherches dans le projet Vidal Recos. Ce partenariat de longue date avec Vidal permet également à l'équipe CISMeF de bénéficier d'un accès à certaines informations incluses dans les bases de données du Vidal.

1.3.1.4 Les besoins

Après une première avancée dans le domaine de l'indexation automatique MeSH avec les travaux d'Aurélie Névéol, l'équipe CISMeF a voulu poursuivre ses efforts dans ce domaine. Ceci a conduit à indexer une partie des ressources (celles considérées de qualité et d'importance moindre) à l'aide de processus automatique [Névéol07b] (niveau 2 et 3 d'indexation). Cette avancée a permis de doubler en peu de temps le nombre de ressources disponibles dans le catalogue CISMeF. La réactivité de l'équipe est ainsi plus grande face aux demandes des utilisateurs et à l'amoncellement de ressources d'intérêt disponibles sur Internet. Le premier besoin est donc de continuer les efforts entrepris en améliorant les méthodes d'indexation acquises et en explorant de nouvelles.

L'équipe a constaté au fil des années une forte montée de l'intérêt pour d'autres terminologies au sein de la communauté hospitalière [28] et des spécialistes. Le deuxième besoin s'exprime donc dans la prise en compte d'autres terminologies au sein du catalogue.

L'une des critiques qui revient le plus souvent au sujet du moteur de recherche CISMeF est la complexité de la recherche d'information qui s'est créée au fur et à mesure des nouveaux développements dans le catalogue. CISMeF travaille donc continuellement à l'amélioration de l'accès à ses informations. Une des améliorations serait de faciliter l'accès à l'information pour les médecins aux différentes bases de données accessibles sur Internet (Medline, Orphanet etc...). En effet, la recherche d'information au cours de l'activité d'un praticien n'est pas encore systématique car

25. Portail accessible ici : `http://doccismef.chu-rouen.fr/servlets/CISMeFBP`.
26. Portail accessible ici : `http://doccismef.chu-rouen.fr/servlets/PIH`
27. Accessible *via* `http://www.umvf.org`
28. On rappelle que l'équipe CISMeF est localement située au sein du CHU de Rouen

elle demande pour le moment d'y consacrer beaucoup de temps.

Enfin, CISMeF est devenu l'un des leaders dans la conception de moteurs de recherche intelligents dans le domaine médical. Son expertise et son expérience sont sollicitées dans la conception de moteurs de recherche spécialisés pour des équipes de recherche et des industriels. Ainsi l'arrivée du dossier patient électronique dans les hôpitaux a entraîné une réelle demande tant pour la structuration que pour la recherche d'information au sein du dossier patient.

1.3.2 Produits et centres d'intérêt de la société Vidal

1.3.2.1 Du papier à l'électronique...

Tout commence au début du XX$^{\text{ième}}$ siècle, les médecins prescrivent alors des «préparations magistrales» que les pharmaciens confectionnent à la demande. Face au succès de certaines préparations, des pharmaciens pensent à fabriquer à l'avance certaines formules, qu'ils proposent directement aux malades et qu'ils font connaître en insérant de la publicité dans des quotidiens. L'industrie pharmaceutique commence à émerger.

C'est dans ce contexte que Louis Vidal créé des fiches pharmacologiques décrivant les médicaments fabriqués de facon industrielle et diffuse ces fiches directement aux médecins, *via* un réseau de visiteurs médicaux. Il créé la société OVP (Office de Vulgarisation Pharmaceutique) en 1911. Le premier dictionnaire des spécialités pharmaceutiques, qui deviendra le dictionnaire Vidal en 1961, apparaît en 1914. Il comporte alors 336 monographies et une classification pharmaceutique. En 1989, OVP s'ouvre à la technologie informatique avec le premier CD-Rom Vidal [29]. En 1995, les produits d'OVP au départ à visée des professionnels de santé et des industries pharmaceutiques s'ouvrent sur le grand public avec le Vidal de la famille.

Vidal SA est passé rapidement de l'édition d'un dictionnaire sur le médicament à la gestion d'une base de données multiplateforme, scientifique et réglementaire s'adressant à tous les professionnels de santé.

L'arrivée du support électronique a permis à Vidal de créer l'une des plus grosses bases de connaissances sur le médicament permettant de nombreux traitements informatiques sur les données qu'elle contient.

L'expertise clé de Vidal réside en un savoir faire dans le domaine de la structuration de l'information de Santé. Cette structuration prend tout son sens en offrant la possibilité, pour l'utilisateur, d'accéder de facon contextuelle à l'information qui l'intéresse. Par ailleurs, les systèmes d'aide à la décision thérapeutique voient leur efficacité s'améliorer grâce à l'usage de données contextuelles sur le médicament.

Aujourd'hui la société Vidal est le spécialiste de l'information de référence sur les produits de santé et des services d'aide à la prescription, à la dispensation et à la délivrance.

29. En 1992, le premier Vidal électronique naît d'une collaboration avec le Dr. Darmoni.

1.3.2.2 Les produits

La société collecte et diffuse [30] l'information de référence réglementaire, administrative, économique et thérapeutique sur différents supports : papier (dictionnaire Vidal, Tarex,...), CD-Rom (VidalCD, VidalExpert,...) et sites Web.

Les données sont rendues plus accessibles grâce à des moteurs de recherche. Les produits électroniques proposent une recherche de spécialités [31] selon plusieurs critères :

- son nom (exemple : «Sectral»)
- les substances qu'elle contient (principe actif ou excipient, exemple : «acebutolol» associé à la spécialité «Sectral»)
- les indications pour lesquelles cette spécialité peut être prescrite (exemple : «diabète insulinodépendant» associé à la spécialité «insuline actrapid»)
- laboratoire de fabrication
- forme/couleur
- par catégories (par la classification thérapeutique Vidal ou l'ATC [32] ou Ephmra [33])

L'utilisateur peut alors consulter la monographie [34] pour la spécialité retrouvée.

Les logiciels Vidal mettent à la disposition des utilisateurs des fonctionnalités de sécurisation de la prescription avec, entre autres, détection des interactions médicamenteuses et proposition d'alternatives thérapeutiques (spécialité appartenant à la même classe pharmacothérapeutique ou dont l'indication thérapeutique est identique). C'est ainsi qu'après une recherche de spécialités que le médecin désire prescrire, il peut procéder à l'analyse de sa prescription médicamenteuse. Au vu de la présence des deux spécialités «Teralithe 400mg en comprimé» et «Advil 400 mg en comprimé», le système va émettre une alerte puisque cela peut entraîner une toxicité pour le patient (voir figure 1.5). Le système peut alors proposer de remplacer l'une des spécialités par une autre qui n'entrainerait aucune interaction (exemple : remplacer l'«Advil» par l'«ALGISEDAL en comprimé»).

L'outil d'aide à la prescription peut aussi prendre en compte l'état physiopathologique d'un patient (grossesse, allaitement, insuffisance rénale, poids, âge, sexe etc...) décrit à l'aide des terminologies standard (CIM10, CISP, DRC [35].). Ces éléments sont liés aux informations contenues sur les médicaments en base afin de créer des alertes de différents niveaux : contre-indications et précautions d'emploi (exemple : la prescription de la spécialité «Sectral» contre-indiquée pour les asthmes sévères, à un malade atteint d'asthme aigu grave (ayant pour code J46 dans la CIM10)). Afin d'aider le médecin dans cette démarche un logiciel d'aide au codage

30. Pour avoir plus de détails sur les différents produits voir `http://www.vidal.fr/`.

31. Une spécialité est la base du médicament, elle peut être commercialisée sous différentes formes et sous plusieurs noms de marque.

32. La classification Anatomique, Thérapeutique et Chimique.

33. La classification de l'European Pharmaceutical Marketing Research Association.

34. Une monographie est élaborée par Vidal et regroupe l'ensemble des informations du Résumé des caractéristiques du produit (RCP) des textes publiés au Journal Officiel et le cas échéant d'autres documents officiels pour une spécialité.

35. DRC : Dictionnaire des Résultats de Consultation publié par la SFMG (Société Francaise de Médecine Générale http ://www.sfmg.org/)

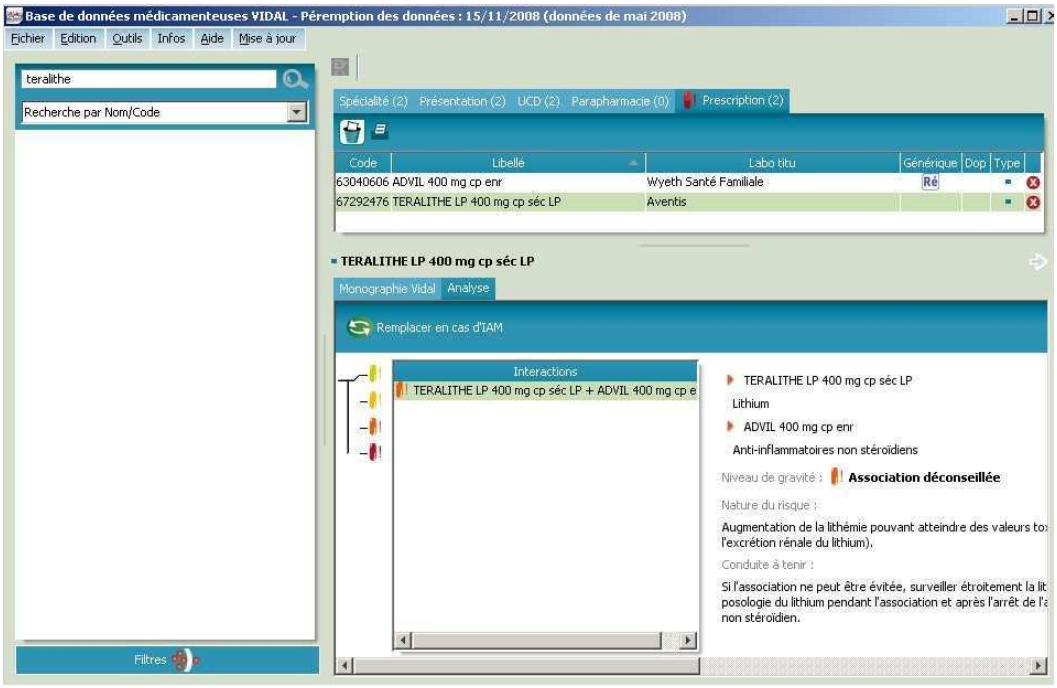

FIGURE 1.5 – Exemple d'une alerte concernant une interaction médicamenteuse détectée à l'aide du logiciel VidalExpert

permet de naviguer dans les terminologies et de rechercher les codes correspondant aux maladies.

Vidal a aussi développé Vidal Recos [36], outil d'aide à la décision thérapeutique, qui diffuse des synthèses de recommandations thérapeutiques [37] pour des pathologies spécifiques.

Pour être au plus près de l'activité des professionnels de santé et permettre un accès très rapide aux informations, des outils ont été développés sur assistants personnels numériques et sur le téléphone portable. Ils sont aussi compatibles avec de nombreux logiciels médicaux disponibles sur le marché comme les logiciels de dossiers patients électroniques.

1.3.2.3 Le recueil des informations sur le médicament

Le dictionnaire Vidal contient des informations sur plus de 5 000 médicaments et 4 900 produits de parapharmacie. Pour chaque médicament, l'information est contenue dans une monographie qui reprend l'information officielle des Résumés Caractéristiques du Produit (RCP) issus de l'Agence Francaise de Sécurité Sanitaire des Produits de Santé (Afssaps) ou de l'EMEA (European Medecines Agency). Afin de commercialiser un nouveau médicament ou spécialité pharmaceutique, le labora-

36. Pour tester une recherche sur 3 «recos» voir `http://www.vidalrecos.fr/pages/index.php`

37. A ce jour, il semble que les deux outils les plus utilisés pour diffuser les recommandations francophones soient CISMEF-Bonnes Pratiques et Vidal Recos.

toire pharmaceutique doit faire auprès de l'organisme habilité (Afssaps [38] en France) une demande d'Autorisation de Mise sur le Marché (AMM) nationale ou européenne pour celui-ci. À la demande d'AMM est associé un dossier comprenant, entre autres, le résultat d'études cliniques visant à démontrer l'intérêt de l'usage du médicament dans le traitement de la pathologie à laquelle il est destiné. Lorsque l'AMM est accordée, elle est accompagnée d'une décision et d'annexes dont le RCP (Résumé des Caractéristiques du Produit) reprenant les données cliniques du médicament, de la notice et de l'étiquetage (Art. L.5121-8 du Code de la Santé Publique). Par la suite, il peut être procédé à une mise à jour du RCP *via* des rectificatifs d'AMM associés à de nouveaux RCP.
Le RCP précise notamment : la dénomination du médicament, la composition qualitative et quantitative, la forme pharmaceutique, les données cliniques (posologie, indications, contre-indications, effets secondaires, précautions d'emploi, etc...). La notice qui accompagne chaque médicament présente l'essentiel des informations du RCP dans un vocabulaire plus accessible pour le patient.

Les RCP sont directement obtenus auprès de l'Afssaps dès leur publication. Les différentes équipes Vidal sont alors chargées de recueillir les informations et de les saisir dans la base de connaissance sur le médicament.

Afin de permettre la sécurisation des prescriptions et l'affichage des données dans les logiciels, l'équipe Données Thérapeutiques Structurées du Vidal est chargée d'indexer manuellement les données cliniques des RCP grâce à des terminologies spécifiques développées en interne.

1.3.2.4 Une priorité : l'innovation en permanence

La société Vidal travaille sans cesse au perfectionnement de ses produits en intégrant de nouvelles fonctionnalités susceptibles d'intéresser les utilisateurs. L'amélioration de la sécurisation de la prescription passe par l'ajout d'alertes contextuelles grâce à l'intégration de nouvelles données sur le médicament.

La société Vidal cherche également à améliorer l'accès aux informations dans ses produits, par exemple en améliorant les supports d'information avec l'XMLisation des RCP, source de l'information traitée.

Des travaux ont été menés afin d'enrichir les terminologies utilisées avec notamment les projets de recherche UMLF et VUMeF (avec l'équipe CISMeF et le laboratoire LERTIM voir section 1.3.1) pour la recherche d'information et l'indexation des RCP.

Dans le même objectif un travail a été réalisé afin de créer une nouvelle terminologie, le TUV (voir section 2.4.2.3) à partir des quatre terminologies d'origines : thesaurus d'indications, contre-indications, précautions d'emploi et effets secondaires, permettant de structurer plus finement les termes afin d'enrichir les connaissances de la base et de les harmoniser en vue d'améliorer les fonctionnalités de recherche et d'alertes dans les produits Vidal. La gestion en est aussi facilitée puisqu'il ne restera qu'une seule terminologie à gérer.

38. Afssaps : Agence francaise de sécurité sanitaire des produits de santé.

1.3.2.5 Les besoins

Une fois la terminologie TUV terminée, il sera nécessaire de la maintenir et de la faire évoluer.

Par ailleurs, d'autres terminologies destinées à l'implémentation de nouvelles alertes voient le jour, ce qui complique d'autant l'indexation. Cette indexation étant liée aux alertes, il est indispensable de ne rien oublier et de ne pas faire d'erreur. En outre, tous les indexeurs n'indexant pas de la même facon, il est aussi important d'arriver à une bonne harmonisation de l'indexation produite. Il devient indispensable d'aider les indexeurs dans l'indexation des RCP.

Vidal souhaiterait aussi proposer une nouvelle fonctionnalité aux médecins qui leur permettrait d'accéder directement aux passages importants du RCP dès lors qu'une alerte s'est produite. Ceci suppose l'existence d'un lien entre l'indexation et la, ou les, portion(s) de textes correspondantes dans le RCP. C'est ainsi que le Vidal s'est penché sur l'indexation assistée (ou semi-automatique) avec le développement d'un nouvel outil de travail pour les indexeurs de l'équipe scientifique, BIBLIS (développé par l'équipe IMAG de l'Université de Grenoble). Au début de cette thèse, ce logiciel était en discussion, les spécifications n'avaient pas encore été concues.

1.3.3 Activités de recherche du LERTIM

1.3.3.1 Domaine de Recherche du LERTIM

La recherche médicale au laboratoire LERTIM [39] s'intéresse à l'élaboration de systèmes d'information hospitaliers [40] performants (adaptés et évolutifs) [Fieschi05].

Le dossier médical informatisé est l'une des composantes du système d'information en réseaux de l'hôpital. Ce dossier permet de recueillir pour chaque patient toutes les informations qui ont trait à son état de santé et à son parcours de soin. En outre, l'informatisation de ce dossier permet :

- de faciliter la coordination des soins et la communication entre les différents professionnels de santé avec un système de prise en charge partagée du patient au sein des différentes structures de soins du réseau.
- de faciliter l'exercice professionnel quotidien par la fourniture d'outils de recherche d'information rapides permettant de rechercher selon plusieurs critères : nature des données (cliniques, biologiques, imagerie), ordre chronologique, nom, âge, lieu de domiciliation, type d'affection.
- l'amélioration de la prise en charge du patient en assurant un soutien auprès du médecin grâce à des outils d'aide à la décision, de contrôle de l'activité, de suivi du processus de soins et de sécurisation des soins.
- l'accès aux connaissances médicales (informations sur la recherche clinique, guides de bonnes pratiques cliniques) pour les médecins ainsi que l'aide à l'apprentissage dans le cadre de la formation médicale continue [41].

39. Le site du laboratoire est accessible ici : `http://cybertim.timone.univ-mrs.fr`.

40. Un système d'information représente l'ensemble des éléments participant à la gestion, au stockage, au traitement, au transport et à la diffusion de l'information au sein d'une organisation.

41. Les pratiques médicales sont en perpétuelle évolution, le médecin est donc obligé tout au long

- de recueillir les données concernant l'activité de l'hôpital (les pathologies traitées par exemple) et son mode de fonctionnement (exemple : mode de prise en charge) afin de procurer à l'hôpital les financements adéquats (tarification à l'activité T2A), d'évaluer la qualité des soins à l'intérieur de l'établissement, de contrôler son activité, et de fournir des données pour la veille sanitaire et les études épidémiologiques. Les données recueillies sont codées (le codage est très semblable à l'indexation si ce n'est que les mots-clés assignés à un document sont exprimés sous forme de codes) à l'aide de terminologies spécifiques : la CIM10 (pour les diagnostics) et la CCAM (pour les actes). Ce codage est appelé codage médico-économique.
- l'accès par le patient à son dossier de santé par Internet où qu'il se trouve lui permet de prendre en charge lui-même certains éléments concernant sa santé (par exemple : messages d'alerte automatiques pour les vaccinations et examens et agenda des consultations).

Les activités de recherche du laboratoire LERTIM consistent à rendre possibles ces différentes tâches concernant le dossier médical informatisé.

L'activité du LERTIM concerne, entre autres, la biostatistique, la représentation des connaissances, l'aide à la décision, les systèmes d'information médicaux et de santé, les systèmes d'information pour la formation à distance et le soutien méthodologique en recherche clinique.

1.3.3.2 Les travaux du LERTIM

L'équipe du LERTIM a mené des travaux pour le dossier médical électronique autour de deux axes :

- Le premier axe consiste à comprendre, expliciter, modéliser, représenter et utiliser les connaissances contenues dans le dossier médical informatisé afin de faciliter leur accès et leur acquisition.
  La plupart des auteurs menant des études épidémiologiques sur la base des banques de données médico-économiques concluent à leur manque de qualité, de validation et leurs lacunes concernant les données cliniques [Deyo94] [Jollis93]. En effet, des diagnostics susceptibles de baisser la tarification seront peu souvent codés. Il s'avère donc nécessaire de compléter ces bases par une indexation complète et descriptive. De plus, grâce à une indexation descriptive de l'intégralité des documents du dossier patient, une recherche d'information telle que celle effectuée dans le moteur de recherche CISMeF serait tout à fait possible. Permettre cette indexation ainsi que le codage des données médico-économiques sous-entend l'usage de terminologies riches ainsi que des connaissances sur le langage médical. Dans cet objectif, un important travail a été réalisé afin de créer de meilleures ressources, pour le traitement des termes biomédicaux permettant ainsi, un meilleur accès aux informations contenues

de sa carrière de maintenir ses connaissances à jour grâce à la formation médicale continue.

dans les parties textuelles des dossiers médicaux électroniques [Avillach08b]. Ces travaux ont été menés dans le cadre du projet UMLF [Zweigenbaum03] et VUMeF [Darmoni03b] en partenariat avec des industriels et d'autres laboratoires dont le Vidal et l'équipe CISMeF. Il est à noter que ces travaux concourrent à l'amélioration de la recherche de connaissances dans la littérature biomédicale et le Web de santé, pour l'aide à l'apprentissage et à la prise de décisions.
Afin d'améliorer la mise en œuvre de connaissances médicales dans un but de sécurisation des soins, des outils permettant le couplage connaissances médicales et informations sur le patient ont été développés. Ces outils visent à améliorer la décision médicale et la prise en charge du patient. Le projet ASTI en 2006 [Bouaud02] se proposait de concevoir et d'évaluer une 2ième génération de systèmes informatisés d'aide à la prescription, définis comme des outils capables d'aider le prescripteur à recourir à la meilleure stratégie thérapeutique en situation clinique.
Une série de projets, les projets ARIANE [Joubert02], COMeDIAS [Joubert03] et WRAPIN [Joubert07a], ont eu pour but de permettre aux professionnels de santé d'accéder à des bases d'information du domaine biomédical (bases de données patients, banque de données sur les médicaments, guides de bonnes pratiques, bibliographie) dans le système d'information de leur entreprise ou sur Internet grâce à un ensemble de services Web en partenariat avec Health On the Net [42]. D'autres projets, comme les projets xGA (multiple (x) Guideline Applications) ont consisté à mettre en œuvre des Guides de Bonnes Pratiques Cliniques informatisés [Dufour05].
Enfin, afin de permettre un meilleur accès à l'information et une meilleure acquisition des connaissances, une partie des travaux de recherche a été réalisée sur la médiation des savoirs au sein du consortium UMVF [Joubert07b]. L'UMVF a pour objectif de favoriser les usages pédagogiques des Technologies de l'Information et de la Communication pour les formations initiales et continues des professionnels de santé.

– Le deuxième axe est le soutien à la recherche clinique et aux recherches en biostatistiques. Les travaux de recherche clinique ont concerné la recherche de facteurs pronostiques notamment en cancérologie avec le projet MEDuS. L'objectif de ce projet était d'évaluer différentes méthodes d'estimation de la survie, de proposer des conseils pratiques aux utilisateurs et de proposer des nouvelles extensions à des modèles existants ou bien de nouvelles techniques d'analyse [Giorgi05].

1.3.3.3 Les besoins

L'équipe du LERTIM travaille à améliorer l'accès aux informations contenues dans les parties textuelles des dossiers médicaux électroniques. Cette amélioration pourrait être obtenue par la structuration des données textuelles contenues dans le

42. Pour plus d'informations sur HON `http://www.hon.ch/`

dossier patient électronique et l'intégration d'un moteur de recherche efficace. Manuellement, il serait très difficile de restructurer toutes les données déjà présentes dans le dossier médical. En effet, le dossier médical informatisé d'un hôpital de plus de 1 000 000 de patients comme Rouen peut contenir plus de 190 000 comptes rendus d'hospitalisation et autant de courriers électroniques et de résultats d'examens. Un outil d'indexation automatique permettant l'indexation du contenu des dossiers médicaux avec un minimum d'interventions humaines serait donc très utile.

Dans le cadre du financement de l'hôpital, les médecins ont l'obligation pour chaque séjour de leurs patients de coder leurs informations médico-économiques (diagnostics à l'aide de la terminologie CIM10 et les actes avec la CCAM). Les études de médecine n'enseignent pas la manière d'indexer des documents à l'aide des terminologies standard. Ce codage est complexe et s'avère très fastidieux pour les médecins qui ont déjà peu de temps pour traiter l'ensemble de leurs patients. Un outil d'aide à l'indexation semi-automatique pour le codage médico-économique permettrait aux médecins de gagner un temps précieux pour une meilleure prise en charge de leurs patients.

1.4 Conclusion

Nous avons pu constater que depuis quelques années le Vidal, le LERTIM et l'équipe CISMeF travaillent sur des problématiques proches : sécurisation de la prescription, structuration de contenus, indexation, création et enrichissement de terminologies, recherche d'information. Ils ont également collaboré sur de mêmes projets (les projets UMLF et VUMeF). Après avoir interrogé les différentes équipes sur leurs besoins, il nous a semblé que l'indexation était au cœur des demandes et devait être le cœur de cette thèse. Cette indexation, pour les besoins de chacun est appliquée à différents types de documents (sites Web, RCP, dossiers médicaux) à l'aide de différentes terminologies dans différents domaines.

L'objectif de notre thèse est de mettre en œuvre des méthodes et de developper des outils susceptibles d'apporter une réponse aux besoins décrits ci-dessus et de s'étendre à d'autres applications. L'indexation doit permettre une meilleure recherche d'information au sein du catalogue CISMeF avec une indexation automatique et semi-automatique des sites Web permettant de recenser dans le catalogue plus de documents plus rapidement. Elle doit par ailleurs permettre d'améliorer la recherche d'information au sein des dossiers électroniques des patients ainsi que d'aider les médecins à produire les codages médico-économiques utiles au calcul des budgets des hôpitaux. Enfin, elle doit optimiser au sein de l'outil BIBLIS chez Vidal l'indexation des RCP pour l'aide à la prescription.

Il nous semble judicieux de construire non pas trois outils mais bien un seul outil capable de réaliser ces différentes tâches. Nous tenterons donc d'explorer un univers encore inconnu pour chaque équipe, celui de l'indexation automatique multi-terminologique, multi-documents et multi-tâches [43]. Nous tenterons aussi d'améliorer

43. Chaque équipe pratiquait une indexation manuelle monoterminologie pour une tâche précise et ne s'intéressait qu'à un seul type de documents.

l'accès aux ressources médicales sur Internet afin d'aider les utilisateurs dans leurs recherches d'information pour l'aide à l'apprentissage et à la décision.

Après cette analyse des besoins, nous allons nous intéresser à l'état de l'art afin de déterminer les solutions envisageables.

Chapitre 2

Analyse du problème

2.1 Introduction

Les besoins étant identifiés, nous allons, dans ce chapitre, examiner l'état de l'art relatif à notre sujet. Deux domaines dans lesquels s'inscrivent ces travaux se dégagent.

Le premier est la recherche d'information électronique, l'indexation des documents étant réalisée à des fins de recherche d'information au sein du dossier patient électronique et du moteur de recherche CISMeF. L'indexation des RCP, elle, n'est pas réalisée à des fins de recherche d'information mais dans un objectif de déclenchement d'alertes de sécurisation. Nous voyons là un deuxième domaine émerger, celui de la construction de bases de connaissances et de systèmes d'aide à la décision. Nous allons dans ce chapitre définir ces deux domaines ainsi que les besoins, usages et accès qui en sont fait par les différents acteurs du monde médical (voir section 2.2).

Ce chapitre présente également la notion d'indexation et sa mise en place dans la réalisation des différentes tâches mises en évidence dans le chapitre 1 (voir section 2.3). La terminologie MeSH et la politique d'indexation des ressources en MeSH au sein de l'équipe CISMeF sont présentées ainsi que le codage médico-économique pour les dossiers patients et les terminologies associées. Suit une présentation de l'indexation des RCP à l'aide des terminologies Vidal (voir section 2.4).

Le sujet et les enjeux posés, nous envisageons de recourir à la construction d'outils d'indexation automatique afin d'améliorer les processus décrits. Nous présentons la notion d'indexation automatique ainsi que les travaux existants dans le domaine et les axes d'améliorations.

2.2 Fondements de la recherche d'information et des bases de connaissances

Le sujet de cette thèse touche deux domaines : la recherche d'information électronique et ses particularités sur Internet ainsi que la construction de bases de connaissances pour les systèmes d'aide à la décision. Nous définissons ces deux do-

maines ainsi que les besoins, usages et accès qui en sont fait par les différents acteurs du monde médical.

2.2.1 Recherche d'information électronique

2.2.1.1 Historique

Les informations médicales peuvent revêtir plusieurs formes : dessins, tableaux ou textes. Nous nous sommes intéressés aux formes textuelles de l'information médicale. Cette information, à l'origine non structurée, est contenue dans des textes : rapports, notes, articles, livres etc... Ces informations sont transcrites par l'écriture afin d'assurer le recueil et la transmission des savoirs.

Avec ces recueils et le besoin de recherche de savoir est née la recherche d'information. Nous définissons la recherche d'information comme l'ensemble des méthodes, procédures et techniques permettant à un individu de sélectionner l'information qui lui semble pertinente dans un ensemble de documents pour répondre à son besoin. Un système de recherche d'information est, dès lors, l'ensemble des modèles et des processus permettant la sélection d'informations pertinentes dans une ou plusieurs documents textuels en réponse aux besoins d'un utilisateur.

Les premiers outils de repérage de l'information datent de plusieurs millénaires [Fayet-Scribe97]. C'est dans l'Antiquité (-4 000 à -3 000 ans avt JC en Basse Mésopotamie) que l'on voit apparaître les premiers tableaux et listes ainsi que les premiers résumés de documents. À la bibliothèque médicale d'Assurbanipal (en -800 à -600 ans avt JC en Mésopotamie), les premiers catalogues, répertoires permettent de réaliser un inventaire des ouvrages et de les répertorier afin de pouvoir les retrouver. Les encyclopédies quant à elles permettent d'organiser les connaissances par thème. Sont apparus ensuite les premières bibliographies et tables de contenu (Rome au 1er siècle), les premiers index (au II^{e} et III^{e} siècle), et les classifications universelles et encyclopédiques (exemple : première édition de la classification de Melvil Dewey (1875)). Les ouvrages sont alors répertoriés, leurs contenus brièvement explicités et le savoir est divisé en classes afin que la recherche d'information soit rendue plus facile et plus rapide. La mécanisation a permis des opérations de tri, classement (par thématique), interclassement avec les catalogues réalisés par listage automatique de références (auteur, date, titre etc...) reportées sur des cartes perforées.

Les références sont des données structurées qui permettent le classement et donc la recherche facilitée des données textuelles qui sont non structurées [Lefèvre00]. Ces données structurées sont appelées les métadonnées ou champs de catalogage. On peut distinguer les données sur la forme (titre, auteurs, date etc...(Dublin Core [Dekkers03]) caractéristiques externes du document) et celles sur la description du contenu (résumé, index). L'opération de catalogage permet à l'utilisateur de rechercher des documents par leur titre, leur auteur ou leur date. Cette opération est importante car la masse d'information médicale est telle que, si le document n'est pas répertorié, il devient introuvable et donc inutilisable. Si l'on ne connaissait ni l'auteur ni le titre de l'ouvrage, la méthode de recherche d'information précédente ne

serait d'aucune utilité puisqu'elle consiste à d'abord sélectionner le thème qui correspond le mieux à l'information recherchée puis à consulter tous les index et les résumés voire tous les contenus des ouvrages si la question est très précise. Cette méthode est bien entendue rendue impossible à cause du volume de données à consulter.

La solution est venue avec l'informatisation et les premières terminologies dédiées :

- L'informatisation a permis, au XX^{e} siècle, de palier les problèmes de la recherche d'information papier : lenteur, non exhaustivité, lenteur de diffusion, problèmes d'archivage, coûts. Les catalogues sont alors devenus centralisés et produits en coopération. L'information médicale contenue dans les ouvrages est alors structurée dans des bases de données mises en mémoire dans les ordinateurs. L'informatisation a aussi permis aux usagers d'interroger cette base de données grâce à un ordinateur dans la bibliothèque ou chez eux grâce à Internet.
- Les thésaurus, apparus au milieu du XXe siècle, sont des terminologies dédiées créées afin de décrire le contenu des documents et de permettre ainsi de compléter les métadonnées existantes dans les bases de données bibliographiques.

De grands fonds documentaires médicaux ont ainsi vu le jour (exemple : la base de données Vidal sur les médicaments, Medline, ou le fonds documentaire du CDSA [1] (Bibliothèque du Centre du droit de la santé)).

2.2.1.2 Types de recherche d'information

Avec l'informatisation, l'utilisateur en quête d'information doit exprimer ses besoins dans une requête. L'outil informatique va analyser cette requête afin de pouvoir y répondre. Il existe plusieurs types de recherche :

- la recherche factuelle : il s'agit d'une recherche très précise. Ce peut être une recherche d'information structurée dans les bases de données sur les métadonnées. La recherche se fait alors sur les champs de la base de données (exemple : «Quels sont les ouvrages écrits par Randal L. Schwartz ? » renvoie les titres des ouvrages correspondants). Cela peut également consister à chercher la réponse à une question précise dans le contenu textuel de la base documentaire (exemple : les systèmes de question-réponse [Jacquemart03] qui peuvent donner la réponse exacte à des questions comme «Quels sont les symptômes de l'angine ? »).

- la recherche documentaire : l'information est envisagée ici du point de vue du document. Le système de recherche d'information dans ce cas va proposer à l'utilisateur une liste de documents dans lesquels il est supposé trouver l'information dont il a besoin après une recherche dans un ou plusieurs fonds de documents plus ou moins structurés. Pour trouver les documents correspondant à la requête il faut que les métadonnées associées aux documents et la requête de recherche soient exprimées dans le même langage (voir figure 2.1).

1. Mis en ligne ici : `http://www.univ.u-3mrs.fr`

On appelle cela le langage d'indexation. Les documents sont préalablement indexés à l'aide de ce langage et la requête sera traduite dans le même langage. Les index des documents stockés en base correspondant le mieux à la requête initiale seront proposés à l'utilisateur. L'indexation permet ainsi d'éviter de passer en revue tous les documents à chaque nouvelle question.
On retrouve ce type de recherche dans des catalogues ou des bases de données bibliographiques sur Internet (CISMeF, Medline [2] ou OMNI [3]).

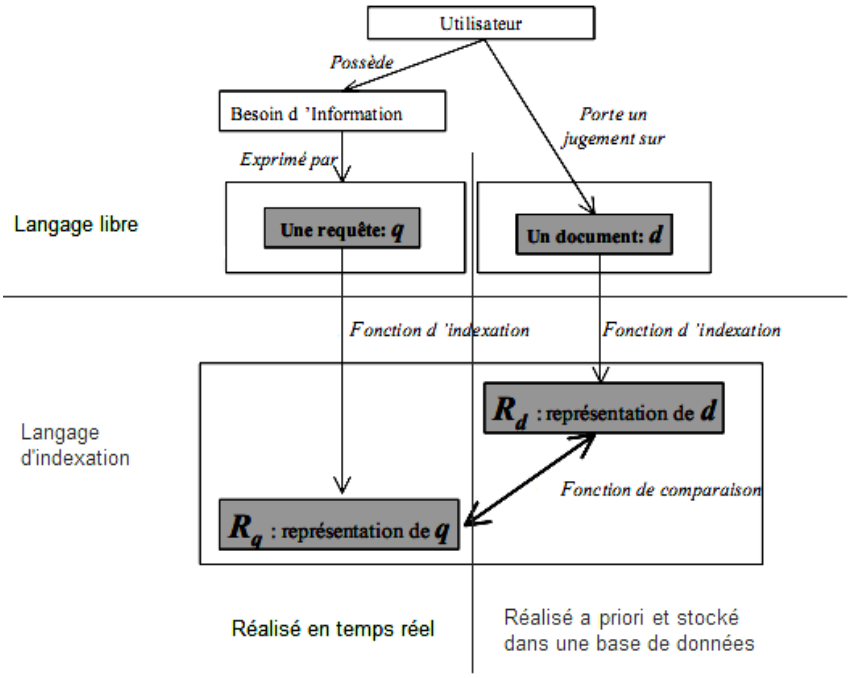

FIGURE 2.1 – Schéma de la recherche documentaire inspiré de [Roussey01]

– la recherche contextuelle [Lefèvre00] : l'évolution actuelle lors de la recherche sur le texte intégral est de non seulement retrouver le ou les documents pertinents, mais aussi de pointer sur la phrase ou la portion de phrase qui constitue une réponse à la question. Elle part d'un mot ou d'un groupe de mots pour aboutir à un texte qui contient les mots en question ou le concept qu'ils représentent.

2.2.2 Particularités de la recherche d'information sur Internet

Internet va fêter, en 2009, ses 40 ans d'existence. Le réseau s'est développé lentement au départ en réponse aux besoins grandissants de communiquer et de partager

2. Base de données bibliographique américaine accessible ici `http://www.ncbi.nlm.nih.gov/pubmed/`

3. Base de données anglaise fournissant des documents Web sur l'éducation et la recherche, site accessible ici `http://www.intute.ac.uk/healthandlifesciences/medicine/`

les travaux des chercheurs grâce à la messagerie et aux serveurs de fichiers. D'abord réservé aux chercheurs, Internet est devenu un instrument de communication ouvert à tous pour échanger, consulter des documents mais aussi en publier. L'arrivée du Web, de l'ordinateur individuel utilisé au travail, dans les lieux publics et à la maison dans les années 90 [Pisani08] [Pillou06] a amplifié le phénomène. Internet connaît ainsi depuis le début des années 90 un développement mondial prodigieux avec un bond de 210% depuis les années 2000. Plus de 20% de la population mondiale (soit 1,5 milliards de personnes) utilisent Internet. La vitalité du réseau s'exprime par une croissance soutenue de l'ordre de 7 millions de pages créées par jour, l'ensemble a dépassé les 10 milliards en 2007 [4]. Le francais est la 5ème langue employée par les utilisateurs (derrière l'anglais, le chinois, l'espagnol et le japonais).

Une des faiblesses d'Internet est que cet espace ouvert, où tout utilisateur devient consommateur et producteur d'information, s'est développé de manière anarchique d'où :

- une quantité énorme de ressources, difficile à évaluer. En effet, la majeure partie des documents ne sont pas répertoriés par les moteurs de recherche généralistes (problème de format, l'absence d'adresse connue). Ce Web «invisible» représente plus de 99% du Web [Bergman01] ;
- l'inexistance d'un index pour le référencement des informations existantes ou leur localisation ;
- des informations instables susceptibles de disparaître avec le temps ;
- tout utilisateur pouvant être producteur d'information, celle-ci peut être de mauvaise qualité puisqu'aucun contrôle n'est réalisé ;
- des informations redondantes [Baeza-Yates99] ;
- des informations hétérogènes (différents formats, supports, langues).

L'Internet est devenu une source majeure d'informations scientifiques et médicales [Schatz97] pour laquelle tous les inconvénients cités ci-dessus sont inacceptables dans le domaine de la santé. En réaction, depuis quelque temps de nombreuses personnes se penchent sur ce monde anarchique pour l'organiser, conscientes que seuls des outils automatiques de recherche peuvent suivre ce rythme de développement. Depuis près de 7 ans, des logiciels «robots» parcourent le réseau de serveurs web pour repérer les pages et en extraire l'information afin de constituer des bases de données.

Pour le professionnel de santé, trouver l'information adéquate sur Internet n'est pas une tâche aisée [Thirion98]. Dans le domaine de la santé, de nombreux travaux ont été entrepris afin de guider les utilisateurs dans leur recherche d'information d'où la multiplication des annuaires et des outils de recherche [Flannery95]. Mais les sites-catalogues ou moteurs de recherches généralistes, comme Google [5] ou Yahoo France [6] ne permettent pas d'obtenir de manière claire et organisée une présentation de l'information disponible en médecine, limitant ainsi son utilisation potentielle. Ces serveurs indexent pourtant un nombre impressionnant de sites médicaux mais l'organisation et la hiérarchie de leurs données ne sont pas adaptées à la médecine.

4. Références de mars 2008 : `http://www.internetworldstats.com/stats.htm`
5. Accessible *via* `http://www.google.fr`
6. `http://www.yahoo.fr`

Des moteurs de recherche fonctionnant sur des bases de données spécialisées ont vu le jour comme Pubmed [7] qui recense 17 millions d'articles scientifiques essentiellement en langue anglaise.

2.2.3 Bases de connaissance et systèmes d'aide à la décision

La société Vidal développe une base de connaissances pour alimenter des outils d'aide à la prescription. Ce type d'outil entre dans la catégorie des systèmes d'aide à la décision.

Depuis les années 70, de nombreux travaux ont été conduits par les communautés d'Intelligence Artificielle et d'Informatique Médicale afin de développer des systèmes d'aide à la décision capables d'améliorer la stratégie diagnostique ou thérapeutique des médecins dans différentes spécialités médicales [Shortliffe76].

L'outil informatique peut apporter une aide directe pour prendre une décision. Il peut faciliter l'accès aux données du patient et améliorer leur représentation (comptes rendus, tableaux de synthèse...). Il peut aussi être capable de faire ressortir rapidement et à partir d'une masse de données hétérogènes et dispersées des informations et des connaissances difficiles à établir par le praticien et qui peuvent confirmer ou infirmer ses choix. L'apport d'un tel système est une complémentarité à l'expertise du médecin. Il est à souligner aussi qu'il constitue également une aide à l'harmonisation des pratiques et à l'auto formation des praticiens.

Les systèmes d'aide à la décision médicale permettent de prédire et prévenir. Ces systèmes peuvent être :

- passifs : le médecin fait appel au système lorsqu'il en a besoin.
- semi-actifs : le système se déclenche de manière automatique (par exemple : le système peut déclencher des alarmes pour signaler des valeurs anormales). Le médecin peut par la suite interagir avec le système.
- actifs : il se déclenche automatiquement sans intervention du praticien.

À partir des informations entrées par l'utilisateur, le système peut alors répondre en donnant un conseil diagnostique ou thérapeutique. Il peut aussi fonctionner en mode critique : l'utilisateur fournit des informations sur le patient et la stratégie mise en œuvre par le praticien, le système peut dès lors émettre des critiques sur les propositions du praticien. Par exemple, les systèmes d'aide à la thérapeutique ont prouvé leur efficacité pour améliorer la qualité des prescriptions médicamenteuses et la réduction des erreurs [Seroussi04].

La décision médicale nécessite la mise en application de connaissances spécifiques à la résolution d'un cas clinique [Degoulet98]. Les informations peuvent être des observations issues de l'examen clinique, des connaissances académiques ou de l'expérience acquise dans l'exercice médical. Ces informations sont stockées dans des bases de connaissances dont le but est de modéliser et stocker sous une forme exploitable par un ordinateur un ensemble de connaissances, idées, concepts ou données et de permettre leur consultation/utilisation. Ces informations peuvent être stockées sous forme de termes provenant de terminologies spécifiques avec leurs réseaux

7. `http://www.ncbi.nlm.nih.gov/pubmed/`

sémantiques. Elles peuvent être entrées à la main, ou issues de procédés d'extraction d'information. Dans notre cas, les données proviennent de l'indexation de documents. Il est nécessaire de mettre à jour de facon régulière la base de connaissance car le domaine de la médecine est un domaine qui évolue en permanence par l'émergence de nouveaux modes de prise en charge des maladies ou de découverte de nouveaux traitements. La BIAM (Banque d'Information Automatisée sur les Médicaments commercialisés en France), Thériaque (base de médicaments du Centre national Hospitalier d'Information sur le Médicament), la BCB (Banque Claude Bernard) et la base Vidal sont les bases de connaissances les plus connues et les plus utilisées dans le domaine du médicament en France.

Une telle base peut être accompagnée de règles (dans ce cas, on parle de base de règles), de faits ou d'autres représentations. Des règles SI-ALORS peuvent être utilisées ainsi que des arbres de décision qui représentent l'ensemble des stratégies thérapeutiques ou diagnostiques du domaine. Un exemple de règle serait «ne pas prescrire la spécialité «Sectral» en cas d'asthme aigu» (exemple repris de la section 1.3.2).

2.2.4 Besoins, usages et accès

L'information recherchée par les spécialistes et le grand public peut être très différente dans le contenu, les supports et la formulation [Chartron92] [Jacquemart05]. Nous distinguons trois groupes de publics pour la recherche d'information médicale : le grand public, les étudiants, et les professionnels de santé.

Les recherches du grand public dans le domaine médical sont dirigées par la curiosité ou la réflexion autour d'un problème personnel ou atteignant un proche. Les patients francais s'orientent de plus en plus vers l'Internet pour rechercher des informations concernant leur pathologie mais aussi leurs droits administratifs et sociaux [8]. Ces informations les aident avant ou, plus souvent, après une consultation médicale. L'information recherchée sera plus synthétique, explicitée et exprimée en langage clair. Le grand public privilégie la facilité d'accès, en revanche le temps d'accès n'est pas un point prioritaire. Les patients privilégieront donc les portails, les logiciels dédiés (tels que les sites CISMeF, HON [9] et Vidal grand public) et les sites d'associations.

Les étudiants s'intéressent prioritairement aux documents didactiques tels que des cours ou des documents plus spécialisés pour apprendre de nouvelles notions ou approfondir leurs connaissances. Ils peuvent utiliser un accès un peu plus spécifique et donc un peu moins facile. Le temps d'accès n'est pas non plus une contrainte. Ils privilégieront les documents électroniques de cours, les sites des universités, les sites dédiés tels que CISMeF et l'UMVF [Darmoni03b] [10].

8. En quelques années, la consultation de sites Web consacrés à la santé a explosé. Depuis sa création en 2000, Doctissimo.fr, leader du secteur a vu son nombre de visiteurs doubler chaque année (4 305 000 personnes ont visité ce site au cours du mois de décembre 2006).

9. WRAPIN (Worldwide online Reliable Advice to Patients and Individuals) `http://www.wrapin.org/`

10. Site du projet accessible ici : `http://www.umvf.prd.fr/`

Les praticiens quant à eux assurent la prise en charge des patients. Ils doivent maintenir leurs connaissances, s'informer des évolutions médicales dans le cadre de la formation continue et répondre aux problèmes rencontrés dans leurs activités professionnelles. Ces informations peuvent conditionner une prise de décision ou une action particulière vis à vis du patient. Ils privilégient les logiciels spécialisés, les sites spécialisés (même en anglais comme Medline [11] ou la National Guideline Clearing House [12]), des outils qui vont les aider dans leur exercice professionnel tels que des logiciels d'aide à la décision (par exemple le logiciel d'aide à la prescription Vidal Expert [13]). Le temps d'accès, là est important car les praticiens peuvent avoir besoin d'informations pour une prise de décision immédiate devant un patient ou, à court terme, avant une opération par exemple. En outre, les praticiens déclarent ne disposer que de 2 minutes en moyenne [Alper01] pour réaliser des recherches. Les recherches sur Internet étant assez longues, elles sont souvent effectuées entre deux rendez-vous ou en fin de journées.

2.3 Définition de l'indexation et du codage

2.3.1 Principe de l'indexation

Nous avons pu constater que l'indexation est utilisée pour la construction de bases de connaissances et pour la recherche d'information.

L'indexation est une représentation extérieure, forcément réductrice du contenu des textes. L'information contenue est alors transférée vers un autre espace de représentation (un langage spécifique) exploitable par un système informatique. La méthode d'indexation dépend du mode de recherche et des applications visées. La notion d'indexation se retrouve dans différents domaines [Lefèvre00] :

- en informatique, l'index qui permet de décrire une base de données est composé des clés d'enregistrement de tous les éléments de la base associés à un pointeur. Ceci permet de localiser plus facilement les données.
- en édition, l'index situé à la fin d'un ouvrage indique les notions importantes développées dans l'ouvrage associées à leur numéro de page d'apparition. Le lecteur peut alors retrouver facilement une notion dans l'ouvrage à partir de l'index.
- en documentation, l'indexation consiste à recenser les concepts (les notions, les sujets) dont traite un document et à les représenter à l'aide d'un langage documentaire. Cette indexation sert à classer et retrouver les documents électroniques dans le cadre de la recherche d'information contextuelle et documentaire. Dans la base documentaire, on retrouve alors pour chaque concept

11. Base de données bibliographique en anglais accessible *via* `http://www.ncbi.nlm.nih.gov/pubmed/`

12. Une ressource publique pour les recommandations de bonne pratique accessible *via* `http://www.guideline.gov/`

13. Pour plus d'informations voir le site de Vidal `http://www.vidal.fr/`

du langage documentaire les emplacements (url par exemple) des documents électroniques qui traitent de ce concept. Nous avons dans notre contexte applicatif deux bases documentaires distinctes :

- Le catalogue CISMeF, base documentaire qui associe à chaque ressource son URL et les termes CISMeF correspondant aux types de la ressource et aux sujets traités dans la ressource.
- Le dossier patient qui pourrait être considéré comme une base documentaire qui associe à chaque document (compte rendu de séjour, courrier des médecins voire résultats d'examens ou radiographies), les termes CIM10, CCAM et SNOMED 3.5 [14] correspondant aux diagnostics, actes et autres éléments médicaux généraux.

Dans le cadre d'une recherche contextuelle, à chaque concept du langage documentaire (voir section suivante) seront associés des index positionnels : adresse du document, numéro de chapitre, de paragraphe, de phrase et position du mot dans la phrase. C'est le principe de la future base de données Vidal qui pour chaque terme du TUV indexé pour une spécialité regroupera le (les) fragment(s) textuel(s) correspondants du RCP.

Nous appellerons cette indexation : «indexation documentaire».

- en analyse de données, l'indexation consiste à recenser certains concepts présents dans un document représentés à l'aide d'un langage fonctionnel (voir section suivante). Cette indexation sert non pas à décrire le document mais à identifier certains concepts à l'intérieur des documents afin de réaliser des traitements informatiques (statistiques, comparaisons, alertes etc...). Dans la base de données, on retrouvera pour chaque entité décrite par le document les concepts qui peuvent lui être associés. Nous appellerons cette indexation : indexation fonctionnelle. L'indexation des RCP à l'aide des thesaurus du Vidal et l'indexation des dossiers médicaux en CIM10 et CCAM sont des indexations fonctionnelles. L'indexation des RCP sert à l'enrichissement de la base de connaissances Vidal qui permet la génération d'alertes dans les logiciels d'aide à la prescription. Au niveau de l'indexation du dossier patient, l'indexation en CIM10 et CCAM des séjours permet aux logiciels groupeurs d'associer de manière statistique un séjour à un coût pour calculer le budget des hôpitaux.

2.3.2 Langage d'indexation

Un langage d'indexation est un langage artificiel.

Dans le cadre de la recherche documentaire, on utilise plutôt le terme de langage documentaire. Celui-ci fournit une représentation formalisée et univoque des documents d'un corpus et des sujets du domaine qui intéressent les utilisateurs. Ce qui permet par la suite de repérer rapidement des documents du corpus qui répondent aux requêtes des utilisateurs. Le MeSH a ainsi été créé pour indexer les articles scientifiques dans le système MEDLARS (système bibliographique biomédical automatisé de stockage et de recherche devenu depuis Medline qui regroupe en 2008 plus de 18

14. Encore peu utilisée en pratique courante en France.

millions d'articles en anglais).

Dans le cadre de l'indexation fonctionnelle, on parle de langage fonctionnel. Celui-ci permet de faire l'inventaire des notions d'un domaine ou pour une tâche précise. Le TUV ainsi que les 4 thésaurus dont il est issu ont été créés pour l'indexation des données thérapeutiques du RCP et la sécurisation de prescriptions du Vidal. De plus, la dixième version de la CIM a été adaptée au codage médico-économique pour décrire l'ensemble des maladies susceptibles d'entraîner un coût pour l'hôpital en France. Enfin, la CCAM a été élaborée uniquement pour la T2A (Tarification à l'activité [Kolher05]) pour décrire les procédures médicales entraînant un coût.

Le rôle du langage documentaire associé à un document lors de la phase d'indexation est double [Salton83] : 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). Un langage fonctionnel, lui, doit surtout être exhaustif par rapport à l'usage qui en est fait et correpondre parfaitement à la tâche demandée.

2.3.2.1 Vocabulaire contrôlé ou libre

Dans l'indexation libre, la forme des termes peut être définie (n-grammes [Halleb97], lemmes, racines etc...) mais les termes n'appartiennent pas à une liste finie. Le vocabulaire utilisé est donc libre. Il peut s'agir de l'ensemble des mots d'une langue. L'ensemble des termes qui peuvent être utilisés n'est pas connu *a priori*. Ce type d'indexation est utilisé dans les moteurs de recherche d'information, par exemple, Google [15] de manière automatique.

Dans le cadre d'une indexation contrôlée, les termes utilisés appartiennent à un langage contrôlé, et donc à une liste fermée. Nous sommes ici le cadre d'une indexation contrôlée puisque tous les termes sont connus à l'avance. Ils sont inclus dans une terminologie, par exemple les terminologies CIM10, CCAM, SNOMED, MeSH et TUV. Des index libres peuvent être extraits pour l'enrichissement de vocabulaires contrôlés ou pour en construire de nouveaux [Charlet06].

2.3.2.2 Un langage pour un objectif

L'indexation n'est pas un but en soi : ce n'est qu'une technique préalable à la recherche d'information et à d'autres types de traitement des informations. Il est important de relier les différentes méthodes d'indexation aux modes de recherche et applications visés.

La méthode ainsi que le langage d'indexation utilisés dépendent de l'objectif à atteindre. L'objectif peut conditionner l'usage des termes ainsi que leurs sens dans le langage d'indexation.

Le langage peut être orienté selon l'utilisateur. Les terminologies que nous utilisons sont très spécialisées. Quelques déclinaisons ont été explorées pour le grand public et les patients par l'équipe CISMeF [**?**].

De plus, il peut aussi adopter le style du langage utilisé dans les documents

15. Moteur de recherche généraliste accessible *via* `http://www.google.fr/`

indexés. Le TUV plus que les autres terminologies possède des libellés tirés directement des RCP (des libellés se rapprochant du langage naturel) contrairement par exemple à la CCAM dont les libellés sont très formatés pour exprimer l'ensemble des conditions d'un acte dans un seul terme.

2.3.2.3 Éléments de représentation

Les informations médicales sont exprimées par tout un chacun en langue naturelle et par écrit en texte libre. Nous nous intéressons ici à la forme écrite qui est le support des informations que nous traitons.

Contrairement au langage informatique, le langage naturel est équivoque [16]. Tout n'est pas exprimé dans un texte (forme implicite), il existe plusieurs facons d'exprimer la même chose (synonymies, paraphrases) ainsi que plusieurs interprétations possibles pour des expressions similaires (ambiguïté et polysémie). De plus, le langage est structuré et permet à partir de concepts élémentaires d'exprimer des concepts plus complexes.

Pour permettre à un outil d'appréhender le langage naturel dans un but d'indexation, il faut tout d'abord lui fournir l'inventaire des termes du langage d'indexation. Il faut également lui permettre d'appréhender le sens de chaque élément ainsi que la formation de termes complexes à partir d'éléments élémentaires. P. Zweigenbaum [Zweigenbaum99] appelle cela le modèle formel. Ce modèle est formé de l'ensemble des termes du langage et des relations qui permettent de relier des concepts généraux à des concepts plus spécifiques, ou de composer des concepts complexes à partir de concepts plus simples. Il existe plusieurs modèles formels, les principaux sont la terminologie et l'ontologie.

2.3.2.3.1 Terminologies

Le mot «terminologie» signifie «ensemble de termes» [Roche05]. La structure et le contenu d'une terminologie sont créés en fonction de l'utilisation qui doit en être faite. Elle est donc généralement créée pour une tâche bien précise. La SNOMED 3.5 [Côté93], la CIM10 [19993], la CCAM [Rodrigues05], le TUV et le MeSH[Douyère04] sont des terminologies.

Dans une terminologie du domaine médical, les concepts du domaine sont normalisés et désignés par des termes précis. La terminologie peut aussi rendre compte des relations qui peuvent exister entre les termes. Les relations de spécialisation-généralisation permettent de hiérarchiser les termes du plus global au plus précis (voir figure 2.2). Un terme plus précis possède toutes les particularités du terme global (au niveau du sens) plus d'autres propriétés qui en font un terme plus spécifique. La définition du terme peut être déduite en partie par des liens que possède le terme avec d'autres termes. Une définition de chaque concept peut aussi être fournie. Une terminologie tente de réduire au maximum les ambiguïtés de sens grâce à sa structure et aux définitions.

À l'intérieur d'une terminologie, les concepts peuvent être désignés par plusieurs

16. Il possède un double sens et peut recevoir plusieurs interprétations.

termes différents (synonymes). Les terminologies peuvent être multilingues, chaque concept peut alors être désigné par plusieurs termes, chacun propre à une langue. Toutes les formes équivalentes sont regroupées sous le même concept. Les concepts peuvent aussi être identifiés par un code numérique ou alphanumérique (un code par concept). Ces codes peuvent refléter la hiérarchie des concepts.

Il existe plusieurs déclinaisons de terminologies :

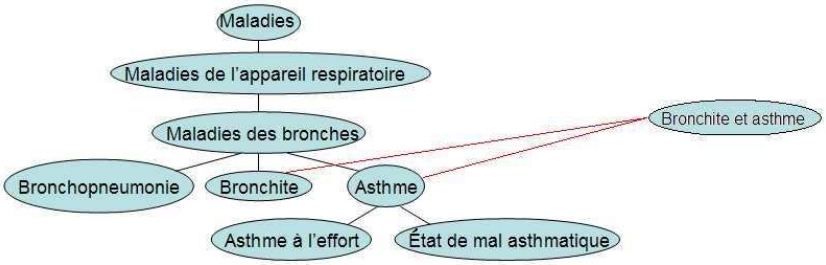

FIGURE 2.2 – Exemple de terminologie ; en noir les relations de hiérarchie (lient un terme général à un terme plus spécifique), en rouge une relation de composition (lie un terme élémentaire à un terme plus complexe)

Vocabulaire contrôlé Un vocabulaire contrôlé est la forme la plus élémentaire d'une terminologie. La signification des termes n'est pas forcément définie et il n'y a pas nécessairement d'organisation logique des termes entre eux.

Classification Une classification est un vocabulaire contrôlé qui a comme particularité d'organiser et hiérarchiser les termes en classes (vocabulaire contrôlé et organisé) [Hoquet05]. Les connaissances sont réparties dans des classes subdivisées en sous-classes plus précises. La CISP (Classification Internationale des Soins Primaires) et l'ATC (classification Anatomique, Thérapeutique et Chimique) sont deux exemples de classification. Dans notre sujet, nous nous intéressons à deux classifications : la CIM10 (voir section 2.4.3.2 pour le détail de cette classification) et la CCAM (voir section 2.4.3.3 pour le détail de cette classification).

Un exemple de classification est la taxinomie, du grec taxis (rangement) et nomos (loi). La taxinomie s'intéresse au classement biologique, en classant les organismes vivants et en les regroupant en entités appelées taxons (familles, genres, espèces, etc...) [Fisher83]. Elle se présente sous la forme d'un arbre, depuis une racine incluant tous les êtres vivants existants ou ayant existé.

Thesaurus Un thesaurus est un vocabulaire contrôlé et organisé [Lefèvre00]. Trois types de relations entre les termes sont considérés : relation hiérarchique (spécialisation - généralisation, tout - partie), relation d'équivalence (synonymes), relation d'association pour les sujets connexes.

Il existe des normes pour l'élaboration des thesaurus monolingues (Norme ISO 2788-1986), multilingues (Norme ISO 5964-1985) et de multiples formats : SKOS

(Spécification en langage RDF développé par le W3C, pour la publication et l'utilisation des thésaurus dans le cadre du Web sémantique) et TMF (Terminology Modelisation Format).

La terminologie MeSH, à laquelle nous nous intéressons, est un thesaurus (voir section 2.4.1.1 pour le détail de ce thesaurus).

Nomenclature Une nomenclature est une terminologie qui vise à recenser tous les termes d'un domaine (exhaustivité). Pour une description précise et fidèle de comptes rendus médicaux, les classifications trop orientées vers un objectif précis se révèlent peu adaptées par rapport à une nomenclature qui fournit un éventail plus varié et plus précis de concepts médicaux.

Une nomenclature est un vocabulaire contrôlé et organisé où les termes peuvent être répartis selon plusieurs axes (ce qui est différent d'une classification généralement monoaxiale). La répartition des concepts en plusieurs axes a pour but additionnel de permettre de composer un concept complexe en combinant des concepts élémentaires pris dans ces axes (exemple : «inflammation, SAI»(axe M), «aigu»(axe G)).

Nous nous intéressons ici à la nomenclature SNOMED (voir section 2.4.3.4 pour le détail de cette nomenclature).

2.3.2.3.2 Ontologie

Une ontologie est un vocabulaire contrôlé, organisé et formalisé [Zweigenbaum95] [Bachimont00]. Elle modélise les concepts, relations et contraintes pour un domaine donné. La relation hiérarchique y est unique : relation «est-un». De plus, il existe des relations sémantiques entre les termes pouvant être associées à des contraintes (voir figure 2.3). En ce sens la terminologie TUV peut s'approcher de la définition d'une ontologie (voir section 2.4.2.3 pour le détail de cette terminologie).

Le format des ontologies est le RDFS (Resource Description Framework Schema) et le OWL (Web Ontology Langage). Des exemples d'ontologies sont les ontologies GALEN [17] (General Architecture for Langage and Nomenclatures [Rector03]) et FMA (Foundational Model of Anatomy) [Rosse03].

2.3.2.3.3 Unification et interopérabilité des terminologies (UMLS)

D.A.B Lindberg, directeur de la NLM, a proposé en 1986, la conception et le développement d'un système de langage unifié ou «Unified Medical Langage System» (UMLS [18]) [Lindberg90] afin d'améliorer l'accès à l'information médicale provenant de sources différentes en permettant aux différentes banques de données de communiquer avec un langage de référence commun. L'UMLS représente une tentative d'approcher au plus près le langage naturel et de lever toutes les ambiguïtés et redondances possibles par une lecture en contexte des documents médicaux. L'UMLS tente de regrouper tous les thesaurus, nomenclatures, et classifications existantes utilisés pour la gestion des données de santé, les bases de données bibliographiques et le

17. Accessible ici `http://www.opengalen.org`

18. Les données de l'UMLS sont accessibles et téléchargeables (sous respect des droits) sur le site UMLSKS `http://umlsks.nlm.nih.gov/`

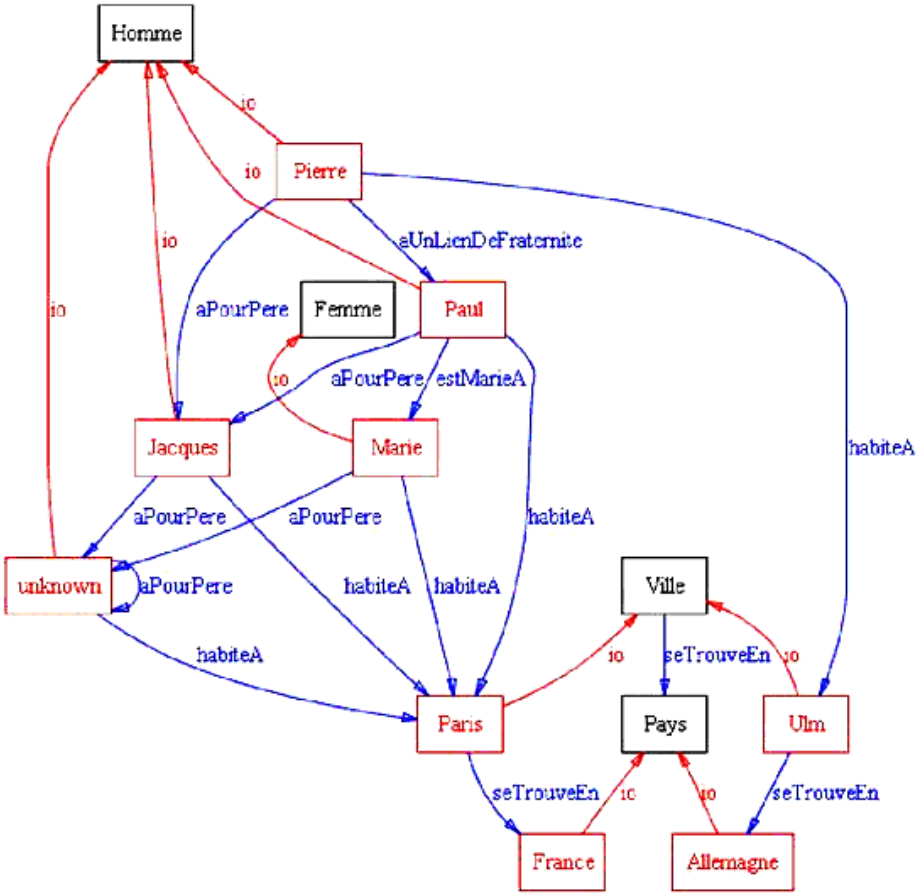

FIGURE 2.3 – Exemple d'une ontologie

dossier patient (plus de 100 terminologies biomédicales dont le MeSH, la SNOMED 3.5 et la CIM10).

L'UMLS est un système qui conjugue trois bases de connaissance : le Metathesaurus (qui regroupe tous les termes), le réseau sémantique (qui regroupe toutes les relations) et le SPECIALIST Lexicon (qui contient les informations syntaxiques, morphologiques et orthographiques).

– **Le Metathesaurus** constitue la base unifiée des concepts médicaux. Il comprend des synonymes, des variations lexicales et des concepts associés afin de dresser la liste de tout le vocabulaire des expressions médicales disponibles. Il a fallu pour créer ce metathésaurus regrouper sous un même concept les différents termes qui l'expriment [Sherertz90] (par exemple : les termes «Atrial Fibrillation» (du MeSH), «Atrial Fibrillation» (de la terminologie PSY), «Atrial Fibrillations» (du MeSH) , «Auricular Fibrillation» (de la terminologie PSY), «Auricular Fibrillations» (du MeSH) appartenant à différentes terminologies doivent être regroupés sous le même concept «Atrial Fibrillation» voir figure 2.4). Chaque concept dans le Metathesaurus a un identifiant unique et permanent (CUI : Concept Unique Identifier). Si un terme MeSH, un terme SNOMED et un terme CIM10 sont associés au même CUI alors c'est qu'ils sont équivalents

en sens (ou synonymes) nous dirons alors qu'ils sont reliés par une relation de transcodage.
À chaque concept correspond : une définition, un terme préférentiel, éventuellement des termes synonymes, des variantes lexicales, un ou plusieurs types sémantiques et un identifiant unique (le CUI).
À chaque terme intégré à partir d'une terminologie extérieure est attribué : un type sémantique, son code dans la terminologie source, le CUI auquel il est associé.
Le Metathesaurus (2007AA) est riche de plus d'1,3 millions de concepts et de 6,4 millions de noms de concepts uniques. Ces concepts sont reliés par 10 millions de relations héritées des terminologies sources et par plus de 2 millions de termes différents (dont 126 000 en langue francaise grâce, entre autre, au projet VUMeF [Darmoni03b] qui avait pour objectif d'augmenter la part du francais dans l'UMLS).
Le Metathesaurus est le creuset de plus de 100 terminologies biomédicales [19] (dont le MeSH, la SNOMED RT, CT et 3.5, la CIM9, CIM9CM et la CIM10) dans 17 langues (pour plus d'informations sur la structure de l'UMLS voir Annexes A1).

Concept (CUI)	Terms (LUIs)	Strings (SUIs)	Atoms (AUIs) * RRF Only
C0004238 Atrial Fibrillation (preferred) Atrial Fibrillations Auricular Fibrillation Auricular Fibrillations	**L0004238** Atrial Fibrillation (preferred) Atrial Fibrillations	**S0016668** Atrial Fibrillation (preferred)	**A0027665** Atrial Fibrillation (from MeSH) **A0027667** Atrial Fibrillation (from PSY)
		S0016669 Atrial Fibrillations	**A0027668** Atrial Fibrillations (from MeSH)
	L0004327 (synonym) Auricular Fibrillation Auricular Fibrillations	**S0016899** Auricular Fibrillation (preferred)	**A0027930** Auricular Fibrillation (from PSY)
		S0016900 (plural variant) Auricular Fibrillations	**A0027932** Auricular Fibrillations (from MeSH)

FIGURE 2.4 – Les concepts de l'UMLS

– **Le réseau sémantique** : alors que le Metathesaurus fournit une liste de tout le vocabulaire des expressions médicales disponibles, le réseau sémantique apporte une structure à ces termes. Cette structure permet notamment de procéder à des regroupements, afin d'englober dans une seule recherche tous les termes se rapportant à une expression donnée. Le réseau sémantique comporte 135 types

19. Il est important dans cette thèse de souligner que les terminologies francaises sont très peu représentées dans l'UMLS (la CIM10 et la SNOMED en francais ne sont pas inclus par exemple).

sémantiques (exemples : «disease or syndrome» et «virus») reliés par 54 relations. Ces relations dénotent des liens de hiérarchie et de non hiérarchie telles que les relations sémantiques réparties en 5 catégories (spatialle, temporelle, conceptuelle, physique et fonctionnelle). À chaque terme du Metathesaurus est associé un (ou plusieurs) type(s) sémantique(s). C'est à travers les concepts et leurs relations ainci que ces types sémantiques, qui sont reliés entre eux dans le réseau sémantique, que les expressions du metathésaurus se retrouvent inscrites dans une structure commune.
Cette structure peut être visualisée comme un graphe dans lequel les concepts sont des noeuds et les liens interconcepts sont les liens entre les nœuds. À chaque type sémantique est associé un identifiant unique, un nombre indiquant sa position dans la hiérarchie et une définition. Pour chaque relation, il existe un identifiant unique, un nombre indiquant sa position dans la hiérarchie, une définition et l'ensemble des types sémantiques qui peuvent être reliés par cette relation.

- **Le SPECIALIST Lexicon** contient les informations syntaxiques, morphologiques et orthographiques nécessaires au traitement automatique de la langue anglaise. Il intègre près de 200 000 entrées lexicales. Chaque entrée possède une forme de base (le lemme), une catégorie syntaxique, un identifiant unique et éventuellement des variantes orthographiques.
Une version francaise du SPECIALIST Lexicon a été réalisée (en partie par l'équipe CISMeF et Vidal) dans le projet UMLF.

2.3.3 L'indexation contrôlée en pratique

L'indexation d'un document comme on l'entend ici consiste à repérer dans celui-ci certains mots ou expressions particulièrement significatifs faisant référence à un terme d'une terminologie dans un contexte donné, et à créer un lien entre ces termes et le texte original.

Il existe un consensus en matière d'indexation [Anderson01] qui consiste en quatre étapes :

1. analyse du texte
2. traduction dans le vocabulaire contrôlé
3. relecture, révision, application de règles d'indexation (optionnel)
4. pour le stockage de l'indexation, il reste à lier dans une base de données les termes d'indexation au document indexé.

En pratique voici ce que l'on peut observer :

L'indexation documentaire consiste à parcourir le document, à repérer et à sélectionner les données caractéristiques du contenu et à retrouver la représentation symbolique qui en est faite dans la terminologie utilisée. L'utilisation de logiciels de navigation et d'interrogation de terminologies peut fournir une aide informatique. Le document peut être lu rapidement afin d'en avoir une compréhension générale ou,

très attentivement, dans le cas où l'indexeur devrait lier manuellement les portions de texte retenues aux termes indexés contenus dans la terminologie. Un travail de synthèse est nécessaire afin de ne sélectionner que les éléments importants pour les faire apparaître dans l'indexation. L'indexation peut être différenciée ou non (elle est alors dite à plat). Une indexation différenciée impose à l'indexeur de ranger les termes par ordre d'importance selon qu'ils décrivent plus ou moins bien l'ensemble ou une partie du document indexé. Les terminologies peuvent être accompagnées de règles d'usages que les indexeurs doivent suivre (exemple : afin de préciser le sens d'un terme, celui-ci peut être associé à un qualificatif pour lequel une association est autorisée [20]). Les indexeurs peuvent aussi élaborer des règles d'indexation communes selon l'usage qui sera fait en interne de leurs indexations. L'étape finale consiste à lier dans une base de données le document et les termes d'indexation.

L'indexation fonctionnelle, quant à elle, consiste à analyser le texte et à repérer quels sont les concepts de la terminologie utilisée présents dans le document. Une lecture très attentive et un effort de compréhension plus précis seront nécessaires. Un travail de synthèse est également utile afin d'éviter les redondances. Là encore, des règles d'indexations peuvent exister. L'étape finale consiste à rentrer de nouvelles connaissances dans la base de connaissance telles que, dans le cadre d'indexation de RCP, les indications, contre-indications, effets secondaires et précautions d'emploi rattachés à une spécialité.

Lors de ce travail d'indexation, il est nécessaire de différencier le thème principal des informations secondaires ou accessoires et décider jusqu'à quel niveau de détail descendre dans la représentation de ces informations. Cette profondeur d'analyse influence les niveaux de bruit et de silence obtenus lors de la recherche. En effet, plus l'indexation d'un document est fournie, plus on entre dans les détails, et plus il y a de risques de prendre en compte des aspects qui sont traités superficiellement dans ce document et qui n'en sont donc pas vraiment caractéristiques : cela engendrera du bruit lors d'une recherche. Par ailleurs, si la description est limitée aux thèmes principaux, sans prise en compte de la variété des sujets traités dans les documents, cela engendrera du silence dans la recherche.

Les termes peuvent être organisés et leurs rôles précisés ou encore structurés dans un véritable réseau sémantique [Coret94].

L'indexation peut présenter une variabilité d'un groupe d'indexeur à l'autre, d'un indexeur à l'autre et également pour un même indexeur à deux instants différents.

L'indexeur peut ne pas avoir de connaissances très pointues dans le domaine sur lequel il travaille. Le temps d'indexation dépendra des connaissances dans le domaine d'indexation de l'indexeur, de l'expérience de celui-ci, de ses connaissances de la terminologie utilisée, de la complexité de cette terminologie, de la longueur et de la complexité du document.

Le codage est une forme d'indexation qui consiste finalement à indexer des codes et non pas les termes associés.

Pour l'instant, ces approches sont propres à l'analyse humaine, l'ordinateur n'est

20. On ne peut pas associer le qualificatif «diagnostic» au terme «bibliothèque» par exemple dans le MeSH.

capable de la simuler que dans une faible mesure.

2.4 Les bases de notre sujet : présentation des tâches d'indexation

La définition des différentes notions abordées étant établie, nous présentons ici les tâches d'indexation exécutées par les différentes équipes afin ensuite de trouver des solutions d'améliorations. Nous décrivons les documents indexés, les terminologies utilisées ainsi que les règles d'indexation appliquées.

2.4.1 Indexation des sites Web médicaux par l'équipe CISMeF

Les ressources dans le catalogue CISMeF sont indexées avec la terminologie CISMeF. Nous allons décrire cette terminologie qui se base sur le thesaurus MeSH ainsi que les règles d'indexation permettant d'associer des termes de cette terminologie à une ressource [21].

