UNIVERSIDADE DA BEIRA INTERIOR
Artes e Letras

Lexique Médical Unifié pour le Portugais (UMLP)

Version corrigée

Isabel Maria Clara Marcelino

Tese para obtenção do Grau de Doutor em
Letras
(3º ciclo de estudos)

Orientador : Prof. Doutor João Malaca Casteleiro
Co-orientador : Prof. Doutor Gaël Harry Dias
Co-orientador : Prof. Doutor José Martinez de Oliveira

Covilhã, 29 de Junho de 2012

Cette thèse a été réalisée dans le cadre du programme de formation de ressources humaines (*Programa Operacional Potencial Humano* – POPH) du Cadre de Référence Stratégique National (*Quadro de Referência Estratégico Nacional* – QREN), et a bénéficié du support financier de la Fondation Portugaise pour la Science et la Technologie (*Fundação para a Ciência e a Tecnologia* – FCT) avec la bourse de Recherche SFRH/BD/29625/2006.

À la mémoire de mes grands-parents et de Madame Vieille.
À ma famille et à mon amour.

Remerciements

Pour les personnes très émotives qui, comme moi, ont énormément de mal à exprimer leurs sentiments sans verser de larmes, il est très difficile de dire à quel point certaines personnes ont eu un impact considérable tout au long des années de réalisation de cette thèse. Mais l'avantage de l'écriture, est que je n'ai pas à m'exprimer à voix haute devant une audience, pour pouvoir remercier qui de droit.

Ces nombreuses années d'étude ont vu défiler au Labtech (notre laboratoire de recherche) de nombreuses personnes qui ont marqué cette longue route. Comme je ne peux pas mentionner tout ce petit monde, je leur dit un grand merci à tous.

Cependant il m'est impossible de ne pas mentionner certaines personnes :

Je tiens tout d'abord à remercier le Professeur Helena Morgadinho, qui m'a toujours orientée dans mes études depuis que j'ai fait mes premiers pas à la faculté. Je la remercie pour son aide précieuse, son soutien et ses nombreux conseils. Mais aussi, car c'est grâce à elle que j'ai fait la connaissance du Professeur Gaël Dias.

Beaucoup plus un ami qu'un directeur de recherche, Gaël Dias m'a fait découvrir une autre facette du TAL. Je tiens à le remercier très sincèrement pour tout ce qu'il m'a appris, pour son aide et son soutien constant, et pour m'avoir toujours redonné du punch dans les moments difficiles. Sans lui, cette thèse n'aurait aboutit.

Je tiens également à remercier mes deux autres directeurs de recherche, les Professeurs João Malaca Casteleiro et José Martinez de Oliveira, sans qui, l'idée de ce sujet de thèse ne serait pas survenue.

Je dois très particulièrement remercier Rumen, pour son amitié, sa patience et sa compréhension dans tous les moments. Mais surtout, pour son soutien inébranlable vis-à-vis de mon travail, pour tout ce qu'il m'a appris en informatique et en LaTeX, et pour tous les programmes qu'il a créés afin de me faciliter la vie dans l'analyse des données. Mille mercis pour tout.

Je ne peux évidemment pas oublier mon collègue et ami Sebastião Pais, une personne exceptionnelle, toujours prête à me conseiller, épauler dans les moments difficiles. Un très grand merci à lui, pour avoir été toujours présent.

Mes amis ont énormément contribué à la résolution de cette thèse grâce à leur soutien constant et nos moments de rigolades pour échapper au stress. Un grand merci, en particulier, à Sylvie et à Séverine.

Pour finir, je tiens à remercier mes parents, sans qui je n'aurais jamais fait d'études supérieures. Donc je leur dois un très grand merci pour m'avoir aidée financièrement et surtout moralement. Je ne peux pas oublier mon frère Fernando et sa petite famille, mon cousin Luis et toute ma famille. Un grand merci à tous pour votre soutien et votre patience pour avoir supporté mes sauts d'humeur constants. Cette thèse vous est entièrement dédiée.

Et je garde le mot de la fin pour remercier du fond de mon cœur mon amour Simão, pour m'avoir supportée et soutenue tout au long de ce travail.

Résumé

La médecine est un domaine caractérisé par une terminologie spécifique. C'est pour cela qu'elle a une position très spéciale, non seulement en raison du nombre impressionnant de termes mis en jeu, mais aussi par les efforts internationaux déjà consacrés pour établir des terminologies standardisées. Ces terminologies, créées principalement de forme manuelle, jouent un rôle clé dans le traitement de l'information et des connaissances médicales. Pour le portugais, il n'existe qu'une seule grande terminologie standardisée, et ce, pour la variante brésilienne.
C'est pour cette raison que notre travail consiste à créer et à maintenir (semi-)automatiquement un dictionnaire électronique unifié et évolutif pour le portugais luso-africain, à partir de diverses terminologies collectées sur la toile, également de forme (semi-)automatique. Le codage de ces terminologies selon un même langage informatique de balisage générique a relevé plusieurs difficultés qui ont dû être résolues, afin de pouvoir faire l'unification intra-terminologique. Cette étape a souligné plusieurs problèmes que nous avons résolu de diverses formes en fonction du type de problème : soit par l'application des règles du Nouvel Accord Orthographique, soit par une distance d'édition.
Ce n'est qu'après être passé par ces différentes étapes que nous pouvons alors fusionner nos données afin de corriger les problèmes d'unification inter-terminologiques alors apparus lors de la fusion. C'est ainsi que nous obtenons le plus grand dictionnaire médical luso-africain de plus de 81.000 termes, où sont également précisées, quand disponibles, les variantes brésiliennes.

Mots-clés

Terminologie médicale portugaise ; thesauri ; unification

Resumo alargado

A integração de sistemas de terminologia padronizada em um sistema de representação unificada de conhecimento para a biomedicina formou uma área chave da investigação nos últimos anos. A *Unified Medical Language System* (UMLS) é a base de conhecimento médico mais conhecida que combina o Metathesaurus, o léxico ESPECIALISTA e a Rede semântica.

No entanto, o UMLS é principalmente dedicado ao idioma Inglês. De fato, apenas algumas línguas são incluídas em seu núcleo, cuja cobertura é muito limitada. Por exemplo, Bodenreider & McCray (1998) mostram que apenas uma mínima percentagem da terminologia médica francesa está incluída no UMLS. Consequentemente, diferentes projectos foram aparecendo, como o UMLF (Zweigenbaum *et al.*, 2003) para o francês e os esforços do Instituto Alemão de Documentação e Informação Médica para produzir dados para a língua alemã para o UMLS original. Mas, a maioria das metodologias utilizadas até agora para construir um UMLS são baseadas em usar o original ou a versão traduzida do thesaurus MeSH (*Medical Subject Headings*), que é o mais importante recurso do Metathesaurus.

No nosso ponto de vista, a fim de construir uma base de conhecimento médico dinâmica, a linguagem médica precisa de ser recolhida por meio de diversos vocabulários médicos controlados, existentes na forma de terminologias, dicionários ou glossários, a fim de obter um leque representando diversas áreas e géneros da medicina. De facto, embora o MeSH seja um recurso valioso, são necessárias constantes actualizações manuais para acompanhar a dinamicidade da linguagem. Demais, a manutenção do MeSH e do UMLS é cara, demorosa e pode não reflectir a realidade da linguagem médica em tempo útil. Além disso, é definida com base em indexação manual, o que pode não reflectir a realidade das relações entre conceitos.

Para evitar tais limitações, propomos construir semi-automaticamente um Metathesaurus médico unificado para o Português chamado UMLP (*Unified Medical Lexicon for Portuguese*). A nossa ideia é, primeiro, construir um léxico unificado baseado em dicionários electrónicos, glossários e taxonomias *online*, na Wikipédia e no *Wikcionário*. Depois, com base nos thesauri criados automaticamente a partir dos recursos, pretendemos construir o Metatesaurus Português.

Para isso, depois de ter codificado todos esses recursos em uma única e mesma linguagem informática, vamos resolver os problemas de unificação intra-terminológicos com a aplicação das regras do Novo Acordo Ortográfico e com o cálculo da distância de Levenshtein. Só depois é que podemos fazer a fusão de todos os nossos dados e resolver os novos problemas de unificação surgidos, para chegar à criação do maior léxico médico unificado para o português luso-africano (com as informações do português do Brasil, sempre que disponíveis), com aproximadamente 81.000 entradas, juntamente com seus respectivos caminhos taxonómicos vindos dos diferentes recursos.

Abstract

Medicine is a field characterized by a terminology. That is why it takes a very special not only because of the impressive number of words involved, but also international efforts already devoted to establish standardized terminologies. These terminologies, created primarily as manual, play a key role in information processing and medical knowledge. For Portuguese, there is only one major standardized terminology in the Brazilian variant. For this reason our job is to create and automatically maintain a unified and scalable electronic dictionary for Portuguese European (while maintaining the Brazilian standard), from various terminologies collected on the Web also form automatically. The coding of these terminologies in a very generic markup language, has identified several challenges that had to be resolved in order to move to the intra-unification of terminology. This stage highlighted several problems we solved in different forms depending on the type of problem : either by applying the rules of the New Orthographic Agreement, or by an edit distance. Only after going through these steps so that we can combine our data to correct the problems of inter-unification of terminology then appeared in the merger. Thus we get the largest dictionary Luso-African about 81.000 words, which are also included, when available, the Brazilian variants.

Keywords

Portuguese medical terminology ; thesauri ; unification

Table des matières

Table des figures

Liste des tableaux

Liste des Acronymes

AO	Accord Orthographique de la Langue Portugaise
CHCB	Centre Hospitalier *Cova da Beira*
CIM	Classification Internationale des Maladies
DeCS	*Descritores em Ciências da Saúde*
DPLP	*Dicionário Priberam da Língua Portuguesa*
DTD	Définition de Type de Document
GMTM	Glossaire Multilingue de Termes Médicaux techniques et populaires en huit langues européennes
MEDLINE	*Medical Literature Analysis and Retrieval System Online*
MeSH	*Medical Subject Headings*
NLM	*National Library of Medicine*
OMS	Organisation Mondiale de la Santé
SNOMED CT	*Systematized Nomenclature of Medicine – Clinical Terms*
TAL	Traitement Automatique des Langues
UBI	Universidade da Beira Interior
UMLS	*Unified Medical Language System*
UMLF	*Unified Medical Lexicon for French*
WWW	*World Wide Web*
XML	*eXtensible Markup Language*

Chapitre 1

Introduction

"A linguagem científica deve ser : exata, para não propiciar equívocos ; simples, para que seja bem compreendida ; concisa, para economizar tempo de leitura e de espaço nas publicações."

Simônides Bacelar

"[Le langage médical] est un art, de savoir parler médecine dans un langage non médical."

Goodman

1.1 Problématique

Depuis plusieurs années, en raison de l'émergence des nouvelles technologies de l'information et de la communication, l'information médicale est devenue de plus en plus disponible et accessible. Le domaine de la médecine contient maintenant une grande quantité de documents électroniques et de ressources linguistiques et terminologiques en différentes langues, telles que corpora, lexiques, thesauri ou ontologies. Pourtant, ce vaste domaine présente certaines particularités. Il se caractérise par la richesse, la complexité et l'actualisation fréquente de son vocabulaire. Cette dynamique contribue largement à la fréquence d'accès à l'information médicale et à la nécessité de l'actualiser. La disponibilité de grandes bases de données d'information médicale, ne garantit pas cependant, la qualité de cette dernière. En effet, à l'exception de l'*Unified Medical Language System* (UMLS) (Humphreys *et al.*, 1998), de nombreuses sources disponibles sur la toile manquent de cohérence, complétude et autorité. C'est une grande préoccupation dans un domaine spécialisé comme la médecine, où la précision et la validité des informations sollicitées sont des critères importants.
Les ressources de base de la langue naturelle sont des ressources cruciales pour l'information médicale. En plus du lexique spécialisé UMLS, qui est un Système d'Unification du Langage Médical conçu pour la langue anglaise, d'autres lexiques médicaux ont été développés, entre autres, pour l'allemand (Widdows *et al.*, 2003) ou le français (Zweigenbaum *et al.*, 2003). Pour la langue portugaise, il existe plusieurs ressources de terminologies médicales, mais la principale est le *Descritores em Ciências da Saúde* DECS (Tardelli, 2007) qui est un lexique médical structuré et trilingue, dérivé de la traduction manuelle du grand thésaurus *Medical Subject Headings* (MeSH), qui fait partie de l'UMLS. Indépendamment d'avoir été créé manuellement et d'être statique (i.e. il n'existe pas d'actualisations systématiques du propre lexique), le DeCS fait référence au portugais variante brésilienne. Par exemple, l'équivalent pour la norme luso-africaine de *Anomalias Congênitas* est le terme *Anomalias Congénitas*. De la même façon, *Cisto*

s'écrit *Quisto* et *Câncer*, *Cancro* dans la norme luso-africaine. Et il couvre des domaines non spécifiques à la Médecine, tels que Phénomènes Sociaux ou encore Économie. Ces caractéristiques laissent une zone d'intervention au niveau scientifique pour la construction d'un dictionnaire cohérent, complet, autoritaire et dynamique pour le portugais dans sa norme luso-africaine.
Ces divergences de la langue portugaise peuvent causer des problèmes de compréhension de l'information médicale. C'est donc pour cela qu'il existe le besoin de créer un lexique médical unifié pour permettre plus d'entente au niveau terminologique entre les variantes luso-africaine et brésilienne du Portugais.

Au Portugal, il existe trois grandes Écoles de Médecine dans lesquelles nous trouvons une grande différence dans la terminologie médicale. Par exemple [1], selon l'École, les termes tels que *Lípidos* ou *Lipídeos* font référence au même concept bien qu'ils aient une graphie différente. Un autre exemple montre que le terme employé pour *anemia* peut se prononcer et s'écrire "anemia" [ɐnimˈiɐ] [2] à Porto et à Coimbra, et "anémia" [ɐnˈɛmiɐ] à Lisbonne. Ces cas ne nécessitent que d'une unification au niveau morphologique ou phonologique. Cependant, des problèmes sémantiques peuvent également se produire. Par exemple, *aborto* et *abortamento* font référence habituellement à la notion de *aborto* à Lisbonne (comme défini dans toutes les terminologies recueillies). Néanmoins, à Porto, *abortamento* n'est pas synonyme de *aborto*, sachant que ce dernier est le produit de *abortamento*. Ces derniers sont des exemples clairs de possibles ambiguïtés dans la communication de l'information médicale.
Par conséquent, il existe une évidente ambiguïté des termes médicaux dans la langue portugaise, aussi bien au niveau de la morphologie, de la phonétique, qu'au niveau de la sémantique. Ambiguïté, qui est un problème évident pour la compréhension et l'interprétation automatique du langage médical.

1.2 Objectifs et travail proposé

Dans ce contexte, le projet de création de l'UMLP (*Unified Médical Lexicon for Portuguese*) est survenu de la nécessité évidente des difficultés trouvées par les professionnels de la santé, en particulier à la Faculté des Sciences de la Santé (FCS) de l'Université de Beira Interior à Covilhã – Portugal.
En effet, la FCS est une école récente qui n'a pas de tradition. Elle doit alors s'appuyer sur les autres écoles plus anciennes pour créer sa propre terminologie ou inventer de nouveaux termes, si ceux qui existent déjà dans les autres écoles ne lui conviennent pas. De plus, d'un point de vue historique, la FCS a été pilote dans une nouvelle forme d'apprentissage et d'évaluation de ses étudiants. Les cours sont assistés par ordinateur ainsi que les examens réalisés sur une plateforme de *e-learning*. Dans ce contexte, dans lequel la correction ou l'évaluation des épreuves est à charge de l'ordinateur, il existe le besoin évident de désambiguïsation lexicale.

Ce travail a donc pour objectif la construction (semi-)automatique d'un lexique médical unifié et évolutif pour le portugais avec ses respectifs thesauri, ou au sens large du terme, de ses respectives structures hiérarchiques. En effet, contrairement à des bases de données créées et maintenues manuellement, telles que l'UMLS ou le DeCS, nous prétendons créer un dictionnaire

1. Ces trois exemples sont tirés de l'Annexe A, où nous présentons une liste de problèmes d'unification répertoriés par le Professeur José Martinez de Oliveira, médecin et professeur à l'hôpital de Covilhã.
2. Transcription phonétique obtenue dans le *Dicionário da Língua Portuguesa Contemporânea* de l'Académie des Sciences de Lisbonne (Academia das Ciências de Lisboa, 2001).

électronique collaboratif, actualisable automatiquement et le plus exhaustif possible, à partir de ressources de la toile, comme les dictionnaires et glossaires en ligne, la Wikipédia [3], et le *Wikciónario* [4]. La langue étant constamment en évolution, un lexique médical doit suivre ce dynamisme en actualisant de forme automatique ses entrées [5] selon les modifications qui se produisent dans la langue portugaise et en particulier dans le langage médical.
Il sera donc important de fournir aux étudiants et aux personnes qui travaillent dans la FCS, des définitions correctes pour chaque terme médical avec une mise à jour constante, grâce à ce dictionnaire informatisé.

1.3 Structure de la dissertation

Afin de rendre compte de tous les aspects importants qui ont permis la création de cette nouvelle ressource terminologique (le plus grand dictionnaire électronique de médecine en portugais luso-africain, avec une méthode d'extraction des données, différente des habituelles extractions manuelles), nous avons divisé notre dissertation en sept chapitres.
Nous avons, dans ce premier chapitre, présenté la problématique de l'objet de l'étude, les objectifs de cette recherche, ainsi que la méthodologie proposée, et pour finir son organisation.
Dans un deuxième chapitre, nous exposons les plus importantes ressources terminologiques existantes du domaine médical, tout d'abord au niveau non-lusophone, puis dans une deuxième partie, au niveau lusophone. Nous séparons ces deux spécifications de langue afin de bien comprendre que la grande partie des terminologies existantes en langue portugaise, sont le fruit de traductions manuelles à partir de ressources en langue anglaise.
C'est dans le troisième chapitre que nous présentons notre base de données. C'est-à-dire, nous expliquons quelles ressources nous avons utilisées et de quelle façon elles ont été recueillies. Ce chapitre est divisé en deux parties, car nous séparons l'extraction manuelle et automatique des données.
Dans le quatrième chapitre, nous expliquons en quel langage informatique les données ont été extraites, et surtout en quel langage unique elles ont été codifiées afin de maintenir une certaine harmonie entre les différentes terminologies recueillies. De plus, dans une deuxième partie, nous faisons un bref récapitulatif de tous les problèmes rencontrés lors de ce codage, aussi bien d'un point de vue général par type de terminologie, comme d'un point de vue plus spécifique en fonction du type de problème.
Ce n'est que dans le cinquième chapitre que nous commençons à parler de l'unification des terminologies, mais seulement d'un point de vue intra-terminologique. Nous entendons par cela, une unification à l'intérieur de chacune des terminologies extraites. Nous présentons alors trois formes d'unification différentes : la première, par l'application des règles du Nouvel Accord Orthographique de la langue portugaise ; la seconde, par l'analyse des termes d'entrées et pour finir, par les définitions.
Après avoir unifié nos termes à un niveau intra-terminologique, nous pouvons dans un sixième chapitre commencer la fusion de toutes les terminologies extraites et alors faire une unification inter-terminologique. Pour cela, le traitement des données est fait de la même forme que pour l'unification intra-terminologique, mais les problèmes rencontrés sont traités à l'aide de différentes règles que nous avons dû créer. Ce n'est qu'après avoir résolu tous ces problèmes d'unification que nous pouvons alors passer au codage final des données, afin de regrouper toute

3. http://pt.wikipedia.org/wiki/
4. http://pt.wiktionary.org/wiki/
5. Nous considérons par "entrée", le terme à définir avec sa définition et ses respectives informations.

la terminologie médicale.
En guise de conclusions et de travail futur, nous faisons un récapitulatif sur la satisfaction des objectifs, puis nous présentons différents abordages qui pourront être réalisés afin de compléter ce dictionnaire.

Chapitre 2

Révision Bibliographique

Le domaine de la Médecine est un des domaines de spécialité les plus importants traité depuis l'apparition des analyses de systèmes d'information médicale[1]. Il est caractérisé par une riche et complexe terminologie qui ne cesse de grandir en raison de la rapide évolution de la recherche conduite dans ce domaine.
C'est la raison pour laquelle, tout d'abord, nous présentons dans ce chapitre quelques-unes des plus importantes ressources terminologiques existantes dans le domaine médical, aussi bien au niveau non-lusophone, comme lusophone. Et pour finir, nous exposerons notre proposition de travail pour le portugais dans ses variantes luso-africaine et brésilienne.

2.1 Ressources terminologiques non-lusophones du domaine médical

Nous montrons dans ce qui suit des ressources terminologiques et ontologiques existantes et conçues pour le domaine de la médecine. Ces ressources ont été construites pour répondre à des besoins spécifiques et divers : le thésaurus MeSH (cf. section 2.1.1) vise à indexer les connaissances médicales pour la recherche d'information dans des bases documentaires, l'UMLS (cf. section 2.1.2) a pour objectif de faciliter le développement des systèmes informatisés afin d'améliorer l'accès à l'information médicale, la CIM (cf. section 2.1.3) permet le codage des dossiers des patients à des fins statistiques, GALEN (cf. section 2.1.4) était un projet qui avait pour but de construire une ontologie médicale généraliste, ORPHANET (cf. section 2.1.5) répertorie tous les noms de maladies rares et la SNOMED-CT (cf. section 2.1.6) est considérée comme la terminologie multilingue des soins cliniques la plus détaillée au monde.

2.1.1 *Medical Subject Headings*

Le thésaurus *Medical Subject Headings*[2] (MeSH) (National Library of Medicine, 2010) est un vocabulaire contrôlé développé par la *National Library of Medicine* (NLM) aux États-Unis d'Amérique et utilisé pour indexer, cataloguer et rechercher des informations et des documents biomédicaux et en rapport à la santé. La première liste officielle de *Subject Headings* publiée par la NLM était, en 1954, intitulée *Subject Headings Authority List*. Ce fut en 1960, avec la création de l'*Index Medicus* qu'est apparu le nouveau et entièrement revu *Medical Subject Headings*.

Le MeSH est une liste structurée de termes médicaux organisés en une arborescence. Au fur et à mesure que nous descendons dans la hiérarchie, les termes sont de plus en plus spécifiques. Ces

1. Par exemple, le *Mycin* (Shortliffe & Buchanan, 1975) a été un des premiers systèmes experts développé dans le domaine médical.
2. `http://www.nlm.nih.gov/mesh/` (accédé le 25 mai 2010)

derniers sont appelés «descripteurs» car ils expriment de forme précise et spécifique le contenu d'un document. Les 25.588 descripteurs (dans l'édition de 2010)[3] sont divisés en 16 branches principales de l'arborescence. Par exemple, la branche [A] correspond à l'"Anatomie" (*Anatomy*), la branche [B] aux "Organismes" (*Organisms*), la branche [C] aux "Maladies" (*Diseases*). Chacune de ces branches contient plusieurs sous-branches qui constituent les différents niveaux de la hiérarchie. Par exemple, [C01] pour la catégorie "Infections bactériennes et mycoses" (*Bacterial Infections and Mycoses*), [C02] pour "Maladies virales" (*Virus Diseases*) ou encore [C03] pour "Maladies parasitaires" (*Parasitic Diseases*).
De plus, chaque terme du thésaurus MeSH est associé à sa définition, à ses synonymes (quand ils existent) et à sa position dans l'arbre (identifiant hiérarchique). Néanmoins, certains descripteurs peuvent apparaître dans plusieurs branches de l'arborescence, c'est-à-dire, qu'un même terme peut appartenir à diverses catégories du MeSH, et donc, peut avoir plusieurs identifiants. Un identifiant est composé d'un numéro alphanumérique : une lettre qui indique la catégorie (comme C = Maladies) et une série de numéros qui précise la position du terme dans la hiérarchie. Par exemple[4], l'identifiant attribué au descripteur "Hépatite C" (*Hepatitis C*) est [C02.440.440], ce qui signifie : [C] pour "Maladie", [C02] pour la catégorie "Maladies virales" (*Virus Diseases*), [C02.440] pour "Hépatites virales humaines" (*Hepatitis, Viral, Human*) et ainsi de suite.

Virus Diseases [C02]
    Hepatitis, Viral, Human [C02.440]
        Hepatitis A [C02.440.420]
        Hepatitis B [C02.440.435]
        Hepatitis C [C02.440.440]

Ce descripteur contient, en plus de ce dernier, d'autres identifiants qui sont [C02.782.350.350] et [C06.552.380.705.440].

Le MeSH est utilisé par de nombreux systèmes de recherche bibliographique, notamment pour indexer des sites et des documents médicaux. C'est le cas, par exemple, de *Medical Literature Analysis and Retrieval System Online* (MEDLINE) (National Library of Medicine, 2011), une base de données bibliographiques recouvrant tous les domaines des sciences de la vie. Cette base[5] est maintenue par la NLM depuis 1966. Elle est devenue la base de données la plus utilisée pour la recherche bibliographique dans le domaine biomédical. MEDLINE contient plus de 15 millions de références bibliographiques provenant d'environ 70 pays, totalisant ainsi plus de 5.000 sources biomédicales distinctes. Toutefois, les résumés, les titres et les descripteurs sont toujours en anglais. D'ailleurs, les articles écrits dans cette langue sont majoritaires dans la base de données, une fois qu'elle représente près de 85% des références. Grâce à cette énorme base de données, la création de corpora médicaux en langue anglaise est facile à réaliser, au contraire du portugais, où le manque de bases de données de textes biomédicaux est très important.

Il existe de nombreuses traductions du MeSH en langues autres que l'anglais. En l'an 2000, sept (Stuart *et al.*, 2000) d'entre elles (allemand, espagnol, finlandais, français, italien, portu-

3. `http://www.nlm.nih.gov/mesh/intro_preface.html` (accédé le 25 mai 2010)
4. Exemple retiré du MeSH Browser (2010 MeSH) : `http://www.nlm.nih.gov/mesh/2010/mesh_browser/MBrowser.html` (accédé le 25 mai 2010)
5. Accessible grâce au moteur de recherche Pubmed sur le site : `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?`

gais et russe) ont incorporé le Metathésaurus de l'UMLS (cf. 2.1.2). En 2009, elles étaient déjà au nombre de onze (Tardelli, 2009). Le fait qu'il soit traduit en de nombreuses langues, fait de lui un outil amplement référencé dans les bibliothèques et autres institutions du monde entier.

2.1.2 *Unified Medical Language System*

L'*Unified Medical Language System* (UMLS)[6], pour Système d'Unification de la Langue Médicale, est actuellement la ressource terminologique de référence dans le domaine de la Biomédecine. Il a été mis en place dans le but d'améliorer l'accès à l'information médicale à partir de sources diverses : bases de données bibliographiques, bases de données d'enregistrements cliniques et bases de données de connaissances médicales (Lindberg & Humphreys, 1990).

Cette ressource, développée et maintenue par la NLM depuis 1986, est le résultat d'une compilation de terminologies provenant de plus de 150 sources de langues et structures différentes, desquelles les plus connues sont le MeSH (un des plus importants thésaurus qui compose le Metathésaurus de l'UMLS) et la SNOMED-CT, lui donnant le statut de Metathésaurus multilingue. Ce dernier contient la terminologie qui résulte de l'union des vocabulaires de ces différentes sources médicales, préservant les relations qui interviennent entre les termes. Cependant, le Metathésaurus n'est pas une ontologie. Il n'a pas été construit dans ce but, et la tentative de réutilisation comme une ontologie s'est soldé par un échec (Charlet *et al.*, 2006)

L'intérêt de l'UMLS réside dans sa large couverture du domaine médical et dans sa disponibilité. En effet, il est composé de plus de 2,2 millions de concepts et 8.2 millions de termes uniques (version UMLS 2010AA)[7] faisant référence à un concept *(unique concept names)* et indique la relation existante entre les concepts. Ces derniers, de plus en plus nombreux, sont liés entre eux par des relations sémantiques héritées des ressources initiales. Ces relations sémantiques sont principalement des relations paradigmatiques, telles que les relations de synonymie[8] ou d'hyperonymie, ainsi que d'autres relations plus spécifiques.
L'UMLS se base également sur le Lexique SPECIALIST (Browne *et al.*, 2000), un lexique général de la langue anglaise, qui contient un grand nombre de termes biomédicaux avec leurs informations morphosyntaxiques.
De plus, l'UMLS a comme troisième grand pilier (voir la figure 2.1)[9], un ample Réseau Sémantique (RS) (Schulze-Kremer *et al.*, 2004; Zweigenbaum, 2004), où les liaisons entre les types sémantiques fournissent la structure pour ce réseau et représentent des relations importantes dans le domaine de la Biomédecine.

Ce réseau contient 133 concepts sémantiques et 54 relations (National Library of Medicine, 2009). Les noeuds du RS sont représentés par les concepts sémantiques, et les liaisons existantes entre les noeuds fournissent les types de relations existantes dans le réseau (voir la Figure 2.2).

6. `http://www.nlm.nih.gov/research/umls` (accédé le 26 mai 2010)
7. Informations trouvées dans le bulletin technique de mai-juin 2010 (`http://www.nlm.nih.gov/pubs/techbull/mj10/mj10_umls_aarelease.html`) (accédé le 26 mai 2010)
8. La synonymie est représentée implicitement par le fait que deux termes étiquettent un même concept.
9. Exemple tiré et traduit du Tutorial de l'UMLS sur :
`http://www.nlm.nih.gov/research/umls/new_users/online_learning/OVR_001.htm`.

