Utilisation du réseau sémantique de l'UMLS pour la définition de types d'entités nommées médicales

Thierry Delbecque* — **Pierre Jacquemart***
Pierre Zweigenbaum*,,*****

* *INSERM, U729*
** *INaLCO, CRIM*
*** *AP-HP, STIM*

{pz,thd}@biomath.jussieu.fr, pierre.jacquemart@free.fr

RÉSUMÉ. *Les entités nommées (EN) sont des objets importants pour les systèmes de Questions-Réponses (QR). Cependant, les types d'EN habituels couvrent des concepts très généraux : dates, lieux géographiques, noms de personnes, etc. Pour un système de QR dédié à la médecine, il serait utile de disposer de types plus spécifiques. Une hiérarchie de types de concepts médicaux est définie dans l'UMLS, une grande base terminologique médicale produite par la NLM. Nous tentons d'évaluer l'utilisabilité de l'UMLS, dans sa partie francophone, comme source de telles entités. Nous réalisons un étiquetage d'un corpus médical par les concepts de l'UMLS et leurs types sémantiques. Puis nous montrons, à travers une étude statistique que les modalités de mise en œuvre de ces nouveaux types d'EN doivent prendre en compte l'origine individuelle des documents explorés lors d'une tâche de QR.*

ABSTRACT. *Named Entities are important concepts, regarding Question-Answering (QA) systems. Nevertheless, Named Entities categories are usually defined in a very broad sense: date, geographical area, and so on. It should quite profitable, for medical QA systems, to benefit from Named Entities especially dedicated to medicine. The UMLS is an important terminological tool created and maintained by the NLM; it comes along with a hierarchical organization of medical concepts. This paper is an attempt to evaluate the French part of UMLS as a resource for a medical-specific Named Entity tagger. We have tagged a set of medical documents, and have shown, using statistical studies that strategies using these new tags in a QA context are to take in account the individual origin of each document.*

MOTS-CLÉS : *UMLS, entités nommées, systèmes de Questions-Réponses, recherche d'information, réseau sémantique, étiquetage, structure thématique, analyse de données, analyse de corpus.*

KEYWORDS: *UMLS, Named Entities, Question-Answering systems, Information Retrieval, Tagging, Semantic Network, Thematic Structure, Data Analysis, Text Mining.*

1. Introduction

La médecine bénéficie aujourd'hui de bases documentaires médicales certifiées et en ligne (accessibles par internet) très riches. Les documents disponibles existent principalement en anglais. Une base classiquement étudiée est la base Medline de résumés d'articles scientifiques du domaine biomédical[1] , dont est extraite par exemple la collection OHSUMED [HER 94] employée dans la tâche de filtrage de TREC-9. Une vaste documentation médicale francophone de qualité est néanmoins également accessible, en particulier à travers le portail CISMeF[2].

La disponibilité de ces ressources offre l'opportunité de se focaliser sur les systèmes de questions-réponses (QR) propres à la médecine, dont différents travaux soulignent l'interêt [ALP 01, ZWE 03]. De tels systèmes devraient apporter des réponses précises et rapides à des questions médicales.

Les « entités nommées » (EN) sont des objets essentiels pour les tâches de recherche et d'extraction d'information [POI 03] ; la plupart des systèmes de QR y ont eux-mêmes recours (voir par exemple [FER 02]). Dans les systèmes dits « à domaine ouvert », les types d'EN référent à quelques familles courantes d'objets du monde ou de concepts communs (personnes physiques, organisations, lieux, dates, etc.). C'est sur ces EN que portent la plupart des questions que l'on trouve dans les évaluations de systèmes de QR à domaine ouvert (TREC, CLEF, tâche générale d'EQueR). Mais si l'on s'intéresse à des questions de domaine spécialisé, le type d'EN utile va varier. Il nous semble donc nécessaire de disposer de types d'entités nommées plus spécifiques aux applications médicales [ZWE 03]. Pour cela, nous proposons de s'appuyer sur des ressources linguistiques et terminologiques du domaine ; par chance, celles-ci ne manquent pas (SNOMED, MeSH, CIM-10, etc. ; voir [ZWE 04]).

L'objet de cet article est de présenter une expérience de définition et de repérage d'entités nommées médicales au sein d'un corpus francophone[3]. La ressource terminologique utilisée a été le *Unified Medical Language System*[4], et plus particulièrement son réseau sémantique. Nous nous sommes contraints de plus à n'utiliser que la partie francophone de l'UMLS.