2.4.1.1 Le thesaurus médical : Medical Subject Heading (MeSH)

La première liste de sujets, la Subject Heading Authority List, a été publiée par la National Library of Medicine (NLM des États Unis dépendant du National Institute of Health) en 1954. Elle était fondée sur la Current List of Medical Literature, le Library's Index-Catalogue et le Quarterly Cumulative Index Medicus Subject Headings de 1940. La première version du MeSH est apparue en 1960 pour indexer les articles scientifiques dans le système bibliographique biomédical automatisé de stockage et de recherche MEDLARS (devenu depuis Medline regroupant aujourd'hui plus de 18 millions d'articles en anglais). Elle est utilisée depuis pour l'indexation et le catalogage par les bibliothèques et d'autres institutions à travers le monde (exemple : CISMeF en France).

Elle a été traduite en 11 langues (francais, anglais, espagnol, portugais...). L'INSERM (Institut National de la Santé Et de la Recherche Médicale) participe à la constitution du MeSH en traduisant celui-ci et en effectuant ses mises à jour en francais à partir du MeSH américain. Une nouvelle version apparaît tous les ans, la dernière en date est la version 2008 [22]. Nous avons utilisé dans nos travaux la version 2007. Un transcodage vers la CIM10 et la CCAM a été réalisé à partir de la version 2007 [Pereira07] par l'équipe CISMeF.

La hiérarchie du MeSH est une hiérarchie à 11 niveaux avec des relations de spécialisation-généralisation et tout-partie divisée en 15 arborescences thématiques

21. Les sites et pages web ou documents numériques sont des documents particuliers que nous appelons ressources.

22. Cette terminologie peut être consultée grâce au MeSH Browser (accessible *via* `http://www.nlm.nih.gov/mesh/MBrowser.html`) de la NLM pour le MeSH américain ou sur le site de l'INSERM (accessible *via* `http://ist.inserm.fr/basismesh/meshv07.html`) pour le MeSH bilingue.

auxquelles correspondent un code spécifique (exemple : l'arborescence thématique «maladie» est associée au code C, voir figure 2.5 pour consulter toutes les arborescences).

À chaque position dans la hiérarchie correspond :

A. Anatomie
B. Organismes
C. Maladies
D. Produits chimiques et médicaments
E. Techniques analytiques, diagnostiques et thérapeutiques équipement
F. Psychiatrie et psychologie
G. Sciences physiques
I. Anthropologie, enseignement, sociologies et phénomènes
J. Technologie aliments et boissons
K. Arts et sciences humaines
L. Sciences de l'information
M. Individus
N. Santé
Z. Emplacements géographiques

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

FIGURE 2.5 – Les 15 arborescences MeSH et un extrait de l'arborescence C

- un terme préféré suivi éventuellement de synonymes. L'ensemble représente plus de 100 000 termes. Il existe plusieurs types de termes : les descripteurs, les qualificatifs et les concepts chimiques supplémentaires. Dans sa version 2007, le MeSH comporte 24 357 descripteurs, 83 qualificatifs et 164 331 concepts chimiques supplémentaires.
- deux codes : un identifiant unique et un code reflétant la place du terme dans l'arborescence (exemple voir figure 2.5 : «amyloïdose» : D000686 et C18.452.090). Un descripteur peut appartenir à plusieurs arborescences, il peut donc avoir plusieurs codes arborescences. Les concepts chimiques élémentaires sont associés à leur numéro CAS [23].
- une définition qui accompagne chaque descripteur.

Les qualificatifs permettent, lorsqu'ils sont combinés à un descripteur, de spécifier davantage le sens du descripteur [Darmoni07] (exemple : «cancer des os/traitement médicamenteux» permet de restreindre le cancer des os (descripteur) au seul aspect du traitement médicamenteux (qualificatif)). À chaque descripteur correspond une liste de qualificatifs auxquels il peut être associé.

De plus il existe deux types de relations :

- la relation «voir aussi» permet de naviguer d'un descripteur à l'autre et de relier des termes proches

23. Le numéro CAS (CAS number ou CAS registry number en anglais) d'un produit chimique, polymère, séquence biologique et alliage est son numéro d'enregistrement unique auprès de la banque de données de Chemical Abstracts Service (CAS), une division de l'American Chemical Society (ACS).

– la relation «ne pas confondre» permet de préciser le sens et de lever les ambiguïtés.

D'autres types de termes sont utilisés pour l'indexation, le catalogage et la recherche en ligne par la NLM : les types de publication (permettent de définir le type des ressources) et les termes géographiques.

2.4.1.2 La terminologie CISMeF une terminologie fondée sur le MeSH

L'équipe CISMeF a adapté le MeSH depuis 1995 pour caractériser davantage les ressources de santé sur l'Internet pour la recherche d'information, l'extraction d'information et la catégorisation. C'est cette terminologie [24] qui est utilisée par l'équipe CISMeF pour indexer les ressources de leur catalogue. Pour les besoins de l'équipe CISMeF, la base des synonymes a été enrichie par les documentalistes en définitions ainsi que de plus de 10 000 synonymes dans le cadre du projet VUMeF (déjà abordé dans le chapitre 1). D'autres types de concepts hiérarchisés ont eux aussi été définis ou étendus : les types de ressources et les métatermes ont été ajoutés [Douyère04]. Une nouvelle relation a aussi été intégrée, la relation «action pharmacologique» qui est une relation descriptive qui indique l'intérêt pratique du composé chimique.

Les types de ressources définissent la nature de la ressource et non pas son contenu comme les mots clés (descripteur ou descripteur/qualificatif) (exemple : le type de ressource «recommandations» est différent du descripteur «recommandations» qui est utilisé pour décrire une ressource qui parle de recommandations) ce qui permet de décrire avec plus de précision une ressource. Ils ont été inspirés des types de publication de la NLM (la National Library of Medicine qui gère la base de données Medline). Ils sont au nombre de 263 et sont accompagnés d'une définition. Le type de ressource peut être utilisé seul afin de décrire la nature de la ressource ou affilié à un descripteur ou une paire descripteur/qualificatif, nous parlons alors de triplet descripteur/qualificatif\type de ressource (exemple : «cancer des os/traitement médicamenteux\matériel enseignement» qui permet de décrire les ressources d'enseignement sur le traitement médicamenteux du cancer des os).

Le thésaurus MeSH dans sa structure d'origine, ne permet pas d'obtenir de vision globale d'une spécialité médicale ce qui peut être utile en matière de recherche d'information. Pour palier à cet inconvénient, l'équipe CISMeF avec l'aide d'experts médicaux a créé manuellement des méta-concepts appelés métatermes car ils permettent une vision plus globale du MeSH en offrant un niveau supplémentaire d'abstraction. Ils correspondent aux spécialités médicales ou aux sciences biologiques (exemple : «cardiologie», «bactériologie»). L'équipe a aussi créé leurs liens sémantiques avec 0 à n descripteurs, qualificatifs, et types de ressources (exemple : le métaterme «cancérologie» est lié au descripteur «vaccins anticancéreux», au qualificatif «radiothérapie» et au type de ressource «service oncologie hôpital») (voir figure 2.6). La terminologie CISMeF comporte 274 métatermes. Les métatermes permettent, lors de la recherche d'information, de prendre en compte tout un ensemble

24. Un accès à cette terminologie est disponible sur le site CISMeF (accessible *via* `http://terminologiecismef.chu-rouen.fr/`).

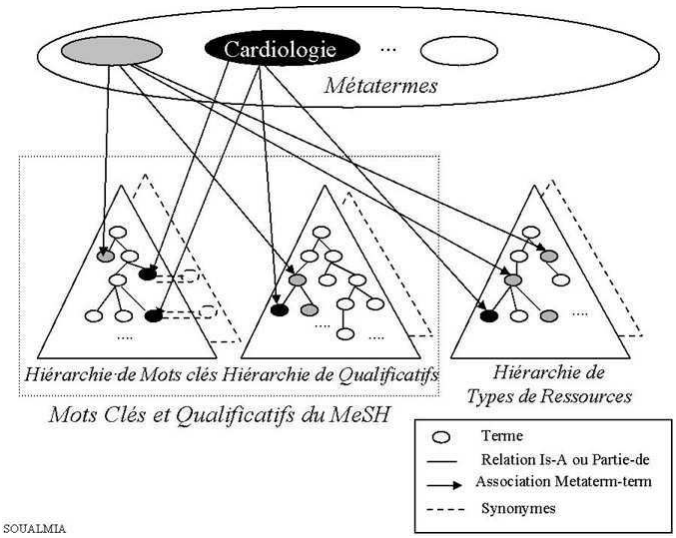

FIGURE 2.6 – Les liens sémantiques entre les métatermes CISMeF et les termes MeSH [Soualmia04]

de descripteurs afin de présenter davantage de ressources à l'utilisateur dans le catalogue CISMeF [Gehanno07].

Certains termes peuvent être à la fois descripteur et qualificatif (exemple : «thérapeutique» voire aussi à la fois qualificatif et type de ressource et descripteur ou qualificatif et métaterme).

2.4.1.3 Règles d'indexation «CISMeFiennes»

Comme nous l'avons dit précédemment, l'indexation d'une nouvelle ressource dans le catalogue CISMeF consiste à créer une notice (voir un exemple figure 1.2) pour cette ressource. Cette notice contient toutes les métadonnées qui permettront au moteur de recherche de présenter et de retrouver la ressource dans CISMeF.

L'indexation consiste à entrer pour une ressource : le titre, les auteurs, l'URL, le format, le site éditeur, le pays d'origine et la date. L'indexeur définit alors le type de ressource, pour cela il choisit 1 à n termes parmi la liste des types de ressources de la terminologie CISMeF. Les types de ressource décrivant plus particulièrement la ressource seront marqués d'un astérisque qui signifie que le type de ressource est «majeur».

Ensuite afin de définir le contenu d'une ressource, un résumé succinct est élaboré par les indexeurs. Enfin, l'indexeur décrit le contenu de la ressource à l'aide de mots clés de la terminologie CISMeF.

Les indexeurs CISMeF privilégient une indexation au plus précis ce qui équivaut dans la terminologie MeSH à ne pas indexer ensemble un père et un fils (sauf exception), mais seulement le plus précis (le fils). Si la ressource comporte les notions d'«asthme» et d'«asthme aigu», c'est «asthme aigu» qui sera indexé. Par contre si le document énumère tous les types d'asthme alors «asthme» sera utilisé pour l'indexation (le père).

Un poids «majeur» peut être apposé à certains mots-clés en y accolant un astérisque. Les mots clés majeurs sont ceux qui décrivent les informations les plus représentatives du document.

L'indexeur utilise le serveur de terminologie CISMeF [25] depuis 2003 pour connaître les termes appropriés à utiliser pour indexer une ressource. Celui-ci permet d'interroger la terminologie grâce à des mots significatifs tapés par l'utilisateur, ainsi que de naviguer à l'intérieur de celle-ci.

L'indexation purement manuelle est réservée aux ressources urgentes (par exemple de nouvelles recommandations pour la bonne pratique) qui doivent être mise en ligne rapidement pour être rapidement accessibles par les médecins.

2.4.1.4 Prémices d'indexation automatique

Pour toute indexation (automatique ou manuelle), l'indexation des métatermes (ou catégorisation en spécialité médicale) se fait de manière automatique [Névéol05a]. Chaque ressource est indexée par une liste de mots clés MeSH, associés ou non à des qualificatifs et par une liste de types de ressources. Par l'intermédiaire des liens sémantiques de la terminologie CISMeF (section 2.4.2), l'algorithme utilisé 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. Un score dit «majeur » est calculé, il correspond au nombre de types de ressource et de descripteurs majeurs à partir desquels le métaterme considéré a été retenu. Les métatermes ayant un score majeur non nul sont dits «majeurs » et sont assignés par un astérisque.

Les ressources moins urgentes sont indexées de manière supervisée. Les indexeurs sont chargés d'indexer manuellement la ressource tout en pouvant s'inspirer du résultat de l'indexation automatique effectuée sur le titre de la ressource.

L'indexation purement automatique sur le titre est réservée aux ressources dont l'utilité et la qualité ne nécessite pas une indexation précise. Ces ressources sont de priorité faible ou leur thème est déjà suffisamment présent dans CISMeF. L'indexation automatique sur le titre est réalisée à l'aide d'un algorithme du sac de mots (voir chapitre 3).

2.4.2 Indexation de l'information pour les médicaments par la société Vidal

2.4.2.1 Le RCP

La définition d'un médicament est précisée en France par l'article L5111-1 du Code de la Santé Publique : «Un médicament est une substance ou composition présentée comme possédant des propriétés curatives ou préventives à l'égard des maladies humaines ou animales». Le médicament est composé d'un ou plusieurs ex-

25. Le serveur de terminologie est accessible *via* l'url : `http://www.chu-rouen.fr/terminologiecismef/`

cipients (substances inertes servant à la formulation de la forme galénique [26] comme l'eau ou le saccharose). Une spécialité est la base du médicament, elle peut être commercialisée sous différentes formes et sous plusieurs noms de marque.

Le Résumé des Caractéristiques du Produit pour une spécialité synthétise les informations du dossier déposé lors de la demande d'AMM notamment sur les indications thérapeutiques, contre-indications, modalités d'utilisation et les effets indésirables. Ces informations sont destinées aux professionnels de Santé (médecins, pharmaciens...) et diffusées par l'Agence Francaise de Sécurité Sanitaire des Produits de Santé (AFSSAPS).

Le RCP comprend plusieurs rubriques distinctes :

- Forme et présentation : présente la forme galénique de la spécialité et ses présentations
- Composition : indique les noms et les quantités des composants constituant le médicament
- Données Cliniques :
  - Indications thérapeutiques : maladie(s) pour le(s)quelle(s) le médicament peut être utilisé
  - Posologie et mode d'administration : doses auxquelles le médicament doit être administré
  - Contre-indications : situation(s) dans le(s)quelle(s) la prise du médicament est dangereuse
  - Mises en garde et précautions d'emploi : situation(s) à considérer lors de la prescription du médicament
  - Interactions avec d'autres médicaments ou substances
  - Grossesse et allaitement : risques éventuels et comportement à adopter en cas de prescription au cours de la grossesse ou en cas d'allaitement d'un nourrisson
  - Conduite et utilisation de machine : comportement à adopter en cas de conduite de véhicule ou d'utilisation de machines
  - Effets indésirables : effets non souhaités, secondaires au traitement par le médicament et aboutissant à un résultat néfaste (gêne, allergie, complications graves, y compris le décès).
  - Surdosage : symptômes et conduite à tenir en cas de surdosage
- Propriétés pharmacologiques :
  - Pharmacodynamique : décrit l'action du médicament sur l'organisme
  - Pharmacocinétique : décrit l'action de l'organisme sur le médicament (vitesse à laquelle le médicament est absorbé, distribué dans l'organisme, métabolisé (transformé), puis éliminé de l'organisme)
  - Sécurité préclinique : les données de sécurité préclinique (la toxicité après des doses répétées, le pouvoir cancérigène...)
- Données pharmaceutiques :
  - Incompatibilités physico-chimiques

26. La forme galénique est la forme d'administration du principe actif au patient (exemple : gélule)

- Conditions de conservation
- Modalités de manipulation

2.4.2.2 Indexation du RCP par le Vidal

La société Vidal exploite, entre autres, les données officielles contenues dans les RCP (Résumé des Caractéristiques du Produit) émis par l'AFSSAPS et le JO (Journal Officiel). Pour chaque spécialité, Vidal recueille, intègre et structure les données économiques, administratives et thérapeutiques. L'un de ses objectifs est de permettre, *in fine*, une sécurisation maximale de la prescription médicale en générant des alertes adéquates et en donnant au prescripteur les informations pertinentes en lien avec le traitement.

Chaque RCP au format PDF est associé par Vidal à des métadonnées sur la forme (spécialité, date, etc...) et sur le contenu à l'aide des thésaurus Vidal (Indications, Contre-indications, Mises en garde, Précautions d'emploi et Effets secondaires) et d'autres terminologies (CIM10, ATC [27], CISP [28]...).

L'indexation manuelle de la partie thérapeutique consiste à indexer avec des termes des thesaurus de différents types. Voici le détail de l'indexation pour chaque partie du RCP :

- Composition
  Cette rubrique peut contenir des informations sur des précautions d'emploi ou contre-indications qui seront alors indexées respectivement avec le type <PE> et <CI>. Les contre-indications et précautions d'emploi peuvent être signalées par des expressions comme «tenir compte de...» ou «réservé à...».
- Posologie et mode d'administration
  Les informations concernant l'état du patient (exemple : «sujet âgé» ou «insuffisant hépatique») seront indexées avec le type <PE>.
- Données Cliniques
  - Indications thérapeutiques : indexées avec le type <INDIC>
  - Contre-indications : indexées avec le type <CI>. Une contre-indication est typée comme «absolue» ou «relative».
  - Mises en garde et précautions d'emploi : indexées avec le type <PE>.
    On distingue deux types de termes d'indexation : ceux concernant tout patient (terrain physiologique donc mise en garde) et ceux liés à un type de patient (état pathologique particulier donc précaution d'emploi).
    Pour les précautions d'emploi sont répertoriés les termes correspondant à des états patients, physiologiques ou pathologiques, susceptibles de générer des alertes (exemple : Insuffisance rénale, Diabète...).
    Enfin une précaution d'emploi peut être liée à une indication : l'indication est alors considérée comme un état du patient et doit être indexée comme telle.
  - Grossesse et Allaitement : indexé avec le type <CI> ou <PE> selon les cas.
  - Conduite et utilisation de machine : indexée si besoin avec le type <PE>

27. Classification Anatomique, Thérapeutique et Chimique maintenue et publiée par l'OMS
28. Classification Internationale Des Soins Primaires

- Effets indésirables : indexés avec le type ‹EII›. La fréquence d'un effet indésirable peut être précisée : très fréquent, fréquent, peu fréquent, rare, très rare.
- Interactions médicamenteuses : peut contenir des termes à indexer avec le type ‹PE›.
- Surdosage : indexé avec le type ‹EII›

Pour compléter l'indexation, des liens dits «contexte d'application» peuvent être créés. Par exemple, une contre-indication a comme contexte une indication ou un terrain (dictionnaire des conditions) ; une précaution d'emploi a comme contexte une indication.

Comme nous avons pu le voir précédemment, il est possible d'indexer des informations d'une rubrique du RCP dans une rubrique différente (exemple : le terme «contre-indiqué en cas d'intolérance génétique au galactose» issu de la rubrique Précaution d'emploi du RCP sera indexé avec le type contre-indication). L'origine de la rubrique est alors mise en commentaire. Il est également possible, en cas de nécessité, d'indexer une propriété clinique absente du RCP ou de ne pas retenir des termes présents dans le RCP.

L'indexation se fait dans l'ordre du RCP et doit contenir au moins une occurrence de chaque type. Si aucun terme ne convient pour une rubrique, un nouveau terme doit être créé manuellement et validé par le gestionnaire de thésaurus.

Afin de maintenir une homogénéité par famille, avant toute indexation, il est nécessaire de connaître l'indexation des autres spécialités de la même classe thérapeutique ainsi que les spécialités indexées par les indications, contre-indications. . .du même groupe.

En cas de besoin, chaque indexeur responsable de l'indexation d'une famille pharmaco-thérapeutique peut rédiger des règles d'indexation (exemple : pour les AINS [29] : ne pas détailler la liste des indications thérapeutiques citées après «notamment»).

Les autres rubriques non indexées sont intégrées avec l'intégralité des données texte du RCP.

2.4.2.3 Thesaurus Unifié du Vidal (TUV)

Au fil des années et des besoins, l'équipe scientifique du Vidal a créé 4 thesaurus : Indications, Contre-indications, Effets secondaires et Précautions d'emploi. Ces thesaurus permettent de décrire les différentes propriétés pharmacologiques et thérapeutiques des spécialités pharmaceutiques [30]. Ces propriétés sont énoncées dans le RCP correspondant à la spécialité.

29. Les anti-inflammatoires non stéroïdiens.

30. Une spécialité pharmaceutique est un médicament qui a un nom commercial. Une même spécialité pourra être commercialisée éventuellement sous un ou plusieurs noms de marque. Elle fait l'objet d'un enregistrement auprès des autorités de santé, et est vendue à un prix déterminé par un laboratoire pharmaceutique. Sous son même nom de marque, il peut exister différentes formes pharmaceutiques et différents conditionnements spécifiques, chacun faisant l'objet d'un enregistrement spécifique.

Ces thesaurus possèdent des transcodages vers la CIM10, le DRC [31] et la CISP [32].

En 2004 a débuté l'uniformisation de ces 4 thesaurus afin de créér un thesaurus unique : le TUV (Thesaurus Unifié Vidal). Cette unification devrait améliorer les fonctionnalités de recherche et d'alertes dans les produits Vidal, enrichir les connaissances de la base et faciliter la gestion grâce à un seul thesaurus. Il s'agit d'une évolution devant aboutir à la construction d'une ontologie, résultat d'une structuration plus fine des termes et de la création de relations sémantiques entre ces termes.

Dans le TUV, les termes sont hiérarchisés. À chaque position dans la hiérarchie se trouve un code ainsi qu'une formulation préférée et éventuellement des synonymes et des variantes lexicales.

Plusieurs types de termes y sont distingués :

- les termes de référence décrivant les propriétés d'une spécialité pharmaceutique. Ils sont utilisés pour l'indexation des RCP et constituent les anciens thésaurus (8 252 termes préférés et 2 728 synonymes ou variantes lexicales, soit 10 980 termes).
- ces termes de référence peuvent être décomposés en termes élémentaires (au nombre de 1 000 pour le moment) (voir figure 2.7).

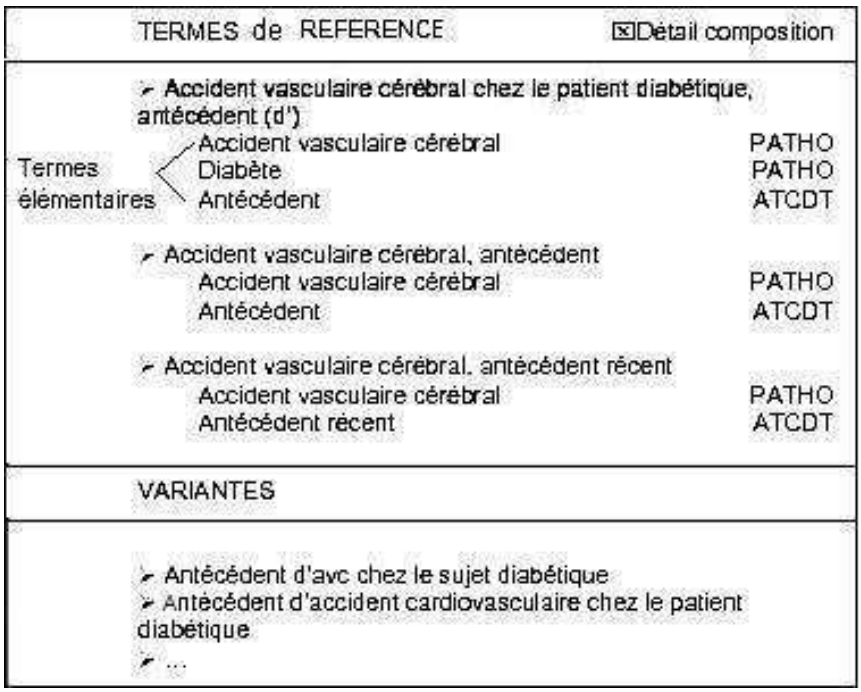

FIGURE 2.7 – Extrait du TUV

Les termes élémentaires peuvent posséder des synonymes et des variantes lexicales. Ces termes élémentaires peuvent être combinés pour former de nouveaux

31. DRC : dictionnaire des résultats de consultation de la SFMG (Société Francaise de Médecine Générale).

32. Classification Internationale des Soins Primaires.

termes de référence (s'ils sont significatifs pour l'indexation des RCP). Les différents types de termes élémentaires sont :
- les états : état primaire ou secondaire («primaire» pour l'état traité, «secondaire» pour l'état pré-existant)
- les compléments (CT) : ce sont des qualificatifs

Chaque terme élémentaire est rattaché à une étiquette sémantique présentant son type et son sens (exemple : le terme élémentaire «sévère» a pour étiquette «CT/NIV-GRAV» qui signifie que le terme est un complément appartenant à la hiérarchie «niveau de gravité», autre exemple, le terme élémentaire «dermatite atopique» a pour étiquette «ETAT/PATHO [Primaire]» ce qui signifie que le terme est un état correspondant à une pathologie primaire).

Tous les termes de référence peuvent être décomposés en un ou plusieurs états et en 0 ou plusieurs compléments (exemple : le terme de référence «Dermatite atopique sévère de l'adulte, traitement de deuxième intention» est constitué des termes élémentaires : «dermatite atopique» (état), «adulte» (état), «sévère» (complément) et «traitement de deuxième intention» (complément)).

Il existe aussi des relations entre les types sémantiques rattachés aux états, telles que «est une complication de».

Ce thesaurus est toujours en court de réalisation, il comporte à ce jour tous les termes de référence et 1 000 termes élémentaires soit 11 980 termes.

2.4.3 Codage de l'information pour les dossiers patients

2.4.3.1 Le codage des dossiers par les professionnels de santé

Les dossiers médicaux papier sont passés progressivement à un dossier électronique du patient dans le début des années 80.

La loi du 31 juillet 1991 portant sur la réforme hospitalière a marqué un tournant. Le PMSI (Programme de Médicalisation des Systèmes d'Information) impose alors aux établissements de santé publics et privés de mettre en œuvre des systèmes d'information [33] capables de recueillir les données concernant leur activité (pathologies traitées par exemple) et leur mode de fonctionnement (exemple : mode de prise en charge) afin de les délivrer à l'État et aux services d'assurance maladie (articles L6113-7 et L6113-8 du code de la santé publique). Ces données sont nécessaires :
- à l'élaboration des cartes sanitaires,
- aux études épidémiologiques (c'est dans cet objectif que la CIM a été initialement élaborée),
- à la détermination des ressources nécessaires à l'établissement (afin de procurer aux établissements de santé les financements adéquats dans le cadre de la tarification à l'activité (T2A)),
- à l'évaluation de la qualité des soins ainsi qu'au contrôle de leur activité et de leur facturation.

33. Un système d'information se compose de l'ensemble des éléments participant à la gestion, au stockage, au traitement, au transport et à la diffusion de l'information au sein d'une organisation.

Des données fausses peuvent les rendre inexploitables et entraîner des problèmes dans le financement de l'hôpital.

Après chaque séjour hospitalier en soins de courte durée (médecine, chirurgie, obstétrique et odontologie (MCO)), un bref compte rendu de l'hospitalisation du patient doit être produit, il est composé d'un compte rendu de séjour dactylographié (voir figure 2.8 pour un exemple de compte rendu de séjour). Celui-ci permet de communiquer de facon précise et concise l'état du patient afin que chaque médecin consultant le dossier puisse avoir une vue synthétique de l'évolution de la maladie au travers des étapes importantes du traitement du patient. Ils peuvent être plus ou moins structurés, allant d'une entête suivie de quelques rubriques à remplir (exemple : Motif d'hospitalisation, Antécédents, Examens cliniques, Traitement de sortie, Conclusion) à un formulaire pré-établi où il suffit de cocher des cases. Le contenu est laissé aux bons soins du rédacteur, il n'y a pas de règles précises ni de vérification *a posteriori*. Il peut être rédigé à l'aide d'un éditeur de texte pour les plus simples ou d'une interface dédiée pour les formulaires.

Après chaque séjour, accompagné du compte-rendu d'hospitalisation, le médecin doit produire le résumé de sortie standardisé (RSS). Il peut être réalisé à partir du compte-rendu d'hospitalisation ou de manière indépendante. Il est composé d'autant de résumés d'unité médicale (RUM) que d'unités médicales fréquentées par le patient pendant son séjour dans le secteur MCO. Ce résumé doit obligatoirement contenir un certain nombre d'informations administratives et médicales (répertoriées dans l'arrêté du 27 et 28 février 2006) qui sont codées pour permettre des traitements informatiques. Les informations administratives pour l'identification du séjour du malade sont le sexe, la date de naissance, le code postal, la date d'entrée et de sortie, le nombre de séances ainsi que les identifiants de séjour, de l'unité médicale, et de l'établissement. Les informations médicales recueillies dans le RSS sont :

- les diagnostics : un diagnostic principal [34], un (des) diagnostic(s) relié(s) [35] et un (des) diagnostic(s) associé(s) [36] significatif(s). Les diagnostics sont codés selon la CIM10 (voir section 2.4.3.2) (voir figure 2.9 pour un exemple de codage de séjour). Certains services utilisent des normes spécifiques à leur discipline imposant un transcodage *a posteriori* en CIM10. Les diagnostics sont codés selon des règles très strictes (les consignes sont disponibles sur le site de l'ATIH [37]) sous peine de ne pas passer les contrôles de l'assurance maladie.
- les actes médicaux sont codés selon la plus récente version en vigueur de la CCAM (voir section 2.4.3.3). Le codage d'un acte CCAM associe obligatoirement son code principal, la phase, l'activité, le nombre d'exécutions de l'acte pendant le séjour. Les autres codes sont facultatifs (extension documentaire, modificateurs, remboursement exceptionnel, etc...). De la même facon les actes

34. Diagnostic ayant mobilisé l'essentiel de l'effort médical et soignant au cours du séjour hospitalier.

35. Tout diagnostic permettant d'éclairer le contexte pathologique, essentiellement lorsque le diagnostic principal n'est pas en lui-même une affection. Le plus souvent, le diagnostic relié correspondra à la maladie causale.

36. Tout autre diagnostic du patient.

37. Accessible ici `http://www.atih.sante.fr/index.php?id=0006500001FF`

Entête Rouen le XX/XX/XXXX

COMPTE-RENDU D'HOSPITALISATION :

Mr XXX Né(e) le : XX/XX/XXXX N° dossier : XXX

Date d'entrée : XXX Date de sortie : XXX Médecin Traitant : XXX

Motif d'hospitalisation : Douleurs thoraciques.

ANTECEDENTS ET HISTOIRE DE LA MALADIE : Légionellose en 2001. Tassement vertébral. Hernie ombilicale. Hernie hiatale. Hypertension artérielle. Hypercholestérolémie. Diabète de type II. Tabagisme à 60 paquets année non sevré.

Le 18/10/2004 apparition d'une douleur thoracique rétrosternale constrictive en étau, au repos, irradiant dans l'épaule et la machoire, durant 5 mn. Le patient consulte en urgence en cardiologie mais refuse l'hospitalisation et repart chez lui. Le 22/10/2004 vers 1 H du matin, récidive de la douleur motivant l'appel du SAMU.

EXAMEN CLINIQUE : 67 kg pour 1,66 m. Tension artérielle 15/8. Bruits du coeur réguliers. Pas de souffle. Pas de frottement. Pas de signe d'insuffisance venticulaire droite ou gauche. Pouls périphériques tous perçus ;

ELECTROCARDIOGRAMME : Rythme sinusal à 72/mn. PR normal. QRS fins. Axe gauche. Onde T négative en D3.

RADIO THORACIQUE : Non faite.

BIOLOGIE : Troponine 0 à plusieurs reprises. Cf. Feuille ci jointe.

évoluTION : Le patient a présenté des récidives douloureuses dans le service de quelques secondes, sans modification ECG ou sans élévation de la troponine au cours de l'hospitalisation.

Epreuve d'effort maquillée sous maximale à 70%, négative électriquement et douteuse cliniquement.

Il est décidé de laisser sortir Mr X avec un traitement médical et de le reconvoquer pour une épreuve d'effort démaquillée à distance.

AU TOTAL : Douleurs thoraciques d'allure angineuse sans modification ECG, sans élévation de la troponine.

Découverte d'une hypercholestérolémie et d'un diabète de type II.

Il devra consulter une diététicienne pour régime diabétique et hypocholestérolémiant. Epreuve d'effort démaquillée en externe.

TRAITEMENT DE SORTIE : KARDEGIC 160 1/j - NITRIDERM 10 1/j - MOPRAL 20 1/j - TAHOR 10 1/j

FIGURE 2.8 – Extrait d'un compte-rendu d'hospitalisation dans le secteur cardiologie de l'hôpital de Rouen

sont codés selon des règles très strictes.

– et d'autres données comme le poids et l'âge gestationnaire pour les nourrissons, l'indice de gravité simplifié [38] (IGS II) et des données documentaires associées.

La codification (ou indexation) est réalisée dans la plupart des hôpitaux manuellement par les médecins en charge du patient ou les secrétaires du service qui n'ont pas vu le patient et qui n'ont pas de connaissances médicales approfondies (ce qui

38. L'indice de gravité est calculé en additionnant des scores. Cet indice permet de prédire le risque de décès à l'admission dans une unité de soins intensifs ou de réanimation ou de surveillance continue. Le risque de mortalité est d'autant plus important que l'indice est élevé.

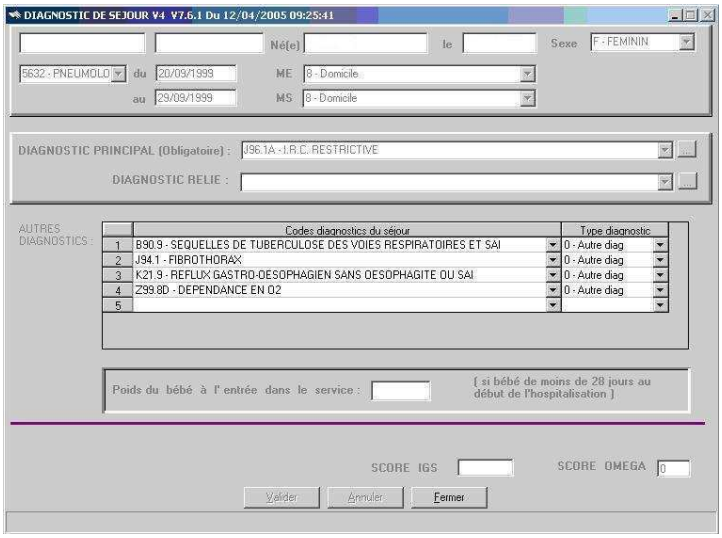

FIGURE 2.9 – Codage CIM10 du compte-rendu d'hospitalisation visualisé à partir du logiciel CDP2, le logiciel de dossier patient électronique créé et utilisé par le CHU de Rouen

peut poser des problèmes de validité des données).

Les informations recueillies permettent par un traitement automatique de classer le RSS dans un GHM (Groupe Homogène de Malades). Un GHM remplit certains critères (diagnostics, actes etc...) liés à un coût, ce qui permet en pratique pour un séjour et pour un malade de connaître le coût associé pour l'établissement. Le codage médico-économique se limite le plus souvent aux codes diagnostiques et actes permettant la classification en GHM.

Les informations recueillies dans le cadre du PMSI sont protégées par le secret professionnel. Les RSS sont ainsi anonymisés en RSA (résumé de sortie anonyme) avant d'être transmis. La transmission des données à l'Agence Régionale de l'Hospitalisation (ARH) se fait mensuellement.

2.4.3.2 Classification statistique Internationale des Maladies et des problèmes de santé connexes 10ème édition (CIM10)

L'origine de la CIM remonte aux années 1850, avec the International List of Causes of Death de W. Farr. Ces travaux reprenaient entre autres ceux de J.Graunt datant de 1700. Elle fut adoptée par the International Statistical Institute en 1893, grâce aux travaux de J. Bertillon qui publie la Nomenclature Internationale des Causes de Décès. Puis, elle a été adoptée par l'Organisation Mondiale de la Santé (OMS) en 1948 [19550]. La neuvième révision CIM9 et son adaptation clinique CIM9-MC ont été rendues plus aptes à servir en matière de statistiques relatives à l'évaluation des soins médicaux. Elle est ainsi utilisée pour le codage médico-économique des dossiers patients à des fins statistiques et budgétaires dans le cadre du PMSI en France. Puis a succédé la 10ème révision [39], la Classification statistique

39. La CIM10 est disponible sur papier en librairie, en version électronique ou sur le site de l'ATIH

internationale des maladies et des problèmes de santé connexes en 1993 [19993], plus adaptée aux statistiques d'assurance maladie et au paiement centralisé des services médicaux. Elle a remplacé en France la CIM9 pour le PMSI (alors que la neuvième révision est toujours utilisée aux États-Unis).

Des extensions de codes de la CIM10 ont été créées pour le PMSI par le PERNNS (Pôle d'Expertise et de Référence National des Nomenclatures de Santé) et l'ATIH (Agence Technique de l'Informatisation sur l'Hospitalisation) pour apporter plus de précisions à certains codes et améliorer le classement en GHM.

Plusieurs langues sont disponibles pour la CIM10 : francaise (dont une version pour la suisse), allemande, anglaise (dont une australienne), néerlandaise, espagnole.

La CIM10 comporte 3 volumes [40], la table analytique qui contient la classification en elle-même, le manuel d'utilisation et l'index alphabétique. La CIM10 est ordonnée en une hiérarchie à héritage simple. Cela signifie que toute entité hiérarchique possède un unique père. La hiérarchie de la CIM10 a jusqu'à 6 niveaux. La CIM a été construite à l'origine de facon statistique, la granularité des différentes branches de la hiérarchie est fonction de la fréquence ou de la gravité des maladies.

La CIM10 est partitionnée en 21 chapitres couvrant l'éventail complet des états morbides, classés par appareil fonctionnel [41] et associés à une lettre (exemple : la lettre E est associée au chapitre «Maladies endocriniennes, nutritionnelles et métaboliques»). Les chapitres sont divisés en groupes, eux-mêmes divisés en sous-groupes composés de catégories à 3 caractères (code composé de 3 caractères) et de sous catégories à 4 caractères, englobant l'ensemble des termes CIM10 (voir figure 2.10). Les catégories à 3 caractères représentent l'unité diagnostique signifiante de base c'est-à-dire le niveau minimum de codification [42]. Enfin des subdivisions peuvent apparaître de manière facultative dans certains chapitres.

Toute position dans la hiérarchie CIM10 est représentée par :

- Un seul code CIM10. Les codes pouvant contenir jusqu'à 5 caractères (ou digits) se décomposent de la manière suivante :
  - Le premier caractère est une lettre majuscule variant de A à Z, (sauf la valeur U). Celui-ci est associé au chapitre.
  - Les caractères 2 et 3 sont numériques de 00 à 99 et désignent une catégorie.
  - Le caractère 4 est toujours précédé d'un point, il est numérique de 0 à 9 et désigne une sous-catégorie.
  - Le caractère 5 est numérique de 0 à 9 et désigne une subdivision.

  Les codes des extensions peuvent comporter des lettres en guise de 5ème caractère et des «+» (exemple : M45.+4, S82.00, E10.8A). Les chapitres, groupes et sous-groupes sont représentés par un code de type intervalle entre les deux catégories les plus extrêmes qu'ils contiennent. Par exemple le chapitre 4 est

`http://www.atih.sante.fr/`.

40. Livres

41. En anatomie, un appareil est un ensemble d'organes dont le fonctionnement concourt à une tâche commune complexe (exemple : appareil digestif).

42. Toutefois de nombreux pays exigent le niveau suivant à 4 caractères comme niveau minimum de codification (c'est le cas de la Suisse par exemple).

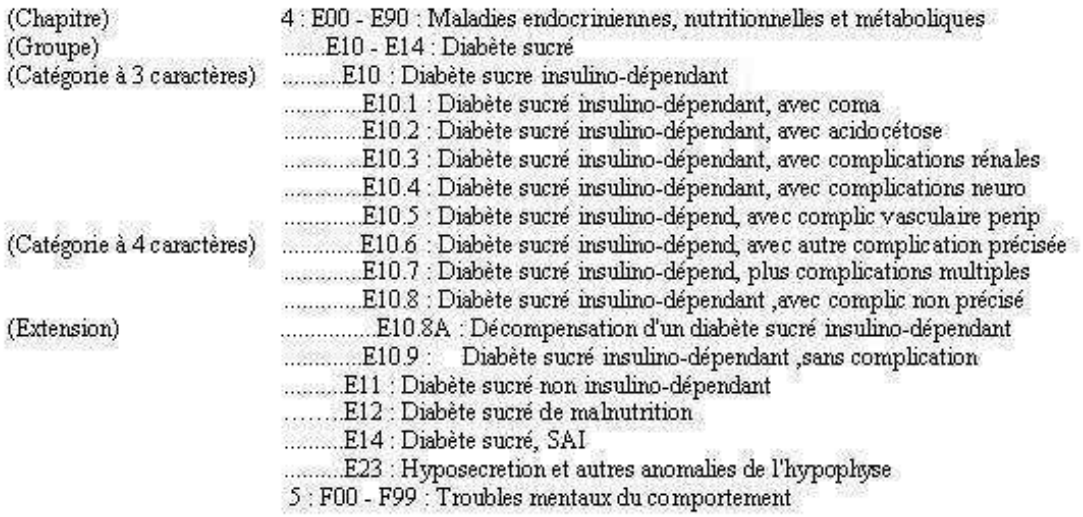

FIGURE 2.10 – Extrait de la classification CIM10

codé par (E00-E90). Tout code possible de la CIM10 (entre A00.00 et Z99.99) n'est pas un code valide. Ainsi, la CIM10 inclut 19 155 codes alphanumériques et 31 222 termes.

- un terme dit «systématique».
- des descripteurs éventuels qui décrivent le terme systématique auquel ils sont associés [43] (voir figure 2.11).

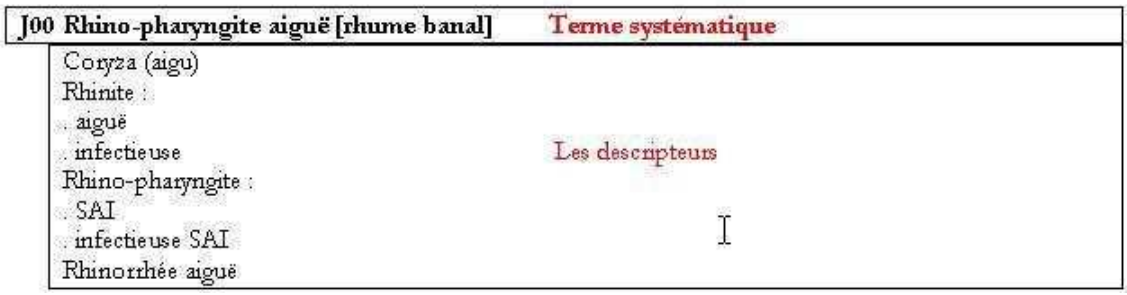

FIGURE 2.11 – Extrait de la classification CIM10 présentant un terme systématique accompagné de ses descripteurs.

- des références : certains termes systématiques peuvent faire référence à des tableaux ou textes. Exemple : le terme systématique «chutes» de code W00-W19 fait référence au terme «tableau des codes du lieu de l'événement», tableau contenu dans le livre de la CIM10 volume 1.
- des inclusions et exclusions : à chaque niveau (chapitre, catégorie, sous catégorie) la CIM10 peut indiquer des inclusions (indiquées par le terme «comprend») ou des exclusions (indiquées par le terme «à l'exclusion de») permettant d'orien-

43. La CIM11 comprendra pour chaque terme une définition.

ter vers une autre partie de la classification (le code de renvoi se trouvant entre parenthèses, voir figure 2.12). Le code excluant et le code exclu sont alors liés par un libellé d'exclusion.

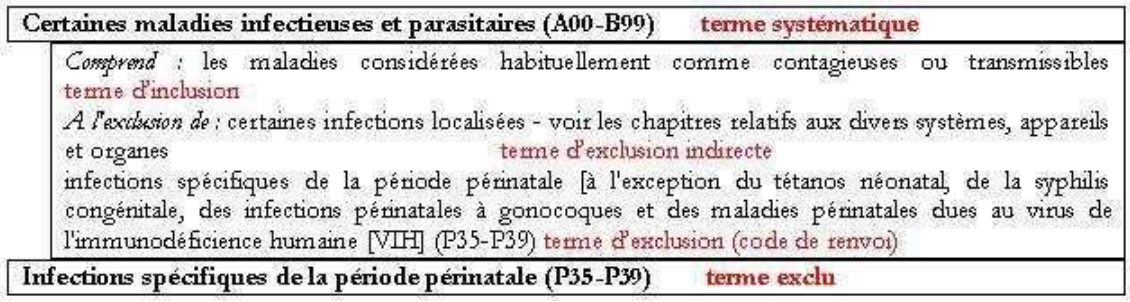

FIGURE 2.12 – Extrait de la classification CIM10 présentant pour un terme systématique les exclusions et inclusions auquel il renvoi.

La CIM10 a prévu des liens horizontaux entre termes de sa hiérarchie, appelés appariements dagues et étoiles ou système de la dague et de l'astérisque. Il permet d'attribuer deux codes à des diagnostics lorsque ceux-ci contiennent des informations concernant à la fois une maladie généralisée initiale et une manifestation localisée à un organe donné qui en est elle-même un problème clinique. Le code primaire est utilisé par la maladie initiale (dague +) et un code supplémentaire facultatif, pour la manifestation (astérisque *). La figure 2.13 montre un exemple.

FIGURE 2.13 – Extrait de la classification CIM10 présentant un exemple d'astérisque systématique.

2.4.3.3 La Classification Commune des Actes Médicaux (CCAM)

La CCAM [Rodrigues05] est le référentiel des actes médicaux qui remplace, pour les médecins, la Nomenclature Générale des Actes Professionnels (NGAP [44]) en secteur libéral, et le Catalogue Des Actes Médicaux (CDAM [45]) en secteur hospitalier français. Elle permet la tarification des actes en médecine libérale.

Élaborée par la CNAMTS (Caisse Nationale d'Assurance Maladie des Travailleurs

44. La NGAP est la nomenclature de médecine ambulatoire.

45. Le CDAM, publié en 1985, a été élaboré par des comités d'experts médicaux coordonnés par la Direction des Hôpitaux. Il répondait à deux objectifs : identifier les actes réalisés pendant le séjour du patient et mesurer la consommation en ressources humaines et matérielles pour réaliser cet acte.

Salariés) et l'ATIH, en étroite collaboration avec les sociétés savantes, la CCAM [46] a été créée afin d'obtenir une liste unique d'actes codés, commune aux secteurs public et privé pour les professionnels de la santé afin de garantir la cohérence des systèmes d'information et de satisfaire les professionnels par l'utilisation d'un seul outil. Elle est destinée à décrire plus précisément chaque acte, à servir de base à la tarification en secteur libéral (cabinets et cliniques) et à l'allocation de ressources aux établissements publics dans le cadre de la tarification à l'activité (T2A).

Elle possède un lien sémantique avec la CIM10, créé par Jacques Chevallier [Chevallier03]. Nous nous sommes intéressée dans cette thèse à la version 6, la version disponible à l'époque de nos premières implémentations. La version la plus récente est la version 13 (9 999 codes) applicable au 28/12/2007. Cette terminologie est peu stable, des mises à jour sont produites tous les 2 voire 3 mois.

La CCAM est une classification purement francaise même si sa structure intéresse de nombreux autres pays tels que le Japon. L'équivalent aux États-Unis de cette classification est la Current Procedural Terminology (CPT). L'équivalent au Canada est la CCI, la Classification Canadienne des Interventions.

Le classement de la CCAM correspond à une logique médicale et se fait par grand appareil et non par spécialité [47]. La CCAM est une hiérarchie à héritage simple organisée en 19 chapitres. Les 17 premiers chapitres sont scindés en deux parties : la première concerne les actes diagnostiques rangés par grande technique puis par organe, la seconde concerne les actes thérapeutiques classés par organe puis par action ; le chapitre 18 regroupe les gestes complémentaires ; le chapitre 19 prend en compte les adaptations pour la CCAM transitoire.

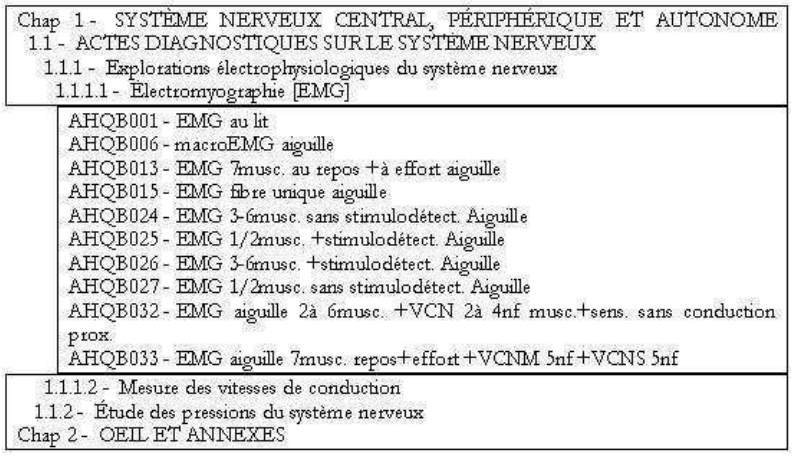

FIGURE 2.14 – Extrait du chapitre 1 de la CCAM

La CCAM est fondée sur le principe de l'acte global : chaque libellé comprend implicitement l'ensemble des gestes nécessaires à la réalisation de l'acte. De plus les

46. La terminologie est disponible sur le site de l'assurance maladie (navigation, recherche et téléchargement sur le site de l'assurance maladie, `http://www.codage.ext.cnamts.fr/codif/ccam/index prsentation.php?p site=AMELI`) ou téléchargeable sur le site de l'ATIH `http://www.atih.sante.fr:80/?id=0003100027FF`

47. Domaine de formation (exemple : cardiologie ou pneumologie.)

libellés sont non ambigus c'est-à-dire sans possibilité d'interprétations divergentes. Elle est aussi bijective c'est-à-dire qu'à un libellé correspond un code et un seul et réciproquement (voir figure 2.14).

La CCAM version 6 comprend 7 389 codes. À chaque libellé de dernier niveau de la CCAM correspond un code à 7 caractères alphanumériques : les 4 premiers sont signifiants (topographie, action, voie d'abord et/ou technique), les 3 derniers constituent un compteur séquentiel.

AA — AA — NNN
Topographie Action Voie d'abord [48] et/ou technique Compteur

- Le premier code constitue le codage du système (exemple : «système respiratoire» (G)).
- La deuxième lettre constitue le codage de l'organe ou de la fonction (exemple : «plèvre» (GG)).
- La troisième lettre correspond au codage de l'action principale du libellé (exemple : «évacuer» (J)).
- La quatrième lettre code le mode d'accès ou la technique utilisée (exemple : «abord ouvert» (A)).
- Chaque code à 4 caractères est affecté d'un compteur à 3 chiffres, pour différencier les actes ayant même code anatomique, même code d'action et même code de voie d'abord ou de technique (exemple : «Évacuation de collection de la cavité pleurale, par thoracotomie sans résection costale»(GGJA002) et «Évacuation de collection de la cavité pleurale, par thoracotomie avec résection costale»(GGJA004)).

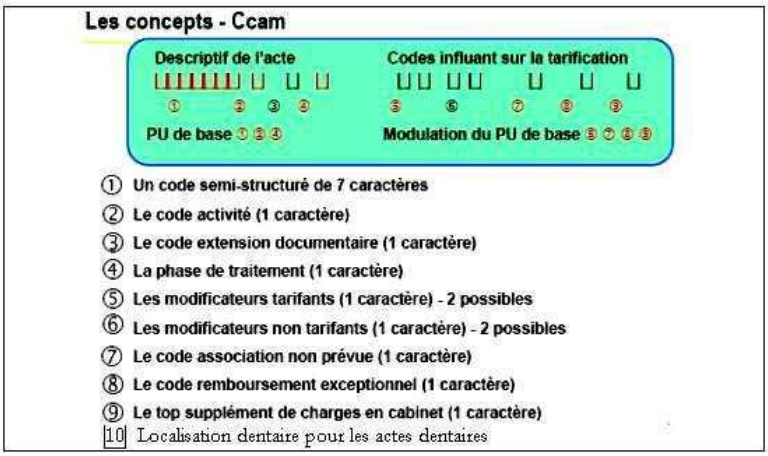

FIGURE 2.15 – Structuration du code CCAM

Des caractères supplémentaires aux codes peuvent être ajoutés, comme le montre la figure 2.15, ceux-ci permettent de :

- décrire l'activité : permet de différencier et énumérer les gestes réalisés au cours d'un même acte par des intervenants différents (valeur de 0 à 5).

48. Voie d'accès pour un acte ou une exploration chirurgicale.

- préciser l'extension documentaire : une lettre qui permet de donner un niveau de détail supplémentaire mais non utile à la tarification (10 valeurs possibles). Exemple : pour le terme «dilatation intraluminale d'une branche de l'aorte abdominale à destinée digestive avec pose d'endoprothèse, par voie artérielle transcutanée» (EDAF005), nous avons entre autres les codes documentaires : «tronc iliaque» (F) et «artère gastrique gauche» (G).
- préciser la phase de traitement : pour distinguer les différentes phases d'un acte en terme de coût et de séjour d'hospitalisation (exemple pour le terme «reconstruction d'un tendon de la main par transplant libre, en deux temps» (MJMA006), il existe deux phases : «reconstruction de la gaine fibreuse digitale avec pose de prothèse provisoire, par abord direct avec ou sans réfection des poulies» (MJMA006 1 1) et «transplant libre de tendon de la main» (MJMA006 1 2)).
- Enfin des codes influant sur la tarification peuvent être juxtaposés :
  - l'application des codes modificateurs indique les circonstances particulières de réalisation de l'acte et peut entraîner une majoration du coût du séjour.
  - un code association qui permet de signaler des associations d'actes non prévues.
  - un code remboursement exceptionnel.
  - un code supplément pour un acte en cabinet (code (C)).

Chaque code est suivi de son tarif en euros et de précisions tarifaires, de caractéristiques générales et de précisions sur le codage et de plus de 20 autres critères divers (voir site de l'assurance maladie).
Plusieurs actes peuvent être associés (4 au maximum). Toutefois, il existe des associations d'actes interdites, elles sont identifiées et listées.

2.4.3.4 La Nomenclature systématique de Médecine humaine et vétérinaire (SNOMED)

Dix ans ont été nécessaires au comité sur la nomenclature et la classification des maladies créé par le College of American Pathologists (CAP) en 1955 pour aboutir à la publication de la SNOP (Systematized Nomenclature of Pathology) une nomenclature fonctionnelle pour les pathologies. En 1973, le Dr Côté fait évoluer la SNOP vers la SNOMED (Systematized Nomenclature of Medicine) [Côté72] qui devient en 1993 [Côté93] la SNOMED version 3.5, appelée aussi SNOMED Internationale, nomenclature pluri-axiale couvrant tous les champs de la médecine et de la dentisterie humaine, ainsi que de la médecine vétérinaire.

Un remaniement de la SNOMED 3.5 avec ajout de descriptions formelles a été effectué afin de créér une terminologie de référence, la SNOMED RT [49] (Reference Terminology) en 1998 [Spackman97] se rapprochant davantage d'une ontologie formelle.

Enfin la SNOMED CT (Clinical Terms) est le résultat de la fusion de la SNOMED RT version 1.1 et des NHS Clinical Terms version 3 (Read Codes) du Royaume-Uni.

49. Ce projet est issu d'une collaboration entre le College of American Pathologists, la société Kaiser Permanente (Health Management Organization) et la Mayo Clinic.

La SNOMED CT [Ame06] est concue pour simplifier la saisie et la recherche de concepts cliniques au sein de systèmes d'information électroniques et pour faciliter leur communication. Son objectif est de rendre les connaissances de soins de santé plus accessibles à toutes les spécialités médicales. Elle contient plus de 400 000 codes, plus d'un million de descriptions et un réseau sémantique constitué de 1 500 000 relations sémantiques que la SNOMED 3.5 ne possède pas. La SNOMED CT est actuellement la nomenclature officielle de la médecine clinique aux États-Unis et dans d'autres pays anglosaxons (Angleterre, Australie, Nouvelle Zélande, Royaume-Uni, Australie) et utilisée dans 38 pays à moindre échelle (Allemagne, Portugal, Suède, Chine, Lituanie etc...), elle est traduite en anglais, allemand et espagnol. La traduction francaise de la SNOMED CT devrait bientôt démarrer grâce au SDO. Elle possède également plus de 10 transcodages vers d'autres terminologies (CIM10, OPCS 4.2, etc...).

La SNOMED 3.5 a été la seule traduite en francais. Cette traduction, réalisée par l'équipe du Centre de recherche en diagnostic médical informatisé (CRDMI) à Sherbrooke, s'est terminée en 2006 en partie grâce au projet VUMeF (déjà abordé dans le chapitre 1). Elle est actuellement traduite en 11 langues (dont francais, espagnol, portugais, chinois, japonais et turc) et renferme des concepts médicaux normalisés. Elle comporte un axe classificatoire qui permet de faire le lien avec la CIM (axe D). La traduction a ainsi été accompagnée par le transcodage en CIM-10. La France a acquis les droits pour cette terminologie en 2007 pour l'indexation des dossiers patients électroniques hospitaliers.

La SNOMED 3.5 est multi-axiale et multi-domaine. Elle comporte onze axes orthogonaux, chaque axe recense les termes d'un sous-domaine de la médecine (exemple : D (diagnostics), T (topographie), M (morphologie) voir figure 2.16). Chaque axe est hiérarchisé en fonction de la spécialisation des termes, qui sont reliés par des relations d'hypo/hyperonymie [50] et méro/holonymie [51]. Par exemple, le concept A-81000 (≪radiation, SAI ;rayonnement ionisant≫) est plus général que le concept A-81020 (≪radiation électromagnétique≫) et que le concept A-81050 (≪rayon-X≫) ; le concept T-61083 (≪salive ; sécrétion de la glande salivaire≫) désigne une partie de T-61000 (≪glande salivaire, SAI≫).

Dans chaque axe, chaque concept est représenté par une série de termes au sein de laquelle on peut distinguer une formulation préférée et des synonymes. Chaque concept de la SNOMED 3.5 recoit un code alphanumérique unique (par exemple, T-01414). Ici les codes reflètent la hiérarchie des termes auxquels ils sont associés : par exemple, A-81000 est plus général (contient moins de chiffres) que A-81020. Le terme préférentiel possède la classe 01, les autres termes la classe 02, 03 ou 05 (voir la figure 2.17 pour un exemple).

Il est possible de combiner des termes provenant d'axes différents (les relations transversales) ce qui permet de composer un concept complexe en combinant des concepts élémentaires pris dans ces axes. La base conceptuelle du codage pluri-

50. Un hyponyme est un mot dont le sens est hiérarchiquement plus spécifique que celui d'un autre.

51. Relation partie/tout.

Axe	Nom de l'axe	NB de Termes
T	Topographie	13 528
M	Morphologie	6 171
F	[dys]Fonctions	20 587
A	artefacts, activités physiques	1 686
L	êtres vivants	26 325
C	produits chimiques	15 940
J	Métiers	2 303
S	contexte social	1 110
D	Diagnostics	42 492
P	Actes	31 980
G	Qualificatifs	1 595
X		363
Total		164 180

FIGURE 2.16 – Les axes de la SNOMED 3.5

axial repose sur la combinaison d'un site anatomique, d'une altération en ce site, d'une cause lorsqu'elle est connue, des effets physio-pathologiques, des circonstances d'apparition et des actions diagnostiques ou thérapeutiques entreprises. L'axe des qualificatifs et termes relationnels (G) contient des concepts supplémentaires servant à qualifier ces concepts ou à préciser leurs liens dans le concept complexe. Par exemple, une «appendicite aiguë» pourra être représentée par la combinaison des concepts «inflammation, SAI»(M-41000), «aigu»(G-A231), «dans»(G-C006), «appendice vermiculaire, SAI»(T-59200), ces termes sont reliés par une relation dite «de référence»au concept «appendicite aiguë».

Code Classe	Terme	Références
D0-10430 01	pemphigoïde, SAI	T-01000
D0-10430 02	pemphigus bénin, SAI	T-01000
D0-10431 01	pemphigoïde bulleux	T-01000, M-51551, M-36760
D0-10432 01	pemphigus bénin des muqueuses	T-00400, M-43000, G-C009, T-AA000, F-01250

Concepts référencés :

« *peau, SAI ; cutané* » (T-01000), « *acantholyse* » (M-51551), « *ampoule, SAI* » (M-36760) « *muqueuse* » (T-00400), « *inflammation chronique, SAI* » (M-43000), « *sans* » (G-C009) « *oeil, SAI* » (T-AA000), « *symptôme, SAI* » (F-01250).

FIGURE 2.17 – Termes, synonymes et références dans la SNOMED 3.5

2.5 Aide à l'indexation

Nous allons étudier les processus d'aide à l'indexation qui peuvent assister les indexeurs humains dans leurs tâches quotidiennes d'indexation précédemment décrites.

2.5.1 Apports de l'indexation automatique et semi-automatique

L'automatisation des tâches d'indexation a un réel intérêt dans un objectif d'aide à l'indexation. Dans la majorité des cas, l'indexation se fait manuellement avec quelques aides informatiques sous forme de formulaires de saisie ou de logiciels d'aide à la navigation. Dans ce contexte, l'automatisation de la tâche d'indexation, de la lecture du document à la proposition d'indexation, serait une aide précieuse.

2.5.1.1 L'indexation automatique

Une indexation produite de manière automatique est plus régulière qu'une indexation produite manuellement. En effet, la variabilité inter-individuelle liée aux indexeurs est alors inexistante puisque face aux mêmes données le programme informatique donnera toujours la même réponse. Elle s'adapte aussi plus facilement aux mises à jour des terminologies. L'indexeur humain habitué à une version aura plus de difficultés à passer à la version suivante alors qu'il suffit simplement de remplacer les données dans la base de données du programme pour qu'elles soient automatiquement prises en compte. Enfin, elle est capable de traiter des masses très importantes de documents en peu de temps à l'inverse de l'indexation humaine. Par exemple pour nos terminologies, l'indexeur doit choisir un ou plusieurs termes parmi une liste de 7 000 à 110 000 termes pour les faire correspondre à la notion qu'il a repérée dans le document. Par exemple, pour indexer une recommandation de bonne pratique, un indexeur CISMeF met en moyenne 1 heure. Les coûts humains sont très élevés d'où l'intérêt de disposer d'outils d'indexation automatiques.

En revanche, l'indexation automatique est plus exhaustive, les programmes informatiques n'ont encore qu'une capacité de synthèse limitée. Ce type d'indexation est aussi sujette aux erreurs dues aux ambiguïtés de polysémie dans les textes [Chartron89].

Face à des volumes importants de documents électroniques à traiter, ce qui est le cas dans nos trois tâches d'indexation, l'indexation automatique serait la méthode la plus appropriée, encore faut-il que celle-ci atteigne une qualité d'indexation équivalente à l'indexation humaine.

2.5.1.2 L'indexation semi-automatique

L'indexation semi-automatique consiste à indexer le document par un programme informatique qui propose au préalable à l'indexeur une indexation, charge à lui de la compléter, de la modifier et enfin, de la valider [Chaumier92].

Il existe différentes méthodes, la première consiste à appliquer dans un premier temps le programme informatique puis lors de l'indexation manuelle, l'indexeur humain peut avoir accès à la proposition d'indexation automatique. Celle-ci peut être considérée comme «valide» ou «à valider». Si elle est valide l'indexeur devra éliminer les termes qu'il ne souhaite pas voir apparaître et compléter la liste avec d'autres termes pour créér l'indexation finale. Si elle est «à valider» l'indexeur devra

sélectionner les termes adéquats et compléter la liste avec d'autres termes pour créér l'indexation finale.

Une seconde méthode consiste à reformuler manuellement le document d'origine afin que les expressions deviennent faciles à analyser pour le programme. Le programme est alors lancé sur le document modifié et finalement l'indexeur valide l'indexation obtenue.

L'indexeur humain peut aussi sélectionner au préalable les portions de texte qu'il veut voir traiter par la machine afin de rendre les traitements plus rapides et diminuer le bruit pouvant être généré par l'indexation automatique.

L'indexation produite possède l'ensemble des qualités de l'indexation humaine et automatique (gain de temps par rapport à l'indexation humaine, désambiguïsation, qualité de l'indexation produite, mise à jour, variabilité faible) sans les défauts.

2.5.2 Méthodes d'évaluation d'outils d'indexation automatique et semi-automatique

Plusieurs critères d'évaluation peuvent être envisagés (voir figure 2.18).

La consistance de l'indexation vise à apprécier la concordance entre des indexations proposées pour un même document par deux indexeurs ou deux méthodes d'indexation différentes [Rolling80]. Idéalement, deux indexeurs différents devraient produire la même indexation pour un même document (consistance inter-indexeur) et un même indexeur devrait produire la même indexation pour un même document à deux moments donnés (consistance intra-indexeur).

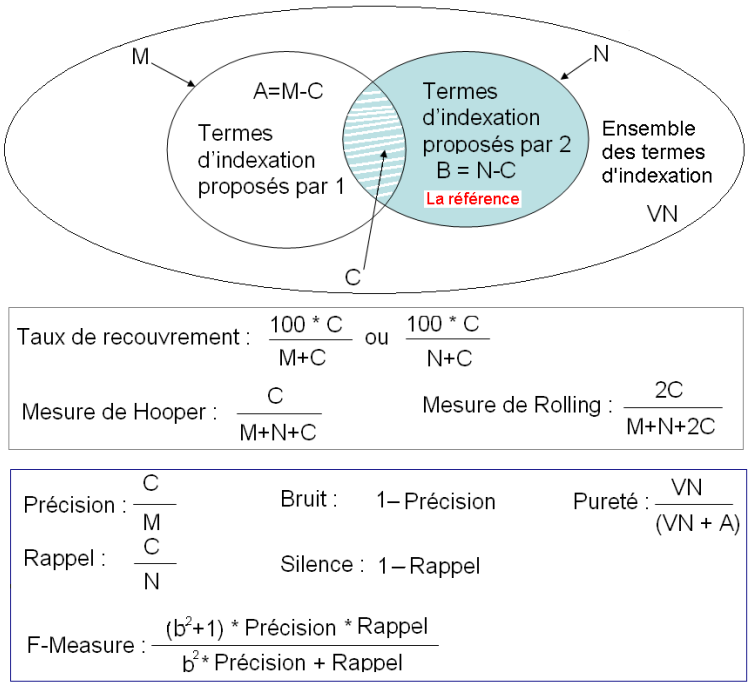

FIGURE 2.18 – Évaluation de l'indexation produite : les mesures de consistances

Plusieurs mesures de consistance existent :

- la mesure de Hooper évalue la proportion de termes proposés par deux indexeurs à la fois, sur l'ensemble des termes proposés par l'un ou l'autre des indexeurs
- la mesure de Rolling accorde un poids supplémentaire aux descripteurs témoignant d'un consensus entre les deux indexeurs
- le taux de recouvrement permet d'évaluer le taux d'accord entre deux listes.

[Berrios02] ont montré que la consistance est meilleure pour un vocabulaire contrôlé.

Il est également possible d'évaluer la qualité d'une indexation, en comparant cette indexation par rapport à une indexation produite par un indexeur expert prise comme référence (ou «gold standard»). Plusieurs mesures sont associées (voir figure 2.18) :

- la précision qui est le rapport du nombre de termes pertinents sur le nombre total de termes proposés.
- le rappel qui est le rapport du nombre de termes pertinents proposés sur le nombre total de termes pertinents.
- la F-mesure qui est une moyenne pondérée harmonique de la précision et du rappel [vanRijsbergen79]. Un paramètre supplémentaire a été introduit par D. Nakache [Nakache05] pour ajouter un poids supplémentaire à la précision ou au rappel selon la tâche que l'on veut évaluer.
- le silence pour évaluer la proportion de termes n'ayant pas été extraits (faux négatifs).
- le bruit pour évaluer la proportion de termes erronés extraits par le système (faux positifs).
- la pureté pour évaluer la proportion d'erreurs d'indexation (extraction d'un terme erroné) évitées par le système [Soergel88].

Pour le résultat de l'indexation automatique le bruit et le silence ont une importance considérable. Du bruit entraînera une perte de temps : pour l'utilisateur qui cherche une réponse parmi un ensemble de documents non pertinents et pour le médecin qui sera distrait par des alertes qui n'ont pas lieu d'être pouvant même entraîner des erreurs de décision. Le silence aboutit à l'impossibilité pour l'utilisateur de retrouver un document pertinent alors que celui-ci aurait dû être proposé et pour le médecin à une absence d'alerte en cas de risque pour le patient lors de sa prescription.

Pour un outil d'indexation semi-automatique[52], le bruit et le silence vont entraîner une perte de temps pour le médecin qui utilise l'outil. Selon le type d'outil, le bruit va obliger l'utilisateur à éliminer ou à préciser les termes non pertinents ou rendre plus difficile la reconnaissance des bons termes d'indexation. Le silence va l'obliger à ajouter les termes manquants.

La qualité peut aussi être évaluée par la validation de l'indexation par un indexeur expert (jugement subjectif de la pertinence des mots clés sélectionnés pour l'indexation ou des documents retournés pour la recherche d'information).

Un des principaux problèmes de ce genre d'évaluation est qu'il n'existe pas d'indexation de référence universelle [Lancaster91]. L'indexation humaine d'un expert

52. Outil proposant une indexation à l'indexeur humain qui doit alors la réviser.

est souvent prise comme référence alors qu'un même document peut être indexé par des ensembles différents de termes qui seront tous corrects. Dans le cadre de groupes d'indexeurs où la tâche d'indexation rencontre une consistance inter-indexeur faible (ce qui est souvent le cas [Funk83a]), la qualité de l'indexation produite automatiquement est souvent sous-estimée. Des études ont été menées afin de proposer des solutions. Une première solution est de considérer comme étalon, le consensus de plusieurs propositions d'indexation manuelle [Wilbur98].

Une deuxième solution consiste à utiliser la similarité sémantique. Dans les différentes évaluations la plupart du temps deux termes provenant de deux indexations différentes sont considérés équivalents si les deux termes sont exactement les mêmes. On peut nuancer cette évaluation en introduisant une mesure de similarité sémantique [Névéol06]. Cette mesure est fondée sur l'hypothèse que les termes possédant le plus de points communs (ancêtres) sont considérés comme étant plus proches. Cette mesure a été inspirée de la mesure de similarité de Dice [Lin98].

La similarité sémantique entre deux ensembles est définie comme suit (voir figure 2.19) :

S(mi,mj) représente l'ensemble des ancêtres partagés par les deux termes mi et mj. «max» représente le maximum et p(m) est la probabilité de trouver m ou l'un de ses descendants indexés dans un corpus. La similarité générée est une valeur entre 0 et 1. La similarité pour deux termes d'arborescences différentes est égale à 0 (aucun ancêtre en commun).

$$SS(l_A, l_B) = \frac{1}{A+B} \times \left(\sum_i \max_j (Sim(m_i, m_j)) + \sum_j \max_i (Sim(m_i, m_j)) \right)$$

$$\text{Sachant que :} \quad Sim(m_i, m_j) = \frac{2 \times \max_{m \in S(m_i, m_j)} [\log(p(m))]}{\log(p(m_i)) + \log(p(m_j))}$$

FIGURE 2.19 – Mesure de similarité

L'indexation peut aussi être évaluée sur différents niveaux de précision ou d'importance, ceci influence les niveaux de bruit et de silence obtenus. Le niveau de précision consiste à définir un niveau dans l'arborescence auquel tous les termes vont être reportés. Par exemple, il peut être reporté à l'ancêtre de niveau 2 ($2^{\text{ème}}$ niveau de la terminologie après la racine) puis de niveau 3 pour évaluer une indexation plus précise. Nous retrouvons ce genre d'étude dans [Névéol05].

Le niveau d'importance consiste à définir un seuil ou à prendre en compte un type de terme particulier. Un seuil peut être défini lorsque l'indexation est rangée, on peut alors décider de ne prendre en compte que les 5 premiers résultats (lorsqu'il y a ou non un score attribué) ou ceux qui ont un score supérieur au seuil (lorsqu'il y a un score). La D-mesure de Nakache [Nakache05] permet aussi d'évaluer la capacité pour un outil d'indexation automatique de proposer en premier les bons termes [Voorhees03]. Nous retrouvons ce genre d'évaluation dans [Névéol05].

2.5.3 Travaux dans le domaine

L'indexation semi-automatique semble une des meilleures solutions candidates pour aider les indexeurs humains dans leurs tâches quotidiennes. Étudions maintenant la littérature afin d'étudier les méthodes actuelles pour proposer ensuite une nouvelle méthode.

2.5.3.1 Différentes approches

Les différentes approches d'indexation automatique contrôlée consistent à déterminer ce qui dans le texte peut renvoyer vers un terme d'indexation.

2.5.3.1.1 Méthodes de classification

Cette première approche consiste à «apprendre» les associations primaires que peut réaliser l'être humain entre deux notions, ici une expression en langue naturelle et un terme d'une terminologie.