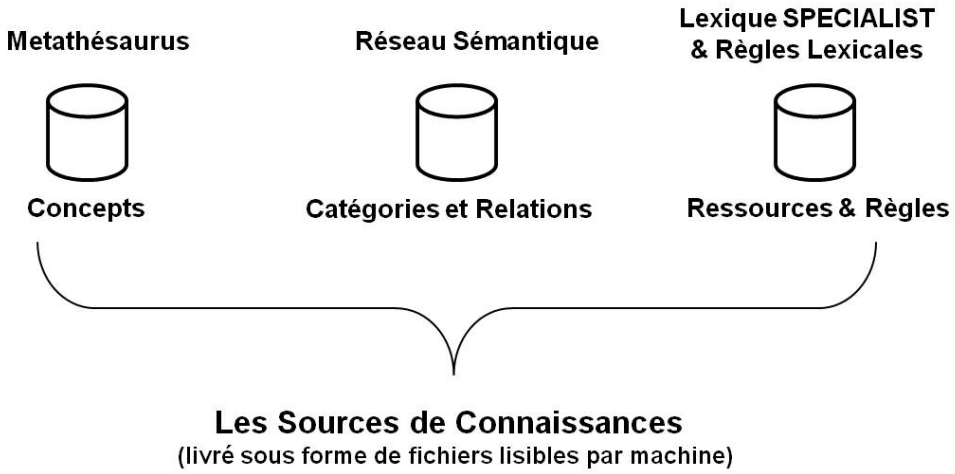


Figure 2.1: Schéma représentant les trois piliers de base de l'UMLS

La principale liaison entre les types sémantiques est la liaison "IS-A" (*is a kind of* = est un type de), qui établit la hiérarchie des types à l'intérieur du réseau et est utilisé pour décider du type sémantique le plus spécifique disponible pour l'attribution d'un concept du Metathésaurus. Par exemple, l'instance "Singe" reçoit le concept sémantique "Mammifère", parce qu'il n'existe aucun concept plus spécifique comme "Primate".
Ce réseau fait de l'UMLS la ressource terminologique du domaine médical la plus amplement exploitée au monde. Elle est très appropriée pour le traitement de l'information biomédicale et donc, elle constitue un outil précieux pour les systèmes de recherche documentaire, notamment pour l'identification, dans des documents médicaux, de concepts spécifiques au domaine biomédical, tels que les gènes, les maladies ou les médicaments.

Quand un concept est additionné au Metathésaurus, il reçoit un identifiant unique et est placé dans la structure du Metathésaurus. Cette dernière dispose de quatre niveaux de spécification [10] :

Concept Unique Identifiers (CUI)
Un concept est un sens. Un sens peut associer de nombreux termes différents. Un des objectifs essentiel de la construction du Metathésaurus est de comprendre le sens prétendu de chaque terme dans chaque vocabulaire d'origine et de relier tous les termes de tous les vocabulaires d'origine qui ont le même sens (les synonymes). Le CUI contient la lettre C, suivi de sept chiffres. (Dans l'exemple de la figure 2.3, le CUI est C0018681.)

Lexical (term) Unique Identifiers (LUI)
Le LUI lie les chaînes de caractères qui sont des variantes lexicales. Ces dernières sont détectées en utilisant un des outils lexicaux de l'UMLS *(Lexical Variant Generator (LVG) program)*. Le LUI contient la lettre L suivie de sept chiffres. (Dans l'exemple de la figure 2.3, nous avons trois variantes lexicales, avec pour chacune d'entre elles, un LUI différent.)

10. L'exemple qui suit a été tiré de :
`http://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_005.htm` (accédé le 26 mai 2010)

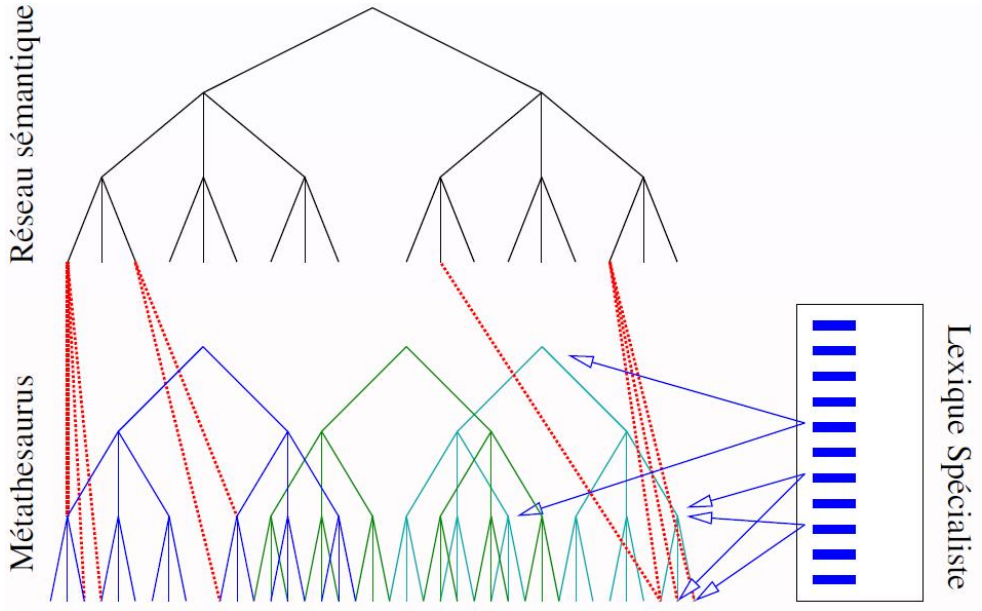


Figure 2.2: Structure de l'UMLS, tirée de Zweigenbaum (2005)

String Unique Identifiers (SUI)
Chaque terme unique (*unique concept name*) ou chaîne de caractères dans chaque langue du Metathésaurus a un identifiant de chaîne de caractères unique et permanent (SUI). Toute variation dans l'ensemble des caractères (majuscules, minuscules ou encore différence de ponctuation) est une chaîne de caractères distincte, avec un SUI différent. Le SUI contient la lettre S suivie de sept chiffres. (Dans l'exemple de la figure 2.3 nous avons quatre chaînes de caractères avec quatre SUI non-identiques.)

Atom Unique Identifiers (AUI)
Les blocs de construction de base ou "atomes" d'où le Metathésaurus est construit sont les termes (*concept names*) ou chaînes de caractères de chacun des vocabulaires d'origine. À chaque occurrence d'une chaîne de caractères dans chaque vocabulaire d'origine est attribué un identifiant unique d'atome (AUI). Si exactement la même chaîne de caractères apparaît plusieurs fois dans un même vocabulaire, par exemple comme un nom alternatif pour des concepts différents, un AUI unique est attribué à chaque occurrence. L'AUI contient la lettre A suivie de sept chiffres. (Dans l'exemple de la figure 2.3 nous avons cinq chaînes de caractères issues de cinq sources différentes, avec cinq AUI différents.)

L'abréviation de la source qui fait référence à chaque chaîne de caractères est indiquée entre parenthèses après la chaîne de caractères.

Cependant, comme la majeure partie des termes introduits dans le Metathésaurus de l'UMLS est en langue anglaise (près de 62% en 2008), l'utilisation de l'UMLS et de son réseau sémantique est plus difficile pour les autres langues. C'est pour cela que l'UMLS a commencé, il y a de nombreuses années, à introduire des termes biomédicaux en d'autres langues que l'anglais. En 2008, le Metathésaurus contenait déjà des termes de dix-sept autres langues dont l'espagnol, le français, le néerlandais, l'italien, le japonais et le portugais. Mais les limitations de l'utilisation de ce système, en prenant comme exemple le français (Bodenreider & McCray, 1998), étaient

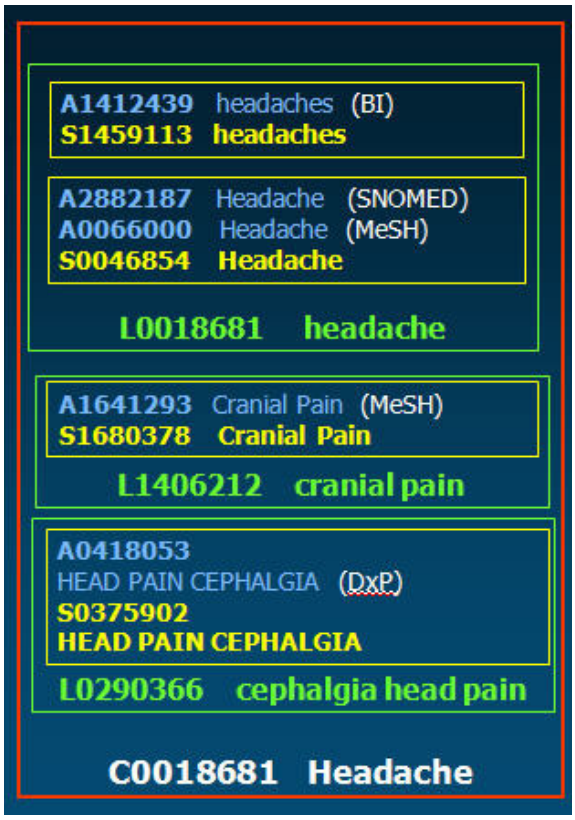


Figure 2.3: Exemple d'identifiants uniques dans le Metathésaurus de l'UMLS

innombrables : traductions partielles, une unique source pour les concepts traduits, ensemble de caractères impropres et absence de règles de correspondance lexicale (*lexical matching*).
En effet, dans cette langue, la terminologie ne recouvrait qu'un très faible pourcentage des concepts de l'UMLS. Cette observation a mené à la création du Lexique Médical Francophone Unifié (UMLF)[11], pour *Unified Medical Lexicon for French* (Zweigenbaum *et al.*, 2003), qui a pour objectif recueillir, unifier et valider les ressources lexicales pour le traitement du français médical. Par une approche monolingue, ce système est destiné à produire un lexique qui contient les variations flexionnelles et dérivationnelles des termes médicaux. Ces informations doivent être encodées dans un format informatique standard afin de favoriser leur intégration dans des systèmes de traitement automatique de la langue médicale.

2.1.3 Classification Internationale des Maladies

La Classification Internationale des Maladies[12] (World Health Organization, 2005), qui a adopté la dénomination "Classification Statistique International des Maladies et Problèmes en rapport à la Santé" lors de sa 10ème révision, étant normalement connue par CIM-10 (en anglais ICD pour *International Classification of Diseases and Related Health Problems*), a été publiée par l'Organisation Mondiale de la Santé (OMS). Elle a été créée en 1990 et pourtant seulement utilisée qu'à partir de 1994.
La CIM-10 est disponible en six langues officielles de l'OMS (anglais, arabe, chinois, espagnol, français et russe) ainsi comme en d'autres 36 langues. Son implémentation au Portugal a eu lieu

11. `http://www-test.biomath.jussieu.fr/umlf/`
12. `http://www.who.int/classifications/icd/en/index.html` (accédé le 26 mai 2010)

avant l'an 2000, et est déjà utilisée dans les statistiques officielles du domaine de la santé en langue portugaise variante brésilienne.
La CIM a pour objectif de classer les maladies, les traumatismes et d'indiquer l'ensemble des motifs de recours aux services de santé. Elle est également utilisée pour identifier les informations de santé sur les causes de mortalité et morbidité dans différents pays. La CIM-10 bénéficie d'actualisations régulières, le numéro 10 correspond à la dernière version d'une série qui a eu ses origines dans la décennie de 1850. Une nouvelle révision de la CIM, prévue pour 2015, est actuellement en cours dans le cadre du projet CIM-11, administré par l'OMS.

La classification dans la CIM-10 est monoaxial et comprend 21 chapitres, desquels 17 concernent des maladies et les 4 restants concernent les signes et les résultats anormaux, les causes de traumatismes, d'empoisonnement ou de morbidité, l'état de santé et les facteurs de recours aux soins. Les maladies sont classées selon plusieurs catégories, telles que : les maladies endocrines (E), les maladies du système nerveux (G), les maladies de l'appareil circulatoire (I), etc., et sont répertoriées selon un degré de gravité. Par exemple, le chapitre des maladies infectieuses recense le plus grand nombre d'entrées car ces maladies sont la première cause de mortalité et de morbidité au monde.
Chaque entrée est identifiée dans la CIM par un code composé de quatre caractères : une lettre qui correspond au chapitre, suivi de deux chiffres pour spécifier les maladies définies à un niveau général, décliné par l'ajout d'un troisième chiffre (derrière un point) pour désigner les diagnostics précis et les formes cliniques. Par exemple, le code A15.9 indique une tuberculose de l'appareil respiratoire, ou encore le code C91.1 désigne une leucémie lymphoïde chronique.

2.1.4 GALEN

Le projet GALEN [13] (*Generalised Architecture for Languages, Encyclopaedias and Nomenclatures in Medicine*), développé à l'Université de Manchester au Royaume-Uni (Rector & Nowlan, 1994), est la première grande initiative à avoir eu comme objectif la construction d'une ontologie pour la médecine. Pour cela, il utilise le langage de représentation GRAIL (*Galen Representation And Integration Language*) (Rector *et al.*, 1997), une variété de descriptions logiques. GALEN vise à faciliter la description des informations cliniques, le codage et le transcodage dans les diverses classifications.
Les concepts primitifs de l'ontologie GALEN sont organisés en arborescence formant un réseau sémantique où les différents concepts sont liés entre eux essentiellement par la relation "IS-A". Cette opération fondamentale donne aux concepts une structure de graphe cyclique dirigé (Zweigenbaum *et al.*, 1996). La version initiale de GALEN (en 1995) comptait sur une hiérarchie de plus de 4.000 concepts. Actuellement, plus de 52.000 entités conceptuelles sont répertoriées, desquelles 818 sont des relations et les autres des concepts. Chaque concept est accompagné d'une déclaration des relations qui doivent ou peuvent le lier à d'autres concepts. Concepts et relations peuvent être combinés librement pour créer de nouveaux concepts structurés.
Dans ce projet, GALEN a associé deux types d'outils : ontologique et linguistique, ce qui rend possible les réutilisations, les reclassifications, les références croisées et les traductions en plusieurs langues à partir des connaissances acquises dans le modèle conceptuel.

13. `http://www.opengalen.org` (accédé le 26 mai 2010)

2.1.5 Orphanet

Orphanet [14] est une base de données sur les maladies rares et les médicaments orphelins en libre accès pour tous publics (Schmidtke, 2007). Elle a été créée en 1997 par le ministère français de la santé (Direction Générale de la Santé) et par l'Institut National de la Santé et de la Recherche Médicale (INSERM). Le portail d'Orphanet a pour objectif de contribuer à l'amélioration du diagnostic, de la prise en charge et des traitements aux patients porteurs de maladies rares.
Ce portail, qui reçoit plus de 20.000 visiteurs par jour, est constitué (i) d'une encyclopédie de maladies rares et de médicaments orphelins, réunissant plus de 8.000 entrées, rédigée par des experts et supervisée par un comité éditorial international ; et par (ii) un répertoire (destiné aux professionnels et malades) des services disponibles dans 35 pays. Ce répertoire inclut des informations sur les consultations expertes, les laboratoires de diagnostic, les projets de recherche en cours et les associations de malades. Ce site permet la recherche de maladies rares par signes cliniques, ainsi comme l'accès au service de mise en contact entre patients, et donne également la possibilité aux patients désirant participer à des recherches cliniques de se faire connaître.

En outre, tous les services d'Orphanet sont accessibles à partir de la page d'accueil du site, disponible en 6 langues : allemand, espagnol, français, anglais et italien. La version portugaise est disponible depuis février 2011. Lors de la dernière mise à jour, la base de données a été enrichie de nouvelles informations sur l'épidémiologie des maladies, sur le mode de transmission et sur les gènes touchés (si cette donnée existe). Et les entrées ont été hiérarchisées selon une classification médicale et scientifique, pour satisfaire tous les utilisateurs.

2.1.6 *Systematized Nomenclature of Medicine – Clinical Terms*

La *Systematized Nomenclature of Medicine – Clinical Terms* (SNOMED CT) [15] est une terminologie clinique détaillée, systématisée et hiérarchisée qui fournit du contenu clinique, afin de rendre expressifs des documents et rapports cliniques (Bodenreider *et al.*, 2007). Elle peut être utilisée pour coder, extraire et analyser des données cliniques. La SNOMED CT (IHTSDO, 2009) est le résultat de la fusion, en 1999, de la *SNOMED – Reference Terminology* (SNOMED RT), développée par le *College of American Pathologists* (CAP) et de la *Clinical Terms Version 3* (CTV3), conçue par le Service National de Santé du Royaume-Uni. Depuis avril 2007, elle est gérée, maintenue et distribuée par l'*International Health Terminology Standards Development Organisation* (IHTSDO), une organisation sans but lucratif établie à Copenhague, au Danemark.
La propre SNOMED *(Systematized Nomenclature of Medicine)* avait débuté en 1965 avec la *Systematized Nomenclature for Pathology* (SNOP), et ensuite elle s'est élargie à d'autres domaines de la Médecine. La SNOMED est une nomenclature de type classification multiaxiale créée pour indexer un ensemble de rapports médicaux. Elle comprend signes et symptômes, diagnostics et procédures ; et son projet unique permet l'intégration complète de toutes les informations médicales dans un dossier médical électronique pour être inséré dans une structure unique de données.

La SNOMED CT est composée de concepts, termes et relations dans l'objectif de représenter de forme précise l'information clinique dans le cadre des soins de santé. Avec des termes pour

14. `http://www.orpha.net` (accédé le 26 mai 2010)
15. `http://www.ihtsdo.org/snomed-ct/` (accédé le 26 mai 2010)

plus de 311.000 concepts uniques accompagnés de définitions basées sur la logique formelle, la SNOMED CT est considérée comme la terminologie clinique disponible la plus complète. Ces concepts sont organisés en 19 niveaux hiérarchiques subdivisés en niveaux multiples de granulosité (du général au spécifique). Les descriptions de concepts (*Concept Descriptions*) sont les termes ou les noms attribués à un concept de la SNOMED CT. Le mot "terme", dans ce contexte, signifie l'expression utilisée pour désigner un concept. Un identifiant de description unique (*unique DescriptionID*) identifie une description. De multiples descriptions peuvent être associées à un concept identifié par un identifiant de concept (*ConceptID*). Il existe près de 800.000 descriptions, incluant les synonymes qui peuvent être utilisés pour faire référence à un concept.

Par exemple [16], le concept "infarctus du myocarde" (*myocardial infarction*) est identifié par un numéro (ConceptID 22298006) auquel sont associés les synonymes : "infarctus cardiaque" (*cardiac infarction*), "crise cardiaque" (*heart attack*) et "infarctus du cœur" (*infarction of heart*) :

- **Nom totalement spécifié** : *Myocardial infarction (disorder)*
  DescriptionID : 751689013
- **Terme préférentiel** : *Myocardial infarction*
  DescriptionID : 37436014
- **Synonyme** : *Cardiac infarction*
  DescriptionID : 37442013
- **Synonyme** : *Heart attack*
  DescriptionID : 37443015
- **Synonyme** : *Infarction of heart*
  DescriptionID : 37441018

Chacune des descriptions mentionnées ci-dessus a un *DescriptionID* unique et elles sont toutes associées à un unique concept (et à un unique *ConceptID* 22298006).

De plus, elle comporte approximativement 1.360.000 liaisons (*links*) ou relations sémantiques entre les concepts de la SNOMED CT, qui permettent la cohérence dans l'extraction et dans l'analyse des données. Ces relations fournissent des définitions formelles et d'autres caractéristiques du concept. Un des types de liaisons est la relation "IS-A". Elle est utilisée pour définir la position du concept à l'intérieur d'une hiérarchie, par exemple :

Diabète sucré (*Diabetes Mellitus*)	IS-A	Dysfonctionnement de la régulation du glucose (*disorder of glucose regulation*)

Les relations "IS-A" peuvent également être nommées relations parents-fils (*parent-child*). Elles sont à la base de la hiérarchie de la SNOMED CT, mais fréquemment nous pouvons trouver des relations morphologiques (*ex., erythema*), topographiques (*ex., left foot*) et étiologiques (*ex., infectious disease*).

Par exemple [17] :

16. Exemple tiré de : SNOMED Clinical Terms®, User Guide, July 2009 International Release, p. 10. (IHTSDO, 2009)
17. Exemple tiré de : SNOMED Clinical Terms®, User Guide, July 2008 International Release, p. 11.

Fracture of tarsal bone (disorder)
- IS_A *Fracture of foot (disorder)*
- FINDING SITE *Bone structure of tarsus (body structure)*
- ASSOCIATED MORPHOLOGY *Fracture (morphologic abnormality)*

La SNOMED CT est une terminologie multilingue et multinationale. Elle comporte un système intégré qui lui permet de gérer différentes langues et dialectes. L'*International Release* dans IHTSDO (2009) inclut un ensemble de concepts et de relations indépendantes de la langue. Aujourd'hui, la SNOMED CT est disponible en anglais américain, en anglais britannique, en espagnol et en danois. Les traductions en langue française, suédoise, lituanienne et diverses autres sont actuellement en cours. Les membres de la IHTSDO planifie aussi de traduire la norme en d'autres langues.
La SNOMED CT travaille également à établir des liaisons explicites (*cross maps*) aux classifications liées à la santé, ainsi qu'à codifier des systèmes utilisés dans le monde entier, comme par exemple, des classifications de diagnostique, telles que la CIM-10 (cf. 2.1.3) ou encore la classification OPCS-4 des interventions (*Office of Population, Censuses and Surveys Classification of Surgical Operations and Procedures (4th revision)* [18]).

2.2 Ressources terminologiques lusophones du domaine médical

La plupart des ressources présentées dans la section précédente, a ses terminologies traduites en portugais, mais principalement pour la norme brésilienne. L'UMLS, un système multilingue, est composé de diverses terminologies traduites en différentes langues, dont le portugais. Cependant, la plus pertinente terminologie médicale portugaise que l'on trouve sur la toile est le DeCS.

2.2.1 Classification Internationale des Soins Primaires

La *Classificação Internacional de Cuidados Primários* (CISP) [Classification Internationale des Soins Primaires] [19] est produite par l'Organisation Mondiale des Collèges Nationaux, Académies et Associés Académiques des Généralistes / Médecins de Famille (WONCA) [20], maintenant connue comme l'Organisation Mondiale des Médecins de Famille. La CISP (Lamberts *et al.*, 1984) est maintenue et actualisée par le Comité International de Classification de Wonca (WICC).
La CISP est utilisée pour la pratique générale/familiale et par les médecins de soins primaires. Elle a été mise à jour en 1993. La CISP a été révisée et publiée dans un livre sous le nom de CISP-2 en 1998 (WONCA International Classification Committee, 1998), et plus tard par voie électronique comme la CISP-2-E en 2000 (Okkes *et al.*, 2000).
La CISP-2 classe les données du patient et l'activité clinique dans les domaines de la médecine générale / de la famille et des soins primaires, en tenant compte de la distribution de

(IHTSDO, 2008)
18. `http://www.connectingforhealth.nhs.uk/systemsandservices/data/clinicalcoding/codingstandards/opcs4`
19. Terminologie présente dans le Metathésaurus de l'UMLS sous le nom de *Portuguese translation of the International Classification of Primary Care (ICPCPOR)*
`http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/ICPC/index.html`
20. WONCA pour *World Organization of National Colleges, Academies and Academic Associates of General Practitioners/Family Physicians*.

fréquence des problèmes rencontrés dans ces domaines. La CISP-2 est un outil épidémiologique utilisé pour classer les données selon trois éléments de la consultation, ou plus généralement de soins primaires : les motifs de rencontre (du point de vue du patient), le diagnostic ou le problème, et les procédures des soins.
Elle a une structure biaxiale et se compose de 17 chapitres (basés sur les systèmes du corps humain avec deux chapitres supplémentaires : un pour les problèmes psychologiques et l'autre pour les problèmes sociaux), chacun divisé en 7 éléments traitant des symptômes et des plaintes (élément 1) ; de diagnostics, de dépistages et de procédures de préventions (élément 2) ; des médicaments, des traitements et des procédures (élément 3), des résultats des tests (élément 4), de l'administration (élément 5), d'aiguillage et d'autres raisons de la rencontre (élément 6) et des maladies (élément 7).

La CISP-2 a été traduite en portugais en 1999 et publiée avec le parrainage de l'*Associação Portuguesa dos Médicos de Clínica Geral* (APMCG). Le livre a été largement distribué aux médecins généralistes travaillant dans le système de la santé publique. Les médecins généralistes, par exemple, utilisent les dossiers traditionnels en papier, alors que les dossiers médicaux électroniques sont rarement utilisés. En conséquence, la CISP est très peu employée dans leur travail quotidien. Toutefois, la CISP-2 est le système de classification le plus fréquemment utilisé dans la recherche, et est également largement diffusé dans les programmes de formation en pratique familiale.

2.2.2 Dictionnaire Médical des Activités Réglementées

Le MedDRA (MedDRA MSSO, 2011), le *Dicionário Médico para Actividades Regulamentares* [Dictionnaire Médical des Activités Réglementées][21] est une terminologie médicale utilisée pour classer les informations liées aux événements indésirables associés à l'utilisation de produits biopharmaceutiques et autres produits médicaux. L'emploi des termes inclus dans le MedDRA pour coder ces informations permet, plus facilement, aux autorités réglementaires et à l'industrie biopharmaceutique d'échanger et d'analyser les informations liées à la sécurité de l'utilisation des produits médicaux (Brown, 2004). Le MedDRA a été développé par la *International Conference on Harmonisation* (ICH) [Conférence Internationale sur l'Harmonisation] et est la propriété de la *International Federation of Pharmaceutical Manufacturers and Associations* (IFPMA) [Fédération Internationale de l'Industrie du Médicament (FIIM)] faisant office de représentant du comité directeur d'ICH.

Le MedDRA est également maintenu en plusieurs langues, parmi lesquelles figurent l'allemand, le tchèque, le chinois, l'espagnol, le français, l'hollandais, le hongrois, l'italien, le japonais, et le portugais.

2.2.3 Terminologie des Effets Indésirables aux Médicaments

La Terminologie des Effets Indésirables aux médicaments de l'OMS (*WHO Adverse Drug Reaction Terminology (WHO-ART)*)[22] est une terminologie hautement raffinée pour le codage de

21. Terminologie présente dans le metathésaurus de l'UMLS sous le nom de *Portuguese translation of the Medical Dictionary for Regulatory Activities (MedDRA)*.
http://www.meddramsso.com/MSSOWeb/index.htm

22. Terminologie présente dans le Metathésaurus de l'UMLS sous le nom de *Portuguese translation of the WHO Adverse Drug Reaction Terminology (WHOPOR)*

l'information clinique liée à des réactions indésirables aux médicaments, utilisée par les pays membres du programme de l'OMS, et dans le monde, par les compagnies pharmaceutiques et les organisations de recherche clinique (Uppsala Monitoring Centre, 2011).
La WHO-ART a été développée pendant plus de trente ans et est maintenue par le *Uppsala Monitoring Centre* (Centre de surveillance d'Uppsala), le Centre de Collaboration de l'OMS pour la surveillance internationale des médicaments (WHO Collaborating Centre for International Drug Monitoring, 2005), pour servir de base rationnelle pour le codage des termes d'effets indésirables.
La structure de la WHO-ART se compose de quatre niveaux hiérarchiques pour désigner les termes (Alj *et al.*, 2005). Parce que les nouveaux médicaments et les nouvelles indications produisent de nouveaux termes d'effets indésirables, la structure de la terminologie est assez souple pour permettre d'incorporer de nouvelles entrées, tout en maintenant sa structure et sans perdre les relations antérieures.
Le premier niveau est : «Catégories de systèmes ou organes» où sont regroupés les termes préconisés pour les effets indésirables (EI) se rapportant à un même système ou organe (Exemple : Troubles de la fréquence et du rythme cardiaque).
Le deuxième niveau est : «Termes plus généraux» où sont les termes préconisés regroupant des affections qualitativement analogues mais quantitativement différentes (Exemple : Tachycardie).
Le troisième niveau est : «Termes préconisés» où sont regroupés les termes utilisés pour caractériser les EI notifiés au système OMS. Ce sont les termes les plus souvent employés à l'entrée des données (Exemple : Tachycardie, palpitations).
Le quatrième niveau est : «Termes d'inclusions» où sont les synonymes des termes préconisés, indiqués par les pays notificateurs. Ils ont pour but d'aider à trouver le terme préconisé correspondant, afin de donner à l'EI notifié, le bon terme de code.
La WHO-ART a été développée en anglais, mais est également traduite pour l'allemand, l'espagnol, le français, l'italien et le portugais.

La WHO-ART couvre la plupart des termes médicaux nécessaires dans la déclaration des réactions indésirables, mais elle est encore assez petite pour permettre de l'imprimer sous forme de listes ce qui la rend facilement utilisable pour les petites entreprises et les centres nationaux.

2.2.4 Medical Wordnet

Comme nous avons pu déjà le constater, la langue biomédicale dispose principalement de sources anglaises. Wordnet (Fellbaum, 1998), une grande base de données électronique lexicales de la langue anglaise, a été initialement conçue comme un modèle à grande échelle de l'organisation sémantique humaine, où les termes et leurs significations sont liés entre eux par des relations sémantiques et lexicales. En résumé, il propose des synonymes pour l'anglais général, mais les ressources correspondantes pour d'autres langues ne sont pas toujours disponibles en libre accès.
Pour le portugais, il existe le WordNet Portugais : WordNet.PT qui se construit depuis 1999 (Marrafa, 2001) au Centre de Linguistique de l'Université de Lisbonne. WordNet.PT a une architecture conceptuelle structurée sur le modèle original de WordNet. Comme les ressources linguistiques disponibles pour le portugais ne conviennent pas suffisamment pour le but de

`http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/WHO/index.html`

construire un WordNet de forme automatique, ce projet est réalisé principalement sur la base du travail manuel. La version actuelle de WordNet.PT contient environ 19.000 expressions lexicales, provenant de différents champs sémantiques. La version disponible en ligne [23] inclut des expressions de différents domaines dont celui de la santé.

Étant donné que WordNet était insuffisant pour le traitement de la langue biomédicale, une nouvelle variante de WordNet a été créée mais cette fois-ci orientée uniquement vers le domaine de la santé (Fellbaum *et al.*, 2006; Smith & Fellbaum, 2004) : le Medical WordNet. Cependant, cette adaptation semble avoir échouée, et n'a pas de traduction pour le portugais.