Notre expérience a consisté à effectuer automatiquement l'étiquetage d'un ensemble important de documents relatifs au domaine médical par des concepts de l'UMLS, ce qui revient au repérage des EN spécifiquement médicales auxquelles nous nous attachons. Ensuite nous avons évalué globalement l'étiquetage ainsi obtenu en mesurant sur des échantillons aléatoires des indicateurs voisins de la précision et du

1. `http://www.ncbi.nlm.nih.gov/entrez/`.
2. `http://www.chu-rouen.fr/cismef/`, [DAR 00].
3. Cette expérience a été menée à l'occasion de la campagne d'évaluation EQueR, tâche médicale, de 2004.
4. UMLS, `http://www.nlm.nih.gov/research/umls/` ; pour une présentation en français, voir [ZWE 04].

rappel[5]. Puis nous avons voulu étudier l'efficacité des EN repérées dans un contexte de QR[6]. Dans la mesure où l'ensemble des nouveaux types d'EN est potentiellement vaste[7], il était difficile lors d'une étude préliminaire comme celle-ci de faire un balayage systématique de tous les cas possibles. Aussi avons-nous eu recours à une analyse exploratoire multidimensionnelle afin d'élire le candidat sur lequel nous concentrerions nos efforts. Lorsque cela a été fait, nous nous sommes placés dans l'hypothèse d'un type particulier de question médicale, et nous avons évalué l'apport du repérage de l'EN dans la recherche d'une réponse, en fonction de la source des documents explorés. Un résultat que nous avons pu ainsi mettre au jour est l'existence d'un gradient d'efficacité le long d'un axe de technicité médicale sur lequel peuvent se projeter les documents, ce phénomène n'étant pas dû à la spécificité du vocabulaire employé, mais bien à des différences de style de langage.

L'UMLS ayant été une ressource essentielle de notre étude, et n'étant pas forcément bien connue en dehors du monde de l'informatique médicale, nous avons jugé utile de commencer par en dresser un rapide portait, en mettant l'accent sur la place (ténue, on le verra) que le français y occupe (section 2). Ce sera également l'occasion d'expliquer ce qui sera utilisé comme étiquettes.

Puis nous détaillerons la méthode que nous avons adoptée, en nous attachant à préciser le corpus utilisé, les procédures d'étiquetage, et les analyses statistiques que nous avons effectuées (section 3).

Les résultats des analyses seront détaillés dans la section 4.

Enfin, nous conclurons en présentant les conséquences pratiques des observations que nous avons pu faire, une discussion critique de l'étude, et les prolongements que nous envisageons à ce travail.

2. L'UMLS

2.1. *Une vaste ressource, hybride et multilingue*

Proposée et maintenue par la NLM (National Library of Medicine), l'UMLS est la ressource terminologique la plus large actuellement disponible pour la médecine. Elle est le résultat de la fusion de plus d'une centaine de thésaurus de différentes langues, dont elle préserve les réseaux de relations entre termes. De ce fait, l'UMLS est qualifié de *métathésaurus*. Les thésaurus originaux les mieux représentés dans la version

5. Nos indicateurs se distinguent toutefois de la précision et du rappel du fait de l'existence de cas où il est difficile de trancher si l'étiquetage doit ou non être considéré comme correct.
6. Avec des contraintes sur la forme que peuvent prendre les réponses : elles doivent être contenues dans une même phrase.
7. Plus d'une centaine, puisque le réseau sémantique de l'UMLS apporte 134 types sémantiques, et 54 relations sémantiques, chaque type et relation constituant une étiquette possible.

que nous avons utilisée[8] sont le MeSH, SNOMED CT (incluant les Read Codes), et SNOMED International, en ce sens qu'ils apportent le plus grand nombre de termes.

L'UMLS organise les termes autour de *concepts* ; il y a ainsi plus de 700 000 concepts, auxquels sont rattachés des ensembles de termes qui les désignent. Enfin, à chaque terme est associé un ensemble de variantes graphiques (les *chaînes*, issues par exemple de l'emploi différent des casses, des ponctuations, etc).

L'UMLS est multilingue : pour un même concept peuvent cohabiter des variantes de termes de langues différentes. L'anglais est très majoritairement représenté, alors que le français ne couvre qu'à peine 2 % des concepts ; de même, en termes de synonymie, là où l'anglais propose en moyenne 2 chaînes pour représenter un concept, le français n'en propose que 1,54[9].

2.2. *Un réseau sémantique, source potentielle de types d'EN*

Tout en maintenant l'ensemble des relations hiérarchiques et transversales proposées par chacune de ses sources, l'UMLS organise l'ensemble des concepts au sein d'un réseau sémantique qui lui est propre [MCC 89]. Celui-ci consiste en 134 *types sémantiques* ; il est de plus structuré par 54 *relations sémantiques*. Types et relations sont hiérarchisés par un lien hiérarchique *is-a*.

Les types sémantiques offerts par l'UMLS semblent de bon candidats comme types d'EN spécifiques au domaine médical. Ils désignent des classes d'objets pertinentes pour le domaine, dont le repérage au sein d'ensembles de documents peut être utile à une tâche d'extraction d'information. Par exemple, le type sémantique *Sign or Symptom* (type T184), lorsqu'il est repéré, peut servir d'indice de localisation pour trouver les réponses à des questions telles que "quels sont les signes digestifs d'une RCH" ; il en va de même pour la plupart des types (*Substance*, *Clinical Drug*, *Enzyme*, *Pathological Function*, *Gene or Genome*, etc.).

Nous avons généralisé cette idée, en envisageant de même les relations sémantiques comme des indices. Ainsi, la relation sémantique *diagnoses*, qui peut par exemple exister entre les types *Sign or Symptom* et *Pathologic Function*, pourrait également servir à localiser les réponses aux questions portant sur des diagnostics.

---

8. Il s'agit de la version 2002-AA. En l'absence d'indication contraire, les chiffres que nous indiquons sont relatifs à cette version.