L'indexation peut être rapprochée de la catégorisation. Indexer revient à classer les documents selon certaines catégories représentées par les termes de la terminologie utilisée [Bertrand93]. Ainsi Sebastiani définit la catégorisation de texte comme l'action de «chercher une liaison fonctionnelle entre un ensemble de textes et un ensemble de catégories (étiquettes, classes)» ce qui est très proche de la définition même de l'indexation. Ainsi les approches de classification automatique de documents textuels ont été utilisées par de nombreux chercheurs afin d'indexer (ou de coder) un document. Cette approche consiste en deux phases principales (voir figure 2.20) :

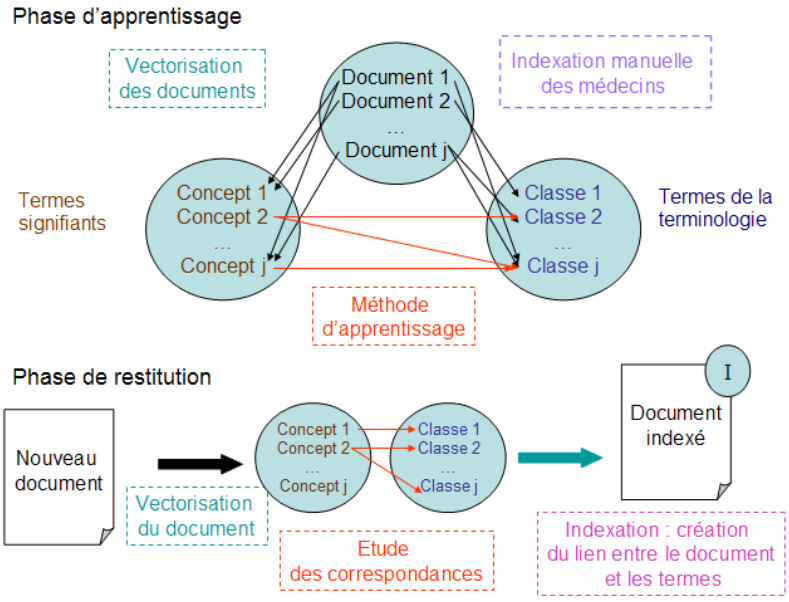

FIGURE 2.20 – Représentation du problème de la classification automatique

- une phase d'apprentissage qui permet d'identifier les relations entre les expressions du document et les codes associés par le codeur humain. Cette phase peut être couplée ou précédée d'une phase de traitement du langage naturel.
  La majorité des outils de classification se fonde sur une représentation vectorielle des documents. Ceci permet de réduire le document textuel à un ensemble de descripteurs significatifs (expressions normalisées appartenant au texte) contenus dans un vecteur. Le fait que les expressions soient normalisées permet de restreindre le nombre d'expressions qui vont être traitées ainsi que de prendre en compte un grand nombre de variations possibles pour cette expression. Les descripteurs sont restreints aux descripteurs les plus discrimants pour le corpus : les plus fréquents et les plus rares sont éliminés (ou très peu pris en compte). La méthode d'apprentissage va consister à lier les descripteurs significatifs pour un ensemble de documents à des termes appartenant au langage d'indexation choisi (les termes dans le cas d'une terminologie). Ces liaisons sont déterminées de manière statistique. Si un descripteur significatif du corpus est souvent associé à un terme (parce que ce terme est souvent indexé pour les documents contenant ce descripteur) alors ce descripteur significatif est lié au terme. Le document contenant ce descripteur sera indexé par ce terme.
  Cette phase est réalisée par des outils d'apprentissage (machine learning). Les méthodes de classification par apprentissage les plus connues sont KPP-V (K Plus Proches Voisins)[Yang94], SVM (Support Vector Machine) [Vapnik95] [Joachims98], LSA (Latent Semantic Analysis) [Deerwester90], LLFS (Linear Least Squares Fit), Naive Bayes [Bayes63]. L'algorithme de CLO3 [Nakache07] obtient de bons résultats puisqu'il améliore de près de 7% les algorithmes analogues.
- une phase de restitution qui permet d'utiliser les correspondances descripteurs/ termes apprises à la phase précédente et stockées dans une base de connaissance pour l'indexation d'un nouveau document. Le nouveau document est analysé, s'il contient un descripteur décrit dans la base alors il est indexé avec le terme correspondant.

Les outil d'indexation automatique utilisant cette approche sont : CIREA [Nakache07] avec l'algorithme CLO3 et le système SMART [Salton89] utilisant le modèle vectoriel.

2.5.3.1.2 Approches TALN (Traitement Automatique du Language Naturel)

La deuxième approche consiste à analyser les associations secondaires réalisées par l'être humain entre deux notions, ici une expression en langue naturelle et un terme d'une terminologie.

Cette approche est associée aux méthodes de TALN pour l'analyse du langage naturel. Le TALN s'appuie sur plusieurs disciplines : la linguistique, l'informatique, les mathématiques (algèbre, logique, statistiques et probabilités), l'Intelligence Artificielle et les sciences cognitives [Cori02]. Tout système de compréhension des langues naturelles doit, par décompositions et analyses successives, transformer la demande

initiale en une formule censée en exprimer le sens. La grande majorité des systèmes de traitement linguistique décomposent les traitements possibles d'un texte selon quatre niveaux, de la compréhension élémentaire à la compréhension globale :

- l'analyse morpho-lexicale se base sur le traitement de la structure des mots
- l'analyse syntaxique se base sur le traitement de la structure des phrases
- l'analyse sémantique se base sur le traitement du sens
- l'analyse pragmatique se base sur le traitement du contexte

Analyse morphologique Elle permet d'identifier les mots du texte. D'abord par identification (ou segmentation) des phrases d'un texte. Puis le texte est découpé en unités lexicales : les mots. Chaque mot peut être identifié par association de sa forme générique (un lemme) et d'une catégorie morphosyntaxique (voir figure 2.21).

Ces méthodes font appel à des traitements lourds, des bases de données volumineuses et nécessitent des réactualisations régulières. Ceci est d'autant plus vrai dans le langage médical où de nouveaux termes apparaissent régulièrement.

Quelques outils : Les outils NOOJ [Silberztein04] et Mmorph [Petitpierre94] permettent une analyse morphologique. Brill [Brill95] et Treetagger [53] sont des systèmes d'étiquetage automatique des catégories grammaticales des mots (compatibles avec FLEMM). FLEMM [Namer00a] est un programme de lemmatisation et d'analyse morphologique du francais.

Analyse syntaxique L'analyse syntaxique traite de la manière dont les mots peuvent se combiner pour former des groupements structurels ainsi que des relations fonctionnelles qui unissent les groupes. Elle se base sur l'analyse morpho-lexicale (voir figure 2.21).

Citons un outil pour le francais : l'analyseur syntaxique SYNTEX [Bourigault00].

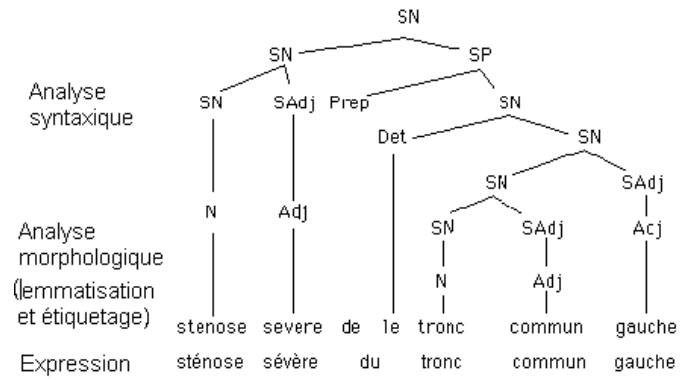

FIGURE 2.21 – Exemple d'analyse morphologique suivie d'une analyse syntaxique (inspiré de [Folch08])

53. Voir le TC Project `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

Analyse sémantique En général, l'analyse sémantique part de l'analyse syntaxique pour déduire le sens de la phrase. Le niveau sémantique est plus complexe à formaliser que les niveaux de traitements précédents. Les analyseurs sémantiques opérationnels sont peu nombreux et concernent des applications très limitées. Nous sommes encore loin de pouvoir couvrir la totalité de la langue. Outre les analyseurs sémantiques [54], l'utilisation d'une terminologie peut permettre d'appréhender le sens d'une phrase par les termes qu'elle contient.

L'outil MENELAS [Zweigenbaum94] contient un analyseur sémantique. On peut citer ici une autre étude celle de Cavazza [Cavazza92].

Analyse pragmatique L'analyse sémantique de phrases, de manière isolée, ne permet pas d'appréhender la signification complète d'un texte, telle que l'humain l'appréhende lors d'un processus de compréhension. Une analyse supplémentaire, l'analyse pragmatique, permet de retrouver des informations implicites liées au contexte d'utilisation des mots. Ces systèmes possèdent une capacité d'inférence [55] [Schank81] [vanDijk90].

Quelques outils : Le prototype Kalipsos d'IBM [Berard-Dugourd89] grâce à une analyse syntaxique et une description conceptuelle permet de résoudre certains liens de sens entre les phrases. Le projet Hélène [Zweigenbaum89] permet l'analyse de l'enchaînement chronologique et causal des faits pour l'analyse de comptes rendus médicaux [Doré92].

Repérer les éléments d'indexation dans un document Les documents sont réalisés pour être lus et compris par des humains et non pour être exploités par des systèmes automatisés, ce qui rend le problème complexe. Afin de déterminer les éléments du document (expressions en langue naturelle) pouvant correspondre morphologiquement (rapprochement au niveau de la forme), syntaxiquement (rapprochement au niveau syntaxique) ou sémantiquement (rapprochement au niveau du sens) à un terme d'une terminologie et, ainsi, réaliser l'indexation du document, il existe plusieurs méthodes (voir figure 2.22).

Utilisation du contenu des terminologies :

Une des méthodes est la construction *a priori* de la liste de l'ensemble des correspondances entre les termes de la terminologie et les expressions en langue naturelle correspondantes. Certaines terminologies contiennent déjà un grand nombre de ces correspondances en liant chaque concept représenté par un terme préféré :

- à ses synonymes (équivalence sémantique) et à ses variantes morphologiques (équivalence morphologique)
- à des références ou compositions (équivalence sémantique) (pour la terminologie SNOMED)
- à des liens de hiérarchies pour des termes proches (proximité sémantique, englobement sémantique)

54. permettent de déduire le sens, de désambiguïser, résoudre les anaphores etc...
55. consiste à tirer une conclusion d'une série de propositions

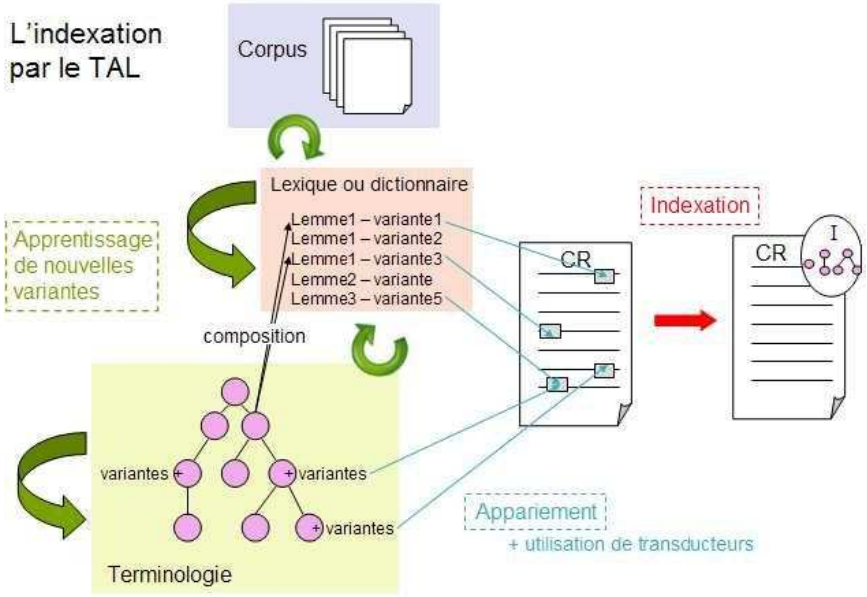

FIGURE 2.22 – L'indexation par les méthodes de TAL

Mais celles-ci sont en nombre insuffisant pour couvrir la réalité.

Le projet VUMeF, qui s'inscrivait dans la suite du projet UMLF, avait pour but d'étendre la part du francais dans le metathésaurus UMLS [56] (projet RNTS 2003 [Darmoni03b]) (collaboration notamment entre l'équipe du LERTIM, la société Vidal et l'équipe CISMeF). Pour le thesaurus MeSH, un gros travail a été réalisé par l'équipe CISMeF [57] afin de définir un maximum de variantes et de synonymes.

Les lexiques :

Pour les termes compositionnels (terme dont le sens est compositionnel, exemple : Sens(infarctus du myocarde)=Sens(infarctus)+Sens(myocarde)), de nouvelles variantes peuvent être découvertes à partir de lexiques simples (exemple : la notion d'«infarctus du myocarde» est complètement dérivable de celle de «infarctus» et de «myocarde»).

Ces genres de lexique sont très complets en anglais pour le domaine médical (CELEX [Burnage90] un lexique pour la langue générale ; le SPECIALIST Lexicon de L'UMLS voir section 2.3.2). En francais, le projet UMLF [Zweigenbaum03] a consisté à créer un lexique médical francophone unifié, ceci à partir de ressources incomplètes et dispersées ([Zweigenbaum90] [Baud92] [Zweigenbaum01]) et en en générant de nouvelles.

De nouvelles variantes à inclure au lexique peuvent être apprises automatiquement à partir des terminologies elles-mêmes [Baud97], [Zweigenbaum98], [Grabar00] ou à partir des lexiques eux-mêmes avec des méthodes :

56. Metathésaurus de l'Unified Medical Langage System contenant plus de 100 terminologies médicale en différentes langues

57. Travail réalisé par l'équipe CISMeF (ajout de plus de 7 000 synonymes), A. Névéol [Névéol05a] et moi-même [Pereira05] (dictionnaire de variantes MeSH)

- d'amorcage à partir de lexiques existants [Gaussier99]
- de décomposition pour les mots composés (exemple : adéno (glande ou ganglion)-myo(muscle)-card(coeur)) [Hathout02a] [Namer00b] [Lovis96].
- à partir de corpus [Xu98], [Jacquemin97], [Hathout02b].

Les grammaires :

Des grammaires morphologiques et syntaxiques peuvent être définies afin de préciser la forme des variantes pour un terme. Ceci peut être très utile pour des termes pouvant prendre des formes multiples (voir figure 2.23).

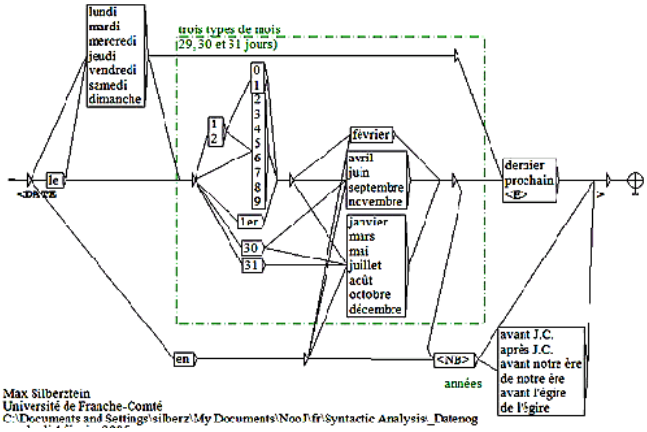

FIGURE 2.23 – Exemple de grammaire syntaxique pour le terme «date»

Ces grammaires sont généralement implémentées sous forme d'automates-dictionnaires (patron d'extraction utilisant des lexiques) [Gaudinat02], [Pouliquen02], [Lovis98] et [Silberztein93].

L'appariement :

L'appariement consiste à faire correspondre une ou des expressions du document à une variante d'un terme (le principe est le même pour la traduction d'une requête par des termes d'une terminologie). Cette mise en correspondance ne prend généralement pas en compte les mots vides (les mots les plus fréquents, susceptibles de fausser la représentation du contenu sémantique du texte. Exemple : «le» ou «de»).

Une expression et un terme sont dits équivalents s'ils sont morphologiquement équivalents (compositions en lemmes égales) ou dérivés (compositions en radicaux ou racines égales) ou proches au niveau de leurs chaînes de caractères ou phonétiquement équivalents ou sont synonymes ou ont de fortes probabilités d'être équivalents (description en N-grammes équivalente) :

- Relier les formes fléchies (exemple : asthme - asthmes) et les formes dérivées (exemple : asthme - asthmatique) à leurs lemmes ou mots de base, accroît la puissance et la souplesse de l'appariement de termes.
- La désuffixation consiste à enlever à un mot son suffixe [58]. Tous les mots dérivés

58. Ce sont les lettres ou syllabes qui s'ajoutent à la fin des mots pour en déterminer la signifi-

obtiennent le même radical (Exemple : diabétique - diabète - diabètes obtiennent le même radical «diabèt»). Les algorithmes de désuffixation les plus célèbres pour l'anglais sont les algorithmes de Porter [Porter80] et de Lovins [Lovins68]. Pour le francais, il existe l'algorithme de Carry [Paternostre02], le «Frenchstemmer» de Lucene [59] utilisé dans les moteurs de recherche sur Internet et l'outil EDA de Nakache [Nakache07].
- La racinisation consiste à obtenir à partir d'un mot sa racine. Une racine est obtenue en éliminant tout préfixe, affixe et suffixe d'un mot.
- La phonémisation consiste à interpréter phonétiquement un mot [Odell18].
- La distance de Levenshtein mesure la similarité entre deux chaînes de caractères. Elle est égale au nombre minimal de caractères qu'il faut supprimer, insérer, ou remplacer pour passer d'une chaîne à l'autre [Levenshtein66a] (d'autres distances existent, nous citons celle-ci qui est la plus connue).
- Dans le projet Vodel, une étude a porté sur la comparaison de termes en étudiant leurs définitions et non plus seulement leurs libellés [Diosan08].
- La méthode des N-grammes permet d'identifier des expressions ayant une forte probabilité d'être synonymes [Bell90]. Le texte à indexer est découpé formant tous les groupes de mots contenant 1 à n mots consécutifs sans ponctuation possible puis tous les mots sont réduits à leurs N premiers caractères. Pour chaque groupe constitué, un score de correspondance avec les termes de la terminologie est calculé.

L'appariement peut s'appuyer sur des éléments syntaxiques, sémantiques ou pragmatiques afin de préciser les conditions d'appariement.

Quelques systèmes utilisent une approche TAL pour l'extraction de termes : NLPAD [Zweigenbaum92], RIME [Berrut90] et LSP-MLP [Sager95].

2.5.3.2 Indexeurs automatiques existants

La majorité des outils d'aide à l'indexation en place aujourd'hui dans les hôpitaux ou les organismes pratiquant une indexation manuelle sont des outils d'aide à la recherche dans les différentes terminologies [Bouchet99] [Berthelot05]. Il en existe beaucoup, les différences se situent dans le type de recherche proposé qui peut aller de la navigation simple dans la hiérarchie à une interprétation plus ou moins intelligente d'une requête de l'utilisateur [60]. Les outils WEBCCAM, WEBCIM de la société Web100t [Lewandowski08], CODAZ (par le Dr P. Frutiger) et l'outil du Dr J. Ruiz sont de bons exemples d'outils d'aide à la recherche intelligents pour la CCAM et la CIM10. Nous pouvons aussi citer ici le serveur de terminologie CISMeF [61] pour la terminologie CISMeF et MeSH [Thirion07].

D'autres outils plus élaborés permettent d'extraire directement les termes d'indexation à partir d'un compte rendu médical. Nous pouvons distinguer trois sortes

cation.

59. voir Snowball.

60. À partir d'une requête de l'utilisateur, l'outil propose les termes de la terminologie les plus adaptés.

61. http ://terminologiecismef.chu-rouen.fr/

d'outils :

- les outils permettant une indexation monoterminologique directe.
  Exemple : CIREA, MeSHMapp, MAIF, Snocode et un outil pour la CIM10 japonaise [Amaraki07] une méthode hybride qui sélectionne la méthode à utiliser par rapport à une entrée donnée.
- les outils permettant une indexation monoterminologie indirecte c'est-à-dire à partir d'un transcodage.
  Exemple : Nomindex (dictionnaire ADM→ MeSH) [Pouliquen02], MedCKARe (ontologie pneumologie→ CIM10).
- les outils permettant une indexation multiterminologique. Ici deux approches peuvent être identifiées :
  Les systèmes produisant une indexation directe pour plusieurs terminologies
  Exemple : HONMeSHMapper et MEDLEE [62] [Friedman04] fonctionnent sur l'UMLS.
  Les systèmes produisant une indexation directe et indirecte pour plusieurs terminologies
  Exemple : MTI fonctionne sur l'UMLS et permet d'indexer en MeSH et CIM9-CM en utilisant tout le Metathesaurus de l'UMLS.

Il existe des outils industriels comme l'outil Snocode et Insight Discoverer Extractor l'outil de la société Témis [63] qui permettent l'extraction de termes MeSH francais et anglais. L'outil de la société Microsoft [64] permet une extraction de termes SNOMED 3.5. Enfin l'outil de HealthLanguage [65] permet une indexation en SNOMED CT.

Nous nous sommes intéressée au fonctionnement des principaux outils et notamment ceux developpés pour le francais pour nos terminologies afin de déterminer notre propre approche.

2.5.3.2.1 MAIF (MeSH Automatic Indexing for French)

Le système MAIF (MeSH Automatic Indexing for French) a été développé par A. Névéol [66] lors de sa thèse au sein de l'équipe CISMeF [Névéol05a]. Ce système permet à partir de l'URL d'une ressource en francais de produire une indexation par des mots clés MeSH francais (mots clés ou paires mot clé/qualificatif). Le texte est traité avec une approche TAL et le titre de la ressource avec une approche k-PPV (voir figure 2.24).

L'**approche TAL** consiste en l'application par le logiciel INTEX [67] d'un dictionnaire et de transducteurs (= patrons d'extraction) MeSH. Ce dictionnaire contient les diverses formes que peuvent prendre en langue naturelle les termes MeSH. Les trans-

62. Medical Language Extraction and Encoding System
63. Voir `http://www.temis.com/`. J'ai pu participer à l'élaboration de leur outil d'indexation en intégrant un module pour l'indexation de la Classification Internationale du Handicap (CIH).
64. Outil présenté à la réunion «Serveurs de terminologies médicales» le 24 septembre 2007 au GDR STIC Santé Thème C
65. http ://www.healthlanguage.com/
66. A. Névéol effectue un post-doc au Lister Hill, NLM. L'équipe CISMeF et moi-même continuons à collaborer activement avec elle (voir liste des publications issues de cette thèse)
67. Logiciel permettant la création et l'application de dictionnaires ainsi que des transducteurs.

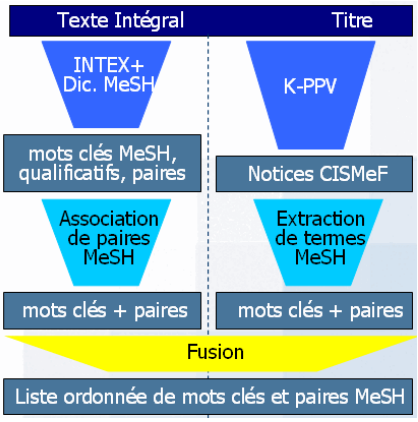

FIGURE 2.24 – Fonctionnement de l'outil MAIF [Névéol05a]

ducteurs permettent de rendre compte de la grande variabilité de certains mots clés (exemple : «adulte d'âge moyen»). Ce dictionnaire a été créé à partir de ressources disponibles sur Internet et dans le milieu de la recherche. Pour la langue générale [68] ces données sont issues des dictionnaires DELA existants et de Lexique23. Pour le langage médical, elles sont issues des données UMLF. Des ajouts ont aussi été effectués de manière manuelle et semi-automatique : bases de synonymes MeSH et CISMeF, traduction automatique, traitement de certaines expressions récurrentes.

La **méthode K-PPV** (K Plus Proches Voisins) reprend l'indexation MeSH de documents dont le titre est proche de celui de la ressource à indexer. Les documents proches contiennent au moins un mot (pertinent) du titre du document à indexer. Pour chaque titre extrait, le calcul d'un score de similarité utilisant la distance de Levenshtein [Levenshtein66b] est calculé afin de ne retenir que les k premiers.

La **fusion des méthodes** consiste à regrouper les indexations produites et à sommer les scores obtenus dans les deux méthodes. Les termes sont enfin rangés dans l'ordre décroissant.

Un **seuil** permet de détecter une rupture dans la continuité des scores et donc dans la pertinence des candidats proposés [Abdallah98].

L'indexation automatique produite par MAIF a été comparée à celle produite manuellement par les indexeurs CISMeF sur le corpus «diabète» de CISMeF comprenant 57 ressources. MAIF a montré une précision de 6,2% et un rappel de 35,3% en prenant en compte les 50 termes MeSH les plus pertinents pour chaque ressource. L'application du seuil permet d'obtenir une précision de 24,2% et un rappel de 7,4%.

MAIF a été comparé à d'autres système d'indexation MeSH : Nomindex[Pouliquen02], HONMeSHMapper[Gaudinat02], MeSHMapp[Ruch03] et MTI (voir section suivante) (voir résultats figure 2.25).

Le système MAIF a été utilisé pour l'indexation en CIM10 de comptes rendus médicaux [Pereira05] [69]. Pour ce faire, il a été couplé à une table de transcodage

68. Langage courant
69. Étude que j'ai menée pendant mon stage de DEA avant la création de F-MTI.

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)**

FIGURE 2.25 – Précision et rappel des systèmes francophones aux rangs fixes 1, 4, 7, 10 et au seuil adaptatif [Névéol05a]

MeSH/CIM10 extraite de l'UMLS afin de transcrire en CIM10 les termes MeSH extraits par MAIF à partir de comptes rendus médicaux. Le système a montré une précision de 15% et un rappel de 28% comparé à une indexation manuelle de 100 comptes rendus médicaux effectuée par des médecins. L'indexation CIM10 été comparée à celle de l'outil industriel SnoCode. La même évaluation sur 100 comptes rendus a montré une précision et un rappel de 26% et 49%.

La médication est directement corrélée aux diagnostics du patient. Une étude a porté sur l'utilisation de la médication pour l'indexation de codes CIM10. Pour chaque médicament prescrit, les liens médicament->groupe d'indication->code CIM10 fournis par la société Vidal ont permis l'extraction de nombreux codes CIM10 potentiels. Ces codes CIM10 sont hiérarchisés grâce à une métrique élaborée lors de cette étude afin de ne garder que les plus probables. L'évaluation sur 100 comptes rendus a montré un rappel de 60% (au rang 0 et 28% au rang 20) et une précision très faible de 3%.

2.5.3.2.2 Medical Text Indexer (MTI)

Medical Text Indexer (MTI) [Aronson00] permet l'indexation semi-automatique en MeSH anglais des articles anglophones de MEDLINE. Dans le cadre de MEDLINE, il traite les titres et les résumés des articles. Il peut aussi indexer en texte intégral. L'indexation automatique produite est proposée à l'indexeur qui clique alors sur les termes qu'il désire garder.

Il associe 3 approches : une approche de Traitement Automatique de la Langue Naturelle implémentée dans le système MetaMap (MM), une méthode utilisant des trigrammes [70], et une approche statistique appelée «PubMed Related Citations» (PRC) tout en utilisant le Metathesaurus de l'UMLS (voir figure 2.26).

MetaMap [Aronson01] permet d'analyser un texte et d'en extraire des termes de l'UMLS. MetaMap opère comme suit :

- découpe le document en groupes nominaux [71] après un étiquetage syntaxique grâce à l'outil Phrasex. Les mots vides sont ici ignorés.
  Exemple : le texte : «The local anesthetic bupivacaine is cardiotoxic...» est

70. Méthode des N-grammes où N=3 (tous les mots sont réduits à leurs 3 premiers caractères)
71. Un groupe nominal est un ensemble de mots groupés autour d'un nom (exemples : une poupée nageait au fil de l'eau).

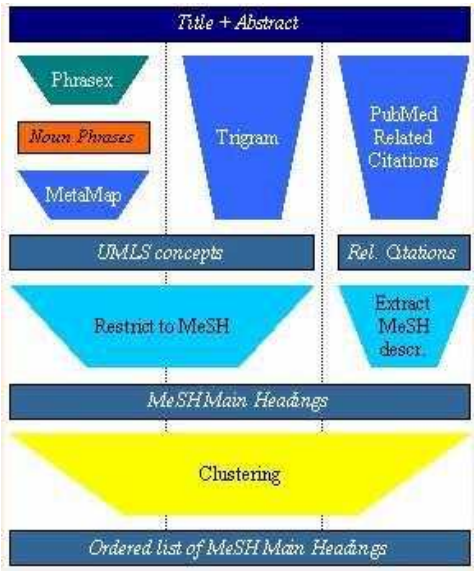

FIGURE 2.26 – Fonctionnement de l'outil MTI [Aronson00]

découpé en «the local anesthetic bupivacaine», «is», «cardiotoxic», ... ;
- génère toutes les variantes, ainsi que certaines variantes de variantes (variantes orthographiques, abréviations, acronymes, synonymes, variantes dérivationnelles et flexionnelles pour chaque mot et chaque combinaison de mots décrits dans le SPECIALIST Lexicon de l'UMLS)
  Exemple : «anesthetics», «anaethetic», «anesthesia»... ;
- extrait les termes candidats du Metathesaurus (donc indépendamment de la terminologie source) contenant au moins l'une de ces variantes
  Exemple : «Bupivacaine», «local anaesthetic», «local anaetheti, NOS» ;
- attribue un score à chaque terme candidat correspondant à la qualité de la correspondance entre les termes candidats et les groupes nominaux dont ils ont été extraits puis range les termes par score ;
- combine les termes candidats liés à un même groupe nominal et calcule à nouveau le score puis sélectionne les candidats ayant le meilleur score. Exemple : «Bupivacaine» et «Local anaethetic» ou «Local anaethetic, Nos».

La **méthode des trigrammes** applique la méthode des N-grammes avec N=3 (voir section précédente). Après l'application de cette méthode, les candidats termes issus du titre de la ressource ou ayant obtenus le meilleur score sont sélectionnés. La fréquence des termes dans le document est aussi calculée.

Le **module «Restrict to MeSH»** trouve tous les mots clés MeSH les plus proches des termes UMLS candidats [Bodenreider00]. Les termes MeSH sont d'abord recherchés parmi les synonymes des termes UMLS. Si aucun synonyme pour un terme UMLS n'est trouvé alors la recherche se fait sur les combinaisons de termes MeSH. Puis le réseau hiérarchique du terme UMLS est exploité afin de trouver un parent lié à un terme MeSH. Enfin si aucun terme MeSH n'est trouvé les relations non hiérarchiques sont exploitées.

L'**algorithme PRC** [Kim01] extrait une liste ordonnée de termes MeSH à partir

d'un titre et d'un résumé d'article en recherchant les articles les 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. Un score est attribué à chaque terme dépendant de sa fréquence et de sa pertinence.

Le **module clustering** permet de générer la proposition d'indexation automatique finale. Tous les termes MeSH candidats extraits par les trois méthodes sont regroupés. Les règles d'indexation MEDLINE sont appliquées :

- les termes sont alors pondérés selon la méthode d'extraction d'origine (poids heuristique de 7 pour MM et 2 pour PRC)
- les termes sont aussi pondérés selon la localisation du groupe nominal d'origine (poids supérieur pour ceux du titre)
- les termes PRC sont éliminés s'il n'y a pas de terme MM plus général

Depuis 2007, MTI est capable d'associer des mots clés MeSH avec des qualificatifs grâce aux travaux d'A. Névéol [Néveol07] issue de l'équipe CISMeF. De plus, un module de désambiguisation utilise les co-occurrences [72] entre les Journal Descriptor [73] et les termes MeSH [Humphrey06]. Une indexation spécialisée pour les ressources du domaine de la génomique existe en utilisant GeneRif [Névéol07a].

L'évaluation de l'indexation produite par MTI sur le titre et le résumé par rapport à une indexation manuelle sur 273 articles de l'équipe Medline a montré une moyenne de 8 termes par article ainsi qu'une précision de 29% et un rappel de 55% pour les mots clés [Aronson04]. Pour les mots clés majeurs (3 en moyenne), les résultats sont de 81% pour la précision et 11% pour le rappel. Les résultats sont différents selon les journaux indexés. De plus, l'interview des indexeurs avait montré que 37% des indexeurs trouvaient que le recouvrement était bon, 53% partiel, 10% insuffisant.

Une autre étude sur 500 articles [Gay05] montre que pour l'indexation du texte intégral le système produit une précision de 31% (-1% par rapport au titre et au résumé), un rappel de 60% (+7%) et une F-mesure de 49.1% (considérant seulement la méthode MetaMap le résultat est 24% - 37% - 32,4%) .

A. Névéol avait comparé MTI à son outil MAIF. Pour cette évaluation, un corpus de ressources parallèles a été utilisé, le corpus parallèle «ENFR» qui comporte 51 ressources CISMeF écrites en 2 langues. Les résultats montrent une supériorité du système MTI (MAIF : Précision 27,2% - Rappel 36,1% - F-mesure 31% et MTI : 33,6% - 61,8% - 43,6%)

Récemment, MTI été appliqué à l'indexation CIM9-CM des documents cliniques. L'outil utilise des méthodes d'apprentissage automatique : SVM et k-PPV et une méthode simple de modèles de correspondance. De plus, il utilise l'outil NegEx [Chapman01, Goldin03] qui permet de trouver les expressions négatives. NegEx a permis de générer un dictionnaire contenant toutes les expressions négatives possibles pour tous les termes du Metathesaurus [Aronson07]. Évalué dans le cadre d'un concours TAL, the Medical NLP Challenge [74], sur un corpus statistiquement normalisé de 1 000 rapports de radiologie, MTI a obtenu une F-mesure de 85%. C'est l'outil

72. Deux termes sont dit co-occurrents s'ils sont retrouvés ensemble dans un corpus. Deux termes souvent co-occurrents ont une forte probabilité d'être reliés par une relation sémantique.

73. Catégorie de journaux par spécialités médicales assez proche des métatermes de CISMeF

74. Voir `http://www.computationalmedicine.org/challenge`

de l'équipe Szeged qui a obtenu les meilleurs resultats avec 89,1% de F-mesure.

2.5.3.2.3 MedCKARe

MedCKARe (Medical Coding by Knowledge Acquisition and Representation) est un outil d'aide au codage développé par A. Baneyx [Baneyx06] dans le cadre du projet PERTOMed. Cet outil permet d'indexer des comptes rendus médicaux en CIM10. Il extrait 337 expressions les plus couramment rencontrées par les pneumologues liées par des relations de transcodage (1 à n) à la classification CIM10. Ces expressions sont modélisées et reconnues à l'aide d'une ontologie du domaine de la pneumologie. Une expression peut être définie dans l'ontologie par une combinaison de deux ou plusieurs concepts primitifs reliés entre eux par une ou plusieurs relations. Le système utilise le dictionnaire Unitex et des patrons lexicosyntaxiques afin de reconnaitre ces combinaisons. La négation est aussi gérée. L'évaluation de cet outil sur un corpus de 500 comptes rendus a montré un rappel de 25% et une précision de 87%. MedCKARe propose aussi une interface dédiée à l'aide au codage.

2.5.3.2.4 CIREA

Un outil d'aide au codage PMSI pour les services de réanimation a été implémenté par D. Nakache [Nakache07] dans le cadre du projet CIREA (Classification Informatique pour la REAnimation [75]). L'outil développé permet d'extraire les codes CIM10 à partir de comptes rendus hospitaliers rédigés en langage naturel. Il utilise un algorithme de classification par apprentissage, l'algorithme CLO3 qui s'inspire à la fois de TF/IDF et de Naive Bayes [Bayes 1763]. Cet algorithme a montré de meilleurs résultats que d'autres méthodes analogues : k-PPV, SVM, Naive Bayes, TF IDF/RM. L'évaluation du système a donné une précision de 43,7% et un rappel de 38,6% pour l'indexation de 10 000 comptes rendus avec un jeu de 30 000 comptes rendus.

2.5.3.2.5 SnoCode

SnoCode est un outil de la société canadienne MedSight [76] qui date de la fin des années 90. Il est destiné à indexer automatiquement les documents cliniques en SNOMED et CIM10. Les informations sur le fonctionnement de l'outil, le stockage des données et les technonologies d'indexation en langage naturel ne sont pas diffusées par la société. Il utilise des méthodes de correspondance et des synonymes pour comparer les séquences de mots du document (jusqu'à 14 mots à la fois) avec la nomenclature SNOMED 3.5 qui a été restructurée afin de permettre des comparaisons rapides et efficaces. Seules les correspondances exactes et les plus longues sont retenues. Le système permet une indexation en SNOMED 3.5 et en CIM10, l'indexation CIM10 étant obtenue par le transcodage SNOMED vers CIM10 qui avait été développé par la SFINM [77].

75. projet faisant parti d'un projet plus vaste, le projet RHEA qui vise à mettre en œuvre des structures informatiques décisionnelles pour les services de réanimation

76. http ://www.medsight-info.com/IndexFr.html

77. Secrétariat Francophone International de Nomenclature Médicale

2.5.4 Notre contribution

L'indexation des ressources Web, des RCP et des dossiers médicaux est de manière générale réalisée à la main à l'hôpital, au Vidal ou sur Internet. Nous proposons d'utiliser des méthodes d'indexation automatique afin d'aider les indexeurs dans ces tâches. Nous proposons de créer un outil multi-tâche, multi-terminologie, et multi-document.

Au vu de l'état de l'art, peu d'outils permettent d'indexer des documents à l'aide de plusieurs terminologies (MAIF, MTI et Snocode). De plus, il n'existe aucun outil d'indexation automatique pour la CCAM et le TUV. Il n'existe pas non plus d'outil d'indexation automatique pour la CIM10 utilisant une méthode TAL avec indexation directe. Enfin, il n'existe pas d'outil d'indexation automatique libre pour la SNOMED 3.5 en francais. En revanche, pour le MeSH les travaux sont nombreux.

Il existe plusieurs types d'indexation pour une terminologie : directe, indirecte et mixte (directe plus indirecte). Aucune évaluation n'a pu montrer quelle était la meilleure méthode.

Les outils utilisent des approches différentes. Les différentes approches présentent des avantages et des inconvénients :

- Les méthodes de classification automatique ont l'avantage de ne pas avoir à analyser le sens d'un texte ou à prendre en compte les règles d'indexation pour une terminologie. En revanche, elles ont pour défaut d'apprendre la facon dont a été indexé un corpus précis pour une tâche précise. Face à l'indexation d'un nouveau document pour une autre tâche, la méthode ne sera pas autant efficace. Dans le cadre de l'indexation CIM10, par exemple, l'outil CIREA a appris à réaliser une indexation médico-économique (comprend des règles spécifiques au classement en GHM des séjours), il serait donc incapable de réaliser une indexation purement descriptive de comptes rendus médicaux (deux tâches différentes). De plus, il peut apprendre sur des associations fausses (qualité du codage faible, et les règles de codage valides une année peuvent ne plus l'être l'année suivante). Le système est donc obligé de réapprendre sans cesse au fur et à mesure des changements de règles ou de nouvelles versions de terminologies. Face à l'ajout de nouveaux termes dans une terminologie, le système n'a aucun élément pour pouvoir les indexer. Pour les terminologies qui évoluent souvent cette approche n'est donc pas du tout adaptée (pour la CCAM ou le MeSH par exemple).
- Les méthodes TAL ont comme avantage de prendre en compte le sens d'un texte et de séparer le processus d'extraction de termes des règles d'indexaction. Un système utilisant cette approche peut donc tout à fait s'adapter à de nouvelles règles d'indexation, à l'indexation de documents de types différents ou à une mise à jour quotidienne de la terminologie qu'il indexe. Le défaut de cette approche est que les ressources nécessaires sont incomplètes. Il faudrait disposer d'un lexique complet pour la langue francaise générale et médicale, et de terminologies complètes (avec toutes les variantes possibles pour chaque terme). De plus ces ressources sont difficiles à obtenir (les méthodes existantes ne sont capables d'extraire que les formes simples (composées de 1 à 2 mots))

et doivent être validées manuellement.

Nous observons aujourd'hui dans les outils d'indexation automatique un usage combiné de ces méthodes. Nous avons choisi de nous intéresser plus particulièrement aux méthodes TAL et non au méthodes statistiques car nous ne sommes pas spécialistes en la matière. Nous ne nous intéressons pas non plus aux méthodes de classification de termes par ordre d'importance ou aux méthodes statistiques telles que k-PPV qui permettent d'utiliser l'indexation de documents proches, sachant qu'A. Névéol a travaillé sur ces méthodes statistiques et qu'elles pourront être *in fine* intégrées dans notre outil (sans compter les travaux de T. Merabti sur les related documents [Merabti08b]).

Nous apporterons notre contribution dans l'enrichissement de terminologies, des lexiques et des grammaires. Nous développerons de nouvelles méthodes d'appariement, ainsi qu'une méthode de création automatique de variantes de termes à partir de corpus.

De plus, l'état de l'art montre que peu d'outils prennent en compte les aspects pragmatiques (Medckare prend en compte la négation, MTI prend en compte les domaines de spécialité). Nous essaierons d'apporter notre contribution dans ce domaine.

Nous proposons de construire un outil d'indexation et d'aide à l'indexation automatique généraliste. Nous contribuerons aussi au développement de nouveaux accès contextuels à l'information médicale.

2.6 Conclusion

L'analyse du contexte et de l'état de l'art nous ont permis d'identifier les domaines ainsi que les tâches d'indexation qui nous préoccupent.

Après analyse de l'état de l'art, nous avons pu définir les limites des travaux d'aide à l'indexation existants. Le chapitre suivant montre notre contribution en matière d'aide à l'indexation avec le développement de F-MTI un outil d'indexation automatique multi-terminologique.

Deuxième partie

F-MTI, un extracteur multi-terminologique pour l'aide à l'indexation

Chapitre 3

Conception de l'extracteur multi-terminologique

3.1 Introduction

Comme exposé dans le chapitre 1, les besoins recouvrant des objectifs d'indexation ont été exprimés par les équipes impliquées dans cette thèse. Nous avons fait le choix de réaliser un outil multi-tâche générique en mesure de reproduire automatiquement les tâches suivantes réalisées habituellement à la main :

- indexation des sites Web en MeSH
- indexation des dossiers médicaux en CIM10, CCAM et SNOMED 3.5
- et indexation des RCP en TUV

Nous avons ainsi développé F-MTI (French Multi-Terminology Indexer), un outil d'indexation automatique multi-document, multi-terminologique et multi-tâche. Nous présentons dans ce chapitre le fonctionnement de cet outil.

3.2 Principe de la multi-terminologie

Cet outil intègre le principe de la multi-terminologie. Ce principe a été inspiré par l'outil d'aide à l'indexation MTI (MeSH Terminology Indexer voir section 2.5.3.2.2). Il consiste à utiliser la totalité du réseau formé par les différentes terminologies considérées et non pas à considérer séparément les terminologies. Comme nous l'avons décrit précédemment, il existe des relations entres ces terminologies. Ces relations sont définies soit à l'intérieur du Metathesaurus de l'UMLS soit créées pour des besoins précis par différents organismes.

Nous nous sommes intéressé plus particulièrement aux relations d'équivalence pure entre ces terminologies. Puisque toutes ces terminologies concernent le même domaine, le domaine médical, certains concepts comme par exemple «asthme» peuvent se retrouver dans plusieurs d'entre elles. Il existe donc entre les différents termes exprimant le concept «asthme» au sein de ces différentes terminologies des relations d'équivalence appelée transcodages («mapping» en anglais).

Les termes liés par une relation de transcodage peuvent être considérés comme des

synonymes ou des variantes lexicales. Ainsi en prenant en compte plusieurs terminologies, nous pouvons répertorier un plus grand nombre de formes textuelles possibles pour un terme, rendant ainsi plus aisée son identification dans un texte.

L'outil MTI utilise ce principe en mettant en œuvre l'ensemble du métathesaurus de l'UMLS (soit plus de 100 terminologies). Dans le fonctionnement, c'est l'outil MetaMap (inclus dans MTI) qui extrait tous les termes du Metathesaurus puis restreint tous les termes extraits aux termes MeSH sémantiquement plus proches pour réaliser une proposition d'indexation MeSH. Tout comme MTI, F-MTI réalise une extraction en deux temps : tout d'abord une extraction des concepts pour les cinq terminologies (CIM10, SNOMED 3.5, CCAM, MeSH, TUV), puis une restriction aux termes de la (ou des) terminologie(s), paramétrée(s) en sortie, sémantiquement équivalents *via* les relations de transcodage.

3.3 Principe de fonctionnement

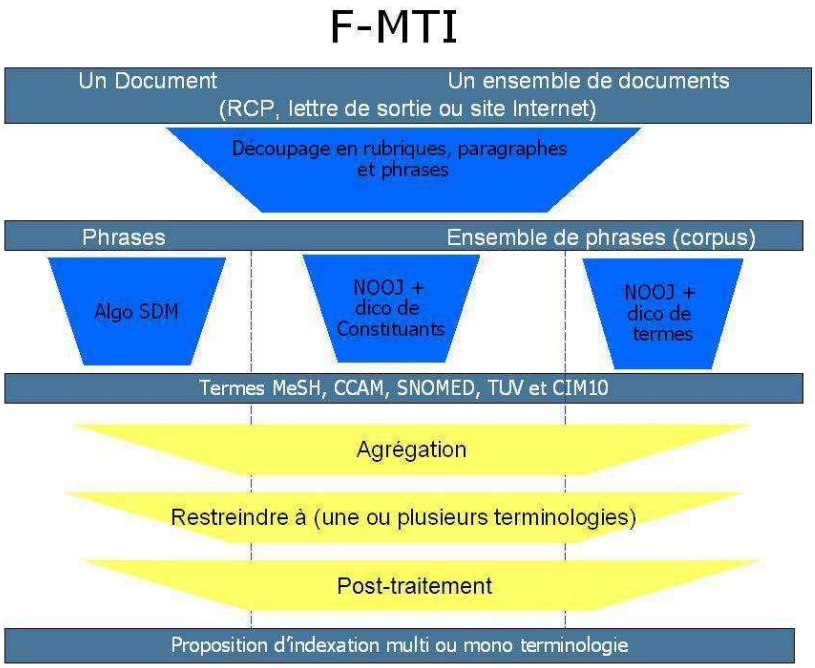

FIGURE 3.1 – Principe de fonctionnement de F-MTI

F-MTI permet une indexation multi-terminologie du texte. Le programme prend en entrée un document ou un ensemble de documents au format texte. Ces documents peuvent être de différentes natures mais un traitement particulier est réalisé pour les comptes rendus hospitaliers, les sites médicaux et les RCP afin de produire une indexation dépendante du type de document.

Le choix des terminologies d'indexation peut être paramétré en entrée mais par défaut les comptes rendus hospitaliers seront indexés en CIM10, CCAM et SNOMED 3.5, les sites Web en MeSH et les RCP en TUV. D'autres paramètres peuvent être

considérés en entrée, nous verrons lesquels par la suite.

La figure 3.1 montre le fonctionnement général de l'outil. L'indexation des documents se fait en plusieurs phases :

- Premièrement, les documents sont découpés en rubriques, paragraphes et phrases. Ce découpage peut être physique ou se limiter à une identification des rubriques, paragraphes et phrases ainsi que leurs emplacements à l'intérieur du document.
- Trois méthodes d'indexation peuvent alors être appliquées : l'algorithme du sac de mots, le dictionnaire de termes et le dictionnaire de constituants. L'outil peut être paramétré afin d'utiliser une ou plusieurs de ces méthodes. Ces méthodes seront décrites dans les sections suivantes.
- Les différents termes issus de ces indexations réalisées par les différentes méthodes sont agrégés et filtrés.
- Enfin des post-traitements sont appliqués afin de proposer une liste de termes d'indexation pour le ou les document(s) à l'utilisateur.

Ces différentes étapes sont décrites dans les sections suivantes.

3.4 Modélisation des terminologies

Afin de permettre à F-MTI d'interroger de facon rapide les cinq terminologies d'intérêt ainsi que les éléments nécessaires aux différentes méthodes, il a fallu dans un premier temps créer une structure de données simple et générique pouvant contenir ces cinq terminologies. De cette structure dépendra le temps d'exécution du programme. La structure doit également être facile à mettre à jour.

Les cinq ressources terminologiques mises en œuvre dans ce projet sont :

- La Classification Internationale des Maladies 10ème édition (CIM10)
- La Classification Commune des Actes Médicaux (CCAM)
- La Nomenclature systématique de médecine humaine et vétérinaire (SNOMED 3.5)
- Le thésaurus médical CISMeF (contenant le Medical Subject Headings (MeSH))
- Le Thésaurus Unifié VIDAL (TUV)

Toutes ces terminologies ont des structures et des particularités différentes. Nous avons dans un premier temps analysé ces structures en modélisant une à une chaque terminologie. Dans un second temps, nous avons élaboré le modèle général à partir de ces modélisations unitaires.

3.4.1 Modèles unitaires

Nous avons modélisé la structure de chaque terminologie à partir des descriptions de chacune faites à la section 2.4. Les éléments définissant la structure de la terminologie ainsi que les liens entre eux ont été identifiés et retranscrits dans un modèle au formalisme UML [1].

1. Ils ont été réalisés à partir du logiciel Poséidon for UML (télechargeable à l'adresse suivante : `http://www.gentleware.com/products.html`)

Nous présentons ici le modèle de la terminologie CISMeF ainsi que celui de la terminologie TUV (pour plus d'informations et pour consulter les autres modèles voir Annexes - Modèles unitaires).

3.4.1.1 Modèle CISMeF

Le modèle de représentation de la terminologie CISMeF déduit de la description faite à la section 2.3.2 est présenté figure 3.2.

Neuf classes ont été identifiées (voir Annexes - Modèles unitaires) :

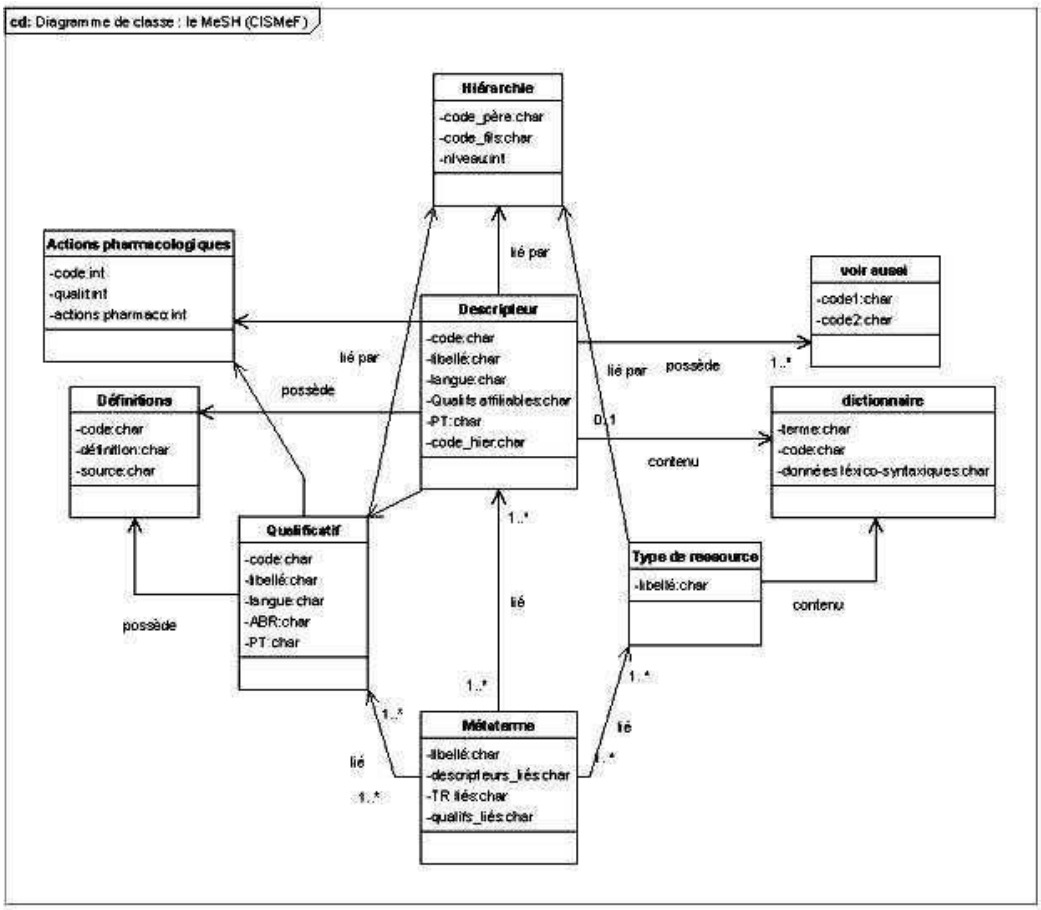

FIGURE 3.2 – Diagramme de classes représentant la structure du MeSH au formalisme UML

- **Classe des descripteurs** : cette classe renseigne les descripteurs du thésaurus.
- **Classe des Qualificatifs** : cette classe renseigne tous les qualificatifs du thésaurus MeSH.
- **Classe des Types de ressources** : cette classe renseigne tous les types de ressources CISMeF.
- **Classe des Métatermes** : cette classe réunit tous les métatermes pouvant être rattachés à un ou plusieurs descripteurs, qualificatifs et types de ressource.
- **Classe Hiérarchie** : cette classe structure la hiérarchie au sein du MeSH.

- **Classe Voir aussi** : cette classe renseigne tous les liens de «voir aussi» entre deux codes MeSH.
- **Classe des Définitions** : cette classe réunit pour chaque code MeSH les définitions auxquelles ils sont rattachés.
- **Classe Dictionnaire** : cette classe indique toutes les variations, flexions, synonymes et leurs classes lexico-syntaxiques pour chaque terme MeSH.
- **Classe des Actions pharmacologiques** : cette classe renseigne tous les liens «action pharmacologique» entre deux termes MeSH.

3.4.1.2 Modèle TUV

Nous présentons ensuite un deuxième modèle, celui de la terminologie TUV (voir section 2.3.2). Ce modèle est présenté figure 3.3 dans un formalisme UML (les noms internes au Vidal ont été conservés). Ce modèle présente 8 classes (voir Annexes -

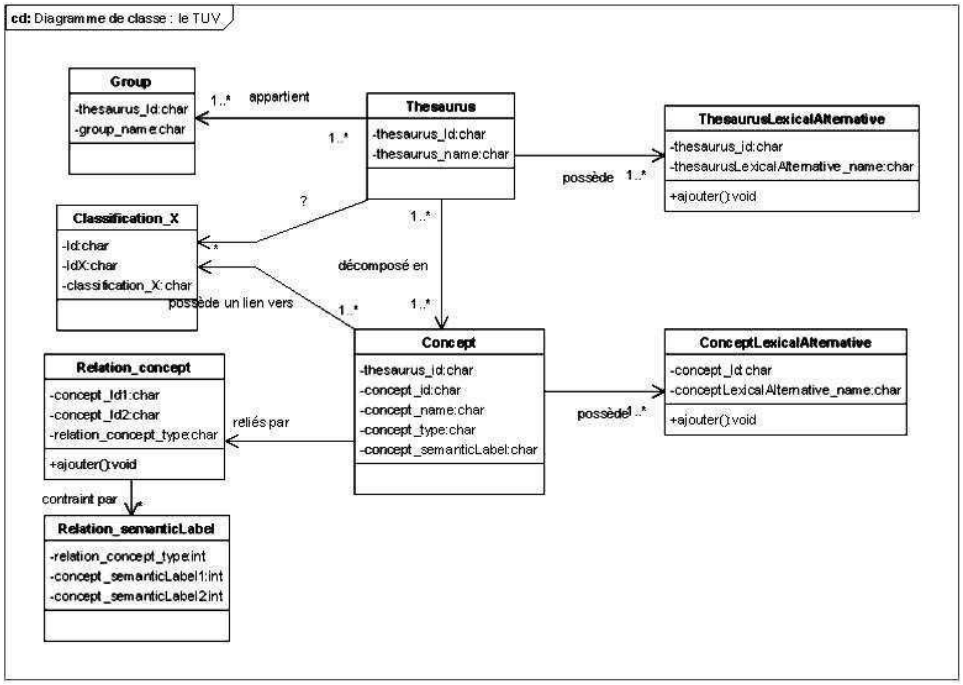

FIGURE 3.3 – Diagramme de classes représentant la structure du TUV au formalisme UML

Modèles unitaires) :

- **Classe des Thesaurus** : cette classe réunit tous les termes de référence du thesaurus TUV.
- **Classe des Concepts** : cette classe réunit tous les termes élémentaires décrivant un terme de référence du TUV.
- **Classe des Group** : cette classe désigne les liens d'appartenance d'un terme d'indexation à un groupe d'indications.
- **Classe des Classification_X** : cette classe renseigne tous les liens reliant un terme de référence ou un terme élémentaire à d'autres terminologies telles que

CIM10, la CISP ou la SFMG.
- **Classe des Thesaurus_Lexical_Alternative** : cette classe indique toutes les variantes lexicales, flexionnelles et synonymiques pour chaque terme d'indexation (terme complexe).
- **Classe des Concept_Lexical_Alternative** : cette classe indique toutes les variantes lexicales, flexionnelles et synonymiques pour chaque terme élémentaire.
- **Classe des Relation_concept** : cette classe renseigne tous les liens sémantiques pouvant relier deux termes élémentaires.
- **Classe des Relation_semanticLabel** : cette classe renseigne tous les liens sémantiques pouvant relier deux étiquettes sémantiques.

3.4.2 Modèle général

Le modèle général doit être simple (pour diminuer le temps d'exécution de F-MTI) et générique (pour inclure les cinq terminologies et permettre d'insérer plus facilement d'autres terminologies dans l'avenir). Nous nous sommes inspiré des tables et des champs définissant la structure du Metathesaurus de l'UMLS [2]. En effet, la structure de l'UMLS contient à ce jour, au sein d'une même structure, plus de 100 terminologies dont la CIM10, la SNOMED 3.5 et le MeSH.

Nous avons tout d'abord identifié tous les attributs et classes communes à toutes les terminologies. Puis, nous avons réalisé des opérations pour certaines terminologies afin de déterminer d'autres attributs et classes en commun et intégrer toutes les données dans le modèle final :
- rassembler des attributs dans un attribut plus général
- ajouter un attribut (la valeur NULL est entrée par défaut pour les attributs non renseignés)
- associer des attributs à une autre classe
- ajouter une classe

Pour ce processus, nous avons décidé de garder certaines structures du Metathesaurus de l'UMLS qui permettent de définir des liens sémantiques et des liens inter-terminologies entres les termes ainsi que les concepts de l'UMLS.
Enfin, il nous a paru important de distinguer d'une part, les variantes lexicales propres à la terminologie et celles incluses dans un dictionnaire et d'autre part, les relations intra et inter-terminologies.

Le modèle général a ainsi été défini selon 7 classes (voir figure 3.4) (voir Annexes - Modèles unitaires) :
- **Classe des Concepts UMLS** : cette classe indique, pour chaque code des différentes terminologies, les liens vers les concepts UMLS (quand ils existent donc seulement pour les codes MeSH (exclut les termes spécifiques CISMeF), CIM10 et SNOMED). Cette classe est inspirée de la table **MRCONSO** (contenant les sources et les noms des concepts dans le Metathesaurus de l'UMLS - voir Annexes A).

2. `http://www.nlm.nih.gov/research/umls/metab.html`

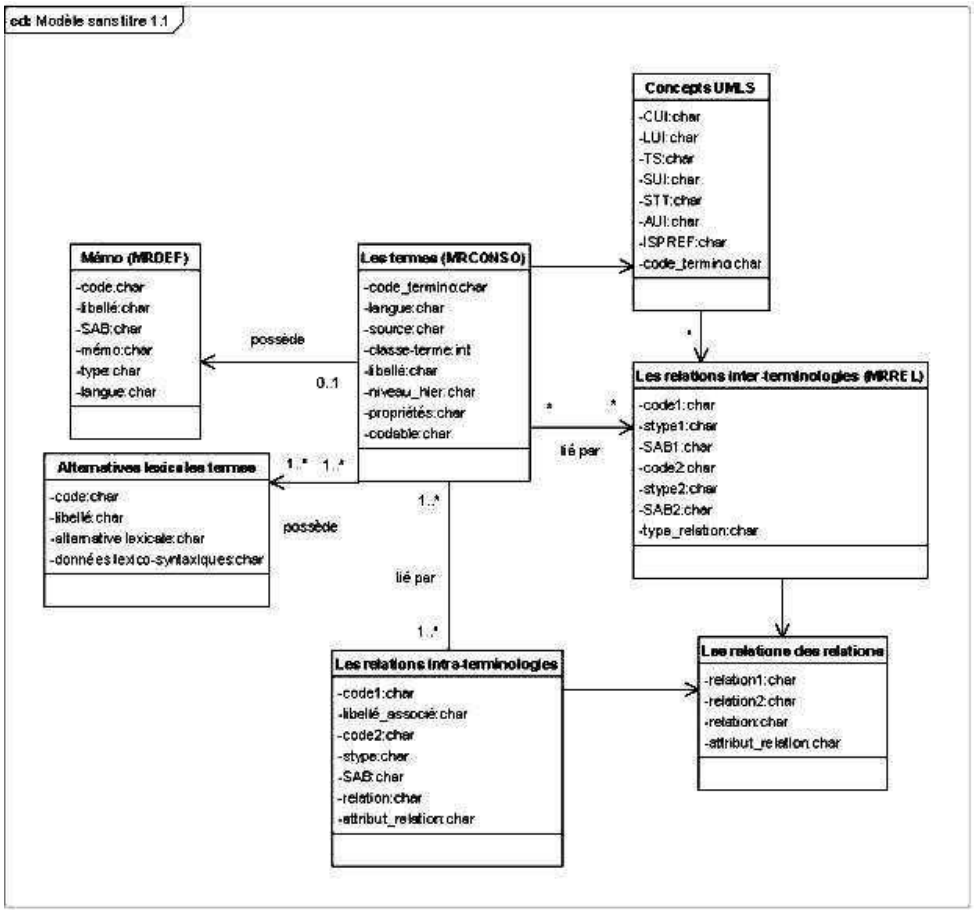

FIGURE 3.4 – Diagramme de classes représentant le modèle général au formalisme UML

- **Classe des Termes** : cette classe réunit tous les termes de chaque terminologie. Cette table a été inspirée de la table MRCONSO (contenant les sources et les noms des concepts dans le Metathesaurus) de l'UMLS. Elle regroupe toutes les classes décrivant les termes pour chaque terminologie : **Termes** de la CCAM, **Descripteur**, **Qualificatif**, **Type de ressource**, **Métaterme** du MeSH, **Termes** de la SNOMED et enfin **Termes systématiques**, **Descripteurs** et **Inclusions** de la CIM10.
- **Classe des Relations inter-terminologies** : cette classe renseigne toutes les relations qui peuvent exister entre deux termes de terminologies différentes. Cette table a été inspirée par la table MRREL (Related Concepts) de l'UMLS. Elle inclut les transcodages entres terminologies : CCAM-MeSH et CCAM_MTCISMeF (voir section 5.8.1), SNOMED-CIM10, TUV-MeSH, TUV-CIM10. Elle intègre aussi toutes les relations inter-terminologiques comprises dans l'UMLS : tels que les liens de transcodage SNOMED-CIM10, SNOMED-MeSH et MeSH-CIM10.
- **Classe des relations** : cette classe précise les relations secondaires qui peuvent exister entre les relations elles-mêmes. Elle est inspirée de la table MRHIER

(Computable Hierarchies) de l'UMLS.
- **Classe des Relations intra-terminologies** : cette classe renseigne toutes les relations qui peuvent exister entre deux termes d'une même terminologie. Cette table a été inspirée par la table **MRREL** (Related Concepts) et **MRHIER** (Computable Hierarchies) de l'UMLS. Elle inclut les classes **Hiérarchie**, **Voir aussi**, **Actions pharmacologiques** du MeSH, **Assiociations médicales** et **Hiérarchie** de la CCAM, **Hiérarchie** et **Références** de la SNOMED, **Hiérarchie**, **Inclusions**, **Dagstar** et **Exclusions** de la CIM10 et enfin **Relation_concept** du TUV. Elle inclut également toutes les relations sémantiques comprises dans l'UMLS pour une même terminologie.
- **Classe des Mémos** : cette classe renseigne toutes les notes et mémos rattachés aux termes des différentes terminologies. Elle inclut les classes **Mémo** et **Références** de la CIM10, **Notes** et **Définitions** du MeSH et **Notes** de la CCAM. Cette classe est inspirée par la table **MRDEF** de l'UMLS.
- **Classe des Alternatives lexicales termes** : cette classe réunit toutes les variations, flexions et synonymes des termes inclus dans le dictionnaire général. Elle inclut la classe **dictionnaire** du MeSH.

3.5 Création de libellés d'indexation

Les différentes méthodes proposées par notre outil F-MTI sont basées sur les libellés des termes de nos terminologies. Ces libellés ne sont pas élaborés, à l'origine, pour faciliter leur indexation. Leur forme est le plus souvent dictée par une stucture logique capable de rendre compte du sens du terme et donc éliminer toute ambiguïté. Elle peut aussi être élaborée afin de faciliter la recherche du terme dans la terminologie. Tous les libellés doivent aussi rendre compte d'une certaine homogénéité.

Une étape nous a ainsi paru nécessaire pour veiller au bon appariement des termes et des phrases. Cette étape consiste à créer, pour chaque libellé de chaque terminologie, un libellé d'indexation qui facilite son indexation.

Il s'agit d'un travail long qui peut être légèrement différent selon les terminologies. Nous l'avons réalisé en guise d'illustration sur la terminologie de la CIM10 (il sera bien entendu nécessaire dans l'avenir de le faire pour les autres terminologies).

Si l'on considère les termes de la CIM10, nous pouvons trouver des expressions comme «sans précision», «sans autre indication» ou «classés ailleurs» qui apparaissent dans certains termes mais ne seront jamais retrouvés dans une phrase. Par exemple, le terme «Angine de poitrine sans autre précision» (de code A10.0) peut être inscrit dans le document, parmi ses nombreuses formes, sous la forme «angine de poitrine». Si le mot «précision» est retenu dans le sac de mots du terme A10.0 avec les mots «angine » et «poitrine», alors l'appariement avec une phrase contenant la notion d'angine de poitrine pourra très rarement être obtenu puisque tous ces éléments ne pourront être retrouvés que dans de rares cas ensemble dans la même phrase. L'expression «sans précision» doit donc être automatiquement éliminée des termes. Ces expressions qui permettent de préciser le sens d'un terme au sein d'une terminologie mais qui empêchent leur indexation doivent être éliminées des termes

avant la création des sacs de mots correspondants. Nous avons ainsi créé des libellés secondaires, dits libellés d'indexation, qui comprennent les libellés d'origine ainsi qu'un à plusieurs libellé(s) alternatif(s) (exemple : les libellés d'indexation de A10.0 sont «angine de poitrine sans autre précision» et «angine de poitrine»). Ce sont ces libellés qui sont pris en compte par les trois méthodes d'indexation.

Nous avons identifié différents types d'expressions à traiter :

- Les éléments de classification tels que «cause de maladie classé en» ou «classés ailleurs» sont inutiles pour l'indexation et seront éliminés grâce à une liste d'expressions dites «vides». Cette liste contient 63 expressions.

- Une forme négative en «non» (exemple : le terme «néphrite tubulo-interstitielle, non précisée comme aiguë ou chronique»). Attention pour les termes comme «rayonnement non ionisant» l'expression «non ionisant» fait partie intégrante du terme et sera retrouvée dans sa forme textuelle. Les premiers cas ont été automatiquement traités grâce à la liste des expressions vides. Les termes du second cas ne sont pas traités.

- Une forme d'exclusion : «sauf», «sans» «SAI» [3], «sans précision», «sans autre indication», «sans mention de confirmation bactériologique», «sans siège/ localisation/ niveau précisé».
  - Les expressions récurrentes ont été recueillies dans la liste des expressions vides puis éliminées pour tous les termes de la CIM10.
  - Les expressions «sans...» peuvent indiquer des éléments de précision pour le terme. Généralement, il existe dans la terminologie le terme avec l'expression inverse «avec...» (exemple : les termes S90.1 et S90.2 «Contusion d'un (des) orteil(s) sans lésion de l'ongle» et «Contusion d'un (des) orteil(s) avec lésion de l'ongle»). Ces cas sont traités en éliminant automatiquement l'expression «sans...». Ces expressions sont toujours en fin de terme, c'est pourquoi le programme informatique élimine le mot «sans» et tout ce qui suit. Le libellé d'indexation du premier terme est «contusion d'un orteil» pour le deuxième terme il est égal au libellé d'origine. Le deuxième terme ne sera retrouvé que s'il est précisé dans la phrase qu'il y a lésion de l'ongle, si rien n'est précisé c'est le premier terme qui sera retrouvé.
  - Les expressions en «sauf» indiquent des exceptions (exemple : S92 «Fracture du pied, sauf la cheville»). La plupart de ces termes possèdent des fils plus précis (ainsi le terme S92 a comme fils les différentes fractures du pied qui ne sont pas de la cheville dont le terme S92.9 «fracture du pied, sans précision»). En cas de fracture du pied le terme S92.9 sera donc indexé, il n'est pas nécessaire ici d'opérer de traitement pour le terme S92.
  - Si le terme ne possède pas de fils alors nous éliminons l'expression «sauf...» de la même facon afin que le terme puisse être indexé.

- La plupart des formes d'exclusion sont accompagnées de formes d'inclusion

3. Abréviation de «sans autre indication».

(exemple : le terme S82 «Fracture de la jambe, y compris la cheville»). Là encore si le terme possède deux fils exprimant la fracture de la jambe et la fracture de la cheville aucun traitement n'est nécessaire. Sinon il est nécessaire de créer deux libellés d'indexation «fracture de la cheville» et «fracture de la jambe». Ceci a été réalisé automatiquement (même méthode que pour les alternatives).

- Les flexions : certaines variations de mots peuvent être explicitées (exemple : pour le terme «plaie ouverte d'un (des) orteil(s) sans lésion de l'ongle»). Les marques de flexion (s), (des), etc. sont éliminées automatiquement afin de créer le libellé d'indexation correspondant.

- Des alternatives du type «ou» ou des synonymes entre parenthèses : ces alternatives peuvent constituer plusieurs libellés d'indexation possibles pour un même terme (exemple : pour le terme «abscence ou perte de désir sexuel» deux libellés d'indexation alternatifs sont créés «abscence de désir sexuel» et «perte de désir sexuel») (autre exemple : pour le terme «pian plantaire humide (pian-crabe)», nous avons deux libellés d'indexation alternatifs «pian plantaire humide» et «pian-crabe»).
  - Les alternatives en «ou» ont été traitées automatiquement puis validées à la main. Le programme permet d'extraire les deux expressions entourant le «ou». Le premier libellé d'indexation conserve la première expression (le «ou» et la deuxième expression sont éliminés). Le deuxième ne conserve que la deuxième expression (la première expression et le «ou» sont éliminés).
  - Les mots entre parenthèses ne sont pas à confondre avec certaines précisions qui sont aussi entre parenthèses et qui sont à conserver (exemple : pour le terme «maladie par VIH à l'origine d'adénopathies généralisées (persistantes)»), ou des élément optionnels ou des alternatives. Il faut donc, dans un premier temps, pour traiter ces termes les faire analyser par un expert qui va déterminer dans quelle catégorie se place le terme. Puis un traitement informatique peut être mis en place pour chaque cas. Nous avons traité une centaine des ces termes en les sélectionnant manuellement puis en les traitant automatiquement mais les autres nécessitent l'intervention d'un expert et seront traités dans le futur.
  - Pour les termes contenant des expressions en «et» (exemple : «Lésions traumatiques superficielles multiples de la cheville et du pied»), nous considérons que cela implique des éléments indissociables, aucun libellé d'indexation alternatif n'est donc créé. Malheureusement dans certains cas, le «et» peut avoir le sens «ou» de la même facon ces cas devront, dans le futur, être repérés par un expert et traités comme des alternatives.

- D'autres expressions peuvent poser problème telles que :
  - «localisation unique» ou «deux doigts ou plus» (exemple : le terme «amputation de deux doigts ou plus (complète) (partielle)»). Ces problèmes ne peuvent être résolus d'une manière simple, automatique et rapide. Ils pour-

raient être traités dans l'avenir, grâce à des transducteurs pour certains mais d'autres solutions restent à envisager pour résoudre l'ensemble de ces cas de manière automatique.

– Pour les expressions de type «autre» (exemple : M20.5 «Autres déformations d'(es) orteil(s)») qui n'ont pas de fils pouvant préciser les «autres» formes, aucune solution, à part celle d'éliminer ce terme si un de ses frères est retrouvé, ne peut être trouvée en utilisant cette seule terminologie. En effet, rien ne nous permet de distinguer dans une phrase où l'expression «déformation de l'orteil» est présente s'il s'agit d'une «Déformation d'(es) orteil(s), sans précision» ou d'une autre déformation. La solution est de rechercher dans les liens de transcodage entre le terme M20.5 et toutes les autres déformations de l'orteil, non répertoriées dans la CIM10, appartenant à la SNOMED 3.5 par exemple. Cette opération est réalisée grâce à la multi-terminologie (voir la section Restriction à une ou plusieurs terminologies).

Les traitements sur la CIM10 ont abouti à la création de 41 258 libellés d'indexation différents (pour 19 155 codes et 31 222 libellés à l'origine).

3.6 Conversions des fichiers

Les documents traités par F-MTI sont de formats différents. Les comptes rendus au CHU de Rouen sont rédigés à l'aide du logiciel Microsoft Word. Ces fichiers sont au format «.doc». Les RCP sont envoyés par l'AFSSAPS, au VIDAL au format «.pdf» (à partir de fichiers Word). Il est prévu, dans le futur, de les envoyer au format XML.

Enfin pour les ressources Web intégrées à CISMeF, le contenu du site qui peut être obtenu à partir de l'URL, peut être de multiples formats (HTML, PDF, PPT etc...).

Le choix d'un format commun et facile à traiter par un programme informatique s'est porté sur le format texte «.txt». Afin de convertir de multiples formats en fichier texte, il existe différents outils tels que : pdftotxt [4], un programme Microsoft Word de conversion des fichiers word en fichier texte [5]. Les fichiers XML sont facilement transformables en texte. En revanche, il n'existe aucun outil de ce style permettant de convertir les fichiers «.ppt», ou les «.pdf» protégés.

3.7 Les unités d'indexation

Comme explicité section 2.5.3.2, certains outils prennent en compte des groupes nominaux. Ces groupes nominaux peuvent être extraits grâce à l'outil SYNTEX [6]

4. Crée par verypdf.com qui concoit des logiciels autour de l'exploitation des PDF. Téléchargement accessible ici : `http://www.verypdf.com/download/download.htm`

5. Accessible *via* le logiciel Microsoft Word

6. Un analyseur syntaxique automatique du francais. Il permet d'analyser les dépendances syntaxiques et ainsi d'extraire des groupes (verbaux, nominaux et adjectivaux)

[Bourigault00].

Il nous semblait important de garder toute la phrase, les verbes pouvant être une source d'information utile et les termes MeSH pouvant être extraits à partir d'informations contenues à la fois dans le sujet et les compléments d'objets, donc dans différents groupes d'une même phrase. Par exemple, pour la phrase «L'enfant a été traité de manière préventive pour des convulsions fébriles », le terme TUV «Convulsion fébrile chez l'enfant, traitement préventif (de la)» ne pourra être extrait qu'en considérant l'ensemble de la phrase comme unité d'indexation. Nous avons ainsi choisi comme unité d'indexation la phrase.

Pour chaque document à indexer, nous avons identifié les phrases ainsi que leurs contextes c'est-à-dire la rubrique et le paragraphe auxquels elles appartiennent.

3.7.1 Identification des rubriques

Connaître pour une phrase la rubrique auquelle elle appartient permet de définir l'emplacement de la phrase dans le document, et également, le contexte d'indexation et certains éléments qui pourront être utiles à l'indexation.

