Caractériser la terminologie des usagers de santé dans le domaine du cancer du sein

Radja Messai*,**, Michel Simonet*
Nathalie Bricon-Souf**, Mireille Mousseau***

*Laboratoire TIMC, Faculté de Médecine, Université Joseph Fourier, Grenoble, France
michel.simonet@imag.fr
http://www-timc.imag.fr
**CERIM, E.A. 2694, Faculté de Médecine, Université de Lille2, Lille, France
radja.messai@univ-lille2.fr
nathalie.souf@univ-lille2.fr
http://cerim.univ-lille2.fr
***CHU de Grenoble, Grenoble, France
mmousseau@chu-grenoble.fr
http://www.chu-grenoble.fr

Résumé. Internet est devenu une source importante d'informations médicales pour les patients et leurs proches : recherche d'informations sur leurs maladies et les dernières recherches cliniques, ainsi que pour y constituer des communautés "numériques" de dialogue et de partage. Cependant, accès à Internet ne signifie pas nécessairement accès à l'information. Le manque de familiarité avec le langage médical constitue un problème majeur pour les usagers de santé dans l'accès à l'information et son interprétation. Ce papier s'inscrit dans la problématique d'étude et de caractérisation de la terminologie des usagers de santé pour pouvoir proposer des services adaptés à leur langage et à leur niveau de connaissances. Le travail réalisé est une ontologie dans le domaine du cancer du sein orientée vers les usagers de santé. Cette ontologie est construite à partir d'un ensemble de corpus de textes représentant deux catégories : les médiateurs et les usagers de santé. Les éléments de cette ontologie ont été analysés en utilisant des méthodes quantitatives et qualitatives sur plusieurs niveaux : termes, concepts et relations.

1 Introduction

L'information médicale à destination du grand public peut avoir un impact considérable sur les plans personnel, social et économique. Cependant, les stratégies de communication de cette information souffrent encore de plusieurs lacunes, et le grand public a peu bénéficié de l'explosion de la recherche médicale, de la littérature scientifique et même des connaissances médicales de base. Deux facteurs principaux peuvent limiter la communication ou le transfert de cette information : l'accessibilité physique et l'accessibilité conceptuelle à l'information. L'accessibilité physique à l'information a été améliorée grâce aux campagnes d'information

de proximité (hôpitaux, médecins, pharmaciens, numéros verts d'information), l'utilisation des médias (télévision, radio, journaux et magazines) et les technologies d'information comme Internet. Cependant, l'accessibilité conceptuelle à l'information, qui correspond à la capacité de trouver, comprendre et interpréter l'information médicale, n'a pas beaucoup progressé.

Dans ce travail de recherche, nous nous sommes intéressés à l'étude de deux communautés : les usagers et les médiateurs de santé. Selon Tse et Soergel (2003), un usager de santé est une personne qui cherche des informations sur des sujets ou des services et produits de santé, dans le but de prendre une décision à propos d'un problème médical personnel. Les médiateurs de santé, même s'ils viennent de diverses professions (santé, journalisme, éducation et communication) et de différents types d'organisations (gouvernementales, privées à but caritatif et commerciales), ils partagent le même objectif : communiquer l'information médicale au grand public. Le type d'information communiquée peut être informationnel, persuasif ou commercial. La recherche que nous avons menée a exploré principalement deux questions :

– Quelles sont les méthodes utiles pour l'identification, la création et l'analyse des termes utilisés par le grand public dans un corpus de textes d'un domaine particulier ?
– Quelles sont les caractéristiques des termes utilisés par les usagers de santé et les médiateurs de santé pour communiquer sur des problèmes d'ordre médical ?

La première étape de ce travail a eu comme résultat la construction d'une ontologie du cancer du sein à partir d'un corpus de textes et qui regroupe les concepts et les termes utilisés par ces deux communautés. Cette ontologie est devenue le noyau d'une application de reformulation de requêtes des usagers de santé. Dans ce papier, nous nous intéressons à la construction de cette ontologie et à l'analyse de ses constituants : termes, concepts et relations.