2.2.5 Diverses terminologies médicales

Au niveau national, différents travaux ont été entrepris depuis plusieurs années, pour tenter de résoudre l'ambiguïté terminologique du domaine de la médecine.

Maria de Lurdes Abrantes Garcia met l'accent sur les problèmes d'instabilité terminologique (Garcia, 1994) dûs aux décalques et équivalents que les chercheurs portugais utilisent à partir de la bibliographie française et anglo-américaine. C'est pourquoi, lors de sa recherche sur la terminologie de la senologie (maladies du sein), elle prévoit de créer un dictionnaire intéractif multilingue (Garcia, 1997) dans ce sous-domaine spécifique de la médecine.
Contente & Magalhães (1997) ont créé un dictionnaire multilingue de médecine en s'appuyant sur les synonymes et les équivalents des termes (Contente, 2004) à partir d'un corpus formé d'ouvrages spécialisés et de dictionnaires récents dans chacune des langues utilisées.

2.2.6 *Descritores em Ciências da Saúde*

Pour la langue portugaise, il existe le *Descritores em Ciências da Saúde* (DeCS) (pour Descripteurs en Sciences de la Santé), un vocabulaire structuré et trilingue (Tardelli, 2007), basé sur des collections de termes organisés pour faciliter l'accès à l'information. Il a été créé, en 1982, par la BIREME [24] pour servir comme un langage unique pour l'indexation d'articles de revues scientifiques, livres, rapports techniques et autres types de matériaux, ainsi que pour être utilisé dans la recherche et la récupération d'information dans des sources d'informations disponibles, telle que MEDLINE.
Il a été développé à partir du MeSH [25], dans l'objectif de permettre l'utilisation d'une terminologie commune pour la recherche en trois langues différentes, proportionnant un moyen consistant et unique pour la récupération d'informations indépendamment de la langue. Ce vocabulaire actuellement disponible en anglais, espagnol et portugais variante brésilienne, est donc totalement compatible avec le MeSH, en raison de son entière traduction manuelle, et est statique, à cause de l'absence d'actualisations systématiques du propre lexique. Il participe au projet de développement de la terminologie unique et du réseau sémantique de la santé, l'UMLS, avec la responsabilité de l'actualisation et de la transmission des termes en portugais et en espagnol.

23. `http://www.clul.ul.pt/wn/`
24. Centre Latino-Américain et des Caraïbes de l'Information en Sciences de la Santé (`http://www.bireme.br`).
25. C'est pourquoi il est introduit dans le Metathésaurus de l'UMLS sous le nom de MSHPOR (*Portuguese translation of the MeSH*)

2.2.6.1 Sa constitution

Outre les domaines qui composent le MeSH, le DeCS comporte (cf. figure 2.4) une terminologie de Santé Publique (*Saúde Pública*), d'Homéopathie (*Homeopatia*), de Science et Santé (*Ciência e Saúde*), ainsi que d'une Vigilance Sanitaire (*Vigilância Sanitária*). Toutefois, le DeCS couvre des domaines non-spécifiques à l'univers de la Médecine, tels que Phénomènes Sociaux ou encore Économie, laissant une zone d'intervention au niveau scientifique : aussi bien au niveau de la dynamicité du lexique, qu'au niveau de l'unification du vocabulaire pour le portugais dans sa norme luso-africaine.

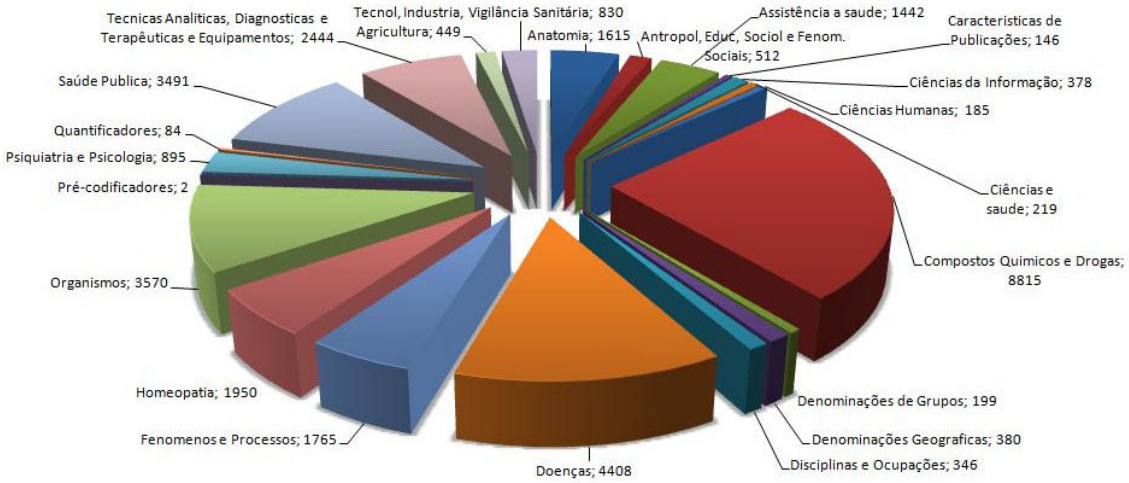


Figure 2.4: Catégories présentes dans le DeCS dans sa version de 2010, tiré de Costa (2010)

Les concepts qui composent le DeCS sont organisés dans une hiérarchie de relations d'hyponymie-hyperonymie entre les termes, permettant donc la recherche de termes plus généraux ou plus spécifiques, ou encore tous les termes appartenant à une même structure hiérarchique (ou branche de l'arbre). Le DeCS est un vocabulaire dynamique totalisant, en 2010, 30.369 descripteurs, sachant que 25.671 viennent du MeSH et 4698 exclusivement du DeCS.

Ce système nous offre plusieurs informations très utiles (cf. figure 2.5), telles que les définitions, les traductions en anglais et en espagnol, les synonymes pour le portugais, les termes médicaux relatifs au terme d'origine ainsi que la position dans la structure hiérarchique ; car tout comme le MeSH, un descripteur peut apparaître plus d'une fois dans la hiérarchie, avec un code hiérarchique différent.
Donc pour chaque entrée du dictionnaire, nous avons une information du ou des chemins possibles pour cette entrée. Par exemple, pour arriver au terme *coração fetal*, il existe deux chemins possibles. Le premier code hiérarchique (A07.541.278) de la figure 2.5 fait référence à l'arbre de la première colonne du tableau 2.1 et le second (A16.378.303) à l'arbre de la deuxième colonne.

2.2.6.2 Ses principaux problèmes

Si nous le comparons avec les deux grands lexiques, celui de l'UMLS (le lexique SPECIALIST) et l'UMLF, qui ont tous deux, un ample lexique créé à partir de la fusion de différentes terminologies médicales existantes dans les différentes variantes de la langue anglaise et française


Pesquisa sobre:CORACAO FETAL
Descritores Encontrados:1
Mostrando:1 .. 1

DeCS

Descritor *Inglês*: **Fetal Heart**

Descritor *Espanhol*: **Corazón Fetal**

Descritor *Português*: **Coração Fetal**

Categoria: A07.541.278
A16.378.303

Definição *Português*: Coração existente no feto de qualquer animal vivíparo. Refere-se ao coração do período pós embrionário e é diferenciado do coração embrionário (CORAÇÃO/embriologia) somente por uma questão temporal.

Figure 2.5: Exemple d'une entrée du DeCS avec deux codes hiérarchiques

Table 2.1: Hiérarchie des deux chemins possibles pour arriver au terme *coração fetal*

ANATOMIA	*ANATOMIA*
Sistema Cardiovascular	*Estruturas Embrionárias*
Coração	*Feto*
Endocárdio	*Feto Abortado*
Coração Fetal	*Líquido Amniótico*
Canal Arterial	*Sangue Fetal*
Tronco Arterial	***Coração Fetal***
	Canal Arterial
	Coxins Endocárdicos
	Tronco Arterial

respectivement ; le DeCS, quant à lui, est une ressource de taille inférieure, et de plus, il a été construit uniquement pour la norme brésilienne du portugais, et non pour la norme luso-africaine, ou pour les deux. Par exemple, le terme : *quisto* existe seulement dans sa variante brésilienne, i.e. *cisto*, qui ne s'emploie pas dans le portugais luso-africain. De plus, l'actualisation que le MeSH réalise annuellement, ainsi que les changements dans les catégories spécifiques au DeCS, exigent une révision et une correction manuelle, et donc coûteuse, de toute la base de données.

2.3 Synthèse sur les ressources terminologiques du domaine médical

Dans cette section, nous avons présenté les principales ressources terminologiques du domaine de la médecine accessibles sur la toile. À la fin de ce panorama, nous pouvons constater que la majeure partie de ces ressources ont été conçues pour la langue anglaise. Des traductions ont effectivement été effectuées pour d'autres langues, dont le portugais, mais principalement dans sa variante brésilienne. De plus, tous ces systèmes sont construits et maintenus manuellement.

En langue portugaise, les corpora et thésauri médicaux sont quasiment inexistants, et la présence de variantes terminologiques au niveau national (selon les différentes écoles de méde-

cine) et international (portugais du Brésil *vs* portugais luso-africain) ne nous facilite pas la tâche. Ces différentes contraintes nous ont donc mené à proposer la création d'un dictionnaire médical unifié en langue portugaise, construit et maintenu de forme semi-automatique. Nous allons maintenant nous intéresser à l'élaboration de la première étape de ce travail, c'est-à-dire la création de la base de données terminologique.

Chapitre 3

Création de la base de données

Dans la langue portugaise, la difficulté de trouver un dictionnaire de médecine en format électronique, ainsi que des corpora médicaux, est très grande. Il existe plusieurs dictionnaires médicaux en support papier (Manuila *et al.*, 2000; Taber, 2000), mais ces outils rendent le traitement automatique de l'information très lent et coûteux. En ce qui concerne les corpora, il n'existe aucun corpus de documents médicaux ou de rapports cliniques accessibles sur la toile. Une des raisons de cette pénurie est la confidentialité des documents. Et même la littérature scientifique produite par les médecins portugais, est la plupart du temps en langue anglaise. Donc, rechercher des documents médicaux en format électronique pour créer un corpus, est une tâche difficile dans la langue portugaise.

C'est la raison pour laquelle nous avons essayé de créer un corpus à partir d'environ 250 sites homologués (i.e. que l'on peut considérer fiables) dans le domaine .pt, tels que les sites de la Société Portugaise de Gynécologie [1], de la Société Portugaise d'Oncologie [2] ou encore de l'Ordre des Médecins [3]. Les principaux problèmes auxquels nous nous sommes confrontés ont été, dans certains cas, le faible contenu pertinent des sites, et d'autres parts, les contenus en langue étrangère. Nous avons alors décidé de ne pas utiliser ce type de ressources pour notre recherche et de nous orienter vers des sources déjà existantes en ligne, telles que dictionnaires ou encyclopédies libres d'accès. Cette situation proportionne un grand vide pour la recherche dans le domaine du Traitement Automatique du Langage (TAL) médical, mais il ouvre parallèlement différents horizons d'études pour les linguistes.

Pour créer notre base de données biomédicale, nous n'avons utilisé que des sources électroniques disponibles gratuitement sur la toile. Sept différentes sources, comprenant dictionnaires et glossaires de la langue portugaise, encyclopédie et dictionnaire libres ainsi qu'un vocabulaire dynamique structuré, nous ont permis de créer une base de données terminologiques large et enrichie morphosyntaxiquement parlant. À cette fin, nous avons développé différents robots d'indexation qui extraient les termes de quatre de ces bases de données terminologiques. En parallèle, nous avons prélevé, manuellement, les termes des trois autres ressources en ligne, en raison de leur faible quantité de contenu et de la facilité d'extraction manuelle de leurs terminologies.

1. http://www.spginecologia.pt/
2. http://www.sponcologia.pt/
3. http://www.ordemdosmedicos.pt/

3.1 Extraction Manuelle

Pour créer notre base de données biomédicale, nous avons employé trois dictionnaires/glossaires électroniques disponibles sur la toile : le dictionnaire *Priberam*, un Glossaire Multilingue ainsi que le glossaire du site du Centre Hospitalier *Cova da Beira*.

3.1.1 *Dicionário Priberam da Língua Portuguesa*

Nous employons le *Dicionário Priberam da Língua Portuguesa* (DPLP)[4] en raison d'une fonctionnalité très intéressante qu'il offre à l'utilisateur. Cette ressource nous permet de rechercher tous les termes qui ont l'information du domaine technique auquel ils font référence. Par exemple, nous pouvons récolter toutes les entrées du dictionnaire qui ont la marque du domaine de la Médecine en utilisant le système de recherche avancée avec "*Med.*", comme nous pouvons le voir dans la figure 3.1 :

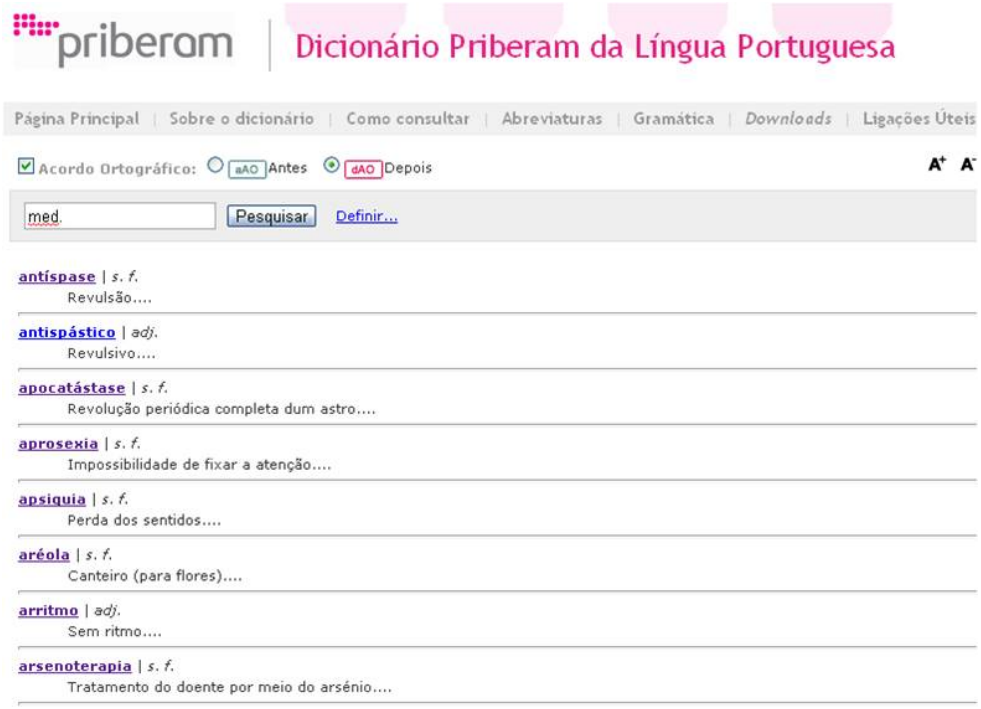


Figure 3.1: Méthode de recherche des termes connotés avec l'information du domaine de la Médecine

Après avoir obtenu la liste des entrées du dictionnaire marquées par le domaine de la Médecine, nous gardons tous ces termes accompagnés de toutes leurs informations complémentaires. Cependant certains termes peuvent comprendre plus d'une définition par entrée. Dans ces cas là, nous ne gardons que les définitions avec la marque du domaine en question. Ainsi, dans l'exemple[5] du tableau 3.1, nous ne gardons que la seconde définition :

Nous adoptons également cette procédure pour tous les sous-domaines de la Médecine, ainsi que ses domaines connexes, qui se trouvent dans ce dictionnaire, en utilisant la même forme de

4. `http://priberam.com/dlpo/dlpo.aspx`
5. Exemple tiré du site : Dicionário Priberam da Língua Portuguesa [en ligne], 2010, `http://www.priberam.pt/dlpo/dlpo.aspx?pal=cordite` [consulté le 03-09-2010].

Table 3.1: Exemple d'une entrée où seulement une des définitions présentées est à garder

cordite
s. f.
1. Uma das muitas variedades de pólvora sem fumo.
2. *Med.* Inflamação das cordas vocais.

recherche avancée : Anatomie ("*Anat.*"), Bactériologie ("*Bacteriol.*"), Biologie ("*Biol.*"), Biochimie ("*Bioquím.*"), Chimie ("*Quím.*"), Chirurgie ("*Cirurg.*"), Cytologie ("*Citol.*"), Embryologie ("*Embriol.*"), Génétique ("*Genét.*"), Histologie ("*Histol.*"), Médecine Vétérinaire ("*Veter.*"), Optique ("*Ópt.*"), Pathologie ("*Patol.*"), Pharmacie ("*Farm.*"), Physiologie ("*Fisiol.*"), Psychanalyse ("*Psican.*"), Psychiatrie ("*Psiquiatr.*"), Psychologie ("*Psicol.*") et Radiologie ("*Radiol.*").

Dans la figure 3.2, nous montrons une entrée telle qu'elle apparaît sur le site. Dans cet exemple, nous avons le terme de l'entrée (*olecrano*), suivi de l'étymologie (*grego olékranon, -ou*), de la catégorie grammaticale et de son genre avec les abréviations *s. m.* pour substantif masculin, du domaine (*Anat.*), de la définition et d'un synonyme (*olecrânio*) introduit par le symbole "=". Toutes les entrées étant organisées de la même façon, l'extraction manuelle de cette terminologie a été simple et assez rapide. De plus, le fait qu'il n'existe pas d'index dans cette ressource, aurait rendu l'extraction automatique plus compliquée.

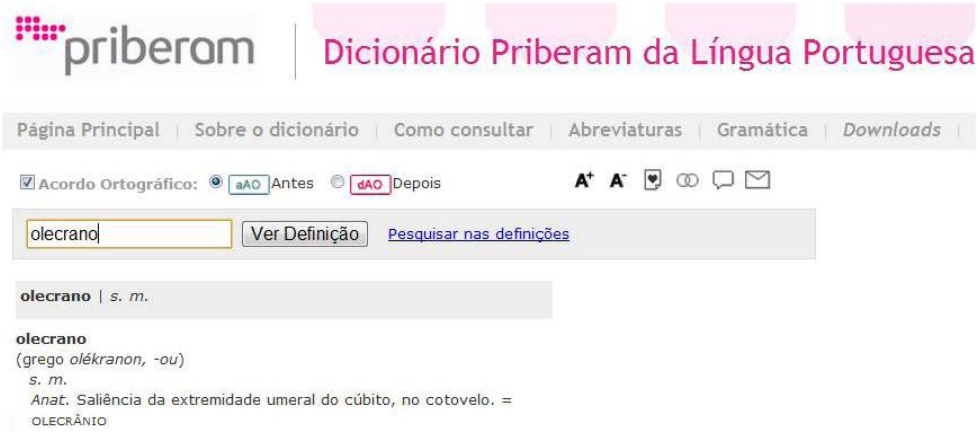


Figure 3.2: Exemple d'une entrée du DPLP

Grâce à cette ressource, nous avons pu obtenir toute la terminologie biomédicale présente dans ce dictionnaire, qui regroupe au total 3.615 entrées avec : définition, catégorie grammaticale, genre, nombre et, quand disponible, l'étymologie du terme, le(s) synonyme(s), antonyme(s), symbole(s), abréviation(s) et adjectif(s) ou autres termes relatifs à l'entrée en question.

3.1.2 Glossaire Multilingue de Termes Médicaux

Sur la toile, nous trouvons également le Glossaire Multilingue de Termes Médicaux techniques et populaires en huit langues européennes (GMTM)[6]. Ce glossaire contient 1.925 termes en portugais pour lesquels nous avons gardé leurs informations morphologiques ainsi que leurs respectives traductions en anglais, espagnol, et français, et quand disponibles les respectives informations complémentaires de chaque traduction (définition, catégorie grammaticale, genre,

6. http://users.ugent.be/~rvdstich/eugloss/welcome.html

etc.).

Le contenu du glossaire peut être présenté de deux formes : sous la forme de glossaire unilingue (cf. figure 3.3), où les entrées sont présentées par ordre alphabétique pour chaque langue individuellement ; ou bien sous la forme de glossaire multilingue (cf. figure 3.4)[7], où les entrées sont présentées en même temps pour toutes les langues, prenant comme terme de référence, le terme anglais.

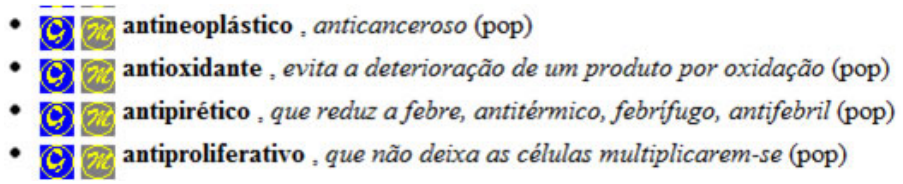


Figure 3.3: Exemple d'une entrée dans la présentation unilingue du GMTM

En comparant ces deux formes de présentation des données, nous voyons que dans la première figure, le terme en gras fait référence au terme technique et le reste au terme familier.

Pour l'entrée *antipirético*, nous obtenons sa définition (*«que reduz a febre»*) ainsi que trois synonymes familiers (*antitérmico, febrífugo, antifebril*). En effet, nous constatons que, *«que reduz a febre»*, n'est pas un synonyme populaire de *antipirético*, mais sa définition.

Avec cette ressource, nous obtenons donc le terme de l'entrée (terme technique), sa définition et/ou son/ses synonyme(s) populaire(s), et parfois également son domaine, sa catégorie grammaticale, son nombre, son genre et son abréviation.
Ces environs 2.000 extractions ont été effectuées manuellement, tout comme pour le DPLP, en raison du faible contenu, mais principalement à cause de la mauvaise organisation de cette ressource, car nous avons dû lire chaque entrée de ce glossaire une par une afin de pouvoir définir quelle était la définition et quels étaient les synonymes populaires. Un système automatique ne serait pas capable de le faire. Cependant, le GMTM est très pertinent, car il nous permet de connaître le terme technique, mais aussi son/ses équivalent(s) populaire(s).

Effectivement, un professionnel de la santé doit connaître le terme technique exact, tout comme il doit également être en mesure d'accéder à sa forme populaire dans la communication avec ses patients. Il convient également de noter que le GMTM fournit parfois le domaine d'un terme. Ainsi, comme pour le DPLP, nous obtenons des informations essentielles pour la future création d'un Metathésaurus.

3.1.3 Glossaire du site du Centre Hospitalier *Cova da Beira*

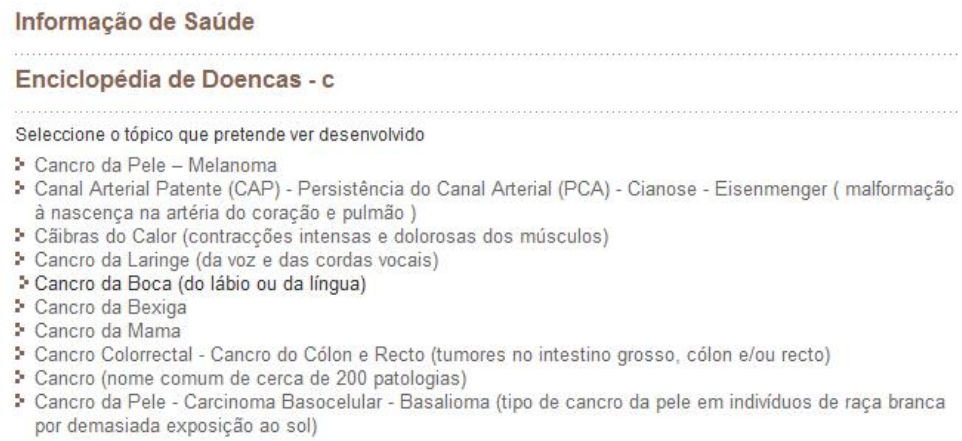
Étant donné que notre travail est principalement destiné à la Faculté des Sciences de la Santé et à l'Hôpital de la *Cova da Beira*, nous avons alors décidé d'utiliser également le glossaire[8]

7. http://users.ugent.be/~rvdstich/eugloss/multi016.html#0158
8. http://www.chcbeira.min-saude.pt/InfoUtente/InfoSaude/?sm\unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox\@tempboxa\hbox{1\global\mathchardef\accent@spacefactor\spacefactor}\accent91\egroup\spacefactor\accent@spacefactor_2

Figure 3.4: Exemple d'une entrée dans la présentation multilingue du GMTM

présent sur le site du Centre Hospitalier *Cova da Beira* (CHCB), qui contient 195 entrées spécifiques organisées par ordre alphabétique (cf. figure 3.5).

Ces entrées sont toujours accompagnées de leurs définitions très développées sur le sujet en question (cf. figure 3.6), ainsi que de leurs synonymes, abréviations et termes relatifs, quand ils sont disponibles.



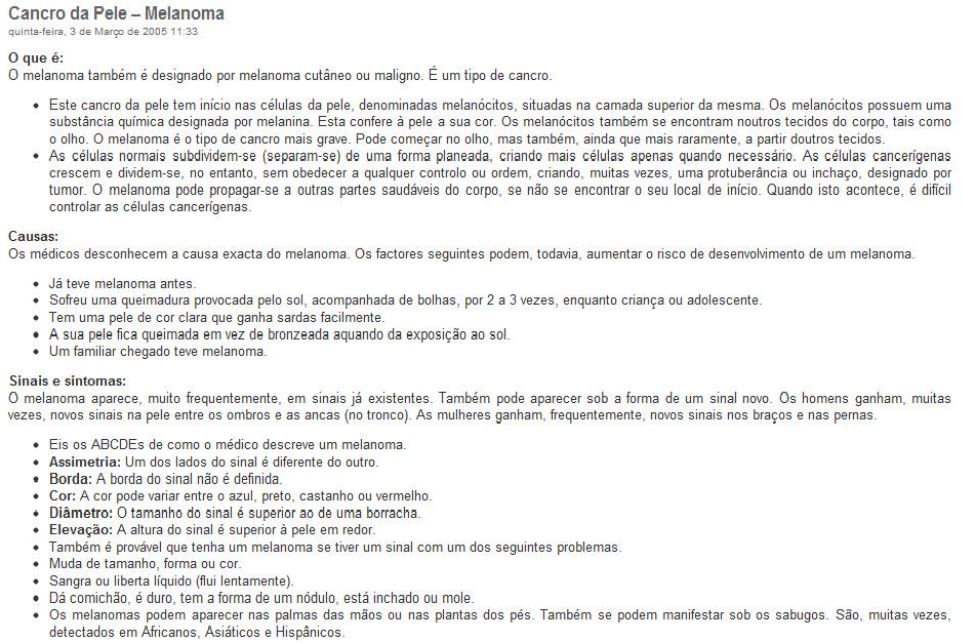
Informação de Saúde

Enciclopédia de Doencas - c

Seleccione o tópico que pretende ver desenvolvido

- Cancro da Pele – Melanoma
- Canal Arterial Patente (CAP) - Persistência do Canal Arterial (PCA) - Cianose - Eisenmenger (malformação à nascença na artéria do coração e pulmão)
- Cãibras do Calor (contracções intensas e dolorosas dos músculos)
- Cancro da Laringe (da voz e das cordas vocais)
- Cancro da Boca (do lábio ou da língua)
- Cancro da Bexiga
- Cancro da Mama
- Cancro Colorrectal - Cancro do Cólon e Recto (tumores no intestino grosso, cólon e/ou recto)
- Cancro (nome comum de cerca de 200 patologias)
- Cancro da Pele - Carcinoma Basocelular - Basalioma (tipo de cancro da pele em indivíduos de raça branca por demasiada exposição ao sol)

Figure 3.5: Exemple d'une entrée dans la présentation alphabétique du glossaire du CHCB



Cancro da Pele – Melanoma

quinta-feira, 3 de Março de 2005 11:33

O que é:
O melanoma também é designado por melanoma cutâneo ou maligno. É um tipo de cancro.

- Este cancro da pele tem início nas células da pele, denominadas melanócitos, situadas na camada superior da mesma. Os melanócitos possuem uma substância química designada por melanina. Esta confere à pele a sua cor. Os melanócitos também se encontram noutros tecidos do corpo, tais como o olho. O melanoma é o tipo de cancro mais grave. Pode começar no olho, mas também, ainda que mais raramente, a partir doutros tecidos.
- As células normais subdividem-se (separam-se) de uma forma planeada, criando mais células apenas quando necessário. As células cancerígenas crescem e dividem-se, no entanto, sem obedecer a qualquer controlo ou ordem, criando, muitas vezes, uma protuberância ou inchaço, designado por tumor. O melanoma pode propagar-se a outras partes saudáveis do corpo, se não se encontrar o seu local de início. Quando isto acontece, é difícil controlar as células cancerígenas.

Causas:
Os médicos desconhecem a causa exacta do melanoma. Os factores seguintes podem, todavia, aumentar o risco de desenvolvimento de um melanoma.

- Já teve melanoma antes.
- Sofreu uma queimadura provocada pelo sol, acompanhada de bolhas, por 2 a 3 vezes, enquanto criança ou adolescente.
- Tem uma pele de cor clara que ganha sardas facilmente.
- A sua pele fica queimada em vez de bronzeada aquando da exposição ao sol.
- Um familiar chegado teve melanoma.

Sinais e sintomas:
O melanoma aparece, muito frequentemente, em sinais já existentes. Também pode aparecer sob a forma de um sinal novo. Os homens ganham, muitas vezes, novos sinais na pele entre os ombros e as ancas (no tronco). As mulheres ganham, frequentemente, novos sinais nos braços e nas pernas.

- Eis os ABCDEs de como o médico descreve um melanoma.
- **Assimetria:** Um dos lados do sinal é diferente do outro.
- **Borda:** A borda do sinal não é definida.
- **Cor:** A cor pode variar entre o azul, preto, castanho ou vermelho.
- **Diâmetro:** O tamanho do sinal é superior ao de uma borracha.
- **Elevação:** A altura do sinal é superior à pele em redor.
- Também é provável que tenha um melanoma se tiver um sinal com um dos seguintes problemas.
- Muda de tamanho, forma ou cor.
- Sangra ou liberta líquido (flui lentamente).
- Dá comichão, é duro, tem a forma de um nódulo, está inchado ou mole.
- Os melanomas podem aparecer nas palmas das mãos ou nas plantas dos pés. Também se podem manifestar sob os sabugos. São, muitas vezes, detectados em Africanos, Asiáticos e Hispânicos.