9. Cette présence dérisoire du français dans le métathésaurus constatée actuellement a motivé la naissance de deux projets dont le but est de remédier à cette situation : UMLF vise à la création d'un lexique complet de termes médicaux comparable au lexique SPECIALIST de l'UMLS ; VUMeF vise à augmenter la présence de chaînes françaises dans l'UMLS, ainsi qu'à y inclure des thésaurus propres au système de santé français tels que la CCAM.

[CANCER]	www.fnclcc.fr et www.fnclcc.com - Fédération Nationale des Centres de Lutte Contre le Cancer
[DOCFRA]	www.ladocfrancaise.gouv.fr et www.ladocumentationfrancaise.fr - La Documentation Française
[AFSSAPS]	afssaps.sante.fr - Agence Française de Sécurité Sanitaire des Produits de Santé (ex Agence du Médicament)
[ANAES]	www.anaes.fr - Agence Nationale d'Accréditation et d'Évaluation en Santé
[ORPHA]	www.orpha.net - Orphanet, serveur d'informations sur les maladies rares
[SENAT]	www.senat.fr - site officiel du sénat français
[CHUROUEN]	www.chu-rouen.fr - CHU de Rouen
[UROUEN]	www.univ-rouen.fr - Université de Rouen, restreinte à sa branche médicale
[CANADA]	www.hc-sc.gc.ca - site bilingue de Santé Canada, (ministère fédéral de la santé), source d'informations générales sur la santé publique au Canada, de statistiques, de conseils, etc.

Tableau 1. *Sources des documents de travail*

3. Matériel et méthode

Nous avons utilisé la version 2002-AA de l'UMLS.

3.1. *Le corpus*

Le corpus sur lequel nous avons travaillé est celui de la compétition EQueR de 2004. C'est un sous-ensemble des documents indexés par CISMeF, faisant partie des domaines listés dans le tableau 1[10].

3.2. *Préparation du corpus*

La préparation des données a été assurée par les traitements suivants :

– téléchargement des fichiers à partir des sources mentionnées précédemment, et nettoyage des données ;

– repérages des termes (par exemple, *« infarctus du myocarde »*, ou *« directive ministérielle »*), et étiquetage des termes par des concepts de l'UMLS francophone (section 3.2.1) ;

– repérage des phrases, et étiquetage des phrases par les types et relations sémantiques de l'UMLS (section 3.2.2).

A l'issue de ces traitements, sur lesquels nous revenons plus tard, on dispose d'un corpus au sein duquel les concepts de l'UMLS francophone ont été repérés, ainsi que de l'ensemble des phrases auxquelles ont été attachés les types sémantiques y apparaissant ; de plus, lorsque deux types sémantiques pouvant être reliés par une une plusieurs relations sémantiques cooccurrent au sein d'une même phrase, chaque relation est attachée à la ladite phrase.

10. Auxquels viennent se joindre les documents référencés *un lien plus loin* sur le même site.

3.2.1. *Repérage des termes et étiquetage par les concepts*

Le repérage de termes de l'UMLS au sein de textes, et les applications qui en découlent, constituent un sujet fréquent dans la littérature anglophone. Pour l'anglais, on peut, à l'instar de MetaMap [ARO 01], profiter d'outils de traitement automatique de la langue offerts par la NLM conjointement à l'UMLS : ceux du Specialist Lexicon [MCC 95], comptant entre autres un lexique, des méthodes de normalisation de termes, etc.

Le français ne dispose pas de telles ressources spécialisées ; dans notre cas, le repérage des syntagmes nominaux (SN) au sein du corpus est effectué sur la base de patrons de parties du discours que nous avons conçus de manière à ce qu'ils soient suffisamment robustes face à des cas de dégradation de la qualité du corpus[11]. Pour cela, l'ensemble du corpus est préalablement étiqueté par TreeTagger [SCH 94][12].

Un fois les SN repérés, la procédure d'étiquetage consiste à chercher, pour chaque mot du syntagme en cours de traitement, les chaînes françaises de l'UMLS dans lesquelles il intervient (en tant que tel, ou comme l'un de ses dérivés). Les chaînes de l'UMLS étant désignées par des identificateurs uniques, les SUI, on obtient, lorsque la recherche individuelle pour chaque mot est achevée, un treillis de chaînes UMLS, dont on ne conserve que les éléments maximaux. Les types des concepts attachés à ces éléments maximaux constitueront les étiquettes du syntagme[13].

3.2.2. *Repérage de phrases, et étiquetage par les relations sémantiques*

L'extraction des phrases du corpus se fait, pour sa part, sur la base du repérage des verbes. Lorsqu'un verbe (ou une séquence verbale, par exemple "être préconisé") est trouvée, des patrons de parties du discours sont utilisés pour isoler la partie de la phrase en amont du verbe et les parties en aval[14]. Ces patrons sont conçus de telle sorte qu'un terme, tel que défini plus haut, ne soit jamais à cheval sur deux phrases.