2 Construction d'une ontologie du cancer du sein à partir d'un corpus de textes

La construction d'ontologies destinées aux usagers de santé n'est pas une tâche facile. Contrairement à la terminologie médicale, qui contient un vocabulaire plus stable et mieux identifiable, le langage des usagers est dynamique et influencé par l'histoire de chaque individu. De ce fait, l'identification du vocabulaire utilisé par les usagers de santé revient d'abord à l'identification de groupes d'usagers qui utilisent plus ou moins les mêmes termes pour parler des mêmes concepts. Les recherches actuelles suggèrent qu'une grande quantité d'expressions utilisées par le grand public est suffisamment stable pour pouvoir constituer un vocabulaire standard (Tse et Soergel, 2003). Dans un contexte plus précis (les utilisateurs d'Internet dans notre cas), et pour une tâche précise (la recherche d'information médicale), il existe un certain consensus, condition nécessaire pour la construction d'une ontologie.

Pour l'acquisition des termes et des connaissances utiles à la construction de cette ontologie, nous avons construit deux corpus de textes qui ont constitué le point de départ de notre processus de construction. Nous nous sommes inspirés des travaux existants de construction de ressources onto-terminologiques à partir de corpus de textes propres à un domaine particulier (Baneyx, 2007; Bourigault et Aussenac-Gilles, 2003).

2.1 Construction du corpus

Pour réaliser ces objectifs, deux ensembles de documents issus du Web ont servi comme sources pour la construction de deux types de corpus :

1. Corpus destiné aux usagers de santé (corpus médiateur[1]), qui consiste en des documents écrits par des professionnels de santé, des journalistes, des communicateurs dans le domaine de la santé ou même des patients avec une longue expérience de la maladie.
2. Corpus des usagers de santé, qui consiste en des messages de discussion écrits par les usagers de deux forums de patients sur le cancer du sein.

La construction du corpus médiateur a été manuelle. Le moteur de recherche *Google* a été utilisé pour générer la liste des pages Web en rapport avec le cancer du sein, en soumettant la requête *cancer du sein*. Nous avons collecté 575 documents en français. Les pages ont été sélectionnées selon des critères qualitatifs : la représentativité du domaine, le public ciblé par le site (grand public ou professionnel de santé), l'auteur de la page (professionnel de santé ou non), le langage utilisé (facile ou difficile). Ces critères nous ont permis de sélectionner les pages Web les plus appropriées pour fournir les termes représentatifs utilisés par les médiateurs de santé. Pour représenter les différents types et modalités de communication médicale, des articles de sites gouvernementaux et commerciaux ont également été sélectionnés.

La construction du corpus des usagers de santé a été automatique en utilisant un parseur développé pour extraire les messages des usagers en s'appuyant sur la structure des sites Web. Nous avons collecté un corpus issu de deux forums de patients atteints du cancer du sein : `Essentielles.net` et le forum de La Ligue Contre le Cancer. Ce corpus contient un ensemble de 9 843 messages.

2.2 Extraction des n-grammes du corpus

L'extraction des termes est *"la tâche de détecter automatiquement, à partir de corpus textuels, des unités lexicales qui désignent des concepts dans des domaines de thématique restreinte"* (Vivaldi et al., 2001, p. 515). Ce problème est difficile et plusieurs techniques ont été proposées pour le résoudre, plus spécialement des méthodes statistiques et linguistiques (Vivaldi et al., 2001; Rousselot et Frath, 2000). Ces dernières comprennent des approches de traitement de corpus par analyse syntaxique et distributionnelle ou par des patrons lexico-synatxiques. Pour que ces approches d'extraction automatique soient efficaces, une connaissance linguistique préalable spécifique au domaine est nécessaire. Que ce soit un système qui utilise des mécanismes linguistiques, probabilistes ou combinés, les caractéristiques des termes spécifiques au domaine doivent être fournies implicitement (i.e., par la qualité du corpus) ou explicitement (e.g., liste de termes de base ou patrons lexico-syntaxiques spécifiques au domaine). Une langue de spécialité bien étudiée peut être un candidat approprié pour de tels systèmes. Cependant, nous pensons que nous ne disposons pas d'assez d'information sur le vocabulaire médical utilisé par les usagers de santé pour profiter de ces approches. Par conséquent, nous avons choisi d'extraire les n-grammes du corpus en nous basant sur la méthode des segments répétés qui consiste à repérer les séquences de mots répétées dans le corpus (Lebart et Salem, 1994).

[1]Mediator corpus : appellation utilisée par Tony Tse en référence aux médiateurs de santé

Les résultats obtenus ont été améliorés par l'utilisation de filtres (Rousselot, 2004) identifiant les mots qui indiquent des frontières de termes, en complément des signes de ponctuation délimiteurs de séquences (filtre "coupant" : verbes courants, adverbes, pronoms relatifs, conjonctions) et ceux qui ne peuvent se trouver aux bornes d'un terme (articles, prépositions).