Figure 3.6: Exemple d'une entrée du glossaire du CHCB

3.2 Extraction Automatique

Après avoir extrait manuellement les données des trois ressources précédentes, nous présentons à présent sous quelle forme nous avons extrait de manière automatique les terminologies des ressources suivantes : le glossaire du site *Médicos de Portugal*, la Wikipédia, le *Wikcionário* et le DeCS. Pour cela, nous avons créé pour chacune des ressources, un robot d'indexation spécifique, c'est-à-dire unique.

Un robot d'indexation (Kobayashi & Takeda, 2000; Wikipédia, 2010) (en anglais *web crawler* ou *web spider*) est un programme informatique qui explore automatiquement sur la toile. Il est généralement conçu pour collecter les ressources (pages *web*, images, vidéos, documents Word, PDF ou PostScript, etc.), afin de les indexer.

3.2.1 Glossaire du site *Médicos de Portugal*

Nous avons extrait les données du glossaire médical rassemblées sur le site *Médicos de Portugal*[9] (MP). En particulier, nous avons vérifié que la majeure partie des termes provienne de la transcription du *Dicionário Médico* de Manuila *et al.* (2000), qui est une référence fondamentale pour les médecins et étudiants en médecine ; mais il est également une aide précieuse pour tous les professionnels de la santé.
Ce glossaire contient 12.828 entrées (terme à définir avec sa définition et ses respectives informations) accompagnées d'informations diverses, telles que : catégorie grammaticale, genre, nombre, définition, synonyme, antonyme, traductions en français, anglais britannique et parfois anglais américain, termes en relation avec le terme défini, etc.

Dans la figure 3.7, nous montrons une entrée simple du glossaire accompagnée de nombreuses informations.

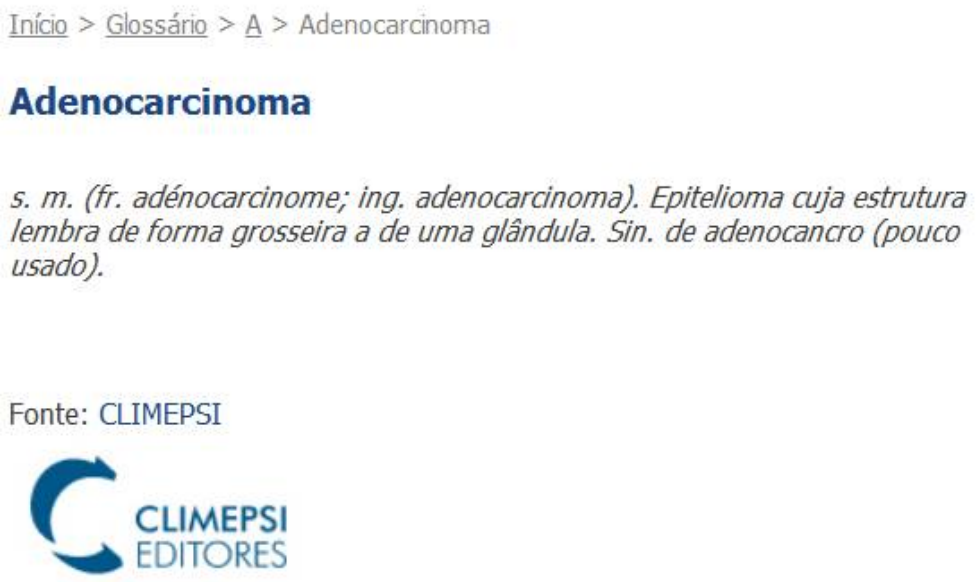


Figure 3.7: Exemple d'une entrée du MP

Dans cet exemple, nous avons le terme de l'entrée (*adenocarcinoma*), la catégorie grammaticale et le genre avec les abréviations *s. m.* pour substantif masculin, la traduction en français et en anglais (*fr. adénocarcinome ; ing. adenocarcinoma*), la définition (*Epitelioma cuja estrutura lembra de forma grosseira a de uma glândula.*), un synonyme (*adenocancro*) introduit par l'abréviation *sin.*, l'information de l'emploi de ce synonyme (*pouco usado*) mis entre parenthèses et finalement l'indication de la source d'où proviennent ces informations.

Bien que l'extraction des données ait été faite de forme automatique à partir de la création d'un robot d'indexation spécifique, il a été nécessaire de post-éditer manuellement chaque entrée de ce glossaire une à une, en raison de nombreuses erreurs orthographiques, aussi bien dans les entrées que dans le contenu des définitions. Erreurs dues certainement à une mauvaise traduction du dictionnaire d'origine en langue française.

9. `http://medicosdeportugal.saude.sapo.pt/`

3.2.2 Wikipédia

Wikipédia est une encyclopédie multilingue, universelle, librement diffusable, disponible sur la toile et écrite par des internautes volontaires de diverses régions du monde. Ce site utilise un outil propre, la technologie wiki, qui permet à toute personne d'améliorer immédiatement quelconque article. C'est cette spécificité qui rend cette ressource intéressante, en raison de son actualisation permanente.

Nous utilisons les ressources de Wikipédia [10] en version portugaise, où les recherches ont été effectuées par un robot d'indexation spécifique, capable de restreindre sa recherche au seul domaine de la Médecine, en partant de la catégorie «*Medicina*» [11] (cf. figure 3.8), et pour parcourir toutes les sous-catégories et articles du domaine de la Médecine.

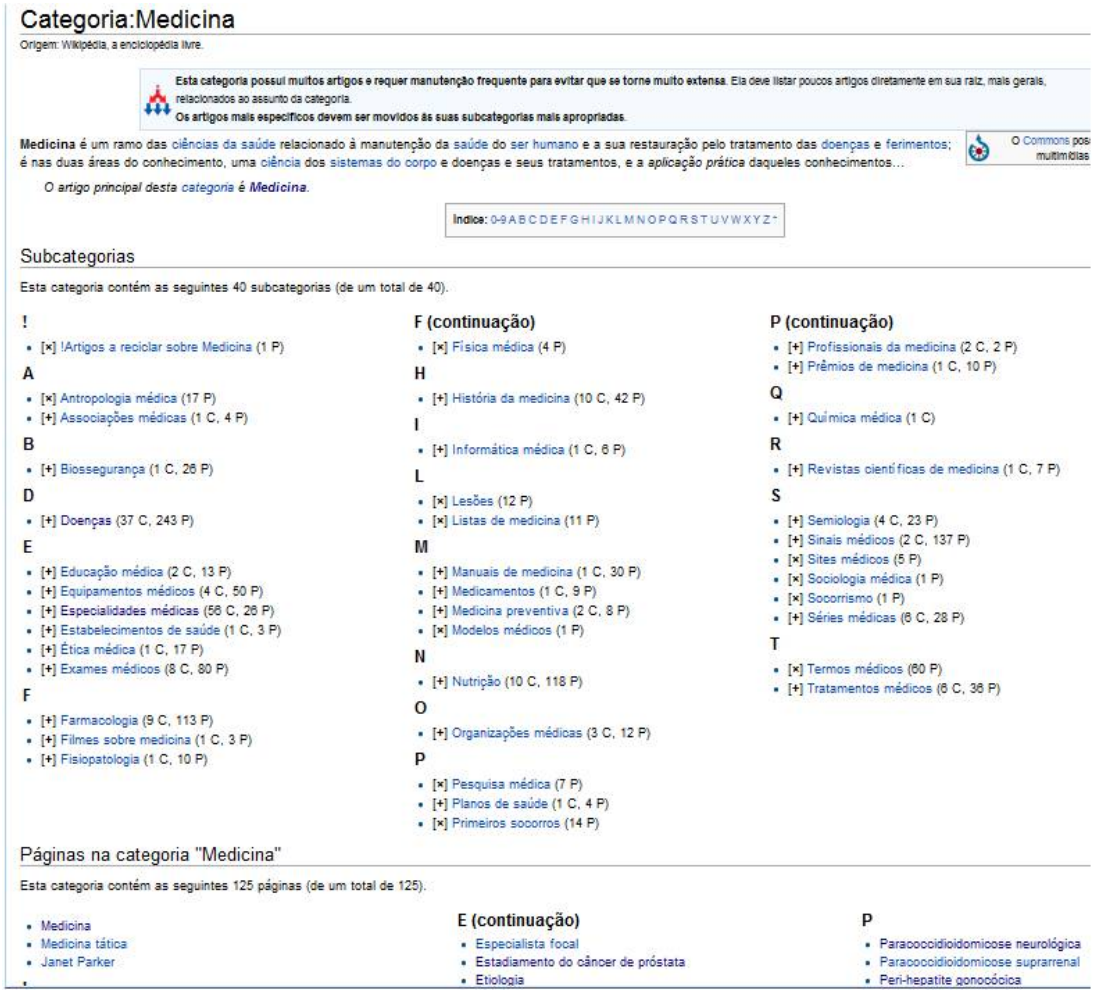


Figure 3.8: Page de la catégorie «*Medicina*» de Wikipédia

De fait, Wikipédia est un graphe orienté dans lequel il existe des cycles. En raison de cette structure, une recherche brute nous conduit "inexorablement" à l'extraction de termes non-médicaux. Pour ce projet a été développé dans Costa (2010), un algorithme de recherche heuristique capable de limiter la recherche de termes à l'intérieur du seul domaine de la Médecine. Grâce à cette fonctionnalité, nous avons obtenu de nombreux termes médicaux, avec leurs respectives définitions.

10. `http://pt.wikipedia.org/wiki/Wikipédia:Página_principal`
11. `http://pt.wikipedia.org/wiki/Categoria:Medicina`

Le fait que Wikipédia soit structurée en catégories, sous-catégories et articles nous permet de créer automatiquement une structure hiérarchique. Par exemple, pour arriver au terme *adenocarcinoma* (adénocarcinome), Wikipédia offre plus d'une possibilité de chemins (cf. figure 3.9).

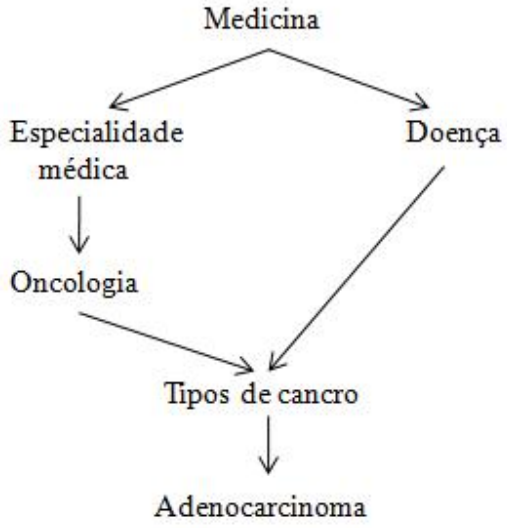


Figure 3.9: Schéma représentant différentes possibilités de chemins pour aboutir à un terme

Afin de construire la structure hiérarchique médicale de Wikipédia, tous les chemins du graphe sont stockés pour chaque terme trouvé. De toute évidence, l'unification de ces voies devra être faite pour atteindre le thésaurus médical de Wikipédia et entrevoir la définition future d'un Metathésaurus. Mais comme nous le verrons dans le chapitre 7, cette tâche est plus complexe qu'initialement prévue.
Les outils de Wikipédia nous ont donc permis de recueillir 7.367 entrées, et nous ont également donné la possibilité d'avoir les définitions ainsi que les traductions des termes médicaux trouvés. Nous gardons ces traductions pour l'anglais, l'espagnol et le français, quand elles sont mentionnées. En effet, les traductions peuvent avoir une influence non négligeable sur l'unification de la terminologie. Par exemple, deux termes avec des variantes différentes en portugais, mais avec la même traduction pour une ou plusieurs langues, augmente notre certitude sur le fait que ces termes puissent être unifiés. Parfois, nous avons également des images ou photographies qui aident à mieux identifier le terme défini. Ces images sont une information importante que nous gardons afin de créer un dictionnaire unifié, évolutif et multimédia.

3.2.3 *Wikcionário*

Le *Wikcionário* [12] (appellation portugaise) est la partie portugaise du projet multilingue Wiktionary® de la Fondation Wikimedia. Ce projet de collaboration mené par un ensemble d'internautes, a vu ses débuts en Mai 2004 et vise à produire un dictionnaire polyglotte libre en portugais, qui contient maintenant 166.012 entrées [13] dans sa version en portugais. Comme dans Wikipédia, chaque internaute peut contribuer, par exemple, en modifiant, corrigeant ou enregistrant une définition.

12. `http://pt.wiktionary.org/wiki/Wikcionário:Página_principal`
13. Informations retirées du site le 24 février 2011.

Initialement créé en tant que complément lexical de Wikipédia, il s'est développé au-delà du simple dictionnaire, et a maintenant pour objectif de décrire tous les mots de toutes les langues, dont le portugais, langue parlée dans plusieurs pays et par diverses communautés dans le monde entier. C'est-à-dire qu'il permet de donner, pour tous les mots lusophones et étrangers, une définition en langue portugaise, ainsi qu'une traduction, mais aussi de donner leur étymologie, synonyme, antonyme, etc.
Cette ressource nous précise fréquemment si le terme d'entrée s'applique au portugais de variante luso-africaine ou au portugais de variante brésilienne, ce qui nous permet de compléter notre dictionnaire avec cette importante information. Tout comme Wikipédia, cet outil est également structuré en catégories, sous-catégories et articles, ce qui nous donne la possibilité d'extraire automatiquement son thésaurus médical et de créer une structure hiérarchique. C'est-à-dire, un robot d'indexation spécifique parcourt le graphe orienté du *Wikcionário* à partir de la catégorie «*Medicina*» [14] (voir la figure 3.10) et extrait tous les termes du domaine médical qui se trouvent annotés par cette catégorie, avec ses définitions, étymologie, synonymes, antonymes et traductions (quand disponible). Au total, nous avons recueilli 2.050 entrées en lien avec la Médecine.


Categoria:Medicina (Português)

Esta categoria contém verbetes relacionados ao tema **medicina** no idioma português.

(200 anteriores) (próximos 200)

Subcategorias

Esta categoria contém as seguintes 3 subcategorias (de um total de 3).

A
- [×] Anatomia (Português) (342 P)

O
- [×] Odontologia (Português) (8 P)

P
- [×] Patologia (Português) (235 P)

Artigos na categoria "Medicina (Português)"

Esta categoria contém as seguintes 197 páginas (de um total de 489).

-
- -ectomia
- -réia
- -terapia

A
- A negativo
- A positivo
- ACHE

C
- CID

a cont.
- apendicítico
- apitoxina
- apoptose
- arcorreia
- asclépio
- assepsia
- astenia
- ataque
- atenuação
- atoxicar
- atrofia

c cont.
- cálculo
- câimbra
- cânula

d
- decesso
- densitometria óssea
- dentista
- depressão
- derrame
- desintoxicado

Figure 3.10: Page de la catégorie «*Medicina*» du *Wikcionário*

Étant donné qu'il s'agit d'un wiki, les informations apportées peuvent être erronées. Aussi convient-il de prendre avec précaution ces informations, tout comme celles de Wikipédia. C'est la raison pour laquelle les dictionnaires en ligne (tel que celui de *Priberam*, par exemple) sont des sources beaucoup plus sûres. Mais comme le *Wikcionário* et Wikipédia sont en constante actualisation, car ils évoluent en même temps que la langue, ce qui fait qu'ils ont une couverture beaucoup plus grande et permettent ainsi une actualisation permanente du lexique médical.

14. `http://pt.wiktionary.org/wiki/Categoria:Medicina_(Português)`

3.2.4 *Descritores em Ciências da Saúde*

Comme nous l'avons vu précédemment dans la section 2.2.6, malgré le fait que cette terminologie se rapporte principalement à la norme brésilienne du portugais, le DeCS est la terminologie médicale de la langue portugaise la plus complète, en raison du fait qu'il soit principalement traduit à partir du MeSH de forme manuelle. Le DeCS a un total de 30.369 descripteurs, dont 25.671 proviennent de la traduction du MeSH [15].
Ce système offre plusieurs informations très utiles, telles que des traductions en anglais et en espagnol, les synonymes pour le portugais et les termes médicaux relatifs. Tout comme Wikipédia, le DeCS nous donne la possibilité de créer automatiquement une structure hiérarchique en raison de son organisation. Pour chaque entrée du dictionnaire, nous avons l'information du ou des chemins possibles pour cette entrée (cf figure 2.5 du chapitre antérieur).

Par conséquent, comme pour Wikipédia et le *Wikcionário* nous avons conçu un robot d'indexation afin de recueillir toutes les données du DeCS ainsi que son thésaurus.

Dans la figure 3.11, nous voyons la liste des catégories présentent dans le DeCS. Cependant, nous avons décidé d'éliminer les catégories suivantes, car leurs contenus s'éloignaient beaucoup trop de notre sujet :

- *denominações geográficas*
- *vigilância sanitária*
- *características de publicações*
- *ciência e saúde*
- *assistência à Saúde*
- *denominações de grupos*
- *ciência da informação*
- *ciências humanas*
- *technologia, indústrias, agricultura*
- *antropologia, educação, sociologia e fenômenos sociais*

Dans cette même figure, nous montrons, pour l'exemple de la catégorie «*Anatomia*» les sous-catégories qu'elle contient. Le robot d'indexation, parcours alors chacune des catégories et entre dans toutes les sous-catégories pour atteindre un article (cf. figure 3.12).
C'est un ainsi que nous recueillons toutes les entrées du DeCS, c'est-à-dire près de 25.688 entrées, accompagnées de leurs chemins hiérarchiques.

3.3 Synthèse sur la création de notre base de données

Par conséquent, nous avons recueilli des termes de sept sources différentes complétés par trois structures hiérarchiques spécifiques qui devront être unifiées pour parvenir à une structure du type Metathésaurus de l'UMLS.
Dans ce chapitre, nous avons proposé une méthodologie de collecte de données à partir de la toile afin de créer la plus grande terminologie médicale existante du Portugais. Toutefois, ces diverses terminologies ont une structure différente que nous avons alors décidé de standardi-

15. Données trouvées sur le site en 2010.

Figure 3.11: Les différentes catégories présentent dans le DeCS

ser en utilisant un codage XML (*eXtensible Markup Language*). Ceci sera l'objet du prochain chapitre.

Les informations contenues dans les plates-formes de collaboration libre d'accès peuvent être incorrectes. Il convient donc de traiter ces informations avec prudence, comme celles de Wikipédia ou du *Wikcionário*. C'est la raison pour laquelle les dictionnaires en ligne (tel que le DPLP, par exemple) sont des sources plus fiables que Wikipédia ou que le *Wikcionário*. Cependant, nous verrons dans les chapitres suivants que de tels dictionnaires présentent également des problèmes de cohérence, contrairement à leur version papier. Mais d'autre part, comme Wikipédia et le *Wikcionário* sont constamment mis à jour, ils évoluent dynamiquement avec la langue. C'est la raison pour laquelle ils ont une couverture terminologique beaucoup plus grande et une mise à jour permanente du lexique médical.

Des études menées par diverses entités ont conclu que le système lui-même (Wikipédia ou *Wikcionário*) est également chargé d'examiner et de corriger tous les cas d'erreurs, car un internaute peut écrire, corriger les erreurs et les fautes d'orthographe, traduire des articles et diffuser des idées ou participer à des discussions pertinentes. Ainsi, des erreurs ou des cas de

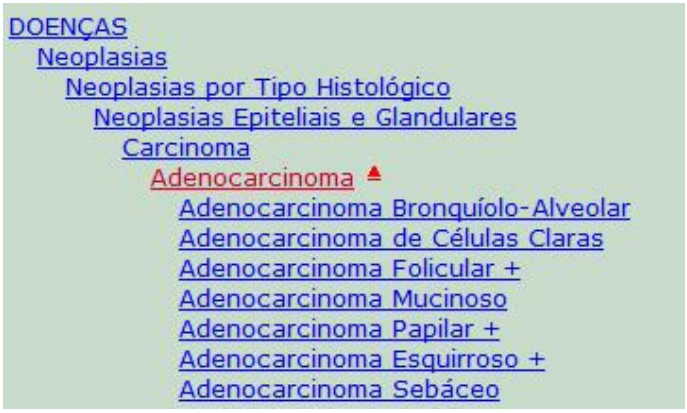


Figure 3.12: Partie de l'arbre du DeCS

«vandalisme» sont généralement corrigés ou éliminés par un internaute collaborateur. En 2005, une étude menée par le journal britannique *Nature* (Giles, 2005) a montré que, même s'il existe des erreurs, la Wikipédia est quasiment au même niveau que l'Encyclopédie Britannique.

Chapitre 4

Codage des données et problèmes rencontrés

Après avoir recueilli toutes ces données relatives à la médecine, nous les avons encodées sous un même langage informatique de balisage générique. Ce dernier répond à certaines règles qui sont explicitées dans ce chapitre. Cependant, ce codage des données a rencontré de nombreux problèmes. Nous allons en citer les principaux dans une seconde partie.

4.1 Codage des données

Nous avons terminé un travail initial de mise en forme des différentes informations. Ce travail consiste à transformer les données recueillies sous format électronique HTML (*HyperText Markup Language*), sous un format XML (de l'anglais *eXtensible Markup Language* (Ross, 2007) qui signifie "langage extensible de balisage") de description des données, permettant alors l'exploration du contenu de chaque base de données. Ci-dessous, nous présentons l'exemple *Adenocarcinoma*, montrer précédemment dans la figure 3.7, dans son format original HTML (figure 4.1) ainsi que sa transformation en format XML (figure 4.2).

```
<h2>Adenocarcinoma</h2>
<br />
<br />
<p><em>s. m. (fr. adénocarcinome; ing. adenocarcinoma). Epitelioma cuja
estrutura lembra de forma grosseira a de uma glândula. Sin. de
adenocancro (pouco usado).</em></p>
<br />
<br />
<br />
<br />
<p>
    Fonte: <a href="http://www.climepsi.pt" target="_blank"
title="CLIMEPSI">CLIMEPSI</a>
</p>
```

Figure 4.1: Format HTML de l'entrée *Adenocarcinoma* du MP

Lors de l'extraction de ce terme, nous constatons que nous obtenons le mot d'entrée, ses informations morphologiques, sa définition, un synonyme accompagné de l'indication de son emploi dans la langue et ses traductions en anglais et en français. En outre, nous avons également d'autres informations complémentaires qui se réfèrent à l'origine de la ressource d'où nous avons extrait l'information (dans ce cas, l'éditeur Climepsi [1]) avec la date de l'extraction de l'information et l'adresse internet du site de cette source. Ces informations sur les droits d'auteur sont indiquées pour chaque entrée de l'UMLP, comme nous le voyons dans la figure 4.3, où

1. http://www.climepsi.pt/

```
<entry id="456">
  <word>Adenocarcinoma</word>
  <source>Médicos de Portugal</source>
  <trusted_source>http://www.climepsi.pt/</trusted_source>
  <url search_date="2008-02-27" type="html"
    >http://medicosdeportugal.saude.sapo.pt/action/10/glo_id/456/menu/2/</url>
  <category>s.</category>
  <number>singular</number>
  <gender>m.</gender>
  <definition>Epitelioma cuja estrutura lembra de forma grosseira a de uma glândula. </definition>
  <synonyms>
    <synonym id="455">
      <word>adenocancro</word>
      <usage>raramente empregado</usage>
    </synonym>
  </synonyms>
  <translation lang="en">
    <word>adenocarcinoma</word>
  </translation>
  <translation lang="fr">
    <word>adénocarcinome</word>
  </translation>
</entry>
```

Figure 4.2: Format XML de l'entrée *Adenocarcinoma* du MP

apparaissent (en surligné) les différents chemins de la structure hiérarchique pour arriver au même terme, *adenocarcinoma*, mais cette fois-ci dans Wikipédia.
Toutes les données récoltées ont été encodées dans le même langage informatique selon une Définition de Type de Document commune (cf. section 4.1.1), pour préserver la cohérence et faciliter l'unification et la manutention de ces sept bases de données. Sa validation a également été effectuée à travers du *software oXygen\XML Editor* [2].

4.1.1 Création des différents XMLs selon une même DTD

4.1.1.1 Définition d'une DTD

La Définition de Type de Document (DTD) est définie dans Goldberg & Éric Jacoboni (2009) [3] de la façon suivante :

> Une DTD est un ensemble de règles permettant de définir un langage à balises personnalisé en XML. Elle se contente essentiellement d'identifier les éléments et leurs attributs. Un document XML qui ne respecte pas les règles définies par sa DTD est considérée comme *non valide* pour ce langage particulier. Le test de validation permet de déterminer rapidement si un document XML respecte ou non les règles [. . .] mises en place pour votre langage.

4.1.1.2 Spécification de la DTD

Tout fichier XML doit être créé et répondre aux règles d'une DTD pour être validé. Nous avons alors créé une DTD unique valable pour les sept XMLs correspondants à chacune de nos ressources. Cette DTD, présentée dans l'Annexe B.1, est explicitée ci-dessous :

2. http://www.oxygenxml.com
3. http://books.google.fr/books?id=KfzQYvg8wHoC

```
<entry id="3421">
  <word>Adenocarcinoma</word>
  <source>wikipedia</source>
  <url doc_date="29 de novembro de 2009." search_date="5 de Dezembro de 2009" type="html"
    >http://pt.wikipedia.org/wiki/Adenocarcinoma</url>
  <paths>
    <path>Medicina\Especialidades médicas\Oncologia\Tipos de câncer</path>
    <path>Medicina\Doenças\Tipos de câncer</path>
  </paths>
  <definition>Adenocarcinoma é um cancro (neoplasia maligna) que se origina em tecido glandular.
    Para ser classificado como um adenocarcinoma, as células não necessariamente precisam fazer
    parte de uma glândula, contanto que elas tenham características secretórias. Esta forma de
    carcinoma pode ocorrer em alguns mamíferos, incluindo humanos. O termo adenocarcinoma é
    derivado de 'adeno', que significa 'pertencente a uma glândula' e 'carcinoma', que descreve um
    cancro que se desenvolveu em células epiteliais. Ele pode se originar inicialmente como um
    adenoma (um tumor glandular que é benigno).</definition>
  <categorias>Tipos de câncer |</categorias>
  <translation lang="en">
    <word>Adenocarcinoma</word>
  </translation>
  <translation lang="fr">
    <word>Adénocarcinome</word>
  </translation>
  <translation lang="sp">
    <word>Adenocarcinoma</word>
  </translation>
</entry>
```

Figure 4.3: Format XML d'une entrée de Wikipédia avec l'indication des différents chemins taxonomiques possibles

<word> (ligne 5) fait référence à l'entrée principale (appelée entrée directe) du dictionnaire qui peut être identifiée par un nombre identificateur **ID** (ligne 6), ou encore par un **type** (ligne 7) de mot, dans notre cas, une famille de mots.

<source> (ligne 8) fait référence au nom de la ressource d'où a été retirée l'information.

<trusted_source> (lignes 9) fait référence au site d'où a été retirée l'information fournie par la ressource[4].

<url> (lignes 10 à 13) fournit diverses informations sur le site de l'origine de la source : l'adresse du site, la date de retrait de l'information (**search_date**), la date du document en ligne (**doc_date**), ainsi que le type de document (**type**).

<etymology> (ligne 14) fournit l'étymologie de l'entrée principale ou du synonyme.

<domain> (ligne 15) nous informe sur le(s) domaine(s) ou sous-domaine(s) auxquels l'entrée fait référence.

<paths> (lignes 16 et 17) indique le(s) chemin(s) utilisé(s) dans la structure hiérarchique pour arriver à l'entrée (cf. figure 4.3).

<category> (ligne 18) fournit la catégorie grammaticale de l'entrée principale ou d'un synonyme.

<number> (ligne 19) fournit le nombre de l'entrée principale, d'un synonyme ou bien d'une forme nominale relative (related_noun).

4. Dans trusted_source, l'information "OPS" signifie Organisation Panaméricaine de la Santé.

<gender> (ligne 20) fournit le genre de l'entrée principale, d'un synonyme, d'une forme adjectivale relative (related_adj) ou bien d'une forme nominale relative (related_noun).

<plural> (ligne 21) indique le pluriel de l'entrée principale, ou d'un synonyme.

<definition> (ligne 22) fournit la définition relative au terme de l'entrée principale.

<image> (lignes 23 et 24) fournit la légende de l'image en question ainsi que son **url**.

<categorias> (ligne 25) montre la ligne des catégories définie par Wikipédia pour chaque entrée de cette ressource.

<synonyms> (lignes 26 à 28) fournit au moins un synonyme de l'entrée principale, qui peut être accompagnée par l'information de son étymologie, catégorie grammaticale, nombre, genre, utilisation dans la langue et abréviation. Peut être identifié par un nombre identificateur **ID** (ligne 28).

<antonym> (ligne 29) fait référence à l'antonyme de l'entrée principale.

<related_adj> (ligne 30) fournit au moins une forme adjectivale relative au terme de l'entrée principale, peut être accompagnée par l'information de son emploi dans la langue, du genre et de sa traduction.

<related_nouns> (lignes 31 à 33) fournit au moins une forme nominale relative au terme de l'entrée principale, peut être accompagnée par l'information de l'utilisation dans la langue, du genre et du nombre. Peut être identifié par un nombre identificateur **ID** (ligne 33).

<related_verb> (ligne 34) fournit au moins une forme verbale relative au terme de l'entrée principale.

<related_word> (ligne 35) fournit au moins un terme relatif au terme de l'entrée principale, peut être accompagnée par l'information de son utilisation dans la langue.

<usage> (ligne 36) fournit diverses informations sur l'emploi du terme auquel l'étiquette se réfère.