Ainsi, dans notre construction, une phrase s'organise toujours autour d'un verbe ou d'une séquence verbale, et peut bénéficier de la projection des types sémantiques préalablement effectuée sur les termes. C'est sur la base de cooccurrences de types sémantiques au sein d'une même phrase que l'on postulera la présence d'un indice de type « relation sémantique ». Par exemple, si dans une même phrase ont été repérés les types sémantiques T184 (*Sign or Symptom*) et T037 (*Injury or Poisoning*), du fait que dans le réseau sémantique ces deux types sont liés par la relation T163 (*diagnoses*),

11. Une grande partie du corpus a dû être convertie du format pdf au format texte, ce qui a généré des défauts qui n'ont pas pu tous être éliminés par notre procédure de nettoyage.
12. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`.
13. Cette méthode permet par exemple de bien poser le concept *œdème papillaire* sur une occurrence de la chaîne "œdème bilatéral papillaire", lorsque cette dernière n'est pas elle-même une chaîne de l'UMLS.
14. Nous avons opté pour une approche de cette nature, du fait que la qualité du corpus, après sa conversion en fichiers textes, ne nous permettait pas toujours de nous appuyer sur des signes tels que la ponctuation pour effectuer la segmentation.

la phrase sera de plus étiquetée avec la relation T163. Lorsqu'un couple de types sémantiques est relié par plus d'une relation, la phrase sera étiquetée avec chacune des relations.

3.3. *Les analyses*

Sur la base de cet étiquetage, nous avons voulu évaluer les aspects suivants.

3.3.1. *Qualité globale de l'étiquetage par les concepts*

Compte tenu de la pauvre couverture du français (source de lacunes), et du fait qu'en même temps, des termes très généraux sont associés à des concepts (source de confusion)[15], nous avons estimé des indicateurs pouvant se rapprocher de la précision et du rappel. Certaines particularités, cependant, les distinguent un peu de ces indicateurs classiques : il a été parfois difficile de décider si une étiquette est légitimement posée, et de même, il arrive qu'une étiquette soit bien posée, mais que l'étiquetage soit malgré tout incomplet. Nous avons dû tenir compte de ces faits lors de nos dénombrements.

3.3.2. *Sélection et évaluation d'une étiquette comme indice*

Nous appelerons *indice* une étiquette rencontrée dans un texte, et qui est effectivement prise en compte par un système de QR hypothétique. C'est donc un type ou une relation sémantique dont la présence dans un texte marque l'intérêt d'un fragment vis-à-vis d'une question.

L'ensemble des types potentiels d'EN que nous offre l'UMLS est vaste ; il n'est pas possible de tous les étudier, aussi avons-nous dû commencer par sélectionner celui que nous allions évaluer. Parmi les critères de choix qui nous ont guidés, deux ont été particulièrement importants :

– la sémantique de l'indice doit pouvoir naturellement servir à localiser la réponse à un type courant de questions médicales ;

– la projection de l'indice sur les phrases doit être en accord avec le sens de la phrase.

Pour le deuxième aspect, nous avons décidé de nous attacher prioritairement aux relations sémantiques, et nous avons profité de la propriété des phrases d'être systématiquement organisées autour de séquences verbales. En partant de l'idée qu'une corrélation existe fréquemment entre le sens d'un verbe utilisé dans une phrase et celui de ladite phrase[16], nous avons établi notre sélection sur la comparaison des proximités

15. Dans l'UMLS français, 9732 mots simples sont rattachés à des concepts ; parmi eux, beaucoup sont très généraux : extension, beauté, faiblesse, faillite, etc.
16. Sauf si le verbe est très fréquent, comme *être* ou *avoir*. Ainsi une phrase dont le verbe est *injecter* aura plus de chance d'être relative à un acte thérapeutique qu'une phrase dont le verbe serait *cueillir*.

entre les verbes et les relations sémantiques, une telle construction s'appuyant sur les cooccurrences des verbes et des relations dans une même phrase. Avec cette méthode, les bons candidats sont les relations dont le sens est bien exprimé par les verbes qui leur sont proches ; pour les autres, nous postulons que la projection de la relation n'est pas fidèle au sens de la phrase.

Différentes approches ont été envisagées pour obtenir ces similarités, dont le recours aux analyses des correspondances [BEN 79] qui conviennent bien à ce genre de problème. Cependant, la dilution des valeurs propres sur les axes des analyses ont fait que ces techniques n'étaient pas appropriées dans notre cas [DEL 04].

Nous avons mis en œuvre une classification ascendante hiérarchique [BEN 84] suivant une approche classique. On projette les relations dans un espace vectoriel dont chaque axe correspond à un verbe. Les coordonnées d'une relation suivant un axe est la fréquence du verbe correspondant dans l'ensemble des phrases comportant cette relation. La classification procède en agrégeant les relations les plus proches dans cet espace, c'est-à-dire celles dont le profil suivant les verbes et le plus similaire. La construction finale aboutit à un dendrogramme, par exemple celui de la figure 1. Pour un groupe donné sa projection relative sur un axe, donc sur un verbe, détermine la contribution de ce verbe à l'excentricité du groupe : plus une contribution est forte, plus le verbe participe à distinguer le groupe par rapport aux autres. Les verbes les plus contributifs seront ceux que nous définirons comme les plus proches du groupe.