Les n-grammes sont classés par ordre de fréquence décroissante. L'extraction a été faite en mode itératif. A chaque itération, les mots les plus productifs et qui ne sont pas représentatifs du domaine ont été ajoutés au filtre. Nous avons gardé pour analyse les expressions régulières avec une fréquence supérieure à 6. A la fin du processus, nous avons obtenu 6 896 termes candidats du corpus médiateur et 11 723 termes candidats du corpus des usagers de santé.

2.3 Sélection des termes candidats du domaine

Définition des termes du domaine La liste des termes et expressions extraite des corpus de textes est l'étape de départ dans la construction de l'ontologie. Cette liste contient encore beaucoup de bruit. Il faut donc filtrer et identifier les termes spécifiques au domaine du cancer du sein. Nous appelons les termes qui nous intéressent *les termes du domaine*. Nous avons utilisé le parseur développé pour l'extraction du corpus des usagers de santé pour extraire des éléments spécifiques dans les pages Web qui forment le corpus médiateur (titres, sous-titres, expressions en gras, expressions en italique, expressions soulignées). Ces éléments ont été reconnus grâce aux balises HTML : <h1>, <h2>, , . . .Cette structure interne des sites Web a permis d'identifier des *termes fondamentaux du domaine*. En effet, l'importance de certains termes peut être présumée au vu de la place qu'ils occupent dans la structure des documents.

Extraction, filtrage et sélection La liste des termes extraits des corpus contient encore beaucoup de bruit. L'identification des termes du domaine doit se faire manuellement. Pour faciliter la tâche de sélection des termes du domaine, nous avons utilisé un concordancier (Bernhard, 2003). Techniquement, le concordancier permet à l'utilisateur de formuler des requêtes sous forme d'expressions régulières dans le corpus et d'afficher toutes les parties du corpus contenant cette expression. Cette vision d'un terme dans son contexte permet en général de déterminer très rapidement son sens.

L'analyse des *termes fondamentaux du domaine* a permis de repérer les grands axes conceptuels typiques du corpus et donc du domaine. Ces axes sont indexés pour pouvoir les utiliser par la suite dans le classement des autres termes. Huit axes ont ainsi été définis : organe, fonction d'organes, substance, pathologie, symptôme, cause, examen, traitement. Ces axes permettent de regrouper les termes du domaine dans des groupes conceptuels *larges* et d'adopter ainsi une méthode de construction descendante.

L'étape suivante consiste à élaborer la structure hiérarchique de l'ontologie. Dans chaque groupe conceptuel, nous avons étudié les termes pour les placer au sein de l'ontologie. Pour chaque terme, un module de recherche développé affiche dans un tableau l'ensemble des autres termes qui le contiennent. Ces termes peuvent avoir des relations avec le premier et permettent ainsi de créer des sous-hiérarchies conceptuelles.

Mise en œuvre des principes différentiels La mise en œuvre des principes différentiels permet de créer la structure hiérarchique de l'ontologie. Cette partie est basée sur les travaux de

Type de correspondance	Pourcentage
Correspondance exacte	83%
Correspondance partielle	3%
Aucune correspondance	14%

TAB. 1 – *Résultats du mapping vers UMLS*

C. Roche et *B. Bachimont* (Roche, 2003; Bachimont et al., 2002). Il convient alors de préciser pour chaque concept les principes différentiels qui le définissent. Ces principes permettent de le placer dans la hiérarchie et de justifier cet emplacement. Ces principes servent aussi de documentation à l'ingénieur de connaissance en cas de mise à jour ou de révision de l'ontologie. Par exemple, le concept *cancer du sein intracanalaire* et le concept *cancer du sein intralobulaire* sont des concepts frères. Le principe différentiel entre ces concepts est relatif à l'emplacement de la prolifération des cellules malignes : les canaux galactophoriques pour le premier et les acini situés dans les lobules pour le deuxième.

2.4 Création des relations

Le concordancier est un bon support pour la recherche de relations sémantiques à partir de leurs formes lexicales dans le corpus. En plus des relations *classiques* telles que *Est_Un*, *Partie_De*, nous avons défini un ensemble de 61 relations. Ce travail s'est appuyé sur les travaux de *Soergel* et *Slaughter* (Slaughter et al., 2006; Soergel et al., 2004). Grâce aux relations sémantiques, nous avons pu modéliser une partie des connaissances et des croyances des usagers de santé. Cette partie sera plus détaillée dans la section analyse des relations.