<abbreviations> (ligne 37 à 39) fournit au moins une abréviation de l'entrée principale, qui peut être accompagné par l'information de son emploi dans la langue. Peut être identifié par un nombre identificateur **ID** (ligne 39).

<symbols> (lignes 40 et 41) fournit au moins un symbole ou sigle (par exemple, pour la nomenclature des enzymes (EC)) de l'entrée principale.

<compound> (ligne 42) indique une autre forme de présentation de l'entrée principale en tant que mot-composé.

<translation> (ligne 43 et 44) fournit au moins une traduction de l'entrée principale, en anglais britannique ou américain, et/ou en espagnol et/ou en français.

Ces étiquettes sont très importantes car elles nous permettent, après l'analyse des données de chaque entrée, de classer un maximum d'informations qui pourront être, par la suite, considérées soit comme des entrées directes (pour le terme principal à définir), soit comme des entrées indirectes (pour les autres termes des autres étiquettes).

4.2 Problèmes rencontrés

Durant le codage des données, nous avons rencontré de nombreux problèmes, aussi bien au niveau de la création automatique des XMLs, qu'au niveau de la création manuelle. Nous allons

voir tout d'abord les problèmes généraux qui ont rendu difficile le codage de chacune des ressources, puis nous montrerons divers exemples selon les problèmes types trouvés.

4.2.1 Principales difficultés de traitement des données rencontrées dans chaque ressource

Que ce soit de forme manuelle ou de forme automatique, chaque ressource a posé au moins un problème de codage lors de la création de son XML. Nous allons alors voir, pour chaque terminologie, les principaux problèmes causés.

4.2.1.1 DPLP

Comme le DPLP n'a pas d'indice de son contenu terminologique, nous avons donc dû rechercher manuellement toutes les entrées en relation avec la Médecine à partir des abréviations des domaines et sous-domaines.
Au départ, un XML (et sa DTD associée) a été créé pour chaque domaine et sous-domaine (cf. liste des domaines dans la section 3.1.1). Ensuite, nous avons décidé de créer un seul XML, qui regroupe tous les XMLs créés à partir du DPLP. Cependant, comme chacun était construit selon une DTD différente, une unification au niveau des DTDs a dû être faite, dans le but de créer également une unique DTD. De plus, des ajustements ont donc dû être effectués au niveau des étiquettes en raison de la création d'une seule DTD.

4.2.1.2 GMTM

Le GMTM étant construit pour différentes langues, nous avons créé un XML pour chacune des langues extraites (l'anglais, l'espagnol, le français et le portugais). Cependant, chaque terme défini dans chacune des langues garde le même numéro identificateur afin de rester toujours lié aux autres. Nous aurions pu procéder de la même forme que pour les autres ressources, c'est-à-dire insérer les traductions dans l'entrée correspondante ; mais comme le GMTM est formé différemment, et que pour toutes les langues, nous avons la possibilité d'avoir une définition, il nous était impossible de mettre toutes ces données dans un même fichier XML sans créer de confusions avec les autres XMLs des autres ressources.

4.2.1.3 CHCB

Le CHCB est un petit glossaire, mais ses entrées sont très bien expliquées et surtout très développées. La longueur des définitions pose problème principalement au niveau du traitement des données pour la recherche de synonymes ou autres.

4.2.1.4 MP

Nous avons rencontré de nombreux problèmes dans l'analyse des données extraites du MP par le robot d'indexation. Le XML créé par ce dernier comportait de nombreuses erreurs dues à une structure non uniformisée des données de départ. Donc pour nous aider à résoudre toutes ces erreurs, nous avons créé une petite interface d'aide à la correction, comme nous pouvons le voir dans la figure 4.4.
Dans cette interface, nous voyons (de droite à gauche) les données de départ à droite telles qu'elles apparaissent sur le site *Médicos de Portugal*, le XML créé automatiquement, et sur la

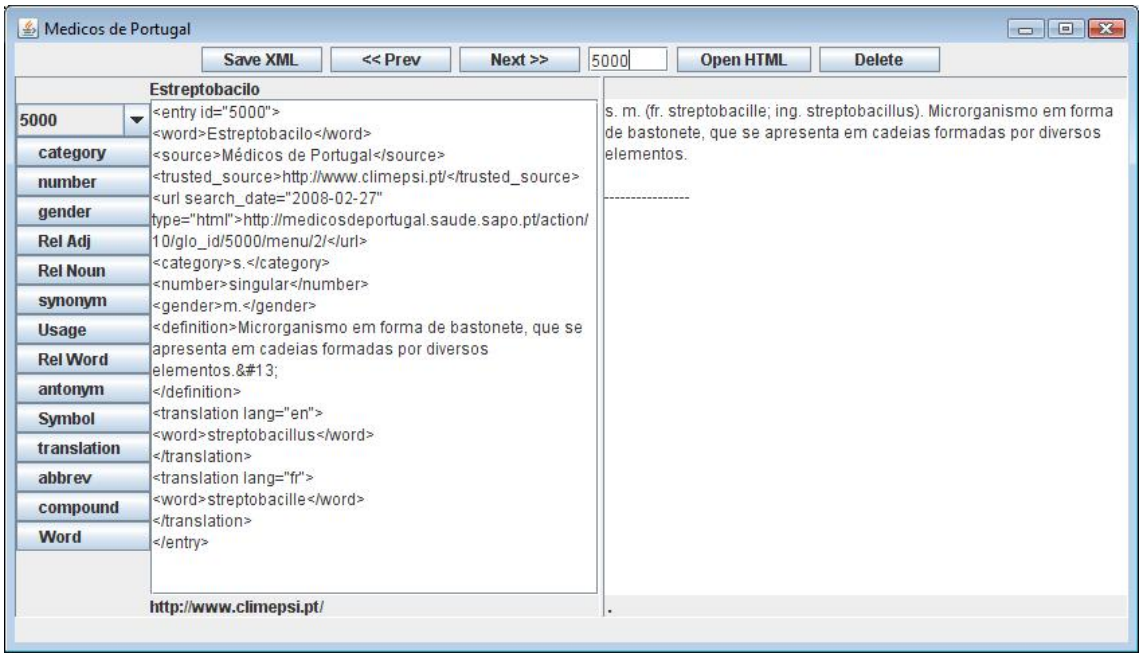


Figure 4.4: Interface créée pour l'aide à la correction du XML du MP

gauche des boutons qui permettent de créer plus rapidement de nouvelles étiquettes dans le XML. Sur le dessus, d'autres boutons nous permettent d'enregistrer les modifications, d'avancer ou reculer dans les entrées du XML, de voir la page HTML des données de départ ou encore d'effacer du XML l'entrée en cours d'analyse.

Les inconvénients qu'il est pertinent de relever sont les suivants : la création manuelle d'une nouvelle entrée dans le XML dûe au fait d'une seule même entrée posséder les définitions de deux termes médicaux ; la classification des synonymes ; la correction de différents problèmes orthographiques et syntaxiques telle que l'absence d'espace entre les termes de l'entrée principale ou de la définition, les nombreuses corrections orthographiques, la correction des traductions avec l'aide des dictionnaires en ligne suivants :

- `http://www.answer.com`
- `http://www.free-medical-dictionary.net`
- `http://dictionnaire.reverso.net`
- `http://www.cnrtl.fr/definition`
- `http://littre.reverso.net`
- `http://medical.dictionary.thefreedictionnary.com`
- `http://www.books.google.fr` ((Nicoulin *et al.*, 2004)).

Cette liste non exhaustive, énumère certaines des difficultés rencontrées lors de la mise en format XML des données du MP.

4.2.1.5 Wikipédia

Avec Wikipédia, le principal grand problème a été la création du robot d'indexation, en raison des nombreuses pages récursives que le site contient. Effectivement, le programme entrait dans des cycles, d'où il ne pouvait plus sortir ; ce qui signifie qu'il est incapable d'extraire toutes les

données et d'aboutir car il entre dans une boucle sans fin.

4.2.1.6 *Wikcionário*

Dans le *Wikcionário*, le problème de la récursivité était moins important, mais le grand inconvénient a été le fait d'avoir énormément de termes relatifs dans chaque entrée. Cela rend les entrées du XML très lourdes, et très souvent sans intérêts.

4.2.1.7 Decs

La principale difficulté du DeCS est dûe à l'existence du pluriel dans de nombreux termes d'entrées principales. Ceci est problématique en raison de la future unification, qui ne pourra pas unifier automatiquement un même terme au singulier et au pluriel.
De plus, nous avons trouvé également des définitions en anglais ; très souvent des parenthèses dans les entrées ou dans d'autres étiquettes (telles que <synonym> ou encore <traduction>), ces dernières ont été traitées quand la résolution de ce problème était claire et évidente. Dans les autres cas, elles sont restées telles quelles afin d'être évaluées par les professionnels de la santé.
Le dernier point gênant à souligner fait référence à la présence de termes d'entrées identiques, mais de domaines différents, raison pour laquelle les définitions respectives soient également différentes.

4.2.2 Problèmes de cohérence intra-terminologique

Nous avons terminé la présentation des problèmes généraux lors de la création de chaque XML, nous allons à présent décrire les difficultés rencontrées pour le remplissage des étiquettes du XML. Ces dernières ont été définies automatiquement au moment de la création des XMLs par les différents robots d'indexation. Il est à noter que très souvent, la construction des données de départ dans chaque ressource diverge et donc le programme ne sait pas choisir le bon contenu et, par conséquent, remplir les différentes étiquettes.
Il a fallu parer à cette difficulté, et pour cela une étude sur un certain pourcentage de chaque base de données, plus spécifiquement sur les définitions, a dû être réalisée. Et ensuite nous avons fait correspondre toutes les données trouvées aux XMLs.

4.2.2.1 Recherche de synonymes

La première étape consiste à analyser tout le texte extrait des sept ressources indépendamment, afin de trouver les différentes amorces qui puissent indiquer la présence d'un ou plusieurs synonymes. Dans le tableau 4.1, nous présentons la liste de toutes ces amorces qui peuvent être des termes, des abréviations, des symboles ou encore des marques de ponctuation.
Pour trouver les possibles synonymes qui suivent ces amorces, nous effectuons, de forme automatique, une recherche dans chaque définition de chaque entrée en utilisant le mode de recherche rapide dans chaque XML pour chaque amorce. Ensuite nous analysons manuellement le contenu des données afin de vérifier s'il s'agit bien d'un ou plusieurs synonymes. Par exemple, dans le MP, avec l'amorce *diz-se também*, nous trouvons :

Diz-se também *vacina DT-TAB*.

Table 4.1: Liste des termes, abréviations et symboles qui indique la présence de synonymes

ou	variante :	nome/designação vulgar
/	variação/ções	nome/designação comercial
=	variedade	conhecido/designado (também) como/por
sin.	sinónimo(s)	também designado/chamado/referido
(. . .)	o mesmo que	também conhecido
chamado(a) também	também é chamado(a)	diz-se também
(também . . .)	comummente chamado	também se diz
denominado	também transcrito como	or vezes referida como
nomeado	também grafado	designação internacional
comercializado	também relatado	designação original
comercializado	nome/designação antiga	antiga designação
nome popular	antigamente chamada de	anteriormente denominada
antigamente	melhor que	anteriormente conhecida como
nomes alternativos	forma preferível a	emprega-se mais vulgarmente
O m. q.	às vezes chamada	vulgarmente chamada/conhecido/designado
em linguagem clínica	em linguagem corrente	;. . . , . . . , . . .

Dans un autre exemple tiré du *Wikcionário*, nous tronvons deux synonymes grâce à l'amorce « ;. . . , . . . , . . . » :

> *exame minucioso de um cadáver, realizado por especialista qualificado, para determinar o momento e a causa da morte **;** necropsia**,** necroscopia.*

De plus, étant donné que nous considérons toujours les termes de la norme brésilienne comme des synonymes de l'entrée directe, alors nous avons recueilli des amorces qui nous permettent de trouver ces variantes (cf. tableau 4.2).

Table 4.2: Liste des termes et abréviations qui indique la présence d'une variante brésilienne

(no Brasil)	(brasileirismo)
(Brasil)	(br)

Nous citons alors un exemple trouvé dans Wikipédia qui nous a permis d'identifier un synonyme de la variante brésilienne grâce à l'amorce «*no Brasil*» :

> *Os antitússicos (antitussígenos, **no Brasil**) são os fármacos utilizados no tratamento sintomático da tosse.*

Ces listes d'amorces nous ont permis de rajouter de nombreux synonymes, et donc d'avoir des XMLs beaucoup plus complets. Mais nous ne pouvons pas oublier de préciser, qu'à chaque fois que ces amorces sont accompagnées d'une information d'indication temporelle (*anteriormente denominada*), de fréquence (*às vezes chamada*) ou autre (*nome popular*), nous insérons toujours cette information dans l'étiquette <usage> qui accompagne l'étiquette <synonym>.

4.2.2.2 Recherche des antonymes

À l'identique de la recherche de synonymes, nous procédons en analysant le texte dans sa totalité, afin de trouver des amorces qui indiquent la présence d'antonymes. Dans le tableau 4.3, nous présentons les amorces utilisées.

Dans le DPLP, nous trouvons, par exemple dans la définition du terme *supinação*, l'amorce «*opõe-se a*» :

Table 4.3: Liste des termes qui indique la présence d'antonymes

em oposição a	opõe-se a	o oposto
o contrário	contrário de	

> P*osição da mão de forma que a palma fique voltada para cima ou para diante (**opõe-se a** pronação).*

Dans Wikipédia, nous présentons un exemple plus complexe qui montre que la vérification manuelle de l'extraction des données est primordiale. Dans le cas contraire, les données recueillies seront fausses :

> ***O contrário do** hipertireoidismo, a falta de hormonas tireoidianas, é conhecido como hipotireoidismo.*

Dans ce cas une analyse précise de la phrase est essentielle, pour pouvoir correctement annoté l'antonyme car si nous prenons le terme qui suit immédiatement l'amorce, l'information extraire sera alors incorrecte.

4.2.2.3 Recherche des formes adjectivales

La recherche des formes adjectivales est beaucoup plus complexe que celle des synonymes ou des antonymes. Dans les textes nous pouvons trouver trois formes d'amorces (cf. tableau 4.4). Elles peuvent apparaître sans parenthèse et à différents endroits de la définition, ce qui complique la recherche automatique.

Table 4.4: Liste des termes ou abréviations qui indique la présence de formes adjectivales

(adjectivo . . .)	(adj. :. . .)	(a. :. . .)

Par exemple, dans le MP, nous trouvons :

> *Estudo das bactérias (estrutura, funções, condições de vida, etc.). V. microbiologia. (**adj**. : bacteriológico).*

Dans cet exemple, l'information de la forme adjectivale est donnée en fin de définition, ce qui nous permet alors de compléter l'étiquette <related_adj>.
Cette manière permet de trouver des formes adjectivales. Il en existe une autre. Nous allons l'illustrer avec l'exemple de la figure 4.5.

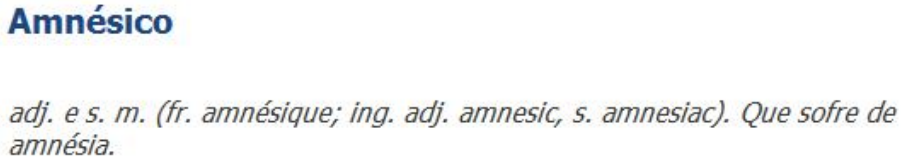


Figure 4.5: Entrée *amnésico* du MP qui montre la présence d'une forme adjectivale

La particularité de cette entrée est qu'elle présente deux catégories grammaticales (adjectif et substantif : *adj. e s. m.*), et de plus elle indique également deux traductions anglaises (une pour le substantif : *s. amnesiac* et l'autre pour l'adjectif : *adj. amnesic*).

La première particularité est fréquente dans les différentes ressources et elle est toujours encodée de la même forme. Tout d'abord, quand ce type d'entrée apparaît, nous constatons que la définition est identique pour les deux formes, donc nous n'avons pas besoin de créer une seconde entrée. Il nous suffit de garder l'information du substantif comme entrée directe, et celle de l'adjectif dans l'étiquette correspondante. Nous prenons le soin néanmoins, de mettre un identificateur à cette forme adjectivale, qui est évidemment le même que l'entrée directe afin de maintenir l'information de départ *adj. e s. m.*. Cette méthode de traitement est expliquée dans la figure 4.6.

```
<entry id="690">
  <word>Amnésico</word>
  <source>Médicos de Portugal</source>
  <trusted_source>http://www.climepsi.pt/</trusted_source>
  <url search_date="2008-02-27" type="html"
    >http://medicosdeportugal.saude.sapo.pt/action/10/glo_id/690/menu/2/</url>
  <category>s.</category>
  <number>singular</number>
  <gender>m.</gender>
  <definition>Que sofre de amnésia.</definition>
  <related_adj>
    <word id="690">Amnésico</word>
    <translation lang="en">
      <word>amnesic</word>
    </translation>
  </related_adj>
  <translation lang="en">
    <word>amnesiac</word>
  </translation>
  <translation lang="fr">
    <word>amnésique</word>
  </translation>
</entry>
```

Figure 4.6: Le XML de l'entrée *amnésico* du MP qui montre la présence d'une forme adjectivale

En ce qui concerne le cas de la seconde particularité, nous devons également l'encoder. Pour cela, nous insérons l'information de la traduction de l'adjectif dans l'étiquette <related_adj> afin d'éviter les confusions avec la traduction du substantif qui se trouve à sa place habituelle.

4.2.2.4 Recherche des formes nominales

Pour trouver des formes nominales, nous avons utilisé principalement les amorces présentent dans le tableau 4.5.

Table 4.5: Liste des termes ou abréviations qui indique la présence de formes nominales

(s. m. : . . .)	(s. f. : . . .)	o profissional . . .
o especialista é/chama-se	o profissional é	

Ces formes sont particulièrement indiquées dans le MP. Prenons deux exemples :

*Que atrasa ou impede a coagulação do sangue. (**s. m.** : anticoagulante.)*

Ce premier exemple fait référence aux types d'amorces les plus fréquemment trouvées pour indiquer une forme nominale. Nous pouvons trouver également de nombreuses informations du type :

*Estudo da alergia e das suas manifestações patológicas. **O especialista chama-se** alergologista.*

Ces types d'amorces existent pour nous informer sur le nom du spécialiste du domaine ou sur le nom du porteur d'une maladie. Par exemple, nous trouvons dans Wikipédia :

O fisioterapeuta onco-funcional deve estar apto para desenvolver suas atividades com pacientes infantis [. . .]

Tem a designação de nefropata a pessoa portadora de algum tipo de nefropatia ou doença nos rins.

Une fois de plus, nous constatons qu'une révision manuelle est fondamentale afin de trouver ces cas particuliers, car une exploration automatique faite uniquement en se fondant sur les amorces est sans succès.

4.2.2.5 Recherche des formes verbales

Pour les formes verbales, nous n'avons pas eu de difficultés car nous n'en trouvons que dans le MP, et elles ont été traitées avec le robot d'indexation.

4.2.2.6 Recherche de termes relatifs au terme de l'entrée

En ce qui concerne les termes relatifs, leur identification fonctionne de la même façon que pour celles des synonymes et des antonymes. Les amorces sont dans le tableau 4.6.

Table 4.6: Liste des termes, abréviations et symboles qui indique la présence de termes relatifs

ver . . .	veja	ver (também)
vd.	cf.	V. este(s) termo(s))
v . . .	v. . . .	Ver este(s) termo(s)

Après avoir recherché de manière automatique les termes qui suivent ces amorces et effectuer leur vérification, nous les classons ensuite dans l'étiquette <related_word>. Montrons à présent deux exemples trouvés dans différentes ressources :

Substância cujo consumo contribui para assegurar o ciclo regular da vida do indivíduo. V. nutriente.

Dans l'exemple suivant, nous découvrons deux synonymes :

*Estado num organismo que, depois de ter estado em contacto com certas substâncias (principalmente proteínas), adquire delas propriedade de reacção, úteis ou não, que são produzidas mesmo em pequenas quantidades. (**Ver** alergia e anafilaxia.)*

De plus chaque fois que nous rencontrons l'expression *distingue(m)-se*, elle nous donne également souvent une information que l'on peut encoder dans cette même étiquette car elle nous informe sur d'autres termes qui sont en lien avec le terme de l'entrée directe, comme :

> ***Distingue-se*** *em geral a audimudez de compreensão e a audimudez de expressão.*

4.2.2.7 Recherche des abréviations

Comme pour la recherche de synonymes et d'antonymes, nous devons analyser le texte afin de trouver des amorces qui relèvent des abréviations. Dans le tableau 4.7, nous présentons les amorces utilisées.

Table 4.7: Liste des termes et abréviations qui indique la présence d'abréviations

abrev.	Por ext.	abreviado
abreviação habitual	abreviação normal	abreviação comun
sigla :	sigla de	acrónimo de

Citons deux exemples trouvés respectivement dans le DPLP ainsi que dans le DeCS :

> *Relação entre a idade mental (determinada através de testes) e a idade real ou cronológica de um indivíduo, multiplicada por cem. (**sigla : Q.I.**).*

> *Volume extra de ar que pode ser inspirado com o máximo de esforço alcançado no final de uma inspiração normal, tranquila. A **abreviação comum** é VRI.*

De plus, si les abréviations trouvées sont accompagnées d'une information sur la forme de son emploi dans la langue, très souvent mis entre parenthèse, alors ces informations sont insérées dans l'étiquette <usage>, de la même forme que pour les synonymes.

4.2.2.8 Recherche des symboles

Comme précédemment, nous recherchons en un premier temps les amorces pour identifier ensuite les symboles. Cependant, dans cette étiquette, nous ne classons pas uniquement des symboles.
Pour ne pas créer une nouvelle étiquette et rendre nos XMLs encore plus lourds et complexes, nous décidons de mettre également sous cette même classification ce qui relève des formules chimiques, des symboles atomiques ainsi que des numéros EC (de l'anglais *Enzyme Commission Numbers*) qui font référence aux numéros du schéma de la classification officielle des enzymes. Les amorces trouvées dans tout le texte sont citées dans le tableau 4.8.

Table 4.8: Liste des termes et abréviations qui indique la présence de symboles

símbolo :	fórmula	símbolo atómico
EC :	fórmula química :	

Nous pouvons citer :

> *Histamina, **formula química** C5H9N3, é amina biogênica envolvida em processos bioquímicos de respostas imunológicas. . .*

> *Um elemento radioativo trivalente e o membro protótipo da família dos actinídeos. Apresenta como **símbolo atômico** Ac.*

4.2.2.9 Recherche de l'étymologie

Lors de notre étude, nous avons remarqué qu'il existait également des références aux étymologies des termes. Les amorces trouvées sont dans le tableau 4.9.

Table 4.9: Liste des termes et abréviations qui indique la présence de l'étymologie

do inglês . . .	do francês . . .
do latim . . .	do grego . . .

Dans le MP, plusieurs exemples font référence à l'étymologie au cœur même des définitions :

*A palavra acne vem **do grego** akme quer dizer "eflorescência", "ponto de elevação".*

*Medo mórbido dos grandes espaços vazios, acompanhado, por vezes, de sensação de vertigem. Ling. : **Do grego** ágora, praça pública.*

4.2.2.10 Recherche du/des domaine(s)

Suite à l'analyse des textes, nous constatons que la marque du domaine est très régulièrement placée entre parenthèses ou crochets avant la définition (premier exemple tiré du DPLP et le second du *Wikcionário*), ou encore entre parenthèse après le terme de l'entrée directe (troisième exemple tiré du DeCS) :

[Anatomia] Relativo à retina.

(Odontologia) remoção cirúrgica da raiz de um dente ou da polpa.

Recognição (Psicologia)
O conhecimento ou percepção que alguém ou algo presente tenha sido encontrado anteriormente.

Cependant, nous trouvons dans les définitions de nombreuses références aux domaines en question avec les amorces citées dans le tableau 4.10.

Table 4.10: Liste des termes et expressions qui indique la présence du domaine

Parte de	Especialidade	Especialidade médica
Ciência	Ramo	Estudo
Campo especial	Campo clínico	Campo de
Disciplina	Área de	Área do conhecimento
Área da medicina	Área de estudo	para a
utilizado na	, na. . .	

Ces amorces d'identification du domaine en question se trouvent dans toutes les ressources recueillies, nous pouvons l'illustrer des exemples suivants :

*A neurociência computacional é a **área da** neurociência que tem por objetivo propor modelos matemáticos.*

*É uma **parte ou ramo da** Biologia e é de grande importância para o homem por seus envolvimentos médicos.*

Dans ces divers cas, nous retirons cette information de l'étiquette <definition> pour la placer dans l'étiquette <domain>.

4.2.3 Principales corrections et modifications effectuées au niveau des XMLs

Lors de l'analyse du codage des données, nous nous apercevons qu'il existe de nombreuses irrégularités, tant au niveau de l'entrée en générale, du terme à définir, de la définition, qu'au niveau des autres étiquettes du XML. Pour résoudre ces difficultés, dans la plupart des cas, la révision manuelle a été l'unique solution.

4.2.3.1 Modification du titre en portugais luso-africain

Wikipédia et *Wikcionário* nous informe très souvent si l'entrée définie fait référence au portugais luso-africain ou bien au portugais brésilien. Grãce à ces informations, nous pouvons faire des modifications au niveau du titre de chaque entrée (c'est-à-dire à l'entrée directe) afin de mettre comme terme principal le luso-africain et le brésilien en synonyme, sans oublier de préciser dans l'étiquette <usage> qu'il s'agit de la variante brésilienne.
Par exemple, le robot d'indexation qui a extrait les données du *Wikcionário* a donc extrait l'entrée *Caxumba*[5]. Cependant en haut de la page correspondante nous sommes informés qu'il s'agit d'un terme de la norme brésilienne, et nous sommes renvoyés à un autre terme : *papeira*, qui est de la norme luso-africaine. Dans ces cas, nous mettons le terme luso-africain en tant qu'entrée directe, et le terme brésilien en synonyme.
Il en est de même avec Wikipédia où le terme *internação compulsória* est défini de la façon suivante :

> *Internamento compulsivo (**português europeu**) ou internação compulsória (**português brasileiro**) é a prática de utilizar meios ou formas legais como parte de uma lei de saúde mental [. . .].*

Dans ce type de cas, nous procédons de la même forme que précédemment, et nous substituons le terme d'entrée par le terme luso-africain et mettons en synonyme le terme brésilien.
Afin de trouver toutes les entrées qui ont ce type d'information, nous avons relevé toutes les amorces qui indiquent si le terme est de la norme luso-africaine ou brésilienne. Ces amorces sont répertoriées dans le tableau 4.11.

Table 4.11: Liste des termes et abréviations qui indique l'origine du terme : soit brésilienne, soit luso-africaine

PT	BR	pt	br
PE	PB	português europeu	português brasileiro
em Portugal	no Brasil	(Brasil)	(Portugal))

4.2.3.2 Modifications effectuées dans les entrées

De nombreuses corrections ont été effectuées au niveau du titre de l'entrée, de la définition, mais également dans les autres étiquettes. Voyons quelques exemples.

5. `http://pt.wiktionary.org/wiki/caxumba`

Au sein du titre (ou de l'entrée directe) :
Dans le titre, nous sommes partis de la simple correction à la plus complexe, c'est-à-dire, que nous avons par exemple corrigé des fautes d'accentuation ou des erreurs de frappe, telle que la présence de virgules après les termes de l'entrée directe : <word>*Balanopostite*,</word>.
Le retrait des virgules était nécessaire car celles-ci causaient des problèmes au niveau des renvois des identificateurs (ID).
Nous avons également modifié totalement les titres lorsque la source présentait, comme entrée directe, une abréviation non commune, par exemple : *HOMA*, est une abréviation dérivée de l'anglais. Dans ce type de cas, nous regardons le début de la définition qui contient, en général, le terme en langue portugaise.

> *O HOMA (do inglês homeostatic model assessment, em português Modelo de avaliação da Homeostase)[. . .].*

Ici, nous choisissons *Modelo de avaliação da Homeostase* comme entrée directe et le terme de départ (*HOMA*) est, quant à lui, rangé sous l'étiquette <abbreviation>, sans oublier d'insérer sous l'étiquette <usage> l'indication qu'il s'agit d'une abréviation dérivée d'une expression anglaise :

```
<abbreviations>
    <abbreviation>
        <word>HOMA</word>
        <usage>do inglês: homeostatic model assessment</usage>
    </abbreviation>
</abbreviations>
```

Cependant, il est possible de rencontrer des abréviations simples dans les entrées directes, comme par exemple : "D". Dans ce cas, l'abréviation est mise dans l'étiquette <abbreviation> et "dioptria" prend sa place en tant qu'entrée directe.
De plus, d'autres modifications sont à effectuer lorsque nous avons des entrées directes du type : <word>xxxx (yyyyy de)</word>. Dans ces cas, nous mettons le groupe nominal en extension dans l'entrée directe, et la forme fléchie dans l'étiquette <compound>. Par exemple pour l'entrée : <word>Malpighi (pirâmides de)</word>, nous mettons cette entrée en extension, ce qui nous donne <word>pirâmides de Malpighi</word>, et plaçons la forme originale fléchie dans l'autre étiquette : <compound>Malpighi (pirâmides de)</compound>, ce qui évite de perdre tout information de départ de toute source confondue.
De la même forme, nous trouvons également des entrées directes du type : <word>xxxx (ou yyyy)</word>. Dans ce type de cas, le premier terme (ou groupe nominal) reste dans l'entrée directe et le second est mis comme synonyme.

Au cœur de la définition :
Dans les définitions, nous avons également fait de nombreuses corrections d'ordre orthographique ainsi que de fautes de frappe. Il arrive que nous ayons eu des erreurs non communes, telle que la répétition de la définition. Cette dernière est en effet répétée deux fois, l'une à la suite de l'autre. Ce type d'erreur a été causé par le robot d'indexation.
De plus, ce programme informatique n'a parfois pas retiré toute la définition relative à une entrée. Le cas a été visible lorsque la définition se termine part deux points, par exemple : <definition> [...] :</definition>. Il a fallu rechercher la fin de la définition dans la ressource en question, de façon manuelle.