Nous avons identifié les rubriques d'intérêt pour les comptes rendus hospitaliers et les RCP :

- Les comptes rendus hospitalier sont formés de plusieurs rubriques (voir section 2.4.3.1), ils peuvent être différentes selon les secteurs d'hospitalisation au sein d'un même hôpital et changer au cours des années. Nous avons identifié ces rubriques pour les comptes rendus du secteur Cardiologie et Pneumologie. Ces secteurs ont été retenus car nous disposions d'un expert dans ces deux domaines. Ces rubriques peuvent aussi êtres exprimées de facon différentes selon les individus, nous avons identifié toutes les déclinaisons rencontrées. Les rubriques identifiées sont les suivantes : Secteur d'hospitalisation, Motif d'hospitalisation, Antécédents médicaux, Examens et Histoire de la maladie, Examens cliniques, Électrocardiogramme, Radio Thoracique, Biologie, Évolution, Conclusion, Traitement, Conduite à tenir.
- Les rubriques d'intérêt pour l'indexation d'un RCP sont (voir section 2.4.2.1) : Indication, Contre-indication, Mise en garde et précautions d'emploi, Grossesse et allaitement, Effets secondaires et Surdosage.
- Pour les sites Web et les autres documents, les rubriques ne peuvent être connues à l'avance. Une partie du programme de F-MTI permet de rentrer de nouvelles rubriques afin qu'elles soient identifiées (il suffit d'ajouter les noms). De plus, un travail avait été réalisé par A. Névéol pour extraire le titre des ressources Web à partir d'une URL [Névéol05a], celui-ci est repris dans nos travaux.

3.7.2 Identification des paragraphes

Connaître pour une phrase, les autres phrases appartenant au même paragraphe permet de définir non seulement l'emplacement de la phrase dans le document mais, aussi, les éléments utiles à l'indexation. En effet, un paragraphe est constitué d'un

certain nombre de phrases reliées entre elles par une unité d'information, une même thématique. Un terme d'une terminologie peut être explicité tout au long d'un même paragraphe et ainsi être extrait automatiquement d'éléments provenant de plusieurs phrases à l'intérieur de celui-ci.

Dans les fichiers XML, un paragraphe étant contenu entre les balises et , il est facile de l'extraire.

Pour les fichiers textes, un paragraphe se termine par un point suivi d'un saut de ligne. La difficulté pour les fichiers textes provient des fichiers issus de conversion : les paragraphes peuvent être complètement désorganisés. De fait, nous avons dû les reconstituer afin de les identifier parfaitement.

3.7.3 Identification des phrases

Des travaux existent sur le découpage en phrase [Pappa04]. Parmis ces travaux, [Friburger00] présente un transducteur [7] INTEX [Silberztein93] pour déterminer la position d'une fin de phrase mais celui-ci ne permet pas de récupérer la phrase mais bien la position de celle-ci. Nous avons donc créé un transducteur NooJ [8] [Silberztein04] (voir figure 3.5), en nous inspirant du transducteur INTEX que nous nous sommes procuré auprès des auteurs. Ce transducteur a été créé en collaboration avec M. Silberztein créateur du logiciel NooJ qui a intégré de nouvelles fonctionnalités dans NooJ afin de rendre la création de ce transducteur possible.

Pour trouver où se situe le début et la fin d'une phrase, nous pouvons nous baser sur la ponctuation. Mais cela n'est pas si simple, le point est un signe ambigu [Silberztein93] [Dister97]. Le transducteur créé permet d'extraire les phrases d'un texte qui correspondent aux critères suivants :

- Le début d'une phrase peut être marqué par :
  - un début de paragraphe (<^>),
  - un tiret ou un guillemet,
  - un tiret suivi d'un guillemet,
  - un mot commencant par une majuscule (<CAP>), entièrement en lettres majuscules (<UPP>) ou un nombre (<NB>),
  - une exception (voir figure 3.6) : les nombres décimaux (exemple : «3.14»), les sigles (exemple : «C.G.T») (voir figure 3.7), les titres de personnes (exemple : M. Henri) (voir figure 3.8) et les abréviations (exemple : «cf. ») (voir figure 3.9) constituent des exceptions car ils présentent un signe de fin de phrase (le point).
- Une phrase peut contenir :
  - n'importe quel mot (<WF>), des nombres (<NB>), certaines ponctuations excepté «.!?!;» qui constituent un signe de fin de phrase (<P-MP= «[.!?;]

7. Un transducteur est un patron d'extraction

8. NooJ a repris et amélioré les fonctionnalités d'INTEX. NooJ est un environment de développement linguistique qui inclut des dictionnaires et des grammaires, et peut traiter des corpus en temps réel. Il permet aux utilisateurs de créer leurs propres dictionnaires, et leurs propres grammaires ainsi que des patrons d'extraction (ou transducteurs) syntaxiques ou morphologiques. Il est téléchargeable *via* `http://www.NooJ4nlp.net/`

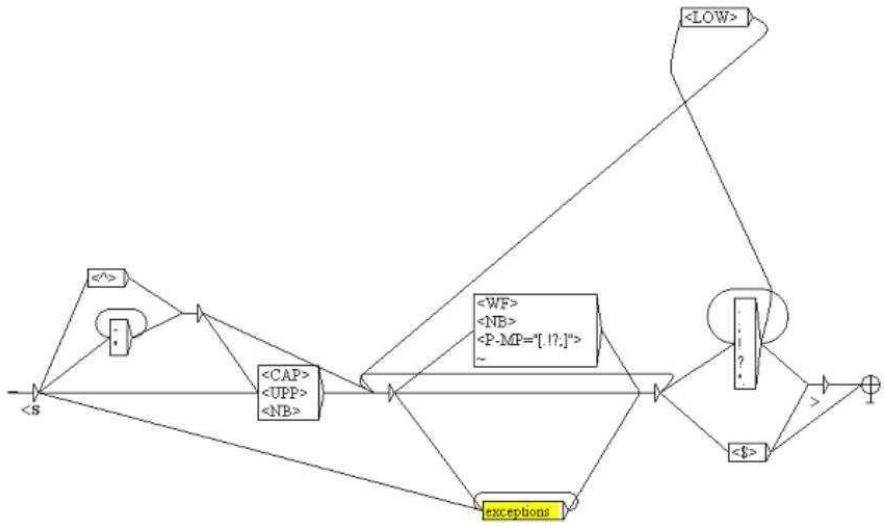

FIGURE 3.5 – Transducteur de phrases réalisé avec le logiciel NooJ

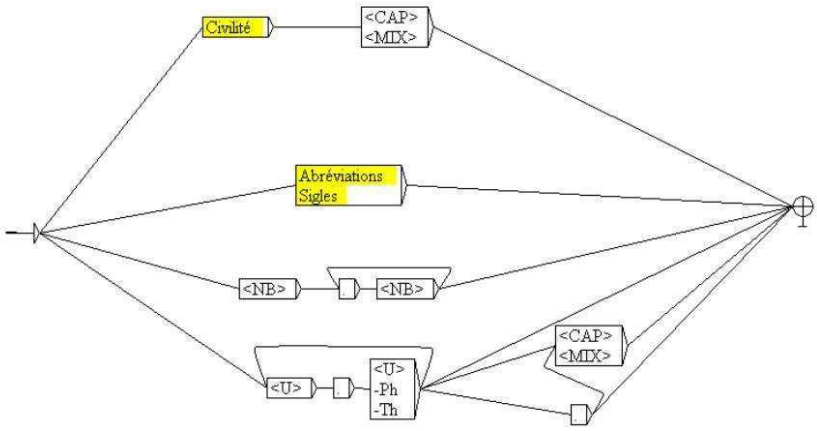

FIGURE 3.6 – Sous-graphe des exceptions réalisé avec le logiciel NooJ

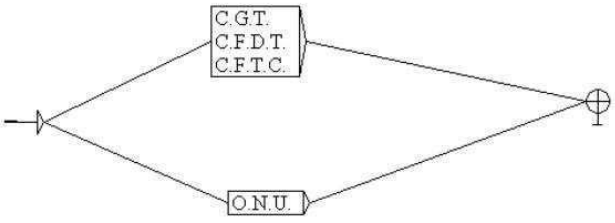

FIGURE 3.7 – Sous-graphe des sigles réalisé avec le logiciel NooJ

»), des caractères spéciaux comme le ~,
  - des exceptions.
- La fin d'une phrase peut être marquée par :
  - une ponctuation de fin de phrase (un point, point-virgule, point d'exclamation, point d'interrogation, guillemet point) sauf si elle est suivie d'un mot

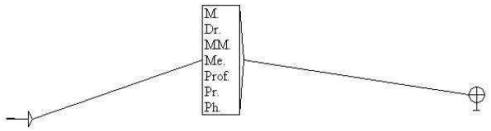

FIGURE 3.8 – Sous-graphe des titres de civilité réalisé avec le logiciel NooJ

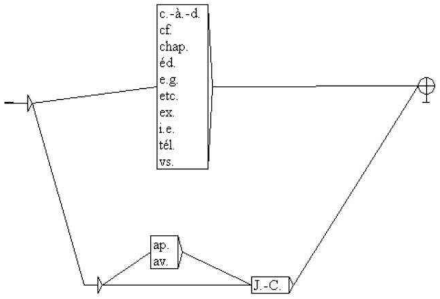

FIGURE 3.9 – Sous-graphe des abréviations réalisé avec le logiciel NooJ

en minuscule,
- une fin de paragraphe (cas des phrases débutant par un tiret)

3.8 Méthodes mises au point

3.8.1 Algorithme du sac de mots

3.8.1.1 Origine

L'algorithme du sac de mots est utilisé pour indexer les documents. Cet algorithme a été utilisé à l'origine par P. Zweigenbaum [Zweigenbaum01] dans le catalogue CISMeF pour retranscrire les requêtes de l'utilisateur, qui sont faites en langage naturel, en termes MeSH et, ainsi permettre au système de proposer des documents correspondant à la requête. Cet algorithme reposait sur des données morphologiques. Il a ensuite été modifié pour de ne plus utiliser de données morphologiques mais la phonémisation [Soualmia04] puis la désuffixation. Nous avons aussi mis en place cet algorithme pour l'indexation automatique des ressources (sur le titre) dans le catalogue CISMeF avec la participaion d'A. Névéol [Névéol07b].

Cet algorithme est efficace pour le traitement des requêtes, nous avons voulu le tester pour l'indexation d'un document (en l'occurrence d'un ensemble de phrases) et non plus d'une requête ou d'un titre. Nous avons aussi voulu le tester pour l'indexation multi-terminologique en CIM10, SNOMED 3.5, CCAM, MeSH et TUV et non plus uniquement en MeSH.

3.8.1.2 Principe de la méthode

Le but est d'apparier des termes issus d'une ou plusieurs terminologies à une phrase. Pour cela, dans un premier temps, nous avons déterminé quels éléments dans la phrase pouvaient nous permettre de reconnaître un ou des termes d'une terminologie (constitution du sac de mots de la phrase). Dans un deuxième temps, nous avons déterminé pour chaque terme des terminologies, les éléments signifiants qui pouvaient être reconnus dans une phrase (constitution du sac de mots du terme). Le même algorithme est appliqué à la phrase comme aux termes des terminologies, les deux entités étant exprimées sur le même plan, nous pouvons alors comparer le sac de mots issus de la phrase et ceux issus des termes et ainsi apparier des termes à la phrase (voir figure 3.10). Ces termes sont alors proposés pour l'indexation de la phrase. Lorsque toutes les phrases du document sont traitées, une proposition préliminaire d'indexation du document peut être faite.

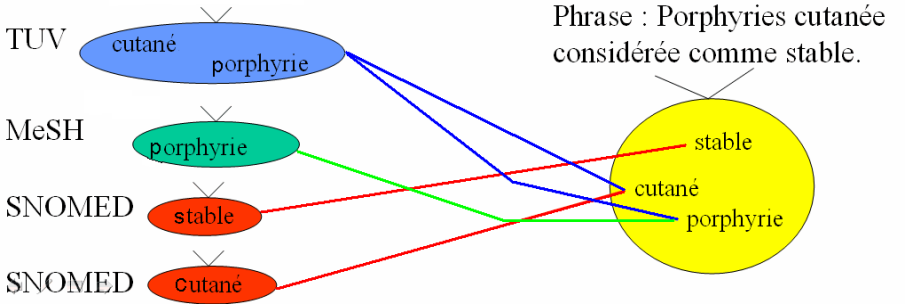

FIGURE 3.10 – Comparaison du sac de mots issus de la phrase et ceux issus des termes

3.8.1.3 L'algorithme du sac de mots

Nous avons modifié l'algorithme utilisé au départ pour la traduction des requêtes [Soualmia04]. Son fonctionnement est le suivant (voir figure 3.11) :

3.8.1.3.1 Constitution des sacs de mots

Le sac de mots contient tous les mots normalisés pertinents d'une phrase ou d'un libellé d'une terminologie dans un ordre indéfini.

Normalisation de la phrase ou du terme : découpage en mots (ou tokenisation)

Il faut d'abord définir ce qu'est un «mot» pour pouvoir les identifier automatiquement. Il est possible d'aborder la question de la définition du mot de deux facons : soit par la définition de critères de segmentation de la phrase en mots, soit par la définition de la structure interne du mot [Molino85]. Ici, nous nous intéressons à la segmentation. Dans le langage courant, un mot est une suite de caractères graphiques formant une unité sémantique et pouvant être distingué par un séparateur (un espace). Cette définition est très sommaire, en fait, beaucoup d'éléments sont à prendre

Algorithme du sac de mots

Entrée : une phrase (avec son contexte : rubrique et paragraphe), Liste_des_mots_vides.txt, Liste_des_expressions_vides.txt, Listes_des_mots_normalisés_non_pertinents.txt, Table_les_termes (avec leurs sac de mots), terminologies_à_utiliser, terminologies_en_sortie, mode_normalisation_mot, Table_les_relations_inter_terminologies.
Sortie : Liste de termes d'indexation

Début
#Ajout du contexte
Si (rubrique = "antécédents") **alors**
phrase= ajout (phrase , " antécédent ")
FinSi

#phase de normalisation de la phrase et découpage en mots
phrase_normalisée<-minusculiser (phrase);
phrase_normalisée <-traduction(phrase_normalisée,ÉÈÊËÀÂÄÎÏÙÛÜÑÒÖÔÓÇ,éèêëaâäîïùûüñòöôóç);
phrase_normalisée <-remplacer(phrase_normalisée,Œ,oe);
phrase_normalisée <-remplacer(phrase_normalisée,œ,oe);
phrase_normalisée <-traduire(phrase_normalisée,éèêëäâäîïùûüòöôóõ,eeeeaaaiiuuuuoooooo);
phrase_normalisée <-enlever_les_doubles_espaces(phrase_normalisée);
phrase_normalisée <-normaliser_les_nombres(phrase_normalisée);
phrase_normalisée <-normaliser_les_unités(phrase_normalisée);
phrase_normalisée <-normaliser_certains_caractères(phrase_normalisée);
phrase_normalisée <-éliminer_ponctuation1(phrase_normalisée);
phrase_normalisée <-éliminer_expressions_vides(phrase_normalisée);
phrase_normalisée <-éliminer_mot_vides(phrase_normalisée);
phrase_normalisée <-éliminer_ponctuation2(phrase_normalisée);

#Découpage en mots
Liste_mots<-découper_entre_les_espaces(phrase_normalisée);

#désuffixation avec l'une des trois méthodes ou lemmatisation (dépend du paramètre en entrée)
Pour chaque mot **de** Liste_mots **faire**
mot_normalisé=mode_normalisation_mot(mot);
FinPour

#enlever les doublons
Liste_mots_normalisés=éliminer_doublons(Liste_mots_normalisés);

#enlever les mots_normalisés non pertinents
ouvrir(Listes_des_mots_normalisés_non_pertinents.txt);
Pour chaque mot_normalisé **de** Liste_mots_normalisés **faire**
Si (mot_normalisé n'existe pas dans Listes_des_mots_normalisés_non_pertinents.txt,) **alors**
Liste_mots_normalisés =éliminer(Liste_mots_normalisés, mot_normalisé);
FinSi
FinPour

#Ranger par ordre alphabétique
Sac_de_mots=ranger par ordre alphabétique(Liste mots normalisés);
#Production des combinaisons
ouvrir(Table_les_termes);
Pour taille=taille(Sac_de_mots) à taille=**1 faire**
tableau_combinaisons=combinaisons(Sac_de_mots,taille);
Pour chaque combinaison de tableau_combinaisons **faire**
Si combinaison = sac_de_mots d'un terme dans Table_les_termes **alors**
proposition d'indexation=ajouter(proposition_d'indexation, code_terme);
#Recherche transcodage
Pour chaque combinaison de tableau_combinaisons **faire**
Si code a des transcodages dans Table_les_relations_inter_terminologies **alors**
proposition_d'indexation=ajouter(proposition_d'indexation, code_terme);
FinSi
FinSi
FinPour
FinPour

Retourner(proposition_d'indexation)
Fin

FIGURE 3.11 – Algorithme du sac de mots

en compte. Voici quelques règles que nous avons adoptées (celles-ci constituent déjà un changement dans l'algorithme d'origine) :

- Un mot peut-être composé, accentué, il peut être un sigle ou un nom propre.
- Les ponctuations ne constituent pas les mots mais sont de bons indicateurs de séparation de mots. Elles seront éliminées en deux temps, excepté pour les tirets qui seront maintenus pour les mots composés.
- Un mot est séparé d'un autre mot par un espace ou une apostrophe (exemple : l'expression «l'angine» contient deux mots : «l'» et «angine»).
- Un nombre est un mot. Il faut donc éliminer les espaces qui peuvent séparer le chiffre des milliers des autres chiffres. De plus, les décimaux peuvent contenir une virgule ou un point qui font partie intégrante du nombre. Il faut donc veiller à ce que cette ponctuation ne soit pas éliminée.
- Nous avons considéré que les mesures pouvaient avoir des formes très diverses et n'avaient de sens qu'en juxtaposant le chiffre et l'unité de mesure. Pour des raisons de normalisation le terme «nombre unité» sera donc considéré comme un seul mot.

Élimination des éléments non pertinents

La complexité de l'appariement (voir section appariement) est directement lié à la taille du sac de mots de la phrase, c'est la raison pour laquelle le sac de mots est réduit aux mots les plus signifiants et pertinents :

- Nous avons éliminé les mots vides. Un mot vide est un mot non significatif figurant dans un texte. En recherche documentaire, les mots vides sont des mots qui sont tellement communs qu'il est inutile de les indexer ou de les utiliser dans une recherche. Les mots vides sont aussi générateurs de bruit, donc il est recommandé de les éliminer (selon la loi de Zipf [Zipf49] et Luhn [Luhn58]). Nous disposions, dans l'équipe, d'une liste de mots vides obtenue à partir de Lexique [9], créée par L. Soualmia et utilisée dans l'algorithme du sac de mots d'origine [Soualmia04]. Cette liste a été entièrement retravaillée afin d'y ajouter des mots vides et d'éliminer les mots pouvant être utiles à l'indexation (comme les termes de l'axe G de la SNOMED qui contient les termes de liaison) et éliminer les mots vides inutiles car rarement retrouvés («boum» ou encore «snyff»). Nous avons consulté plusieurs bases de données sur Internet pour trouver de nouveaux mots vides. Les mots vides considérés sont :
  - les pronoms possessifs (exemple : «mon»)
  - les conjonctions (exemple : «mais»)
  - les déterminants (exemple : «du»)
  - les interjections (exemple : «diantre»)
  - les prépositions (exemple : «durant»)
  - les pronoms personnels (exemple : «il»)
  - les pronoms possessifs (exemple : «leur»)
  - les pronoms relationnels (exemple : «auquel»)

9. Lexique fournit une base de données lexicales avec des estimations de fréquences et des formes fléchies accessibles *via* `http://www.lexique.org`

  - les symboles et locutions (exemple : «ainsi»)

  En plus des mots vides, il existe des expressions vides (exemple : «tout d'abord»). Une liste d'expressions vides a ainsi été crée et ajoutée à la liste des mots vides. La liste des mots vides est ordonnée afin d'éliminer en premier lieu les expressions les plus longues.

  La liste d'origine contenait 1 422 mots vides. La nouvelle liste contient 1 267 entrées.
- Lors de l'appariement toutes les combinaisons de mots sont générées les doublons sont donc inutiles et aussi éliminés du sac de mots.
- Dans le sac de mots présentant les mots signifiants d'une phrase que l'on désire indexer, certains mots sont non pertinents car jamais retrouvés dans aucun terme appartenant aux terminologies utilisées. Nous avons ainsi préparé la liste complète des mots normalisés (stèmes ou lemmes - voir section désuffixation et lemmatisation) présents dans au moins un terme des différentes terminologies. Les stèmes sont au nombre de 61 274 pour l'ensemble des cinq terminologies et sont typés selon leurs terminologies d'origine. Lors de l'élaboration du sac de mots de la phrase, les mots vides appartenant à notre liste et les lemmes ou stèmes n'appartenant pas à notre liste seront éliminés afin d'éliminer les ambiguïtés et pour ne pas surcharger le sac de mot pour une exécution rapide du programme.

Normalisation de la phrase ou du terme : désuffixation ou lemmatisation

En informatique, il est difficile pour un programme de savoir que deux mots, l'un issu d'une phrase et l'autre d'un terme d'une terminologie, sont deux formes textuelles d'un même mot. C'est la raison pour laquelle une normalisation des mots est nécessaire.

Les mots sont tout d'abord rendus à leurs formes minuscules. On élimine ainsi les variations dues à la position dans la phrase (mot débutant la phrase), aux différents usages d'écriture [10] ou aux normes d'écriture pour les différentes terminologies. Par contre ils sont un bon indicateur des noms propres (mots invariants) et sigles qui demanderaient un traitement particulier. Il serait intéressant de prendre en compte ces formes particulières dans une prochaine version de notre outil (voir discussion et perspectives).

Les caractères spéciaux doivent aussi être normalisés comme par exemple les formes attachées «oe».

L'algorithme du sac de mots utilise, dans CISMeF, la désuffixation dans un but de recherche d'information. La désuffixation cherche à rassembler les différentes variantes d'un mot autour d'un stème (ou radical) (exemple : «passer», «passe», «passes», «passa», «passant» ont le même stème «pass»). Nous pouvons ainsi traiter à la fois des cas relevant de la flexion (exemple : bactérie - bactéries) et de la dérivation (exemple : asthme - asthmatique). La technique repose généralement sur une liste de suffixes et un ensemble de règles de désuffixation construites *a priori*

10. En effet, entre les «usages actuels» et «les bons usages» des majuscules, il existe de grandes différences, comme le montre cet article `http://perso.univ-lyon2.fr/~poitou/Typo/t03.html`

qui permettent pour un mot de trouver son stème. L'algorithme de désuffixation utilisé dans CISMeF a été développé en interne (par B. Dahamna). Nous testerons trois méthodes de désuffixation (voir section 4.2.1) : l'algorithme de CISMeF, l'algorithme de Carry [Paternostre02] et le FrenchStemmer de Lucene [11] [Cutting04].

Une alternative à la désuffixation est la lemmatisation. La lemmatisation d'un mot consiste à en prendre la forme canonique : pour un verbe, ce verbe est mis à l'infinitif, pour les autres mots le mot est mis sous la forme masculin/ singulier [12]. Ici, «passe» et «passes» ont le même lemme «passe». Dans l'autre cas, «passer», «passa» et «passant» sont assignés au lemme «passer».

Les outils permettant la lemmatisation doivent, dans un premier temps, définir les données lexico-syntaxiques du mot avant d'être en mesure de déterminer le lemme de ce mot. Nous utiliserons dans cette catégorie le Sémiographe (de la société Mémodata) [13].

Selon la méthode, les accents peuvent être éliminés ou gardés. Lorsqu'ils sont pris en compte, ils permettent de discriminer des mots de sens différents (exemple : «sur» et «sûr»). Lorsqu'ils sont éliminés, ils permettent de rapprocher certaines formes telles un adjectif et un nom ou une forme conjuguée et un adjectif (exemple : «dégénère» et «dégénéré»).

Nous comparons ces deux méthodes de normalisation (désuffixation par rapport à lemmatisation) dans le cadre de l'indexation automatique (voir section 4.2.4).

D'autres méthodes existent (comme la phonémisation voir section 2.5.3.1.2), le choix s'est porté sur ces deux méthodes car nous disposions d'outils pour permettre leurs applications. En outre, ces méthodes ont des particularités différentes que nous voulions tester dans le cadre de l'indexation automatique. Enfin, les RCP, dossiers médicaux et ressources Web de qualité ne présentant que peu de fautes d'orthographe (contrairement aux requêtes entrées par les utilisateurs dans le catalogue CISMeF), la phonémisation ne nous a pas semblé être la méthode adéquate.

3.8.1.3.2 Appariement phrase/termes appartenant aux terminologies

Les termes et la phrase sont, par cette méthode, exprimés de la même facon : un ensemble de mots normalisés où l'ordre n'est plus pertinent. En programmation ceci revient à ranger par ordre alphabétique les mots normalisés constituant le sac de mots. Plusieurs termes cibles peuvent être nécessaires pour couvrir les différents lemmes ou stèmes d'une phrase.

Algorithmiquement, cela se traduit par la constitution de toutes les combinaisons (de taille 1 à n) de stèmes ou lemmes contenus dans le sac de mots de la phrase. Puis chaque combinaison est recherchée dans l'ensemble des sacs de mots pour chaque terme des différentes terminologies déterminés à l'avance et stockés dans notre base de données multi-terminologiques (voir section 3.4.2). Lorsqu'un sac de mots d'un

11. `http://lucene.apache.org/`

12. Les entrées d'un dictionnaire, comme le Larousse ou le Petit Robert par exemple, sont lemmatisées.

13. Utilisé dans le cadre du projet Vodel (http ://vodel.insa-rouen.fr/) issu d'une collaboration entre l'équipe CISMeF, la société Mémodata, le laboratoire Laseldi et la société EADS et le Sinequa Labs.

terme a été identifié dans la phrase alors le terme ainsi que les éléments d'information l'entourant (code, langue etc...) est ajouté à la proposition d'indexation finale (voir figure 3.12 pour un exemple).

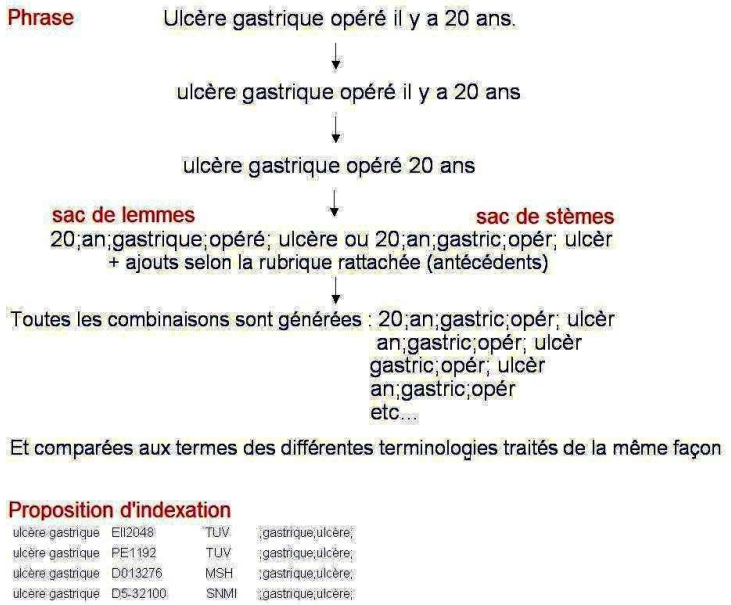

FIGURE 3.12 – Exemple d'indexation par l'algorithme du sac de mots d'une phrase extraite d'un compte-rendu d'hospitalisation

3.8.1.4 Implémentation

La méthode du sac de mots a été implémentée en Perl[Wall01]. Nous avons choisi ce langage informatique car il est parfaitement adapté pour toutes les tâches liées à la manipulation de chaînes de caractères (pour la normalisation et le découpage en mots cela est très utile). Il permet de stocker et récupérer les données dans une table de hachage de manière extrêmement rapide. Il existe de nombreux modules Perl (site CPAN), c'est-à-dire de nombreuses fonctions déjà implémentées. Enfin les outils permettant l'implémentation et l'exécution de programme Perl sont gratuits et disponibles sur Internet.

3.8.1.5 Algorithmique

L'accès à un terme dans une table de hachage a une complexité de O(1) en moyenne, quel que soit le nombre de termes dans la table, ce qui est très rapide.

La complexité de l'algorithme du sac de mots tient surtout à la complexité de la génération de l'ensemble des combinaisons pour la liste des mots signifiants de la phrase (complexité factorielle en $O(n!^2)$).

Dans notre algorithme, pour une phrase constituée de 25 mots signifiants, l'appariement en considérant des combinaisons de 6 mots peut prendre plus d'une minute. Nous nous sommes donc limité pour des soucis de temps de calcul à 5 mots signifiants

pour un terme. Seul les termes qui ont un sac de mots de moins de 6 mots pourront donc être indexés par notre algorithme du sac de mots. Ainsi même face à une phrase longue le programme mettra un temps raisonnable.

3.8.1.6 Points forts et points faibles de la méthode du sac de mots

La méthode du sac de mots est basée sur les mots. Cette méthode a l'avantage d'être simple. Lorsqu'elle utilise la désuffixation elle ne nécessite que peu de ressources : une table des suffixes et des règles à appliquer suffisent.

Contrairement à l'algorithme disponible dans l'équipe CISMeF, toutes les combinaisons de mots sont autorisées ce qui permet de retrouver dans la phrase «L'enfant de 5 ans et l'adulte sont asthmatiques», les termes «enfant de 5 ans asthmatique» et «adulte asthmatique». Alors que dans l'ancien algorithme seul le premier terme était retrouvé car l'indexation des termes les plus longs était privilégiée [14].

Cette méthode permet de trouver des termes dont l'ordre des mots n'est pas respecté dans la phrase. Par exemple, le terme TUV «enfant diabétique» est indexé pour la phrase «Nous avons décelé un diabète chez cet enfant». Malheureusement, cette méthode peut induire des erreurs en indexant un terme dont les mots peuvent être éloignés dans la phrase et ne pas correspondre au même terme. Exemple, pour la phrase «Ce médicament est contre-indiqué pour l'enfant diabétique et l'adulte asthmatique» l'algorithme du sac de mots indexe les termes «enfant diabétique», «enfant asthmatique», «adulte diabétique» et «adulte asthmatique». Les termes «enfant asthmatique» et «adulte diabétique» sont faux. Ceci peut-être amélioré en acceptant une distance limite entre 2 mots afin de privilégier les combinaisons de mots localement proches. Cette amélioration sera exploitée dans une version ultérieure de l'algorithme.

De plus, il est difficile d'identifier les négations pour cette méthode (voir section 3.9.1).

Et, l'indexation est limitée aux termes de moins de 6 mots signifiants et ne peut donc se faire sur l'ensemble des termes de nos terminologies.

3.8.2 Méthode du dictionnaire de termes

3.8.2.1 Méthode des dictionnaires DELA

Cette méthode est inspirée de l'approche TAL utilisée dans l'extracteur MeSH, MAIF [Névéol05a]. Dans le système MAIF, l'extraction des termes MeSH se fait à l'aide d'un dictionnaire de termes au format DELA. Le dictionnaire de termes contient les formes textuelles des termes : leurs dérivations (exemple : asthme - asthmatique), flexions (exemple : bactérie - bactéries) et synonymes. Le format de ce dictionnaire est inspiré du format DELA :

14. L'algorithme cherche d'abord les termes couvrant n mots puis n-1 mots puis n-2 mots etc...À chaque itération, si un terme est trouvé ses mots sont éliminés du sac de mots. Le mot «asthmatique» est ainsi éliminé après l'obtention du terme «enfant de 5 ans asthmatique», aux itérations suivantes il est ainsi impossible d'obtenir le terme «adulte asthmatique».

FormeTextuellePossibleDuTerme,LibelléDuTerme,InformationsDiverses

L'application de ce dictionnaire se fait *via* l'outil INTEX [Silberztein93] pour la recherche de termes d'indexation des ressources Web.

Nous avons voulu réappliquer cette méthode qui s'est révélée être efficace dans F-MTI pour nos cinq terminologies. Malheureusement, la constitution d'un dictionnaire est très fastidieuse à réaliser à la main. La création du dictionnaire DELA du thésaurus MeSH (22 995 termes dans sa version 2005) a constitué une part très importante de la thèse d'A. Névéol [Névéol05a]. Il nous a donc semblé très important de rendre la réalisation de ce genre de dictionnaire la plus automatique possible pour les terminologies SNOMED Internationale (environ 108 000 termes), CISMeF (25 000 termes dans sa version 2007), CIM10 (32 000 termes) et TUV (11 980 termes). Pour ce faire les résultats de nombreux travaux antérieurs (voir section suivante) ont été intégrés dans le dictionnaire de termes de F-MTI. De plus, nous avons élaboré une méthode permettant de recueillir automatiquement des variantes pour nos termes à partir de corpus.

Nous avons testé cette méthode sur les termes du TUV, avec l'idée sous-jacente de l'appliquer aux autres méthodes en cas d'obtention de bons résultats.

3.8.2.2 Variantes provenant de précédents travaux

Dans ce dictionnaire DELA, nous avons tout d'abord répertorié l'ensemble des variantes de termes connues de la terminologie TUV. Cela peut être des variantes flexionnelles, dérivationnelles ou des synonymes.

Exemple, pour le terme de référence TUV «affection des voies biliaires» ayant comme synonyme «affection de la vésicule biliaire» nous avons intégré dans le dictionnaire les entrées :
affection des voies biliaires, affection des voies biliaires,176+CC+PE+scientifique+TUV
affection de la vésicule biliaire, affection des voies biliaires,176+CC+PE+scientifique+TUV

Pour chaque entrée, il est indiqué le code (dans notre exemple «176»), le type (dans notre exemple, CC : concept complexe ou CE : concept élémentaire), l'étiquette sémantique (dans notre exemple «PE+scientifique») et la terminologie source (dans notre exemple, le TUV).

Pour compléter cette première liste, nous avons exploré les variantes lexicales et dérivationnelles créées lors de précédents travaux. Nous avons ainsi analysé le lexique médical unifié francophone créé dans le projet UMLF [Zweigenbaum03], le dictionnaire MeSH réalisé par A.Névéol [Névéol05a], et les lexiques créés dans le projet VUMeF [Darmoni03b]. Les variantes rattachées à des libellés équivalents TUV ont ainsi été recueillies et ajoutées au dictionnaire de termes.

3.8.2.3 Recueil automatique de nouvelles variantes

Les grammaires morphologiques et syntaxiques permettent de préciser la forme des variantes pour un terme (voir section 2.5.3.1). Nous avons utilisé ces grammaires afin de définir pour chaque terme un patron d'extraction capable d'extraire dans un

corpus [15] ses variantes dérivationnelles, flexionnelles et synonymiques qui viendront compléter le dictionnaire de termes TUV [16].

Un grand nombre de dérivations, flexions ou synonymes d'un terme ne sont que le reflet des dérivations, flexions ou synonymes des mots signifiants qui le composent reliés par des mots de liaison (principe que l'on retrouve dans la méthode du sac de mots). Par exemple, les formes textuelles du terme «diminution des facteurs de coagulation» peuvent être représentées par le transducteur présenté à la figure 3.13 (<diminution >, <facteur> et <coagulation> correspondent aux dérivations, flexions et synonymes des lemmes «diminution», «facteur» et «coagulation» contenus dans le dictionnaire de lemmes ; <MVP> est le dictionnaire des mots de liaison (983 mots vides sélectionnés pour cette tâche dont le tiret)). Ce transducteur ne prend pas en compte l'ordre des mots.

Ces transducteurs utilisent un dictionnaire de lemmes (38 219 entrées) qui

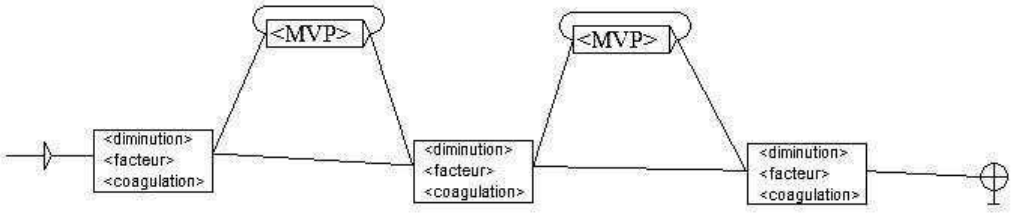

FIGURE 3.13 – Exemple de transducteur morphologique réalisé avec le logiciel NooJ pour le terme «diminution des facteurs de coagulation»

contient pour chaque lemme, identifié dans la terminologie TUV, ses flexions, dérivations et synonymes (asthmes,asthme,X). Ce dictionnaire a été créé à partir des dictionnaires médicaux et généraux :

- Morphalou [17] : ce lexique contient 590 020 formes fléchies associées à leurs lemmes.
- Lexique 3 [18] : Lexique 3 est une base de données qui fournit [19] pour 137 405 formes du francais le lemme associé (55 000 lemmes).
- MeSH [Névéol05a] : ce dictionnaire contient 44 856 variantes pour la terminologie MeSH.
- UNITEX [20] : possède un dictionnaire pour le francais de 683 824 mots avec leurs lemmes (102 073 lemmes).

15. Ensemble de documents

16. J'ai été aidée dans l'implémentation de cette tâche par Nicolas Rozanes, étudiant en master à L'INALCO

17. Le lexique Morphalou est un lexique ouvert des formes fléchies du francais. Les données initiales de Morphalou proviennent du TLFnome, la nomenclature du Trésor de la Langue Francaise. Voir `http://www.cnrtl.fr/lexiques/morphalou/`

18. Voir `http://www.lexique.org/`

19. Il fournit aussi les représentations orthographiques et phonémiques, la catégorie grammaticale, le genre et le nombre, les fréquences

20. UNITEX est un système de traitement de corpus qui permet de nombreux traitements proches de ceux proposés par NooJ. Ce système possède de nombreuses ressources téléchargeables sur l'Internet. Voir `http://www-igm.univ-mlv.fr/~unitex/`

- NooJ [Silberztein04] : le système NooJ comprend un dictionnaire pour le francais.
- UMLF [Zweigenbaum03] : dictionnaire médical de 23 141 formes fléchies associés à leurs lemmes.
- VUMeF [Darmoni03b] : dans le cadre de ce projet 2 742 variantes de concepts Vidal ont été produites.
- Le dictionnaire intégral du Sémiographe [Dutoit00] : il comprend 540 000 mots avec leurs lemmes et synonymes.

Toutes les variantes pour les unités de dosage et les chiffres ont complété ce dictionnaire de lemme.

L'application du transducteur de la figure 3.13 à un ensemble de documents nous permet d'extraire les variantes : «diminution des facteurs de la coagulation» et «diminution du facteur de coagulation». Ces variantes découvertes dans le corpus pourront venir compléter le dictionnaire de termes avec les entrées suivantes :

diminution des facteurs de la coagulation, diminution des facteurs de coagulation,1443+CE+ETAT ANOMALIE DES EXAMENS DE LABO+TUV

diminution des facteurs de la coagulation, diminution des facteurs de coagulation,1443+CE+ETAT ANOMALIE DES EXAMENS DE LABO+TUV

3.8.2.4 Constitution des transducteurs

Un ensemble de 33 719 termes provenant du Vidal (termes, concepts, variantes et synonymes TUV ainsi que les termes de recherche et les groupes d'indication) a été traité.

La constitution d'un transducteur dans le logiciel NooJ s'effectue manuellement. Afin de traiter notre ensemble important de termes, nous avons développé une méthode automatique permettant de générer les 33 719 transducteurs (voir figure 3.15).

Les termes sont, dans un premier temps, traités par l'algorithme du sac de mots afin de définir la liste des lemmes pour chacun. Pour chaque terme, un fichier (fichier_terme) est créé automatiquement contenant l'ensemble des lemmes. Le nom du fichier contient la taille du sac de lemmes ainsi que l'identifiant du terme TUV (exemple : 3_1223.txt).

Nous avons ensuite créé manuellement 12 transducteurs [21] génériques dépendants du nombre de lemmes (voir figure 3.14). Pour chaque transducteur, le remplissage pour un nouveau terme est automatique. La procédure a été enregistrée à l'aide d'un outil d'enregistrement de séquences [22], Action Recorder [23].

La construction des transducteurs se fait à l'aide du logiciel WinMacro [24] qui va

21. 12 est la taille maximale du nombre de lemmes pour nos termes
22. Enregistrement des actions de la souris et du clavier
23. Voir `http://www.maxxiweb.com/logiciel/utilitaire/divers/action-recorder/`
24. WinMacro simule des actions courantes de l'utilisateur telles que la saisie au clavier ou le positionnement des fenêtres. Il prend en charge des tâches plus complexes telles que la copie de fichiers. Plus d'une cinquantaine d'actions sont disponibles. L'intérêt du logiciel est qu'il permet de modifier le code source d'une séquence enregistrée (voir `http://www.01net.com/telecharger/windows/Utilitaire/planificateurs et lanceurs/fiches/1452.html`).

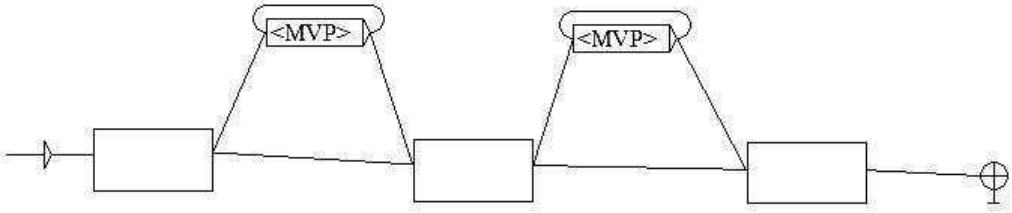

FIGURE 3.14 – Transducteur générique à 3 lemmes

pour chaque terme de notre ensemble :

- Ouvrir l'application NooJ qui permet de construire les transducteurs
- Ouvrir le transducteur générique correspondant au nombre de lemmes du terme dans l'application NooJ
- Ouvrir le fichier_terme du terme
- Exécuter la séquence enregistrée pour ce transducteur générique (grâce à ActionRecorder). La séquence consiste à :
  - copier/coller le contenu du ficher terme dans le transducteur
  - enregistrer le transducteur en NbLemme_CodeTerme.nog
  - fermer le logiciel NooJ

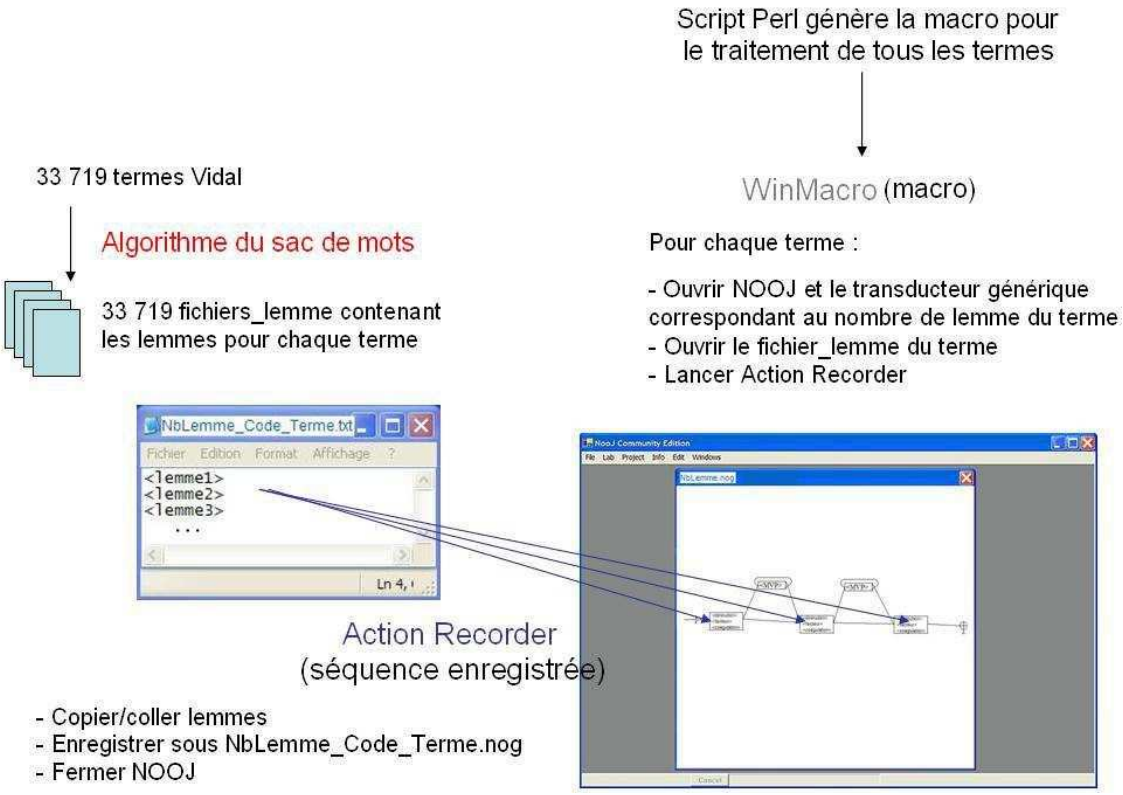

FIGURE 3.15 – Constitution automatique des transducteurs

3.8.2.5 Corpus utilisés

Les transducteurs ont été appliqués grâce au logiciel NooJ sur un ensemble de documents afin de récupérer de nouvelles variantes potentielles. Le corpus devait être composé de documents médicaux et être assez volumineux pour pouvoir extraire de nombreuses variantes. Les documents devaient être aussi de qualité pour ne pas récupérer de mauvaises variantes (avec des fautes d'orthographes ou des formes inconnues du jargon médical) entrées par les auteurs.

Les variantes recherchées étant en majorité des données thérapeutiques, nous avons inclus dans le corpus l'ensemble des RCP disponibles chez Vidal (14 104). Nous avons ajouté à cela des documents médicaux : 100 comptes rendus d'hospitalisation et l'ensemble du corpus CISMeF (plus de 40 000 ressources).

Ce corpus a été créé grâce au logiciel NooJ, ce logiciel peut prendre en compte plusieurs formats de fichiers dont le texte, le format XML et PDF qui constituent nos documents.

3.8.2.6 Résultats pour le TUV

L'application des transducteurs a permis de générer 3 633 092 variantes. Parmi ces variantes, 3 243 325 respectent l'ordre des lemmes du terme de départ et 336 918 sont constituées des lemmes dans le désordre. Un filtre a permis d'éliminer les variantes avec des mots en double [25] (soit 52 849 variantes éliminées).

Enfin, les variantes déjà existantes dans les terminologies du Vidal ou en double sont éliminées.

Au final, 7 800 variantes ont été recueillies grâce à cette méthode dont 1 007 concernaient le TUV.

Une validation manuelle des 1 007 variantes par un expert [26] a permis de valider 550 variantes (soit 55%).

Voici quelques exemples de variantes validées :
grossesses normales,grossesse normale,5250+CC+TUV
diabète non-insulino-dépendant,diabète non insulino-dépendant,2600+CC+TUV
pneumocoque et les infections,infection à pneumocoques,5419+CC+indic+PHR+TUV
antécédents de fracture,fracture_ antécédent,2543+CC+TUV
yeux infectés,infections des yeux,530+CC+TUV
antécédents récents d'infarctus du myocarde,infarctus du myocarde_ antécédent récent (d'),3589+CC+TUV

Voici quelques exemples de variantes rejetées :
âge du sujet,sujet âgé,6253+CC+TUV
augmentation de la charge,augmentation du poids,624+CE+ETAT+PATHO+TUV

25. En effet, la faiblesse de nos transducteurs est qu'ils permettent de générer des variantes avec des lemmes représentés plusieurs fois. Exemple pour le transducteur 3.13 si le corpus contient cette variante «diminution diminution de la coagulation», la variante est retrouvée.
26. M. Korshia, pharmacienne et gestionnaire du thésaurus chez Vidal.

maladie à cette période,maladie périodique,3543+CE+ETAT+MALADIE+TUV
hémorragique d'un accident,accident hémorragique,28+CE+scientifique+TUV

3.8.2.7 Création de nouvelles variantes

Une facon simple d'obtenir des variantes supplémentaires est de générer automatiquement les variantes flexionnelles (pluriels et singuliers) pour chaque variante déjà répertoriée. Nous avons produit ces variantes pour les termes de deux mots et moins leur construction étant simples.

Un script Perl permet de générer ces variantes (voir algorithme figure 3.16) :

Nous avons ainsi généré 4 279 variantes non répertoriées dans notre dictionnaire

```
Pour chaque variante de moins de trois mots du dictionnaire de termes faire
  découper_en_mots(variante)
  Pour chaque mot faire
    Si (mot ≠ mot_invariant[2]) Alors
      Si (mot fini par un "s") Alors
        enlever_en_fin_de_mot("s", mot)
      Sinon
        Si (mot fini par un "eaux") Alors
          remplacer_en_fin_de_mot("eaux", "eau", mot)
        Sinon
          Si (mot fini par un "aux") Alors
            remplacer_en_fin_de_mot("aux", "al", mot)
          Sinon
            Si (mot fini par un "al") Alors
              remplacer_en_fin_de_mot("al", "aux", mot)
            Sinon
              Si (mot fini par un "ail") Alors
                remplacer_en_fin_de_mot("ail", "aux", mot)
              Sinon
                Si (mot fini par un "aux") Alors
                  remplacer_en_fin_de_mot("aux", "al", mot)
                FinSi
              FinSi
            FinSi
          FinSi
        FinSi
      FinSi
    Sinon
      enlever_en_fin_de_mot("s", mot);
    FinSi
variante = ajouter (mot, variante)
Si variante ∈ dico_termes Alors
  Afficher(variante)
FinSi
```

[2] "gros", "abcès", "infarctus", "abces", "accès", "acces", "anticorps", "anus", "colapsus", "chez", "comas", "lupus", "psoriasis", "virus", "utérus", "gris", "à", "a", "b", "foetus", "collapsus", "décès", "herpès", "b12", "b6", "es", "ds", "ks", "pps", "poids"

FIGURE 3.16 – Algorithme de génération de variantes flexionnelles

de terme.

Ces variantes potentielles ont été validées par moi-même et notre expert M. Korshia. Sur 4 279 seulement 328 variantes ont été éliminées (soit 7,7% - exemple «astérixis» pour «asterixi»).

Le dictionnaire final TUV contient 40 266 variantes (pour 11 980 termes).

3.8.2.8 Indexation par le dictionnaire de termes

L'indexation d'un document par le dictionnaire de termes consiste à appliquer, grâce au logiciel NooJ (voir section 3.8.2.3), le dictionnaire de termes au corpus à indexer (en une seule fois).

Le fichier obtenu contient pour chaque variante retrouvée dans le corpus :

- le nom du fichier à partir duquel elle a été extraite
- sa position dans le document (les positions des caractères de début et de fin)
- son entrée dans le dictionnaire (*Variante,LibelléDuTerme,CodeTUV+Informations Divers*)

Puis le transducteur pour le découpage en phrases est appliqué. De la même facon, nous obtenons un fichier avec les phrases identifiées pour chaque document et leurs positions.

Ces deux fichiers permettent de générer une proposition d'indexation pour chaque document avec pour chaque phrase les libellés et codes des termes TUV associés.

3.8.2.9 Points forts et points faibles de la méthode du dictionnaire de termes

La méthode du dictionnaire de termes est plus rapide et plus fiable que la précédente.

En effet, l'application d'un dictionnaire dans NooJ est indépendante de la taille du dictionnaire. Le temps d'application est donc quasi instantané pour un document.

Ce temps varie selon le nombre de documents à indexer. Pour un corpus de 10 000 documents (de 5 pages chacun), quelques petites minutes suffisent. L'application du dictionnaire étant exécutée en une seule fois.

Les variantes sont validées en amont, ce qui lors de l'indexation permet de générer un minimum d'erreurs ce qui n'est pas le cas pour la méthode du sac de mots ou celle de la méthode des constituants (voir section suivante).

Malheureusement, la qualité de l'indexation dépend du nombre de variantes répertoriées dans le dictionnaire. Le nombre de variantes pour chaque terminologie est encore insuffisant pour couvrir l'ensemble des variantes existantes. En outre, notre méthode d'obtension de variantes a demandé plusieurs mois d'exécution.

3.8.3 Méthode du dictionnaire de constituants

3.8.3.1 Principe de la méthode

La méthode, explicitée ci-dessus, peut être implémentée différemment en prenant en compte les éléments constitutifs du terme et non les mots seuls ou le terme dans sa globalité.

Le principe est d'indexer un terme pour une phrase si celle-ci contient tous les constituants associés à ce terme.

Un constituant est défini comme toute variante incluse dans un terme. Exemple, le terme «angine de poitrine sévère» comprend plus de 6 constituants : «angine», «angines», «poitrine», «sévère», «aigu» «angor» etc. ...

Les entrées du dictionnaire de constituants pour ce terme sont : *angine,angine de poitrine sévère,411+CC+pe+PHR++TUV+1*

angines,angine de poitrine sévère,411+CC+pe+PHR++TUV+1
poitrine,angine de poitrine sévère,411+CC+pe+PHR++TUV+1
poitrines,angine de poitrine sévère,411+CC+pe+PHR++TUV+1
sévère,angine de poitrine sévère,411+CC+pe+PHR++TUV+1
sévères,angine de poitrine sévère,411+CC+pe+PHR++TUV+1
aigu,angine de poitrine sévère,411+CC+pe+PHR++TUV+1
aiguë,angine de poitrine sévère,411+CC+pe+PHR++TUV+1
angor,angine de poitrine sévère,411+CC+pe+PHR++TUV+2

Une première version du dictionnaire des constituants des termes a été réalisée. Celui-ci ne contient que les constituants de 1 mot et les constituants équivalents aux termes.

Afin de définir les constituants de 1 mot, nous avons répertorié pour nos cinq terminologies tous les lemmes associés et leurs variations, flexions et synonymes grâce à l'analyse des dictionnaires existants (voir section 3.8.2.3).

Dans l'avenir, une deuxième version contiendra les constituants de plus de 1 mot et de poids supérieur à 1. Ceux-ci peuvent être obtenus en cherchant les inclusions dans les lexiques dont nous disposons.

3.8.3.2 Indexation à l'aide du dictionnaire de constituants

L'indexation des phrases d'un ensemble de documents par le dictionnaire de constituants consiste à appliquer grâce au logiciel NooJ le dictionnaire de constituant au corpus à indexer (en une seule fois).

Il faut ensuite pouvoir déterminer pour chaque phrase si elle contient tous les constituants requis pour un ou des terme(s) des terminologies.

Afin de réaliser cela, une note est ajoutée à chaque constituant afin de définir sa couverture en matière de lemmes pour le terme associé. Ici le constituant «angor» a un poids de 2 puisqu'il couvre les lemmes «angine» et «poitrine». Les autres ont un poids de 1.

Dans notre base de données multi-terminologique est répertorié pour chaque terme son nombre de lemmes. Ainsi il est indiqué que le terme «angine de poitrine sévère» a un poids de 3. Pour indexer une phrase avec le terme «angine de poitrine sévère», il faut avoir une couvrance parfaite de l'ensemble des lemmes du terme, donc atteindre un poids de 3 pour ce terme.

Prenons un exemple :

Indexation de la phrase : « Le patient est atteint d'un syndrome sévère, le syndrome de Down accompagné d'asthme.»

Après application du dictionnaire de termes grâce à l'outil NooJ, il a été retrouvé les constituants suivants :

syndrome, syndrome de Down, TUV+PATHO+ms
syndrome, syndrome de Wolfram, TUV+PATHO+ms
syndrome, syndrome de Down, TUV+PATHO+ms
syndrome, syndrome de Wolfram, TUV+PATHO+ms

Down, syndrome de Down, TUV+ PATHO +np
asthme, asthme,TUV+PATHO+1+ms

Après élimination des doublons et ajout des poids pour chaque terme nous obtenons :

- un poids de 2 pour «syndrome de Down»
- un poids de 1 pour «syndrome de Wolfram»
- et un poids de 1 pour «asthme»

Seuls les termes ayant un poids égal au nombre de lemmes le constituant sont indexés pour la phrase. Donc seuls les termes «syndrome de Down» et «asthme» seront indexés pour cette phrase.

Pour que cette méthode fonctionne, il faut que tous les constituants pour un terme soient uniques et non inclus dans un autre constituant. Un autre filtre doit donc être appliqué avant le calcul du poids pour chaque terme. Ce filtre élimine tout constituant inscrit dans un autre constituant et dont le poids est plus faible que celui-ci [27].

3.8.3.3 Points forts et points faibles de la méthode du dictionnaire de constituants

Par rapport aux deux autres méthodes citées précédemment, la méthode du dictionnaire de constituants permet de prendre en compte un plus grand nombre de variantes potentielles.

De la même manière que pour la méthode du dictionnaire de termes, le temps d'indexation est rapide. En revanche, la taille du dictionnaire est limitée pour le logiciel NooJ. Afin de poursuivre nos travaux pour cette méthode et ajouter l'ensemble des constituants, il faudra changer de logiciel ou de méthode (ici les travaux de E. Prieur pourront être utilisés [Prieur07]).

3.9 Prise en compte des contextes

3.9.1 Prise en compte des négations

Il est important dans l'indexation d'un document non seulement de repérer tous les termes présents mais aussi d'identifier parmi eux ceux qui sont inclus dans une négation ou une exception. Ceci est important pour l'indexation de comptes rendus ou de RCP [28] puisqu'il est nécessaire de déterminer les éléments à écarter (maladies, effets secondaires). Par exemple, pour la phrase «Aucune suspicion d'accident vasculaire cérébral», les termes D020521 (MeSH), D3-89550 (SNOMED), 61 (TUV), I64 (CIM10) «accident vasculaire cérébral» doivent être indexés avec un type «négatif».

Plusieurs outils permettant d'identifier des négations sont cités dans la littérature. La plupart de ces systèmes se basent sur les expressions et les conjonctions marquant

27. Ce filtre utilise une table d'inclusion qui indique, pour chaque couple de constituants, le constituant à éliminer si les deux sont retrouvés pour la même phrase

28. Mais ceci n'a aucune utilité pour l'indexation de ressources Web puisque même si le sujet est traité de manière négative il est traité dans la ressource donc il doit être indexé.

la négation. Les travaux [Chapman01] et [Elkin05] listent ces expressions pour l'anglais (exemple : «absence of» ou «except»). D'autres permettent de les apprendre grâce à des méthodes d'apprentissage automatique [Averbuch04]. Pour le francais, A. Baneyx a développé une méthode simple, un transducteur permettant de détecter les formes négatives pour les maladies, symptômes et signes [Baneyx06].

Comme nous avons pu le voir, la négation et les exceptions ont d'abord été gérées dans les termes d'indexation, eux-mêmes, grâce à l'élaboration de libellés d'indexation (voir section 3.5).

Pour l'identification de négations dans la phrase, nous nous sommes fondée sur les méthodes de TAL citées ci-dessus. Voici comment sont repérées les négations pour nos trois méthodes :

- Méthode du sac de mots : les expressions marquant la négation sont le plus souvent éliminées du sac de mots car elles font parties des mots vides (exemple : «pas» et «sans»). L'une des solutions est, lors du découpage en mots de la phrase, de repérer les expressions négatives (exemple : «pas de»). Le mot qui suit ce genre d'expressions négatives peut être éliminé du sac de mots. La méthode étant peu efficace nous ne l'avons pas implémentée.
- Méthode du dictionnaire de termes : pour cette méthode des transducteurs permettent de détecter les termes impliqués dans une négation ou une exception (voir figure 3.17, 3.18, 3.19).

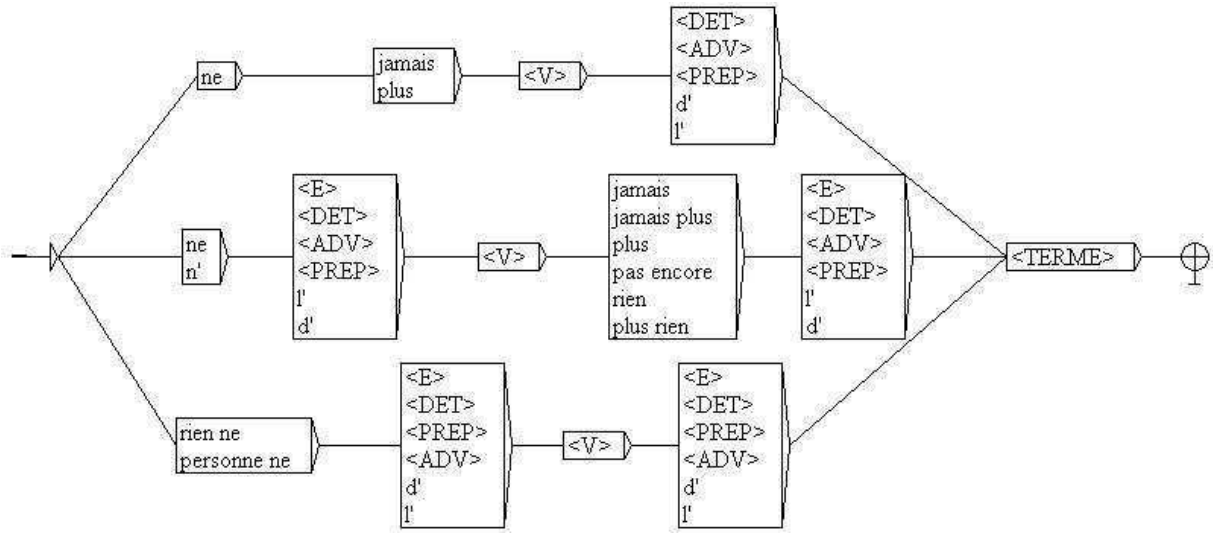

FIGURE 3.17 – Transducteur permettant d'identifier les termes associés à un verbe négatif

- méthode du dictionnaire de constituants : des transducteurs équivalents à la méthode précédente peuvent être utilisés afin de détecter les constituants à ne pas prendre en compte (<TERME> est remplacé par <CONSTITUANT>).

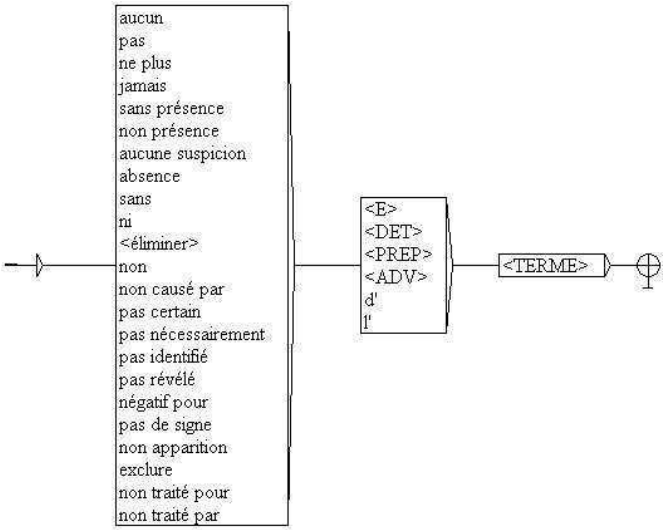

FIGURE 3.18 – Transducteur permettant d'identifier les termes associés à des expressions négatives antérieures

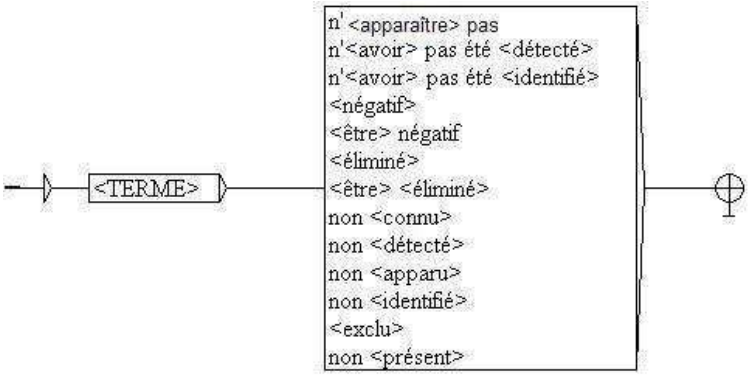

FIGURE 3.19 – Transducteur permettant d'identifier les termes associés à des expressions négatives postérieures

3.9.2 Prise en compte des rubriques

La rubrique dans laquelle se trouve la phrase à indexer est un élément du contexte à prendre en compte.

En effet, les éléments du contexte permettent de préciser certaines notions qui peuvent dès lors être rattachées à un terme d'une terminologie. Par exemple, pour la rubrique «antécédents», l'intégralité des phrases de cette rubrique va porter sur les antécédents du patient. Il est difficile de repérer dans les phrases que les maladies concernées sont des antécédents, soit parce que ces phrases ne sont qu'une énumération de maladies soit parce que le seul élément qui montre que cela est un antécédent est la conjugaison du verbe au passé.

Nous présentons la facon dont cela a été implémenté dans les différentes méthodes :

– Méthode du sac de mots : pour chaque phrase de la rubrique, il est ajouté au sac de mots correspondant le lemme ou stème du mot «antécédent» (voir figure

3.12 pour un exemple)
- Méthode du dictionnaire de constituants : le constituant «antécédent» est ajouté au traitement de chaque phrase appartenant à la rubrique.
- Méthode du dictionnaire de termes : pour cette méthode, une méthode à base de règles peut être envisagée.
  Exemple : Si (rubrique=«antécédents» et «tumeur maligne» appartient à termes_indexés) Alors indexer «Antécédent de tumeur maligne». Cette méthode nécessite de définir toutes les règles et de les valider par un expert. Cette méthode pourra être envisagée dans l'avenir.

3.10 Fusion des indexations produites par les trois méthodes

Les trois méthodes (algorithme du sac de mots, dictionnaire de termes et dictionnaire de constituants) ont été créées afin d'être complémentaires.

Tous les termes indexés par les trois méthodes sont donc agrégés afin d'avoir une indexation la plus complète possible.

Les termes pourraient être pondérés selon la méthode d'obtention. La méthode du dictionnaire de termes extrayant des variantes validées, les termes obtenus grâce à cette méthode pourraient être assignés d'un poids supplémentaire (2 au lieu de 1 pour les autres méthodes).

Pour l'instant, notre outil ne propose qu'une méthode simple d'agrégation mais dans l'avenir la méthode pourra être étendue. Par exemple, la proposition d'indexation de la méthode du dictionnaire de termes pourra permettre d'éliminer des termes proposés par les autres méthodes.

3.11 Restriction à une ou plusieurs terminologies

Les termes sont ensuite restreints aux termes équivalents appartenant aux terminologies d'indexation choisies par l'utilisateur.

Afin de récupérer les termes proches, nous utilisons les différents transcodages existant entre nos cinq terminologies qui fournissent des liens de synonymie et d'équivalence :
- Les transcodages entre les terminologies MeSH, CIM10 et SNOMED sont extraits du Metathesaurus de l'UMLS (version 2007ac). Ces transcodages sont bidirectionnels. Un autre transcodage, cette fois unidirectionnel[29], entre la SNOMED et la CIM10 (SNOMED->CIM10) créé par le SFINM a aussi été utilisé.
- Le transcodage unidirectionnel CCAM->MeSH créé dans l'équipe CISMeF par P.Massari (voir section 5.8.1)
- Le transcodage CIM10-CCAM de TOTHEM [Chevallier03]
- Le transcodage unidirectionnel TUV->MeSH créé par CISMeF et validé par la société Vidal

29. Terme A->les termes C+D+E d'une autre terminologie. Mais C->D n'est pas valide.

– et le transcodage unidirectionnel TUV->CIM10 créé par Vidal

Après quelques expérimentations, il s'est avéré que de nombreux transcodages n'étaient pas adaptés. Le sens n'est parfois pas respecté après transcodage. C'est le cas des transcodages TUV->CIM10 et CCAM-CIM10 qui ont été réalisés pour des tâches précises au sein des organismes. Ces tâches ne correspondaient pas à un besoin d'équivalence en sens. Ces transcodages n'ont donc pas été implémentés dans F-MTI.

Les autres transcodages sont implémentés dans la table «Les_relations_inter_terminologiques» de notre base de données multi-terminologique.

La méthode est appliquée après fusion des termes obtenus par les différentes méthodes d'indexation. Elle n'utilise que les transcodages impliqués par les terminologies d'indexation choisies. Par exemple, si l'utilisateur choisit d'indexer son document à l'aide de la terminologie CIM10, seuls les transcodages MeSH->CIM10 et SNOMED->CIM10 seront appliqués. Seuls les termes CIM10 seront proposés à l'utilisateur en fin de parcours.

L'utilisation des transcodages permet de compléter une indexation existante. Pour l'indexation de la phrase «Ulcère gastrique opéré il y a 20 ans.» (voir figure 3.12), le transcodage permet de compléter l'indexation par les termes K25.9, D5-32422, D013270 et C16.9 (voir figure 3.20).

ulcère gastrique	EII2048	TUV	;gastrique;ulcère;
ulcère gastrique	PE1192	TUV	;gastrique;ulcère;
ulcère gastrique	D013276	MSH	;gastrique;ulcère;
ulcère gastrique	D5-32100	SNMI	;gastrique;ulcère;
ulcère de l'estomac	non précisé comme étant aigu ou chronique, sans hémorragie ni perforation	K25.9	CIM10	;aigu;chronique;estomac;ulcère;
ulcère gastrique sans hémorragie ni perforation ou obstruction	D5-32422	SNMI	;gastrique;ulcère;
estomac	D013270	MSH	;estomac;
tumeur maligne estomac, sans précision	C16.9	CIM10	;estomac;malin;tumeur;

FIGURE 3.20 – Complément d'indexation apporté par le transcodage

3.12 Post-traitement

Le post-traitement consiste à générer l'indexation finale pour toutes les phrases d'un document ainsi que l'indexation finale pour le document.

Il comprend plusieurs étapes :

– élimination des doublons (même termes ou un terme et son synonyme de la même terminologie)
– application des règles d'indexation :
  – les règles générales :
    – Nous privilégions une indexation au plus précis. Les termes les plus précis sont donc privilégiés par rapport aux termes moins précis qui sont éliminés. Ainsi si, dans notre proposition d'indexation, un terme et son fils sont

retrouvés alors le terme père est éliminé. De même, les sacs de mots sont analysés pour chaque terme indexé. Les termes ayant un sac de mots inclus dans un autre sont éliminés.

- les règles spécifiques à chaque terminologie :
  - Pour la CIM10 : notre indexation CIM10 est purement descriptive et non médico-économique, elle n'intègre donc pas les règles de codage PMSI.
  - Pour le MeSH : si un terme et un qualificatif qui lui est affiliable sont indexés alors ils sont appariés. Tous les qualificatifs n'étant pas appariés sont éliminés. De plus certains termes ne sont pas utilisés pour l'indexation, parce qu'ils sont susceptibles d'être indexés pour la plupart des ressources alors que leur indexation n'est pertinente que dans de rares cas (exemple : «conseil», «maladie», «médecine», «informatique»). Il en existe 18, une nouvelle liste est en cours d'élaboration. Ces termes sont éliminés de la proposition d'indexation.
  - Pour les autres terminologies : pour la SNOMED et le TUV ces terminologies n'ayant encore jamais été indexées en routine aucune règle d'indexation n'existe. Pour la CCAM, seules des règles de codage médico-économique existent, aucune pour un codage descriptif.

L'indexation finale pour un document consiste à appliquer le post-traitement à l'union de toutes les indexations finales pour toutes les phrases qu'il contient.

3.13 Paramètres et éléments en sortie

3.13.1 Paramètres

F-MTI peut indexer un document à l'aide d'une ou plusieurs des trois méthodes selon le choix de l'utilisateur.

Il inclut un paramétrage spécifique pour les RCP, les comptes rendus d'hospitalisation et les sites web. S'il est indiqué que les documents en entrée sont des RCP, l'indexation produite par F-MTI tiendra compte des rubriques usuelles du RCP et utilisera le TUV. Pour les sites Web l'indexation s'effectuera en MeSH et pour les comptes rendus d'hospitalisation les rubriques sont prises en compte et elle s'effectuera en CIM10, CCAM et SNOMED.

L'utilisateur peut choisir d'effectuer une indexation mono ou multi-terminologique. Dans les deux cas, les terminologies prises en compte peuvent être paramétrées.

3.13.2 Sortie

En sortie, F-MTI génère deux propositions d'indexation, l'une pour chaque document l'autre pour chaque phrase. Pour chaque document sont présentés : rubrique, paragraphe, phrase et l'indexation proposée. L'indexation fournit les termes indexés avec leur source, leur type, leur découpage en lemmes ou stèmes et leur taille :

NomRubrique|N°paragraphe|Phrase|N°phrase|Type_terme|Codes_terme|Langue| Terminologie|Taille|Libellé_Terme|Type|Propriétés|Codable|Découpage_lemmes_ou_stèmes
ANTECEDENTS|3|asthme|2|1|G-0003|FRE|SNMI|1|antécédents
de|G|NULL|O| ;antécédent ;
ANTECEDENTS|3|asthme|2|1|D001249|FRE|MSH|1|asthme|D|C08.127.108 ;C08.381.495.
108 |O| ;asthme ;
ANTECEDENTS|3|asthme|2|1|J45.9|FRE|CIM10|1|asthme,sans précision|S|NULL|
O| ;asthme ;

3.14 Conclusion

Ce chapitre a permis d'exposer le fonctionnement de l'outil F-MTI. Plusieurs méthodes ont été implémentées afin de réaliser une indexation multi-document, multi-terminologique et multi-tâches [30].

Dans le chapitre suivant, nous évaluons l'indexation produite par l'outil F-MTI pour les différentes tâches d'indexation décrites au départ.

30. Par multi-tâche, nous entendons la capacité de F-MTI à indexer un même document avec une même terminologie mais pour des tâches différentes. Par exemple, F-MTI pourrait réaliser une indexation CIM10 d'un compte rendu à visée médico-économique ou bien à visée descriptive. Ceci est possible grâce aux règles d'indexation du post-traitement qui seront différentes selon la tâche visée.

Chapitre 4

Évaluation de l'indexeur multi-terminologique

4.1 Introduction

Nous avons procédé à différentes évaluations, la première consiste à évaluer différentes méthodes de désuffixation afin de déterminer la meilleure méthode à intégrer dans F-MTI.

Les cinq évaluations suivantes portent sur les performances du F-MTI «en situation». Ainsi l'indexation produite à l'aide de la CIM10, de la CCAM et de la SNOMED pour les comptes rendus d'hospitalisation est évaluée. Nous évaluons aussi l'indexation des ressources Web à l'aide du MeSH et des RCP à l'aide du TUV.

Notre outil a finalement été comparé à un autre outil d'indexation automatique en SNOMED 3.5 : SnoCode.