4. Résultats

4.1. *Étiquetage par les concepts*

Le tableau 2 donne une mesure de la densité d'étiquetage. Dans ce tableau, un *token* désigne soit un mot, soit une ponctuation ; un *mot plein* désigne un mot auquel TreeTagger a affecté l'une des catégorie NOM, NAM (nom propre), ADJ, ADV, ou VER. Nous attachons aux syntagmes nominaux une attention toute particulière, dans

Item	Nombre
Occurrences de tokens	21744788
Occurrences de mots pleins	8091838
Mots pleins distincts	98837
Lemmes distincts	27015
Occurences de syntagmes nominaux	4101404
Syntagmes nominaux distincts	391966
... étiquetés	147007
... non étiquetés	246959

Tableau 2. *Densité de l'étiquetage au sein du corpus*

la mesure où les projections se font sur eux. On constate que 37 % des syntagmes différents sont étiquetés.

Le silence est estimé sur la base d'un échantillon aléatoire de 300 syntagmes non étiquetés. Parmi cet échantillon, nous avons dénombré les cas de :

faux négatifs - les termes pour lesquels nous étions convaincus d'une connotation médicale qui aurait justifié un étiquetage ; par exemple, "curage ganglionnaire" ;

vrai négatifs - les termes pour lesquels nous étions convaincus de l'absence d'une telle connotation ; par exemple "cadre stratégique national" ;

indécision - les termes pour lesquels le doute nous a empêché de trancher ; par exemple "enfermement familial", ou "entourage ultraviolent".

Le tableau 3a donne le résultat de ces estimations.

Item	Taux
Vrais négatifs	45 %
Faux négatifs	25 %
Indécision	30 %

(a) silence

Item	Taux
Acceptés	52 %
Partiels	32 %
incorrects	16 %

(b) précision

Tableau 3. *Estimation du "silence" et de la "précision" de l'étiquetage*

Nous avons procédé à des estimations similaires sur l'ensemble des termes étiquetés. Ainsi sur un échantillon de 300 syntagmes avons-nous tâché de dénombrer les étiquetages :

acceptés qui sont ceux pour lesquels nous avons considéré qu'il rendent bien le sens principal du syntagme ;

partiels qui ne reflètent qu'un sens incomplet ; par exemple, l'étiquetage de "atteinte gastrointestinale" par S0276149.C0012634(T047)[17] est incomplet par le fait qu'il ne se rapporte qu'à "atteinte", et que le silence est fait sur le modificateur "gastrointestinal" ;

incorrects qui reflètent un sens autre que celui du syntagme ; par exemple "démarche institutionnelle" où l'étiquette S0232145.C0016928(T033 T040), attachée à "démarche" fait émerger les types T033 (*finding*) et T040 (*organism function*)[18], alors qu'ici "démarche" est l'acte de "faire une démarche".

Les résultats de cette estimation figurent dans le tableau 3b. Nous notons que sur la base de ces échantillons, environ la moitié des syntagmes ont bénéficié d'une décision valide.

4.2. *Relations sémantiques et têtes verbales*

La question de l'adéquation entre, d'une part, la projection des relations sémantiques sur les phrases et, d'autre part, le sens de ces phrases, est un point important, en

17. Le type T047 correspond à *disease or syndrome*.
18. Au sens par exemple de la *démarche tabétocérébelleuse*.

particulier pour choisir l'étiquette à étudier. Pour l'aborder, nous avons procédé à la classification ascendante hiérarchique (**[CAH-RL-VERB]**) des relations sémantiques dans l'espace des têtes verbales, ainsi que nous l'avons vu dans la section 3.3.2, ceci afin de percevoir le rapport entre verbes (supposés corrélés avec le sens de la phrase) et relations sémantiques induites par l'étiquetage. Le dendrogramme obtenu est re-