2.5 Mapping vers UMLS et CHV

Au cours de cette étape, nous avons cherché à relier les concepts de notre ontologie à ceux d'UMLS (Unified Medical Language System) et à ceux de CHV (Consumer Health Vocabulary). L'intérêt de ce travail est double. D'une part, UMLS est le point d'entrée à plusieurs terminologies médicales et le mapping vers cette ressource facilite le mapping vers plusieurs d'autres terminologies. D'autre part, comme UMLS est une ressource destinée aux professionnels de santé, ce mapping peut être le point de départ d'une étude comparative entre les deux types de ressources.

Le mapping a été effectué manuellement. Pour chaque concept, la liste de ces termes a été traduite en anglais. Ensuite, les termes en question ont été soumis aux serveurs d'UMLS 2008AA[2] et CHV[3]. Seuls les cas de correspondance exacte ont été considérés. Par exemple, le concept *Adenopathie_Tumorale* n'a pas de correspondance exacte dans UMLS. Le concept *Adenopathie* est suggéré par le serveur d'UMLS comme correspondance partielle. Ce dernier n'a pas été pris en compte dans le processus de mapping. Le tableau 1 montre les résultats de ce mapping :

En plus de ces résultats, nous avons fait cet ensemble de remarques :

[2] `https://login.nlm.nih.gov/cas/login?service=http://umlsks.nlm.nih.gov/uPortal/Login`

[3] `www.consumerhealthvocab.org`

	Usager de santé	Médiateur	INFACE
Moyenne des caractères par terme	21,5	22,8	27,4
Moyenne des mots par terme	3,1	3,0	3,8

TAB. 2 – *Longueur des termes*

– 5 concepts ont des correspondances multiples dans UMLS. Par exemple, *Cancer de l'ovaire* est désigné par les deux concepts *Ovarian carcinoma* et *Malignant neoplasm of ovary* dans UMLS. Le concept *Sein* peut être également désigné par les deux concepts *Breast* et *Entire breast*.
– 2 paires de concepts ont une correspondance unique dans UMLS. Les concepts *Mammographie* et *Mammogramme* sont alignés vers le concept *Mammography* dans UMLS. La même chose est observée pour la paire de concepts *Primipare* et *Primiparité* et le concept *Primiparity* dans UMLS.

Ces résultats montrent des différences entre les deux terminologies au niveau des termes et au niveau des concepts. Ces différences vont être analysées dans la section analyse des concepts.

2.6 Résultats obtenus

Nous avons défini l'ensemble des concepts de l'ontologie et les termes du domaine (les synonymes), ainsi que les relations sémantiques qui les relient. Nous avons utilisé le logiciel PROTÉGÉ pour l'édition de l'ontologie. L'ontologie obtenue a été traduite en langage OWL, qui répond à nos besoins en termes d'expressivité et de maniabilité.

Nous avons pu identifier 1 287 concepts désignés par 2 783 termes français. L'ensemble des concepts et des termes a été revu par un médecin oncologue au CHU de Grenoble et deux cadres infirmiers du service d'oncologie.

3 Analyse des résultats

3.1 Analyse des termes

Des recherches actuelles utilisent la longueur des termes comme substitut de la complexité des termes dans les calculs de "*la lisibilité*" (readability) des documents (Zeng et al., 2007; Rosemblat et al., 2006; Gemoets et al., 2004). Nous avons voulu comparer la longueur des termes qui proviennent des deux corpus ainsi que les termes d'une ontologie de cancer du sein destinée aux professionnels de santé développée au cours du projet européen INFACE [4]. Pour cela, nous avons développé un outil qui a effectué les calculs suivants : moyenne des caractères par terme et moyenne des mots par terme pour chaque source.

Bien que les termes dans le corpus médiateur soient de plus d'un caractère plus longs que les termes issus du corpus des usagers de santé, ils sont pratiquement identiques en nombre de mots par terme. Les termes issus de l'ontologie destinée aux professionnels contiennent plus

[4]Visual Interfaces for Timely retrieval of Patient-Related Information, projet du 5ème PCRD, Sept. 2002 – Août 2004 `http://www.inface.org`.