4.2 Évaluations réalisées

4.2.1 Évaluation de différentes méthodes de désuffixation

4.2.1.1 Principe

La méthode de l'algorithme du sac de mots implémentée dans F-MTI nécessite un algorithme de désuffixation. À l'origine, cette méthode utilise un algorithme de désuffixation produit par l'équipe CISMeF pour la traduction des requêtes en termes MeSH dans le moteur de recherche Doc'CISMeF. Cependant, cet algorithme est connu pour être très simple et restreint aux suffixes les plus courants. Il existe, par ailleurs, plusieurs outils libres d'utilisation mais très peu ont été évalués et aucun n'a été testé à ce jour dans le domaine médical.

Les termes médicaux sont très particuliers. Plus que dans d'autres domaines, il se trouve de nombreux mots de composition savante formés à partir de radicaux, de préfixes ou de suffixes (exemple : «hépatite» composé à partir de «hépa» (pour foie) et du suffixe «ite» (pour inflammation)). Ainsi, certaines racines d'usage strictement médical ne se retrouvent que dans les mots du domaine (exemple : «ecto-

mie»). En outre, les mots peuvent être empruntés au grec, au latin (exemple : «*in vitro* »), à l'anglais (exemple : «overdose» pour surdosage) ou à l'allemand avec une prédominance pour le grec. On trouve aussi des expressions comportant des noms propres avec notamment les maladies éponymiques (exemple : «maladie d'Alzheimer»). Le vocabulaire médical fait aussi état de nombreux néologismes [1] pour identifier les nouveaux concepts issus de nouvelles découvertes. Les termes peuvent aussi contenir de nombreux sigles (exemple : «ph»), symboles (exemple : «Na» pour sodium), unités (exemple :«g» pour gramme), multiples ou fractions d'unités (exemple : «kilo» pour multiplier par 1000), des symboles mathématiques, des lettres grecques. Enfin, il existe aussi de nombreux mots composés avec trait d'union.

Tout ceci peut complexifier la désuffixation, et certains algorithmes peuvent être mieux adaptés que d'autres.

Nous avons donc comparé trois méthodes de désuffixation :

- l'algorithme CISMeF : l'algorithme traite à tour de rôle des suffixes d'une liste (63 suffixes, voir la liste des traitements figure 4.1). Le traitement consiste à éliminer ou remplacer les suffixes rencontrés dans certaines conditions. Les conditions portent sur la taille du mot, le suffixe ou le mot. L'ordre de traitement des suffixes implique de traiter les suffixes les plus long en premier.
  Par exemple , après application de la règle 1, le mot «angines» devient «angine» qui devient «angin» (le stème) après application de la règle 4. Cet algorithme a été choisi dans notre évaluation car nous l'avions à notre disposition. De plus, la comparaison à d'autres algorithmes, nous permettrait éventuellement d'améliorer le moteur de recherche Doc'CISMeF.

TB suffixes.txt

```
ORDRE|SUFFIXE|TAILLE_MIN_MOT|SUFFIXE_NORMALISE|EXEMPLES|SAUF_SUFFIXES|SAUF_MOTS
1|s|5||||
2|er|5||||
3|r|5||||
4|e|5|||canne|
5|aux|5||hormonaux|eaux|
8|eaux|5|eau|vaisseaux||
6|eux|6||osseux||
7|al|5||grippal||
9|issement|8||vieillissement||
10|issant|8||vieillissant||
11|iv|5||digestive||
12|if|5||digestif||
```

FIGURE 4.1 – Quelques règles de désuffixation pour l'algorithme CISMeF

- l'algorithme de Carry [Paternostre02] : il constitue une adaptation francaise de l'algorithme de Porter qui traite les mots de la langue anglaise [Porter80]. Cet algorithme a été réalisé par M. Paternostre dans le cadre du projet de recherche GALILEI [2] en 2002.
  Cet algorithme se déroule en diverses étapes par lesquelles les suffixes sont

1. Fabrication de nouveaux mots ou utilisation de mots habituels avec une signification nouvelle
2. Generic Analyser and Listener for Indexed and Linguistics Entities of Information, l'algorithme est téléchargeable gratuitement sur le site du projet `http://www.galilei.ulb.ac.be`

traités à tour de rôle, en utilisant des règles et des conditions comme l'algorithme précédent (482 règles, voir liste figure 4.2). De la même facon ici, l'ordre des étapes est établi pour que ce soit le suffixe le plus long qui détermine la règle à appliquer.
Les différences principales, outre le nombre de règles appliquées, sont les condictions prise en compte. Pour les auteurs, chaque mot du francais peut être réduit à cette formule : [C] $(VC)^m$ [V] où (VC) est répété un nombre «m» de fois (C = consonne, V = voyelle, les crochets marquent des événements optionnels). Les conditions portent sur la valeur de «m» [Porter80].

TBsuffixescarry.txt

m	SUFFIXE	SUFFIXE_NORMALISE
0	issaient	
0	ellement	el
0	issement	
0	alement	al
0	eraient	
0	iraient	
0	eassent	
0	ussent	
0	amment	
0	emment	
0	issant	

FIGURE 4.2 – Quelques règles de désuffixation pour l'algorithme de Carry

– et le Frenchstemmer de Lucene [3] [Cutting04] : réalisé par Patrick Talbot, celui-ci s'inspire aussi des travaux de Porter.
Cet algorithme se déroule en 6 étapes : élimination des suffixes standard, traitement des suffixes verbaux, traitement des suffixes résiduels, traitement des formes particulières, traitement des caractères doubles et des accents. Pour chaque étape, une liste de règles est appliquée dépendant d'une ou plusieurs conditions. Ici aussi les conditions sont particulières. Les auteurs prennent en compte 3 régions pour un mot : RV, R1 et R2. RV est le mot. R1 est la région après la première non-voyelle suivie d'une voyelle ou la fin du mot. R2 est l'équivalent de R1 pour R1. Par exemple, pour le mot «fameusement» RV = «fameusement», R1 = «eusement» et R2 = «ement». Les conditions portent sur ces régions, sur leurs présences ou les caractères les précédant ou les suivant (voir exemple figure 4.3).
De la même facon ici, l'ordre des étapes est établi pour que ce soit le suffixe le plus long qui détermine la règle à appliquer.
Celui-ci a été choisi car il est utilisé dans des travaux en cours chez Vidal, il a donc paru intéressant de le comparer aux autres pour mesurer l'impact des différents algorithmes.

3. Lucene est un moteur de recherche libre écrit en Java qui permet d'indexer et de rechercher du texte. C'est un projet open source de la fondation Apache mis à disposition sous licence Apache. Il est également disponible pour les langages Ruby, Perl, C++, PHP. Pour plus d'informations voir : `http://lucene.apache.org/`

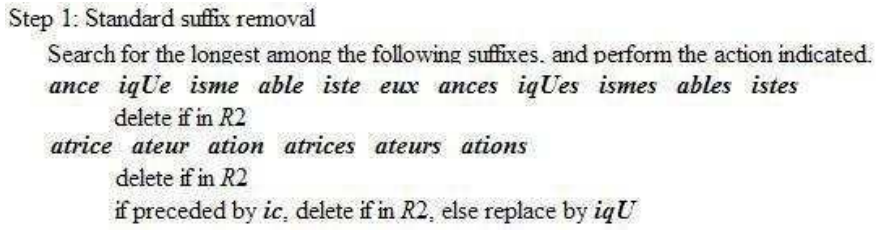
Step 1: Standard suffix removal
Search for the longest among the following suffixes, and perform the action indicated.
ance iqUe isme able iste eux ances iqUes ismes ables istes
delete if in *R2*
atrice ateur ation atrices ateurs ations
delete if in *R2*
if preceded by *ic*, delete if in *R2*, else replace by *iqU*

FIGURE 4.3 – Quelques règles de désuffixation pour le FrenchStemmer de Lucene

Le but ici est de déterminer la meilleure méthode de désuffixation pour le domaine médical. La meilleure méthode de désuffixation est celle qui produit le meilleur radical, capable d'être associé à tous les mots d'une même famille. Par exemple, «asthme» «asthmes» «asthmatique» et «asthmatiques» sont quatre mots composant une même famille. Ils doivent donc tous être associé au même radical, un radical possible étant «asthm».

Pour l'indexation automatique, la désuffixation doit pouvoir apparier les mots courants ou médicaux d'une phrase (provenant d'un RCP ou d'un compte rendu médical ou d'un site Web médical) avec des mots pour la plupart médicaux appartenant à des terminologies médicales. Nous avons essayé de recréer ce phénomène dans notre évaluation.

4.2.1.2 Éléments d'évaluation

Nous avons pris comme éléments d'évaluation la liste de l'ensemble des mots signifiants (sans les mots vides) composant le TUV. Le choix du TUV a été orienté car c'est une des terminologies implémentées dans F-MTI et qui semble contenir plus de mots de types différents (unités, mots anglais, latin, grec etc...) que les autres terminologies. De plus, cette évaluation intéressait le Vidal pour de futurs produits.

Tous les mots signifiants du TUV ont d'abord été extraits puis désuffixés à l'aide des trois algorithmes de désuffixation. Nous avons ainsi identifié 5 463 mots médicaux et généraux sur 84 968 dont les radicaux étaient différents pour au moins une des trois méthodes. Ensuite, pour ces mots, nous avons mesuré la pertinence de chaque stème par rapport à une référence.

Cette référence a été constituée à partir de plusieurs sources médicales et générales (dictionnaire repris d'une précédente étude voir section 3.7.3). Tous ces dictionnaires ont permis de constituer 8 404 familles de mots (soit 707 108 mots en tout). Une famille de mots est constituée par tous les mots partageant le même thème morphologique et un sens commun présents dans les dictionnaires (exemple : «asthme», «asthmes», «asthmatique» et «asthmatiques» font partie de la même famille).

Enfin, la pertinence de chaque radical pour chaque mot du TUV, est calculée en comparant les familles de mots créées par ce radical par rapport aux familles de référence. Pour définir la famille de mots pour chaque radical, nous avons réalisé la liste de l'ensemble des mots répertoriés dans la référence qui ont été désuffixés grâce aux trois algorithmes. Pour chaque algorithme, les mots ayant le même radical seront rassemblés dans la même famille.

L'évaluation a consisté à mesurer la précision et le rappel en comparant les familles de mots créées pour chaque algorithme par rapport aux familles de référence (voir figure 4.4).

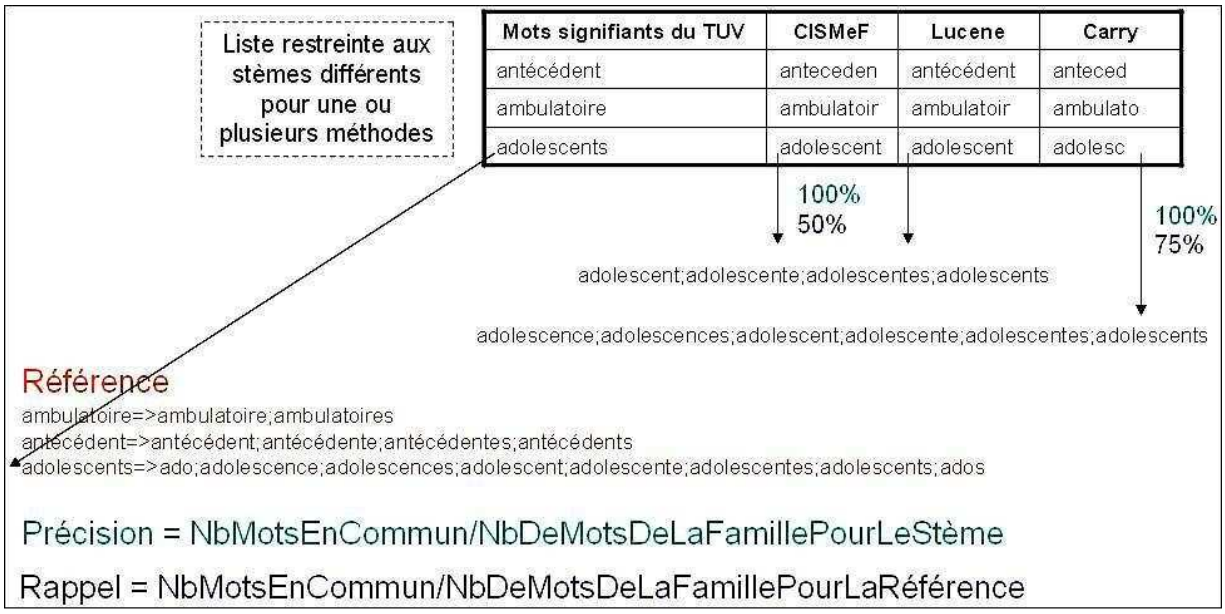

FIGURE 4.4 – Protocole d'évalution des trois méthodes de désuffixation

4.2.1.3 Résultats de la comparaison des trois algorithmes de désuffixation

Algorithmes	Précision	Rappel	F-measure
Algorithme CISMeF	70,9%	69,4%	70,4%
Algorithme de Carry	59,3%	76,3%	66,7%
Frenchstemmer de Lucene	81,4%	74,7%	77,9%

FIGURE 4.5 – Résultats de l'évaluation des trois algorithmes pour les mots du TUV par rapport au dictionnaire de référence

Les résultats présentés à la figure 4.5 montrent que l'algorithme de Carry produit le meilleur rappel avec 76,3% (v.s 74,7% pour l'algorithme de Lucene et 69,4% pour l'algorithme de CISMeF). En revanche, c'est l'algorithme de Lucene qui produit la meilleure précision avec 81,4% (v.s 70,9% pour l'algorithme de CISMeF et 59,3% pour l'algorithme de Carry) et la meilleure F-mesure [4] avec 77,9% (v.s 66,7% pour l'algorithme de Carry et 70,4% pour l'algorithme de CISMeF).

4. Moyenne pondérée de la précision et du rappel.

4.2.1.4 Discussion

Bien que le principal avantage de ces outils réside dans leur simplicité, l'absence de contraintes linguistiques fortes engendre néanmoins des erreurs de surracinisation (exemple : le stème «nat» apparie à la fois «nature» et «nation») ou de sous-racinisation (exemple : le stème «adaptat» empêche l'appariement des formes «adapter» et «adaptation»). Cette remarque est confirmée par nos résultats puisque l'algorithme Lucene, qui prend en compte le plus de contraintes sur la forme du mot, obtient les meilleurs résultats.

Cette évaluation s'est faite dans le cadre de la terminologie TUV, nous étendons cette hypothèse aux autres terminologies puisque 80% des lemmes sont communs entre le TUV et les quatre autres terminologies.

Au niveau du temps d'exécution, pour la désuffixation de 30 000 mots, l'algorithme de Lucene met 12 min, celui de Carry, 15 min 30 et l'algorithme de CISMeF, 11 min 30 (bien sûr cela est largement dépendant du langage utilisé pour implémenter les trois méthodes, ici le Perl). Le FrenchStemmer de Lucene propose donc en matière de temps d'exécution des résultats tout à fait corrects.

Nous choisissons donc l'algorithme Lucene comme algorithme de désuffixation pour F-MTI. Cet algorithme sera aussi intégré dans Doc'CISMeF.

Il faut ajouter que l'impact de la désuffixation sur les performances des systèmes de recherche d'information est cependant discuté [Moreau].

En outre, il existe d'autres méthodes d'évaluation de ce genre d'algorithme telles que le nombre moyen de mots, le niveau de compression obtenu, le nombre moyen de caractères supprimés ou la distance de Hamming [Paice96]. Mais celles-ci ne mesurent pas l'algorithme en situation d'indexation. La méthode que nous avons développée nous semble donc mieux adaptée à notre problématique.

4.2.2 Évaluation de l'extraction de termes CIM10 et CCAM pour les dossiers patients

4.2.2.1 Méthode d'évaluation

Nous avons souhaité évaluer l'outil F-MTI dans le cadre du codage des comptes rendus d'hospitalisation francais en CIM10 [Pereira08b] et en CCAM. Nous avons appliqué la méthode de l'algorithme du sac de mots seule méthode alors implémentée lors de cette évaluation.

4.2.2.2 Corpus d'évaluation

Nous avons extrait au départ 1000 comptes rendus. Parmis ceux-ci 206 comptes rendus se sont révélés être des courriers ou des comptes rendus d'hospitalisation ne respectant pas les rubriques identifées ou le codage en CIM10 et CCAM n'a pu être raccroché à ceux-ci. F-MTI a donc été évalué sur un corpus de 794 comptes rendus d'hospitalisation, 490 provenant de séjours en Cardiologie et 304 provenant de séjours en Pneumologie effectués au CHU de Rouen. Nous avons choisi ces secteurs car ils font partie du domaine d'expertise de notre expert en codage (Dr P. Massari). Ces

dossiers concernent 794 patients différents, ayant effectué un séjour en 2007. Ils ont été extraits du logiciel de gestion de dossier patient électronique du CHU de Rouen nommé CDP2 [Massari00] (1 080 384 patients et 182 808 comptes rendus d'hospitalisation en 2005).

Un compte rendu d'hospitalisation détaille les antécédents du patient, les examens qu'il a subi, les actes réalisés, les résultats et la prescription de médicaments. Ces résumés sont tapés à la sortie du patient de l'unité de soin par les médecins en charge du patient ou les secrétaires dans le secteur où ont été effectués les soins. Puis ces comptes rendus sont codés en CIM10 et en CCAM dans une période plus ou moins courte après la sortie. Ce codage, répondant à un objectif budgétaire, est réalisé en conformité avec les règles médico-économiques en vigueur (voir section 2.4.3.1). Nous avons récupéré ces codages.

F-MTI ne produit qu'une indexation purement descriptive du document. Nous avons donc en plus demandé à un médecin expert du codage d'indexer manuellement de manière descriptive 100 comptes rendus d'hospitalisation tirées au hasard parmi les 794 (50 provenant de séjours effectués en Cardiologie et 50 de séjours en Pneumologie). Cet expert était en aveugle quant à l'indexation médico-économique préalablement réalisée par les médecins et à l'indexation automatique produite par F-MTI.

4.2.2.3 Mesures d'évaluation

La précision et le rappel ont été utilisés afin de mesurer les performances du F-MTI. La proposition d'indexation produite automatiquement par l'outil a été comparée à celle effectuée manuellement et de manière médico-économique par les médecins pour les 794 comptes rendus d'hospitalisation. De plus, elle a été comparée à l'indexation manuelle descriptive produite par l'expert pour 100 comptes rendus d'hospitalisation.

De plus, nous avons identifié différents niveaux d'indexation, du moins précis au plus précis en prenant en compte le nombre de digits des codes CIM10. Pour un code CIM10, chaque digit supplémentaire ajoute un niveau de précision supplémentaire. Par exemple, le terme présenté par le code A03 («shigellose») est plus général que le terme («Shigellose à Shigella dysenteriae») associé au code A03.0. Nous avons considéré le nombre de digits en commun dans notre calcul. Par exemple, si F-MTI extrait le code A03 et que le médecin code A03.0 alors nous considérons que nous avons une correspondance de 3 digits. Il y a jusqu'à 5 digits dans un code CIM10, le 5e digit étant généralement dédié aux codes extensions de la CIM10.

Dans un premier temps, nous nous sommes intéressée à tous les codes CIM10 extraits. Puis, nous avons voulu nous pencher sur les performances du F-MTI en matière d'extraction de diagnostics et de symptômes. Nous avons utilisé pour cela les types sémantiques de l'UMLS (voir section 2.3.2.3.3). Chaque code CIM10 dans l'UMLS est associé à un type sémantique dont «diagnosis» (pour diagnostic) et «symptom» (pour symptôme). Au moment de l'évaluation de l'extraction de diagnostics ou de symptômes, nous n'avons pris en compte que les codes de diagnostics (5 025 codes concernés) ou de symptômes (221 codes concernés).

Dans chaque secteur, les médecins codent généralement leurs comptes rendus à l'aide d'une liste restreinte de codes CIM10. Ces listes contiennent généralement les codes classants [5] selon leur spécialité. Nous avons ainsi restreint nos évaluations aux codes de chaque secteur : Cardiologie (326 codes concernés) et Pneumologie (317). Pour la CIM10 et la CCAM, l'assignation de la spécialité concernées pour chaque code a été effectuée par l'équipe CISMeF [Massari08] (voir section 5.8.1 et 7.3).

4.2.2.4 Résultats de l'extraction de termes CIM10 pour les dossiers patients

Nombre de digits pris en compte	Précision	Rappel
1	30,0	90,1
2	15,0	72,1
3	6,8	51,0
4	3,5	30,8
5	3,4	29,7

FIGURE 4.6 – Résultats de l'évaluation de l'indexation automatique produite par F-MTI comparée à l'indexation manuelle médico-économique des 794 comptes rendus

La première évaluation (voir figure 4.6) montre une précision de 3,4% et un rappel de 29,7% par rapport à une indexation médico-économique. De plus, nous pouvons constater que plus l'indexation considérée est précise plus la précision et le rappel diminuent passant d'une précision de 30% à 3,4% et d'un rappel de 90,1% à 29,7%.

Nombre de digits pris en compte	Précision	Rappel
1	26,3	93,0
2	11,7	75,8
3	6,3	57,9
4	3,2	37,0
5	3,0	35,7

FIGURE 4.7 – Résultats de l'évaluation de l'indexation automatique produite par F-MTI comparée à l'indexation manuelle médico-économique des 490 comptes rendus de Cardiologie

5. Classant pour les GHM, voir section 2.4.3.1

Nombre de digits pris en compte	Précision	Rappel
1	36,1	85,5
2	20,3	66,1
3	7.6	40,0
4	4,1	20,7
5	4,0	19,9

FIGURE 4.8 – Résultats de l'évaluation de l'indexation automatique produite par F-MTI comparée à l'indexation manuelle médico-économique des 304 comptes rendus de Pneumologie

Les résultats sont différents selon le secteur choisi (voir figures 4.7 et 4.8). Ainsi l'évaluation de l'indexation automatique pour les comptes rendus de Cardiologie montre une précision de 3,0% et un rappel de 35,7% pour 5 digits. Alors que pour les comptes rendus de Pneumologie, on obtient 4,0% pour la précision et 19,9% pour le rappel.

	CARDIO		PNEUMO	
Nombre de digits pris en compte	Précision	Rappel	Précision	Rappel
5	15,4	76,7	51,3	75,4

FIGURE 4.9 – Résultats de l'évaluation de l'indexation automatique produite par F-MTI comparée à l'indexation manuelle médico-économique en ne considérant que les diagnostics et les termes reliés à la spécialité «cardiologie» ou «pneumologie» selon le secteur d'origine des comptes rendus

Une deuxième évaluation ne prenant en compte que les termes reliés à la spécialité «pneumologie» ainsi qu'aux types «diagnostics» et «symptômes» pour l'indexation des comptes rendus de Pneumologie a été effectuée. Pour les comptes rendus de Cardiologie, restreints aux termes reliés à la spécialité «cardiologie», la précision obtenue est de 15,4% et le rappel de 76,7% pour l'extraction de diagnostics. Pour les comptes rendus de Pneumologie, restreints aux termes reliés à la spécialité «pneumologie», la précision obtenue est de 51,3% et le rappel de 75,4% pour l'extraction de diagnostics.

L'indexation des symptômes dans le secteur de la Cardiologie montre une précision de 41,0% et un rappel de 96,1% (voir figure 4.10). Dans le secteur de la Pneumologie, nous avons une précision de 39,3% et un rappel de 97,5%.

La dernière évaluation a été effectuée sur 100 comptes rendus indexés de manière

	CARDIO		PNEUMO	
Nombre de digits pris en compte	Précision	Rappel	Précision	Rappel
5	41,0	96,1	39,3	97,5

FIGURE 4.10 – Même évaluation en ne considérant que les symptômes

	Indexation médico-économique			Indexation descriptive		
Nombre de digits pris en compte	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
5	2,6	38,0	4,9	3,7	32,9	5,8

FIGURE 4.11 – Résultats de l'évaluation de l'indexation automatique produite par F-MTI comparée aux indexations humaines médico-économiques et descriptives des 100 comptes rendus d'hospitalisation

médico-économique par les médecins et de manière descriptive par un expert. Les résultats montrent une précision de 2,6% et un rappel de 38,0% (voir figure 4.11) comparés à une indexation médico-économique et une précision de 3,7% et un rappel de 32,9% par rapport à une indexation descriptive.

4.2.2.5 Discussion

Indexation descriptive et médico-économique L'indexation proposée par F-MTI devrait être plus proche d'une indexation descriptive que d'une indexation médico-économique (F-mesure de 5,8% vs. 4,9%) puisqu'il extrait l'ensemble des termes qu'il peut trouver dans un compte rendu sans appliquer les règles de codage du PMSI. Cependant, le meilleur rappel étant obtenu par rapport à une indexation médico-économique (38,0% vs. 32,9%), il est difficile de trancher.

Dans son activité quotidienne, notre expert code également ses comptes rendus de manière médico-économique aussi étant habitué à utiliser certains codes cela a pu avoir un impact sur son indexation descriptive.

Analyse de l'indexation produite par F-MTI Les résultats mettent en évidence une précision très faible (de l'ordre de 3%). Ceci est dû au grand nombre de codes proposés par F-MTI (26 codes en moyenne par compte rendu) comparé au nombre de codes assignés par le médecin (3 codes en moyenne par compte rendu). De plus, le rappel montre que le système n'est capable d'extraire qu'environ un tiers des codes de référence.

L'analyse des erreurs d'indexation produite par F-MTI montre que :

– F-MTI indexe l'ensemble des termes contenus dans les comptes rendus alors que

le médecin ou l'expert ne code que ce qu'il juge important ce qui entraîne une faible précision. Il est difficile pour F-MTI de déterminer quels codes sont les plus importants parmi tous ceux qu'il a extrait. Les codes les plus importants ne sont pas plus représentés dans les comptes rendus médicaux que les autres. L'hypothèse selon laquelle les codes les plus importants sont présents dans la conclusion n'a pas non plus été vérifiée puisque les résultats obtenus n'ont pas été significatifs [Pereira05]. Il est donc important d'injecter des connaissances médicales pour les identifier.

L'une des solutions est d'éliminer les redondances : les diagnostics et leurs symptômes ou différentes formes du même diagnostic ou bien encore la manifestation et la maladie initiale. Le médecin ne code généralement pas les symptômes ou les manifestations associées sauf s'ils ne sont associés à aucun diagnostic. La règle suivante pourrait être appliquée : si deux codes CIM10 coexistent l'un étant un symptôme ou la manifestation de la maladie représentée par le second code alors le code du symptôme ou la manifestation est éliminé. En revanche, un symptôme ou une manifestation non associée à une maladie sera conservé. La CIM10 contient les relations manifestations/ maladies, ce sont les relations dagues/ astérisque (voir section 2.4.3.2). Mais elle ne contient pas les relations «symptôme de» ou «diagnostic de». On retrouve ces relations dans la SNOMED CT qui est reliée par des relations de transcodage, à la CIM10. Un travail a été réalisé très récemment par un doctorant de l'équipe CISMeF [Merabti08a] pour transposer ces relations de la SNOMED CT à la CIM10. Une future version de F-MTI intégrera ces règles et ces relations.

Une autre solution peut être d'utiliser les actes médicaux (les co-occurrences entre les codes CIM10 et CCAM et les liens sémantiques entre eux) ou les prescriptions. Ceci peut permettre d'identifier les diagnostics importants qui ont demandé un acte médical ou une médication appropriée. De futures évaluations testeront ces méthodes.

Enfin une dernière solution serait d'intégrer les travaux de P. Avillach [Avillach08a] utilisant les relations sémantiques de l'UMLS afin de déterminer les diagnostics les plus importants.

- les erreurs liées au contexte : les problèmes de négation induisent des erreurs d'indexation. Les négations ne sont pas prises en compte par l'algorithme du sac de mots. S'il est noté dans un compte rendu «Le patient n'a pas d'asthme» le terme «asthme» est indexé par F-MTI alors qu'il ne l'est pas par l'indexeur humain. Ceci contribue à faire diminuer la précision. Les autres méthodes développées prenant en compte la négation, le dictionnaire de termes et le dictionnaire de constituants pourront pallier ce problème.

  Un autre contexte pose problème : le contexte d'un diagnostic touchant un proche ou une autre personne de la famille. De la même facon que nous avons pris en compte la négation, cet autre contexte devra être géré comme le fait Chapman [Chapman07].
- la faible qualité des comptes rendus. Un compte rendu mal rédigé ou peu précis entraîne une baisse de la mesure du rappel. Les comptes rendus contiennent des abréviations ou des fautes d'orthographe qui ne permettent pas à un système

automatique de coder le code CIM10 associé alors que celui-ci est codé par le médecin. Les comptes rendus présentent aussi un manque de précision dans les diagnostics voire même l'absence de certains diagnostics. Certains diagnostics sont codés par les médecins alors qu'ils ne figurent pas dans les comptes rendus d'hospitalisation. Ceci peut être le cas lorsque le médecin code le séjour du patient qu'il a traité sans consulter le compte rendu qu'il a auparavant rédigé ou qu'une autre personne de son équipe a rédigé.
- les compétences en matière d'indexation des codeurs sont parfois en cause. Les médecins sont le plus souvent peu ou pas formés à l'indexation des comptes rendus. Les comptes rendus ne sont pas toujours codés par la personne qui a pris en charge le patient. Ils peuvent être codés par un autre médecin ou une secrétaire ce qui peut entraîner des erreurs d'indexation.
- des erreurs liées à la méthode du sac de mots (voir section 3.8.1). De plus, des termes lexicalement proches peuvent être des maladies différentes.
- des problèmes de transcodage qui lient des concepts non équivalents possédant des degrés de précision différents. Le transcodage CIM10-MeSH a été analysé. Nous avons pu mettre en évidence un nombre important de différences de granularité ou de sens entre les concepts liés par ces relations de transcodage. De plus les transcodages ne lient pas les codes extensions (codes à 5 digits) puisque ce transcodage est issu de la CIM10 anglaise qui ne contient pas ces codes.
- F-MTI ne peut «raisonner» comme un médecin et ne peut pas associer des éléments provenant de différents endroits du compte rendu. Il sera donc nécessaire d'implémenter des règles médicales afin d'indexer correctement les comptes rendus.

Qualité de l'indexation différente selon les secteurs La qualité de l'indexation automatique CIM10 dépend du secteur dans lequel celle-ci est effectué. Ainsi, une précision plus élevée (4,0% vs. 3,0%) et un rappel plus faible (19,9% vs. 35,7%) sont obtenus pour les comptes rendus de Pneumologie. Cette disparité peut s'expliquer par la différence de contenu. Les rubriques renseignées ne sont pas les mêmes (exemple : les comptes rendus de Pneumologie contiennent le motif d'hospitalisation à la différence des comptes rendus de Cardiologie). Les médecins ont des facons différentes de rédiger les comptes rendus selon leur formation. Les méthodes de codage varient selon le secteur médical. À Rouen, les cardiologues ne codent que le strict nécessaire pour le PMSI. Alors que les pneumologues codent tout même ce qui semble peu important. On trouve aussi beaucoup de malades polypathologiques en Pneumologie ce qui peut expliquer le nombre de codes important. Les cardiologues codent ainsi en moyenne 1.4 codes pour leurs comptes rendus et les pneumologues 5.8.

L'indexation des symptômes et des diagnotics La restriction aux termes reliés à la spécialité et aux termes «diagnostics» et «symptômes» montre des résultats intéressants avec un rappel de plus de 75% pour l'indexation des diagnostics et de

96% pour l'indexation des symptômes. Comme il a été fait état d'un nombre trop élevé de codes présentés aux indexeurs humains par F-MTI nous pouvons imaginer leur présenter l'indexation automatique sous différentes vues pour une aide à l'indexation semi-automatique. Les diagnostics et les symptômes pourraient être deux vues.

Méthodes d'évaluation Le pourcentage d'erreur inter-indexeurs est de plus de 10% pour le 3ème digit du code CIM10 et de 25 à 30% pour le 4ème digit [6]. D'après D. Nakache, le niveau de reproductibilité totale est très faible (18% de consensus total) entre deux indexations humaines [Misset05]. Ces observations permettent de soulever le problème de l'évaluation qui peut expliquer les résultats. Le codage n'étant pas reproductible, il faudra réaliser une analyse qualitative fine des résultats obtenus par l'algorithme.

Une évaluation à plus grande échelle permettrait peut-être de faire pencher la balance de manière plus significative. Un corpus plus important de comptes rendus indexés par plusieurs indexeurs humains pourrait être envisagé. La référence serait alors le consensus de plusieurs indexations humaines.

Une évaluation qualitative manuelle est en cours par notre expert qui pour chaque code indexé indique si celui-ci est pertinent ou non.

Comparaison à d'autres systèmes d'indexation automatique Pour le même corpus de 100 comptes rendus, l'outil MAIF [Névéol05a] couplé à un transcodage MeSH-CIM10 (voir section 2.5.3.2) obtient une précision de 15% et un rappel de 28% [Pereira05]. Ceci permet de comparer un système d'indexation multi-terminologique pour la CIM10 et un système d'indexation indirecte en CIM10. L'indexation multi-terminologique obtient un meilleur rappel mais elle produit une précision plus faible.

Le système MTI [Aronson00] donne une F-mesure de 85% sur un corpus statistiquement normalisé de 1 000 comptes rendus de radiologie. Un corpus statistiquement normalisé permet d'obtenir des résultats très élevés mais très éloignés de la réalité. Une version francaise de MTI pourrait être envisagée afin de comparer les résultats (voir discussion).

Le système CIREA [Nakache07] produit une meilleure précision et un meilleur rappel. Afin de confirmer ces résultats, il faudrait évaluer ces deux outils sur le même corpus.

Enfin l'outil MedCKARe [Baneyx06] produit de meilleurs résultats pour la Pneumologie mais il est incapable d'indexer des diagnostics d'autres secteurs. De la même facon ceci devra être confirmé par l'évaluation de ces deux outils sur le même corpus.

Bénéfices Le système F-MTI peut traiter un compte rendu en 1/2 seconde [7]. Ces performances permettent une indexation automatique en temps réel. Le temps nécessaire au codage diagnostique manuel étant largement influencé par celui nécessaire

6. Observation par rapport à plusieurs articles
7. Intégration à un serveur 4 cœurs

à la recherche des codes dans la terminologie, ceci permettrait de libérer du temps pour les praticiens. Une évaluation de F-MTI dans le cadre d'une indexation semi-automatique devra être réalisée.

Résultats de l'indexation CCAM La même étude a été réalisée pour l'indexation des comptes rendus en CCAM. Malheureusement F-MTI éprouve de grandes difficultés à extraire les termes CCAM. Ceux-ci sont très complexes (exemple représentatif de l'ensemble des termes de la terminologie : HPMA003 «Réparation de perte de substance par lambeau pédiculé de grand omentum [épiploon], en situation extraabdominale»). 85% des termes CCAM contiennent plus de 5 mots ce qui rend difficile leur extraction à partir d'une phrase. Ces termes nécessiteraient la création de libellés d'indexation, de transducteurs ainsi que de règles utilisant les termes des autres terminologies afin de recouper plusieurs éléments provenant de différents endroits du compte rendu. Les co-occurrences et les liens Tothem CIM10-CCAM ainsi que les éléments de l'ontologie Galen pourraient être utilisés [Rodrigues05].

Perspectives L'indexation, ici, a été réalisée grâce à l'algorithme du sac de mots. L'indexation à l'aide du dictionnaire de termes et de constituants, des comptes rendus en CIM10 sera évaluée.

D'autres comptes rendus provenant de plusieurs hôpitaux et d'autres secteurs pourront être utilisés afin de rendre les résultats indépendants du CHU de Rouen.

4.2.3 Évaluation de l'extraction de termes SNOMED pour les dossiers patients

Nous avons, par la suite, souhaité évaluer notre outil dans le cadre de l'extraction de termes SNOMED pour les comptes rendus à l'aide de l'algorithme du sac de mots. Pour ce faire, les performances du F-MTI ont été comparées [Pereira08a] à celle d'un outil commercial canadien SnoCode [8] (voir section 2.5.3.2), seul outil d'indexation automatique pour la SNOMED 3.5 qui, à notre connaissance, existe pour le francais.

4.2.3.1 Méthode d'évaluation

Au départ de cette étude, nous voulions comparer le résultat de l'indexation automatique produite par les deux outils F-MTI et SnoCode par rapport à une indexation SNOMED réalisée manuellement par un expert (Dr A. Buemi), sur l'échantillon des 100 comptes rendus utilisés dans l'évaluation CIM10. Cela aurait été, en France, la première expérience d'indexation manuelle de comptes rendus en SNOMED 3.5.

Les 100 comptes rendus ont été présentés à l'expert qui, suite à l'indexation d'un seul compte rendu, a démontré qu'une indexation manuelle était beaucoup trop fastidieuse et prendrait beaucoup trop de temps. Il lui a fallu plusieurs heures (8 heures) pour indexer un seul compte rendu de 3 pages. L'explication est liée à la complexité de la SNOMED 3.5 (voir discussion).

8. http ://www.medsight-info.com/IndexFr.html

Face à ce constat, il a été nécessaire de trouver un autre moyen de comparer ces deux outils. La projection des codes SNOMED vers une autre terminologie moins complexe et qui puisse être manuellement indexée a semblé être la solution la plus simple. La CIM10 déjà utilisée pour l'indexation des 100 comptes rendus choisis nous permet de comparer ces deux outils en terme d'extraction de maladies.

Nous avons donc, pour chaque ensemble de codes SNOMED produit par les deux outils, transcodé ces codes en leurs équivalents CIM10.

Tout d'abord, les deux résultats d'indexation générés par F-MTI et SnoCode ont été comparés sans référence avec des mesures simples. Puis les deux indexations ont été transcodées en CIM10 et comparées aux résultats de l'indexation manuelle descriptive réalisée par l'expert (voir section précédente).

Le transcodage réalisé par les deux systèmes est différent. SnoCode utilise le transcodage francais créé par la SFINM. F-MTI utilise la somme de deux transcodages SNOMED-CIM10 : celui de l'UMLS 2007AA et celui produit par la SFINM. C'est la raison pour laquelle nous avons réalisé deux évaluations : une avec les différents transcodages et l'autre avec l'utilisation, pour les deux outils, du même transcodage en l'occurrence celui utilisé par F-MTI.

4.2.3.2 Corpus d'évaluation

Nous avons repris les 100 comptes rendus indexés en CIM10 (voir section 4.2.2).

4.2.3.3 Mesures d'évaluation

Nous avons utilisé la mesure de Hooper (voir section 2.5.2) pour comparer les deux ensembles de codes SNOMED produits par les deux outils. Cette mesure est habituellement utilisée pour mesurer la consistance de l'indexation entre deux indexeurs humains. Nous l'utilisons ici afin de comparer les résultats de nos deux indexations automatiques, en considérant F-MTI et SnoCode comme deux indexeurs potentiels.

Nous avons également calculé le recouvrement de chaque ensemble l'un par rapport à l'autre.

Enfin, nous avons mesuré la précision et le rappel pour comparer à la référence CIM10 les codes SNOMED transcodés en CIM10 pour les deux outils.

4.2.3.4 Résultats de l'extraction de termes SNOMED pour les dossiers patients

La figure 4.12 montre que SnoCode extrait moitié moins de codes que F-MTI (54,9 vs 100.3). La moitié des codes SNOMED extraits par SnoCode a aussi été extraite par F-MTI (voir figure 4.13).

Les figures 4.12 et 4.14 présentent les résultats de la comparaison des deux outils après transcodage vers la CIM10. Le changement du type de transcodage produit des résultats différents. Le nombre de codes moyen extraits par compte rendu est passé de 7 à 17 codes extraits par SnoCode (vs F-MTI 26,5 codes et 4,2 pour l'indexeur humain).

F-MTI extrait beaucoup trop de codes par rapport à SnoCode et à l'indexation

	Nombre moyen de codes SNOMED par compte-rendu	Nombre moyen de codes CIM10 par compte-rendu en considérant les transcodages d'origine	Nombre moyen de codes CIM10 par compte – rendu en utilisant le même transcodage
F-MTI	100,3	26,5	26,5
Snocode	54,9	6,5	17,1
Indexeur humain	-	4,2	4,2

FIGURE 4.12 – Nombre moyen de codes par compte rendu

Pourcentage de codes F-MTI couvrant les codes Snocode	29,9%
Pourcentage de codes Snocode couvrant les codes F-MTI	51,5%
Mesure de Hooper	31,3%

FIGURE 4.13 – Évaluation des recouvrements des codes SNOMED extraits par les deux outils

	Transcodages différents		Transcodage semblables	
	Précision	Rappel	Précision	Rappel
F-MTI	4,4	30,7	4,4	30,7
Snocode	15,0	22,2	6,1	24,7

FIGURE 4.14 – Comparaison des deux outils avec et sans le même transcodage CIM10

manuelle, ce qui donne une précision très faible 4,4%. SnoCode produit une meilleure précision 15% et 6,1% avec le même transcodage. Les scores se rapprochent beaucoup lorsque l'on utilise le même transcodage. F-MTI produit un meilleur rappel (30,7% vs 22,2%) et une plus faible précision (4,4% vs 6,1%) par rapport à SnoCode.

4.2.3.5 Discussion

Comparaison entre SnoCode et F-MTI Il n'est pas surprenant que le nombre de codes générés par les deux systèmes varie grandement (moyenne de 54.9 codes SNOMED pour SnoCode vs. 100.3 pour F-MTI ; moyenne de 17.1 codes CIM10 pour SnoCode vs. 26.5 pour F-MTI). Ces variations sont dues au fait que SnoCode se base seulement sur les codes SNOMED alors que F-MTI se fonde sur 4 autres terminologies pour générer des codes SNOMED.

Dans la figure 4.13, la mesure de Hooper montre que les deux outils produisent des indexations aussi différentes que peuvent l'être deux indexations humaines (31,3%). À titre de comparaison à la NLM, les indexeurs manuels génèrent une mesure de Hooper de 39% pour l'indexation MeSH [Funk83b]. D'après les figures 4.12 et 4.14, nous pouvons envisager que les principales différences de résultats entre les deux ou-

tils sont liées aux différences de transcodage SNOMED-CIM10 utilisés. L'application du même transcodage que celui utilisé par F-MTI, a abouti à une diminution de 8,9% de la précision et une augmentation du rappel de 2,5%.

La projection des codes SNOMED vers la CIM10 a montré que, comparé à une indexation manuelle, SnoCode produisait une meilleure précision (+2%) et un plus faible rappel (-6%) en terme d'extraction de maladies. Les résultats peuvent être considérés comme assez proches alors que nous comparons un système mono-terminologique de plus de 20 ans d'expérience et un système multi-terminologique de seulement 6 et qui peut encore beaucoup évoluer. Sachant que SnoCode est un outil déjà commercialisé et en place dans certains hôpitaux, nous pouvons considérer que les résultats obtenus par F-MTI sont relativement satisfaisants.

Analyse des résultats L'analyse de l'indexation produite par F-MTI met en évidence quelques erreurs :

- L'extraction de termes non pertinents pour l'indexation, par exemple les termes de l'axe G de la SNOMED contenant les qualificatifs et termes de relations qui n'ont aucun sens lorsqu'ils ne sont pas reliés aux autres termes SNOMED.
- F-MTI (tout comme SnoCode) ne permet pas de relier des termes appartenant à différents axes de la SNOMED lors de leur indexation. Il n'existe pas de règles d'indexation à ce sujet. Il est donc nécessaire d'implémenter des règles afin d'indexer correctement les comptes rendus médicaux.
- Certains termes sont incorrectement retrouvés car l'extraction par la méthode du sac de mots ne permet pas de respecter l'ordre des mots. Des améliorations doivent être apportées dont l'implémentation de l'analyse sémantique des phrases.
- Le problème des transcodages qui ne relient pas systématiquement des concepts de sens strictement équivalent avec parfois des degrés de précision différents. Les transcodages devront donc être revus, par la suite, avec plus d'attention par nos équipes afin d'éliminer les transcodages inadéquats et ainsi faire diminuer le bruit généré par F-MTI.
- Le problème des redondances entre termes extraits : les diagnostics et leurs symptômes ou différentes formes du même diagnostic ou bien encore la manifestation et la maladie initiale. Les relations «symptôme de» et «diagnostic de» sont présentes dans la SNOMED CT qui est reliée par des relations de synonymie à la SNOMED 3.5 dans l'UMLS (car reliés aux même concepts UMLS - voir section 2.3.2.3). Un travail a été réalisé par un doctorant de l'équipe CISMeF [Merabti08a] pour transposer les relations «symptôme de» et «diagnostic de» de la SNOMED CT à la SNOMED 3.5. Une future version de F-MTI intégrera ces règles et ces relations.
- Le problème du contexte : antécédents, autre membre de la famille touchée, négations etc.... Des améliorations au niveau de l'analyse du contexte, avec par exemple des transducteurs pourront être implémentées.
- F-MTI ne peut raisonner comme un médecin et par exemple, associer des idées provenant de différentes parties du texte. Un système de règles pourra être utile

ici.
- Les problèmes de formulation : il existe un manque de précision au niveau des diagnostics non décrits dans les comptes rendus. Les médecins devront être invités à mieux décrire l'état de leur patient.

L'évaluation Cette approche d'évaluation consistant à employer un transcodage vers d'autres terminologies moins complexes pourra facilement être appliquée pour d'autres évaluations où l'indexation manuelle est difficile par exemple pour la SNOMED CT qui est beaucoup plus complexe que la SNOMED 3.5 et qui possède des liens d'équivalence avec la CIM10 dans l'UMLS.

Un expert n'indexe manuellement pas plus de 5 codes par compte rendu. En revanche, un outil automatique indexe dix fois plus de codes. Ce qui amène à la réflexion suivante : faut-il tout coder dans un compte rendu médical ? Tout y est-il important ? Lors d'une consultation le médecin préférera ne consulter que les éléments importants comme les maladies en cours pour une lecture rapide. Dans le cadre du budget, les termes d'indexation sont souvent limités aux codes classants (voir section 2.4.3.1). En revanche, dans un contexte de recherche d'information, d'analyse de données ou d'alertes, nous pensons qu'une extraction complète des concepts présents dans le compte rendu et décrits dans la terminologie est préférable.

Une évaluation secondaire qualitative sur les codes extraits par F-MTI sera effectuée par un expert en assignant à chaque code une étiquette «pertinent» «non pertinent» et «peu pertinent» (comme réalisé dans la section 4.2.4 pour le MeSH).

L'indexation SNOMED : une tâche complexe La nomenclature SNOMED 3.5 contient sept fois plus de termes et est 11 fois plus complexe que la CIM10 du fait de la possibilité de combinaison des termes provenant des 11 axes. De plus il n'existe à ce jour aucune règle d'indexation concernant la SNOMED 3.5. Vu le peu de temps dédié à la tâche d'indexation manuelle en SNOMED, nous pouvons imaginer que cette dernière ne pourra jamais être réalisée sans une assistance informatique ou une restriction très sévère des termes utilisés. Ces observations peuvent être transposées à l'indexation en SNOMED CT celle-ci renfermant plus de 370 000 concepts et 1 000 000 termes (presque trois fois plus que la SNOMED 3.5) et plus de 1 300 000 relations (dans sa version 2007).

4.2.4 Évaluation de l'extraction de termes MeSH pour les sites Web

Nous avons ensuite procédé à l'évaluation de F-MTI dans le cadre de l'indexation de documents dans CISMeF [Pereira08c] en utilisant l'algorithme du sac de mots.

4.2.4.1 Méthode d'évaluation

Comme nous l'avons vu précédemment, l'indexation automatique des documents en MeSH dans CISMeF est réalisée sur le titre des documents par un outil utilisant un

algorithme de sac de mots proche de celui de F-MTI [Névéol07b]. Nous avons voulu ici montrer la plus-value de l'utilisation de F-MTI pour réaliser cette indexation [**?**].

Nous avons évalué quelle était la méthode de normalisation (lemmatisation ou désuffixation) de mots la plus adaptée à notre problématique. Nous avons évalué aussi l'apport de l'approche multi-terminologique.

4.2.4.2 Le corpus d'évaluation

Pour réaliser cette évaluation, nous avons extrait l'ensemble des ressources CISMeF indexées manuellement dans le catalogue (soit 18 804 ressources en 2007). Nous avons choisi de constituer un corpus conséquent représentatif de l'activité de CISMeF. Les indexeurs avaient, lors de l'intégration des documents du corpus dans le catalogue, enregistré pour chacun en base : le titre, les types de ressource ainsi que les mots clés MeSH (les métadonnées du Dublin Core [Dekkers03]). Les types de ressource ont été sélectionnés manuellement à partir de la liste des types de ressource CISMeF. Les mots-clés MeSH (descripteurs et paires descripteurs/qualificatifs) ont été sélectionnés manuellement à partir de la liste des descripteurs CISMeF (incluant le MeSH) et des qualificatifs. Pour rappel, la terminologie CISMeF contient 24 357 descripteurs et 83 qualificatifs dans sa version 2007 mais le corpus qui a été constitué sur 13 ans a été indexé avec les versions du MeSH en application au moment de l'intégration de chaque ressource. A chaque mot-clé, l'indexeur a apposé un poids «majeur» en y accolant une astérisque ou «mineur» sans astérisque dépendant de sa capacité à décrire le contenu du document. Un mot-clé très représentatif du contenu de la ressource est considéré comme majeur (mineur sinon).

4.2.4.3 Mesures d'évaluation

Grâce au calcul de la précision et du rappel, nous avons déterminé la qualité de l'indexation MeSH effectuée automatiquement par F-MTI par rapport à l'indexation MeSH faite manuellement qui est considérée comme la référence.

F-MTI a été appliqué successivement avec différents paramètres :

1. F-MTI mono-terminologie incluant la désuffixation
2. F-MTI mono-terminologie incluant la lemmatisation
3. F-MTI multi-terminologies incluant la désuffixation
4. F-MTI multi-terminologies incluant la lemmatisation

Nous avons, par ailleurs, calculé les performances en considérant séparément trois catégories de termes :

- Les mots-clés (MC) : descripteurs MeSH ou paire descripteur/qualificatif. L'association descripteur/qualificatif est prise en compte (exemple : «cancer du sein» et «cancer du sein/prévention et contrôle» sont considérés comme non équivalents).
- Les descripteurs (D) : les descripteurs MeSH sans les qualificatifs qui peuvent leur être associés (exemple : «cancer du sein» et «cancer du sein/prévention et contrôle» sont considérés comme équivalents). Pour les descripteurs, nous avons

choisi d'évaluer en plus l'indexation automatique sur trois types de ressource différents reliés aux trois cibles majeures du catalogue CISMeF (les professionnels de santé, les étudiants et les patients) ; les types de ressource associés étant respectivement : «recommandations», «matériel et enseignement» et «patient» (ainsi que leurs fils).
- Les descripteurs majeurs (*D) : seuls les descripteurs, sans les qualificatifs qui peuvent leur être associés, assignés d'une astérisque sont pris en compte (exemple «*Pharyngite»).

Nous avons aussi réalisé une deuxième évaluation permettant de mesurer cette fois la qualité de l'indexation obtenue par F-MTI.

Cette évaluation est secondaire et reprend les résultats obtenus à la première évaluation. Nous avons extrait pour 1 000 ressources, tous les mots-clés considérés comme faux dans la première évaluation (c'est-à-dire les mots-clés extraits automatiquement mais non assignés par les indexeurs humains). Nous avons ensuite demandé à l'un de ces indexeurs d'associer à chaque mot-clé une appréciation sur l'impact qu'aurait l'indexation de ce mot-clé pour la ressource à des fins de recherche d'information. Trois types d'appréciation ont été assignés : «bon impact» «impact négatif» ou «impact mineur». Le corpus de 1 000 ressources était constitué de 200 ressources portant le type de ressource «recommandation», 400 pour le type de ressource «matériel et enseignement», 300 pour le type de ressource «patient» et 100 pour tous les autres types de resources confondus. Ces ressources ont été tirées au hasard afin de respecter les proportions du corpus d'origine.

4.2.4.4 Résultats de l'extraction de termes MeSH pour les sites Web

4.2.4.4.1 Comparaison entre la lemmatisation et la désuffixation

En comparant les résultats du F-MTI incluant la désuffixation à ceux du F-MTI incluant la lemmatisation (voir figures 4.15 et 4.16), on observe dans la plupart des cas que la précision est un peu plus basse et le rappel un peu plus élevé dans le cas de la désuffixation (moins 0,8% pour la précision et plus 0,4% pour le rappel dans le cadre de l'évaluation de l'indexation produite par F-MTI mono-terminologie comparée à l'indexation manuelle en prenant en compte seulement les descripteurs sur l'intégralité du corpus).

4.2.4.4.2 Résultats pour F-MTI multi-terminologies

Lorsque l'on compare F-MTI multi-terminologies par rapport à F-MTI mono-terminologie incluant la désuffixation, les résultats montrent une augmentation du rappel de 0,5% et une diminution de la précision de 3,5% (voir figures 4.15 et 4.16). Pour la lemmatisation, les résultats montrent que ce système multi-terminologique produit une baisse de la précision de 1,6% et une augmentation du rappel de 1%.

4.2.4.4.3 Résultats concernant les différents types de ressource

Lorsque l'on considère les résultats selon le type de la ressource (recommandations, enseignement et patient), les variations sont importantes. Les résultats produits par F-MTI multi-terminologique incluant la désuffixation montrent :

- 44,4% de précision et 25,7% de rappel pour les ressources d'enseignement
- 39,9% de précision et 18,7% de rappel pour les recommandations
- 38,3% de précision et 27,8% de rappel pour les ressources patients

Ces variations peuvent être reliées au nombre moyen de descripteurs MeSH assignés manuellement pour chaque type de ressource : 5,5 pour les ressources d'enseignement (vs. F-MTI : 2,1), 9,3 pour les recommandations (vs. F-MTI : 2,9) et 3,5 pour les ressources patient (vs. F-MTI : 1,5).

		Performances Précision (%) – Rappel (%)	
Type de termes	Type de ressources	(a) Mono/ désuffixation	(b) Mono/ lemmatisation
Mots-clés	Tous	29,4 - 13,0	28,3 - 12,1
Descripteurs	Tous	37,7 - 21,3	38,8 - 20,7
	Recommandations	43,7 - 17,9	47,4 - 16,9
	Enseignement	51,6 - 24,7	51,9 - 24,8
	Patient	42,4 - 27,5	43,7 - 25,9
Descripteurs majeurs	Tous	36,0– 36,4	37,7 - 35,6

FIGURE 4.15 – Performances du F-MTI mono-terminologie comparé à l'indexation manuelle sur les différents corpus

		Performance Précision (%) – Rappel (%)	
Type de termes	Type de ressources	(c) Multi / désuffixation	(d) Multi / lemmatisation
Mots-clés	Tous	25,9 - 13,5	26,7 - 13,1
Descripteurs	Tous	35,5 - 23,1	26,8 - 22,4
	Recommandations	39,9 - 18,7	42,3 - 17,3
	Enseignement	44,4 - 25,7	45,7 - 24,4
	Patient	38,3 - 27,8	38,9 - 26,4
Descripteurs majeurs	Tous	30,5 - 38,1	31,5 - 37,6

FIGURE 4.16 – Performance de F-MTI multi-terminologie comparé à l'indexation manuelle sur les différents corpus

4.2.4.4.4 Résultats concernant les différents types de termes

En comparant les résultats selon les différents types de termes (mots-clés, descripteurs, descripteurs majeurs), nous observons que F-MTI extrait de manière plus efficace les descripteurs majeurs, puis les descripteurs, et finalement les mots-clés.

Pour les descripteurs majeurs, F-MTI multi-terminologique incluant la désuffixation produit une précision de 30,5% et un rappel de 38,1%.

4.2.4.4.5 Résultats de l'indexation qualitative

L'analyse secondaire réalisée sur 1 000 ressources par un indexeur CISMeF a montré que 4,5% des descripteurs automatiquement assignés et considérés comme faux dans la première évaluation ont été considérés comme ayant un «bon impact», 79,6% un «impact négatif» et 15,9% un «impact mineur».

4.2.4.5 Discussion

Lemmatisation ou désuffixation ? Les résultats ont montré que les deux algorithmes de lemmatisation et de désuffixation produisent des résultats assez proches. Cependant, la lemmatisation donne une meilleure précision mais un rappel plus faible du fait de la sous analyse de variantes de termes. Le choix dépend donc de la tâche à effectuer, une tâche qui privilégie un minimum de bruit ou un silence minimum.

La lemmatisation est meilleure en terme de rappel et de précision pour la mono-terminologie. Ceci est inhabituel mais possible dans certains cas. Par exemple, pour le titre «Rapport concernant le symposium sur le syndrome d'alcoolisme foetal et les effets de l'alcool sur le foetus», F-MTI extrait les liens «syndrome d'alcoolisme foetal» et «alcoolisme» et «alcools». Dans le processus, nous filtrons les termes dont le sac de mots est inclus dans un autre sac de mots d'un autre terme : avec la lemmatisation «alcoolisme» est rejeté et pour la désuffixation «alcool» et «alcoolisme» sont rejetés. Ainsi la désuffixation donne une meilleure précision que la lemmatisation.

Mono-terminologie ou multi-terminologie ? Les performances du F-MTI mono-terminologie vs. F-MTI multi-terminologie sont assez proches en terme de précision et de rappel.

L'utilisation d'un système multi-terminologique permet d'exploiter un réseau sémantique plus large composé de plusieurs terminologies. L'accès à un réseau sémantique plus important permet *a priori* d'extraire plus de termes. Les résultats montrent pour un système multi-terminologique un meilleur rappel et une précision inférieure comparé à un système mono-terminologique.

La baisse de précision est due aux erreurs de transcodage indépendamment de l'outil F-MTI. Il est important dans notre méthodologie que tous les transcodages ne relient que des termes qui ont strictement le même sens. Les transcodages bidirectionnels CIM10-MeSH et SNOMED-MeSH de l'UMLS ont été analysés. Nous avons pu mettre en évidence un nombre important de différences de granularité ou de sens entre les concepts liés par ces relations de transcodage. Nous espérons obtenir une meilleure précision après élimination des erreurs de transcodage.

Impact sur l'indexation CISMeF La politique de recherche d'information de l'équipe CISMeF consiste à proposer à l'utilisateur peu de ressources mais très ciblées plutôt qu'une grande quantité de ressources qui demanderaient à l'utilisateur

de passer du temps à faire le tri (contrairement à Pubmed). En terme d'indexation, cela se traduit par le choix de favoriser une meilleure précision plutôt qu'un bon rappel, c'est pourquoi, en ce basant sur cette évaluation, la lemmatisation devrait être utilisée dans F-MTI. Malheureusement, nous ne devons pas oublier les considérations techniques. La lemmatisation demande un temps d'exécution deux fois supérieur à celui de la désuffixation. De plus l'installation et l'interrogation du Sémiographe [9] complexifient le procédé. En pratique, le gain de précision obtenu avec la lemmatisation n'est pas assez significatif pour justifier de l'augmentation de la complexité technique de l'algorithme. L'équipe CISMeF a donc décidé de garder la désuffixation comme méthode de normalisation de mot dans son environnment de production.

Qualité de l'indexation L'indexeur doit prendre en compte, selon Lancaster [Lancaster91] : 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. Une évaluation secondaire de la qualité de l'indexation produite est donc nécessaire.

L'analyse secondaire de l'indexation multi-terminologique par un indexeur CISMeF a montré l'intérêt de F-MTI comme aide à l'indexation manuelle. 4,5% des descripteurs MeSH évalués ont été considérés comme ayant un impact positif sur la recherche d'information. Ces termes n'ont pas été assignés manuellement et auraient dû l'être. 15,9% des descripteurs évalués ont été considérés comme ayant un impact mineur, ils auraient pu être assignés à la ressource en plus des descripteurs assignés manuellement.

Nous trouvons que 79,6% des termes extraits par F-MTI considéré comme du bruit (car non indexés par les indexeurs humains) étaient effectivement du bruit et avaient un impact négatif. Donc nous pouvons penser que la précision de F-MTI est en fait meilleure que ce que nous avions évalué.

Nous avons prévu plusieurs changements pour améliorer les performances du F-MTI : la correction des transcodages et l'utilisation d'éléments de contexte et de règles d'indexation. Ranger les termes par ordre d'importance permettra de diminuer le bruit.

F-MTI incluant la multi-terminologie et la désuffixation sera bientôt intégré à l'environment de production de CISMeF.

Comparaison à d'autres outils MTI [Aronson00] produit une précision de 29% et un rappel de 55% pour l'indexation des titres et résumés d'articles Medline. La précision est du même ordre que celle obtenue par F-MTI par contre le rappel est meilleur. Pour l'indexation de descripteurs majeurs, il obtient une précision de 81% et un rappel de 11%, donc une plus faible précision et un bien meilleur rappel que pour F-MTI. Une comparaison entre les outils MTI et F-MTI sur un corpus parallèle bilingue devra être réalisée afin de confirmer ces résultats.

9. L'outil de lemmatisation que nous avons utilisé.

MAIF [Névéol05a] obtient une précision de 6,2% et un rappel de 35,3% pour l'indexation de ressources CISMeF. La précision semble plus faible et le rappel meilleur, ceci devra être confirmé par une comparaison sur un corpus CISMeF identique.

4.2.5 Évaluation de l'extraction de termes TUV pour les RCP

4.2.5.1 Méthode d'évaluation

Le thesaurus TUV sera bientôt finalisé, et mis en place pour l'indexation des RCP au sein du Vidal. Jusqu'à présent les RCP ont été indexés à l'aide des quatre thésauris (indications, contre-indications, précautions d'emploi et effets secondaires). Le but, ici, est de simuler ce prochain mode d'indexation et d'évaluer les performances que pourrait apporter l'outil F-MTI. Nous avons utilisé pour l'indexation, la méthode du dictionnaire de termes.

4.2.5.2 Le corpus d'évaluation

Nous avons extrait un corpus de 5 191 RCP indexés manuellement par les indexeurs de l'équipe scientifique du Vidal avec les quatre anciens thésauri du Vidal. Ces RCP étaient au format PDF il a donc fallu les convertir en texte grâce au programme pdftotxt.

Il a fallu réaliser la table de transcodage ancien thesaurus - TUV. Nous avons créé celle-ci en croisant les informations issues de différentes bases de données du Vidal ainsi qu'en ajoutant les différents éléments qui pouvaient être manquants. Le TUV n'étant pas terminé cette table est strictement limitée aux termes de référence TUV existants qui constituent l'ensemble des termes d'indexation possibles (les concepts élémentaires n'étant pas utilisés pour l'indexation). Elle contient 7 834 correspondances entre les termes des quatre anciens thesaurus et les termes de référence du TUV.

En transposant l'indexation des anciens thesaurus pour chaque RCP en TUV, nous obtenons un corpus de 5 191 RCP indexés en TUV avec leur type d'indexation correspondant au thesaurus d'origine (contre-indications, indications, effets secondaires ou precautions d'emploi).

4.2.5.3 Mesures d'évaluation

Nous avons appliqué le dictionnaire de termes pour le TUV (créé à la section 3.7.3) sur ce corpus. Nous avons aussi appliqué les transducteurs pour identifier les négations (voir section 3.9.1). De plus, nous avons appliqué un patron d'extraction NOOJ pour identifier les rubriques du RCP et leurs localisations afin d'obtenir, pour chaque terme, la rubrique correspondante (qui correspondra pour nous au type du terme) (voir section 2.4.2.2) :

- À chaque terme de la rubrique «Indications» est associé le type indication (‹INDIC›).

- A chaque terme de la rubrique «Contre-indications» est associé le type contre-indications (‹CI›).
- A chaque terme des rubriques «Effets indésirables» et «Surdosage» est associé le type Effets secondaires (‹EII›).
- A chaque terme des rubriques «Précautions et Mise en garde» et «Conduite et utilisation de machine» est associé le type Précautions d'emploi (‹PE›).

Nous avons ainsi pu calculer la précision et le rappel, en comparant cette indexation produite automatiquement avec l'indexation manuelle TUV obtenue après transcodage.

Nous avons mesuré la précision et le rappel en considérant différentes catégories :

- chaque type de terme séparément (indications, contre-indications, effets secondaires, précautions d'emploi). Les codes TUV assignés automatiquement à la rubrique «indications» donc au type ‹INDIC› sont comparés aux codes TUV assignés manuellement à un RCP avec le type ‹INDIC›.
- en considérant tous les types de terme (moyenne de la précédente évaluation)
- en ne tenant pas compte des types de terme. Tous les codes TUV assignés automatiquement aux quatre rubriques sont comparés aux codes TUV assignés manuellement à un RCP avec l'un des quatre types en ne tenant pas compte du fait qu'ils appartiennent ou non au même type.

4.2.5.4 Résultats de l'extraction de termes TUV pour les RCP

Rubriques concernées	Précision	Rappel
Indications	48,1%	21,7%
Contre-indications	46,1%	23,5%
Effets secondaires	77,0%	59,4%
Précautions d'emploi	28,4%	49,3%
Total pour les quatre rubriques	52,9%	46,2%
Total en ne tenant pas compte des rubriques	57,6%	43,4%

FIGURE 4.17 – Résultats de l'évaluation de l'extraction de termes TUV à partir d'un corpus de RCP

Les performances du F-MTI montrent une précision de 57,6% et un rappel de 43,4% comparé à l'indexation manuelle (voir figure 4.17).

Lorsque l'on considère les performances selon les rubriques, les résultats sont très différents selon le type de terme considéré. Les meilleures performances sont obtenues pour les effets secondaires avec une précision de 77% et un rappel de 59,4%.

L'indexation automatique de F-MTI pour les indications est limitée avec un rappel de 21,7% ; celle pour les précautions d'emploi est très bruitée avec une précision de 28,4%.

4.2.5.5 Discussion

Performances générales Les performances obtenues sont satisfaisantes. Elles sont largement supérieures à celles obtenues par l'indexation des autres terminologies (voir sections précédentes). Ceci peut s'expliquer par le fait que contrairement aux autres terminologies les thesaurus Vidal ont été créés à partir du contenu des RCP. Les libellés des termes sont donc tout à fait en accord avec ce qui peut être trouvé dans les RCP. L'indexation en est largement facilitée.

Performances différentes selon les rubriques Les résultats sont différents selon les types considérées. En effet les termes sont plus ou moins longs et complexes selon les types. Ainsi les termes de type <INDIC> et <CI> sont plus complexes que les autres. Ils sont donc plus difficiles à extraire d'où un rappel plus faible.

Toutes les règles d'indexation (voir section 2.4.2.2), n'ont pas pu être intégrées. Certaines rubriques n'ont pas été prises en compte : «Composition» et «Posologie et mode d'administration» et «interactions médicamenteuses» qui peuvent contenir des termes «PE». De plus, nous avons considéré que chaque rubrique ne pouvait contenir que des termes d'un seul type alors que ce n'est pas vrai pour toutes les rubriques :

- Grossesse et Allaitement : termes indexés avec le type <CI> ou <PE> selon les cas.
- Précaution d'emploi et Mise en garde : contient parfois des termes <CI>.

Analyse des erreurs d'indexation L'analyse des résultats montre que la majorité des erreurs est due à l'insuffisance des variantes présentes dans le dictionnaire de termes pour le TUV. Ces variantes pourraient être retrouvées par une autre méthode que la méthode de l'algorithme du sac de mots.

De nombreuses erreurs sont liées à la conversion des documents PDF en texte avec des problèmes de retour à la ligne et, ainsi, de non reconnaissance de certains termes. La restitution des titres de rubrique est parfois mauvaise ce qui entraîne une mauvaise affiliation des rubriques. Les tableaux ne sont pas restitués alors qu'ils peuvent contenir des termes à indexer.

Perspectives Afin d'améliorer les résultats, nous envisageons d'intégrer toutes les règles d'indexation suivantes :

- liens contexte d'application : pour compléter l'indexation, des liens dits «contexte d'application» peuvent être créés automatiquement. Par exemple, une contre-indication peut avoir comme contexte une indication. Ceci peut être traité à l'aide de transducteurs traduisant les liens existants (exemple : «ne pas <PE> en cas de <INDIC>»). Ceci permettra de faire la distinction

entre plusieurs types lorsqu'un terme est indexé dans une rubrique pouvant contenir des termes de types différents.
- les fréquences pour les termes ‹EII› peuvent être ajoutées automatiquement à l'indexation. Les expressions à identifier peuvent être ajoutées au dictionnaire de terme (exemple : «très fréquent»).
- indexation des rubriques non prises en compte ici : «Composition», «Posologie et mode d'administration» et «interactions médicamenteuses» qui peuvent contenir des termes ‹PE›.

De plus, les travaux concernant l'XMLisation des RCP étant achevés, ceci résoudra les problèmes de conversion, de tableaux et améliorera la reconnaissance des rubriques. Dans cette indexation aucun transcodage n'a été utilisé, car il n'existe aujourd'hui aucun transcodage vers le TUV (indexation mono-terminologique). Le TUV pourrait être intégré dans un futur proche au metathésaurus de l'UMLS par l'équipe Vidal.

4.3 Conclusion

Nous avons effectué différentes évaluation de F-MTI qui ont permis de juger de ses performances. Le chapitre suivant présente les applications qui peuvent être faites de l'outil.

Chapitre 5

Applications du F-MTI

5.1 Introduction

Après l'évaluation de notre outil F-MTI, nous voyons dans ce chapitre les différentes mises en application envisagées.

5.2 Applications pour l'indexation semi-automatique de RCP : BIBLIS

5.2.1 Présentation de l'outil BIBLIS

Comme explicité à la section 1.3.2, la société Vidal avec l'aide du laboratoire IMAG de Grenoble travaille sur un outil d'indexation semi-automatique nommé BIBLIS. BIBLIS permet l'indexation des RCP en utilisant le TUV. Le développement de cet outil est fondé sur de précédents travaux du laboratoire IMAG, notamment sur l'outil Noésis, un outil pour l'annotation textuelle et conceptuelle de documents [Patriarche05]. Nous présentons BIBLIS car il est prévu par la société Vidal d'intégrer à celui-ci F-MTI [1] afin de proposer aux indexeurs humains une proposition d'indexation automatique pour les documents qu'ils indexent.

À l'avenir, à l'arrivée d'un nouveau RCP, l'indexeur sera invité à l'indexer à l'aide de l'outil BIBLIS. L'outil permet de visualiser le RCP ainsi que les différentes terminologies nécessaires à son indexation dont le TUV (et d'autres terminologies comme le dictionnaire ATC etc...). Cet outil permet de réaliser une indexation manuelle classique : sélection d'un terme d'une terminologie et indexation du RCP avec ce terme (création d'un lien entre le terme et le document). Les fonctionnalités principales proposées par BIBLIS afin de faciliter l'indexation des RCP sont (voir figure 5.1) :

- **navigation facilitée** à l'intérieur du RCP et dans les différentes terminologies (visualisation des différentes propriétés pour chaque terme)

1. Références : spécifications de l'outil BIBLIS rédigées par R. Patriarche (Timc) et B. Plaisantin (Vidal). Toute l'équipe scientifique ainsi que moi-même avons participé à la réflexion autour de ces spécifications.

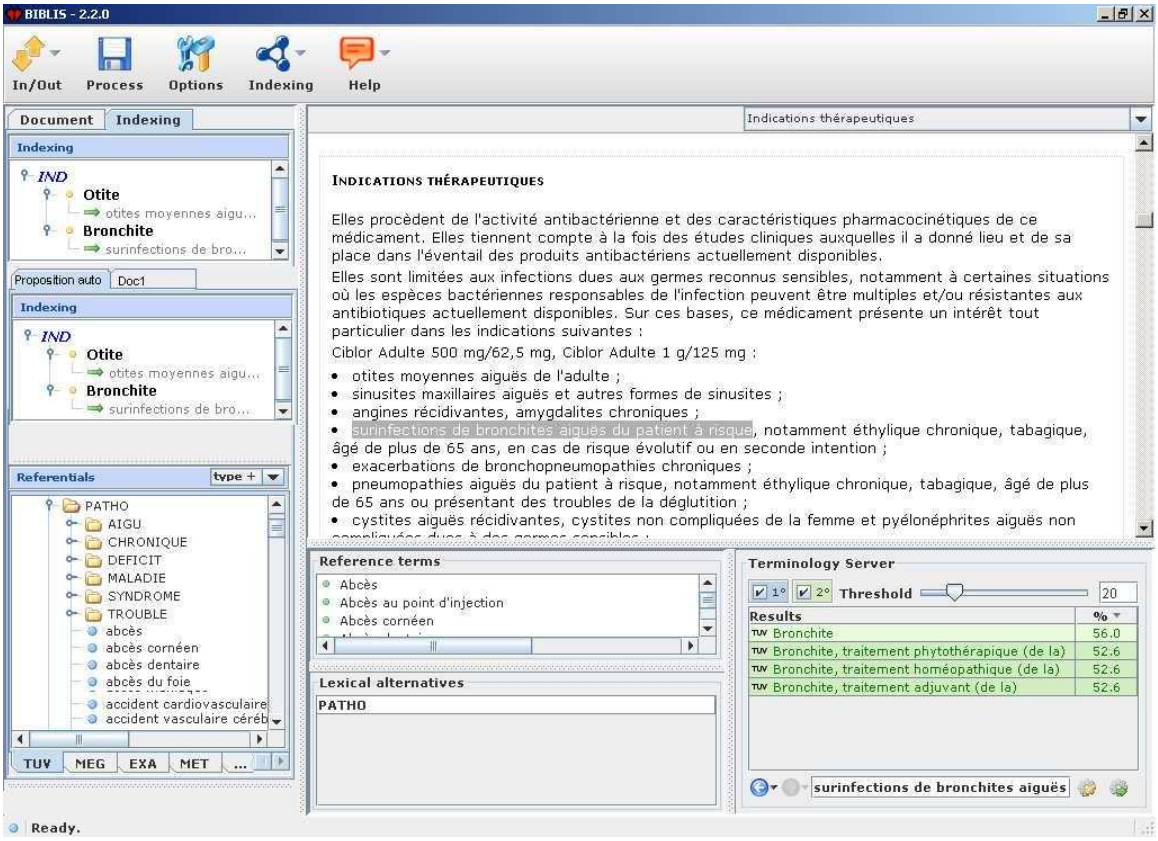

FIGURE 5.1 – Interface de l'outil d'indexation semi-automatique BIBLIS

- **proposition en temps réel de termes d'indexation automatique** à partir d'un fragment de texte du RCP ou d'une requête tapée par l'utilisateur grâce au serveur terminologique. Les termes retrouvés par le serveur de terminologies à partir de la requête sont rangés par ordre de pertinence par rapport à la requête (un score définit le pourcentage de couverture du terme par rapport à la requête)
- **attribution du type d'indexation** (exemple : ‹INDIC› pour «indication») au terme d'indexation choisi (le type portant le nom de la rubrique est proposé en première intention)
- **création du lien entre les termes d'indexation et le fragment textuel** du document contenant l'information indexée et sa localisation dans le RCP
- **visualisation de la couverture du document traité** (visualisation de tous les fragments indexés et de leurs positions dans le RCP)
- **création des liens** «contexte d'application» : il s'agit d'une mise en garde ou d'une indication liée à un terme indexé.
- **ajouts de commentaires**
- **possibilité de supprimer et d'ajouter un terme de l'indexation**
- **possibilité d'indexer des tableaux**
- **possibilité de réutiliser les indexations de documents traitant de spécialités proches**. Les indexations de documents proches peuvent être utiles

pour l'indexation d'un nouveau document. En effet, le RCP peut être un rectificatif ou une reprise complète d'un RCP d'une spécialité précédemment traitée. Les documents considérés comme proches doivent être sélectionnés manuellement par l'indexeur.