Nous avons également retiré certaines expressions qui font référence à un renvoi dans le propre site de la source, par exemple : "*(ver itens específicos nesse site)*", "*que estão escritos em itens específicos deste site*", "*e está descrita em outro item desse site*", "cada uma delas apresentadas em itens específicos desse site" ou encore "(ver a seguir)". Mais si nous trouvons une information du type : "(ler Hipertireoidismo e Hipotiroidismo)" et que cette information existe dans les entrées que nous avons extrait de la ressource, alors ces deux références sont mises dans l'étiquette <related_word>. Il en est de même pour l'expression "*Ver diversas formas de fetor*", où nous avons retiré cette expression de la définition pour mettre sous l'étiquette <related_word> les différentes entrées en question.

Dans les autres étiquettes
Dans les autres étiquettes, des erreurs de dédoublement de termes existent. Nous présentons un exemple du MP dans la figure 4.7 :

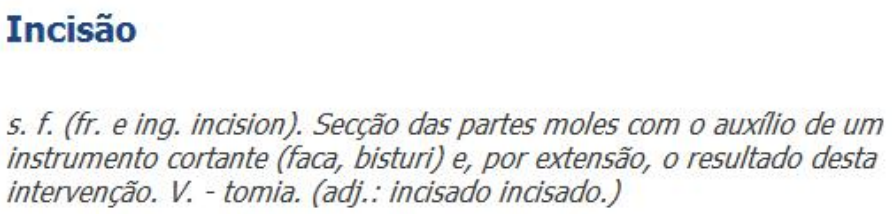


Figure 4.7: Exemple d'une entrée à dédoublement de termes

Finalement, l'erreur ne vient pas du robot d'indexation, mais de la propre ressource. Par conséquent après extraction des données, nous avons corrigé toutes les étiquettes <related_adj> qui ont un dédoublement du terme adjectival.
De plus, nous avons eu le cas de figure de répétitions entre synonymes et termes relatifs, ou entre antonymes et termes relatifs. La solution trouvée a consisté à éliminer les termes relatifs, puisque les termes sont déjà mentionnés soit en tant que synonyme, soit en tant qu'antonyme.
Quand nous avons un terme qui peut être soit masculin, soit féminin, alors nous le codifions dans les XMLs avec : "n." dans l'étiquette <gender>. Il en est de même avec les termes qui peuvent être soit singulier, soit pluriel, ils sont codifiés "n." dans <number>.

4.2.3.3 Création de nouvelles entrées

Après avoir analysé minutieusement les différentes entrées recueillies, nous nous apercevons que nous pouvons diviser certaines entrées, afin d'avoir seulement une définition par entrée.
Dans un premier exemple tiré du *Wikcionário*, nous montrons un cas où une définition est insérée à l'intérieur d'une autre définition, il s'agit de la définition du terme *anestesiologista* :

> *que ou pessoa especialista em anestesilogia, especialidade médica que estuda os métodos e medicamentos anestésicos.*

Nous constatons qu'à l'intérieur de la définition du terme *anestesiologista*, il existe la définition du terme *anestesiologia*. La décision a été prise de créer une nouvelle entrée avec le terme et la définition appropriée, quand ce même terme n'existe pas dans la ressource en question.
Un autre exemple, est celui de la numérotation à l'intérieur d'une définition. Par exemple le terme *alérgico* dans le MP :

1) adj. : que se refere à alergia ou que dela resulta. Ex. : conjuntivite alérgica.
2) adj. e s. m. : que é sujeito à alergia.

Nous avons ici deux définitions distinctes qui montrent, en plus, une différence au niveau de la catégorie grammaticale du terme à définir. Mais nous pouvons trouver le problème de ne pas avoir de numérotation pour distinguer les différentes définitions. Par conséquent, il faut que l'analyse soit faite avec prudence pour éviter les erreurs, comme dans la définition de *fissura* dans le MP où nous distinguons trois définitions différentes, pour trois termes différents dans une même entrée. Ce qui signifie qu'à partir de cette entrée, nous obtiendrons donc trois entrées différentes :

Fenda anatómica ou patológica. ***Fissura óssea : fenda, sem fractura total, de um osso (também designada, em linguagem corrente, racha)****. Fissura cutânea : greta ou rágada. (adj. : fissurado.)*

Dans le DPLP, plusieurs des définitions comportaient la ponctuation " ;". Ce symbole représentait une séparation entre les différentes définitions possibles pour un même terme. Devant ce cas, nous avons créé une nouvelle entrée pour chaque définition.
Dans le DeCS, certaines définitions présentent plusieurs définitions de sources différentes, par exemple :

Constituição ou afecção do corpo que fazem com que os tecidos reajam de maneira especial a determinados estímulo extrínsecos, consequentemente tendendo a tornar o indivíduo mais suscetível a determinadas doenças que o normal. (Tradução livre do original : MeSH) Afecção na qual existe uma diminuição da resistência de um indivíduo frente a determinada doença ou intoxicação e que se experimenta com dose a exposições inferiores às habitualmente nocivas para o resto da população. (Fonte : Tesauro REPIDISCA, CEPIS/OPS/OMS, para o conceito Suscetibilidade)

Comme pour les autres exemples, nous créons une nouvelle entrée pour chaque source mentionnée dans la définition.

Nous pourrions encore citer de nombreux exemples tous différents les uns des autres, au sujet de la création de nouvelles entrées. Mais notre principal objectif est de faire comprendre que ce procédé est nécessaire en raison de la future unification. En effet, chaque ressource ne contient pas les mêmes définitions pour chaque terme, c'est-à-dire, deux ressources peuvent définir le même concept nº1, mais une seule peut avoir la définition d'un second concept. Et si les définitions n'étaient pas séparées, nous ne pourrions pas, par la suite, unifier les termes qui ne définissent pas un même concept.

4.2.3.4 Élimination d'entrées

Certaines entrées des ressources ont dû être éliminées, soit parce qu'elles étaient marquées par un domaine autre que celui de la Médecine et de ses sous-domaines, soit parce que la définition n'était pas marquée par un des domaines de la médecine et sans lien avec notre sujet.
Nous présentons dans l'exemple suivant deux entrées synonymes tirées du MP :

Trichomonas intestinalis
sin. de Pentatrichomonas hominis.

> ***Pentatrichomonas hominis***
> *Espécie de Trichomonas que vive habitualmente no intestino grosso do homem. Parasita anódino quando em pequeno número, pode provocar irritação intestinal e diarreia em caso de infecção maciça (tricomoníase intestinal). Sin. de Trichomonas intestinalis.*

Dans la première, nous avons seulement l'indication d'un synonyme, dans la seconde, plus complète, nous constatons qu'elle a comme synonyme le terme de la première entrée. Nous pouvons donc éliminer l'entrée qui ne présente que le synonyme. Ce type de phénomène arrive également avec les termes relatifs ainsi qu'avec les abréviations.

Dans le GMTM, nous éliminons directement les entrées qui ne présentent aucune information, du type :

```
    <word>/</word>
    <synonyms>
        <synonym>
            <word>/</word>
        </synonym>
    <synonyms>
ou
    <word>?</word>
    <synonyms>
        <synonym>
            <word>?</word>
        </synonym>
    </synonyms>
```

4.2.3.5 Rajout de traductions

Dans la continuité de notre analyse sur les définitions, nous remarquons la présence de traductions au sein même des définitions, par exemple :

> *Colêmese(francês : cholémèse ; inglês : cholemesis) é a presença de biles no vômito.*

Dans ces cas, nous avons retiré ces informations de la définition, pour les insérer sous les étiquettes appropriées, ici, <translation lang="en"> et <translation lang="fr">.

4.2.3.6 Problèmes de renvoi par le numéro identificateur

Nous avons inséré de forme automatique les identificateurs (id) pour les étiquettes : <synonym>, <antonym>, <related_adj>, <related_noun>, <related_verb>, <related_word> et <abbreviation>, afin de créer une meilleure harmonie entre les différentes entrées des terminologies. Regardons la figure 4.8.
Ces identificateurs font référence au numéro de l'entrée à laquelle ils correspondent. Dans cet exemple, le synonyme existe déjà en tant qu'entrée directe dans l'entrée numéro 457. Par ailleurs, nous pourrions tout à fait éliminer une des entrées puisqu'elles sont synonymes et qu'elles définissent un même concept. Dans la majorité des cas, nous ne l'avons pas fait afin de préserver toutes les informations recueillies, telles que l'étymologie, les traductions, etc.

```
<entry id="5352">
  <word>Fibroadenoma</word>
  <category>s.</category>
  <number>singular</number>
  <gender>m.</gender>
  <definition>Adenoma caracterizado pela presença de abundante
    tecido conjuntivo fibroso.</definition>
  <synonyms>
    <synonym id="457">
      <word>adenofibroma</word>
    </synonym>
  </synonyms>
</entry>
```

Figure 4.8: Entrée qui comporte un synonyme avec identificateur

Cependant, le fait d'insérer automatiquement les id, pose parfois certains problèmes. Quand un même terme (dans le sens de même graphie) possède plusieurs définitions, donc plusieurs entrées, le programme informatique ne le reconnaît pas. Nous avons eu recours à un autre programme informatique afin de relever ces ambiguïtés et pour nous faciliter la vérification manuelle.
Avec les identificateurs, nous rencontrons un autre type de difficulté : il arrive qu'un terme renvoie à une entrée grâce au numéro id, mais le synonyme en question n'est pas le terme de l'entrée directe auquel il renvoie. En effet, il arrive qu'un id fasse référence à une étiquette autre que celle de l'entrée directe.
De plus, nous constatons lors de cette analyse sur les renvois liés au numéro identificateur, l'importance de l'orthographe des termes. En effet, s'il existe une erreur orthographique ou autre, le programme ne reconnaît pas le terme, et donc ne lui associe pas de numéro identificateur, d'où l'importance d'une unification lexicale correcte.

4.3 Récapitulatif quantitatif relatif aux entrées

Dans cette section, nous voulons simplement montrer la quantité d'informations que nous avons extrait des données des sept terminologies afin de remplir les différentes étiquettes possibles de nos XMLs. Dans le tableau 4.12, nous présentons pour chacune de nos terminologies traitées, les entrées directes, les entrées indirectes et les traductions.
Nous entendons par entrée directe, le terme principal à définir ; par entrée indirecte, tout terme présent dans une des étiquettes qui fasse référence au : pluriel, synonyme, antonyme, abréviation, symbole, forme adjectivale, forme nominale, forme verbale, et terme relatif. Toutes ces entrées indirectes pourraient très bien être des entrées directes.

4.4 Synthèse du codage des données

Dans ce chapitre, nous avons présenté notre façon de coder les sept terminologies extraites en format HTML. Après avoir transformé ces données en un même langage XML, soit de forme

6. Le remplissage de cette étiquette est expliqué dans la section 6.2.1.

Table 4.12: Tableau récapitulatif du nombre d'entrées possibles pour chaque base de données

		MP	DPLP	GM	Wikip	Wikc	DeCS	CHCB
Entrées directes		12828	3615	1925	7367	2050	25688	195
Entrées indirectes	pluriels [6]	25	3	5	20	12	1856	11
	synonymes	2152	361	935	3065	867	37502	78
	antonymes	22	3	0	3	42	0	0
	abréviations	427	6	1	531	7	2237	27
	symboles	139	40	0	120	3	686	0
	related_adj	1133	0	0	1	1	0	0
	related_noun	112	0	0	50	3	0	0
	related_verb	18	0	0	0	0	0	0
	related_word	274	3	0	60	4155	567	26
Traductions	anglais (BR)	10960	0	1826	5470	1020	25685	0
	anglais (USA)	6	0	0	0	0	0	0
	espagnol	0	0	1941	3465	848	25685	0
	français	12428	0	1921	3586	752	0	0

manuelle, soit de forme automatique, nous avons dû analyser minutieusement les données encodées afin de résoudre toute irrégularité. Ce travail a demandé énormément de temps et de minutie en raison des nombreux détails qui, s'ils étaient laissés au dépourvu, causeraient énormément d'erreurs dans le dictionnaire final.
Maintenant que nos sept XMLs sont créés sous un bon format, nous pouvons commencer à présent l'unification intra-terminologique.

Chapitre 5

Unification des sources

Les divergences de la langue portugaise peuvent poser des problèmes de compréhension au niveau de l'information médicale. Pour cette raison, il est nécessaire de créer un lexique médical unifié pour permettre plus d'entente au niveau terminologique entre les variantes du portugais, c'est-à-dire luso-africaine et brésilienne (cf. annexe C.1).
Une grande partie des difficultés rencontrées peuvent être résolues par l'application des règles du "Nouvel Accord Orthographique de la Langue Portugaise". Cependant, pour nous aider à trouver les problèmes d'unification restants et afin de les résoudre, nous utiliserons une formule mathématique appelée Distance de Levenshtein, qui va nous permettre de trouver à la fois les termes d'entrées proches, et les définitions identiques ou similaires.

5.1 Unication par l'Accord Orthographique

5.1.1 Définition de l'Accord Orthographique

Après avoir recueilli et encodé chacune de nos sept bases de données, nous passons à présent à la première étape de l'unification : l'application du Nouvel Accord Orthographique de la Langue Portugaise (AO) (Assembleia da República, 1991).

Signé le 16 Décembre 1990 par des représentants officiels d'Angola, du Brésil, de Cap Vert, de Guinée-Bissau, de Mozambique, du Portugal et de São Tomé e Príncipe (Timor-Leste n'a rejoint l'Accord qu'en 2004 après avoir obtenu son indépendance.), cet accord est un traité international qui prétend instituer une orthographie officielle unifiée pour la langue portugaise, afin d'être utilisée par tous les pays de langue officielle portugaise.
L'AO a pour objectif explicite de mettre fin à l'existence de deux normes orthographiques officielles divergentes, l'une au Brésil et l'autre dans les autres pays de langue officielle portugaise, contribuant ainsi à augmenter le prestige international de cette langue.
L'AO (Casteleiro & Correia, 2008; Janssen, 2008) privilégie le critère phonétique au critère étymologique, et n'interfère ni dans les différences orales, ni dans les variations grammaticales ou lexicales, et ne concerne que la graphie de l'écriture.
Pour éviter une grande partie des conflits orthographiques entre les normes cultes luso-africaine et brésilienne, nous appliquons donc toutes ces règles à nos différentes ressources.

De forme brève et concise, les nouvelles règles de l'AO de 1990 qui s'appliquent à notre terminologie peuvent se diviser en quatre groupes :

(i) La suppression graphique des consonnes muettes, est une règle qui existe uniquement dans

la norme culte luso-africaine car elle s'applique déjà à la norme culte brésilienne depuis longtemps :

reação (et pas *reacção*)
injetar (et pas *injectar*).

(ii) La suppression de certains accents graphiques dans les deux normes :

tranquilizante (et pas *tranqüilizante* au Brésil)
traqueia (et pas *traquéia* au Brésil)
paranoia (et pas *paranóia* dans les deux normes).

(iii) La réduction et la systématisation de l'emploi du tiret des deux côtés de l'Atlantique :

ultrassonografia (et pas *ultra-sonografia*)
autoimunidade (et pas *auto-imunidade*)
hiporreflexia (et pas *hipo-reflexia*)
céu da boca (et pas *céu-da-boca*).

(iv) Cependant, la coexistence de la double graphie de certains mots se maintient selon :
(1) la prononciation de certaines séquences consonantiques (comme : -ct-, -pç- ou -gd-). Dans :

olfactivo/olfativo
concepção/conceção
amígdala/amídala

(2) l'accentuation dans l'articulation des mots ou dans le timbre de certaines voyelles, comme : les voyelles ***e*** et ***o*** suivies d'une consonne nasale ***m*** ou ***n***. Comme :

gémeo/gêmeo
vómito/vômito
pénis/pênis
clónus/clônus.

5.1.2 Application de l'Accord Orthographique

Dans notre recherche, l'AO est plus un outil de correction que d'unification. En effet, un grand nombre de problèmes d'unification peut être résolu avec la simple application des règles de cet accord.

Ce que l'on peut nommer comme le "premier niveau d'unification" permet de résoudre des problèmes simples tels que ceux présentés dans le tableau 5.1.

Table 5.1: Exemples de problèmes d'unification simples résolus avec l'application des règles de l'AO.

Norme luso-africaine [1]	Norme brésilienne	Après l'AO
Diarreia	Diarréia	Diarreia
Fractura	Fratura	Fratura
Tracto	Trato	Trato
Enjoo	Enjôo	Enjoo
Adenóide	Adenóide	Adenoide
Sanguínea	Sangüínea	Sanguínea
Injecção	Injeção	Injeção

Nous vérifions donc que, grâce aux règles de l'AO, il n'existe plus aucune difficulté d'unification pour les problèmes illustrés dans le tableau 5.1. Cependant, l'AO permet dans certains cas, la coexistence des graphies des deux variantes, comme présentée dans le tableau 5.2.

Table 5.2: Exemples de coexistence des deux graphies dans l'univers de la langue portugaise.

Norme luso-africaine	Norme brésilienne
Olfacto	Olfato / Olfacto
Contracetivo	Contraceptivo
Oxigénio	Oxigênio
Insónia	Insônia
Fémur	Fêmur
Abdómen	Abdômen
Bebé	Bebê

Danc ces cas, l'unification devra être résolue par d'autres formes que nous verrons dans le chapitre suivant.

Ces règles de l'AO ont donc été appliquées aux terminologies de chacune des sept ressources avec toutes ses spécificités qui ne sont pas ici listées. Pour cela, nous avons recherché dans chacune d'entre elles, toutes les entrées directes et indirectes qui contenaient une erreur relative aux règles de l'AO. Par exemple, dans Wikipédia et dans le DeCS (ressources qui présentent plus de termes dans la norme brésilienne), nous avons trouvé des problèmes de suppression d'accent graphique[2]. Pour corriger ces erreurs, nous avons collecté tous les termes avec l'indication des règles citées ci-dessus, et corrigé de la manière suivante : la variante brésilienne et/ou la graphie antérieure à l'AO de 1990 sont synonymes de l'entrée principale dans la norme luso-africaine. Nous montrons dans les tableaux 5.3 et 5.4, l'exemple pour le terme *sialorreia*.

Table 5.3: Entrée *sialorréia* avant l'application des règles de l'AO

```
<entry id="4945">
   <word>Sialorréia</word>
   <definition>Sialorréia é a secreção
   abundante de saliva.</definition>
</entry>
```

Table 5.4: Entrée *sialorreia* après l'application des règles de l'AO

```
<entry id="4945">
   <word>Sialorreia</word>
   <definition>Sialorréia é a secreção
   abundante de saliva.</definition>
   <synonyms>
      <synonym>
         <word>Sialorréia</word>
         <usage>variante brasileira antes do AO
         de 1990</usage>
      </synonym>
   </synonyms>
</entry>
```

Sachant que cette règle s'applique seulement aux termes de la norme brésilienne, nous pouvons indiquer que le terme mis en synonyme est le terme de la "norme brésilienne avant l'AO"

1. Nous entendons par norme luso-africaine et norme brésilienne, les graphies antérieures à l'AO de 1990.

2. Pour cette règle de suppression d'accents graphique, nous avons effectué en tout, (pour les sept terminologies) près de 1.500 corrections.

(*variante brasileira antes do AO de 1990*).
Dans l'exemple précédent, nous montrions l'application de l'AO dans le terme d'entrée. Dans celui-ci, nous illustrons un cas plus complexe dans lequel l'entrée et son/ses synonyme(s) doit/doivent être modifié(s). Dans les tableaux 5.5 et 5.6 nous montrons la modification pour l'entrée *cartilagem aritenoide*.

Table 5.5: Entrée *cartilagem aritenóide* avant l'application des règles de l'AO

```
<entry id="343">
   <word>Cartilagem aritenóide</word>
   <definition>As cartilagens aritenóides são . . .
   </definition>
   <synonyms>
      <synonym>
         <word>Aritenóide</word>
      </synonym>
   </synonyms>
</entry>
```

Table 5.6: Entrée *cartilagem aritenoide* après l'application des règles de l'AO

```
<entry id="343">
   <word>Cartilagem aritenoide</word>
   <definition>As cartilagens aritenóides são . . .
   </definition>
   <synonyms>
      <synonym>
         <word>Cartilagem aritenóide</word>
         <usage>grafia anterior ao AO de 1990
         </usage>
      </synonym>
      <synonym>
         <word>Aritenoide</word>
      </synonym>
      <synonym>
         <word>Aritenóide</word>
         <usage>grafia anterior ao AO de 1990
         </usage>
      </synonym>
   </synonyms>
</entry>
```

Chaque fois que nous faisons une modification avec les règles de l'accord orthographique, nous mettons toujours le terme avant l'AO en synonyme avec l'indication, dans l'étiquette <usage>, qu'il s'agit de la "graphie avant l'application de l'AO" (*grafia anterior ao AO de 1990*).

Dans le cas des doubles graphies acceptées par l'AO pour chacune des normes (cf. exemples du tableau 5.2), la graphie de la norme luso-africaine est toujours mise comme entrée principale et nous gardons la graphie relative à la norme brésilienne en synonyme avec l'indication dans l'étiquette <usage> de *variante brasileira*. Nous illustrons cette situation dans les tableaux 5.7 et 5.8.

Dans l'annexe D, nous présentons le nombre de corrections effectuées grâce aux règles de l'AO, dans chaque étiquette de chacune des sept ressources.

Avant de réaliser quelconque unification basée sur la fusion des sept terminologies, nous décidons de vérifier la cohérence de chacune des ressources, individuellement, après l'application de l'AO.

Table 5.7: Entrée *nervo trigêmeo* avant l'application des règles de l'AO

```
<entry id="536">
   <word>Nervo trigêmeo</word>
   <definition>O nervo trigêmeo constitui . . .
   </definition>
</entry>
```

Table 5.8: Entrée *nervo trigémeo* après l'application des règles de l'AO

```
<entry id="536">
   <word>Nervo trigémeo</word>
   <definition>O nervo trigêmeo constitui . . .
   </definition>
   <synonyms>
      <synonym>
         <word>Nervo trigêmeo</word>
         <usage>variante brasileira</usage>
      </synonym>
   </synonyms>
</entry>
```

5.2 Unification par entrées

Après avoir appliqué les règles de l'AO aux entrées directes et indirectes des ressources des sept méthodologies, nous essayons de voir la cohérence de chacune d'elles, et plus précisément nous vérifions s'il existe des entrées répétées. Ce travail de vérification est nécessaire car ce type de problème n'existe pas dans des dictionnaires papier mais peut souvent arriver dans des ressources électroniques.
Comme nous avons pu le constater dans la section 4.2.2 du chapitre précédent, ce type de problème se produit fréquemment dans les sources électroniques, par conséquent toute cette phase de vérification est essentielle dans les dictionnaires papier.
Un dictionnaire papier de qualité permettra par la suite d'obtenir un dictionnaire automatique de qualité équivalente.

5.2.1 Définition de la Distance de Levenshtein

Pour faciliter cette recherche d'erreurs possibles, nous utilisons le calcul de la Distance de Levenshtein (Levenshtein, 1966), également appelée Distance d'édition (plus courant sous le nom anglophone : *Edit-Distance*), de manière à trouver des graphies semblables de forme automatique.

La distance de Levenshtein entre deux séquences de caractères X et Y mesure le nombre minimum d'opérations d'édition nécessaires pour convertir X en Y. Les opérations d'édition standard sont la substitution (*medi<u>t</u>ina* por *medicina*), l'insertion (*medi<u>tc</u>ina* por *medicina*) et la suppression (*medi<u>i</u>na* por *medicina*) d'un symbole, auxquelles certains systèmes ajoutent la transposition de deux symboles adjacents (*me<u>ci</u>dina* por *medicina*).

5.2.2 Application de la Distance de Levenshtein aux entrées directes

La distance de Levenshtein est très utile pour trouver des entrées identiques (sans aucune modification entre deux séquences de caractères, tout comme des entrées très proches, avec des possibles erreurs orthographiques ou encore des entrées au pluriel. À cette fin, nous avons appliqué cette mesure à toutes les entrées de chaque ressource individuellement, avec un nombre d'opérations allant jusqu'à trois. Dans les tableaux 5.9 e 5.10, nous illustrons un cas où

la distance d'édition a constaté une similitude entre *antibiótico* et *antibióticos* dans le MP.

Table 5.9: Entrée *antibiótico* avant l'unification

```
<entry id="259">
  <word>Antibiótico</word>
  <trusted_source>http://www.grupogci.net/
  </trusted_source>
  <definition>Substância química que
  interfere na capacidade da bactéria
  funcionar normalmente.</definition>
</entry>
```

Table 5.10: Entrée *antibióticos* avant l'unification

```
<entry id="260">
  <word>Antibióticos</word>
  <trusted_source>http://www.grupogci.net/
  </trusted_source>
  <definition>Substância química que
  interfere na capacidade da bactéria
  funcionar normalmente (lembre-se, as
  bactérias são organismos vivos). Pode inibir o
  seu crescimento (antibiótico bacteriostático)
  ou matar as bactérias (antibiótico
  bactericida). . . </definition>
</entry>
```

Dans cet exemple, nous avons d'un côté, une entrée au singulier accompagnée d'une brève définition, et de l'autre côté, une entrée au pluriel avec une définition beaucoup plus détaillée et qui insère complètement la définition de la première entrée.
Quand nous sommes confrontés à ce type de difficultés, nous gardons le terme d'entrée au singulier et la définition la plus détaillée, comme nous pouvons le vérifier dans la figure 5.11.

Table 5.11: Nouvelle entrée *antibiótico* après l'unification

```
<entry id="259">
  <word>Antibiótico</word>
  <trusted_source>http://www.grupogci.net/</trusted_source>
  <definition>Substância química que interfere na capacidade da bactèria funcionar
  normalmente (lembre-se, as bactèrias são organismos vivos). Pode inibir o seu crescimento
  (antibiótico bacteriostático) ou matar as bactérias (antibiótico bactericida). . . </definition>
</entry>
```

La distance de Levenshtein est traditionnellement appliquée pour déterminer le degré de similitude entre les mots. Vérifier si un dictionnaire est cohérent passe par vérifier toutes ses entrées. C'est pourquoi, dans l'annexe C.2, nous montrons quelques exemples des résultats obtenus lors de l'application de la distance d'édition aux entrées directes du DPLP.

5.2.3 Exemple d'analyse des résultats obtenus

Quand la distance de Levenshtein trouve deux termes proches ou identiques, nous devons tout d'abord, vérifier si les deux séquences ont réellement une similitude graphique. Par exemple :

ânus id="717" *vs* íris id="312" $\Rightarrow$ aucune similitude

afacia id="4353" *vs* afasia id="6119" $\Rightarrow$ existence de similitude

ooferectomia id="4653" *vs* ooforectomia id="4654" $\Rightarrow$ existence de similitude

Ensuite nous devons analyser les définitions pour savoir si les termes définissent un même concept. Par exemple :

afacia id="4353" *vs* afasia id="6119" ⇒ concepts différents
ooferectomia id="4653" *vs* ooforectomia id="4654" ⇒ concepts identiques

Après avoir trouvé des séquences semblables qui définissent un même concept, nous vérifions à l'aide de dictionnaires électroniques sur CD-Rom ou en ligne (par exemple : Houaiss, Infopédia (geral), Infopédia (termos médicos)), ou encore dans Google (seulement si les termes sont inexistants dans les dictionnaires précédents), si l'un des termes à une graphie erronée.
Si l'une des graphies est erronée, l'entrée correspondante est éliminée. Mais si les deux graphies sont correctes, et si les définitions sont identiques ou semblables, un des termes est mis en synonyme et son entrée correspondante est éliminée.
De plus, si l'une des entrées contient certaines informations (définition, synonymes, traductions, etc.) que l'autre ne comporte pas, ces informations sont également rajoutées à l'entrée gardée.

De nombreuses corrections ont été faites aux sept XMLs grâce à l'application de la distance d'édition. Nous en avons un récapitulatif dans le tableau 5.12, où nous mentionnons pour chaque ressource, le nombre total de paires trouvées par cette mesure. Ces listes comportaient énormément de bruit, et après avoir retiré ce surplus d'informations non nécessaires, nous avons garder les paires qui, pour nous, sont susceptibles d´être synonymes (colonne "Problèmes potentiels"). Parmi ces dernières, nous mentionnons celles qui était réellement un problème et que nous avons résolu (colonne "Problèmes réels").

Table 5.12: Distance de Levenshtein appliquée aux entrées de chaque ressource

	Nombre Total	Problèmes Potentiels	Problèmes Réels
DPLP	2290	132	17
MP	6917	54	38
GMTM	1009	1	1
DeCS	7689	101	1
Wikipédia	2178	118	22
Wikcionário	728	350	3
CHCB	49	3	3
Total	20860	840	85

5.2.4 Analyse de l'unification par entrée

Tout d'abord, nous avons appliqué la distance de Levenshtein aux entrées de chaque base de données avec une différence maximale de trois lettres, et nous avons obtenu les résultats indiqués dans le tableau 5.12. (Nous montrons une partie de la liste des termes obtenus avec leurs terminologies correspondantes, dans l'annexe E.1.)
Dans ce tableau, les "problèmes potentiels" se réfèrent aux paires qui sont susceptibles d'avoir deux entrées synonymes, comme indiqué dans le tableau 5.13 avec des exemples du MP. Dans l'exemple de la première ligne nous obtenons une fausse paire, c'est-à-dire après avoir examiné le contenu des deux entrées, nous concluons que *ácido linoleico* n'est pas synonyme de *ácido linolénico*. Cependant, dans la deuxième ligne, nous avons, avec le mot en gras (*antibióticos*) de la deuxième colonne, un cas d'unification par entrée, qui se réfère à la colonne "problèmes réels" du tableau 5.12. En particulier, nous notons dans ce dernier que le GMTM, le DeCS et le *Wikcionário* ne présentent quasiment aucun problème d'unification avec la distance de Levenshtein appliquée aux entrées, contrairement au MP et au DPLP, terminologies que nous pensions

plus fiables.