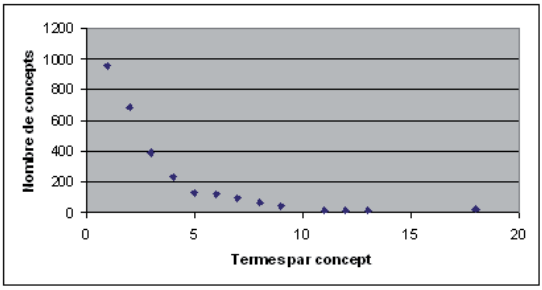

FIG. 1 – *Distribution des termes par concept*

de mots et un nombre plus élevé de caractères. Les usagers de santé utilisent généralement des expressions descriptives au lieu du terme médical. Ceci implique l'utilisation de beaucoup de mots, par exemple : *enlever le sein* pour *mastectomie* ou *perte de cheveux* pour *alopécie*. Cependant, cette comparaison présente une limitation. L'ontologie INFACE ne contient pas les mêmes concepts que notre ontologie. Elle a un niveau de granularité beaucoup plus élevé et contient par conséquent des termes très longs qui désignent des concepts très spécifiques, comme par exemple : *reflux de la lymphe tissulaire dans les tissus mammaires*.

3.2 Analyse des concepts

Le mapping de l'ontologie vers UMLS a révélé plusieurs choses intéressantes. Plusieurs concepts ont des correspondances multiples dans UMLS. Par exemple, le concept *Cancer de l'ovaire* a été relié aux deux concepts UMLS *Ovarian carcinoma* et *Malignant neoplasm of ovary*. D'autres concepts de notre ontologie, au contraire, ont été alignés au même concept UMLS. Par exemple, les concepts *Mammographie* et *Mammogramme* sont alignés vers le concept *Mammography* dans UMLS, UMLS ne faisant pas la différence entre les deux. Dans notre cas, nous les avons séparés car le Mammogramme est le résultat de la Mammographie : une image radiologique pour le premier et une technique d'examen radiologique pour le deuxième. Ces cas montrent des anomalies de conceptualisation dans UMLS.

3.3 Analyse des termes-concepts

Analyse de la variabilité expressive des concepts Le langage naturel est flexible et fournit plusieurs manières d'exprimer la même notion. Cette "variabilité expressive" peut être modélisée par le nombre de termes qui désignent le même concept (Tse et Soergel, 2003). Nous avons voulu étudier cette variabilité expressive pour savoir quel est le type de concepts qui a la plus grande variabilité expressive. La moyenne de la variabilité expressive est de 2,16 pour l'ensemble de l'ontologie. La distribution de la fréquence totale des termes par concept est illustrée dans la figure 1. La distribution des termes suit une courbe de type Zipf, avec la majorité des concepts qui possèdent un seul terme.

Pour étudier le type de concepts ayant la plus grande variabilité expressive, nous avons utilisé la classification des descriptions médicales de *Blois* pour classer les concepts dans des

Les niveaux de Blois
0 : Patient comme un entier.
-1 : Parties majeures du patient : e.g., abdomen, tête.
-2 : Sytèmes physiologiques : e.g., système cardiovasculaire.
-3 : Parties du système, ou organes : e.g., vaisseau, cœur.
-4 : Parties d'organe, ou tissu : e.g., myocarde, moelle osseuse.
-5 : Cellule : e.g., cellule épithéliale, lymphocyte.
-6 : Parties des cellules : e.g., membrane cellulaire, noyau de la cellule.
-7 : Macromolécule : e.g., enzyme, protéine.
-8 : Micromomolécule : e.g., glucose, acide ascorbique.
-9 : Atomes ou ions : e.g., sodium, calcium.

TAB. 3 – *Les niveaux hiérarchiques des descriptions médicales de Blois*

groupes sémantiques (Blois, 1984). La classification de *Blois* (voir Tableau 3) a été construite dans le but de représenter les maladies par un ensemble d'attributs sur l'ensemble des classes. Bien que l'ambiguïté existe sur tous les niveaux d'attributs, *Blois* a observé que l'ambiguïté est particulièrement remarquée dans les niveaux supérieurs où "il s'agit des objets de tous les jours . . .et des processus de la vie quotidienne . . ." (p.61).