- **auto-apprentissage de l'outil** : si le fragment textuel lié au terme ne fait pas partie des variantes lexicales du terme celui-ci peut être proposé comme nouvelle variante. Une nouvelle variante ou un nouveau terme ne sera effectif qu'après validation par le gestionnaire de thésaurus.

5.2.2 Intégration de F-MTI dans l'outil BIBLIS

Dans l'outil BIBLIS, F-MTI permettra à l'indexeur de consulter avant le démarrage de sa propre indexation une proposition d'indexation automatique du document qu'il va indexer. F-MTI sera donc appliqué en amont, au moment où le document est recu par l'équipe données thérapeutiques.

Le fichier de sortie de F-MTI a été formaté au format d'entrée de BIBLIS. Ce fichier contient les termes proposés pour l'indexation du RCP, avec leurs types, ainsi que les fragments et localisations correspondants. Les fragments textuels seront soit la phrase dans laquelle a été trouvée le terme, soit les mots du sac de mots ayant permis l'appariement au terme d'indexation.

L'indexeur ouvrant BIBLIS pour indexer un nouveau RCP aura accès à la proposition d'indexation automatique de F-MTI et pourra dès lors choisir de garder certains termes, puis pourra les préciser en ajoutant certains contextes.

L'outil BIBLIS est capable de définir de nouvelles variantes au fur et à mesure de nouvelles indexations. Ces nouvelles variantes seront intégrées à l'outil F-MTI qui au fur et à mesure pourra évoluer et donner une meilleure indexation.

Avant toute indexation et afin de maintenir une homogénéité par famille, il faut connaître : l'indexation des autres spécialités de la même classe thérapeutique et quelles sont les spécialités indexées par les indications, contre-indications. . .du même groupe (voir section 2.4.2.2). La méthode statistique k-PPV (k Plus Proches Voisins) utilisée par A.Névéol dans ses travaux [Névéol05a] a montré que l'on pouvait utiliser l'indexation de documents proches afin de compléter une indexation automatique. Cette méthode déjà implémentée dans MAIF sera intégrée dans F-MTI.

Les documents proches peuvent être un rectificatif ou une reprise complète d'un RCP d'une spécialité précédemment traitée. Ces documents ont donc des portions de textes communs. Le découpage en phrase de F-MTI pourrait permettre d'identifier les documents partageant un ensemble de phrases communes et ainsi de proposer automatiquement des documents proches. Les travaux de T. Merabti qui permettent de classer les documents proches par une méthode mixte : statistique et sémantique pourront être aussi réutilisés [Merabti08b].

L'intégration opérationnelle de F-MTI sera finalisée à la suite de cette thèse en partie par mes soins.

5.2.3 Évaluation de l'apport de BIBLIS et de F-MTI (*via* BIBLIS) à l'indexation humaine

L'avis préliminaire des indexeurs de l'équipe Vidal est pour l'instant favorable à l'intégration de l'outil F-MTI à BIBLIS. L'équipe estime que cela facilitera son travail d'indexation, cependant ceci reste à évaluer dans leur pratique quotidienne.

Une première évaluation permettra d'analyser l'apport de l'outil BIBLIS pour l'indexation quotidienne de RCP. Cette évaluation consistera à comparer l'indexation produite avec l'outil et sans l'outil sur un corpus de RCP, les indexations étant produites par le même indexeur sur deux périodes proches (pour maximiser la consistance).

Une deuxième évaluation permettra d'évaluer l'apport de la proposition d'indexation automatique de F-MTI dans l'outil BIBLIS. Les indexeurs seront alors invités à indexer le RCP sans consultation de l'indexation F-MTI puis à consulter l'indexation de F-MTI et mesurer la quantité et la qualité des changements effectués après cette consultation.

5.3 Indexation automatique de dossiers patients

L'outil F-MTI pourra être utilisé pour l'indexation automatique des dossiers patients.

5.3.1 Aide au codage pour le recueil de données médico-économique

Les médecins ont de moins en moins de temps pour coder les dossiers de leurs patients. F-MTI pourrait être intégré à des logiciels de gestion de dossiers patients électroniques afin d'aider l'utilisateur dans le codage des maladies et des actes médicaux. F-MTI pourrait être intégré avec une interface spécifique ou de manière discrète dans l'éditeur de texte, par exemple, dans Microsoft Word qui est utilisé par les médecins et secrétaires au CHU de Rouen pour rédiger leurs comptes rendus hospitaliers. Microsoft Word permet de créer des macros [2], d'appeler des programmes et de créer de nouveaux boutons sur la barre d'outils. F-MTI pourra alors être utilisé après sélection d'une portion de texte jugée pertinente ou importante (ou l'ensemble du document) pour lequel il présentera l'indexation possible en fin de document (voir les étapes 1, 2 et 3 de la figure 5.2). Ce mode de fonctionnement est similaire à celui de l'outil SnoCode.

Comme nous l'avons constaté, l'outil F-MTI réalise une indexation descriptive, il pourrait être couplé à des outils médico-économiques en post-traitement pour réaliser un codage médico-économique pour le PMSI. Il existe des outils d'aide à l'indexation médico-économique permettant de naviguer dans les terminologies et permettant de simuler la fonction groupage afin de déterminer le coût d'un séjour (exemple : l'outil

2. Programmes permettant de commander les fonctionalités de logiciel.

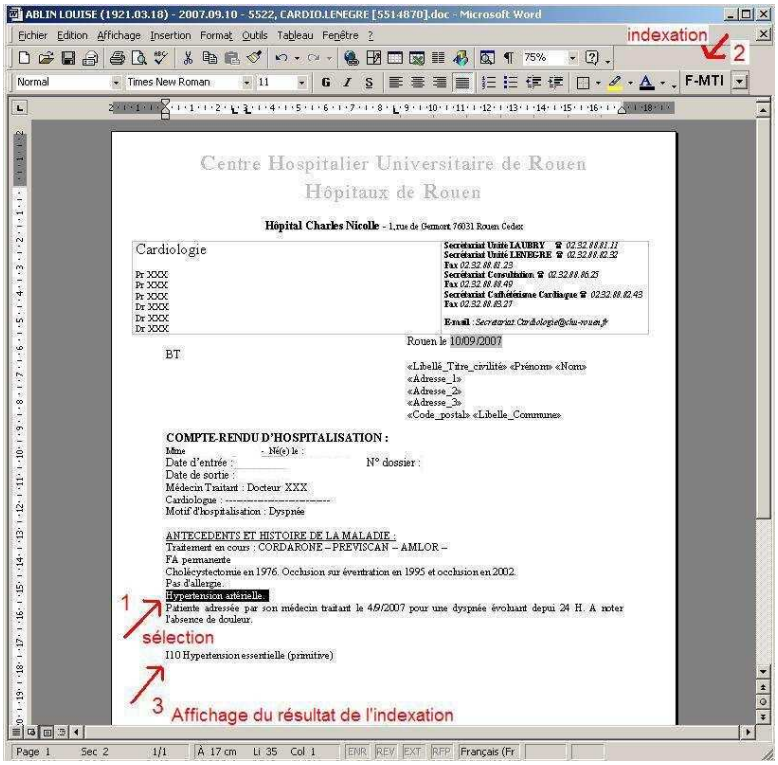

FIGURE 5.2 – Interface Word avec intégration du bouton F-MTI

WebFG de la société WEB100T). Aucun de ces outils ne permet d'appliquer les recommandations de codage de l'ATIH [3]. Ce type d'outil qui reste encore à développer pourrait alors améliorer la qualité du codage médico-économique, ainsi que la reproductibilité du codage, libérer du temps pour les praticiens, faire correspondre la valorisation financière du séjour avec le coût réel, et rendre le codage conforme aux règles qui sont difficiles à appréhender par les médecins.

5.3.2 Structuration des informations du dossier patient

Seuls les éléments nécessaires pour le recueil de données médico-économiques sont pour le moment structurés (diagnostics et actes utiles à la classification des séjours dans des groupes de tarification). Pourtant, il y a un réel besoin de structurer l'ensemble des informations du dossier patient électronique pour réaliser tous les traitements informatiques nécessaires en vue :

- d'une présentation claire du dossier médical du patient
- d'améliorer la prise en charge des patients (génération d'alertes dans le cadre du suivi du patient)
- d'aider le médecin à prendre des décisions (outils d'aide à la décision)
- de contrôler l'activité
- de rechercher des informations

3. Agence Technique de l'Information sur l'Hospitalisation

- de produire des données pour les études épidémiologiques
- de communiquer des données entre professionnels de santé

L'indexation descriptive de l'intégralité du contenu du dossier patient, des images incluses, avec des terminologies adaptées au contenu permettrait de structurer l'ensemble des informations. L'idée principale est de structurer *a posteriori* des dossiers patients non structurés (pour l'indexation de l'ensemble des comptes rendus du CHU de Rouen F-MTI mettrait environ 4 jours). L'indexation permettrait aussi de modéliser les liens sémantiques entres les différents éléments du dossier patient. Le codage médico-économique produit pour le moment est très nettement insuffisant. La faisabilité d'une structuration complète reste à discuter car aucune terminologie ne permet encore de prendre en compte l'ensemble de ces données [Nachimuthu07] [Campbell97]. En incluant à F-MTI les travaux de F.Florea sur l'indexation des images [Florea07a], et en intégrant à F-MTI l'ensemble des terminologies médicales (LOINC [4], MedDRA [5], WhoArt [6], etc...) une part importante des données d'un dossier patient pourrait être indexée.

L'indexation produite par F-MTI pourrait être utilisée comme suit pour la réalisation des différentes tâches :

- **une présentation claire du dossier médical du patient** pour les médecins et les patients (voir section 6.3)
- **aider le médecin à prendre des décisions** en améliorant les outils d'aide à la décision.
- **contrôler l'activité** La cohérence des données peut être contrôlée.
  Nous avons montré dans une autre étude [Pereira05] que le codage des médicaments pouvait aider à l'indexation de comptes rendus en CIM10. Il peut aussi mettre en évidence des incohérences entre prescriptions et diagnostics (exemple : un médicament ayant été prescrit pour un diagnostic non renseigné, ou un diagnostic n'étant traité par aucune médication). Ces travaux pourront être repris pour l'analyse des données produites par F-MTI.
  Dans le même orde d'idée, nous pouvons contrôler plusieurs paramètres, par exemple si chaque acte correspond bien à un diagnostic (en utilisant la table de transcodage CCAM-CIM10 voir section 5.8, ou des tables de co-occurrence).
- **rechercher des informations**
  La structuration des données facilite la recherche d'information. Le médecin peut rechercher quel médicament est le plus utilisé au sein de l'hôpital pour le traitement de l'asthme, les patients donneurs d'organes, quels patients ont besoin d'une appendicectomie pour pouvoir organiser les opérations etc...ce

4. La terminologie LOINC (Logical Observation Identifiers Names and Codes) permet de décrire les observations produites en laboratoire

5. L'objectif de MedDRA (Medical Dictionary for Drug Regulatory Activities) est de décrire toutes les étapes du développement des médicaments et les problèmes liés aux affaires réglementaires. Il inclut des termes pour la description des effets indésirables médicamenteux, les indications, les signes et symptômes, l'histoire familiale, les examens de laboratoire et les interventions chirurgicales.

6. La terminologie WhoArt (World Health Organization - Adverse Reaction Terminology) décrit les effets secondaires pour les médicaments.

qui peut rendre plus aisé son activité de tous les jours.
Il peut rechercher aussi à l'intérieur du dossier d'un patient quels sont les éléments qui se rapportent aux traitement de son asthme ou à quel moment a eu lieu son dernier bilan sanguin voir même comparer à deux instants t des résultats de biologie etc. ... Lorsque le dossier du patient est volumineux ou que celui-ci est atteint d'une maladie chronique cela peut aider à mieux prendre en charge ce patient. Le patient peut aussi retrouver des informations dans son propre dossier.
Une réflexion sur la conception d'un outil de type Google pour l'accès des patients à leurs données de santé a été apporté par C. Quantin [7]. Nous proposons, quant à nous, l'indexation des données par F-MTI couplée à un moteur de recherche de type CISMeF et à des stratégies de recherche adaptées. Une thèse a été lancée très récemment sur ce sujet en septembre 2008 dans l'équipe CISMeF (doctorant Ahmed-Diouf).
Cette recherche d'information peut être associée à de nombreux filtres. Nous proposons au chapitre 6 une méthode permettant de filtrer des informations par spécialité médicale.

- **produire des données pour les études épidémiologiques**
  L'épidémiologie étudie les facteurs influencant la santé et les maladies des populations humaines. Ce type d'étude nécessite de recueillir un maximum de données sur l'état de santé de chaque individu appartenant à l'échantillon de la population étudiée. Les acteurs du monde de l'épidémiologie se plaignant de la pauvreté des bases de données médico-economiques, là encore une indexation complète des informations aurait un grand impact.
- **communiquer des données entre professionnels de santé**
  Dans le cadre du DMP (Dossier Médical Personnel) dont le but est de mettre en place un dossier unique national pour chaque patient, un langage commun est indispensable. Ce langage commun ou tout au moins pivot envisagé pour l'instant est la SNOMED 3.5.

Le besoin d'un tel outil se fait sentir auprès des professionnels de santé. La littérature relate de nombreux travaux dans plusieurs pays [Fujii07]. Une phase de mise en oeuvre dans les hôpitaux pourrait être mise en place prochainement puisque l'ASISP [8] a lancé un appel d'offre pour la conception d'un extracteur de termes SNOMED.

5.3.3 Production de résumés et rédaction assistée de documents

Un médecin rencontrant un nouveau patient pour la première fois aura besoin pour affiner son diagnostic et assurer le suivi des soins, de connaître le parcours médical de ce patient. Autrefois était utilisé le carnet de santé, petit livret papier permettant en 2 minutes de voir les principaux faits marquants du parcours de santé

7. Présentation EMOIS2008

8. Agence des Systèmes d'Information de Santé Partagés, organisme chargé de mettre en oeuvre le DMP (Dossier Médical Personnel)

du patient. Depuis le passage au dossier électronique le médecin est contraint de consulter tous les documents décrivant les séjours du patient ou la fiche de synthèse de tous les séjours du patient quand elle existe. Pour les patients ayant effectué plus de 5 séjours à l'hôpital une consultation rapide devient impossible. Lorsque le DMP permettra de stocker un dossier national pour un patient, les données du dossier patient pourront être étendues à tous les établissements où il aura séjourné ce qui accroîtra d'autant le volume des dossiers.

Pour faciliter la gestion de ces informations, un résumé automatique pourrait être créé par F-MTI [9]. Ce résumé contiendrait : les principaux diagnostics en cours ou les plus récents, la médication en cours, les allergies, quelques informations administratives, les antécédents familiaux et les événements prévus (voir figure 5.3).

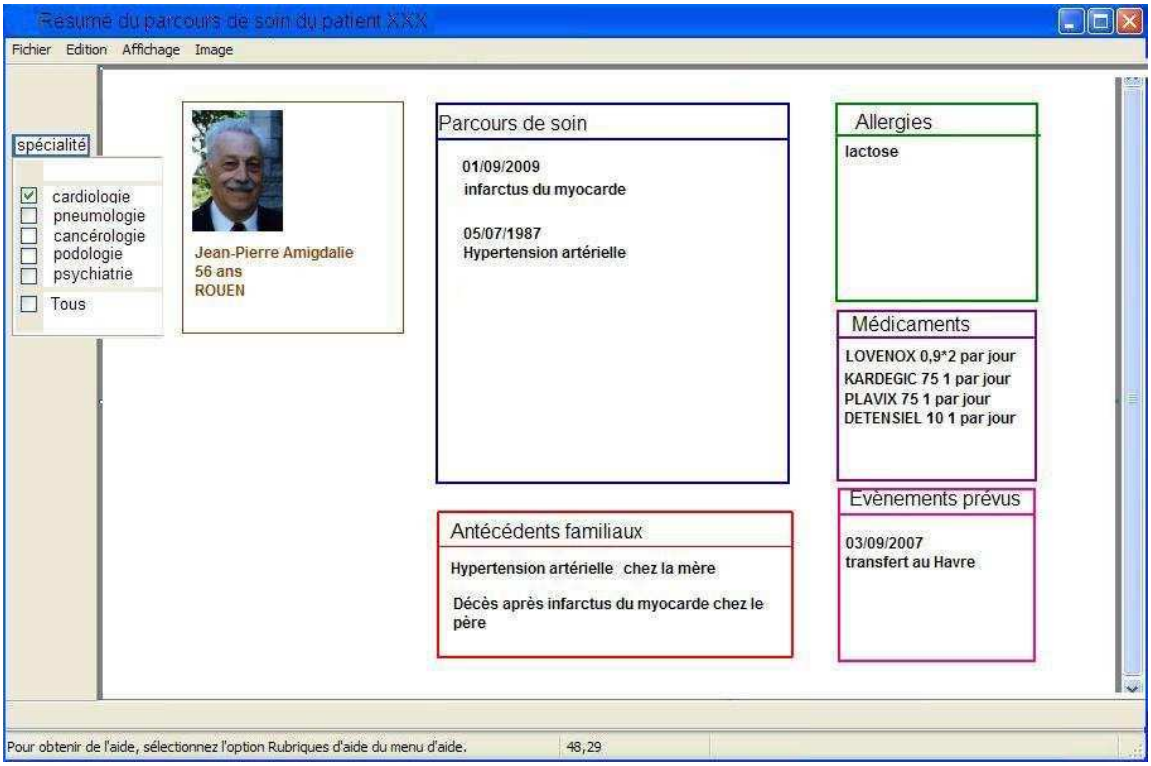

FIGURE 5.3 – Maquette d'une interface pour la présentation de résumés automatiques

Chaque acte et diagnostic serait extrait par F-MTI et reliés aux comptes rendus correspondants grâce à un hyperlien. Les données pourraient être présentées par problème ou/et par ordre chronologique.

Un transducteur ou un dictionnaire spécifique peuvent être utilisés pour l'extraction de dates. Un transducteur NOOJ existe déjà. Des dictionnaires ont été créés par P. Bramsen [Bramsen06] pour extraire les éléments liés au temps ou à l'enchaînement des événements comme la conjugaison ou les conjonctions de subordination anglaises (exemple : «after»). Un moyen de relier les termes aux dates correspondantes serait

9. C. Lovis s'était déjà intéressé à la création de résumés à partir du codage CIM10 [Lovis96].

d'associer chaque date à chaque terme ayant été extrait à partir de la même phrase ou du même paragraphe et d'ordonner les événements selon les conjonctions de subordination retrouvées.

Ils pourraient être aussi restreintes à une spécialité médicale. Ce résumé serait créé à partir de l'ensemble des comptes rendus rédigés pour un patient au cours de ses séjours à l'hôpital.

L'exercice du résumé consiste aussi à déterminer ce qui est important et pertinent dans le cadre du suivi du patient de ce qui ne l'est pas. Là est la difficulté et aucun outil n'est encore au point pour la surmonter. Afficher toutes les allergies, les problèmes récents ou chroniques et les médications associées, tous les antécédents familiaux recensés ainsi que les événements prévus dont la date est inférieure à la date du jour semble être une première piste à creuser. L'interface pour les résumés devrait être améliorée, implémentée et validée avec les professionnels de santé.

Nous pouvons imaginer de la même facon une génération automatique de lettres à partir d'une indexation : une lettre destinée à un patient (les synonymes patients seront privilégiés) ou à un collègue professionnel de santé (les termes techniques peuvent être conservés).

Un des principaux problèmes identifiés comme rendant difficile l'indexation est la rédaction même du compte rendu qui n'est pas adaptée pour sa propre indexation. Les documents sont rédigés en langage libre, ils sont alors difficiles à appréhender pour l'ordinateur et par des humains ayant peu de connaissances du domaine. Une rédaction assistée de documents permettrait la rédaction de documents structurés et adaptés. Les tournures pourrait être imposées afin de faciliter l'indexation et faciliter la lecture pour les autres utilisateurs (voir chapitre 6 pour faciliter la lecture par les patients). Par exemple, contraindre l'utilisateur à ne pas utiliser d'abréviations ou lui proposer, dès qu'une abréviation est détectée, de la remplacer par le terme exact ou, s'il y a ambiguïté, de préciser le terme correspondant ce qui permettra d'éliminer à la source les ambiguités. Cet éditeur de texte contrôlé devra répondre en temps réel, il pourra utiliser l'outil F-MTI ; certaines améliorations et fonctionnalités seront à envisager pour rendre cet éditeur opérationnel.

5.4 Indexation automatique de ressources Web

Vu les performances obtenues par F-MTI pour l'indexation automatique des sites Web, il devrait remplacer l'algorithme du sac de mots qui fonctionnait jusqu'à ce jour pour l'indexation automatique en MeSH des titres de ressources dans CISMeF (voir section 3.8.1). Il pourrait aussi remplacer ce même algorithme pour le traitement des requêtes tapées par les utilisateurs dans le moteur de recherche CISMeF.

Seule la terminologie MeSH est aujourd'hui utilisée pour l'indexation des ressources alors que d'autres terminologies pourraient améliorer cette indexation et permettre une recherche plus précise et plus adaptée selon les utilisateurs. À titre d'exemple, la CCAM est mieux adaptée à la description des actes médicaux que le MeSH. Une recherche de ressources concernant des actes médicaux restera très généraliste avec le MeSH alors qu'elle sera très précise avec la CCAM. De plus, les

professionnels de santé amenés à utiliser de plus en plus des terminologies spécifiques dans leur quotidien professionnel sont familiarisés avec certaines terminologies et seraient plus disposés à rechercher de l'information avec ces terminologies là. À ce titre, CISMeF souhaiterait passer d'un univers mono-terminologique à un univers multi-terminologique en indexant les documents à l'aide de plusieurs terminologies. Les terminologies d'intérêt pour CISMeF sont celles traduites en francais et les plus usitées dans le monde médical :

- le MeSH (Medical Subject Headings) et la terminologie CISMeF [Douyère04], les terminologies de bases de la recherche d'information
- la SNOMED 3.5 [Côté93] (Systematized Nomenclature of Medicine) la terminologie choisie par la France pour structurer les dossiers médicaux
- la CIM10 [19993] (Classification statistique International des Maladies et des problèmes de santé connexes (version 10))
- la CCAM [Rodrigues05] (Classification Commune des Actes Médicaux)
- la CISP2 [Lamberts87] (Classification Internationale des Soins Primaires 2$^{\text{ième}}$ version)
- le DRC [SFMG96] (Dictionnaire des Résultats de Consultation)
- la CIF/CIH [WHO] (Classification Internationale du Fonctionnement, du handicap et de la santé)
- la terminologie de MedlinePlus [10] (Base de données bibliographiques de la NLM)
- les concepts et le réseau sémantique de l'UMLS [Aronson01] (Système de Langage Médical Unifié) permettant l'interopérabilité entre plus de 100 terminologies
- et d'autres terminologies adaptées à la recherche de médicaments comme les noms commerciaux [11], les DCI [12] et les codes CIP [13], CIS [14], ATC [15] et CAS [16].

Ce virage a déjà été amorcé pour le catalogue CISMeF dans le cadre du projet PSIP (Patient Safety through Intelligent Procedures in medication, voir section 5.7) avec l'intégration des terminologies sur les médicaments [Letord] pour la création d'un portail d'information sur le médicament [17] (PIM). Le moteur de recherche Doc'CISMeF permet pour le moment de rechercher des codes CAS, CIS et ATC dans les titres et sous-titres des ressources.

Il va être très rapidement possible d'indexer automatiquement les ressources à l'aide de toutes les terminologies autour du médicament, puisque celles-ci sont en cours d'intégration dans F-MTI par S. Sakji (équipe CISMeF/LERTIM).

Pour les autres terminologies, le passage devrait se faire progressivement avec

10. `http://www.nlm.nih.gov/medlineplus/`
11. Données Vidal
12. Dénomination Commune Internationale
13. Code Identifiant de Présentation
14. Code d'Identification de la Spécialité
15. Classification Anatomique, Thérapeutique et Chimique
16. Chemical Abstract Service
17. PIM est le résultat d'une collaboration entre l'équipe CISMeF et la société Vidal. Il est accessible ici : `http://doccismef.chu-rouen.fr/servlets/PIM`

l'aide du projet Interstis (démarré en 2007 voir section 5.6).

5.5 Outil d'aide à l'indexation généraliste

F-MTI est un outil d'indexation automatique multi-document, multi-terminologique et multi-indexation capable d'indexer tout document texte à l'aide de cinq terminologies : CIM10, CCAM, SNOMED, TUV et MeSH.

Pour une indexation plus précise d'autres documents, les rubriques à indexer peuvent être spécifiées à F-MTI.

A priori n'importe quelle terminologie pourrait être indexée par F-MTI. Pour rajouter une terminologie, il suffit de :

- l'intégrer à la base de données multi-terminologique de F-MTI (analyser la structure de la terminologie et définir les ressemblances avec le modèle général de la base de données de F-MTI et intégrer l'ensemble dans les différents champs prévus)
- produire le sac de mots pour chaque terme (une fonction y est dédiée dans F-MTI)
- inclure dans la partie du code de F-MTI les règles d'indexation liées à cette terminologie et à la tâche effectuée

Ces étapes sont assez faciles même si elles sont dépendantes de la complexité de la terminologie à ajouter. Pour une meilleure indexation d'une nouvelle terminologie, la méthode de création du dictionnaire de variantes peut être appliquée.

5.5.1 Interface adaptée

Voici dans l'idéal comment nous imaginons notre futur outil d'aide à l'indexation générique. Les fonctionnalités ont été inspirées de nos travaux, de l'outil BIBLIS, et d'autres travaux (voir l'interface proposée figure 5.4) :

- une navigation facilitée à l'intérieur des documents à indexer (elle sera d'autant plus facile que la structure du document est précisée au départ dans l'outil, une fonctionnalité pourrait être dédiée)
- une navigation facilitée dans les différentes terminologies ainsi qu'une visualisation des différentes propriétés et liens inter et intra terminologies pour chaque terme (le serveur SMTS pourra être utilisé ici voir section 5.6)
- proposition de termes d'indexation automatique à partir d'un fragment de texte du RCP ou d'une requête tapée par l'utilisateur grâce au serveur terminologique. Les termes retrouvés par le serveur de terminologies à partir de la requête sont rangés par ordre de pertinence par rapport à la requête. Ce serveur pourrait être amélioré en combinant les méthodes du serveur de BIBLIS et de F-MTI.
- création du lien entre les termes d'indexation et le fragment textuel du document contenant l'information indexée et sa localisation.
- visualisation de la couverture du document traité (concerné par l'indexation)

- création des liens entres les termes (intra et inter-terminologies) : combinaisons de termes provenant d'axes différents pour la SNOMED, contextes pour les termes du TUV, associations mot clé/qualificatif/type de ressource pour le MeSH, associations des termes CCAM aux codes supplémentaires
- ajouts de commentaires pour un terme indexé
- possibilité de supprimer et d'ajouter un terme de l'indexation
- un terme peut être relié à plusieurs fragments textuels
- possibilité d'indexer des tableaux et des figures grâce aux outils de F. Florea. D'autres formats de documents pourraient être traités.
- possibilité de réutiliser les indexations de documents proches. Les documents proches pourront être déterminés par la méthode k-PPV d'A. Névéol ou par la méthode de related articles de T. Merabti ou par la fonctionnalité de comparaison de documents qui pourraient découler de F-MTI. Nous pouvons aussi envisager une méthode combinée

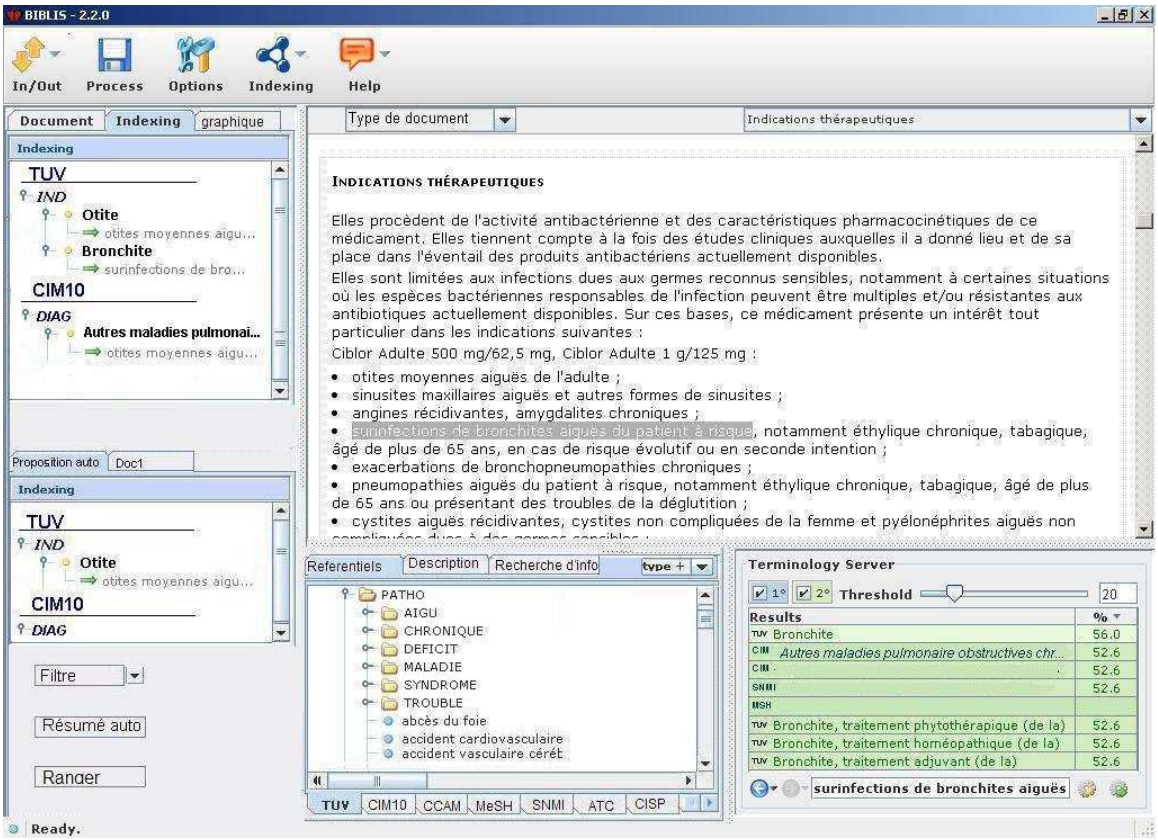

FIGURE 5.4 – Maquette d'une interface pour le logiciel d'aide à l'indexation multi-terminologique

- proposition d'ajout de nouveaux termes référents ou de nouvelles variantes. Si le fragment textuel lié au terme ne fait pas partie des variantes lexicales du terme celui-ci peut être proposé comme nouvelle variante (auto-apprentissage de l'outil).

- vues différentes sur l'indexation grâce à des filtres automatiques : par axe pour la SNOMED, par type pour le TUV, par diagnostic/symptôme pour la CIM10, par type de termes MeSH (qualificatifs, métatermes, type de ressources, mots-clés). Tous les types de termes pour chaque terminologie ainsi que les types sémantiques de l'UMLS peuvent être repris ici.
- association des éléments descriptifs de la ressource (date, titre, etc...)
- génération d'un résumé automatique avec les phrases les plus importantes, ou pour chaque rubrique les termes indexés. Le contenu du résumé pourra être paramétré.
- recherche d'information à partir d'un ou de plusieurs termes appartenant aux terminologies au travers de bases de connaissances sur Internet (CISMeF, Intute, Pubmed etc...). Les requêtes seront automatiquement traduites pour chaque site.
- ranger les termes par importance : la méthode de P. Avillach ainsi que celle de A. Névéol pourront être reprises et combinées ici.
- visualisation graphique de l'indexation : visualisation de l'indexation à plat ou de manière graphique telle que les icônes VCM de JB.Lamy [Lamy06] pour le TUV [18], ou visualisation en arbre créée par C.Abi Chahine de l'équipe CISMeF pour le MeSH et qui pourra être étendue aux autres terminologies.

5.5.2 Perspectives

Nous voudrions faire valider cette interface et les fonctionalités proposées, par des professionnels de santé et des indexeurs experts. Ceci pourra conduire à l'implémentation de cet outil d'aide à l'indexation générique.

Une version plus élaborée pourra introduire la fonctionalité d'indexation «à la volée». Ce genre d'indexation est à l'étude pour l'outil Snocode (pour la terminologie SNOMED 3.5) et pourrait être exploitée dans notre outil en indexation multi-terminologique. L'indexation «à la volée» consiste en l'indexation en temps réel du document au moment même où celui-ci est en train d'être rédigé. La rédaction peut être manuelle ou dictée à voix haute grâce à des outils de reconnaissance vocale [Happe03].

5.6 Intégration à un serveur multi-terminologie

Il existe un besoin fort pour un serveur multi-terminologie pour des internautes spécialistes de l'une ou l'autre des terminologies médicales francophones (documentalistes notamment), mais aussi des professionnels des traitements de l'information médicale, soucieux d'obtenir une source terminologique complète.

Le projet InterSTIS [19] (Interopérabilité Sémantique des Terminologies dans les

18. en cours de mise en place chez Vidal

19. Projet ANR-07-TECSAN-010-02. Les partenaires de ce projet sont : Le LERTIM de Marseille, l'équipe CISMeF du CHU de Rouen et du LITIS, l'INSA de Rouen, la société Vidal, la société Mondeca, le LIMSI, le DSPIM, le LabSTIC, la société Mémodata, le CHU de Saint Etienne et de

Systèmes d'Information de Santé Francais), débuté en 2007, a pour but d'améliorer et d'accroître l'interopérabilité sémantique entre les terminologies dans les systèmes d'information de santé francais.

Le projet propose la création d'un Serveur Multi-Terminologique en Santé [20] (SMTS voir figure 5.5) qui permet l'accès centralisé et aisé aux informations telles que libellés, définitions, liens entres les termes, etc.... Les terminologies d'intérêt sont celles traduites en francais et les plus usitées dans le monde médical (les mêmes qu'à la section précédente, voir le schéma 5.5).

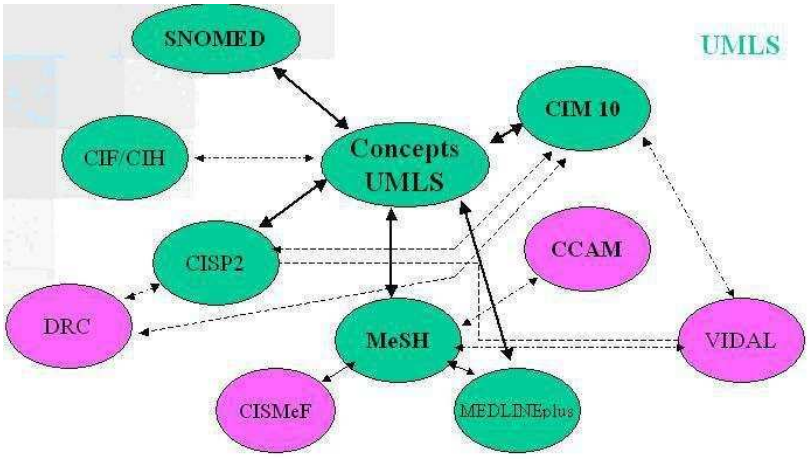

FIGURE 5.5 – Liste des principales terminologies médicales en langue francophone intégrées au SMTM et les relations entre elles (en rose : terminologies non intégrées au métathesaurus de l'UMLS)

Le projet consiste à développer une interface web [21] proposant notamment des fonctions de recherche dans les terminologies multilingues [22] (voir figure 5.6). Nous proposons d'utiliser F-MTI afin de traduire les requêtes des utilisateurs en termes appartenant aux différentes terminologies (la mise en place et l'adaptation de F-MTI à cette tâche sera réalisée dans le cadre de la thèse de S. Sakji et T. Merabti au sein de l'équipe CISMeF et pourra utiliser les technologies de Semantic Mining d'Oracle [23]).

Ce projet a demandé la création d'une base de donnés multi-terminologique [24] à même de recevoir les terminologies concernées (ainsi que d'autres terminologies

Lille et l'organisation HON (Ch).

20. Les fonctionnalités offertent par le SMTS s'apparentent à celles disponibles à partir du serveur de terminologie CISMeF (accessible via `http://www.chu-rouen.fr/terminologiecismef/`) qui prend en compte la terminologie CISMeF (incluant le thesaurus MeSH).

21. Une première version a été developpée dans le cadre d'un projet PIC (projet universitaire de 5e année)

22. notamment l'anglais et l'espagnol

23. Le Semantic Mining d'Oracle permet de créer des requêtes en SPARQL, le langage d'interrogation des ontologies

24. Les étudiants ont été co-encadrés par moi-même pour cette étape : présentation des différentes terminologies et aide pour la modélisation

éventuelles dans le futur). La structure de la base de données a été contrainte par le fonctionnement de la plateforme [25]. Le modèle généré est différent du modèle de base de données multi-terminologique de F-MTI dans le sens ou sa structure a été éclatée. Cependant une fonction permet de regénérer les tables conformes au modèle de F-MTI et utiles au fonctionnement de F-MTI [26]. En corollaire, il sera plus aisé d'intégrer les terminologies du SMTS manquant à F-MTI. La mise à jour des terminologies sera automatisée. F-MTI pourra ainsi bénéficier de cette fonctionnalité. Comparé

FIGURE 5.6 – Recherche sur le terme «Acute myocardial infarction» dans le SMTM

aux serveurs de terminologies industriels existants (DTS (Distributed Terminology System) de la société Apelon [27] et LExPlorer de la société Health Language [28]), ce serveur de terminologies offre des fonctionnalités plus importantes. Un autre serveur de terminologies médicales est en cours de réflexion dans le groupe hospitalier du Havre. Celui-ci est plus axé applications métiers du dossier patient électronique afin que les applications utilisent les mêmes référentiels [Briquet07].

25. Les technologies utilisées sont celles de la plateforme ITM (Intelligent Topic Manager) de la société Mondeca (http ://www.mondeca.com/fr/index.htm). ITM est une plateforme logicielle pour la gestion de référentiels métier, taxonomies, thésaurus, terminologies, bases de liens, bases de connaissances, catalogues, portails sémantiques, basée sur les technologies des ontologies (format SKOS : Simple Knowledge Organisation System et OWL : Web Ontology Language) et du Web 3.0

26. Les tables existantes sont trop nombreuses et la structure trop complexe pour que F-MTI fonctionne de manière optimale.

27. Pour plus de renseignements : `http://www.apelon.com/products/dts.htm`

28. Pour plus de renseignements : `http://www.healthlanguage.com/p&s software.html`

5.7 Optimisation de la prescription informatisée (PSIP)

Les effets indésirables (sévères) liés aux médicaments s'observent dans 6% des séjours hospitaliers entraînant au moins 10 000 décès en France (98 000 aux USA) par an. Ceci constitue un problème majeur de santé publique.

Dans ce contexte, le projet PSIP [29] (Patient Safety Through Intelligent Procedures in medication), débuté en 2008, a pour objectif de mieux recenser et connaître les effets indésirables liés aux médicaments dans le contexte hospitalier. Le projet propose de développer des méthodes innovantes destinées à contextualiser l'information et les alertes dans un nouveau système d'aide à la prescription.

Le système d'information hospitalier présente des fonctionnalités permettant de gérer le circuit du médicament. Le circuit du médicament est un des processus de soins les plus transversaux et structurants dans les établissements de santé. Chaque étape du circuit – prescription, dispensation, administration – est source d'erreurs potentielles pouvant mettre en jeu la sécurité des patients. Ces fonctionalités sont reliées au CPOE (Computerised Provider Order Entry). Ce système intègre des fonctions d'aide à la décision : suggestions de dosage, rappels automatiques (changements de doses par exemple), vérifie les interactions médicamenteuses et les allergies, communication entre tous les acteurs du circuit.

A partir des données extraites des CPOE, le projet va tenter de déterminer, par des outils de data mining, les situations à risque pour le patient, ceci sous forme de règles (voir figure 5.7).

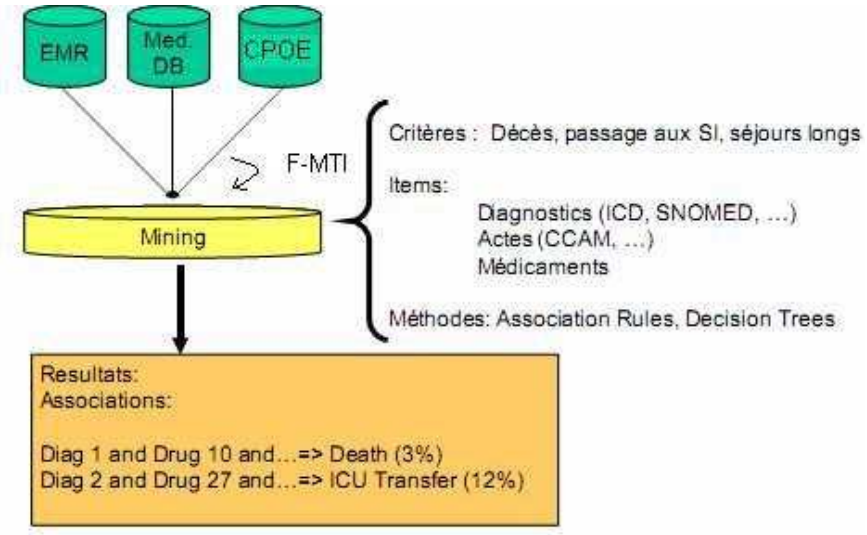

FIGURE 5.7 – Principes du projet

29. Projet FP7 ICT-2007.5.2. Les partenaires du projet sont : les CHU de Lille, de Rouen, de Denain et les Hôpitaux de Copenhague, les sociétés Oracle, IBM Danemark, Medasys, Vidal, KITE solutions et Ideea Advertasing et les universités UMIT (Autriche), AUTH (Grèce) et AAU (Danemark).

Certains hôpitaux ne possèdent pas de CPOE, et quand il existe, les données structurées du dossier patient (contenant des informations sur les prescriptions, dispensations et administrations de médicaments aux patients) sont incomplètes. Il est à souligner que les outils de data mining donneront de meilleurs résultats si les données au départ sont structurées et standardisées dans un langage commun.

C'est à partir de ce constat que l'équipe CISMeF et la société Vidal ont décidé d'utiliser l'outil F-MTI afin de compléter et de standardiser ces données à l'aide de terminologies adaptées et de développer le module «Semantic Mining» [30] du projet PSIP. Les connaissances extraites du Semantic Mining seront reversées en entrée des outils de Data Mining.

Dans cet objectif, il sera ajouté aux terminologies existantes dans F-MTI, les terminologies francaises et danoises suivantes : les noms des spécialités ainsi que les noms commerciaux, les codes ATC et les INN (International Nonproprietary Name, DCI). Ce travail est en cours de réalisation par S. Skaji, I. Kergourlay avec mon aide au sein de l'équipe CISMeF. Ces données sont fournies par le Vidal. De nouveaux modèles de comptes rendus pourront être aussi intégrés à F-MTI afin d'être en mesure de traiter des comptes rendus provenant d'hôpitaux et de secteurs différents.

5.8 Aide au transcodage

F-MTI permet de retrouver, à partir d'une requête ou d'une phrase, des termes appartenant à différentes terminologies. Une méthode identique pourrait être utilisée dans le cadre du transcodage automatique, afin de déterminer, à partir d'un terme, les autres termes appartenant à d'autres terminologies auxquels il renvoit.

5.8.1 CCAM-MESH

Nous avons testé cette hypothèse dans le cadre du transcodage CCAM-MeSH réalisé pour des besoins d'accès contextuel dans le dossier patient électronique (voir chapitre 6). Ce transcodage a été réalisé à la main par un expert du codage CCAM (P. Massari de l'équipe CISMeF) et par un expert du thesaurus MeSH (B. Thirion de l'équipe CISMeF). Nous avons profité de l'occasion pour réaliser un autre transcodage entièrement automatique grâce à F-MTI. Nous avons pu ainsi comparer ces deux transcodages (manuel et automatique) [Pereira07] [Pereira] et évaluer l'outil F-MTI dans une tâche de transcodage automatique.

L'indexation manuelle a consisté à analyser chaque terme CCAM et à l'associer à :

- 0 ou plusieurs termes MeSH :
  L'utilisation du modèle GALEN [Rodrigues05] donne une signification au code lui-même par les quatre lettres qu'il contient (voir chapitre 2.4.3.3), les deux premières correspondent à une région anatomique, la troisième à l'action, la quatrième à la voie d'abord. C'est à partir de ces significations et notament de la région anatomique et de la voie d'abord que l'expert a défini manuellement

30. Il est vraisemblable que je continue à travailler sur F-MTI dans le cadre du projet PSIP

les mots-clefs MeSH, et ainsi développé et validé un transcodage CCAM-MeSH. Chaque terme CCAM a été assigné à 1 ou plusieurs termes MeSH (4.8 (ecart type +/-3.5) codes MeSH en moyenne par code CCAM). Les mots clés MeSH assignés appartenaient à 2 des 15 catégories MeSH (A et E) correspondant aux termes techniques, anatomiques et diagnostiques. Par exemple : pour le code BACA008, F-MTI a assigné les termes MeSH : «procédés chirurgicaux »(technique), «sourcil»(anatomie) et «plaies et lésions traumatiques»(diagnostic).

- 1 ou plusieurs métatermes :
  La CCAM est classée par grands appareils et non par spécialités ce qui ne permet pas d'emblée de définir un métaterme pour les codes. La spécialité n'est pas non plus spécifiée explicitement dans le libellé. L'assignation s'est faite grâce aux connaissances de l'expert.

L'indexation automatique a consisté pour chaque terme CCAM à :

- utiliser F-MTI et notament la méthode du sac de mot [31] sur les libellés CCAM. Plusieurs termes cibles peuvent être nécessaires pour couvrir les différents mots d'un terme. Nous avons ainsi extrait les mots-clefs MeSH contenus dans chaque libellé CCAM. Par exemple, pour le code BACA008 «Suture de plaie du sourcil», l'expert a assigné les termes MeSH : «sourcil» (anatomie) et «plaies et lésions traumatiques» (diagnostic).
- ces mots-clefs MeSH sont reliés aux métatermes par des liens sémantiques (voir section 2.4.1.2). Nous avons ainsi pu déterminer les métatermes associés à chaque ensemble de termes MeSH pour chaque libellé CCAM. Pour un terme CCAM, les métatermes peuvent être nombreux (15 alors que l'expert a associé en moyenne 1.18 métatermes par libellé CCAM). Plusieurs mots-clefs MeSH d'une même liste peuvent être associés au même métaterme, nous avons décidé arbitrairement de calculer la fréquence pour chaque métaterme obtenu et de ne prendre que les deux métatermes les plus fréquents pour chaque liste de métatermes. De plus, nous avons pris en compte les associations de métatermes (exemple : chirurgie + neurologie = neurochirurgie).
- dans une deuxième étude, nous avons utilisé les mots-clefs MeSH associés manuellement aux libellés CCAM par l'expert pour retrouver les métatermes reliés (la même étude a été réalisée pour l'assignation automatique de métatermes pour la CIM10 - Voir Annexes). De la même facon nous n'avons pris en compte que les deux métatermes les plus fréquents et les associations de métatermes.

5.8.2 Évaluation

La comparaison de ces deux transcodages (ou «indexations») a consisté à calculer la précision et le rappel. Le transcodage manuel a été considéré comme la référence. D'une part nous avons réalisé cette évaluation en ne prenant en compte que les mots clés MeSH (voir figure 5.8). La similarité sémantique (voir section 2.5.2) a été intégrée dans le calcul de la précision et du rappel afin de définir la proximité des deux transcodages.

31. La raison du choix ce cette méthode est qu'elle seule était implémentée au moment de l'étude.

D'autre part, nous avons réalisé l'évaluation en ne prenant en compte que les

Catégories	Performances Précision (%) – Rappel (%)	
Anatomie	58	13
Techniques	40	19
Autres	10	5

FIGURE 5.8 – Résultats de la comparaison entre le trancodage effectué par l'expert et celui produit par F-MTI

métatermes (voir figure 5.9). La hiérarchie des métatermes n'étant pas très développée nous avons décidé de ne pas utiliser la mesure de similarité sémantique ici.

Pour la CCAM (7 389 codes) :

Manuellement	Automatiquement 1	Automatiquement 2
▪ 8 698 paires libéllés CCAM/métaterme	▪ 13 946	▪ 15 400
▪ 0 à 4 métatermes par code CCAM	▪ 0 à 11	▪ 1 à 10
▪ Moyenne de 1,18 métatermes pour chaque libéllé	▪ 1,89	▪ 2,08
▪ Pour 126 libéllés aucun métaterme n'a été associé	▪ 1 150	▪ 0

Précision	21%	29%
Rappel	28%	38%

FIGURE 5.9 – Résultats de la comparaison entre le trancodage effectué par l'expert et celui produit par F-MTI

5.8.3 Discussion

Les objectifs de ce travail étaient d'étudier la possibilité de générer un transcodage automatique entres deux terminologies. Cette étude a montré qu'il était difficile de produire un transcodage de manière manuelle ou automatique entres deux terminologies dédiées à des tâches différentes. Cette difficulté est due à une faible adéquation sémantique entre la terminologie CCAM et le MeSH, et au fait que l'algorithme du sac de mot ait été développé pour une indexation descriptive et non dans un but de classification d'actes techniques.

Les transcodages manuels et automatiques ont montré des différences. Les méthodes automatiques peuvent générer plus de termes que l'expert.

L'algorithme du sac de mots est une méthode purement lexicale et ne permet pas de déduire des éléments implicites alors que l'expert en est capable.

Au niveau de l'assignation des métatermes, la méthode des transcodages a donné les meilleurs résultats avec des taux de précision et de rappel de l'ordre de 50% et 60% pour la CIM10 et de 30% et 40% pour la CCAM. La méthode du sac de mots

est purement lexicale et est, en pratique, la plus intéressante, car elle ne nécessite aucune indexation manuelle. En revanche, elle montre de moins bons résultats. Voici listées ci-dessous quelques constatations pouvant expliquer les résultats :

- L'expert a assigné des métatermes dans un objectif de recherche dans un dossier médical fondé sur la pratique médicale, alors que les méthodes automatiques se fondent sur les relations métaterme CISMeF – mots clés MeSH qui avaient été originellement utilisées dans un objectif de recherche documentaire dans CISMeF.
- Les métatermes utilisés sont proches des spécialités médicales dont les contours ne sont pas toujours très bien définis et dépendent de pratiques «locales». Une grande variabilité inter-expert dans l'assignation de ces métatermes est, dans ce cadre, tout à fait vraisemblable.
- Certains mots clés sont retrouvés dans plusieurs arborescences MeSH, liées sémantiquement à plusieurs métatermes. Certains de ces métatermes peuvent ne pas s'appliquer pour certains actes ou maladies très spécifiques.
- L'expert choisit parfois d'englober les différents concepts inclus dans les libellés dans un métaterme beaucoup plus général.
- Le transcodage CIM10/MeSH peut produire des termes MeSH plus précis ou plus globaux que ceux utilisés originellement dans les libellés CIM10.
- Seul 8,9% de la CIM10 est transcodable en MeSH, il n'est donc pas possible de générer automatiquement les métatermes associés à tous les termes de la CIM10 avec cette technique. Néanmoins, parmi les 1 000 codes CIM10 les plus codés au CHU de Rouen, 53,5% sont transcodables en MeSH et appartiennent à notre table, ces 1000 codes couvrent 82% des comptes rendus d'hospitalisation du CHU de Rouen.
- Le choix de ne prendre que les deux métatermes les plus fréquents pour les assignations automatiques peut également être une explication. Certains métatermes ne sont pas pris en compte parce que les termes MeSH auxquels ils sont rattachés étaient lexicalement moins présents dans le libellé ou au niveau des liens entre les mots clés MeSH et les métatermes. La fréquence n'est peut-être pas le bon critère de sélection des métatermes, une pondération des métatermes ou des mots clés pourraient être plus performante.

Dans notre évaluation, certains termes considérés comme faux, parce qu'ils ont été reconnus automatiquement mais oubliés dans l'indexation manuelle, pourraient être rajoutés à l'indexation manuelle. Il est envisagé de procéder, dans une future étude, à une validation secondaire qui marquerait ce type de métaterme. Nous pourrions ensuite dans une deuxième série de comparaisons entre les assignations manuelles et automatiques ajouter ces métatermes à l'indexation manuelle.

Dans une future étude, nous pourrions également étudier la répartition des résultats par métatermes ou appliquer l'algorithme du sac de mot sur les libellés de la CIM10, ce qui donnerait peut être de meilleurs résultats puisque l'adéquation terminologique entre la CIM10 et le MeSH est plus grande que celle entre le MeSH et la CCAM, le MeSH ayant été créé à la base à partir de la CIM.

Une autre tentative de transcodage automatique a été réalisée chez Vidal entre

une terminologie icônographique VCM [Lamy06] et les termes du TUV, mais ceci n'a pas donné de bons résultats car les libellés VCM contiennent des notions très générales.

Notre méthode permet de d'obtenir un transcodage unidirectionnel les termes de la terminologie indexée étant le point de départ. Plusieurs études ont montré que l'on pouvait utiliser un outil d'indexation automatique pour déterminer des transcodages [Min06]. Il existe des méthodes lexicales et sémantiques utilisant le réseau sémantique de l'UMLS [Fung05].

5.9 F-MTI multilingue

F-MTI pourrait aisément être appliqué à d'autre langues sous réserve de disposer :
- d'une terminologie traduite dans cette langue (terminologie qu'il faudra intégrer à la base de données multi-terminologique)
- d'une liste de mots vides de la langue
- d'un outil de désuffixation dans la langue désirée

Un exemple de langage possible est l'anglais avec l'intégration du MeSH anglais, de nombreuses listes de mots vides ont déjà été développées par d'autres équipes et l'algorithme de Porter permet une bonne désuffixation.

5.10 Conclusion

Nous avons proposé plusieurs applications possibles de notre outil F-MTI. F-MTI sera intégré pour réaliser les tâches d'indexation au sein des trois équipes LERTIM, Vidal et CISMeF. Il sera aussi utilisé dans plusieurs projets (Interstis, PSIP). D'autres applications ont été envisagées comme l'aide à l'indexation semi-automatique généraliste, l'indexation multilingue, la structuration du dossier patient, et le transcodage automatique.

Chapitre 6

Discussion

Nous résumons ici les principaux résultats obtenus et évoquons les différentes perspectives.

6.1 Discussion générale des résultats obtenus

L'outil F-MTI a été évalué sur différents axes.

Nous avons montré les performances de notre outil dans la réalisation de trois tâches d'indexation :

- indexation des sites Web en MeSH
- indexation des dossiers médicaux en CIM10, CCAM et SNOMED
- indexation des RCP en TUV

Un des résultats les plus importants de cette thèse a été d'objectiver la différence des résultats d'évaluations selon : (a) la tâche d'indexation, (b) la terminologie, (c) le corpus, (d) le type de document au sein du corpus (e) les rubriques au sein du document.

Les résultats sont différents selon la tâche d'indexation considérée allant d'une précision de 3.4% et d'un rappel de 29.7% pour l'indexation des comptes rendus en CIM10 à une précision de 57.6% et un rappel de 43.4% pour l'indexation des RCP en TUV.

Nous avons pu montrer que les résultats étaient aussi différents selon la terminologie d'indexation considérée. Pour l'indexation des comptes rendus médicaux, l'algorithme du sac de mots a obtenu une précision de 3.4% et un rappel de 29.7% pour la CIM10 alors que pour la CCAM, il n'a pas été capable de produire d'indexation pertinente.

Les résultats dépendent du type de document formant le corpus. Dans notre travail, nous avions évalué des corpus comprenant des ressources Internet, des comptes rendus d'hospitalisation et des RCP. Les comptes rendus ayant été les plus difficiles à indexer. De plus, des différences existent dans un même corpus pour des types de documents différents. Dans l'étude sur le thésaurus MeSH et le corpus CISMeF, les résultats ont été très différents selon le type de ressources étudié, passant d'une précision de 44.4% et un rappel de 25.7% pour les ressources pédagogiques à une

précision de 39.9% et un rappel de 18.7% pour les recommandations. Ils sont aussi différents pour différentes rubriques d'un même document. Pour l'indexation des RCP en TUV, nous avions une précision de 28.4% et un rappel de 49.3% pour les précautions d'emploi et une précision de 77.0% et un rappel de 59.4% pour les effets secondaires.

Enfin les résultats dépendent de l'objectif visé. Pour l'indexation des comptes rendus d'hospitalisation les résultats sont différents selon que l'on considère une indexation médico-économique ou bien descriptive des comptes rendus en CIM10.

6.2 D'où l'importance de...

Ces résultats montrent l'importance de disposer de terminologies adaptées à la tâche d'indexation automatique visée. Les libellés doivent être clairs, sans ambiguité et représentatifs du contenu des documents à indexer. La terminologie doit également faire état de l'ensemble des variantes pouvant être rencontrées. Toutes les règles d'indexation doivent être explicitées selon la tâche à effectuer.

La rédaction des documents doit aussi être précise et comporter un minimum de formulations ambiguës ou complexes. Comme le montrent certains corpus statistiquement élaborés pour l'évaluation de méthodes d'indexation (the Medical NLP Challenge [1]), les résultats peuvent être très impressionnants (proches de 90% de F-mesure) lorsque les documents sont bien rédigés.

Mais tout cela ne suffit pas, pour une indexation automatique de qualité, l'outil doit être capable de prendre en compte le contexte, les éléments implicites et de «raisonner» sur des connaissances médicales.

Enfin, il lui faut encore être capable de synthétiser les informations recueillies et reconnaitre ce qui est important de ce qui ne l'est pas.

Tout cela laisse à penser qu'une bonne indexation entièrement automatique est un objectif difficilement atteignable [Wehrli88]. Je pense pour ma part qu'avec les efforts de chaque acteur, nous pouvons tendre à atteindre cet objectif :

- terminologues pour l'amélioration des terminologies et le développement de règles d'indexation propre à la terminologie
- indexeurs pour la formation à l'indexation, l'apprentissage des terminologies utilisées, et le développement de règles d'indexation pour les tâches visées
- auteurs de documents destinés à être indexés pour la formation à la rédaction
- informaticiens pour le développement d'outils d'indexation automatique plus performants

Pour une bonne évaluation de ce genre d'outil, il est nécessaire de disposer d'une indexation manuelle de référence de qualité ce qui n'est, pour l'instant, pas le cas. En effet, disposer de corpus assez importants de documents indexés avec la même version d'une terminologie et selon les mêmes règles reste très difficile. Ajouter à cela des documents de qualité associés à une indexation manuelle issue d'un consensus de plusieurs individus experts est mission impossible. Comme le dit Lancaster, le

1. Voir `http://www.computationalmedicine.org/challenge`

problème concernant l'évaluation d'une indexation est qu'il n'existe pas de référence universelle [Lancaster91]. Une évaluation manuelle de l'indexation par rapport à un objectif visé par plusieurs experts est bien plus juste mais est très chronophage.

6.3 Différentes méthodes

Au cours de cette thèse, nous avons développé trois méthodes : la méthode de l'algorithme du sac de mots, le dictionnaire de termes et le dictionnaire de constituants. Deux de ces méthodes ont été évaluées, la troisième étant dans l'état actuel très proche en terme de résultats à ceux de l'algorithme du sac de mots.

L'utilisation préférentielle de la lemmatisation ou de la désuffixation n'a pas été démontrée, ce choix dépendant de l'objectif à atteindre.

Concernant l'apport d'une approche multi-terminologique par rapport à une approche mono-terminologique, les résultats ne sont pas tranchés. Le rappel est meilleur pour une approche multi-terminologique mais la précision en est impactée. Les causes principales sont les transcodages et la difficulté de déterminer parmi tous ces codes ceux qui sont plus importants. Cela étant, nous pensons que cette approche est bien plus intéressante du fait de la quantité plus importante d'informations pouvant être prise en compte pour l'indexation de documents.

6.4 Comparaison à d'autres outils

À notre connaissance, F-MTI est le premier outil multi-terminologique pour le francais. Il constitue une avancée comparé à d'autres outils :

- Il est le seul outil pour le francais à réaliser une indexation directe TAL en CIM10.
- Il constitue une toute première tentative d'indexation automatique pour la CCAM.
- Il est le second outil pour l'indexation en SNOMED 3.5 après SnoCode (un outil commercial).
- Il est le seul outil à intégrer la terminologie TUV.
- Il est le seul outil à s'intéresser à l'indexation automatique des RCP.

La comparaison à d'autres outils a été discutée. F-MTI comparé aux outils SnoCode et MAIF donne des résultats satisfaisants.

Par rapport à d'autres outils en francais comme CIREA ou MEDCKARE, il apporte une réelle plus value en permettant une indexation descriptive sur l'ensemble de la CIM10.

L'outil le plus approchant pour l'anglais, MTI, prend en compte un plus grand nombre de terminologies (plus de 100 issues de l'UMLS alors qu'il n'en existe que 10 disponibles pour le francais) et comprend des méthodes à la fois statistiques et TAL.

En matière de performance MTI traite 4 000 articles (titre + résumé) chaque nuit. À l'heure actuelle, F-MTI permet de traiter 2 000 comptes rendus d'hopitalisation en 1 heure (sur un serveur 4 cœurs) ce qui laisse entrevoir d'autres applications

industrielles.

Tout comme ces outils, F-MTI va être intégré dans un logiciel d'aide à l'indexation.

6.5 Perspectives

6.5.1 Amélioration de l'outil

Certaines améliorations sont d'ores et déjà envisagées : amélioration des transcodages, meilleure aggrégation des proposition d'indexation de nos différentes méthodes, insérer les constituants de poids supérieur à 1, implémentation de transducteurs pour les termes compliqués, créer des règles médicales (ajout des relations SNOMED CT), combinaison de termes SNOMED, ajouter les rôles des termes, élargissement des notions de contexte implémentées, traitement des ambiguïtés, analyse sémantique, présentation des informations (résumés), associations d'idées provenant de différentes localisations dans le compte rendu, calcul de scores.

A l'occasion de cette thèse, les collaborations de l'équipe CISMeF avec la NLM (et le centre de recherche du Lister Hill [2] en particulier), créateur de MTI, ont perduré. Elles vont s'intensifier ces prochaines années puisque nous envisageons d'implémenter les méthodes de MetaMap [3] pour le francais pour optimiser les résultats de F-MTI. Ces travaux se dérouleront dans le cadre d'une autre thèse.

6.5.2 Poursuite des travaux

Les travaux doivent être poursuivis, d'autres évaluations sont nécessaires comme la comparaison de nos méthodes d'indexation et l'évaluation des performances lorsque plusieurs méthodes sont combinées. Ceci pourra se faire avec les corpus déjà constitués et en considérant l'indexation d'une ou de plusieurs terminologies.

6.5.3 Ouverture importante pour les différentes équipes

6.5.3.1 Un CISMeF multi-terminologique

Cette thèse a ouvert une véritable révolution stratégique au sein de l'équipe CISMeF avec le passage d'une stratégie mono-terminologique à une stratégie multi-terminologique (L'organisation des projets passe de la figure 1.4 à 6.1). Dès à présent, trois autres thèses, dans la continuité de celle-ci, ont débuté pour explorer cette nouvelle voie de recherche :

- Travaux sur l'interopérabilité sémantique inter et intra-terminologies (T. Merabti). Ces travaux visent à développer des méthodes pour améliorer et étendre les transcodages existants. Ces travaux ont pour l'instant permis de transposer

2. Grâce à A. Névéol, ancienne doctorante de l'équipe CISMeF et postdoctorante depuis 2 ans et demi au Lister Hill.

3. Outil d'extraction de termes inclus dans MTI.

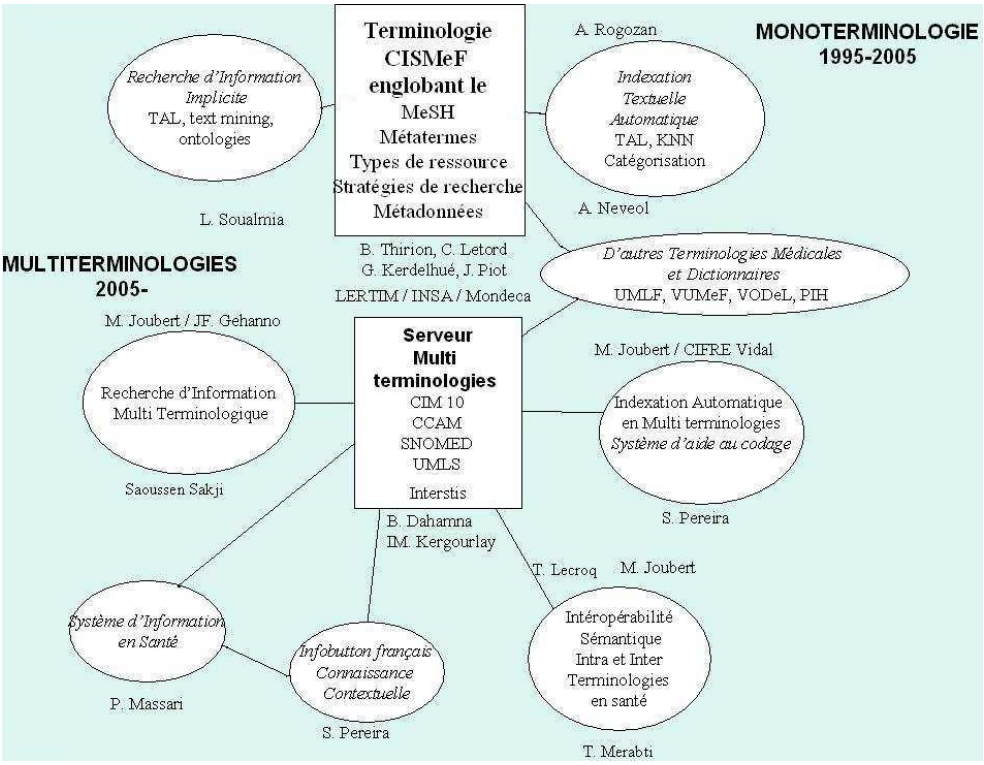

FIGURE 6.1 – Nouvelle organisation des projets de l'équipe CISMeF

les liens sémantiques de la SNOMED CT à la CIM10, à la SNOMED 3.5 et au MeSH [Merabti08a]. Une étude est en cours pour intégrer la CCAM à l'UMLS.
- Mise en place d'une recherche d'information multi-terminologique (S. Sakji) s'appuyant sur notre indexation multi-terminologique
- Poursuite des travaux sur la recherche d'information multi-terminologique pour le dossier patient électronique (A. Diouf)

Cette thèse a aussi été l'occasion d'une collaboration poussée avec le Dr P. Massari qui rejoint l'équipe pour continuer de développer des applications pour les systèmes d'information hospitaliers et de santé.

6.5.3.2 Une aide à l'indexation et des perspectives de logiciels hospitaliers pour Vidal

Les résultats sont encourageants pour l'indexation des RCP en TUV. F-MTI va également intégrer un logiciel d'aide à l'indexation semi-automatique, BIBLIS. Cet outil sera utilisé par tous les indexeurs de l'équipe données thérapeutiques du Vidal.

Cette thèse ouvre pour le Vidal des perspectives en matière d'exploitation d'autres terminologies médicales pour des alertes toujours plus performantes. La collaboration entre données du dossier médical électronique et logiciels d'aide à la prescription va pouvoir être étendue.

6.5.3.3 Vers un dossier patient plus structuré et une aide au codage pour les médecins - LERTIM

Cette thèse a permis de faire un nouveau pas vers l'élaboration de systèmes d'information hospitaliers performants (adaptés et évolutifs) et notamment pour la création d'un Dossier Médical Personnel (DMP). Les thèses de S. Sakji, T. Merabti et A. Diouf en cotutelle avec le laboratoire LERTIM permettront de poursuivre cet axe de recherche.

Une meilleure structuration des dossiers patients électroniques avec une indexation descriptive ouvre des perspectives dans des voies de recherche connues comme la création automatique de synthèses médicales, de résumés automatiques, l'aide au codage médico-économique et d'autres moins connues comme la rédaction assistée de documents.

6.5.4 Vers d'autres projets communs

Les collaborations entre la société Vidal et les équipes LERTIM et CISMeF continuent, trois projets ont déjà débuté InterStis, PSIP et Aladin :

- Le projet Interstis (Interopérabilité Sémantique des Terminologies dans les Systèmes d'Information de Santé Francais voir section 5.6), débuté en 2007, va permettre le développement d'un Serveur Multi-Terminologique en Santé (SMTS) (avec S.Sakji), pendant de notre outil F-MTI pour l'extraction automatique. Toutes les terminologies de santé incluses dans le SMTS seront intégrées dans l'outil F-MTI. Les terminologies suivantes sont en cours d'intégration : DRC, CISP2. En 2009 sont programmées, après leurs intégration préalable dans le SMTS, l'ajout des terminologies suivantes au sein de F-MTI : LOINC, MedDRA et Who-Art.
- Le projet PSIP (Patient Safety Through intelligent Procedures in medication voir section 5.7), débuté en 2008, a pour objectif l'optimisation de la prescription informatisée. Dès à présent, dans le cadre du projet européen PSIP, les noms commerciaux et internationaux des médicaments sont intégrés par S. Sakji au F-MTI version 2.
- L'outil F-MTI version 2 sera également réutilisé et adapté à une nouvelle problématique, les infections nosocomiales, dans le projet ALADIN-DTH (Assistant de Lutte Automatisé et de Détection des Infections Nosocomiales à partir de Documents Textuels Hospitaliers - ANR TecSan 2008)

Dans PSIP et dans Aladin, l'outil développé dans ma thèse fera l'objet d'améliorations en terme de couverture terminologique et technologique.

Il est probable que je continue à travailler sur F-MTI dans le cadre de ces trois projets.

Troisième partie

Contribution à l'accès aux connaissances

Chapitre 7

Conception d'outils et mise au point de méthodes pour l'accès aux connaissances

7.1 Introduction

Après nous être intéressé à l'indexation, nous présentons notre contribution en matière d'accès aux connaissances. Nous avons vu que les professionnels de santé, les patients et les étudiants avaient besoin dans leur quotidien d'informations de santé, que ce soit dans le cadre de l'apprentissage de nouvelles connaissances, d'aide à la décision ou de suivi de son état de santé pour les patients (voir section 2.2.4).

L'accès à ces informations n'est pas toujours aisé, or pour chacun et plus particulièrement le médecin, les informations doivent être rapidement consultables. En effet, les informations sur Internet ne sont pas toujours référencées et lorsqu'elles le sont, elles sont contenues dans de nombreuses bases de connaissances. En outre, il n'est pas toujours aisé de trouver une information compréhensible par l'utilisateur (langue, formulation).

L'objectif ici est d'aider tout acteur dans sa recherche d'information dans le domaine de la santé en offrant des accès simplifiés afin qu'il accède à la bonne information, au bon moment.

Access to the right information, at the right time for the right person.

La prise en compte du contexte rend cela possible. Nous proposons donc plusieurs méthodes et leurs applications afin de proposer des accès contextuels prenant en compte la demande, le profil et la langue de l'utilisateur ainsi que l'existence du contenu recherché. Nous présentons trois types d'accès contextuel liant différents types de données :

- à partir du dossier patient vers les banques d'information en ligne multilingues
- au sein du dossier patient
- à partir d'une banque d'information en ligne vers d'autres banques d'information en ligne en francais ou en d'autres langues

7.2 Accès contextuel à la connaissance à partir du dossier patient

7.2.1 Accès aux connaissances à partir du dossier patient

Autrefois, seuls les médecins et les étudiants en médecine avaient le droit de consulter les dossiers de leurs patients. Ce n'est plus le cas aujourd'hui puisque la loi [1] permet aux patients d'accéder à leurs dossiers médicaux et donc aux comptes rendus ainsi qu'au codage de leurs données. Cet accès est dédié à la personne concernée ou son représentant légal, un intermédiaire, ou les ayants-droit après un décès. Le patient peut être seul ou accompagné dans sa consultation. Une première phase d'expérimentation du DMP (Dossier Médical Personnel) en janvier 2007 a montré que les patients étaient intéréssés par cet accès puisque sur 1 330 patients, 10% se sont connectés à leur dossier consultant essentiellement les données générales (23% des documents consultés), les synthèses (19%) et les comptes rendus de consultation (11%) [GIP-DMP07].

Le contenu des dossiers médicaux est complexe, cette ouverture à un large public pose de nombreux problèmes. Chaque acteur a des besoins spécifiques (voir section 2.2.4), une bonne compréhension des informations contenues dans le dossier du patient nécessite des connaissances médicales pointues ce qui n'est pas forcément le cas pour les étudiants ou les patients [Keselman07] [Zeng-Treitler07]. Malheureusement la plupart de ces demandes restent sans réponse [Covell85] [Ely05]. Il y a donc un besoin important d'informations auquel le dossier médical ne répond pas aujourd'hui. Un des moyens de se documenter est de poser des questions sur sa pathologie à son médecin ou un collègue médecin, ce qui demande d'y consacrer du temps, de se déplacer voire même représenter un certain coût. Un autre moyen est de consulter les documentations existantes (livres) voire, ce qui est aujourd'hui très répandu, chercher une information médicale sur Internet. Comme nous l'avons vu dans la section 2.2.2, une quantité importante d'informations existent sur Internet pour répondre aux besoins. En revanche, le temps nécessaire à une recherche peut s'avérer long, de plus trouver une information de qualité sur Internet est compliqué et demande aux médecins de travailler sur plusieurs supports (leur logiciel de dossier patient et un navigateur Internet).