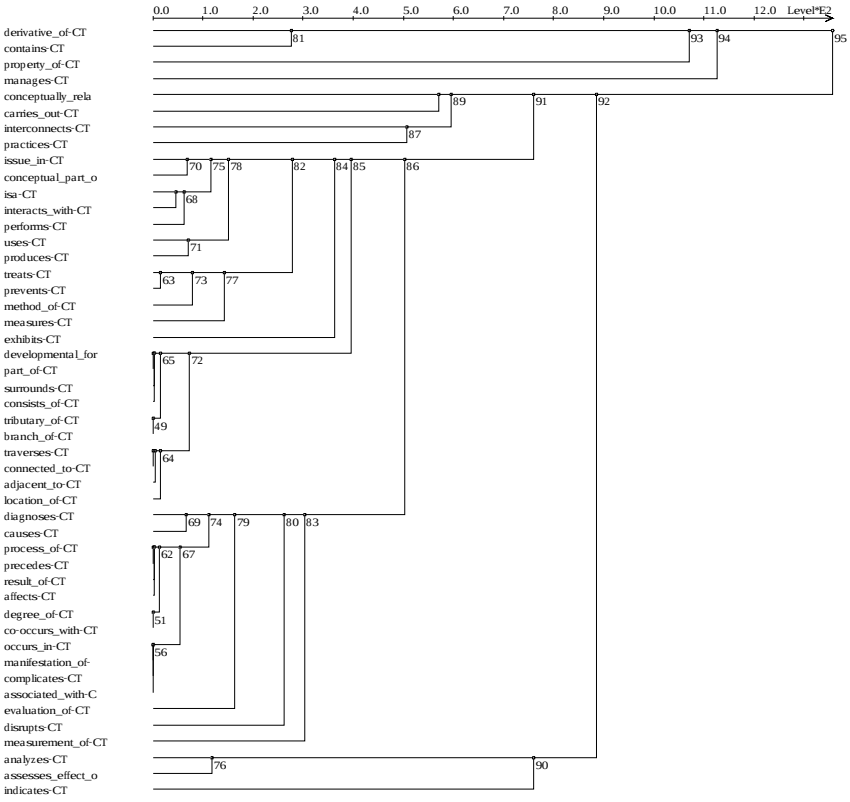

Figure 1. *Dendrogramme de **[CAH-RL-VERB]** relations sémantiques classées par les têtes verbales.*

présenté sur la figure 1. A sa lecture, on constate l'apparition d'une hiérarchie très différente de celle proposée par le réseau sémantique de l'UMLS. Cependant, on peut distinguer la formation de quelques groupes, dont les plus remarquables sont, en les désignant par leurs numéros de nœud :

(72) qui se subdivise trés distinctement en les nœuds **(64)** et **(65)**. Ce regroupement est constitué des relations qui héritent[19] de T132 (*physically-related-to*) et T189

19. Au sens de l'UMLS.

(*spatially-related-to*). Nous disons que c'est un regroupement propre à la description de structures physiques (anatomie). Au sein de **(72)**, **(64)** se spécialise dans les relations de type spatial (*traverses*, *adjacent-to*, *location-of*), et **(65)** se spécialise dans les relations de type structurel (*part-of, consists-of, tributary-of, branch-of*) ;

(67) qui se subdivise également très distinctement en **(56)** et **(62)**, regroupe les relations qui indiquent un lien de cause à effet : *process-of, precedes, result-of, affects*, *occurs-in*, *manifestation-of, complicates* ;

(69) qui regroupe uniquement *diagnoses* et *causes* ;

(63) qui regroupe uniquement *treats* et *prevents*.

La construction de ces groupes cohérents est un signe positif. En effet, n'étant due qu'à des proximités entre profils par rapports aux têtes verbales, elle pose l'hypothèse que les étiquettes (relations sémantiques) de ces groupes sont souvent projetées conformément à leurs sens.

Le dépouillement du dendrogramme nous a incité à nous pencher sur le groupe de relations {*treats, prevent*} dont le sens est également un avantage en sa faveur (premier critère de sélection évoqué dans la section 3.3.2). Ainsi, le tableau 4 énumère les verbes qui contribuent le plus à son excentricité ; nous dirons que ce sont les verbes qui particularisent le plus les phrases où l'on a fait apparaître cette relation, par rapport aux autres phrases. Les verbes ont été conservés de manière à atteindre 30 % de l'excentricité du groupe. Cette liste nous semble favorable : on *envisage* tel traitement pour telle pathologie, on *traite* tel cas par tel acte, l'apparition de tel signe *justifie* le recours à telle procédure, etc. L'ensemble de ces critères nous a fait élire la relation *treats* comme un bon candidat pour des investigations supplémentaires.

Verbe	Part d'excentricité	Part d'excentricité cumulée
envisager	0,0599	0,0599
traiter	0,0465	0,1064
justifier	0,0360	0,1424
discuter	0,0331	0,1755
donner	0,0295	0,2050
modifier	0,0275	0,2325
rédiger	0,0264	0,2589
proposer	0,0241	0,2830
recommander	0,0209	0,3039

Tableau 4. *Verbes les plus contributifs au couple {*treats, prevent*}*

4.3. *La relation* treats *comme type d'entité nommée*

Les analyses que nous venons de décrire nous ont conduits à privilégier la relation *treats* comme type d'entité nommée à étudier en premier lieu [20]. Nous montrons ici

20. La relation *treats* relie des procédures thérapeutiques, des substances pharmaceutiques ou du matériel médical avec des pathologies, des symptômes, des empoisonnements ou des problèmes

les résultats de cette étude, qui a consisté à mesurer l'efficacité de cette étiquette, en fonction des sources des documents explorés, et dans le contexte d'un type de question médical précis.

Par un retour au texte, nous avons ainsi mesuré la qualité de l'étiquetage par la relation *treats*. Le critère que nous avons retenu pour cela est la précision obtenue lorsque la présence de l'étiquette *treats* est utilisée comme unique indice pour extraire des phrases suffisantes en tant que telles pour répondre à une question de type "quel est le traitement envisagé pour ..."[21]. Autrement dit, sont comptées positivement les phrases dans lesquelles apparaît sans ambiguïté l'association entre un cas clinique (symptôme, profil de malade) et un acte ou produit thérapeutique ; nous avons *a contrario* considéré comme cas négatifs les phrases où soit était absente toute référence à un traitement précis, soit une telle référence était présente, mais l'un des deux termes (objet ou instrument du traitement) était absent de la phrase parce qu'entendu par le contexte. Par exemple, l'extrait suivant est positif :

> "*Chez les patients au stade de cirrhose , l' utilité du traitement par l' interféron - alpha sur la survie et ou la prévention des complications de la cirrhose (carcinome hépatocellulaire notamment) est démontrée.*"

alors que celui-ci est négatif :

> "*L' ostéoporose peut être secondaire à un état pathologique ou à la prise de certains médicaments.*"

Dans ce dernier exemple, la relation *treats* existe bel et bien entre les termes "médicament" et "ostéoporose", mais "médicament" est un terme trop générique pour valider cet exemple. D'autres cas négatifs peuvent surgir par exemple lorsque la relation, supposée sur la base de la cooccurrence de deux termes, n'existe pas en réalité dans la phrase, comme cela se produit dans le fragment fictif "... en cas de diabètes chez des consommateurs d'aspirine ...".