Comme les usagers de santé sont plus familiers avec les objets et les processus de tous les jours (niveaux supérieurs) que les cellules et les molécules (niveaux inférieurs), *Blois* a établi que les usagers de santé vont plutôt utiliser les termes de ces niveaux que ceux des niveaux inférieurs. La hiérarchie a été divisée en deux niveaux globaux : "expérience quotidienne", défini comme les niveaux 0 (usager de santé) à -4 (Partie d'organe ou tissu), et "niveau technique", allant du niveau -5 (cellule) au niveau -9 (atomes ou ions). C'est à dire que les cinq niveaux supérieurs concernent les objets et les entités qui peuvent être directement observées ou expérimentées dans la vie de tous les jours. Par exemple, les organes et les tissus sont identifiables par l'observation directe de notre corps (e.g., l'oeil ou la peau) ou des produits animaliers. Par contre, les cellules, les bactéries et les virus, qui ne sont pas directement observables, peuvent cependant être détectés indirectement à travers leurs effets (e.g., lait caillé ou viande avariée). Les associations au niveau de l'expérience quotidienne semblent logiques et simples. Cependant, aux niveaux inférieurs, les effets peuvent ne pas être notables et la chaîne du raisonnement causal devient plus longue.

Nous avons pris un échantillon au hasard de 100 concepts. 50 concepts avec une variabilité expressive > 5 et 50 autres concepts avec une variabilité expressive ≤ 5. Nous avons rencontré des difficultés pour classer les concepts au sein de cette classification. La classification a été conçue pour classer les maladies mais l'ontologie contient d'autres concepts que des maladies, tels que les traitements, les examens cliniques et les concepts liés aux conditions psychologiques et sociales des patients. Ces concepts ont été classés au sein de la hiérarchie selon leur cause, leur effet ou ce sur quoi ils agissent. Par exemple, le concept *Tumorectomie* a été classé au niveau -4 (celui des parties d'organes) car la tumorectomie est appliquée sur la partie de l'organe touchée par la tumeur. Pareillement, le concept *Antibiotique* a été classé au niveau -5 car les antibiotiques sont administrés contre les bactéries, et les bactéries sont au niveau cellulaire. Le tableau 4 montre la distribution des concepts selon les niveaux de classification de *Blois*.

Niveaux	(VE $\leq$ 5)	(VE > 5)
0	6	4
-1	9	2
-2	4	7
-3	1	9
-4	4	17
-5	2	8
-6	2	0
-7	8	3
-8	9	0
-9	5	0

TAB. 4 – *Distribution des concepts selon leur niveau de variabilité expressive sur les niveaux de Blois*

Les concepts avec une variabilité expressive élevée concernent généralement des concepts du niveau supérieur "expérience quotidienne", plus spécialement le niveau -4. Il s'agit donc d'un niveau de concepts qui peuvent être rencontrés facilement dans la vie de tous les jours mais qui commencent à être à la frontière des concepts techniques. Les concepts concernent généralement des procédés médicaux ou des symptômes. Ils désignent généralement des concepts que l'on commence à connaître quand on côtoie la maladie. Les usagers de santé, quand ils ne connaissent pas le terme médical exact, décrivent le procédé médical (*enlever le sein* au lieu de *mastectomie*) ou le symptôme qu'ils ressentent (*mal de tête* au lieu de *céphalée*). Comme ce type de descriptions n'est ni normalisé ni limité, la production des termes est plus élevée. Par contre, les concepts avec une variabilité expressive faible concernent soient des concepts connus et bien appréhendés par les usagers de santé, comme le nom des organes (e.g., foie, poumon, ...) ou des concepts très spécifiques aux professionnels de santé (e.g., Tamoxifène, Bilirubine, ...).

Analyse de recouvrement entre les deux terminologies Nous avons comparé les deux terminologies issues des corpus médiateurs de santé et grand public pour déterminer le degré de recouvrement conceptuel et terminologique entre les deux. La comparaison a été conduite en deux étapes :

1. Le recouvrement conceptuel : déterminer les concepts qui sont communs aux deux terminologies et les concepts propres à chacune.
2. Le recouvrement terminologique : Pour les concepts communs, déterminer le nombre de termes communs aux deux terminologies.

Les listes des termes et des concepts qui proviennent de chaque corpus ont été comparées. Le tableau 5 montre les résultats de cette comparaison :

Les deux terminologies partagent un grand nombre de concepts. En étudiant de plus près les concepts propres à chaque terminologie, nous avons pu faire ces constatations :

– Les concepts propres à la terminologie des médiateurs de santé concernent généralement des concepts médicaux très spécifiques, surtout d'ordre anatomique. Par exemple : *Tubercules de Montgomery*.