Nous proposons ici un accès facilité aux connaissances, en évitant les recherches fastidieuse sur Internet, en proposant des ressources adaptées aux différents besoins, et en évitant la multiplication des supports. Cet accès, inspiré de l'InfoButton de Cimino [Cimino97], est contextuel et se fait directement à partir du dossier patient vers des bases de connaissances de qualité sur l'Internet.

1. la loi N°2003-303 du 4 mars 2002 relative aux droits des malades et à la qualité du système de santé, et le décret N°2002-637 du 29 avril 2002 apportent une réforme importante au sein de l'arsenal législatif.

7.2.2 Accès contextuel

Notre projet a été inspiré par le «Knowledge coupling» [Cimino97] c'est-à-dire que des informations spécifiques issues du dossier patient sont couplés avec des connaissances médicales spécifiques pour donner une connaissance adaptée «au bon moment, à la bonne personne». Cette connaissance prend en compte un double contexte : le contexte du patient (diagnostics, actes médicaux) et le type d'utilisateur (médecin, étudiant, patient).

Les connaissances sont recherchées sur l'Internet, sur des sites spécialisés dans la recherche en santé, 50 sites Web provenant des gouvernements de pays francophones, d'organisation de santé nationale, des facultés de médecine et d'odontologie qui ont été définis par l'équipe CISMeF comme étant de qualité. Les ressources sont filtrées selon le profil de l'utilisateur : recommandations pour les professionnels de santé, ressources pédagogiques pour les étudiants et documents spécifiques pour les patients. L'utilisateur peut également choisir le type de connaissances qu'il recherche. Par exemple, le médecin ayant un rôle fondamental d'infomédiation [2] il voudra rechercher des informations sur un diagnostic spécifique pour un patient qui lui en aurait fait la demande.

Nous avons développé un outil permettant d'accéder à des connaissances médicales contextualisées (3 dimensions : le profil de l'utilisateur, le diagnostic ou l'acte, l'existence de ressources) et potentiellement applicable à n'importe quel logiciel de dossier patient.

7.2.3 Développement

L'outil mis au point s'inspire du bouton d'information (InfoButton) imaginé par Cimino en 1997 [Cimino97]. Ce bouton, intégré dans les systèmes cliniques, permettait aux utilisateurs, en un seul clic, d'interroger les ressources d'informations en ligne en utilisant les données du patient. Pour accéder aux ressources appropriées, l'utilisation de la terminologie Medical Entities Dictionary (MED) traduisait les données du patient concernées par la demande de l'utilisateur en une forme reconnue par les ressources. L'infoButton est un outil de recherche d'information qui prévoit à l'avance les questions qu'un utilisateur peut se poser ainsi que les ressources d'information sur Internet dont il peut avoir besoin dans un contexte particulier. En pratique, l'infoButton doit mener l'utilisateur le plus près possible de la réponse à sa question grâce à un minimum d'interaction entre l'utilisateur et l'ordinateur [Del Fiol06].

Nous avons donc créé deux boutons d'information contextuels et personnalisés, destinés à anticiper les besoins d'information des utilisateurs, dans les fiches des codages du séjour du patient à l'hôpital, contenant les diagnostics et actes médicaux et dans la fiche de synthèse. La fiche de synthèse regroupe toutes les informations issues de tous les séjours effectués par le patient à l'hôpital. Ces fiches ont été jugées par un médecin expert (P. Massari [3]) ainsi que dans la littérature [GIP-DMP07] comme des endroits stratégiques de consultation et de possibles besoins d'information.

2. Le médecin joue le rôle d'intermédiaire informateur entre le monde médical et le patient
3. Médecin intégré à l'équipe CISMeF

Le premier bouton crée un accès direct vers le site CISMeF (voir section 1.3.1). Les requêtes adressées au site sont personnalisées et contextuelles. La requête regroupe deux informations majeures : le diagnostic codé en CIM10 ou l'acte codé en CCAM (nécessite un transcodage CIM10→ MeSH et CCAM→ MeSH) pour lesquels des informations supplémentaires sont recherchées. Le type de l'utilisateur est connu grâce à son login. Ainsi, le médecin sera dirigé vers des ressources de type recommandations, l'étudiant en médecine vers des ressources pédagogiques et les patients vers des ressources spécifiques. Une liste de documents appropriés est ainsi fournie par CISMeF à partir de la liste existante des codes CIM 10 et codes CCAM présents dans la fiche des codages du compte rendu d'hospitalisation et dans la fiche de synthèse du dossier patient.

Le deuxième bouton crée un accès vers d'autres sites spécialisés dans la recherche en santé. Ceux-ci sont catégorisés selon le type de connaissances recherchées : recommandations, matériel pédagogique, spécifique patient, bases de données bibliographiques, santé publique, essais cliniques, maladies rares, outils de recherche en santé et outils de recherche généralistes. Ils sont aussi classés selon la langue : sites francais et anglais (voir figure 12 - Annexes). Les différents sites et bases de connaissances accessibles en ligne sont : CISMeF [4], National guidelines clearinghouse [5] (NGC), Medline / PubMed [6], MedlinePlus [7], NLMGateway [8], BDSP [9], Clinical trials [10], Orphanet [11], Hon [12], Intute [13], HealthInSite [14], Google [15].

Pour accéder aux ressources appropriées avec ces deux boutons d'information,

4. Accessible ici `http://www.chu-rouen.fr/cismef/`. CISMeF global, CISMeF patient, CISMeF recommandations et CISMeF pédagogie.

5. Base de données recensant les recommandations de langue anglaises à destination des professionnels de santé. Accessible ici `http://www.guideline.gov/`

6. Base de données bibliographiques. Accessible ici `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi`

7. Site en anglais, à destination des patients et du grand public, mis en place par la NLM et pointant sur des sites de qualité. Accessible ici `http://medlineplus.gov/`

8. Porte d'entrée permettant une recherche simultanée sur plusieurs bases de données de la NLM : PubMed, MEDLINEplus, HSBD...Accessible ici `http://gateway.nlm.nih.gov/gw/Cmd`

9. Banque de Données Santé Publique, résultat d'un réseau francais de coopération pour la mise en ligne de sources d'information en santé publique. Accessible ici `http://www.bdsp.tm.fr/`

10. ClinicalTrials.gov fournit des informations sur les essais cliniques chez l'homme. Accessible ici `http://clinicaltrials.gov/`

11. Orphanet est un serveur d'information en libre accès pour tout public sur les maladies rares et les médicaments orphelins. Accessible ici `http://www.orpha.net/`

12. Fondation Health On the Net (La Santé sur Internet) est une fondation dont l'objectif est de promouvoir le développement et les applications de nouvelles technologies d'information notamment dans les domaines de la médecine et de la santé. Accessible ici `http://www.hon.ch/`

13. Intute est un portail de ressources de qualité en santé, pour les étudiants et professionnels de santé. Accessible ici `http://omni.ac.uk/`

14. HealthInSite est un portail de ressources de qualité en santé et concernant essentiellement le diabète, le cancer, l'asthme et la santé mentale. Accessible ici `http://www.healthinsite.gov.au/`

15. Google est le moteur de recherche sur Internet le plus utilisé dans le monde aujourd'hui. Accessible ici. Un partenariat avec CISMeF a permis de restreindre l'accès de Google à une liste de sites de qualité pour le domaine médical (`http://www.google.com.custom?hl=fr&lr=lang fr&client=google-coop-np&cof=AH`) et pour les médicaments (`http://www.google.com/coop/cse?cx=015430007758165987576\%3Ab3cmgan4uas&hl=fr`).

il est nécessaire de traduire la requête de l'utilisateur c'est-à-dire traduire les diagnostics codés en CIM10 et les actes codés en CCAM en une forme compatible avec l'indexation des ressources. L'indexation des ressources, pour tous ces sites, utilise la terminologie MeSH (voir section 2.4.1.1) (outre pour leur contenu de qualité, c'est la raison pour laquelle nous les avons sélectionnés). Pour ce faire, nous avons utilisé le transcodage CCAM MeSH (voir section 5.8.1) et CIM10 MeSH extrait du Metathesaurus de l'UMLS (version 2004AC voir section 2.3.2).

La table ainsi obtenue (voir figure 7.1) contient plusieurs termes MeSH possibles pour un même code CIM10 : terme préféré, synonymes et terme correspondant à une plage CIM10 (ex : A15-A19.9). Nous avons décidé de ne pas considérer les sy-

CODE CIM10	TERME MeSH	NBRECOMMANDATION	NBENSEIGNEMENT	NBPATIENT
A00	Choléra	7	7	1
A00.9	Choléra	7	7	1
A15-A19.9	tuberculose	40	29	3
D70	Agranulocytose	3	4	2
F99-F99.9	Troubles mentaux	168	182	174
F99	Troubles mentaux	168	182	174

FIGURE 7.1 – Extrait de la table de transcodage CIM10/MeSH intégré au DEP

nonymes, seulement les termes préférés (les synonymes sont explorés au moment de la requête sur les sites interrogés). Si le code CIM10 est transcodable en plusieurs termes MeSH (un terme préféré plus un terme MeSH regroupant une plage de codes CIM10) le terme préféré est choisi en priorité. La table finale contient 1 629 transcodages CIM10→ MeSH, ce qui est peu par rapport aux 18 000 codes CIM10 existants (environ 10%).

Les deux boutons sont présentés à côté de chaque code CIM10 et CCAM qui ont été renseignés par les médecins. Nous avons appelé le premier bouton, le bouton «CISMeF» et le deuxième, le bouton «plus d'infos». Les deux boutons d'information contextuels et personnalisés ne sont visibles pour l'utilisateur qu'à côté des termes CIM10 et CCAM pour lesquels il existe une connaissance adaptée dans CISMeF ou sur un des sites de la page Web. Pour contrôler cela, nous avons ajouté des colonnes dans la table de transcodage qui indiquent pour chaque terme MeSH issu du transcodage CIM10 et CCAM le nombre de ressources spécifiques pour les étudiants, les patients et le nombre de recommandations dans CISMeF. Le principe sera le même pour les catégories des sites présents sur la page Web.

La contextualisation appliquée est formée de 4 dimensions (voir figure 7.2) :

- l'apparition des boutons se fait seulement après vérification du statut de l'utilisateur et n'est disponible que pour les patients, médecins, et étudiants
- le diagnostic demandé doit aussi être présent et sous la bonne forme
- le terme CIM10 ou CCAM doit être transcodable en MeSH
- des documents appropriés pour l'utilisateur doivent être disponibles sur CISMeF pour le premier bouton et sur au moins un des sites de la page Web pour le deuxième

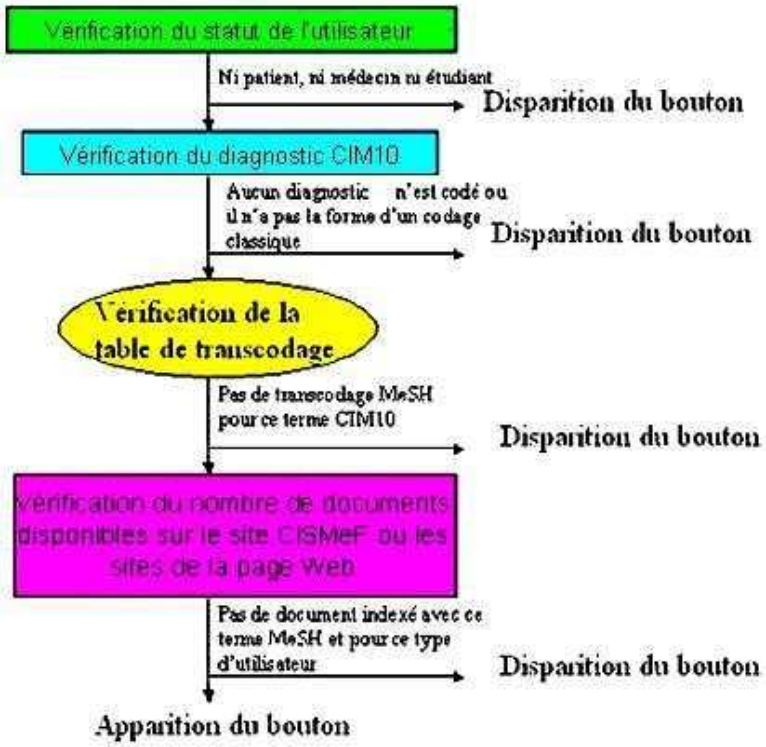

FIGURE 7.2 – Traitements réalisés pour déterminer l'apparition des deux boutons

S'il est présent, comme décrit précédemment, l'utilisateur peut alors cliquer sur le bouton contextuel associé à un diagnostic (code CIM10) ou un acte (code CCAM) décrit dans le dossier patient, pour obtenir des informations sur celui-ci. Pour le premier bouton, la page de CISMeF correspondant à la requête apparaît alors. Cette requête est le fruit de l'association du statut et du terme MeSH à partir de la table de transcodage sous la forme d'une URL adaptée (voir figure 7.3).

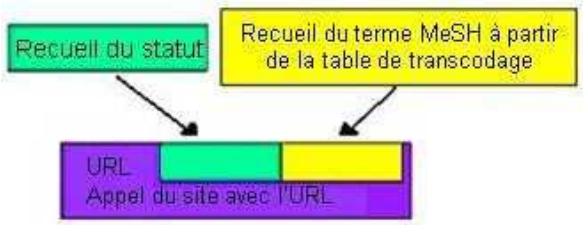

FIGURE 7.3 – Traitements réalisés après avoir cliqué sur le bouton CISMeF ou l'un des sites de la page Web

Pour le deuxième bouton, la page Web des autres sites apparaît seulement pour les sites où des ressources adaptées sont disponibles. L'utilisateur n'a plus qu'à choisir la catégorie qui l'intéresse (des documents pour le patient, des recommandations etc.) , la langue qui lui convient (anglais / francais) et le site qu'il préfère. Chaque site a son propre moyen d'interrogation que l'utilisateur ne maîtrise pas

forcément, certains permet l'utilisation de booléens (OR, NOT etc...) d'autres non. L'expertise de l'équipe CISMeF a permis d'élaborer pour la cinquantaine de sites un modèle de requêtes approprié pour chacun (exemple : une requête d'un utilisateur interprétée par le moteur de recherche CISMeF comme équivalente au terme MeSH «asthme/prévention et contrôle» sera transformée en «asthma/PC[MeSH Terms] OR ((("asthma, bronchial"[Tiab] OR "asthmas"[Tiab] OR "asthmas, bronchial"[Tiab] OR "bronchial asthma"[Tiab] OR "bronchial asthmas"[Tiab]) AND ("PC"[Tiab])) NOT MEDLINE[SB])» si l'utilisateur approfondit sa recherche en cliquant sur le site Pubmed).

Pour une démonstration, vous pouvez consulter l'Annexe Démonstration.

7.2.4 Valorisation industrielle

Nos boutons d'information contextuels ont été valorisés [16] à l'université de Rouen puis commercialisés par la société privée IS@S [17] [Darmoni08]. Un bouton d'information spécifique aux professionnels de santé en secteur privé est en cours de développement. En février 2008, les boutons d'information ont été présentés à des médecins du secteur privé ainsi qu'à des petits hospitaux privés (n<100 lits). Un groupe de cliniques privées teste le produit que nous avons appelé «French Info Button». Plusieurs industriels dans le secteur des systèmes d'information de santé ont récemment visité l'hôpital de Rouen afin de tester les boutons contextuels en environnement réel.

Les tables de transcodage devront être mises à jour à chaque nouvelle version des terminologies impliquées.

7.2.5 Perspectives

A plus long terme, nous voudrions appliquer le même principe de connaissance contextuelle à partir d'un compte rendu textuel.

Les boutons seront alors accessibles sur la barre d'outils du logiciel permettant la rédaction et la lecture du compte-rendu d'hospitalisation (voir figure 7.4 avec l'apparition du bouton de recherche d'information dans la barre d'outil du logiciel Microsoft Word). Ce bouton donnera l'accès à l'indexation CIM10 et CCAM produite par F-MTI et pour chaque terme, l'accès aux connaissances contextuelles correspondantes sur Internet.

Un profil plus élaboré pourrait aussi permettre de renseigner d'autres caractéristiques comme le secteur d'activité du médecin, ou pourrait permettre de renseigner plusieurs profils pour permettre au médecin de rechercher de l'information pour lui-même ou pour transmettre à son patient.

16. Ils ont fait l'objet d'un brevet universitaire

17. Très Petite Entreprise innovante travaillant dans l'ingénierie santé-sociale. Grâce à la loi Allegre de 1999, les 9 co-auteurs de ce projets (l'équipe CISMeF) ainsi que l'Université recevront des fonds par la companie IS@S. Le prix des boutons contextuels a été estimé à 5-10 € par lit d'hôpital.

Nous pourrions aussi imaginer une diffusion sélective et ciblée d'information avec l'avertissement de l'utilisateur de l'apparition de nouvelles informations sur un ou plusieurs sujets selon son profil (par courriel par exemple).

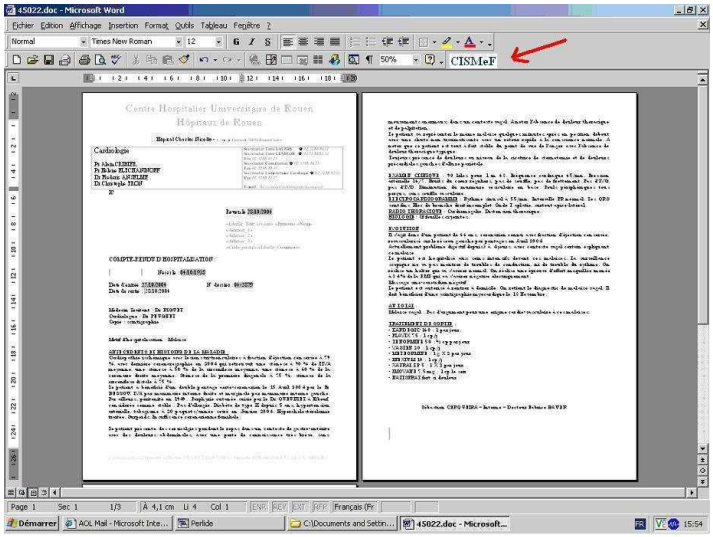

FIGURE 7.4 – Compte-rendu d'hospitalisation provenant du service de Cardiologie du CHU de Rouen avec le bouton CISMeF dans la barre d'outil

De nombreuses études ont montré que l'InfoButton fournit des réponses dans les unités de soins de manière satisfaisante, avec un temps satisfaisant et avec un haut niveau de satisfaction des utilisateurs [Cimino06] [Maviglia06]. Un exemple de succès de l'infobutton a montré une utilisation de plus de 80 000 fois pour 3 590 utilisateurs en 5 ans à l'Intermountain Healthcare [Del Fiol07].

Dans la littérature, des améliorations ont été apportées à l'Infobutton, avec l'utilisation de bases de connaissances liant les éléments du contexte avec des besoins d'information liés à des ressources. Ce qui, en pratique, permet de proposer à l'utilisateur des liens directs vers les ressources [Li07]. Une étude récente utilise des méthodes d'apprentissage automatique afin de prédire la ressource qui sera sélectionnée par un utilisateur dans un contexte particulier afin de ne présenter que les plus probables à l'utilisateur [Del Fiol07]. Le temps de recherche de l'utilisateur qui doit rechercher parmi plusieurs ressources possibles est ainsi réduit. Nous pourrons appliquer ces méthodes dans une prochaine version.

7.3 Recherche par spécialité médicale

Dans les dossiers médicaux électroniques, les informations du patient sont le plus souvent classées par date et par séjour ce qui ne facilite pas la recherche d'information par les professionnels de santé et les patients surtout face à un dossier important avec de nombreuses informations et de nombreux séjours. Pour améliorer cette recherche d'information, le dossier médical «orienté problème» a été introduit en 1963 [Weed68] mais il est encore peu appliqué (surtout en France [Falcoff99]) du fait de la

structuration particulière des données du patient qu'elle nécessite [Lundsgaarde81]. Cette structuration implique une saisie des données par les professionnels de santé plus complexe, ce qui entraîne, encore ici, un problème de temps.

Chaque séjour est lié à des codes CIM10 et éventuellement CCAM et à un ou plusieurs comptes rendus médicaux. Une solution serait d'implémenter des vues adaptées aux besoins de l'utilisateur en mettant en œuvre des outils terminologiques.

C'est ce qui a été réalisé par un clinicien, P. Massari et le chef des documentalistes de l'équipe CISMeF, B. Thirion, en appliquant les métatermes CISMeF [18] (voir section 5.8.1) aux terminologies du dossier patient francais.

Ces «super-concepts» ont été adaptés à la CIM10 et à plusieurs classifications d'actes médicaux : la CCAM [Rodrigues05] (utilisée depuis 2005), le CDAM (le Catalogue Des Actes Médicaux utilisés avant la CCAM) pour les actes thérapeutiques et diagnostics et l'ADICAP (l'Association pour le Développement de l'Informatique en Cytologie et Anatomo-Pathologie) pour les examens d'anatomie-pathologie. Sur 123 métatermes CISMeF, 66 ont été réutilisés ici (soit 54%). Les liens sémantiques ont été créés manuellement pour chaque super-concept (de 0 à n relations) vers les classifications CIM10, CCAM, CDAM et ADICAP (voir figure 7.5). Exemple, le métaterme

Classification	Nombre de codes	Nombre de liens sémantiques	Min-Max par code
ICD10	10 505	13 650	1-3
CCAM	7 389	12 538	1-5
CDAM	7 699	13 508	1-4
ADICAP	279	372	0-3

FIGURE 7.5 – Liens sémantiques entre les super-concepts et les différentes classifications

«cardiologie» est lié au code CIM10 I50.0 «Insuffisance cardiaque congestive», au code CCAM DZQM006 «Échographie-doppler transthoracique du coeur et des gros vaisseaux» et au code ADICAP BHCZ «Biopsie endomyocardique».

F-MTI a été appliqué afin de créer automatiquement ces liens (voir section 5.8 [Pereira07]).

L'utilisation de métatermes pour réaliser des requêtes sur des consultations cliniques nécessite l'indexation des différents services de consultation (exemple : l'unité d'échocardiographie a été liée sémantiquement à deux métatermes «cardiologie» et «imagerie diagnostique»).

L'implémentation de ces super-concepts permet à l'utilisateur de filtrer les données selon une ou plusieurs spécialités médicales créant ainsi une vue sur les données adaptée à son activité sur les données. La vue affiche uniquement les séjours, actes

18. On rappelle que les métatermes correspondent à des spécialités médicales (exemple : «cardiologie»), des types d'actes médicaux (exemple : «chirurgie») ou des sujets de santé (exemple : «diagnostic», «thérapie»). La liste est disponible *via* l'URL suivante `http://doccismef.chu-rouen.fr/liste des meta termes anglais.html`.

médicaux et diagnostics liés aux métatermes séléctionnés. Le cardiologue voudra ne consulter que les informations concernant son domaine, la Cardiologie, ou seulement les comptes rendus pour un acte particulier comme un acte de chirurgie pour son patient (voir figure 7.6). La vue lui permet ainsi de gagner un temps précieux sans avoir à connaître la date de l'acte passant ainsi de 5 à moins de 2 minutes de recherche.

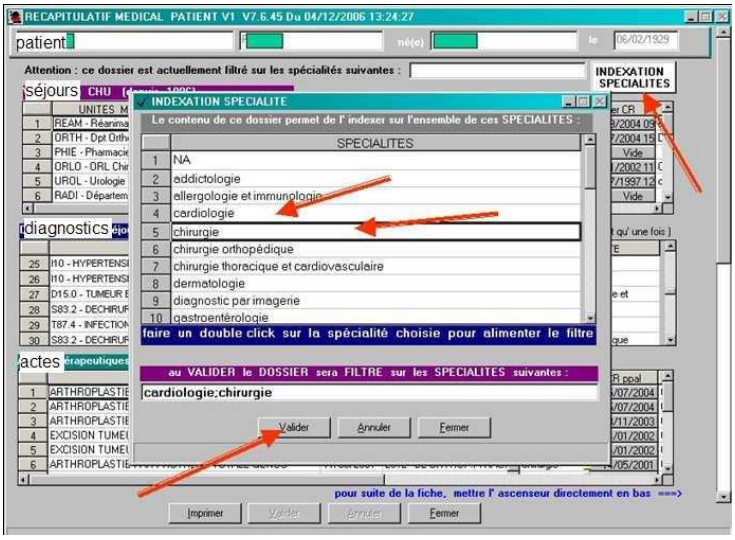

FIGURE 7.6 – Recherche par spécialité dans la fiche de synthèse d'un patient dans le logiciel CDP2

L'évaluation a été réalisée par des médecins, et plusieurs spécialistes (cardiologues, pneumologues, gastroentérologues) [Massari08]. La recherche classique orientée «chronologie» et la recherche orientée «spécialités» pour les comptes rendus du dossier patient électronique à Rouen, CDP2, ont été comparées. Un tiers de ces comptes rendus contient plus de 20 séjours et plus de 20 actes médicaux enregistrés. L'évaluation a montré des résultats considérés satisfaisants pour l'équipe CISMeF et les médecins rouennais même si une vision d'ensemble de l'état du patient est parfois nécessaire dans certains cas. C'est ainsi que cet outil de vue par «spécialité» a été intégré dans un environment de production dans le dossier patient électronique du CHU de Rouen en mai 2007. Cette vue est actuellement utilisée quotidiennement par les médecins avec des réactions positives [19].

L'efficacité des vues orientées a été observée par plusieurs auteurs [Doré95], [Zeng99]. Plus récemment, une deuxième génération de ce type d'outil utilise une ontologie pour définir la structure orientée «problème» du dossier patient ainsi que les concepts fondamentaux qui y sont rattachés [Elisabeth02]. D'autres outils utilisent une vision graphique des problèmes avec la représentation des épisodes liés à chaque problème sur une échelle de temps [Brainbridge96] ou par un schéma du corps humain représentant les régions atteintes par les problèmes médicaux du patient [Sundvall07] ou les travaux de J.B. Lamy [Lamy06]. Pour chaque patient, le

19. Cet outil a été acquis par la société IS@S

dossier peut être présenté par problème et/ou par ordre chronologique et/ou par spécialité (voir section 6.3).

7.4 Recherche contextuelle dans VidalRecos

La nécessité de maîtriser les données actuelles de la science et de respecter les référentiels en vigueur constitue l'une des bases de l'exercice professionnel pour un médecin. Le site VidalRecos [20] est un outil d'aide à la décision thérapeutique. Il constitue aussi un outil pédagogique pour les étudiants en médecine ou en pharmacie et pour les médecins dans le cadre de la formation médicale continue. Il diffuse des synthèses de recommandations thérapeutiques, appelées les «recos» résultant de la synthèse des recommandations thérapeutiques émanant de la HAS, de l'AFSSAPS et des sociétés savantes pour les situations médicales les plus fréquentes en médecine de ville. Des arbres décisionnels résument chacune des démarches thérapeutiques du diagnostic au traitement. Des grades de recommandation donnent le niveau de preuve scientifique chaque fois que cela est possible. En outre, pour chaque pathologie, tous les médicaments indiqués dans le traitement de celle-ci sont listés.

L'accès aux recommandations se fait grâce à un moteur de recherche, par domaine thérapeutique ou par ordre alphabétique des recommandations. Actuellement 125 recommandations Vidal sont disponibles. L'utilisateur peut aussi taper une requête en texte libre. Toutes les recommandations dont le titre correspond à la requête sont proposées.

Pour aider les utilisateurs à étendre leurs recherches, nous avons créé un accès contextuel afin de lier VidalRecos à d'autres bases de connaissances sur les recommandations [21]. Le choix s'est porté sur des sites de référence et de qualité où les documents sont soigneusement répertoriés facilitant ainsi la recherche. Les sites indexant les documents à l'aide de la terminologie MeSH et publiant des recommandations francophones - le site CISMeF - et étrangères pour les principaux sites médicaux internationaux - PubMed, NHS, NGC, Intute et CMA Infobase - ont été séléctionnés.

Chaque «reco» est liée à un ou plusieurs termes de recherche (plus de 3 000 termes de recherche). Afin de permettre l'interrogation du site CISMeF, chaque terme de recherche a été traduit en une requête CISMeF à l'aide de mots clés MeSH et d'opérateurs (exemple : la «recos» qui porte le titre «Ménopause : traitement hormonal» est liée au terme de recherche «traitement hormonal substitutif» qui a été traduit par la requête CISMeF «menopause.mc ET traitement hormonal substitutif.mc») (voir l'onglet «approfondir - recommandations francophones» figure 7.7). Ces correspondances sont produites manuellement par l'équipe CISMeF, revues par l'équipe Vidal et mises à jour régulièrement.

A partir de ces requêtes a été créé le transcodage terme de recherche Vidal/ termes MeSH [22] (2 947 correspondances). Ce transcodage permet l'approfondissement de la requête ainsi que l'interrogation des autres sites (voir l'onglet «appro-

20. Pour tester une recherche sur 3 «recos» voir `http://www.vidalrecos.fr/pages/index.php`
21. Ce travail est issu d'une collaboration Vidal-CISMeF
22. Tâche à laquelle j'ai participé dans l'équipe Vidal

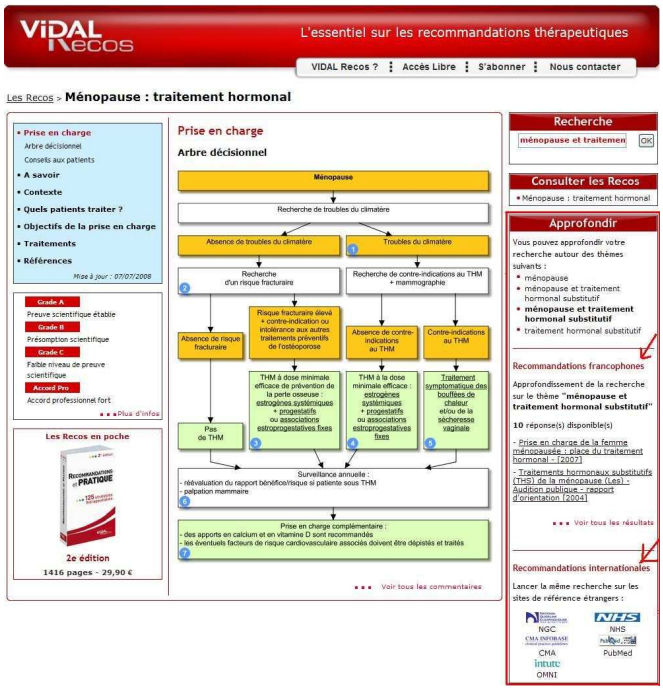

FIGURE 7.7 – Site VidalReco

fondir - recommandations internationales» figure 7.7). Pour chaque site (PubMed [23], NHS [24], NGC [25], Intute [26] et CMA Infobase [27]) un modèle de requête adapté a été créé par l'équipe CISMeF [28] (ce sont les mêmes modèles qui sont discutés dans la section précédente). Ce qui permet à l'utilisateur souhaitant approfondir sa rechercher d'accéder directement aux recommandations francophones et internationales en un clic.

7.5 Recherche translangue

Un patient recherchant une information en santé aura à sa disposition une multitude de ressources sur Internet. Il fait alors face à un problème : interroger des moteurs de recherche qui utilisent un langage qui ne lui est pas adapté [29] (exemple : le patient va rechercher des informations sur le mal de tête alors que son problème est identifié comme étant une «céphalée»). Les ressources ne sont pas toutes adaptées à son niveau de compréhension (vocabulaire trop technique, connaissances faibles du

23. Accessible ici `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi`
24. `http://www.nhs.uk/Pages/homepage.aspx`
25. Accessible ici `http://www.guideline.gov/`
26. Accessible ici `http://www.intute.ac.uk/`
27. Accessible ici `http://www.cma.ca/index.cfm/ci id/54316/la id/1.htm`
28. Il existe un contrat de coopération commerciale entre Vidal et l'équipe CISMeF pour le projet VidalRecos
29. Le langage courant et souvent très éloigné, dans la forme, des termes très pointus du monde médical [Tse03]

domaine) et écrites dans une langue différente de la sienne.

En matière de recherche d'information adaptée aux patients, il est nécessaire d'interroger des sites dont le contenu est adapté au niveau de compréhension des patients. De plus, il faut pouvoir traduire une requête en langage patient écrite en francais, en anglais [Néveol06]. Le passage d'une langue à une autre en matière de recherche d'information s'appelle la recherche translangue. La recherche d'information translangue peut être considérée comme une extension de la recherche d'information monolangue [Chiao04]. En pratique, elle peut être abordée de deux facons. La première est la traduction des documents dans la langue de la requête, malheureusement les méthodes de traduction automatique ne sont pas encore assez performantes et la masse de documents sur Internet est trop importante. La deuxième approche est la traduction de la requête.

Le site CISMeF utilise cette approche en permettant à l'utilisateur de rechercher des documents en francais à partir d'une requête tapée en francais et en anglais. Il propose aussi, en résultat d'une requête, des liens vers des catalogues (majoritairement) anglophones de qualité en santé, offrant ainsi aux utilisateurs la possibilité d'approfondir leurs recherches. En choisissant d'étendre sa recherche sur l'un de ces sites, l'utilisateur voit sa requête entrée en francais dans CISMeF traduite automatiquement en anglais. Ceci est rendu possible grâce à l'utilisation par tous ces sites d'un thesaurus multilingue, le thesaurus MeSH.

L'approfondissement de la recherche dans CISMeF sur d'autres catalogues n'est pour l'instant disponible que pour les ressources adaptées aux médecins (type de ressources : «recommandations»). Un travail similaire reste à réaliser pour l'accès à des ressources destinées aux patients (type de ressources : «patient»).

L'équipe CISMeF a créé en francais une liste de 531 synonymes patients rattachés aux termes MeSH [30] (431 termes MeSH sont impliqués) (exemple : «tabagisme passif» est un synonyme patient du terme MeSH «pollution fumée tabac»). Ces synonymes permettent de traduire au sein du catalogue une requête en langage patient en termes MeSH ce qui permet d'améliorer la recherche d'information [Plovnick04].

MedlinePlus [31] est un site en anglais à destination des patients et du grand public mis en place par la NLM. L'équipe MedlinePlus a créé en anglais 698 sujets de santé [32] afin de catégoriser leurs ressources. Plus tard, afin de rendre le site intéropérable avec d'autres catalogues, ces termes ont été reliés à 1 ou n termes MeSH (1130 en tout) (exemple : «health topic AIDS» est lié au mot clé MeSH «Acquired Immunodeficiency Syndrome and HIV infections»).

Ces termes patients ont été développés indépendamment en francais et en anglais par les équipes CISMeF et MEDLINEplus. Grâce à la traduction francaise du MeSH réalisée par l'INSERM, les liens entre termes MeSH francais et anglais sont déjà disponibles. Les efforts pour enrichir le MeSH avec des termes patient en francais (synonymes patient de CISMeF) et en anglais (MEDLINEPlus topics) a conduit à la création de liens sémantiques entre les termes patients et les termes MeSH dans

30. Nous lions des termes professionnels (MeSH) à leurs équivalents en langage courant.
31. Accessible ici `http://medlineplus.gov/`
32. Appelés aussi Consumer Health Information (CHI) terms

chaque langage (voir figure 7.8). Grâce à ces liens existants nous avons pu induire les liens qui existaient entre les termes patients en anglais et en francais. Par exemple, lié au terme patient anglais «second-hand smoking», nous trouvons le terme MeSH anglais «tobacco pollution», et son équivalent francais «pollution fumée tabac». Il existe un terme patient lié au terme MeSH francais, «tabagisme passif». Nous pouvons donc induire la relation d'équivalence entre les termes patient «tabagisme passif» et «second-hand smoking».

280 liens de ce type ont été créés soit 129 liens contextuels validés.

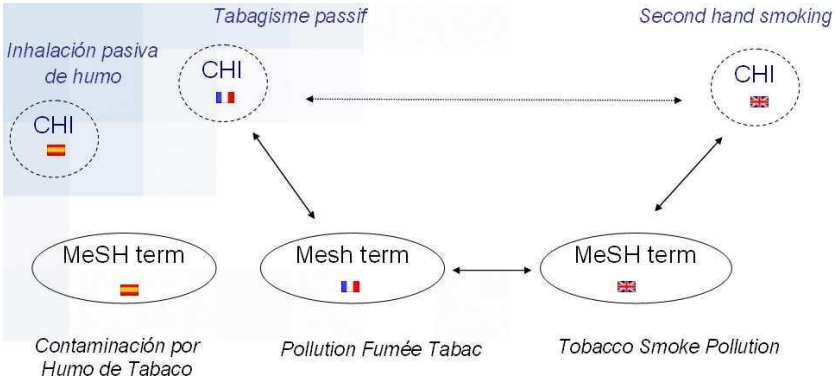

FIGURE 7.8 – Création de liens d'équivalence entre les termes patients en anglais et en francais

Ainsi, la requête dans CISMeF «tabagisme passif» concernant des ressources patient, retourne des documents indexés avec le terme MeSH «tabagisme passif» accompagné d'un lien contextuel permettant à l'utilisateur d'étendre sa recherche aux documents indexés par le terme MeSH «secondhand smoking» dans MedlinePlus (voir figure 7.9).

La méthode appliquée facile à réaliser utilise des données existantes. Cette méthode est une méthode générique qui pourrait être appliquée à d'autres langages comme l'espagnol et le chinois [Lu05] qui possèdent des termes patients rattachés au MeSH. Pour l'anglais une méthode d'extraction à partir de corpus parallèles (techniques et didactiques) permettent de définir des synonymes en langage courant pour les termes de l'UMLS [Elhadad07].

Salton a montré que la traduction d'une requête (anglais vers allemand) montre une performance élevée en matière de recherche d'information similaire à un système monolangue [Salton73].

Les méthodes de traduction de requête que nous avons proposées sont basées sur des thésaurus multilingues. Une étude a montré que le thesaurus MeSH donnait de meilleur résultats en terme de traduction automatique de requête [Ruch04] avec une meilleure désambiguïsation de termes difficiles. D'autres méthodes existent comme la traduction automatique de requête utilisant des dictionnaires ou des méthodes basées sur des corpus. Des travaux ont montré qu'une combinaison de ces deux méthodes peut améliorer les performances ou l'extraction de nouvelles traductions [Déjean05].

Notre étude permet la délivrance de connaissances de manière contextuelle entre

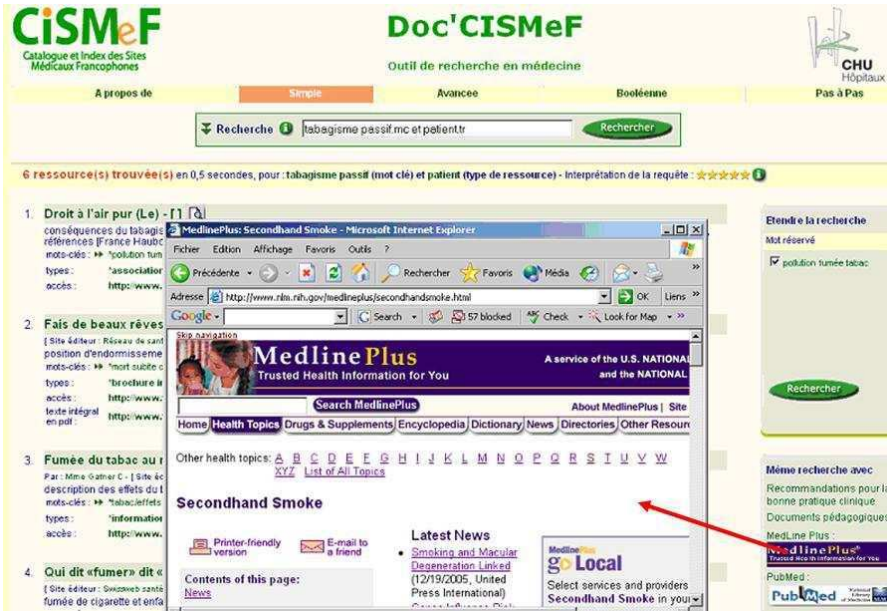

FIGURE 7.9 – Recherche d'information translangue sur le site MedlinePlus

deux banques d'informations CISMeF-patient et MedlinePlus. Cet accès a été repris dans un autre système d'information, le dossier électronique du patient (voir section 6.2) afin d'aider les patients à comprendre et à rechercher de l'information sur les données électroniques les concernant.

CISMeF n'est pas le seul site à proposer ce genre de recherche translangue. Les outils PICO et BabelMeSH permettent une recherche translangue pour MEDLINE/Pubmed en plusieurs langues : espagnol, francais, portugais, japonais, italien, allemand et russe [Liu06] [Fontelo07].

7.6 Discussion/Conclusion

Nous avons créé plusieurs accès contextuels :

- Un accès de type «InfoButton» à partir du dossier patient vers des bases de connaissances sur l'Internet donnant aux utilisateurs un accès, lorsqu'ils en ont besoin, à de l'information disponible et adaptée à leur profil (patient, médecin ou étudiant). L'outil développé a été mis en place au CHU de Rouen, valorisé auprès de l'Université et vendu à une société.
- Un deuxième accès de type «vue» a été concu au sein du dossier patient afin de filtrer les diagnostics et actes pour un patient en fonction de la spécialité médicale intéressant l'utilisateur. Là encore ce travail a été intégré au logiciel de gestion de dossiers patients du CHU de Rouen et vendu à une société.
- Un troisième accès de type «approfondissement de la recherche» qui permet à un utilisateur lors d'une recherche sur un moteur de recherche d'accéder à d'autres documents afin d'approfondir sa recherche sur d'autres sites de qualité a été réalisé. Ce système a été mis en place sur le site VidalRecos pour la

recherche de recommandations de bonnes pratiques.
- Un quatrième accès de type CLIR [33] a été élaboré pour aider les utilisateurs dans leur recherche d'information dans une langue qu'ils ne maîtrisent pas complètement. Ce système a été mis en place sur le site CISMeF.

Ces accès permettent, à des niveaux différents, et dans des domaines bien précis, d'accéder «à la bonne information, au bon moment et pour les bonnes raisons [34]».

L'accès à «la bonne information» est réalisé en prenant en compte la demande de l'utilisateur qui veut accéder à une information spécifique de qualité et qui est adaptée à son profil (son métier, son niveau de compréhension du domaine, sa langue, son pays d'origine, son niveau dans la pratique d'autres langues).

L'accès «au bon moment» nécessite la mise en œuvre de cet accès à un moment pertinent et quand l'utilisateur en a besoin (placé à des niveaux stratégiques au niveau de l'interface, là on aura besoin d'informations et où il sera intuitif pour lui d'aller en chercher) et au moment où c'est pertinent (disponibilité de l'information).

Plus on prend en compte d'éléments du contexte plus l'utilisateur gagnera du temps et moins il sera découragé dans sa recherche d'information car le système ne lui proposera que les documents correspondant au plus près à sa demande. Il est possible d'imaginer la prise en compte d'autres éléments du contexte : le temps dont on dispose (dans ce cas l'utilisateur préférera les documents de type résumé), l'endroit où l'on se trouve (information valide dans le pays d'origine) etc.... Le type de document et le pays de diffusion du document sont déjà référencés dans de nombreuses bases de connaissances telles que CISMeF, pour rendre cet accès contextuel disponible, il manque ici un profil utilisateur plus détaillé.

Il serait intéressant de mesurer la qualité et l'apport pour les utilisateurs de ces différents accès. Nous pourrions étudier par exemple la qualité des documents proposés ou par un mode d'interview la satisfaction de l'utilisateur [Gutnik07]. Ce genre d'outil pourrait être amélioré en donnant accès directement à l'information et non pas à un document contenant l'information recherchée. Les systèmes de question-réponse permettent l'accès à des informations précises [Berard-Dugourd89], nous pourrions les améliorer en ajoutant des éléments de contexte tels qu'étudiés ici. Ces éléments de contexte permettent en outre de désambiguiser et de préciser la question posée.

Nous pourrions aussi imaginer un profil rédigé en texte libre par l'utilisateur qui pourrait lui permettre de se décrire. L'outil F-MTI serait alors utilisé pour extraire les termes MeSH inclus et, à partir de règles, pour établir une stratégie de recherche d'information médicale contextuelle.

33. Cross-Langage Information Retrieval
34. *Access to the right information, at the right time for the right reason.*

Chapitre 8

Conclusion générale

Nous souhaitons dans ce chapitre réaliser le bilan de cette thèse.

Notre problématique initiale était d'aider les indexeurs dans leurs tâches d'indexation manuelle :

- l'indexation des ressources Web à l'aide du MeSH dans l'équipe CISMeF
- l'indexation des RCP à l'aide du TUV dans l'équipe données thérapeutiques de la société Vidal
- l'indexation des dossiers médicaux à l'aide de la CIM10, de la CCAM et de la SNOMED 3.5

Pour ce faire, nous avons développé un outil d'indexation automatique, F-MTI. Cet outil est capable de réaliser l'indexation de n'importe quel document à l'aide d'une ou plusieurs terminologies et permet une indexation des documents considérés dans nos différentes tâches.

Il a la particularité, contrairement à d'autres outils existants pour le francais, de réaliser une indexation multi-terminologique.

Il a demandé le développement d'une base de données multi-terminologique.

Trois méthodes d'indexation complémentaires ont été développées : la méthode du sac de mots, le dictionnaire de termes et le dictionnaire de constituants. Ces méthodes ont été associées à la création de libellés d'indexation pour chaque terme de chaque terminologie et une méthode d'extraction automatique de variantes lexicales à partir de corpus afin d'optimiser leurs performances.

Afin de tenir compte du contexte lors de l'indexation (négations, rubriques, paragraphes), nous avons ajouté certaines méthodes.

Deux de ces méthodes ont été évaluées dans la réalisation des tâches qui nous concernaient.

Trois méthodes de désuffixation ont également été comparées. Le FrenchStemmer de Lucene est apparu comme le meilleur outil pour le langage médical.

Enfin, l'outil a été comparé à d'autres outils d'indexation donnant des résultats satisfaisants.

Les applications potentielles de F-MTI au sein des trois équipes ont été envisagées. Ainsi, l'outil sera ainsi intégré, pour la société Vidal, dans l'outil d'aide à l'indexation des RCP, BIBLIS. Au sein du moteur de recherche CISMeF, il sera utilisé pour l'indexation automatique et semi-automatique des ressources Web à l'aide

de plusieurs terminologies. Dans un dossier patient électronique, cet outil permettra une aide à l'indexation médico-économique, pour le calcul du budget des hôpitaux, et descriptive pour la structuration des dossiers patients.

F-MTI sera utilisé dans plusieurs projets de recherche :

- Interstis pour la recherche de termes dans un serveur multi-terminologies
- PSIP pour la collecte de données pouvant permettre d'optimiser la sécurisation de prescriptions
- Aladin pour la détection des infections nosocomiales à partir de documents textuels hospitaliers

Nous avons envisagé et testé d'autres applications de notre outil. Celles-ci sont l'aide au transcodage, l'indexation multilingue, l'aide à l'indexation généraliste, la constitution de résumés automatiques et l'aide à la rédaction pour lesquelles les travaux seront poursuivis.

D'autres travaux ont consisté à créer des outils et mettre au point des méthodes pour permettre aux utilisateurs d'accéder à la bonne information, au bon moment.

C'est ainsi qu'un accès de type «InfoButton» permet à partir du dossier patient d'accéder à des bases de connaissances sur Internet donnant aux utilisateurs un accès, lorsqu'ils en ont besoin, à de l'information disponible et adaptée à leur profil (patient, médecin ou étudiant). L'outil développé a été mis en place au CHU de Rouen, valorisé auprès de l'université et vendu à une société.

Un deuxième accès de type «vue» a été concu au sein du dossier patient afin de filtrer les diagnostics et actes pour un patient en fonction de la spécialité médicale intéressant l'utilisateur. Là encore ce travail a été mis en place au CHU de Rouen et vendu à une société.

Un troisième accès de type «approfondissement de la recherche» qui permet à un utilisateur, à partir d'un moteur de recherche, d'approfondir sa recherche sur d'autres sites de qualité a été réalisé. Ce système a été mis en place sur le site VidalRecos pour la recherche de recommandations de bonnes pratiques.

Enfin, un quatrième accès de type CLIR a été élaboré pour aider les utilisateurs dans leurs recherche d'information dans une langue qu'ils ne maîtrisent pas complètement. Ce système a été mis en place sur le site CISMeF.

Au cours de cette thèse, nous avons pu répondre aux besoins des différentes équipes. Un important travail dont nous avons pu identifier les contours reste encore à réaliser afin d'obtenir une indexation automatique de qualité. La suite est déjà assurée avec des thèses en cours, et des projets à venir. Il est vraisemblable que je continue à travailler la réalisation de ces projets.

Au travers de tous les travaux réalisés au cours de cette thèse, nous avons pu parfaire nos connaissances dans le domaine du traitement automatique du langage, de la multi-terminologie et les appliquer au travers de réalisations concrètes.

Annexe A

Annexes

A.1 UMLS

CUI	Unique identifier for concept
LAT	Language of term
TS	Term status
LUI	Unique identifier for term
STT	String type
SUI	Unique identifier for string
ISPREF	Atom status - preferred (Y) or not (N) for this string within this concept
AUI	Unique identifier for atom - variable length field, 8 or 9 characters
SAUI	Source asserted atom identifier [optional]
SCUI	Source asserted concept identifier [optional]
SDUI	Source asserted descriptor identifier [optional]
SAB	Abbreviated source name (SAB)
TTY	Abbreviation for term type in source vocabulary
CODE	Most useful source asserted identifier
STR	String
SRL	Source restriction level
SUPPRESS	Suppressible flag. Values = O, E, Y, or N

exemple :
C0001175|ENG|P|L0001175|VO|S0010340|Y|A0019182||M0000245|D000163|MSH|PM|D00
0163|Acquired Immunodeficiency Syndromes|0|N||

FIGURE A.1 – Description des champs de la table MRCONSO

L'UMLS est constitué de plusieurs bases de données :
- Les concepts et leur source sont stockées dans la base de données MRCONSO (voir détail figure A.1).
- Les attributs (MRSAT, MRDEF, MRSTY, MRHIST)
- Les relations (MRREL (détail voir figure A.2), MRCOC, MRCXT, MRHIER, MRMAP, MRSMAP)
- Les données sur le Métathesaurus (MRFILES, MRCOLS, MRDOC, MRRANK, MRSAB, AMBIGLUI, AMBIGSUI, CHANGE/MERGEDCUI, CHANGE/MERGEDLUI, CHANGE/DELETEDCUI, CHANGE/DELETEDLUI, CHANGE/DELETEDSUI, MRCUI)

– Les index (MRXW-BAQ, MRXW-DAN, MRXW-DUT, MRX-ENG, MRXW-FIN, MRXW-FRE, MRXW-GER, MRXW-HEB, MRXW-HUN, MRXW-ITA, MRXW-NOR, MRXW-POR, MRXW-RUS, MRXW-SPA, MRXW-SWE, MRXNW-ENG, MRXNS-ENG)

CUI1	Unique identifier of first concept
AUI1	Unique identifier of first atom
STYPE1	The name of the column in MRCONSO.RRF that contains the identifier used for the first concept or first atom in source of the relationship
REL	Relationship of second concept or atom to first concept or atom
CUI2	Unique identifier of second concept
AUI2	Unique identifier of second atom
STYPE2	The name of the column in MRCONSO.RRF that contains the identifier used for the second concept or second atom in the source of the relationship
RELA	Additional (more specific) relationship label (optional)
RUI	Unique identifier of relationship
SRUI	Source asserted relationship identifier
SAB	Abbreviated source name of the source of relationship
SL	Source of relationship labels
RG	Relationship group
DIR	Source asserted directionality flag. Y indicates that this is the direction of the relationship in its source
SUPPRESS	Suppressible flag. Values = O, Y, E, or N
CVF	Content View Flag

Exemple :
C0002372|A0022284|AUI|RB|C0002371|A0022279|AUI||R01983351||MSH|MSH|||N||

FIGURE A.2 – Description des champs de la table MRREL

A.2 Modèles unitaires pour la base de données multi-terminologique

A.2.1 Modèle CISMeF

Le modèle de représentation de la terminologie CISMeF déduit de la description faite à la section 2.3.2 est présenté figure 3.2. Neuf classes ont été identifiées :

– **Classe des descripteurs**
But : Cette classe renseigne les descripteurs du thésaurus.
Les attributs :
L'attribut **code** désigne le code et l'attribut **code_hier** les codes arborescences (de 1 à n) MeSH pour le descripteur.
Le libellé du descripteur est inscrit dans l'attribut **libellé** avec la langue dans lequel il est exprimé *via* l'attribut **langue** (anglais ou francais).
L'attribut **qualifs_affiliables** renseigne les codes des qualificatifs affiliables pour le descripteur (de 0 à n).
Enfin, l'attribut **PT** permet d'indiquer le statut du terme (PT : terme préféré, S : synonyme).

- **Classe des Qualificatifs**
  But : Cette classe renseigne tous les qualificatifs du thésaurus MeSH.
  Les attributs :
  Le code, le libellé et la langue du qualificatif sont désignés par les attributs **code**, **libellé** et **langue**.
  L'attribut **ABR** permet de préciser en plus l'abréviation pouvant être utilisée pour exprimer le qualificatif. Et l'attribut **PT** renseigne le statut du terme (PT : terme préféré, S : synonyme).

- **Classe des Types de ressources**
  But : Cette classe renseigne tous les types de ressources CISMeF.
  Les attributs :
  Un attribut suffit, celui qui désigne le libellé du type de ressource, **libellé**.

- **Classe des Métatermes**
  But : Cette classe réunie tous les métatermes pouvant être rattachés à un ou plusieurs descripteurs, qualificatifs et types de ressource.
  Les attributs :
  L'attribut **libellé** désigne le libellé du métaterme.
  Les attributs **descripteurs_liés**, **TR_liés** et **qualifs_liés** permettent de renseigner tous les codes descripteurs, les types de ressources et les codes qualificatifs pouvant être reliés au métaterme.

- **Classe Hiérarchie**
  But : Cette classe structure la hiérarchie au sein du MeSH.
  Les attributs :
  L'attribut **code_père** désigne le code MeSH du père et l'attribut **code_fils** désigne le code MeSH de son fils.
  De plus, l'attribut **Niveau** permet de préciser le niveau du lien père-fils (niveau 1 : père-fils, niveau 2 : grand père-fils).
  Commentaires : La hiérarchie MeSH est complexe, nous pouvons avoir de 1 à n fils pour un père et de 1 à n pères pour un fils.

- **Classe Voir aussi**
  But : Cette classe renseigne tous les liens de «voir aussi» entre deux codes MeSH.
  Les attributs :
  Les attributs **code1** et **code2** permettent de renseigner les deux codes liés par un lien de «voir aussi».
  Commentaires : Il existe de 0 à n liens «voir aussi» pour un code MeSH.

- **Classe des Définitions**
  But : Cette classe réunie pour chaque code MeSH les définitions auxquelles ils sont rattachés.
  Les attributs :

L'attribut **code** désigne le code MeSH auquel s'applique la définition et les attributs **définition** et **source** renseignent la définition ainsi que sa source.
Commentaires : Il existe de 0 à n définitions pour chaque code MeSH.

- **Classe Dictionnaire**
  But : Cette classe indique toutes les variations, flexions, synonymes et leurs classes lexico-syntaxiques pour chaque terme MeSH.
  Les attributs :
  L'attribut **terme** désigne les variations lexicales, fonctionnelles ou synonymiques pour le code MeSH et l'attribut **données lexico-syntaxiques** leurs données lexicales (ex : maladie) ou syntaxiques (ex : nom féminin pluriel).
  Enfin, l'attribut **code** renseigne le code du terme MeSH dont les variations sont indiquées.

- **Classe des Actions pharmacologiques**
  But : Cette classe renseigne tous les liens «action pharmacologique» entre deux termes MeSH.
  Les attributs :
  Les attributs **code** et **action pharmaco** désignent le code MeSH du descripteur ainsi que le code MeSH précisant son action pharmacologique. L'attribut **qualif** renseigne le qualificatif précisant le sens du code descripteur.
  Commentaires : Il existe de 0 à n liens «action pharmacologique» pour chaque code MeSH.

A.2.2 Modèle TUV

Ce modèle est présenté figure 3.3. Ce modèle présente 8 classes :

- **Classe des Thesaurus**
  But : Cette classe réunie tous les termes de référence du thesaurus TUV.
  Les attributs :
  Les attributs **thesaurus_id** et **thesaurus_name** désignent le code et le libellé du terme d'indexation TUV.

- **Classe des Concepts**
  But : Cette classe réunie tous les termes élémentaires décrivant un terme de référence du TUV.
  Les attributs :
  Les attributs **concept_id** et **concept_name** indiquent le code et le libellé du concept. Des attributs permettent ensuite de décrire le type du concept : **concept_type** renseigne le type (état ou complément) et **concept_semanticLabel** désigne le type sémantique (pathologie, physiologie etc...).
  Enfin, l'attribut **thesaurus_id** est le code du terme de référence décrit par le concept.
  Commentaires : Il existe de 1 à n termes élémentaires décrivant un terme de

référence.

- **Classe des Group**
  But : Cette classe désigne les liens d'appartenance d'un terme d'indexation à un groupe d'indications.
  Les attributs :
  L'attribut **thesaurus_id** désignant le code TUV du terme d'indexation est ainsi lié à un groupe d'indication décrit par l'attribut **group_name**.
  Commentaires : Un terme d'indexation peut être rattaché à 0 à n groupes d'indications.

- **Classe des Classification_X**
  But : Cette classe renseigne tous les liens reliant un terme de référence ou un terme élémentaire à d'autres terminologies telles que CIM10, la CISP ou la SFMG.
  Les attributs :
  L'attribut **id** désignant le code du terme TUV (terme de référence ou terme élémentaire) est associé à un **idX**, code d'une autre terminologie indiqué par la source **classiffication_X**.
  Commentaires : Un terme du TUV peut être transcodé en 0 à n codes d'autres terminologies.

- **Classe des Thesaurus_Lexical_Alternative**
  But : Cette classe indique toutes les variantes lexicales, flexionnelles et synonymiques pour chaque terme d'indexation (terme complexe).
  Les attributs :
  Au **thesaurus_id** désignant le code TUV du terme d'indexation peut être associé un **thesaurusLexicalAlternative_id** qui indique le code de la variante du terme d'indexation et à un **thesaurusLexicalAlternative_name**, le libellé de la variante.
  Commentaires : Le libellé du terme de référence est considéré comme une variante possible. Un terme de référence peut être relié à 1 à n variantes.

- **Classe des Concept_Lexical_Alternative**
  But : Cette classe indique toutes les variantes lexicales, flexionnelles et synonymiques pour chaque terme élémentaire.
  Les attributs :
  De même, au **concept_id** désignant le code TUV du terme élémentaire peut être associé un **concept_Lexical_Alternative_id** qui indique le code de la variante du terme élémentaire et à un **concept_Lexical_Alternative_name**, le libellé de la variante.
  Commentaires : Le libellé du terme élémentaire est considéré comme une variante possible. Un terme élémentaire peut être relié à 1 à n variantes.

- **Classe des Relation_concept**

But : Cette classe renseigne tous les liens sémantiques pouvant relier deux termes élémentaires.
Les attributs :
Les deux termes élémentaires désignés par les codes **concept_id1** et **concept_id2** sont liés dans une relation sémantique.
De plus, l'attribut **relation_concept_type** renseigne sur le type de la relation sémantique (exemple : «symptôme» et «père-fils»).
Commentaires : Il peut exister pour un même terme plusieurs relations sémantiques vers d'autres termes TUV.

- **Classe des Relation_semanticLabel**
  But : Cette classe renseigne tous les liens sémantiques pouvant relier deux étiquettes sémantiques.
  Les attributs :
  Les deux attributs **relation_semanticLabel1** et **relation_semanticLabel2** renseignent les deux étiquettes sémantiques impliquées dans la relation **relation_concept_type**.

A.2.3 Modèle de la CIM10

Ce modèle a été inspiré par la représentation formelle de la classification CIM10 en entités et relations de l'OFS (Office Fédéral de la statistique) [OFS06]. Ce modèle comporte 9 classes (voir figure A.3), voici quelques indications :

- **Classe des Termes systématiques** : cette classe définit tous les termes systématiques de la classification CIM10.
  Quelques commentaires : la terminologie source des termes CIM10 est donnée par l'attribut **source** (FR_OMS, EN_OMS, GE_DIMDI, GE_AUTO, FR_CHRONOS, ICD10DUT, ICD10AMAE, ICD10AM, ICD10AE, ICD10). L'attribut **niveau**, quand à lui, définit le niveau du code CIM10 (C - chapitre, G - bloc U- sous-bloc, K - catégorie, S - sous-catégorie, D - subdivision ou descripteur, L - local).
- **Classe des Descripteurs** : cette classe définit tous les descripteurs décrivant les termes systématiques de la classification CIM10.
  Quelques commentaires : Il y a de 0 à n descripteurs pour chaque terme de la classification CIM10.
- **Classe des Références** : cette classe définit toutes les références liées à des termes systématiques et descripteurs de la classification CIM10.
- **Classes des Inclusions** : cette classe défini quels sont les libellés de type «comprend »associés à certains termes systématiques de la classification.
  Quelques commentaires : l'attribut **code** désigne le code CIM10 tu terme systématique et l'attribut **libellé** le libellé du terme inclus.
- **Classes des Exclusions** : cette classe identifie pour un terme systématique toute exclusion d'un autre terme.
  Quelques commentaires : l'attribut **code désigne** le code CIM10 du terme excluant. Les attributs **code_exclu**, **libellé** et **type_exclusion** désignent le

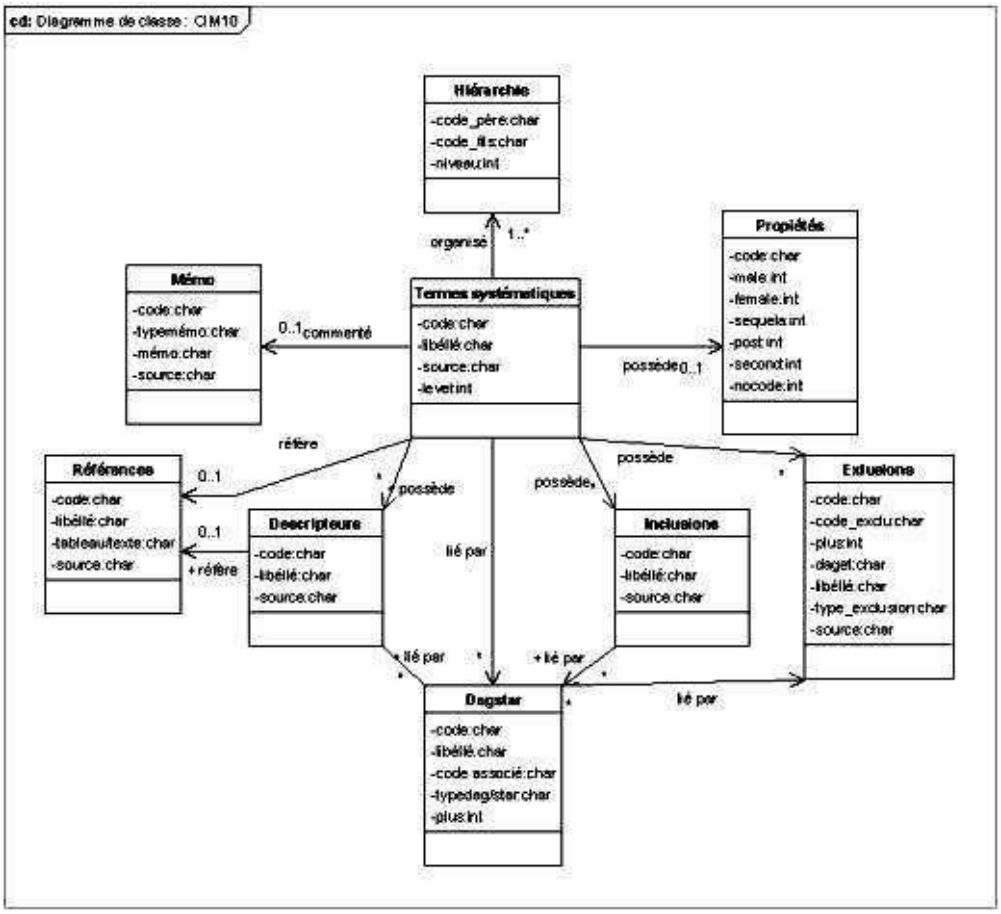

FIGURE A.3 – Diagramme de classes représentant la structure de la CIM10 au formalisme UML

code CIM10 du terme exclu, son libellé ainsi que son type d'exclusion (directe : D , indirecte : I pas de code_exclu pour celui-là). L'attribut **typedag/star** désigne le type de renvoi dague ou astérisque si elle s'applique à une exclusion avec un **plus** si la relation est de type dag astérisque +.

- **Classes des Dagstar** : cette classe explicite tous les appariements dagues et astérisques de la classification CIM10.
  Quelques commentaires : les attributs **code** et **libellé** désignent le code et le libellé du terme de départ (descripteur, du terme systématique, de l'exclusion ou de l'inclusion) de l'appariement dague et étoile. L'attribut **code_associé** désigne le code du terme CIM10 apparié au précédent. L'attribut **typedag/star** désigne le type de renvoi dague ou astérisque avec un drapeau **plus** signifiant l'extension de l'appariement portant sur une catégorie à la sous-catégorie adéquate.
- **Classe des Mémo** : cette classe indique les glossaires ou notes qui peuvent être rattachés aux termes systématiques CIM10.
- **Classe des Propriétés** : cette classe réuni pour chaque terme différents attributs de type booléen (sexe, séquelles, états après, non valable comme diagnostic

principal, non codable). **Quelques commentaires** : L'attribut **female** qualifie les termes réservés exclusivement au sexe féminin. L'attribut **sequela** qualifie les termes réservés exclusivement au sexe masculin. L'attribut **post** qualifie les termes réservés aux désordres après une opération. L'attribut **second** qualifie les termes non utilisables comme diagnostic principal. L'attribut **nocode** qualifie les termes non codables, c'est-à-dire pour lesquels il existe un terme plus approprié plus bas dans la hiérarchie. Pour le CIM10 : la valeur O (pour oui) sera attribuée aux termes de dernier niveau, N pour les autres. Tous les codes ayant un descendant ne sont pas codables, soit 1849 termes. L'attribut second n'est pas renseigné il devra être complété.

A.2.4 Modèle de la CCAM

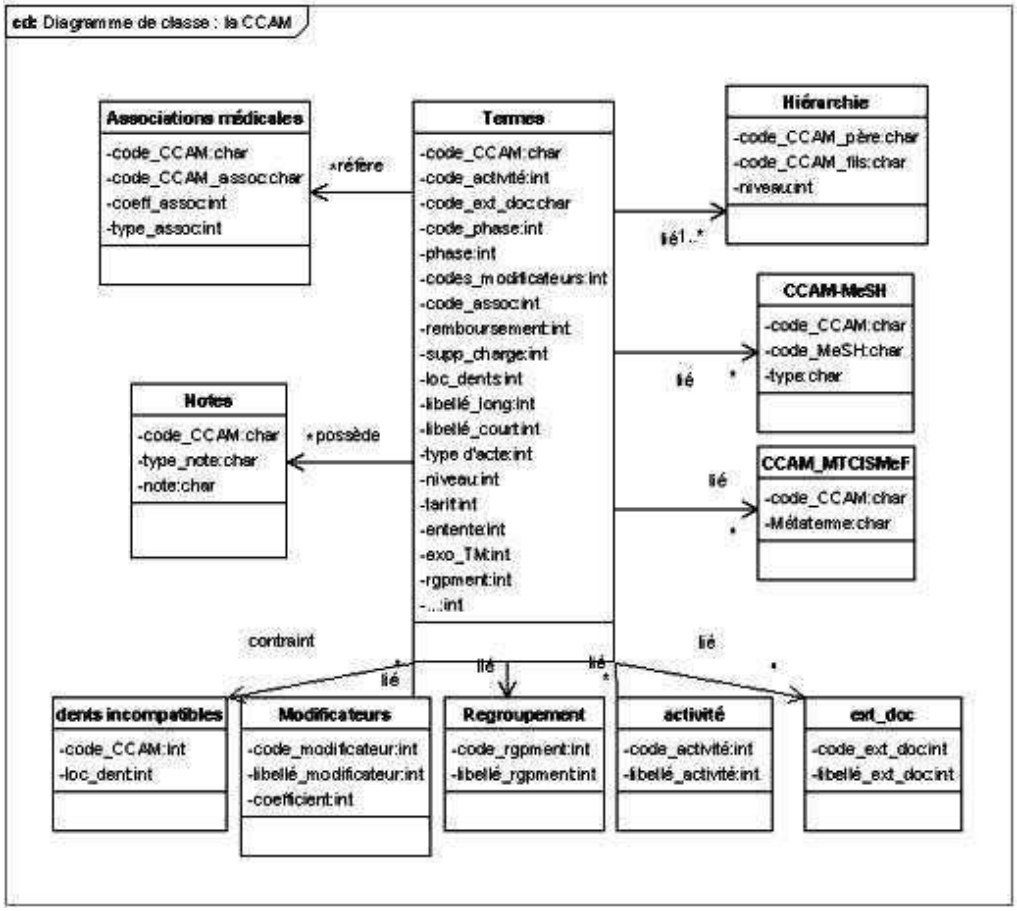

FIGURE A.4 – Diagramme de classes représentant la structure de la CCAM au formalisme UML

Nous avons aussi élaboré le modèle de la CCAM (voir figure A.4). Celui-ci comporte 11 classes, voici quelques indications :

– **Classe des Termes** : Cette classe définit tous les termes de la classification CCAM. **Quelques commentaires** : L'attribut **code_activité** identifie

les actions réalisées par un même acte par différents intervenants. L'attribut **code_ext_doc** collecte les extensions documentaires pour ce terme. L'attribut **code_phase** spécifie le numéro de la phase (ex : 1e phase...). L'attribut **phase** indique la phase de traitement pour cet acte. L'attribut **code_modificateurs** collecte les codes modificateurs pouvant s'appliquer au code (4 maximum séparés par un point virgule). L'attribut **code_assoc** permet de signaler des associations d'actes non prévues. L'attribut **remboursement** renseigne sur le remboursement possible sous condition ou non pour cet acte (N ou O). L'attribut **supp_charges** indique si un supplément au tarif est appliqué en cas d'acte en cabinet (C si oui). L'attribut **loc_dents** renseigne les localisations de dents traitées si acte dentaire (6 maximum séparées par un point virgule). Les attributs **libellé_long** et **libellé_court** spécifient le libellé long et le libellé court pour ce code. L'attribut **type_d'acte** indique le type de l'acte (AI : acte isolé, P : procédure, AC : geste complémentaire). L'attribut **niveau** indique le niveau du code dans la hiérarchie. L'attribut **tarif** indique le tarif pour cet acte. L'attribut **entente** indique si l'acte est soumis à une entente préalable (O ou N). L'attribut **exo_TM** indique si l'acte peut être exonéré et dans quelles conditions. L'attribut **rgpment** désigne le code regroupement de l'acte. L'attribut **...** indique tous les autres champs qui peuvent être insérés dans la table et non présentés précédemment.

- **Classe des Modificateurs** : cette classe comprend la liste de tous les modificateurs pouvant être reliés à n'importe quel code CCAM.
  Quelques commentaires : l'attribut **coefficient** indique le coefficient appliqué au tarif pour ce modificateur. 10 codes possibles.
- **Classe des Propriétés** : cette classe comprend la liste de tous les codes regroupement pouvant être rattaché à un code CCAM.
  Quelques commentaires : 15 codes possibles.
- **Classe des Activité** : cette classe comprend la liste de tous les codes activité pouvant être relié à n'importe quel code CCAM.
  Quelques commentaires : 6 codes possibles.
- **Classe des Ext_doc** : cette classe comprend la liste de toutes les extensions documentaires pouvant être reliées à n'importe quel code CCAM.
  Quelques commentaires : 10 codes possibles.
- **Classe des Associations médicales** : cette classe indique toutes les associations de codes (code CCAM+code activité) permises et non permises pour un code CCAM (voir annexe n°13).
  Quelques commentaires : l'attribut **code_activité** représente le code de l'activité du code associé. L'attribut **coeff_assoc** indique le coefficient de l'association appliqué au tarif. L'attribut **type_assoc** permet de signaler si l'association est permise ou non.
- **Classe des Notes** : cette classe indique les notes qui peuvent être rattachés aux termes CCAM (voir annexe n°11).
  Quelques commentaires : l'attribut **type_note** indique le type de la note («à l'exclusion de...», «comprend...», «inclut...»etc...).
- **Classe des Dents incompatibles** : cette classe indique les localisations de

dents incompatibles avec l'acte pratiqué.
Quelques commentaires : l'attribut **loc_dent** indique les localisations de dents incompatibles avec l'acte désigné précédemment.

- **Classe CCAMMeSH** : cette classe contient le transcodage CCAM-MeSH qui a été réalisé par Philippe Massari (voir chapitre 6 de la thèse)
  Quelques commentaires : l'attribut **type** qualifie le type du code MeSH (technique,...).
- **Classe CCAMMTCISMeF** : cette classe réuni pour chaque code CCAM les métatermes qui y sont rattachés.

A.2.5 Modèle SNOMED 3.5

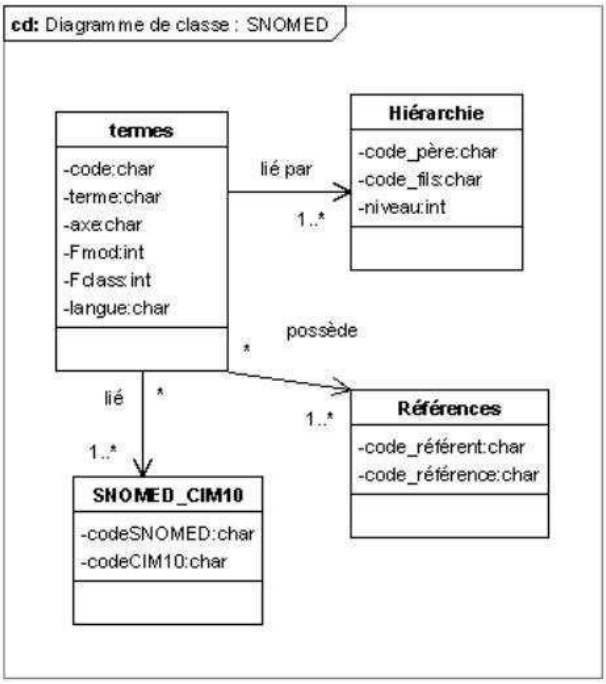

FIGURE A.5 – Diagramme de classes représentant la structure de la SNOMED au formalisme UML

Enfin, nous avons réalisé le dernier modèle celui de la SNOMED 3.5 (voir figure A.5). Celui-ci comporte 4 classes, voici quelques indications :

- **Classe des Termes** : cette classe comprend tous les termes de la nomenclature SNOMED. **Quelques commentaires** : l'attribut **axe** renseigne l'axe auquel appartient le terme (de 1 à 11). L'attribut **Fmod** (F,B) indique la source du terme. La lettre F signifie que ce terme est utilisé principalement en France, mais pas au Québec. La lettre B signifie que c'est un code de Bethesda. L'attribut **Fclass** indique la classe du terme (0 : terme de section ou chapitre, 1 : terme préféré, 2 : synonyme, 3 : variante lexicale).
- **Classe des Références** : cette classe réuni pour chaque code SNOMED, les codes SNOMED auxquels ils réfèrent.