Table 5.13: Exemples d'unification avec l'application aux entrées de la distance de Levenshtein

Entrée	Terme proche avec *Edit-Distance*
ácido linoleico	ácido linolénico
antibiótico	**antibióticos** antiluótico antimicótico antimitótico antipirótico antipsicótico antitóxico

En appliquant ce calcul à toutes les entrées de chaque terminologie extraite, nous avons trouvé une série de sept problèmes. Tout d'abord, il existe des termes identiques dans le MP, avec des entrées tout à fait identiques, comme par exemple, *unidade hounsfield* (id="10796") et *unidade hounsfield* (id="6657"), dans les figures 5.1 et 5.2. Dans ces cas, lorsque nous trouvons deux entrées parfaitement identiques, nous ne gardons que celle qui donne le plus d'informations. Mais dans ce cas précis, nous avons dû rajouter une étiquette dans le XML afin de pouvoir retirer les parenthèses du terme d'entrée (cf. figure 5.3 par exemple, dans ce cas :

```
<entry id="6657">
<word>Hounsfield (unidade)</word>
<source>Médicos de Portugal</source>
<trusted_source>http://www.climepsi.pt/</trusted_source>
<url search_date="2008-02-27" type="html">http://medicosdeportugal.saude.sapo.pt/action/10/glo_id/6657/menu/2/</url>
<definition>Em tomodensitometria, designa a unidade de medida da densidade de um quadrado elementar (píxel).
As unidades Hounsfield repartem-se arbitrariamente por uma escala de - 1000 (o ar) a + 1000 (o osso denso),
encontrando-se a água no 0. (Hounsfield, Sir Godfrey Newbold, investigador inglês, prémio Nobel de 1971, co-inventor
com Allan Mac Leod Cormack do tomodensitómetro, 1919-.)
</definition>
<translation lang="en">
<word>Hounsfield unit</word>
</translation>
<translation lang="fr">
<word>unité Hounsfield</word>
</translation>
</entry>
```

Figure 5.1: Entrée *unidade Hounsfield* (1) avant l'unification

Ensuite, nous rencontrons des problèmes de nombre, à savoir la coexistence de la version au singulier et au pluriel dans une même source. Ces problèmes apparaissent dans le MP et le DeCS. Par exemple, comme nous avons pu voir précédemment dans l'exemple du MP (cf. tableaux 5.9 et 5.10), nous avons *antibiótico* (id="260") et *antibióticos* (id="259"). Il est toujours important, cependant, de vérifier les définitions. En effet, des différences peuvent exister entre les entrées au pluriel et au singulier, comme nous pouvons le voir avec l'exemple de *bexiga* et *bexigas* dans les figures 5.4 et 5.5, où *bexiga* fait référence à la vessie et *bexigas* aux boutons provoqués par la variole ou la varicelle (emploi populaire).

Généralement, comme dans le premier cas, lorsque nous appliquons l'*Edit-Distance* sur les entrées qui définissent un même concept et que nous avons, d'un côté, un terme d'entrée au singulier et de l'autre un terme d'entrée au pluriel, nous conservons alors, le terme au singulier

```
<entry id="10796">
<word>Unidade Hounsfield</word>
<source>Médicos de Portugal</source>
<trusted_source>http://www.climepsi.pt/</trusted_source>
<url search_date="2008-02-27" type="html">http://medicosdeportugal.saude.sapo.pt/action/10/glo_id/10796/menu/2/</url>
<definition>Em tomodensitometria, unidade de medida de densidade de um quadrado elementar (píxel). Estas unidades
  de medida repartem-se por uma escala de -1000 (ar) a +1000 (osso denso), na qual a água corresponde a 0.
  [Hounsfield, Sir Geoffrey, investigador britânico, prémio Nobel de Medicina e Fisiologia, com Allan MacLeod Cormack
  (1979), 1919-.]
</definition>
<translation lang="en">
    <word>Hounsfield unit</word>
  </translation>
<translation lang="fr">
    <word>unité Hounsfield</word>
  </translation>
</entry>
```

Figure 5.2: Entrée *unidade Hounsfield* (2) avant l'unification

```
<entry id="6657">
<word>Unidade Hounsfield</word>
<source>Médicos de Portugal</source>
<trusted_source>http://www.climepsi.pt/</trusted_source>
<url search_date="2008-02-27" type="html">http://medicosdeportugal.saude.sapo.pt/action/10/glo_id/6657/menu/2/</url>
<definition>Em tomodensitometria, designa a unidade de medida da densidade de um quadrado elementar (píxel).
As unidades Hounsfield repartem-se arbitrariamente por uma escala de - 1000 (o ar) a + 1000 (o osso denso),
encontrando-se a água no 0. (Hounsfield, Sir Godfrey Newbold, investigador inglês, prémio Nobel de 1971, co-inventor
com Allan Mac Leod Cormack do tomodensitómetro, 1919-.)
</definition>
<compound>
<word>Hounsfield (unidade)</word>
</compound>
<translation lang="en">
<word>Hounsfield unit</word>
</translation>
<translation lang="fr">
<word>unité Hounsfield</word>
</translation>
</entry>
```

Figure 5.3: Entrée *unidade Hounsfield* après l'unification

avec la définition la plus détaillée et avec le plus d'informations complémentaires (catégorie grammaticale, synonymes, traductions, etc.). Si nous ne savons pas quelle définition choisir, alors nous gardons les deux pour que les professionnels de la santé puissent décider plus tard sur notre application collaborative.
Alors que, dans le deuxième cas, si les définitions ne sont pas semblables car elles définissent des concepts différents, alors nous gardons les deux entrées sans effectuer aucune modification.

Troisièmement, nous trouvons fréquemment des entrées très semblables, avec les mêmes informations, comme par exemple dans les tableaux 5.14 et 5.15, avec acinésico (id="72") et acinético (id="73"). Dans ce cas, en plus d'avoir une graphie très proche, ces entrées ont la même etymologie, le même domaine, la même catégorie grammaticale et la même définition. Nous gardons alors une seule entrée, et l'autre est insérée dans l'étiquette <synonym> (cf. tableau 5.16).

Quatrièmement, nous avons trouvé des fautes d'orthographe, en particulier dans le MP. Par exemple, *ascardíase* (id="14") et *ascaridíase* (id="1224"), où la graphie de la première entrée n'existe pas (nous avons vérifié la graphie de ces deux termes sur les dictionnaires en ligne

```
<entry id="1509">
  <word>Bexiga Urinária</word>
  <source>Médicos de Portugal</source>
  <trusted_source>http://www.climepsi.pt/</trusted_source>
  <url search_date="2008-02-27" type="html"
    >http://medicosdeportugal.saude.sapo.pt/action/10/glo_id/1509/menu/2/</url>
  <definition>Reservatório com parede muscular, no qual a urina, que corre pelos ureteres (a
    partir do rim) se deposita nos intervalos entre as micções. Durante a micção, a urina deixa
    a bexiga pela uretra. A bexiga está situada na cavidade pélvica, atrás da sínfise púbica
    (no homem por cima da próstata, por cima e à frente do recto; na mulher, à frente do útero
    e da vagina). A sua capacidade fisiológica média é de cerca de 300 ml. Os orifícios
    ureterais situam-se nos dois ângulos superiores do trígono vesical, enquanto o orifício
    uretral (colo da bexiga) ocupa o seu ângulo inferior. A mucosa que reveste a cavidade
    vesical, lisa na criança, torna-se areolar no adulto, por hipertrofia das fibras
    musculares, e pode assumir, no idoso, um aspecto de bexiga de colunas.&#13; </definition>
  <synonyms>
    <synonym id="1507b">
      <word>bexiga</word>
      <usage>em linguagem clínica corrente, diz-se muitas vezes</usage>
    </synonym>
  </synonyms>
  <related_adj>
    <word id="3793">vesical</word>
  </related_adj>
  <related_word>
    <word id="2411">cístico</word>
  </related_word>
  <translation lang="en">
    <word>urinary bladder</word>
  </translation>
  <translation lang="fr">
    <word>vessie urinaire</word>
  </translation>
</entry>
```

Figure 5.4: Entrée *bexiga*

Table 5.14: Entrée *acinésico* avant l'unification

```
<entry id="72">
    <word>acinésico</word>
    <source>Priberam</source>
    <url search_date="2007-11-29"
         type="html">http:</url>
    <etymology>do Gr. a, priv. + kínesis
ou kinetikós, movimento</etymology>
    <domain>
        <word>Medicina</word>
    </domain>
    <category>adj.</category>
    <definition>calmante.</definition>
</entry>
```

Table 5.15: Entrée *acinético* avant l'unification

```
<entry id="73">
    <word>acinético</word>
    <source>Priberam</source>
    <url search_date="2007-11-29"
         type="html">http:</url>
    <etymology>do Gr. a, priv. + kínesis
ou kinetikós, movimento</etymology>
    <domain>
        <word>Medicina</word>
    </domain>
    <category>adj.</category>
    <definition>calmante.</definition>
</entry>
```

ainsi que sur le moteur de recherche Google [3]), et leurs définitions sont très semblables. Parfois, lorsque les entrées ont peu d'informations, nous devons vérifier si le terme est correct

3. http://www.google.com

```
<entry id="3209">
  <word>Varíola</word>
  <source>Médicos de Portugal</source>
  <trusted_source>http://www.climepsi.pt/</trusted_source>
  <url search_date="2008-02-27" type="html"
    >http://medicosdeportugal.saude.sapo.pt/action/10/glo_id/3209/menu/2/</url>
  <category>s.</category>
  <number>singular</number>
  <gender>f.</gender>
  <definition>Doença infecciosa grave, muito contagiosa, outrora epidémica e frequentemente
    mortal, devida a um poxvírus e caracterizada por um exantema que passa por diversas fases
    (pápulas, vesícula, pústulas umbilicadas). Em seguida a uma vasta campanha de erradicação
    realizada pela OMS a partir de 1965, a doença desapareceu totalmente. O vírus variólico
    conserva-se unicamente em dois laboratórios a título de amostra de referência. As
    vacinações, antigamente obrigatórias, já não se fazem.&#13; </definition>
  <synonyms>
    <synonym id="1510">
      <word>bexigas</word>
      <usage>nome desusado e popular</usage>
    </synonym>
  </synonyms>
  <translation lang="en">
    <word>smallpox</word>
    <word>variola</word>
  </translation>
  <translation lang="fr">
    <word>variole</word>
  </translation>
</entry>
```

Figure 5.5: Entrée *bexigas*

Table 5.16: Entrée *acinético* après l'unification

```
<entry id="73">
    <word>acinético</word>
    <source>Priberam</source>
    <url search_date="2007-11-29" type="html">http:</url>
    <etymology>do Gr. a, priv. + kínesis ou kinetikós, movimento</etymology>
    <domain>
        <word>Medicina</word>
    </domain>
    <category>adj.</category>
    <definition>calmante.</definition>
    <synonyms>
      <synonym>
        <word>acinésico</word>
      </synonym>
    </synonyms>
</entry>
```

dans d'autres sources. Dans ce cas, l'entrée du terme erroné est éliminée.

Cinquièmement, dans le DPLP et le MP il existe des entrées directes qui sont déjà des synonymes pour d'autres entrées. Par exemple, dans le DPLP, l'entrée *pediatra* (id="2335") a déjà le synonyme *pediatro* qui est, lui aussi, une entrée (id="2336"). Et pour ce dernier cas d'ailleurs, il

n'a pas de définition, mais seulement le synonyme *pediatra*. Dans ce cas de figure, nous conservons l'entrée qui a toutes les informations, à savoir la définition et ses synonymes, l'autre est alors éliminée.

Sixièmement, dans le DPLP nous avons des entrées dont les différences sont le domaine et la définition. Par exemple, *viro* (id="3293") du domaine de la biologie et *vírus* (id="3295") du domaine de la médecine. Dans ce cas, lorsque nous avons des domaines et/ou des définitions différentes, il n'y pas d'unification et les deux entrées coexistent.

Enfin, il existe dans toutes les terminologies, à l'exception du GMTM, des problèmes qui demeurent non résolus. Par exemple, dans le MP, nous avons *cancro da pele* (id="1927") et *cancro de pele* (id="16"). Dans ces cas, nous ne savons pas quel terme choisir, et nous gardons les deux entrées. Cependant dans le chapitre suivant, nous allons trouver d'autres méthodes d'analyse des données pour nous aider à résoudre ce type de difficultés problématiques intra et inter-terminologiques.

5.3 Unification par définitions

Toutefois, la distance de Levenshtein peut être également utilisée pour comparer les définitions des termes et ainsi trouver automatiquement des potentiels synonymes, dont l'encodage XML sera unifié, c'est-à-dire, il existera une et une seule entrée (la plus utilisée) avec la liste de ses synonymes. Dans ce cas, la distance d'édition calcule le nombre de mots à substituer, insérer ou éliminer entre définitions pour arriver à une même définition. Ce troisième traitement peut être vu comme une troisième unification en terme de codage.

5.3.1 Application de la Distance de Levenshtein aux définitions

Dans le cas suivant tiré du DPLP, nous identifions une unification par définition entre les termes *viral* et *vírico* qui montrent une distance de Levenshtein élevée en termes de mots, mais pas en termes de définition. En outre, ils ont la même catégorie grammaticale, les mêmes traductions et la seconde entrée a la première entrée comme synonyme, comme illustré dans les tableaux 5.17 e 5.18.

Dans ce cas, nous pouvons alors éliminer la première entrée (id="4374") parce que son information est déjà contenue dans l'entrée synonyme, ce que nous obtenons comme résultat au tableau 5.19.
Généralement, lorsque nous appliquons l'*Edit-Distance* sur les définitions, si les entrées ont, en plus de définitions très similaires, d'autres informations complémentaires identiques, nous gardons comme entrée la plus complète, et nous mettons en synonyme le terme de l'autre entrée.

Nous avons ensuite appliqué la distance de Levenshtein à toutes les définitions de chaque base de données avec une différence allant jusqu'à dix mots, et nous avons obtenu les résultats indiqués dans le tableau 5.20.

Comme dans le tableau 5.12, les "problèmes potentiels" font référence aux entrées qui sont

Table 5.17: Entrée *viral* avant l'unification

```
<entry id="4374">
  <word>Viral</word>
  <category> adj.</category>
  <definition>Relativo aos vírus. Ex. : infecção
  vírica, inactivação vírica.</definition>
  <translation lang="en">
     <word viral</word>
  </translation>
  <translation lang="fr">
     <word>viral</word>
  </translation>
</entry>
```

Table 5.18: Entrée *vírico* avant l'unification

```
<entry id="4373">
  <word>Vírico</word>
  <category>adj.</category>
  <definition>Relativo aos vírus. Ex. : infecção
  vírica, inactivação vírica.</definition>
  <synonyms>
    <synonym>
      <word>Viral</word>
    </synonym>
  </synonyms>
  <translation lang="en">
     <word>viral</word>
  </translation>
  <translation lang="fr">
     <word>viral</word>
  </translation>
</entry>
```

Table 5.19: Nouvelle entrée *vírico* après l'unification

```
<entry id="4373">
  <word>Vírico</word>
  <category>adj.</category>
  <definition>Relativo aos vírus. Ex. : infecção vírica, inactivação vírica.</definition>
  <synonyms>
    <synonym>
      <word>viral</word>
    </synonym>
  </synonyms>
  <translation lang="en">
     <word>viral</word>
  </translation>
  <translation lang="fr">
     <word>viral</word>
  </translation>
</entry>
```

Table 5.20: Distance de Levenshtein appliquée aux entrées de chaque ressource

	Nombre Total	Problèmes Potentiels	Problèmes Réels
DPLP	231	36	7
MP	426	45	20
GMTM	106	11	2
DeCS	1670	357	0
Wikipédia	136	26	4
Wikcionário	202	23	6
CHCB	0	0	0
Total	2771	498	39

susceptibles d'avoir deux définitions semblables, mais qui peuvent finalement ne pas être synonymes. Et la colonne des "problèmes réels", nous montre le nombre de paires que nous avons trouvé, et qui avait une synonymie des définitions, et par conséquent, des entrées.

Nous pouvons alors constater que l'application de cette distance d'édition nous a permis de trouver des définitions identiques, et donc, des entrées synonymes que nous avons alors unifiées.

5.3.2 Analyse de l'unification par définition

En appliquant la distance de Levenshtein aux définitions des sept terminologies, nous avons trouvé une série de six problèmes.
Tout d'abord, il existe des entrées dans les différentes ressources, sauf dans le DeCS, qui ont en plus de la définition, toutes les autres données supplémentaires identiques, comme par exemple, dans Wikipédia (cf. figures 5.6 et 5.7), *telodendro* (id="5935") et *nódulo de Ranvier* (id="6247").

```
<entry id="5935">
 <word>Telodendro</word>
 <source>wikipedia</source>
 <url type="html" search_date="5 de Dezembro de 2009" doc_date="6 de setembro de 2009."
  >http://pt.wikipedia.org/wiki/Telodendro</url>
 <paths>
  <path>Medicina\Especialidades médicas\Neurologia</path>
  <path>Medicina\Especialidades médicas\Neurologia\Sistema nervoso\Neurociência\Neurologia</path>
 </paths>
 <definition>A não continualidade da bainha forma espaçamentos isentos de mielina, os nódulos de
  Ranvier. Isto facilita um movimento mais ágil do impulso que vai ocorrendo em saltos, já que o
  impulso só se propaga com a presença de mielina.</definition>
 <categorias>Sistema nervoso | !Esboços sobre histologia |</categorias>
 <translation lang="en">
  <word>Nodes of Ranvier</word>
 </translation>
 <translation lang="fr">
  <word>Nœud de Ranvier</word>
 </translation>
 <translation lang="sp">
  <word>Nodo de Ranvier</word>
 </translation>
</entry>
```

Figure 5.6: Entrée *telodendro*

Dans ces situations, nous gardons l'une des entrées, et mettons le mot d'entrée de la seconde en synonyme sans autre condition, comme le montre la figure 5.8. De plus, nous voyons également dans cette figure, que nous en avons profité pour regrouper les différents chemins taxonomiques.

Deuxièmement, il y a des entrées qui ont seulement la définition et aucune autre information similaire. Par exemple, *hister-* (id="6559") et *metr-* (id="10291a"). Dans ces cas, nous préférons maintenir les deux entrées, sans les unifier, étant la plupart du temps des problèmes de terminologie technique.

Troisièmement, dans Wikipédia les définitions identiques ne peuvent être unifiées en raison du peu de détails de la définition, ce que nous mettons en relief dans les tableaux 5.21 et 5.22. L'*artéria cerebelar superior* (id="396") a la même définition que l'*artéria cerebral anterior* (id="397"), *artéria cerebral média* (id="398") et *artéria cerebral posterior* (id="399").

```
<entry id="6247">
 <word>Nódulo de Ranvier</word>
 <source>wikipedia</source>
 <url type="html" search_date="5 de Dezembro de 2009" doc_date="6 de setembro de 2009."
  >http://pt.wikipedia.org/wiki/N%C3%B3dulo_de_Ranvier</url>
 <paths>
  <path>Medicina\Especialidades médicas\Neurologia\Sistema nervoso</path>
  <path>Medicina\Especialidades médicas\Neurologia\Sistema nervoso\Neurociência\Sistema
  nervoso</path>
 </paths>
 <definition>A não continualidade da bainha forma espaçamentos isentos de mielina, os nódulos de
  Ranvier. Isto facilita um movimento mais ágil do impulso que vai ocorrendo em saltos, já que o
  impulso só se propaga com a presença de mielina.</definition>
 <categorias>Sistema nervoso | !Esboços sobre histologia |</categorias>
 <translation lang="en">
  <word>Nodes of Ranvier</word>
 </translation>
 <translation lang="fr">
  <word>Nœud de Ranvier</word>
 </translation>
 <translation lang="sp">
  <word>Nodo de Ranvier</word>
 </translation>
</entry>
```

Figure 5.7: Entrée *nódulo de Ranvier*

```
<entry id="6247">
 <word>Nódulo de Ranvier</word>
 <source>wikipedia</source>
 <url doc_date="6 de setembro de 2009." search_date="5 de Dezembro de 2009" type="html"
  >http://pt.wikipedia.org/wiki/N%C3%B3dulo_de_Ranvier</url>
 <paths>
  <path>Medicina\Especialidades médicas\Neurologia\Sistema nervoso</path>
  <path>Medicina\Especialidades médicas\Neurologia\Sistema nervoso\Neurociência\Sistema
  nervoso</path>
  <path>Medicina\Especialidades médicas\Neurologia\Sistema nervoso</path>
  <path>Medicina\Especialidades médicas\Neurologia\Sistema nervoso\Neurociência\Sistema
   nervoso</path>
 </paths>
 <definition>A não continualidade da bainha forma espaçamentos isentos de mielina, os nódulos de
  Ranvier. Isto facilita um movimento mais ágil do impulso que vai ocorrendo em saltos, já que o
  impulso só se propaga com a presença de mielina.</definition>
 <categorias>Sistema nervoso | !Esboços sobre histologia |</categorias>
 <synonyms>
  <synonym>
   <word>Telodendro</word>
  </synonym>
 </synonyms>
 <translation lang="en">
  <word>Nodes of Ranvier</word>
 </translation>
 <translation lang="fr">
  <word>Nœud de Ranvier</word>
 </translation>
 <translation lang="sp">
  <word>Nodo de Ranvier</word>
 </translation>
</entry>
```

Figure 5.8: Entrée *nódulo de Ranvier* après l'unification

Par ailleurs, les traductions sont différentes.
Dans ce cas, seul un spécialiste de la santé nous aidera à unifier ou pas ces entrées. Nous avons donc décidé de les sauvegarder dans leur totalité.

Table 5.21: Entrée *artéria cerebral anterior*

```
<entry id="397">
  <word>Artéria cerebral anterior</word>
  <definition>A artéria cerebral anterior é
  uma artéria da cabeça.</definition>
  <translation lang="en">
     <word>Anterior cerebral artery</word>
  </translation>
  <translation lang="fr">
     <word»Artère cérébrale antérieure</word>
  </translation>
  <translation lang="sp">
     <word»Arteria cerebral anterior</word>
  </translation>
</entry>
```

Table 5.22: Entrée *artéria cerebral média*

```
<entry id="398">
  <word>Artéria cerebral média</word>
  <definition>A artéria cerebral média é uma
  artéria da cabeça.</definition>
  <translation lang="en">
     <word>Middle cerebral artery</word>
  </translation>
  <translation lang="sp">
     <word»Arteria cerebral media</word>
  </translation>
</entry>
```

Quatrièmement, il existe des entrées du DPLP où seul le nom du domaine est différent. Par exemple, *loba* (id="1894") avec le domaine Vétérinaire et *sarcoma* (id="2775a") en Médecine. Nous ne pouvons pas unifier les deux entrées en joignant les deux domaines, car les utilisateurs ne peuvent pas utiliser un terme de médecine vétérinaire pour parler de pathologies humaines.

Cinquièmement, nous avons constaté des différences étymologiques dans le MP. Dans une entrée, on se réfère au latin et grec. Par exemple, *mico-* (id="10341") pour le grec et *fung-* (id="5634") pour le latin. Nous avons ici gardé les deux entrées.

Table 5.23: Entrée *fung-*

```
<entry id="5634">
  <word>Fung-</word>
  <definition>Pref. de origem latina que
  exprime relação com os fungos.</definition>
  <retated_word>
     <word id="10341">mico-</word>
  </related_word>
</entry>
```

Table 5.24: Entrée *mico-*

```
<entry id="10341">
  <word>Mico-</word>
  <definition>Pref. de origem grega que
  exprime relação com os fungos</definition>
  <synonyms>
    <synonym>
      <word>miceto-</word>
    </synonym>
  </synonyms>
  <retated_word>
     <word id="5634">fung-</word>
  </related_word>
</entry>
```

Et enfin, il existe dans le MP et le GMTM, des entrées qui ont des différences de traduction. Par exemple, *dermatológico* (id="470") et *epidérmico* (id="606"). Dans ces cas, nous préférons garder les deux entrées, car des traductions différentes peuvent révéler des sens différents.

5.4 Synthèse de l'unification intra-terminologique

La distance de Levenshtein appliquée aux entrées et aux définitions, a permis l'identification et la solution de nombreux problèmes d'unification. Maintenant que nous avons résolu ces difficultés d'unification par l'application de l'AO et le calcul de la distance de Levenshtein, nous passons à l'étape suivante : la fusion des sept sources extraites.

Chapitre 6

Création de l'UMLP

Après avoir résolu les problèmes d'unification intra-terminologique grâce à l'application de l'AO et de la distance de Levenshtein, nous allons maintenant passer à l'étape la plus intéressante de ce travail de recherche, qui est la création de l'UMLP.
Pour cela, un programme informatique nous a permis de fusionner les données des sept sources, afin de pouvoir traiter toutes les informations extraites et mieux corriger les problèmes d'unification encore présents à un niveau inter-terminologique. Finalement nous pourrons insérer les données corrigées et unifiées dans un nouveau XML.

6.1 Fusion des sources

Un programme informatique créé par le Professeur Rumen Romalyiski, permet de regrouper, dans un seul fichier, toutes les données contenues dans les étiquettes des XMLs. Plus précisément, ce système nous a donc donné la possibilité de fusionner toutes les entrées, directes et indirectes sans aucune répétition. Nous obtenons ainsi une longue liste de termes uniques, non répétés, comme nous pouvons le voir dans l'exemple du tableau 6.1.
Dans la première colonne, nous avons le terme non répété, et dans les colonnes suivantes, les différentes sources et leur respective position dans la source mentionnée, soit comme entrée directe, soit comme entrée indirecte. Dans ce dernier cas, il est toujours précisé, en abrégé, à quelle étiquette fait référence cette entrée indirecte (syn pour <synonyme>, rAdj pour <related_Adj>, rW pour <related_Word>...).

De cette forme, nous obtenons une liste de 91.715 termes uniques. Cependant, nous avons restreint le programme informatique afin que soient ignorés dans cette liste, les termes concernant les graphies antérieures à l'AO, les variantes brésiliennes, et les pluriels (comme nous le verrons dans la section 6.2.1). Ainsi nous obtenons une liste de 83.701 termes uniques. Nous en présentons un petit extrait dans l'annexe E.2.
Cependant, après avoir fusionné les termes de toutes nos bases de données, nous constatons que différents types de problèmes d'unification persistent.

6.2 Unification simple des données fusionnées

Nous avons restreint la liste précédente à près de 84.000 termes, afin de nous faciliter la tâche en ce qui concerne la prochaine étape, c'est-à-dire, le traitement des données de cette liste. En effet, le fait d'avoir retiré de cette liste toutes les entrées directes ou indirectes qui font référence aux graphies antérieures à l'AO, aux variantes brésiliennes et aux pluriels, nous évite

Table 6.1: Exemple de termes uniques du futur UMLP

Terme unique	Sources et respectives positions							
antiácido	MedPt id="923"	glo id="126"	Wpdia id="11056"	Wtnry id="1293"	Wtnry id="1293a"	Decs id="18678"		
antidepressivo	MedPt syn id="939"	glo id="139"	Wpdia id="10832"	Wtnry id="1413"	Wtnry id="1413a"	Decs id="18453"		
cateterismo	MedPt id="2136"	glo id="305"	Pri syn id="2907"	Wpdia id="29"	Wtnry id="1533"	Decs id="19896"		
convulsivo	MedPt rA id="3090"	MedPt rA id="3090a"	glo syn id="639"	glo syn id="642"	Wtnry rW id="1583b"	Decs id="18462"		
curetagem	MedPt id="3552"	glo id="444"	Pri id="851"	Pri id="851a"	Wpdia id="5029"	Decs id="20262"		
dieta	MedPt id="3963"	MedPt id="3963a"	Wpdia id="4224"	Wtnry id="315"	Wtnry id="316"	Decs id="19716"		
emoliente	MedPt id="4505"	MedPt rN id="4505"	glo id="576"	Pri id="1061"	Pri id="1061a"	Decs id="18443"		
imunização	MedPt id="6773"	glo id="926"	Pri id="1738"	Wpdia id="5346"	Wtnry rW id="1"	Decs id="19444"		
incidência	MedPt id="6814"	MedPt id="6814a"	MedPt id="6814b"	glo id="937"	Wpdia id="4256"	Decs id="19647"		
mortalidade	MedPt id="11592"	glo id="1178"	Wpdia id="4260"	Wpdia id="4260a"	Wpdia id="4260b"	Decs id="19635"		
órbita	MedPt id="9017"	MedPt id="9017a"	Pri id="2209"	Wpdia id="550"	Wtnry rW id="2435"	Decs id="199"		
risco	Decs id="19608"	Decs id="19608a"	Decs id="19608b"	Decs id="19608c"	Decs id="19608d"	Decs id="19608e"		
transplante	MedPt id="11904"	MedPt id="11904a"	glo syn id="1737"	Wpdia syn id="5578"	Wtnry id="76"	Decs id="18709"	Decs id="19955"	
vacinação	MedPt id="2959"	glo id="1780"	Wpdia id="5496"	Wtnry id="82"	Wtnry id="82a"	Decs id="19447"		

d'avoir à traiter ces problèmes déjà résolus dans l'analyse intra-terminologique de chaque ressource (cf. section 4.2.2 et chapitre 5).
Pour résoudre ces nombreux problèmes d'unification restants, nous utilisons de nouveau la distance de Levenshtein, avec une distance d'édition de un. Mais cette fois-ci, nous perfectionnons ce calcul en le combinant à une autre distance développée par le Professeur João Paulo Cordeiro (Cordeiro, 2011). Ces deux distances seront appliquées d'une part à tous les termes de cette liste, puis aux définitions des paires à réels problèmes.

6.2.1 Unification par entrée

Nous avons donc, d'une part, appliqué la distance de Levenshtein et nous avons obtenu 8.955 problèmes potentiels ; et d'autre part, avec la distance du Prof. João Paulo Cordeiro nous avons trouvé 7.026 problèmes potentiels.
Nous constatons alors, que nous avons une différence de près de 2.000 problèmes. De plus certains problèmes réels d'unification sont présents dans une des distances et pas dans l'autre, et réciproquement, comme le montre l'exemple du tableau 6.2.