	Commun	Spécifique usager de santé	Spécifique médiateur
Concepts	1254	8	25
Termes	2238	289	182

TAB. 5 – *Comparaison entre les deux terminologies*

– Les concepts propres à la terminologie des usagers de santé concernent soit des concepts d'ordre social et psychologique, soit des concepts généraux. Par exemple : *Remboursement des maladies*, *Produit de beauté*.

En étudiant les termes propres à chaque terminologie, nous avons pu faire les constatations suivantes :

– Termes très spécifiques : comme dans le corpus des médiateurs de santé il y a des descriptions de procédés médicaux, de maladies et du corps humain, les médiateurs de santé utilisent des termes très spécifiques à la médecine que les usagers de santé n'emploient pas entre eux.
– Fautes d'orthographe : les termes médicaux sont parfois difficiles à retenir. Les usagers de santé interprètent parfois mal les mots qu'ils entendent. Ils utilisent par conséquent des homophones et des mots du langage général morphologiquement similaires. Par exemple : *limphocyte* (lymphocyte), *maladie de Hojkin* (maladie de Hodgkin).
– Termes abrégés : les termes médicaux sont généralement longs. Les usagers de santé utilisent des formes abrégées et non standardisées telles que les abréviations et les coupures. Par exemple : *neurochir* (neurochirurgien), *chimio* (chimiothérapie).
– Définition/Description : ignorant le terme spécifique, les usagers de santé peuvent définir le concept ou décrire ses propriétés. Par exemple : *enlever le sein* (mastectomie), *diluant de sang* (anticoagulant).
– Exemples : ne connaissant pas le terme spécifique, les usagers de santé peuvent utiliser un concept particulier de la classe qui représente *un exemple* (concept plus spécifique) du concept en question. par exemple : *aspirine* (antalgique).
– Néologismes : les usagers créent parfois de nouveaux mots, utilisés parfois au sein d'un groupe de patients. Par exemple : *cancerinette* (jeune femme atteinte d'un cancer), *mort-vivant* (patient en fin de vie).

3.4 Analyse des relations

Les usagers de santé, en plus de leur problème de terminologie médicale, rencontrent parfois des problèmes pour comprendre comment les concepts médicaux sont reliés entre eux. Parmi les relations créées, on trouve la *Relation_X* qui modélise le fait qu'il existe une relation entre deux concepts mais qui ne peut pas être spécifiée. Cette relation est utilisée pour définir un lien entre des concepts que les usagers de santé relient sans qu'il y ait une justification médicale suffisante derrière. Par exemple, le concept *Pilule_Contraceptive* est reliée au concept *Cancer_Sein* parce qu'un certain nombre d'usagers de santé croient qu'il y a un lien entre les deux, bien que les études scientifiques n'en aient établi aucun. Par conséquent, nous avons choisi de les relier en utilisant *Relation_X*. Nous avons fait de même pour les concepts mal appréhendés par les usagers de santé. Par exemple, le concept *Vaginite* est relié au concept plus général *Maladie_Vagin* par les deux relations *Est_Un* et *Relation_X*. Les usagers de santé

associent la vaginite à tous les problèmes du vagin, bien que la vaginite ne représente que l'inflammation de la muqueuse du vagin. L'utilisation de ce type de relation pour modéliser ce type de phénomènes permet de ne pas altérer la structure bien fondée et stable de l'ontologie. Cette relation peut également changer ou disparaître au fil du temps et la modéliser de cette manière facilite la mise à jour de l'ontologie.

4 Conclusion

La procédure développée dans ce travail a montré quelques-unes des différences qui existent entre la terminologie des usagers de santé et celle des professionnels de santé. Nous avons pu constater que la principale différence réside au niveau des termes. Par conséquent, séparer le développement des ontologies destinées aux usagers de santé de celles des professionnels de santé n'est pas, à notre avis, justifiable. Il serait plus utile de reprendre des ressources destinées aux professionnels de santé et d'y ajouter des termes utilisés par les usagers de santé. Par contre, une réflexion plus profonde est indispensable sur un format de représentation de ce type de ressources. Une annotation du type des termes selon leur *technicité* pourrait faciliter l'utilisation de la même ressource selon le public des utilisateurs cibles.