Les mesures ont été faites sur la base d'échantillons propres à chaque source constitutive de notre corpus ; les résultats figurent dans le tableau 5 dans lequel σ est une estimation de l'écart-type du taux de positifs dans le corpus de travail pour la source considérée[22]. La lecture de ce tableau nous montre que la relation *treats* a un comportement d'étiquetage très différent en fonction de la source des documents ; les cas extrêmes sont **[SENAT]** (faible présence et faible précision) et **[CANCER]** (présence plus importante, et bonne précision).

Les sources utilisées ont des vocations distinctes : **[SENAT]**, qui est une sous-partie du site du Sénat indexée par CISMeF et dont le contenu plutôt législatif porte

---

anatomiques. Dans le métathésaurus, plus de 100 000 couples de concepts sont établis sur cette relation.

21. Qui constitue l'un des types de questions courantes en médecine générale pour lesquelles un système de QR constituerait un support appréciable aux praticiens dans leur pratique quotidienne [ALP 01].

22. Pour **[AFSSAPS]** et **[ORPHA]** l'écart-type est rigoureusement nul du fait que toutes les phrases étiquetées sont dans l'échantillon.

	Nombre de phrases	Nombre de phrases étiquetées '*treats*'	Échantillon	positifs (%)	σ (%)
SENAT	199372	1265 (0,6 %)	200	10	2,1
CANADA	90986	2743 (3,0 %)	200	16	2,6
CHUROUEN	10232	230 (2,2 %)	200	19	2,8
UROUEN	14799	621 (4,2 %)	200	20	2,8
AFSSAPS	5187	202 (3,9 %)	202	20	0,0
ANAES	125659	4174 (3,3 %)	200	22	2,9
ORPHA	1460	25 (1,7 %)	25	27	0,0
CANCER	47356	2325 (4,9 %)	200	32	3,3

Tableau 5. *Précision de* treats, *pour les associations traitements-cas*

par exemple sur des thèmes de santé publique (environnement, toxicologie, ...) est ouvert au grand public ; **[CANADA]** est un portail grand public lui aussi, mais à thématique médicale plus prononcée. A contrario, **[CANCER]** est un site très spécialisé, à contenu technique. Le tableau 5 fait apparaître l'existence d'un gradient de précision de l'étiquetage par *treats* (selon le critère de positivité que nous avons fixé plus tôt) suivant l'axe **sites généraux-sites spécialisés**[23]. Il est important de noter que les écarts relevés dans le tableau ne proviennent pas de différences dans la quantité d'étiquettes *treats* posées, mais relèvent bien de différences de niveaux de langage à travers les corpus.

5. Discussion

La présence de l'étiquette *treats* ne constitue pas un critère suffisant pour estimer la fiabilité d'un fragment de texte à apporter une réponse à une question ; le tableau 5 fait ressortir des précisions globalement faibles. D'autres éléments (mots clés, verbes, synonymes...) sont toujours nécessaires. Néanmoins il est apparu que l'apport de l'étiquette est sensible à la source des documents étiquetés, puisque sur des corpus à forte technicité (par exemple pour **[CANCER]**), la précision peut dépasser 30 %.

Cette étude nous a donc appris que la mise en œuvre d'un étiquetage par l'UMLS dans un contexte de QR doit prendre en compte la provenance des documents au sein desquels la recherche est effectuée. Par exemple, pour des sources à vocation générale, le recours à ces étiquettes serait désactivé, et réactivé au contraire pour des sources à connotation scientifique. Une autre voie serait de mettre au point un système de pondération fonction du type de documents. Le rôle important de la source documentaire est à rapprocher de travaux sur l'étiquetage de rapports de radiologie [HUA 03], lorsque les auteurs remarquent l'importance que prennent le type de rapports indexés et la zone dans le rapport, dans la qualité d'une indexation par des concepts UMLS.

23. Des analyses de correspondances effectées sur l'ensemble des documents nous ont permis d'affiner cette idée, en dévoilant la structure globale du corpus apportée par l'étiquetage [DEL 04].

D'autre part, les types sémantiques de l'UMLS se projettent différemment en fonction de l'origine du document. Des analyses de correspondances montrent que l'étiquetage obtenu confère au corpus une structure cohérente vis-à-vis des vocations respectives de chaque source [DEL 04]. Cette structure pourrait fournir des indices supplémentaires utiles.