– **Classe SNOMED_CIM** : cette classe comprend tous les liens de transcodages entre un terme SNOMED et un terme CIM10.
Quelques commentaires : l'attribut **source** renseigne la source du code CIM (CIM9, CIM10, code CIM9 supplémentaire ou code CIM10 supplémentaire).

A.3 Modèle général

Le modèle général a ainsi été défini selon 7 classes (voir figure 3.4) :

– **Classe des Concepts UMLS**
But : Cette classe indique, pour chaque code des différentes terminologies, les liens vers les concepts UMLS (quand ils existent donc seulement pour les codes MeSH (exclu les termes spécifiques CISMeF), CIM10 et SNOMED). Cette classe est inspirée de la table **MRCONSO** (contenant les sources et les noms des concepts dans le Metathésaurus de l'UMLS - voir Annexes A).
Les attributs :
Tous les codes répertoriés dans l'UMLS pour les termes CIM10, MeSH ou SNOMED sont répertoriés ici : l'identifiant unique du concept UMLS **CUI**, l'identifiant unique du terme **LUI**, l'identifiant unique de la chaîne de caractères **SUI**, l'identifiant unique de l'atome **AUI** ce qui est généralement le code de dernier niveau dans l'UMLS et, enfin, l'identifiant unique du terme dans la terminologie le **code-termino**.
Une série d'attribut permettent aussi de définir le type du terme au sein du métathésaurus : le type du terme **TS**, de la chaîne de caractères **STT**, et **ISPREF** qui indique si le **AUI** est le préféré ou non.

– **Classe des Termes**
But : Cette classe réunie tous les termes de chaque terminologie. Cette table a été inspirée de la table MRCONSO (contenant les sources et les noms des concepts dans le Metathésaurus) de l'UMLS. Elle regroupe toutes les classes décrivant les termes pour chaque terminologie : **Termes** de la CCAM, **Descripteur**, **Qualificatif**, **Type de ressource**, **Métaterme** du MeSH, **Termes** de la SNOMED et enfin **Termes systématiques**, **Descripteurs** et **Inclusions** de la CIM10.
Les attributs :
L'attribut **code-termino** renseigne le code du terme dans la terminologie source.
Modifications apportées : nous avons ajouté un code pour les Types de ressources et les Métatermes du MeSH.
L'attribut **langue** indique la langue dans laquelle le libellé est exprimé.
Commentaire : Les cinq terminologies ont été intégrées en francais dans F-MTI, donc ici l'attribut pour tous les termes est égal à «FRE» pour francais.
L'attribut **source** désigne la terminologie dont est issu le terme («CIM10», «SNMI», «MeSH» ou «CCAM»).
L'attribut **classe_terme** indique la classe du terme (0 : terme de section ou

chapitre, 1 : terme préféré, 2 : synonyme, 3 : variante lexicale, 4 : abréviation, 5 : descripteur, 6 : inclusion). Il correspond à l'attribut «F class» de la SNOMED et permet de renseigner l'attribut **PT** du MeSH ainsi que toutes les formes particulières pour les différentes terminologies (les libellés courts CCAM et les abréviations des qualificatifs du MeSH en valeur 4 (abréviation) et les termes CIM10 notés 5 : descripteur et 6 : inclusion).
Commentaires : le MeSH ne possède pas de terme de section ou de chapitre et la CCAM ne renseigne que des termes préférés.
L'attribut **libellé** renseigne le libellé du terme.
L'attribut **niveau_hier** renseigne le niveau du terme dans la hiérarchie de la terminologie. Ce qui correspond à l'attribut **level** de la CIM10, **niveau** de la CCAM et **axe** de la SNOMED.
Les **niveau_hier** Q - qualificatif, D - Descripteur, TR - type de ressource, MT - métaterme ont été créés pour le MeSH et CC - concept complexe, CE - concept élémentaire pour le TUV.
L'attribut **propriétés** renseigne les propriétés des termes.
Plusieurs valeurs séparées par des « ;» peuvent être indiquées.
Les **propriétés** M - male, F - female, S - sequela, P - post, S - second ont été créées pour la CIM10. Pour la CCAM sont renseignés ici les codes influant sur la tarification (code_activité, code_ext_doc, code_phase, codes_modificateurs, code_assoc, remboursement, supp_charge, loc_dents).
Pour le MeSH, nous avons renseigné ici les codes arborescences des termes MeSH séparés par des « ;». Il n'existe pas de propriétés pour les termes de la SNOMED, l'attribut sera donc «NULL».
Enfin, l'attribut **codable** renseigne si le code peut être indexé ou non. Dans toutes les terminologies, on retrouve des termes pouvant être indexés et d'autres non.
Modifications apportées : Pour la CCAM, la valeur «N» pour «non» sera attribuée aux termes de chapitre ou de section, «O» pour les autres. Pour la SNOMED, la valeur «N» sera attribuée aux termes de chapitre ou de section, «O» pour les autres. Enfin pour le MeSH, la valeur «N» sera attribuée aux qualificatifs (seuls ils ne peuvent pas être codés), «O» pour les autres.

– **Classe des Relations inter-terminologies**
But : Cette classe renseigne toutes les relations qui peuvent exister entre deux termes de terminologies différentes. Cette table a été inspirée par la table MRREL (Related Concepts) de l'UMLS. Elle inclut les transcodages entres terminologies : CCAM-MeSH et CCAM_MTCISMeF (voir section 5.8.1), SNOMED-CIM10, TUV-MeSH, TUV-CIM10. Elle intègre aussi toutes les relations inter-terminologiques comprises dans l'UMLS : tel que les liens de transcodage SNOMED-CIM10, SNOMED-MeSH et MeSH-CIM10.
Les attributs :
Les attributs **code1** et **code2** désignent les deux codes impliqués dans la relation. Les attributs **STYPE1** et **STYPE2** indiquent chaque type de code impliqué dans la relation (valeurs : AUI, CODE ou CUI).

Les attributs **SAB1** et **SAB2** désignent les terminologies sources de chaque code (valeurs : SNMI, TUV, UMLS, CCAM, CIM10 ou MeSH).
Enfin, l'attribut **relation** renseigne le type de la relation liant les deux codes.
Modifications apportées : les relations : «transcodage» et «appartenance à un groupe» ont été ajoutées pour le TUV.

– **Classe des Relations**
But : Cette classe précise les relations secondaires qui peuvent exister entre les relations elles-même. Elle est inspirée de la table MRHIER (Computable Hierarchies) de l'UMLS.
Les attributs :
Les attributs **relation1** et **relation2** désignent les relations impliquées et l'attribut **type_relation** indique le type de relation qui existe entre ces 2 relations.
Enfin, l'attribut **attribut_relation** renseigne le type sémantique de la relation.

– **Classe des Relations intra-terminologies**
But : Cette classe renseigne toutes les relations qui peuvent exister entre deux termes d'une même terminologie. Cette table a été inspirée par la table **MRREL** (Related Concepts) et **MRHIER** (Computable Hierarchies) de l'UMLS. Elle inclue les classes **Hiérarchie**, **Voir aussi**, **Actions pharmacologiques** du MeSH, **Assiociations médicales** et **Hiérarchie** de la CCAM, **Hiérarchie** et **Références** de la SNOMED, **Hiérarchie**, **Inclusions**, **Dagstar** et **Exclusions** de la CIM10 et enfin **Relation_concept** du TUV. Elle inclut également toutes les relations sémantiques comprises dans l'UMLS pour une même terminologie.
Les attributs :
Cette classe a la même structure que celle des relations intra-terminologiques.
L'attribut **attribut_relation** renseigne le type sémantique de la relation (attributs **niveau** pour les relations «père_fils», «type_assoc» du MeSH et «typedag/star»de la CIM10).
L'attribut **libellé_associé** indique le libellé lié à la relation. Celui-ci correspond aux attributs **libellé** pour les «exclusions» et «dagstar» de la CIM10 et les qualificatifs pour les «actions pharmacologiques» du MeSH.
Modifications apportées : les relations «exclusions», «exclusions systématiques», «dagstar» de la CIM10, «références» de la SNOMED, «associations médicales» de la CCAM, «regroupement» de la CCAM, «voir aussi», «MT/TR», «MT/D», «MT/Q», «D/Q», «actions pharmacologiques» du MeSH ont été ajoutées à celles de l'UMLS.
La valeur NULL sera attribuée pour les autres terminologies et relations.

– **Classe des Mémos**
But : Cette classe renseigne toutes les notes et mémos rattachés aux termes des différentes terminologies. Elle inclut les classes **Mémo** et **Références** de la CIM10, **Notes** et **Définitions** du MeSH et **Notes** de la CCAM. Cette classe est inspirée par la table **MRDEF** de l'UMLS.

Les attributs :
L'attribut **code** désigne le code du terme de la terminologie source **SAB** rattaché au mémo **mémo**.
L'attribut **type** précise le type du mémo («glossaire», «note», «référence», «infotarif »). Enfin, l'attribut langue précise la **langue** dans laquelle est exprimé le mémo.
Modifications apportées : les autres attributs reliés aux termes CCAM (**exo_TM**, **tarif**, **entente** etc...) considérés comme purement informationnels ont été ajoutés. Ils seront séparés par un « ; ».

– **Classe des Alternatives lexicales termes**
But : Cette classe réunie toutes les variations, flexions et synonymes des termes inclus dans le dictionnaire général. Elle inclut la classe **dictionnaire** du MeSH.
Les attributs :
Les attributs **code** et **libellé** désignent le code du terme ainsi que son libellé.
Les attributs **alternative_lexicale** et **données_lexico_syntaxiques** renseignent les variations, flexions et synonymes du terme ainsi que les données lexicales et syntaxiques.

A.4 CIM10-Métatermes MeSH

F-MTI permet de retrouver, à partir d'une requête ou d'une phrase, des termes appartenant à différentes terminologies. Une méthode identique pourrait être utilisée dans le cadre du transcodage automatique, afin de déterminer, à partir d'un terme, les autres termes appartenant à d'autres terminologies auxquels il renvoit. Dans un deuxième temps (après avoir testé pour la CCAM voir section 5.8.1), nous avons testé cette hypothèse pour l'assignation de métatermes à la CIM10.

Les métatermes ont été définis manuellement par un expert (P. Massari) en utilisant la hiérarchie de la nomenclature. Pour chaque sous-chapitre de dernier niveau, il a été défini un ou plusieurs métatermes lorsqu'ils s'appliquaient aux codes sous-jacents. Dans un certain nombre de cas des métatermes ont été définis au niveau des codes, soit en complément, soit quand aucun n'était adapté à tous les codes d'un chapitre (voir figure A.6).

1)

Codes CIM10	Metaterme
A00	infectiologie
A00	bactériologie
A15	infectiologie
A15	pneumologie
A15	bactériologie

2)

Code CIM10	Metaterme
A15.5	otorhinolaryngologie

FIGURE A.6 – Assignation manuelle de métatermes aux codes CIM10

Automatiquement, nous avons utilisé le transcodage CIM10-MeSH (transcodage extrait du metathesaurus de l'UMLS [13]). Cette table permet de retrouver à partir d'un code CIM10 le ou les mots clefs MeSH supposés équivalents au terme CIM10.

Cette méthode est limitée puisque tous les codes CIM10 n'ont pas d'équivalent en MeSH. Seul 8.9% des codes CIM10 sont transcodables. Et à partir des relations termes MeSH - métatermes de la terminologie CISMeF, nous avons obtenu la liste des métatermes reliés à ces termes MeSH.

De la même facon nous avons calculé la précision et le rappel [Pereira07] (voir figure A.7). Seulement 110 métatermes ont été pris en compte.

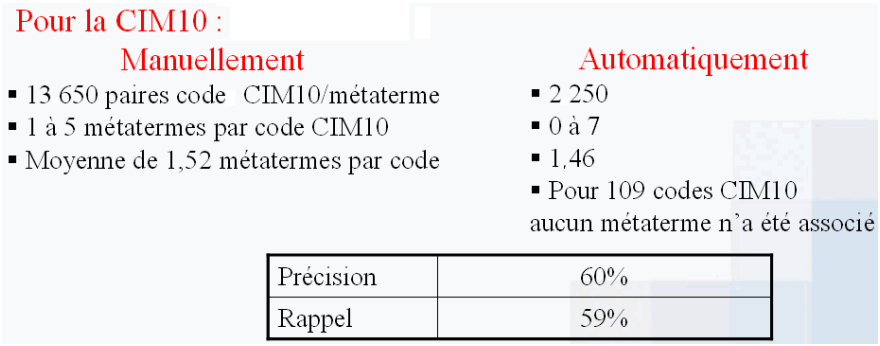

FIGURE A.7 – Résultats de la comparaison entre le transcodage manuel et automatique

A.5 Démonstration

Les boutons contextuels ont été créés et intégrés en environment de test au logiciel CDP2, logiciel du dossier patient électronique actuellement en place au CHU de Rouen [Massari00]. Ce logiciel présente les dossiers sous forme d'une arborescence événementielle (aux normes HISA [1] : Patients-Episodes-Séjours-Actes). Ils ont été programmés en VB (Visual Basic) language informatique utilisé dans CDP2. La recherche est dépendante du profil de l'utilisateur, on montre donc 3 exemples : connexion en tant que médecin, en tant qu'étudiant et en tant que patient, ceci à partir de la fiche des diagnostics CIM10 et de la fiche de synthèse.

Connexion en tant que médecin :

La connexion au logiciel se fait grâce à une fenêtre de connexion (voir figure A.8). Cette identification permet de connaître le profil de l'utilisateur (ici un médecin). Pour atteindre la fiche des diagnostics d'un patient, il faut tout d'abord sélectionner le service du patient (Dermatologie, Cardiologie...) puis le patient et le séjour d'intérêt.

Dès l'ouverture de la fiche des diagnostics de séjour (voir figure A.9), nous pouvons observer que le bouton CISMeF n'est pas apparu pour le diagnostic principal «choléra» ayant pour code CIM10 A00.1 car celui-ci n'a pas d'équivalence MeSH, il n'est donc pas trouvé dans les tables de transcodage. Le diagnostic relié, l'«agranulocytose» (D70), est lui, trouvé, son terme MeSH est «agranulocytose». Des recommandations à destination du médecin existent dans CISMeF (3 ressources trouvées).

1. Healthcare Information System Architecture

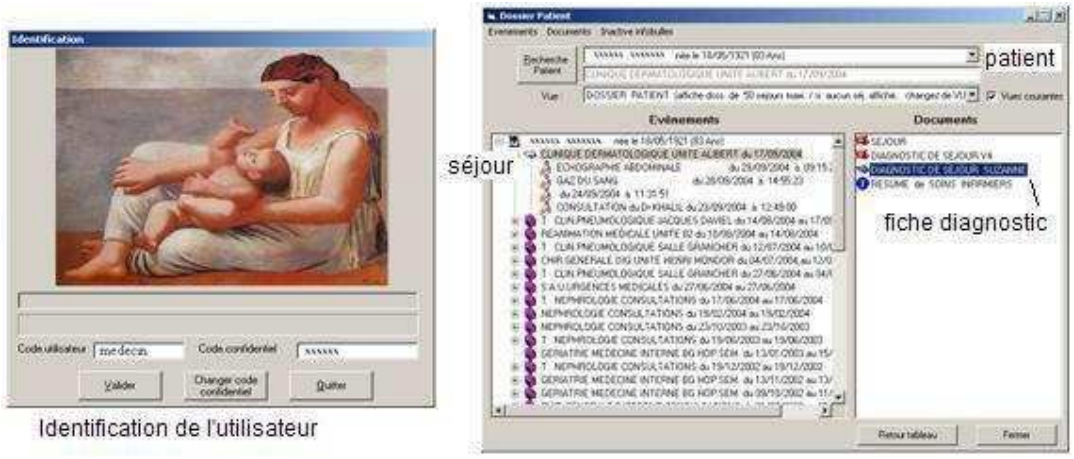

FIGURE A.8 – Ecran de connexion de l'utilisateur au logiciel CDP2 et accès aux diagnostics séjours d'un patient

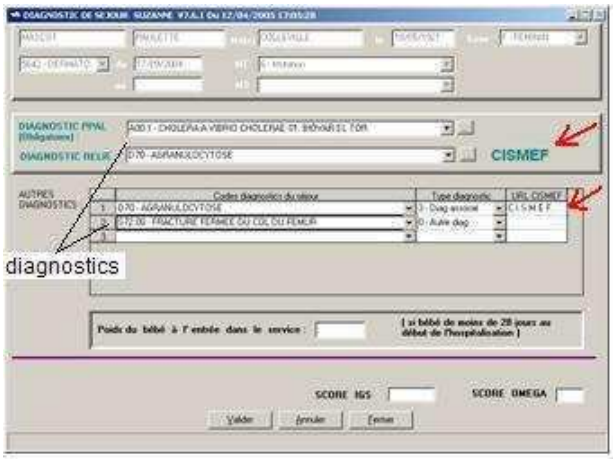

FIGURE A.9 – Codages CIM10 du compte-rendu d'hospitalisation avec le bouton CISMeF pour le diagnostic «agranulocytose»

Il en est de même pour les diagnostics secondaires. Lorsque que l'on appuit sur le bouton CISMeF la requête : «agranulocytose.mc et recommandations.tr» est lancée (mc=mots clés, tr=type de ressource) (voir figure A.10). «Agranulocytose» est le terme MeSH transcodé du terme CIM10 «Agranulocytose» codé D70. «recommandations» est sélectionné car l'utilisateur est un médecin. Une liste de documents appropriés et personnalisés est alors proposée sur le site.

Connexion en tant que patient :

Si l'utilisateur est un patient et qu'il consulte le même dossier et le même diagnostic, il sera dirigé vers la page CISMeF correspondant à la requête : «Agranulocytose.mc et patient.tr» (voir figure A.11).

Connexion en tant qu'étudiant :

FIGURE A.10 – Page CISMeF avec les listes des documents correspondant à la requête «Agranulocytose.mc et recommandations.tr»

FIGURE A.11 – Page CISMeF avec la liste des documents correspondant à la requête «Agranulocytose.mc et recommandations.tr»

De même, si l'utilisateur est un étudiant, et qu'il clique sur le bouton CISMeF à côté du diagnostic «troubles mentaux» dont le code est F99, la requête «troubles mentaux.mc et matériel pédagogique.tr» est lancée.

Le bouton de recherche d'information a aussi été développé pour la fiche de synthèse qui récapitule pour un patient l'ensemble de ses séjours à l'hôpital avec les codes diagnostics et actes médicaux associés (voir figure A.13).

Le deuxième bouton, quant à lui, permet d'accéder à d'autres sites de qualité en santé (voir page web figure A.14) classés par catégories et langues. Chaque lien vers un site spécialisé donne l'accès direct à la page contenant tous les documents pertinents correspondant au diagnostic d'intérêt, la requête ayant été traduite automatiquement.

FIGURE A.12 – Page CISMeF avec les listes des documents correspondant à la requête «troubles mentaux.mc et matériel pédagogique.tr»

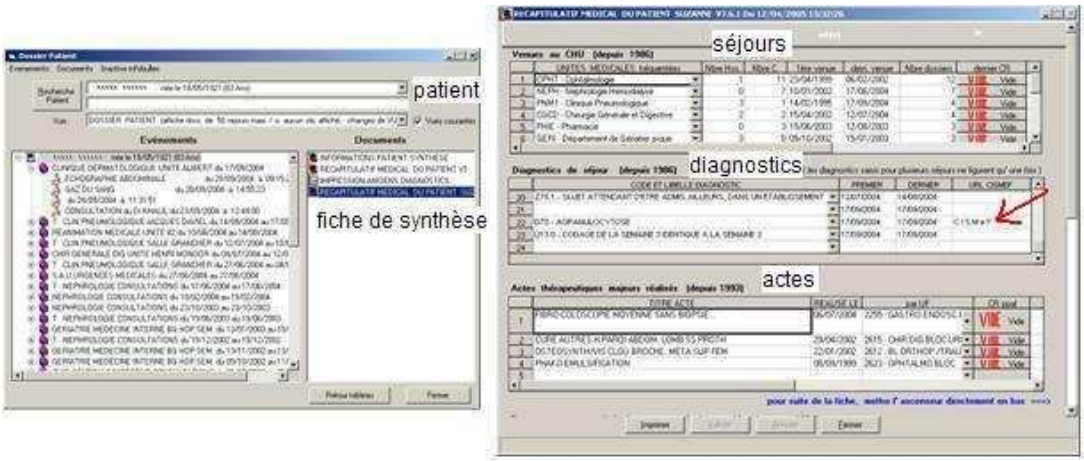

FIGURE A.13 – Accès à la fiche de synthèse appelée fiche récapitulative dans le DEP et à la fiche de synthèse avec le bouton CISMeF pour les diagnostics de séjour (tableau du milieu)

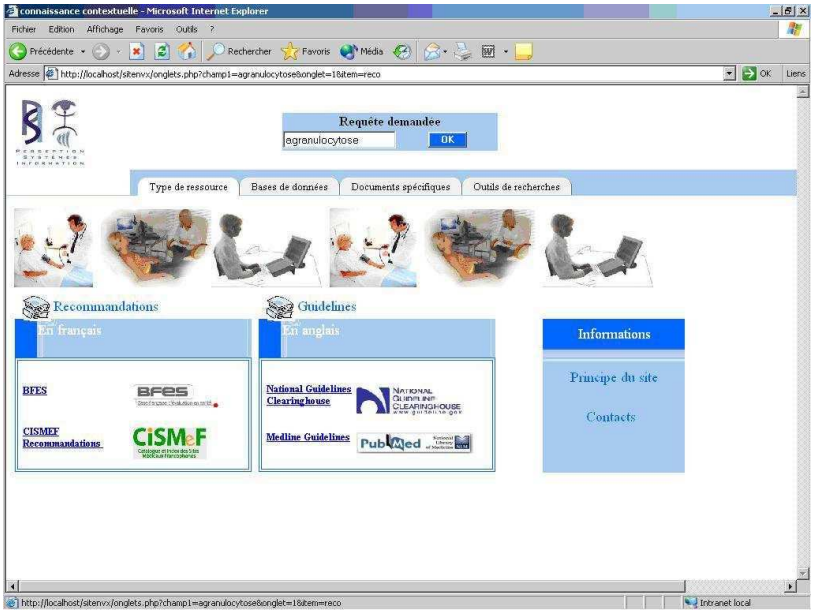

FIGURE A.14 – Page Web contenant les principaux sites de recherche en santé sur Internet

Bibliographie

[19550] Manuel de classement statistique international des maladies, traumatismes et causes de décès. Sixième révision des nomenclatures internationales de maladies et causes de décès adoptée en 1948, volume 1 & 2, index alphabétique. Technical report, Organisation mondiale de la santé Genève, 1950.

[19993] CIM-10 : Classification statistique internationale des maladies et des problèmes de santé connexes, dixième révision, volume 1. Technical report, Organisation mondiale de la santé Genève, 1993.

[Abdallah98] Abdallah I. Segmentation et codage de signaux de parole par critères entropiques. Ph.D. thesis, Université du Maine, 1998.

[Alper01] Alper B., Stevermer J., White D., Ewigman B. Answering family physicians' clinical questions using electronic medical databases. *J Fam Pract*, 2001 ; 50(11) : 960–965.

[Amaraki07] Amaraki E., Imai T., Kajino M., Miyo K., Ohe K. Statistical Selector of the Best Multiple ICD-coding Method. *Stud Health Technol Inform*, 2007 ; 645–649.

[Ame06] SNOMED CT Abstract Logical Model , Representational Forms. Technical report, College of American Pathologists, November 2006.

[Anderson01] Anderson J., Perez-Carballo J. The nature of indexing : how humans and machines analyze messages and texts for retrieval. Part 1 : Research, and the nature of human indexing. *Information Processing and Management*, 2001 ; 2(37) : 231–254.

[Aronson00] Aronson A., Bodenreider O., Chang F., Humphrey S., Mork J., Nelson S., Rindflesch T., Wilbur J. The NLM Indexing Initiative. *AMIA Annu Symp Proc*, 2000 ; 17–21.

[Aronson01] Aronson A.R. Effective mapping of biomedical text to the UMLS metathesaurus : the Metamap program. *AMIA Annu Symp Proc*, 2001 ; 17–21.

[Aronson04] Aronson A.R., Mork J.G., Gay C.W., Humphrey S.M., Rogers W.J. The nlm indexing initiative's medical text indexer. *Stud Health Technol Inform*, 2004 ; 268–272.

[Aronson07] Aronson A., Bodenreider O., Demner-Fushman D., Wah Fung K., Lee V., Mork J., Névéol A., Peters L., Rogers W. From Indexing the Biomedical Literature to Coding Clinical Text : Experience with MTI and Machine Learning Approaches. BIONLP, Biological, translational, and clinical language processing. 2007 105–12.

[Averbuch04] Averbuch M., Karson T., Ben-Ami B., Maimond O., Rokachd L. Context-Sensitive Medical Information Retrieval. *Stud Health Technol Inform*, 2004 ; 282–286.

[Avillach08a] Avillach P., Joubert M., Fieschi D. Improving the quality of the coding of primary diagnosis in standardized discharge summaries. *Health Care Management Science*, 2008 ; 147–151.

[Avillach08b] Avillach P., Joubert M., Fieschi M. Improving the quality of the coding of primary diagnosis in standardized discharge summaries. *Health Care Management Science*, 2008 ; .

[Bachimont00] Bachimont B. Ingénierie des Connaissances : Évolutions récentes et nouveaux défis, chapter Chapitre 19 : Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances, 305–323. L'Harmattan, 2000.

[Baeza-Yates99] Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval. CM Press Books, 1999.

[Baneyx06] Baneyx A. Construire une ontologie de la pneumologie : aspects théoriques, modèles et expérimentations. Ph.D. thesis, Université Pierre et Marie Curie - PARIS 6, 2006.

[Baud92] Baud R., Rassinoux A., Scherrer J. langugage processing and semantical representation of medical texts. *Methods Inf Med*, 1992 ; 31 : 117–25.

[Baud97] Baud R., Lovis C., Rassinoux A., Michel P., Scherrer J. Extracting knowledge from an international classification. Proceedings of MIE'97. IOS Press, 1997 .

[Bayes63] Bayes T. An essay towards solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London*, 1763 ; 53.

[Bell90] Bell T., Cleary J., Witten I. Text compression. *NJ : Prentice Hall*, 1990 ; .

[Berard-Dugourd89] Berard-Dugourd A., Fargues J., Landau M., Rogala J. Un système d'analyse de texte et de question/réponse basé sur les graphes conceptuels. *Informatique et Gestion des Unités de Soins, Paris : Springer-Verlag*, 1989 ; 1 :223–33.

[Bergman01] Bergman M.K. The Deep Web : Surfacing Hidden Value. *The Journal of Electronic Publishing*, 2001 ;

7, Issue 1 : http ://quod.lib.umich.edu/cgi/t/text/text–idx ?c=jep ;view=text ;rgn=main ;idno=3336451.0007.104.

[Berrut90] Berrut C. Indexing medical reports : The rime approach. *Inf Process Manage*, 1990 ; 26(1) : 93–109.

[Berthelot05] Berthelot G., Mazars P., Sanou M. Codage du dossier patient à usage médico-économique. Recension des outils, algorithmes d'optimisation économique. Master's thesis, Université Paris V, 2005.

[Bertrand93] Bertrand A. Compréhension et catégorisation dans une activité complexe : l'indexation de documents scientifiques. Ph.D. thesis, Université de Toulouse le Mirail., 1993.

[Bodenreider00] Bodenreider O. Using UMLS semantics for classification purposes. *AMIA Annu Symp Proc*, 2000 ; 86–90.

[Bouaud02] Bouaud J., Séroussi B., Dréau H., Falcoff H., Riou C., Joubert M., Simon C., Simon G., Venot A. ASTI, un système d'aide à la prescription médicamenteuse basé sur les guides de bonnes pratiques. *Informatique et Santé*, 2002 ; .

[Bouchet99] Bouchet C. Comment choisir un outil d'aide au codage. *Le magazine de la médecine électronique MEDCOST*, 1999 ; .

[Bourigault00] Bourigault D., Fabre C. Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de grammaire*, 2000 ; 25 : 131–151.

[Brainbridge96] Brainbridge M., Salmon P., Rappaport A., Hayes G., Williams J., Teasdale S. The Problem Oriented Medical Record - just a little more structure to help the world go round ? *Proceedings of the Annual Conference of The Primary Health Care Specialist Group of the British Computer Society*, 1996 ; http ://www.phcsg.org/main/pastconf/camb96/mikey.html.

[Bramsen06] Bramsen P., Deshpande P., Keok Lee Y., Barzilay R. Finding Temporal Order in Discharge Summaries. *AMIA Annu Symp Proc*, 2006 ; 81–85.

[Brill95] Brill E. Transformation-based error-driven learning and natural langage processing : A case study in part-of-speech tagging. *Computational Linguistics*, 1995 ; 21(4) : 543–65.

[Briquet07] Briquet D.E. L'interopérabilité sémantique au GHH. *Coder l'information médicale du Dossier de Santé Informatisé GDR STIC Santé Thème C*, 2007 ; .

[Burnage90] Burnage G. CELEX - A Guide for Users. *Nijmegen : Centre for Lexical Information, University of Nijmegen*, 1990 ; .

[Campbell97] Campbell K., Carpenter P., Sneiderman C.e.a. Phase II Evaluation of Clinical Coding Schemes : completeness, taxonomy,

mapping, definition and clarity. *J Am Med Inform Assoc*, 1997 ; 4 : 238–251.

[Cavazza92] Cavazza M., Doré L., Zweigenbaurn P. Model-based natural language understanding in medicine. *Stud Health Technol Inform*, 1992 ; 1356–1361.

[Chapman01] Chapman W., Bridewell W., Hanbury P., Cooper G., Buchanan B. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed inform*, 2001 ; 34 : 301–10.

[Chapman07] Chapman W., Dowling J., Chu D. ConText : an algorithm for identifying contextual features from clinical text. *Actes de BioNLP2007 : Biological, translational, and clinical language processing*, 2007 ; 81–88.

[Charlet06] Charlet J., Bachimont B., Jaulent M. Building medical ontologies by terminology extraction from texts : an experiment for the intensive care units. *Comput Biol Med*, 2006 ; 36(7-8) : 857–70.

[Chartron89] Chartron G., Dalbin G., Monteil M., Verillon M. Indexation manuelle et automatique : dépasser les oppositions. *Documentaliste*, 1989 ; 26(4-5).

[Chartron92] Chartron G. De l'information spécialisée à l'information élaborée : problèmes de modélisation. *8e congrès SFSIC*, 1992 ; 462.

[Chaumier92] Chaumier J., Dejean M. L'indexation assistée par ordinateur, principes et méthodes. *Documentaliste*, 1992 ; 29(1).

[Chevallier03] Chevallier J., Griesser J., Brunel L. Tothem, un outil d'aide au codage selon la CIM10. *EMOIS2003*, 2003 ; .

[Chiao04] Chiao Y. Extraction lexicale bilingue à partir de textes médicaux comparables : application à la recherche d'information translangue. Ph.D. thesis, Université Pierre et Marie Curie - Paris VI, 2004.

[Cimino97] Cimino J., G. E., Zeng Q. Supporting Infobuttons with Terminological Knowledge. *J Am Med Inform Assoc*, 1997 ; 4 (Suppl) : 528–532.

[Cimino06] Cimino J. Use, usability, usefullness, and impact of an infobutton manager. *AMIA Annu Symp Proc*, 2006 ; 151–5.

[Coret94] Coret A., Menon B., Schibler D., Terrasse C. Un système d'indexation structurée à l'INIST. *Documentaliste*, 1994 ; 31(3).

[Cori02] Cori M., Léon J. La constitution du TAL, Étude historique des dénominations et des concepts. *TAL*, 2002 ; 43(3) : 21–55.

[Covell85] Covell D., Uman G., Manning P. Information needs in office practice : are they being met ? *Ann Intern Med*, 1985 ; 103(4) : 596–9.

[Côté72] Côté R. From SNOP to SNOMED - A Challenge for the Medical Record Librarian. *Bulletin of the Canadian Association of Medical Record Librarians*, December 1972 ; 5,no1.

[Côté93] Côté R., Rothwell D., Patolay J., Beckett R., Brochu L., eds. The Systematized Nomenclature of Human and Veterinary Medicine : SNOMED International. Technical report, College of American Pathologists, 1993.

[Cuggia07] Cuggia M., Darmoni S., Garcelon N., Soualmia L., Bourde A. Doc'UMVF : two search tools to provide quality-controlled teaching resources in French to students and teachers. *Int J Med Inform*, 2007 ; 76, Number 5-6 : 357–362.

[Cutting04] Cutting D., Hatcher E., Gospodnetic O. Lucene in Action. Manning Publications, 2004.

[Darmoni98] Darmoni S., Leroux V., Daigne M., B. T., Santamaria P., Duvaux C. Critères de qualité de l'information de santé sur l'Internet. *Santé et Réseaux Informatiques Informatique et Santé Springer Verlag France*, 1998 ; 162–74.

[Darmoni02] Darmoni S., Thirion B., Platel S., Douyère M., Mourouga P., Leroy J. CISMeF-patient : a French counterpart to MEDLINEplus. *J Med Libr Assoc*, 2002 ; 90 : 248–253.

[Darmoni03a] Darmoni S.J., Amsallem E., Haugh M., Lukacs B., Leroux V., Thirion B., Weber J., Boissel J.P. Level of evidence as a future gold standard for the content quality of health resources on the internet.. *Methods Inf Med*, 2003 ; 42 : 220–225.

[Darmoni03b] Darmoni S.J., Jarrousse E., Zweigenbaum P., Le Beux P., Namer F., Baud R., Joubert M., Vallée H., Côté R.A., Buemi A., Bourigault D., Recource G., Jeanneau S., Rodrigues J.M. VUMeF : extending the French involvement in the UMLS Metathesaurus. *AMIA Annu Symp Proc*, 2003 ; 824.

[Darmoni07] Darmoni S., Thirion B., Ionut-Florea F., Rogazan A., Letord C., Kerdelhué G., Dacher J. Affiliation of a resource type to a MeSH term in a quality-controlled health gateway. *Stud Health Technol Inform*, 2007 ; .

[Darmoni08] Darmoni S., Pereira S., Névéol A., Massari P., Dahamna B., Letord C., Kedelhué G., Piot J., Derville A., Thirion B. French Infobutton : an academic and... business perspective. *AMIA Annu Symp Proc*, 2008 ; en cours de publication.

[Deerwester90] Deerwester S., al. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 1990 ; 41 : 391–407.

[Degoulet98] Degoulet P., fieschi M. Informatique médicale. 1998.

[Dekkers03] Dekkers M., Weibel S. State of the Dublin Core Metadata Initiative. *D-Lib Mag*, 2003 ; v9 n40.

[Del Fiol06] Del Fiol G., Rocha R., Clayton P. Infobuttons at Intermountain Healthcare : Utilization and Infrastructure. *AMIA Annu Symp Proc*, 2006 ; 180–4.

[Del Fiol07] Del Fiol G., Haug P. Use of Classification Models Based on Usage Data for the Selection of Infobutton Resources. *AMIA Annu Symp Proc*, 2007 ; 171–5.

[Deyo94] Deyo R., Taylor V., Diehr P., Conrad D., Cherkin D., Ciol M., Kreuter W. Analysis of automated administrative and survey databases to study patterns and outcomes of care. *Spine*, 1994 ; 19 : 2083S–2091S.

[Diosan08] Diosan L., Rogozan A., Pécuchet J. Information Retrieval in Biomedicine : Natural Language Processing for Knowledge Integration, chapter Automatic Alignment of Medical Terminologies with General Dictionaries for an Efficient Information Retrieval. IGI publisher, Ed., 2008.

[Dister97] Dister A. Champs Linguistiques, chapter Problématique des fins de phrase en traitement automatique du francais. Duculot, 1997.

[Déjean05] Déjean H., Gaussier E., Renders J., Sadat F. Automatic processing of multilingual medical terminology : applications to thesaurus enrichment and cross-language retrieval. *Artif Intell Med*, 2005 ; 33 : 111–124.

[Doré92] Doré L., Cavazza M., Zweigenbaum P., J.F. B. Analyse pragmatique pour la compréhension de comptes rendus d'hospitalisation. *Informatique et Santé, Paris, Springer-Verlag France*, 1992 ; 5 : 139–152.

[Doré95] Doré L., Lavril M., Jean F., Degoulet P. An object oriented computer-based patient record reference model. *Proc Annu Symp Comput Appl Med Care*, 1995 ; 377–81.

[Douyère04] Douyère M., Soualmia L., Névéol A., Rogozan A., Dahamna B., Leroy J., Thirion B., Darmoni S. Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. *Health Info Libr J*, Dec 2004 ; 21(4) : 253–261.

[Dufour05] Dufour J. Contribution à l'amélioration de la décision : Intégration des guides de bonnes pratiques cliniques informatisés dans la pratique médicale. Ph.D. thesis, Université de la Méditérranée, 2005.

[Dutoit00] Dutoit D. Quelques opération texte-sens et texte-sens-texte utilisant une sémantique linguistique universaliste a priori. Ph.D. thesis, Université de Caen, 2000.

[Elhadad07] Elhadad N., Sutaria K. Mining a Lexicon of Technical Terms and Lay Equivalents. *Proceedings of BIONLP*, 2007 ; 49–56.

[Elisabeth02] Elisabeth B., Oystein N., Anders G. Ontologies for knowledege representation in a computer-based patient record. *14th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'02)*, 2002 ; 114.

[Elkin05] Elkin P., Brown S., Bauer B., Husser C., Carruth W., Bergstrom L., Wahner-Roedler D. A controlled trial of automated classification of negation from clinical notes. *BMC Medical Informatics and Decision Making*, 2005 ; 5–13.

[Ely05] Ely J., Osheroff J., Chambliss M., Ebell M., Rosenbaum M. Answering physician's clinical questions : obstacles and potential solutions. *J Am Med Inform Assoc*, 2005 ; 12(2) : 217–24.

[Falcoff99] Falcoff H. Le dossier orienté problème existe, je l'ai rencontré. *Informatique et Santé*, 1999 ; 11.

[Fayet-Scribe97] Fayet-Scribe S. Chronologie des supports, des dispositifs et des outils de repérage de l'information. 1997.

[Fieschi05] Fieschi M. Vers le dossier médical personnel. Les données du patient partagées : un atout à ne pas gâcher pour faire évoluer le système de santé. *Revue Droit Social*, 2005 ; 80–90.

[Fisher83] Fisher J., Rey R. De l'origine et de l'usage des termes taxinomie-taxonomie. *Documents pour l'histoire du vocabulaire scientifique*, 1983 ; V : 97–113.

[Flannery95] Flannery M. Cataloging Internet resources. *Bull Med Libr Assoc*, 1995 ; 83(2) : 211–5.

[Florea07a] Florea F. Indexation et recherche d'information combinée texte et image dans le catalogue de santé CISMeF. Ph.D. thesis, INSA de Rouen, 2007.

[Florea07b] Florea F., Buzuloiu V., Rogozan A., Bensrhair A., Darmoni S. automatic Image Annotation Combining the Content and the Context of Medical Images. *Proc International Symposium on Signals, Circuits and Systems ISSCS 2007*, 2007 ; 1 : 1–4.

[Folch08] Folch H., Habert B. Proximités de comportement syntaxique entre les mots. *JADT2008*, 2008 ; 295.

[Fontelo07] Fontelo P., Liu F., Leon S., Anne A., Ackerman M. PICO Linguist and BabelMeSH : Development and Partial Evaluation of Evidence-based Multilanguage Search Tools for MEDLINE/PubMed. *Stud Health Technol Inform*, 2007 ; 817–21.

[Friburger00] Friburger N., Dister A., Maurel D. Améliorer le découpage en phrase sous INTEX. *In Actes des troisièmes journées Intex, Revue Informatique et Statistiques dans les sciences humaines 36*, 2000 ; 1-4 : 181–200.

[Friedman04] Friedman C., Shagina L., Lussier Y., Hripcsak G. Automated Encoding of Clinical Documents Based on Natural Language Processing. *J Am Med Inform Assoc*, 2004 ; 11 : 392–402.

[Fujii07] Fujii H., Yamagishi H., Ando Y., Tsukamoto N., Kawaguchi O., Kasamatsu T., Kurosaki K., Osada M., Kaneko H., Kubo A. Structuring of Free-Text Diagnostic Report. *Stud Health Technol Inform*, 2007 ; 669–85.

[Fung05] Fung K., Bodenreider O. Utilizing the UMLS for Semantic Mapping between Terminologies. *AMIA Annu Symp Proc*, 2005 ; .

[Funk83a] Funk M., Reid C., McGoogan L. Indexing consistency in MEDLINE. *Bull Med Libr Assoc*, 1983 ; 2(71) : 176–183.

[Funk83b] Funk M., Reid C., McGoogan L. Indexing consistency in MEDLINE. *Bull Med Libr Assoc*, 1983 ; 176–83.

[Gaudinat02] Gaudinat A., Boyer C., Baujard V., Ruch P. 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*, 2002 ; .

[Gaussier99] Gaussier E. Unsupervised learning of derivationnal morphology from inflectional lexicons. ACL Workshop on Unsupervised Methods in Natural Langage Learning. 1999 .

[Gay05] Gay C., Kayaalp M., Aronson A. Semi-Automatic Indexing of Full Text Biomedical Articles. *AMIA Annu Symp Proc*, 2005 ; 271–5.

[Gehanno07] Gehanno J., Thirion B., Darmoni S. Evaluation of Metaconcepts for Information Retrieval in a Quality-Controlled Health Gateway. *AMIA Annu Symp Proc*, 2007 ; 269–273.

[Giorgi05] Giorgi R., Payan J., Gouvernet J. RSURV : a function to perform relative survival analysis with S-PLUS or R. *Comput Biol Med*, 2005 ; .

[GIP-DMP07] GIP-DMP. Dossier Médical Personnel : premiers éléments de l'étude auprès des acteurs de la phase d'expérimentation. *rapport présenté au COR*, 30 janvier 2007 ; .

[Goldin03] Goldin I., Chapman W. Learning to detect negation with 'not' in medical texts. *Proc Workshop on Text Analysis and Search for Bioinformatics, ACM SIGIR*, 2003 ; .

[Grabar00] Grabar N., Zweigenbaum P. A genral method for sifting linguistic knowledge from structured terminologies. *J Am Med Inform Assoc*, 2000 ; 7(suppl) : 310–4.

[Gutnik07] Gutnik L., Collins S., Currie L., Cimino J., Patel V. Infobuttons : a study of usability. *Stud Health Technol Inform*, 2007 ; 1481.

[Halleb97] Halleb M., Lelu A. Hypertextualisation automatique multilingue à partir des fréquences des n-grammes. *Hypertextes et hypermédias*, 1997 ; 1(2-3-4) : 275–287.

[Happe03] Happe A., Pouliquen B., Burgun A., Cuggia M., Le Beux P. Automatic concept extraction from spoken medical reports. *Int J Med Inform*, 2003 ; 70(2-3) : 255–63.

[Hathout02a] Hathout N., Namer F., Dal G. An experimental constructional database : the MorTAL project. *Many morphologies, Cambridge Mass, Cascadilla Press*, 2002 ; 178–209.

[Hathout02b] Hathout N., Namer F., Dal G. An experimental constructional database : The Mortal project. *Cascadilla Press*, 2002 ; 178–209.

[Hoquet05] Hoquet T., al. Linné et la classification des plantes. Les fondements de la botanique. Vuibert, Paris, 2005 .

[Humphrey06] Humphrey S., Rogers W., K. K., D. D.F., Rindflesch T. Word Sense Disambiguation by Selecting the Best Semantic Type Based on Journal Descriptor Indexing : Preliminary Experiment. *Journal of the american society or information science and technology*, 2006 ; 57(1) : 96–113.

[Jacquemart03] Jacquemart P., Zweigenbaum P. Towards a medical question-answering system : a feasibility study. *Stud Health Technol Inform*, 2003 ; 95 : 463–468.

[Jacquemart05] Jacquemart P. Accès à l'information textuelle médicale : de la recherche d'information aux systèmes de question réponse. Ph.D. thesis, Université de Paris 5, 2005.

[Jacquemin97] Jacquemin C. Guessing morphology from terms and corpora. *Actes 20th ACM SIGIR*, 1997 ; 156–67.

[Joachims98] Joachims T. Text categorization with Support Vector Machines : Learning with many relevant features. *Proceedings of the Tenth European Conference on Machine Learning (ECML'98), Springer Verlag*, 1998 ; 137–142.

[Jollis93] Jollis J., Ancukiewicz M., De Long E., Pryor D., Muhlbaier L., Mark D. Discordance of databases designed for claims payment versus clinical information systems. Implications for outcomes research. *An Intern Med*, 1993 ; 119 : 855–857.

[Joubert02] Joubert M., S. A., Fieschi D., Fieschi M. ARIANE : un moteur de recherche de deuxième génération dans le domaine de la santé. *Informatique et Santé*, 2002 ; .

[Joubert03] Joubert M., Dufour J., Aymard S., Falco L., Staccini P., Fieschi M. Le Projet CoMeDIAS : Accès à des Bases de Données

Hétérogènes au Moyen de Services Internet. *Informatique et Santé*, 2003 ; .

[Joubert07a] Joubert M., Gaudinat A., Boyer C., Fieschi M., members H.F.C. WRAPIN : a tool for patient empowerment within EHR. *Stud Health Technol Inform*, 2007 ; 129 : 147–51.

[Joubert07b] Joubert M., Le Beux P., Darmoni S., Fieschi M. Evaluation de l'indexation des documents de l'Université Médicale Virtuelle Francophone. *IPM*, 2007 ; .

[Keselman07] Keselman A., Slaughter L., Smith C., Hyeoneui K., Divita G., Browne A. Towards Consumer-Friendly PHRs : Patient's Experience with Reviewing their Health Records. *AMIA Annu Symp Proc*, 2007 ; 399–403.

[Kim01] Kim W.and Aronson A., Wilbur W. Automatic mesh term assignment and quality assessment. *AMIA Annu Symp Proc*, 2001 ; 319–323.

[Kolher05] Kolher F., Toussaint E. La T2A, les pôles et la contractualisation interne. Quels modèles en hospitalisation de court séjour ? *Journées Francophones d'Informatique médicale*, 2005 ; .

[Lamberts87] Lamberts H., Wood M. International Classification of Primary Care (ICPC). Oxford University Press, 1987.

[Lamy06] Lamy J. Conception et évaluation de méthodes de visualisation des connaissances médicales : mise au point d'un langage graphique et application aux connaissances sur le médicament. Ph.D. thesis, Université Paris 6, 2006.

[Lancaster91] Lancaster F. Indexing and abstracting in theory and practice. Champaign, IL, 1991.

[Lefèvre00] Lefèvre P. La recherche d'information du texte intégral au thésaurus. Hermes Science, sept 2000.

[Letord] Letord C., Sakji S., Pereira S., Dahamna B., Kergourlay I., Darmoni S. Un portail d'information sur le médicament en Europe Drug Information Portail in Europe.

[Levenshtein66a] Levenshtein V. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 1966 ; 10 : 707–710.

[Levenshtein66b] Levenshtein V.I. Binary codes capable of correcting deletions, insertions and reversals. *B Sov Phys Dokl*, 1966 ; 6 : 707–710.

[Lewandowski08] Lewandowski E. De nouveaux outils informatiques au service du PMSI. *Paroles d'expert M DH Magazine*, 2008 ; 118 : 67.

[Li07] Li J., Cimino J. Auditing Dynamic Links to Online Information Resources. *AMIA Annu Symp Proc*, 2007 ; 448–52.

[Lin98] Lin D. An information-theoretic definition of similarity. *In Proc Int Conf on Machine Learning*, 1998 ; 296–304.

[Lindberg90] Lindberg D., Humphreys B. The UMLS Knowledge Sources : Tools for Building Better User Interfaces. *Proceedings of the 14th annual SCAMCANDEEE Computer Society Press*, 1990 ; 121–125.

[Liu06] Liu F., Fontelo P., Ackerman M. BabelMeSH : Developpement of a Cross-Language Tool for MEDLINE/Pubmed. *AMIA Annu Symp Proc*, 2006 ; 1012.

[Loisel07] Loisel A., Chaignaud N., Kotowicz J. Designing a Human-Computer Dialog System for Medical Information Search. *Proc IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology Workshops*, 2007 ; 350–353.

[Lovins68] Lovins J. Developpement of a stemming algorithm. *Mechanical Translation and Computional Linguistics*, 1968 ; 11 : 22–31.

[Lovis96] Lovis C. Codage médico-économique des diagnostics et procédures. Ph.D. thesis, Université de Genève, 1996.

[Lovis98] Lovis C., Raud R., Rassinoux A., Michel P., J.R. S. Medical dictionaries for patient encoding systems : a methodology. *Artif Intell Med*, 1998 ; 14 : 201–14.

[Lu05] Lu W., Lin S., Chan Y., Chen K. Semi-automatic construction of the Chinese-English MeSH using web-based term translation method. *AMIA Annu Symp Proc*, 2005 ; 475–9.

[Luhn58] Luhn H. The automatic creation of literature abstracts. *IBM Journal of research and development*, 1958 ; 2 : 159–165.

[Lundsgaarde81] Lundsgaarde H., Fisher P., Steele D. Human problems in computerized medicine. *University of Kansas Publications in Anthropology*, 1981 ; 12.

[Massari00] Massari P., Fuss J. Dossier patient informatisé du CHU de Rouen : migration des anciennes applications vers C-PAGE Dossier Patient. *Gestions hospitalières*, 2000 ; 395 : 316–320.

[Massari08] Massari P., Pereira S., Thirion B., Derville A., Darmoni S. Use Of Super-Concepts To Customize Electronic Medical Records Data Display. *Stud Health Technol Inform*, 2008 ; 136 : 845–850.

[Maviglia06] Maviglia S., Yoon C., Bates D., Kuperman G. KnowledgeLink : Impact of context-sensitive information retrieval on clinician's information needs. *J Am Med Inf Assoc*, 2006 ; 13 : 67–73.

[Mayer03] Mayer M., Darmoni S., Fiene M., Köhler C., Roth-Berghofer T., Eysenbach G. MedCIRCLE : collaboration for Internet rating, certification, labelling and evaluation of health information on

the World-Wide-Web. *Stud Health Technol Inform*, 2003 ; 95 : 667–672.

[Merabti08a] Merabti T., Pereira S., Lecroq T., Joubert M., Darmoni S. Inheritance of SNOMED CT Relations between concepts to two Health Terminologies (SNOMED International and ICD-10). *Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED)*, 2008 ; 118.

[Merabti08b] Merabti T., Pereira S., Letord C., Lecroq T., Dahamna B., Joubert M., Darmoni S. Searching Related Resources in a Quality-Controlled Health Gateway : a feasibility Study. *Stud Health Technol Inform*, 2008 ; 136 : 205–210.

[Min06] Min Z., Baofen D., Weeber M., Van Ginneken A. Mapping OpenSDE Domain Models to SNOMED CT. *Methods Inf Med*, 2006 ; 4–9.

[Misset05] Misset B., Metais E., Nakache D., Dumont S., De Lassence A., Darmont M., Garrouste Orgeas B., Mourvillier M., Adrie C., Pease S., Costa de Beauregard M.A., Stocco C. Reproductibilité du codage. *in 33ème congrès de la SRLF (Société de Réanimation de Langue Française), Cnit Paris*, 2005 ; .

[Molino85] Molino J. Où en est la morphologie ? *Langages*, 1985 ; 78 : 5–40.

[Moreau] Moreau F., Claveau V., Pascale S. Intégrer plus de connaissances linguistiques en recherche d'information peut-il augmenter les performances des systèmes ? *Actes de la 4ème Conférence en recherche d'informations et applications, (CORIA'07), St Etienne* ; .

[Nachimuthu07] Nachimuthu S., Lau L. Practical issues in using SNOMED CT as a reference terminology. *Stud Health Technol Inform*, 2007 ; 129(Pt 1) : 640–4.

[Nakache05] Nakache D., Metais E., Timsit J. Evaluation and NLP. *proceedings of DEXA Database and Expert System Application*, 2005 ; 626–632.

[Nakache07] Nakache D. Extraction automatique de diagnostics à partir de comptes rendus médicaux textuels. Ph.D. thesis, Conservatoire des Arts et Métiers, 2007.

[Namer00a] Namer F. FLEMM : un analyseur flexionnel du francais à base de règles. *Traitement Automatique des Langues*, 2000 ; 41(2) : 523–47.

[Namer00b] Namer F., Dal G. GéDériF : automatic generation and analysis of morphologically constructed lexical resources. *Proceedings of the Second International Conference on Language, Resources and Evaluation*, 2000 ; 1447–1454.

[Néveol05] Néveol A., Mork J., Aronson A., Darmoni S. Evaluation of French and English MeSH Indexing Systems with a Parallel Corpus. *AMIA Annu Symp Proc*, 2005 ; 565–9.

[Néveol06] Néveol A., Pereira S., Soualmia L., Thirion B., Darmoni S. A method of cross-lingual consumer health information retrieval. *Stud Health Technol Inform*, 2006 ; 124 : 601–608.

[Néveol07] Néveol A., Shooshan S., Humphrey S., Rindflesh T., Aronson A. Multiple approaches to fine-grained indexing of the biomedical literature. *Pacific Symposium on Biocomputing*, 2007 ; 12 : 292–303.

[Névéol05a] Névéol A. Automatisation des tâches documentaires dans un catalogue de santé en ligne. Ph.D. thesis, INSA de Rouen, 2005.

[Névéol05b] Névéol A., Mork J., Aronson A., Darmoni S. Evaluation of French and English MeSH indexing systems with a parallel corpus. *AMIA Annu Symp Proc*, 2005 ; 565–569.

[Névéol06] Névéol A., Zeng K., Bodenreider O. Besides Precision & Recall : Exploring Alternative Approaches to Evaluating an Automatic Indexing Tool for MEDLINE. *AMIA Annu Symp Proc*, 2006 ; 589–593.

[Névéol07a] Névéol A., Mork J., Aronson A. Automatic Indexing of Specialized Documents : Using Generic vs. Domain-Specific Document Representations. *BIONLP : Biological, translational and clinical language processing*, 2007 ; 183–190.

[Névéol07b] Névéol A., Pereira S., Kerdelhué G., Dahamna B., Joubert M., Darmoni S. Evaluation of a simple method for the automatic assignment of MeSH descriptors to health resources in a French online catalogue. *Stud Health Technol Inform*, 2007 ; 129 : 407–11.

[Odell18] Odell M., Russell C. The soundex coding system. *US Patents*, 1918 ; .

[OFS06] OFS O.f.d.l.s. Définition en entités et relations de la CIM10. *La CIM10 par l'OFS*, 2006 ; .

[Paice96] Paice C. Method for evaluation of stemming algorithms based on error counting. *Journal of the American Society for Information Science*, 1996 ; 47 : 632–349.

[Pappa04] Pappa A., Bernard G., Ouekeradi H. Détection automatique de frontières de phrases - Un système adaptatif multi-langues. *Permanent online Journal of Information and Communication Technologies, ISDM (Informations, Savoirs, Décisions et Médiations)*, 2004 ; 13.

[Paternostre02] Paternostre M., Francq P., Lamoral J., Wartel D., M. S. Carry, un algorithme de désuffixation pour le francais. *urlhttp ://siculbacbe/research/is/galilei/carry*, 2002 ; .

[Patriarche05] Patriarche R., Gedzelman S., Diallo G., Bernhard D., Cyr-Gabin B., Ferriol S., Girard A., Mouries M., Palmer P., Simonet A., Simonet M. Noesis Annotation Tool : un outil pour l'annotation textuelle et conceptuelle de documents. *Ingénierie des Connaissances IC'2005*, 2005 ; 15–16.

[Pereira] Pereira S., Massari P., Darmoni S. Evaluation of a method for automatic mapping between French procedure terminology (CCAM) and MeSH. Non publié, mais sera soumis dans un prochain congrès.

[Pereira05] Pereira S. Evaluation de plusieurs méthodes d'optimisation du codage médico-économique. Master's thesis, Université Paris 5, 2005.

[Pereira07] Pereira S., Massari P., Joubert M., Darmoni S. Utilisation de métatermes pour la recherche d'information dans les dossiers médicaux. *In Actes des journées Francophones d'Informatique Médicale*, 2007 ; .

[Pereira08a] Pereira S., Massari P., Buemi A., Dahamna B., Serrot E., Joubert M., Darmoni S. Evaluation of two French SNOMED indexing systems with a parallel corpus. *Poster 3rd international conference on Knowledge Representation in Medicine (KR-MED)*, 2008 ; .

[Pereira08b] Pereira S., Massari P., Joubert M., Serrot E., Darmoni S. Exploring Multi-terminology Indexing of Discharge Summaries. *Poster MIE2008*, 2008 ; .

[Pereira08c] Pereira S., Névéol A., Kerdelhué G., Serrot E., Joubert M., Darmoni S. Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue. *AMIA Annu Symp Proc*, 2008 ; 586–590.

[Petitpierre94] Petitpierre D., Russel G. Mmorph - The Multext Morphology Program. *Technical Report, ISSCO*, 1994 ; .

[Pillou06] Pillou J. Tout sur les Réseaux et Internet. Dunot, 2006.

[Pisani08] Pisani F., Piotet D. Comment le web change le monde : L'alchimie des multitudes. VILLAGE MONDIAL, 2008.

[Plovnick04] Plovnick R., Zeng Q. Reformulation of Consumer Health Queries with Professional Terminology : A Pilot study. *J Med Internet Red*, 2004 ; 6(3) : e27.

[Porter80] Porter M. An algorithm for suffix stripping. *Program*, 1980 ; 14 (3) : 130–137.

[Pouliquen02] Pouliquen B. Indexation de textes médicaux par indexation de concepts, et ses utilisations. Ph.D. thesis, Université Rennes 1., 2002.

[Prieur07] Prieur E. Méthodes et structures de données pour l'indexation et la détection de répétitions dans les séquences biologiques : les vecteurs de suffixes. Ph.D. thesis, Université de Rouen, 2007.

[Rector03] Rector A., Rogers J., Zantra P., Van der Haring E. OpenGalen : Open Source Medical Terminology and Tools. *AMIA Annu Symp Proc*, 2003 ; 982.

[Roche05] Roche C. Terminologie et ontologie. *Language*, 2005 ; 157.

[Rodrigues05] Rodrigues J., Trombert Paviot B., Martin C., P. V. Integrating the Modelling of EN 1828 and Galen CCAM Ontologies with Protégé : towards a Knowledge Acquisition Tool for Surgical Procedures. *Stud Health Technol Inform*, 2005 ; 116 : 767–72.

[Rolling80] Rolling L. Indexing consistency, quality and efficiency. *Information Processing and Management*, 1980 ; 69–77.

[Rosse03] Rosse C., Mejino J.J. A reference ontology for biomedical informatics : the Foundational Model of Anatomy. *J Biomed Inform*, 2003 ; 36(6) : 478–500.

[Roussey01] Roussey C. Une méthode d'indexation sémantique adaptée aux corpus multilingues. Ph.D. thesis, INSA de Lyon, 2001.

[Ruch03] Ruch P., Baud R., Geissbühler A. Learning-free text categorization. *Proc AIME 2003 - LNAI 2780*, 2003 ; 119–204.

[Ruch04] Ruch P. Query translation by Text Categorization. *Proceedings of the 20th international conference on Computational Linguistics COLING*, 2004 ; .

[Sager95] Sager N., Lyman M., Nhhn N., Tick L. Medical language processing : Applications to patient data representation and automatic encoding. *Methods of Information in Medicine*, 1995 ; 34 : 140–146.

[Salton73] Salton G. Experiments in multilingual information retrieval. *Information Processing Letters*, 1973 ; 2(1) : 6 – 11.

[Salton83] Salton G., M.J. M. Introduction to modern information retrieval. 1983.

[Salton89] Salton G. Automatic text processing : The transformation, analysis, and retrieval of information by computer. *Reading, MA : Addison-Wesley*, 1989 ; .

[Schank81] Schank R., Riesbeck C., eds. Inside Computer Understanding. *Hillsdale,New Jersey : Lawrence ErIbaurn Associates*, 1981 ; 259–307.

[Schatz97] Schatz B. Information Retrieval in Digital Libraries : Bringing Search to the Net. *Science*, 1997 ; 275 : 327–34.

[Seroussi04] Seroussi B., Bouaud J., Dreau H., Falcoff H., Venot A. Modalités d'interaction avec des systèmes d'aide à la décision

médicale par alerte ou à la demande pour délivrer des recommandations : une étude préliminaire dans le cadre de la prise en charge de l'hypertension. *IC 2004, 15es journées francophones d'ingénierie des connaissances*, 2004 ; 65–76.

[SFMG96] SFMG. Dictionnaire des Résultats de concultation. *Doc Rech Mec Gen*, 1996 ; 47–48.

[Sherertz90] Sherertz D., Olson N., Tuttle M., ErIbaum M. Source Inversion and Matching in the UMLS Metathesaurus. *Proceedings of the 14th annual SCAMC, IEEE Computer Society Press*, 1990 ; 141–145.

[Shortliffe76] Shortliffe E. Computer-Based Medical Consultation : MYCIN. *New York Elsevier*, 1976 ; .

[Silberztein93] Silberztein M. Dictionnaires électroniques et analyse automatique de textes : le système INTEX. Masson, Paris, 1993.

[Silberztein04] Silberztein M. NooJ : an oriented object approach. J. Royauté, M. Silberztein, editors, INTEX pour la Linguistique et le Traitement Automatique des Langues. Presses Universitaires de Franche-Comté, 2004 .

[Soergel88] Soergel D. Indexing and retrieval performance : the logical evidence. *Journal of American Society for Information Science*, 1988 ; 39(3) : 161–176.

[Soualmia03] Soualmia L., Barry C., Darmoni S. Knowledge-Based Query Expansion over a Medical Terminology Oriented Ontology. Artif Intell Med : 9th Conference on Artificial Intelligence in Medicine in Europe, AIME. 2003 209–213.

[Soualmia04] Soualmia L. Etude et Evaluation d'Approches Multiples d'Expansion de Requêtes pour une Recherche d'Information Intelligente : Application au Domaine de la Santé sur Internet. Ph.D. thesis, INSA de Rouen, 2004.

[Spackman97] Spackman K., Campbell K., Côté R. SNOMED RT : a reference terminology for health care. *AMIA Annu Symp Proc*, 1997 ; 640–4.

[Sundvall07] Sundvall E., Nyström M., Forss M., Chen R., Peterson H., Ahlfeldt H. Graphical Overview and Navigation of Electronic Health Records in a Prototyping Environmen Using Google Earth and openEHR Archetypes. *Stud Health Technol Inform*, 2007 ; 1043–7.

[Thirion98] Thirion B., Darmoni S. Les sites médicaux francophones sur Internet : le devoir d'ingérence des bibliothèques. *Bulletin des Bibliothèques de France*, 1998 ; 42–5.

[Thirion04] Thirion B., Douyère M., Soualmia L., Dahamna B., Leroy J., Darmoni S. Metadata element sets in the CISMeF Quality-

Controlled Health Gateway. *International Conference on Dublin Core and Metadata Applications*, 2004 ; .

[Thirion07] Thirion B., Pereira S., Névéol A., Dahamna B., Darmoni S. French MeSH Browser : a cross-language tool to access MEDLINE/PubMed. *AMIA Annu Symp Proc*, 2007 ; 1132.

[Tse03] Tse T., Soergel D. Exploring medical expressions used by consumers and the media : An emerging view of consumer health vocabularies. *AMIA Annu Symp Proc*, 2003 ; 674–98.

[vanDijk90] van Dijk T., Kintsch W. Strategies of Discourse Comprehension. *New York : Academic Press*, 1990 ; 664.

[vanRijsbergen79] van Rijsbergen C. Information Retireval. Butterworths. *London*, 1979 ; .

[Vapnik95] Vapnik V. The Nature of Statistical Learning Theory. *Springer*, 1995 ; .

[Voorhees03] Voorhees E. Evaluatiing the evaluation : Edmonton. *Proceesings of HLT-NAACL*, 2003 ; 181–188.

[Wall01] Wall L. Programmation en Perl, 3e édition. Broché, 2001.

[Weed68] Weed L. Medical records that guide and teach. *N Engl J Med*, 1968 ; 10(2)278(12) : 652–7.

[Wehrli88] Wehrli E. Medical linguistics software tools for prospective production. *In : Scherrer JR, Côté RA & Mandil SH, eds Computerized natural medical language processing for knowledge representation Amsterdam : Elsevier Science*, 1988 ; 67–72.

[WHO] WHO W.H.O. International Classification of Functionning, Didability and Health. URL : http ://www.who.int/classifications/icf/fr/.

[Wilbur98] Wilbur J. The knowledge in multiple human relevance judgements. *ACM*, 1998 ; 102–115.

[Xu98] Xu J., Croft B. Corpus-based stemming using co-occurence of word variants. *ACM Transactions on Information Systems*, 1998 ; 16(1) : 61–81.

[Yang94] Yang Y., Chute G. An example-based mapping method for text categorization and retrieval. *ACM Transactions on Information Systems*, 1994 ; 12(3) : 252–277.

[Zeng-Treitler07] Zeng-Treitler Q., Kim H., Goryachev S., Keselman A., Slaughter L., Smith C. Text Characteristics of Clinical Reports and their Implications for the Readability of Personal Health Records. *Stud Health Technol Inform*, 2007 ; 1117–21.

[Zeng99] Zeng Q., Cimino J. Evaluation of a system to identify relevant patient information and its impact on clinical information retrieval. *AMIA Annu Symp Proc*, 1999 ; 642–6.

[Zipf49] Zipf G. Human behavior and principles of feast effort. 1949.

[Zweigenbaum89] Zweigenbaum P., Bachimont B., Bouaud J., Cavazza M., Doré L. Hélène Compréhension de comptes rendus d'hospitalisation. *Informatique et Gestion des Unités de Soins Paris : Springer-Verlag*, 1989 ; 1 :257–68.

[Zweigenbaum90] Zweigenbaum P., Cavazza M. Deep sentence understanding in a restricted domain. *Proc 13 th COLING, Helsinki*, 1990 ; 82–4.

[Zweigenbaum92] Zweigenbaum P., Cavazza M., Doré L., Bouaud J., Sedlock D. Natural language processing of patient discharge summaries (NLPAD) – extraction prototype. *In Jaap Noothoven, IOS Press, Amsterdam*, 1992 ; 277–286.

[Zweigenbaum94] Zweigenbaum P., consortium MENELAS. MENELAS : an access system for medical records using matural language. *Comput Methods Programs Biomed*, 1994 ; 45 : 117–20.

[Zweigenbaum95] Zweigenbaum P., Bachimont B., Bouaud J., Charlet J., Boisvieux J. Issues in the structuring and acquisition of an ontology for medical language understanding. *Methods Inf Med*, 1995 ; 34(1-2) : 15–24.

[Zweigenbaum98] Zweigenbaum P., Courtois P. Acquisition of lexical resources from SNOMED for medical language processing. *Proc 9th World Congress on Medical Informatics*, 1998 ; 586–90.

[Zweigenbaum99] Zweigenbaum P. Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances. *Innovation Stratégique en Information de Santé*, 1999 ; (2–3) :27–47.

[Zweigenbaum01] Zweigenbaum P., Darmoni S., Grabar N. The contribution of morphological knowledge to French MeSH mapping for information retrieval. *Journal of the American Medical Informatics Association*, 2001 ; 8 (suppl) : 796–800.

[Zweigenbaum03] Zweigenbaum P., Baud R., Burgun A., Namer F., Jarrousse E., Grabar N., Ruch P., Le Duff F., Thirion B., Darmoni S. UMLF : construction d'un lexique médical francophone unifié. *In Actes des 10 Journées Francophones d'Informatique Médicale*, 2003 ; .

Publications personnelles

A.6 Publications internationales à comité de lecture

[Pereira08] Pereira S., Névéol A., Kerdelhué G., Serrot E., Joubert M., Darmoni S.J. Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a french online catalogue. AMIA Annu Symp Proc (in press), 2008.

[Pereira06] Pereira S., Névéol A., Massari P., Joubert M., Darmoni S.J. Construction of a semi-automated ICD-10 coding help system to optimize medical and economic coding. Proceedings of MIE2006, Stud Health Technol Inform. 2006 ;124 :845-50.

[Massari08] Massari P., Pereira S., Thirion B., Derville A., Darmoni S.J. Use of super–concepts to customize electronic medical records data display. Stud Health Technol Inform. 2008 ; 136 :845–850.

[Merabti08] Merabti T., Pereira S., Lecroq T., Joubert M., Darmoni S.J. Inheritance of SNOMED CT relations between concepts to two health terminologies (SNOMED International and ICD10). Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED). 2008 ;118.

[Merabti08] Merabti T., Pereira S., Letord C., Lecroq T., Dahamna B., Joubert M., Darmoni J. Searching Related Resources in a Quality Controlled Health Gateway : a Feasibility Study. Proceedings of MIE2008, Stud Health Technol Inform, Volume 136, Pages 235–240, 2008

[Névéol07] Névéol A., Pereira S., Kerdelhué G., Dahamna B., Joubert M., Darmoni S.J. Evaluation of a simple method for the automatic assignment of MeSH descriptors to health resources in a french online catalogue. Proceedings of Medinfo2007, Stud Health Technol Inform. 2007 ; 129 :407-11.

[Névéol06] Névéol A., Pereira S., Soualmia F.F., Thirion B., Darmoni S.J. A method of cross-lingual consumer health information retrieval. Proceedings of MIE2006, Stud Health Technol Inform. 2006 ; 124 :601–608.

A.7 Publications nationales à comité de lecture

[Pereira07] Pereira S., Massari P., Joubert M., Darmoni S. Utilisation de métatermes pour la recherche d'information dans les dossiers médicaux. In

Actes des journées Francophones d'Informatique Médicale. 2007.

[Pereira07] Pereira S., Névéol A., Massari P., Darmoni S., Joubert M. Evaluation de plusieurs terminologies médicales pour optimiser l'aide au codage médico-économique par analyse automatique de dossiers électroniques de patients. In Acte EMOI2006. 2006.

A.8 Posters nationaux et internationaux

[Pereira08] Pereira S., Massari P., Joubert M., Serrot E., Darmoni S.J. Exploring multi-terminology indexing of discharge summaries. Poster MIE2008. 2008.

[Pereira08] Pereira S., Massari P., Buemi A., Dahamna B., Serrot E., Joubert M., Darmoni S.J.. Evaluation of two french snomed indexing systems with a parallel corpus. Poster 3rd international conference on Knowledge Representation in Medicine (KR-MED). 2008.

[Pereira07] Pereira S. Indexation multi-terminologies. Journée des doctorants du laboratoire LITIS. 2007.

[Pereira05] Pereira S., Névéol A., Massari P., Darmoni S.J. Evaluation de plusieurs terminologies médicales pour optimiser l'aide au codage médico-économique par analyse automatique de dossiers électroniques de patient. ASTI2005, Clermont Ferrand. 2005.

[Darmoni08] Darmoni S.J., Pereira S., Névéol A., Massari P., Dahamna B., Letord C., Kedelhué G., Piot J., Derville A., Thirion B.. French info- button : an academic and. . .business perspective. AMIA Symp., en cours de publication, 2008.

[Thirion07] Thirion B., Pereira S., Névéol A., Dahamna B., Darmoni S.J. French MeSH browser : a cross-language tool to access Medline/Pubmed. AMIA annual symposium, page 1132, 2007.

A.9 Autres communications

[Pereira08] Pereira S., Serrot S., Joubert M., Darmoni S.J. Extraction de concepts multi-terminologiques. Journée des doctorants LITIS. 2008.

[Pereira07] Pereira S., Darmoni S.J.. Diffusion et mise en oeuvre des recommendations de pratique clinique : Les GBP des textes essentiellement. cours de Master santé publique, université Paris 5. 2007.

[Pereira07] Pereira S., Serrot S., Joubert M., Darmoni S.J. Extraction de concepts multi-terminologiques. Séminaire CISMeF. 2008.

[Pereira07] Pereira S., Serrot S., Joubert M., Darmoni S.J. Extraction de concepts multi-terminologiques du dossier médical. Journée ≪Serveurs de terminologies médicales pour le codage du dossier patient : mythes et limites ≫. 2007.

[Névéol07] Névéol A., Pereira S., Lortal G., Darmoni S.J. Using NooJ for the analysis of medical text. *NOOJ2007*

[Pereira05] Pereira S., Névéol A., Massari P., Darmoni S.J. Évaluation de plusieurs terminologies médicales pour optimiser l'aide au codage médico-économique par analyse automatique de dossiers électroniques de patient. Santé Publique, Lille. 2005.

A.10 Rapports

[Pereira08] Pereira S. Comparaison des serveurs de terminologies existants. *Rapport interne Vidal.* 2008.

[Dahamna07] Dahamna B., Pereira S., Darmoni S.J. Fiche de proposition de sujet PIC. INSA de Rouen. 2007

A.11 Valorisation

[Pereira06] Pereira S., Thirion B., Kerdelhué G., Letord C., Dahamna B., Névéol A., Piot J., Darmoni S.J. Connaissance contextuelle et personnalisée. Valorisation auprès de l'université de Rouen.

A.12 Non encore publiés

[Pereira] Pereira S., Massari P., Buemi A., Dahamna B., Serrot E., Joubert M., Darmoni S.J. Evaluation of two French SNOMED indexing systems with a parallel corpus.

[Letord] Letord C., Sakji S., Pereira S., Dahamna B., Kergourlay I., Darmoni S. Un portail d'information sur le médicament en Europe.

Table des figures