Une des principales limitations des approches de construction par corpus est liée à la difficulté d'assigner correctement un sens aux termes repérés dans le corpus. Comme seuls les textes et non pas leurs auteurs sont disponibles, trouver le sens des termes est limité à l'interprétation de l'ingénieur des connaissances de l'intention de chaque auteur. Haas et Hert ont souligné ce problème : *"Même si les mots des utilisateurs peuvent être vus, l'intention derrière leur utilisation, ou ce que réellement veut l'utilisateur (le contenu et le contexte de l'utilisateur), ne peut être connu"* (Haas et Hert, 2002, p. 44). L'utilisation des messages de patients sur des forums réduit cette limitation en fournissant plus de contexte par rapport aux travaux existants et qui s'appuient sur des requêtes d'utilisateurs sur des moteurs de recherche ou des sites Web. Cependant, l'intention réelle peut être mieux cernée en utilisant des approches interactives.

5 Remerciements

Ce travail a été financé par La Ligue Nationale Contre le Cancer, La Fédération Hospitalière de France et l'Association Grenobloise d'Aide à la Recherche en Oncologie (AGARO).

Références

Bachimont, B., A. Isaac, et R. Troncy (2002). Semantic commitment for designing ontologies : A proposal. In *13th International Conference on Knowledge Engineering and Knowledge Management*, Volume Lecture Notes in Artificial Intelligence, pp. 114–121. Springer.

Baneyx, A. (2007). *Construire Une Ontologie De La Pneumologie : Aspects Théoriques, Modèles Et Expérimentations*. Ph. D. thesis, Université Paris 6.

Bernhard, D. (2003). Ontology building based on text corpora. Master's thesis, Institut National Polytechnique de Grenoble.

Blois, M. S. (1984). Information and medicine. Berkeley, CA : University of California Press.

Bourigault, D. et N. Aussenac-Gilles (2003). Construction d'ontologies à partir de textes. In *TALN*, Batz-sur-Mer.

Gemoets, D., G. Rosemblat, T. Tse, et R. Logan (2004). Assessing readability of consumer health information : an exploratory study. In *Medinfo*, Volume 11, pp. 869–873.

Haas, S. W. et C. A. Hert (2002). Finding information at the u.s. bureau of labor statistics : overcoming the barriers of scope, concept, and language mismatch. *Terminology 8*(1), 31–56.

Lebart, L. et A. Salem (1994). *Statistique textuelle*. Dunod, Paris.

Roche, C. (2003). The differentia principle as a cornerstone of ontology. In *Knowledge Management and Philosophy Workshop in WM 2003 Conference*, Luzern.

Rosemblat, G., R. Logan, T. Tse, et L. Graham (2006). Text features and readability : Expert evaluation of consumer health text. In *Medical Internet. MEDNET*.

Rousselot, F. (2004). L'outil de traitement de corpus likes. In *In Actes de TALN 2004*.

Rousselot, F. et P. Frath (2000). Terminologie et intelligence artificielle.

Slaughter, L. A., D. Soergel, et T. C. Rindflesch (2006). Semantic representation of consumer questions and physician answers. *Int J Med Inform 75*(7), 513–529.

Soergel, D., T. Tse, et L. Slaughter (2004). Helping healthcare consumers understand : an "interpretive layer" for finding and making sense of medical information. *Medinfo 11*(Pt 2), 931–935.

Tse, T. et D. Soergel (2003). Exploring medical expressions used by consumers and the media : An emerging view of consumer health vocabularies. In *AMIA Annu Symp Proc*, pp. 674–678.

Vivaldi, J., L. Marquez, et H. Rodriguez (2001). Improving term extraction by system combination using boosting. In *In Proceedings of ECML*, pp. 515–526.

Zeng, Q., H. Kim, S. Goryachev, A. Keselman, L. Slaughter, et C. Smith (2007). Text characteristics of clinical reports and their implications for the readability of personal health records. *Stud Health Technol Inform 129*, 1117–1121.

Summary

The Internet has become an important source of medical information for patients and their family members: search for information about their diseases and recent clinical research, building numeric communities for exchange and sharing of information and of personal experience. However, access to the Internet does not mean access to information. The lack of familiarity with the medical language is a major problem for health consumers in information access and understanding. The aim of this paper is to analyse and characterize the terms used by non-professionals during their discourse on medical topics in order to propose services adapted to their language and to their level of knowledge. The result of this work is a health consumer ontology in the breast cancer field, which is based on two types of text corpora : health mediators and health consumers. The elements of this ontology have been analysed on several levels: terms, concepts and relations.