Nous pensons qu'une étude systématique similaire à celle qui a été faite sur *treat* mérite d'être menée sur l'ensemble des relations sémantiques, pour avoir une appréciation plus complète. Ainsi, l'émergence des nœuds **(63)** et **(69)** de **[CAH-RL-VERB]** est prometteuse, car elle ouvre la perspective de l'existence de marqueurs de zones à thématique diagnostique ou thérapeutique, marqueurs en provenance soit des têtes verbales, soit des relations sémantiques. De tels marqueurs sont bien adaptés à des questions telles que "quelle est la cause de ce symptôme", ou "que faire face à ce signe", mais leur capacité en ce sens doit être validée au cas par cas.

Il nous manque encore des indicateurs très importants, que nous n'avons pas pu estimer. En particulier, le rappel, qui indique la part des documents trouvés au sein de l'ensemble ciblé, est absent de notre étude, et nous pensons qu'il sera faible.

Nous avons procédé, parallèlement à l'étude quantitative, à des analyses qualitatives manuelles d'échantillons du corpus de travail. Il en ressort que les concepts de l'UMLS permettent d'autant moins de représenter l'information que l'on s'éloigne du cœur du domaine médical ; sur des textes généraux, lorsque l'on trouve un terme de l'UMLS à projeter, il arrive que le sens dans l'UMLS de ce terme soit distinct du sens entendu dans le texte. D'une manière générale, dans l'échantillon que nous avons observé, nous avons constaté que les types sémantiques de l'UMLS ne couvrent que partiellement les concepts rencontrés. En effet, les concepts et relations UMLS sont restreints strictement au domaine médical. On se trouve contraint de *déformer le sens des termes* UMLS pour représenter le sens des phrases. De nombreux énoncés textuels se rapprochent de la langue générale. Ainsi, il est difficile de représenter certaines formes avec le métathésaurus.

Un autre problème évident dans le cadre qui nous occupe est celui des anaphores, sur lequel nous avons fait l'impasse en nous intéressant uniquement aux cas où la réponse pouvait être contenue intégralement dans une même phrase. L'analyse qualitative nous a montré à quel point cette limitation pouvait être pénalisante.

Dans la suite que nous projetons de donner à ces travaux figurent donc, à coté de l'analyse *post mortem* suite à la campagne EQuER du système de QR associé à cette étude, le renforcement des techniques de traitement automatique des langues, et une systématisation de l'étude quantitative des étiquettes.

6. Bibliographie

[ALP 01] ALPER B. S., STEVERMER J. J., WHITE D. S., EWIGMAN B. G., « Answering Family Physicians' Clinical Questions Using Electronic Medical Databases », *J Fam Pract*, vol. 50, n° 11, 2001, p. 960–965, `http://www.jfponline.com/content/2001/11/`

jfp_1101_09600.asp.

[ARO 01] ARONSON A. R., « Effective Mapping of Biomedical Text to the UMLS Metathesaurus : The MetaMap Program », vol. 8, 2001.

[BEN 79] BENZÉCRI J.-P., « *Correspondances* », vol. 1, Dunod, 1979.

[BEN 84] BENZÉCRI J.-P., « *La taxinomie* », vol. 1, Dunod, 1984.

[DAR 00] DARMONI S. J., THIRION B., LEROY J. P., DOUYÈRE M., BAUDIC F., PIOT J., « CISMeF : a structured Health resource guide for healthcare professionals and patients », C.I.D., 2000.

[DEL 04] DELBECQUE T., « Structuration de corpus médicaux par l'UMLS. Utilisabilité comme source d'entités nommées pour les systèmes de Questions-Réponses. », Rapport de DEA, Informatique Médicale, 2004, Université Paris 5.

[FER 02] FERRET O., GRAU B., HURAULT-PLANTET M., ILLOUZ G., JACQUEMIN C., « *Quand la réponse se trouve dans un grand corpus* », vol. 7 de *Ingénierie des systèmes d'information*, Hermes Sciences, 2002.

[HER 94] HERSH W. R., BUCKLEY C., LEONE T. J., HICKAM D. H., « OHSUMED : An interactive retrieval evaluation and new large test collection for research », *17th ACM SIGIR*, 1994, p. 192–201.

[HUA 03] HUANG Y., LOWE H. J., HERSH W. R., « A Pilot Study of Contextual UMLS Indexing to Improve the Precision of Concept-based Representation in XML-structured Clinical Radiology Reports. », vol. 10, n° 6, 2003, p. 580–587.

[MCC 89] MCCRAY A. T., « The UMLS semantic network », IEEE, 1989, p. 503–507.

[MCC 95] MCCRAY A. T., NELSON S. J., « The Semantics of the UMLS Knowledge Sources », vol. 34, n° 1/2, 1995.

[POI 03] POIBEAU T., « *Extraction automatique d'information* », Hermes Sciences, 2003.

[SCH 94] SCHMID H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *International Conference on New Methods in Language Processing*, Manchester, UK, 1994, p. 44–49.

[ZWE 03] ZWEIGENBAUM P., « Question Answering in Biomedicine », DE RIJKE M., WEBBER B., Eds., *Workshop on Natural Language Processing for Question Answering, EACL 2003*, Budapest, 2003, ACL, p. 1–4.

[ZWE 04] ZWEIGENBAUM P., « L'UMLS entre langue et ontologie : une approche pragmatique dans le domaine médical », vol. 18, 2004, p. 111–137.