Table 6.2: Exemple de problèmes réels présents dans une seule distance

Distance de Levenshtein	Ergotamina	MedPt id="4791"	Wpdia id="11161"	Decs id="11476"
	Ergotaminas	Decs id="11468"		
Distance de Cordeiro	Fobia	MedPt id="5467"	glo id="753"	Wpdia id="2824"
	Fobias	Decs syn id="21354"		

Dans le premier exemple, le problème d'unification *ergotamina/ergotaminas* n'apparaît que grâce à la distance de Levenshtein, et il n'apparaît pas dans les résultats de l'autre distance ; et réciproquement, avec le problème de *fobia/fobias*.
Pour nous permettre d'obtenir de meilleurs résultats, nous concaténons ces deux distances pour ne perdre aucune donnée fondamentale, et nous obtenons alors 11.101 problèmes potentiels d'unification.

Ensuite, une analyse manuelle a été faite pour retirer le bruit de cette longue liste. Nous entendons par "bruit", toutes les paires qui n'ont aucun rapport l'une avec l'autre (cf. exemple 1 du tableau 6.3). Malgré que tous les termes trouvés soient graphiquement parlant très proches du fait de la distance d'édition n'avoir été faite qu'à un, il faut toujours vérifier si la paire définit un même concept. Si c'est le cas, la possibilité d'avoir un problème d'unification persiste, si non, la paire peut donc être retirée de la liste. Cette vérification manuelle est donc très importante et permet d'alléger la longueur de la liste à traiter.
Également après réflexion, toutes les paires synonymes, related_word, related_adj ou abréviations d'une même entrée, ont été considérées comme bruit (cf. exemple 2 du tableau 6.3), car nous ne souhaitons pas modifier l'interprétation des différents auteurs des sept sources.

Après avoir retiré tout le bruit contenu dans la liste, nous obtenons alors une liste beaucoup moins étendue, avec seulement 2587 problèmes d'unification à résoudre.

Ensuite, pour un meilleur traitement des données, nous trions toutes ces paires par types de problèmes. Nous avons alors séparé les données en 10 types différents :

Table 6.3: Exemples de paires trouvées avec la Distance de Levenshtein, sur toutes les entrées directes et indirectes

Exemple 1	**Olheira**	Wpdia id="3072"		
		Olmeira	Decs syn id="23997"	
	Ódio	Wpdia id="9731"		
		Ópio	MedPt id="8999"	Decs id="17874"
	Medo	Wpdia id="5845"	Decs id="21713"	
		Meso	MedPt id="10203"	Pri id="1990"
		Mudo	MedPt id="11648"	Wtnry id="489"
Exemple 2	**hemafobia**	Wpdia syn id="2890"		
		hemofobia	Wpdia syn id="2890"	
	Canal Espinal	Decs syn id="220"		
		Canal Espinhal	Decs syn id="220"	
	hagioterapia	Wtnry rW id="1425"		
		higioterapia	Wtnry rW id="1425"	
	acidósico	MedPt rAdj id="399"		
		acidótico	MedPt rAdj id="399"	
	IL-23p19	Decs abbr id="16971"		
		IL23p19	Decs abbr id="16971"	

1. Problèmes de variantes PT/BR.

   Exemple : *planeamento familiar* (MP) / *planejamento familiar* (DeCS)

2. Problèmes liés à l'application de l'AO.

   Exemple : *olfação* (MP) / *olfacção* (DeCS)

3. Problèmes de double graphie acceptée par l'AO. [1]

   Exemple : *psicoléptico* (MP) / *psicolético* (GMTM)

4. Problèmes d'accents dans une même ressource.

   Exemple : *pênfigo* (vários) / *penfigo* (DeCS)

5. Problèmes de singulier/pluriel dans une même entrée.

   Exemple : *ouvido* (DeCS) / *ouvidos* (DeCS)

6. Problèmes de singulier/pluriel dans une même ressource.

   Exemple : *transplante de células* (DeCS) / *transplantes de células* (DeCS)

7. Problèmes d'entrée principale et de synonyme dans une même entrée.

   Exemple : *lisossomos* (DeCS) / *lisossomas* (DeCS)

8. Problèmes d'entrée principale et de synonyme dans une même ressource et/ou différentes ressources.

   Exemple : *espinal* (GMTM) / *espinhal* (MP)

9. Problèmes d'accents dans différentes ressources.

   Exemple : *diástase* (MP, DPLP) / *diastase* (DeCS)

10. Problèmes de singulier/pluriel dans différentes ressources.

    Exemple : *vitamina* (MP, Wikipédia) / *vitaminas* (DeCS)

Table 6.4: Nombre de paires par type de problèmes réels

Types de problèmes	Nombre de problèmes
1	27
2	117
3	268
4	23
5	334
6	68
7	922
8	303
9	104
10	318

Dans le tableau 6.4, nous pouvons voir combien de problèmes réels nous avons alors trouvés pour chaque type de problèmes cités ci-dessus.

Certaines de ces difficultés présentées ont été résolues simplement : comme par exemple, les problèmes d'AO restants, les problèmes de double graphie acceptée par l'AO et les problèmes des variantes luso-africaine/brésilenne. Ces trois types de cas ont été solutionnés de la même forme que spécifiée antérieurement. De plus nous avons également résolu facilement les problèmes de termes singuliers et pluriels dans une même entrée. Quand ces derniers apparaissent tous deux en tant que synonymes (cf. partie gauche du tableau 6.5), nous insérons alors l'étiquette <plural> à l'intérieur de l'étiquette <synonym>, comme nous pouvons le voir dans la partie droite de ce même tableau.

Table 6.5: Exemple de résolution des pluriels entre synonymes

```
<synonyms>
  <synonym>
    <word>osso</word>
  <synonym>
  <synonym>
    <word>ossos</word>
  <synonym>
<synonyms>
```

```
<synonyms>
  <synonym>
    <word>osso</word>
    <plural>ossos</plural>
  <synonym>
<synonyms>
```

Cependant, il peut également se produire qu'un des termes singulier ou pluriel soit le terme de l'entrée principale, et l'autre synonyme. Il est alors décidé ce qui suit :
- si le terme singulier est synonyme, alors nous le mettons comme terme principal, et insérons le terme pluriel dans l'étiquette <plural>, comme le montre le tableau 6.6.
- si l'entrée principale est déjà au singulier, alors cette dernière reste telle quelle, et le synonyme pluriel est déplacé dans l'étiquette <plural>.

1. Vérification faite dans le *Vocabulário da Língua Portuguesa* (VOLP) de l'Académie Brésilienne de Lettres (disponible gratuitement en ligne sur `http://www.academia.org.br/abl/cgi/cgilua.exe/sys/start.htm?sid=23`) et dans celui publié par l'éditeur Porto Editora (disponible gratuitement en ligne sur `http://www.infopedia.pt/default.jsp`) avec l'orientation scientifique de João Malaca Casteleiro, qui incorporent les Bases de l'Accord Orthographique de la Langue Portuguaise approuvé à Lisbonne le 12 octobre 1990.

Table 6.6: Exemple de résolution des pluriels entre l'entrée principale et les synonymes

```
<entry id="22866">
   <word>oncogenes</word>
   <definition>[. . . ]</definition>
   <synonyms>
      <synonym>
         <word>oncogene</word>
      <synonym>
   <synonyms>
</entry>
```

```
<entry id="22866">
   <word>oncogene</word>
   <plural>oncogenes</plural>
   <definition>[. . . ]</definition>
</entry>
```

Grâce à cette analyse par entrée, nous avons également découvert des problèmes d'erreurs orthographiques que nous avons corrigées dans les sources respectives. Par exemple, nous constatons dans le tableau 6.7, un oubli de la marque du pluriel. Cette erreur a été corrigée dans le XML correspondant, dans ce cas, dans le XML de Wikipédia.

Table 6.7: Exemple d'erreur orthographique trouvée avec la combinaison des deux distances

Síndrome das **Perna** Inquietas	Wpdia syn id="3285"
Síndrome das **pernas** inquietas	Decs id="7305"

6.2.2 Unification par définition

A partir des 2587 problèmes potentiels, nous avons fait un autre traitement de vérification en appliquant la distance d'édition aux définitions de ces paires, avec une distance de 10 mots. Cette analyse complémentaire a été faite seulement afin de comparer si avec la distance d'édition sur les définitions, nous trouvions les même résultats qu'avec ce même calcul appliqué précédemment aux entrées.

6.3 Unification complexe des données fusionnées

Après avoir classé par types les problèmes réels, et résolu les difficultés les plus simples restées en instance lors de l'analyse intra-terminologique, nous pouvons maintenant commencer à unifier les données restantes, c'est-à-dire les 1527 problèmes réels, étant donné que nous avons solutionné les dernières difficultés d'AO, de double graphie, de variante luso-africaine/brésilienne, de pluriels dans une même entrée et d'erreurs orthographiques.
Pour cela, nous utilisons deux méthodes : l'unification en fonction de l'origine de la source, et l'unification en fonction de son usage, de sa fréquence. Afin de pouvoir effectuer ce classement, nous décidons d'organiser les ressources, c'est-à-dire :

- Sont considérées comme sources à prédominante luso-africaine (annotée "PT"), le DPLP, le MP et le CHCB.
- Sont considérées comme sources à prédominante brésilienne (annotée "BR"), le DeCS, Wikipédia et *Wikcionário*.
- Est considéré neutre, la source GMTM (annotée "N").

6.3.1 Unification en fonction de l'origine de la source

Nous commençons par appliquer des règles que nous avons créées afin de résoudre certains des problèmes en fonction de l'origine de la source. Les règles utilisées sont :

Règle 1 :

> Si l'on a d'un côté au moins une source PT, et d'un autre côté au moins une source BR et/ou N, alors nous choisirons le terme où se trouve la source PT.

Règle 2 :

> Si l'on a d'un côté au moins une source PT "id" (c'est-à-dire l'entrée directe), et d'un autre côté au moins une source PT "autre que id" (c'est-à-dire une entrée indirecte), alors nous choisirons le terme où se trouve la source qui fait référence à l'entrée directe.

Pour comprendre notre analyse des données, prenons les exemples du tableau 6.8, dans lequel sont présentés différents types de problèmes.

Table 6.8: Exemples résolus grâce aux règles d'analyse en fonction de l'origine de la source

Exemple 1	Histerosalpingografia	Wpdia id="5044"	
	Histerossalpingografia	MedPt id="6572"	Decs id="20946"
Exemple 2	Epididimo	Decs id="386"	
	Epidídimo	MedPt id="4726"	Wtnry id="2206"
Exemple 3	Geriatra	Pri id="1514"	Wtnry id="106"
	Geriatro	Pri syn id="1514"	
Exemple 4	Língua Plicada	MedPt syn id="277"	
	Língua Plicata	Decs syn id="6403"	

Dans les deux premières illustrations, nous constatons que nous avons d'un côté une source BR et de l'autre côté une PT et une BR. Dans ces deux cas, nous appliquons la Règle 1, et le terme prédominant est alors celui où se trouve la source PT, c'est-à-dire : *Histerossalpingografia* et *Epididimo*.
Dans la troisième, nous avons un PT de chaque côté, mais l'un est une entrée directe, et l'autre une entrée indirecte. Dans ce cas, nous appliquons la Règle 2 et le terme prédominant est alors celui où se trouve le terme de l'entrée directe de la source PT, c'est-à-dire : *Geriatra*.
Et dans la dernière, nous avons d'un côté une source BR et de l'autre côté une PT, mais tous deux sont des synonymes. Même dans ce type de cas, c'est toujours le terme PT qui prédomine, donc nous gardons *Língua Plicada*.

Cependant, aucune donnée n'est effacée afin de conserver les droits d'auteurs. Lors de ces changements pour créer un dictionnaire à prédominante luso-africaine, sont créés de nouveaux XMLs avec la mention prédominante PT, où l'entrée directe est le terme à prédominante luso-africaine, et les autres sont toujours conservés et mis en synonyme.

6.3.2 Unification en fonction de son usage

Après avoir appliqué les règles qui nous montrent sans difficultés le terme à prédominante luso-africaine, nous allons maintenant utiliser des règles qui vont nous permettre d'analyser les

problèmes restants. Nous appelons cette analyse “par usage”, car chaque règle va nous mener à l’utilisation du moteur de recherche Google, afin de garder le terme qui a une fréquence plus grande en spécifiant pour chaque terme de la paire les domaines des sites .pt et .br. Les règles employées sont :

Règle 3 :

> Si l’on a d’un côté au moins une source PT “id” (c’est-à-dire l’entrée directe), et d’un autre côté au moins une autre source PT “id”, alors nous devrons analyser leur fréquence sur le moteur de recherche Google.

Règle 4 :

> Si l’on a d’un côté au moins une source BR “id” (c’est-à-dire l’entrée directe), et d’un autre côté au moins une source BR, alors nous devrons analyser leur fréquence sur le moteur de recherche Google.

Règle 5 :

> Si l’on a d’un côté une/des source(s) N et/ou BR, et d’un autre côté également une/des source(s) N et/ou BR, alors nous devrons analyser leur fréquence sur le moteur de recherche Google.

Après l’utilisation de ces règles, nous appliquons la formule (6.1) aux paires à problèmes :

$$\frac{HITS(x|PT)}{HITS(x'|PT)}, \frac{HITS(x|BR)}{HITS(x'|BR)} \quad (6.1)$$

Dans cette formule, ***x*** fait référence à l’alternative 1 e ***x’*** fait référence à l’alternative 2. Après avoir calculé chaque terme de l’équation (6.1), nous pouvons alors décider quel est le terme le plus utilisé dans la norme luso-africaine. Par exemple, si le premier terme de l’équation (6.1) est plus grand (en terme de fréquence) que le second terme, il sera alors plus probable que ***x*** soit la variante luso-africaine. Si c’est le contraire, ***x’*** sera alors le terme le plus probable de la norme luso-africaine. En cas d’égalité, l’indécision continuera. Cependant, nous devons toujours garder à l’esprit le grand problème auquel nous sommes confrontés avec cette méthode, sur la toile il existe beaucoup plus de documents de sites dans le domaine .br que dans le domaine .pt. Mais comme il s’agit de termes très spécifiques, cette formule est en général une très bonne solution pour découvrir le terme à prédominante luso-africaine.

Ces cinq règles nous ont permis de solutionner les manques qui persistaient. Cependant, lors de l’application des ces dernières, nous devons toujours faire attention aux paires où apparaissent plusieurs définitions dans une même source pour un seul mot. Car ces cas-là définissent en général des termes homographes, de définitions différentes, qui ne doivent absolument pas être unifiés.
Afin de faciliter le travail du linguiste, et de ne pas à avoir à utiliser toutes ces règles manuellement, un petit programme informatique qui applique automatiquement ces règles dans l’ordre défini ci-dessus, a été créé. Ce programme utilisé dans les paires à problèmes non résolus, permet l’application des règles et donne la solution finale, c’est-à-dire le meilleur terme à employer.

Maintenant nous pouvons encoder nos données nettes de difficultés afin d'obtenir un XML unique contenant toutes nos données.

6.4 Codage des données unifiées

En raison de la complexité de nos données, la création d'un XML unique n'est pas une tâche des plus simples, c'est pourquoi nous décidons de créer un XML général qui va chercher l'information dans chaque XML spécifique. Nous montrons dans la figure 6.1, deux petits exemples d'entrées du XML de l'UMLP afin de mieux comprendre sa création.

```
<umlp>
  <entry id="1">
    <word>Ablação</word>
    <origine_id>
      <prib id="12"/>
      <medpt id="296"/>
      <wikip id="4"/>
      <wikit id="6"/>
      <wikit id="7"/>
    </origine_id>
    <origine_syn>
      <glo id="698"/>
    </origine_syn>
  </entry>
  <entry id="2">
    <word>Aborto Clonal</word>
    <usage>Cuidado!!! Este termo médico pode pertencer unicamente à variante brasileira.</usage>
    <origine_syn>
      <decs id="22090"/>
    </origine_syn>
  </entry>
</umlp>
```

Figure 6.1: Exemple de deux entrées du XML unique de l'UMLP

Pour que ce XML soit valide en raison de ses nouvelles étiquettes, nous avons du créer une nouvelle DTD que nous pouvons consulter dans l'annexe B.2.

Dans ce premier exemple du XML de l'UMLP, nous voyons que l'entrée *Ablação* appelle plusieurs entrées directes des sources (<origine_id>), mais également un synonyme (<origine_syn>) qui est dans un autre dictionnaire non mentionné dans les entrées directes référencées. Donc le programme qui accompagne cet XML, va chercher toutes les informations (définitions, synonymes, antonymes, etc.) de chaque entrée citée, mais également celles du synonyme.
Dans le second exemple, comme cette entrée ne fait référence qu'à une seule source, qui est le DeCS, nous insérons alors une information dans l'étiquette <usage> : «ALERTE!!! Ce terme médical peut appartenir exclusivement à la variante brésilienne.» Cette information est insérée à chaque fois qu'une entrée de ce nouveau XML n'appelle que des sources à prédominante brésilienne.

6.5 Synthèse de la création de l'UMLP

Pour aboutir à une terminologie unifiée, nous avons dû fusionner les données des sept bases de données de départ, afin de pouvoir traiter toutes les entrées directes et indirectes extraites et de résoudre les problèmes d'unification. Finalement, la dernière phase a été celle du nouveau codage de toutes ces données.
Afin de rendre ce nouveau dictionnaire plus intéressant et dynamique, il a été créée une plate-forme interactive avec accès par mot de passe après enregistrement, pour que toute la communauté médicale de l'UBI puisse proposer des corrections, des modifications ou encore de nouvelles entrées. Si l'utilisateur insère une erreur (que ce soit dans une entrée existante ou bien dans une nouvelle entrée créée), elle sera évaluée et corrigée par le reste de la communauté qui sera informée des modifications faites.

Chapitre 7

Conclusions et Travail Futur

"La recherche est un processus sans fin dont on ne peut jamais dire comment il évoluera. L'imprévisible est dans la nature même de la science."

François Jacob

"La recherche doit avant tout être un jeu et un plaisir."

Pierre Joliot

7.1 Satisfaction des Objectifs

Dans ce travail de recherche, nous avons abordé le sujet de l'unification dans la terminologie médicale. La médecine étant un domaine de spécialité, sa terminologie contient un vocabulaire complexe et spécifique. C'est cette caractéristique qui rend très intéressant le travail d'unification.
En portugais, nous nous sommes rendu compte qu'il existe de nombreuses différences dans la terminologie médicale. Afin de comprendre ce langage et ses problèmes, nous avons fait un tour d'horizon sur les principaux systèmes et terminologies existantes au niveau médical. Nous constatons alors, que les difficultés d'unification de cette terminologie spécifique ne sont pas seulement sujettes à la langue portugaise. Effectivement, si nous prenons comme exemple l'anglais ou le français, ces langues ont elles aussi rencontré des problèmes d'unification de leur terminologie médicale. Mais, ses grands systèmes d'unification fonctionnent principalement par un travail manuel des données et par des traductions de terminologies existantes dans d'autres langues. C'est la raison pour laquelle notre objectif principal, lors de la construction semi-automatique de ce dictionnaire unifié, était d'éviter au maximum le travail manuel, ce qui est donc moins coûteux en main d'oeuvre, et de préparer un système d'actualisation facile et rapide.
C'est la raison pour laquelle nous n'avons utilisé que des ressources disponibles gratuitement sur la toile. Ainsi, les actualisations seront beaucoup plus simples et pourront être faites de forme automatique.

Nous constatons que ce travail de recherche a été interdisciplinaire. Au delà de la linguistique et surtout de la lexicographie, nous avons eu une très grande présence de l'informatique, du fait de vouloir limité au maximum le travail manuel et habituel des lexicographes, afin de créer un dictionnaire différent, aussi bien au niveau de sa construction que de son utilisation. Le fait de n'utiliser que des ressources présentes sur la toile et principalement savoir les synchroniser, montrent bien l'importance du domaine de l'informatique dans cette thèse. De plus, bien évidemment, la médecine a sa place principale dans cette étude en raison de son langage de spécialité à la fois très intéressant et complexe. Le travail en équipe est donc primordial, en raison du fait d'avoir choisi un sujet interdisciplinaire.

7.2 Travail Futur

Il est très difficile de terminer une thèse sur ce sujet, car si nous ne mettons pas de point final à notre recherche, nous aurons toujours de nouvelles informations à rajouter. Dans un domaine aussi intéressant que la médecine, nous pouvons penser à de nombreux travaux futurs, qui puissent améliorer notre dictionnaire.

Nous avons, par exemple, pensé à rajouter des exemples à chacune des entrées de l'UMLP, afin de les rendre plus accessibles et compréhensibles.
D'un point de vue plus technique, nous pouvons également, après la fusion des terminologies, effectuer une unification des entrées avec une distance d'édition d'au moins deux, afin de trouver plus de problèmes d'unification, et donc de rendre ce dictionnaire plus cohérent. Pour éviter le bruit, il faudrait créer un programme qui ne garderait que les distances d'édition à deux différence et élimerait celle à un, déjà effectuée dans le chapitre 6.
En terme d'unification, nous pouvons également employer d'autres méthodes d'unification, telles que l'unification par traduction. C'est-à-dire, en unifiant nos traductions, nous pouvons très bien obtenir des entrées synonymes qui seront par la suite unifiées ou reliées par numéro identificateur.
Lors de l'application des différentes règles que nous avons créées, particulièrement pour celles qui utilisent la fréquence sur le moteur de recherche *Google*, quand la fréquence d'usage des deux termes est très proche l'une de l'autre, nous pouvons faire une analyse de leurs traductions ou de leurs étymologies, afin de trouver le terme le plus approprié.
Également au niveau des règles, quand nous avons une confrontation de deux termes luso-africain, nous pouvons approfondir l'analyse du sujet afin de découvrir à quelle école doctorale appartiennent les termes en question.
De plus nous pouvons aussi mettre en place le Metathésaurus grâce aux structures hiérarchiques déjà extraites lors de l'analyse de chacune des sources.
Un autre travail très important, est celui de la programmation du système qui permettra l'actualisation (semi-)automatique de tout notre dictionnaire, à chaque fois qu'une des bases de données est modifiée.

Et pour finir, le plus important est de mettre en ligne une application pour le dictionnaire médical évolutif et collaboratif, ainsi que de le rendre accessible sur *smartphone*, car élèves, médecins et spécialistes de la médecine n'ont pas toujours à porté de main un ordinateur.

Bibliographie

Academia das Ciências de Lisboa (2001). *Dicionário da Língua Portuguesa Contemporânea*, vol. 1 e 2. Verbo, Lisboa, Portugal. 2

Alj, L., Benkirane, R. & Soulaymani-Bencheikh, R. (2005). Terminologie des Effets Indésirables. Disponible sur `http://www.who.int/entity/medicines/areas/quality_safety/safety_efficacy/trainingcourses/1TeminologieEI.pdf`. 16

Assembleia da Républica (1991). Resolução da Assembleia da Républica n.º 26/91 : Aprova, para ratificação, o Acordo Ortográfica da Língua Portuguesa. Diário da República n.º 193, SÉRIE I-A, 23 de Agosto. Disponível em : http://dre.pt/pdfgratis/1991/08/193A00.pdf. 55

Bodenreider, O. & McCray, A.T. (1998). From French vocabulary to the Unified Medical Language System : A preliminary study. *Medinfo 1998*, 670–674. xi, 9

Bodenreider, O., Smith, B., Kumar, A. & Burgun, A. (2007). Investigating subsumption in SNOMED CT : An exploration into large description logic-based biomedical terminologies. *Artificial Intelligence in Medicine*, **39**, 183–195. 12

Brown, E.G. (2004). Using MedDRA : Implications for Risk Management. *Drug Safety*, **27**, 591–602. 15

Browne, A.C., McCray, A.T. & Srinivasan, S. (2000). The Specialist Lexicon. Research Report, Lister Hill National Center for Biomedical Communications - National Library of Medecine. 7

Casteleiro, J.M. & Correia, P.D. (2008). *Atual - O novo acordo ortográfico*. Texto Editores, Lisboa, Portugal. 55

Charlet, J., Bachimont, B. & Jaulent, M.C. (2006). Building medical ontologies by terminology extraction from texts : An experiment for the intensive care units. *Computers in Biology and Medicine*, **36**, 857 – 870, special Issue on Medical Ontologies. 7

Contente, M. (2004). Terminocriatividade Sinónima e Equivalência Interlinguística em Medicina. Doutoramento em Linguística, especialidade de Lexicologia. 17

Contente, M. & Magalhães, J. (1997). Dictionnaire multilingue de médecine – vers une dictionnaire d'apprentissage. **42**, 114–120. 17

Cordeiro, J.P. (2011). Rule induction for sentence reduction. Doutoramento em Engenharia Informática. 75

Costa, R. (2010). *Escul@pio : Uma plataforma colaborativa de acesso ao UMLP*. Master's thesis, Universidade da Beira Interior. xvii, 18, 28

Fellbaum, C. (1998). *WordNet : An Electronic Lexical Database*. The MIT Press, illustrated edition edn. 16

Fellbaum, C., Hahn, U. & Smith, B. (2006). Towards new information resources for public health : from wordnet to medicalwordnet. *J. of Biomedical Informatics*, **39**, 321–332. 17

Garcia, M.L.A. (1994). A terminologia médica. In *IV^o Simpósio Ibero-Americano de Terminologia RITerm – Terminologia e Desarrollo*, 8. 17

Garcia, M.L.A. (1997). Projet de dictionnaire interactif multilingue de termes médicaux. **42**, 110–113. 17

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, **438**, 900–901. 33

Goldberg, K.H. & Éric Jacoboni (2009). *Manuel de prise en main de XML*. Le Programmeur, Pearson Education France. 36

Humphreys, B.L., Lindberg, D.A., Schoolman, H.M. & Barnett, G.O. (1998). The Unified Medical Language System : an informatics research collaboration. *J Am Med Inform Assoc*, **5**, 1–11. 1

IHTSDO (2008). SNOMED CT User Guide – July 2008. International Release, The International Health Terminology Standards Development Organisation. 14

IHTSDO (2009). SNOMED CT User Guide – July 2009. International Release, The International Health Terminology Standards Development Organisation. 12, 13, 14

Janssen, M. (2008). *Vocabulário em Mudança*. Editorial Caminho, Lisboa, Portugal. 55

Kobayashi, M. & Takeda, K. (2000). Information retrieval on the web. *ACM Computing Surveys*, **32**, 144–173. 26

Lamberts, H., Meads, S. & Wood, M. (1984). International Classification of Primary Care : a multipurpose classification. Presentation at the International Epidemiological Association Meeting. 14

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, **10**, 707–710. 59

Lindberg, D.A.B. & Humphreys, B.L. (1990). The UMLS knowledge sources : tools for building better user interfaces. In *Proceedings of the 14th Annual Symposium on Computer Applications in Medical Care*, 121–125. 7

Manuila, L., Manuila, A., Lewalle, P. & Nicoulin, M. (2000). *Dicionário Médico*. Climepsi Editores, Lisboa, Portugal, tradução José Nunes de Almeida. Consultoria científica e revisão científica Dr. João Alves Falcato. 21, 27

Marrafa, P. (2001). WordNet do Português : uma base de dados de conhecimento linguístico. Lisboa : Instituto Camões. 16

MedDRA MSSO (2011). Welcome to MedDRA and the MSSO. Http ://www.meddramsso.com/index.asp. 15

National Library of Medicine (2009). *UMLS® Reference Manual [Internet]*. Disponível em http://www.ncbi.nlm.nih.gov/books/NBK9679. 7

National Library of Medicine (2010). Medical Subject Headings. Disponível em http://www.nlm.nih.gov/mesh/meshhome.html. 5

National Library of Medicine (2011). MEDLINE Fact Sheet. Disponível em http://www.nlm.nih.gov/pubs/factsheets/medline.html. 6

Nicoulin, M., Manuila, L., Lewalle, P. & Papo, T. (2004). *Dictionnaire Médical*. Masson, Paris, France, `http://books.google.com/books/about/Dictionnaire_médical_Manuila.html?id=89V53WqN5CgC`. 40

Okkes, I.M., Jamoulle, M., Lamberts, H. & Bentzen, N. (2000). ICPC-2-E. The electronic version of ICPC-2. Differences with the printed version and the consequences. *Fam Pract 2000*, **17**, 101–106. 14

Rector, A.L. & Nowlan, W.A. (1994). The GALEN project. *Computer methods and programs in biomedicine*, **45**, 75–78. 11

Rector, A.L., Bechhofer, S., Goble, C., Horrocks, I., Nowlan, W. & Solomon, D. (1997). The GRAIL concept modeling language for medical terminology. *Artificial Intelligence in Medicine*, **9**, 139–171. 11

Ross, T. (2007). *XML-Managing Data Exchange*. Global Media. 35

Schmidtke, S.A.J. (2007). Networking for rare diseases : a necessity for Europe. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz*, **50**, 1477–1483, doi :10.1007/s00103-007-0381-9. PMID 18026888. 12

Schulze-Kremer, S., Smith, B. & Kumar, A. (2004). Revising the UMLS Semantic Network. In *Proceedings of Medinfo 2004*, 1700. 7

Shortliffe, E.H. & Buchanan, B.G. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences 23 (3-4)*. 5

Smith, B. & Fellbaum, C. (2004). Medical wordnet : a new methodology for the construction and validation of information resources for consumer health. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, 371–382, Association for Computational Linguistics, Stroudsburg, PA, USA. 17

Stuart, N., Schopen, M., Schulman, J.L. & Arluk, N. (2000). An Interlingual Database of MeSH Translations. In *Proceedings of the 8th International Conference on Medical Librarianship*. 6

Taber (2000). *Dicionário Médico Enciclopédico-Taber*. Editora Manole, tradução José Nunes de Almeida. Consultoria científica e revisão científica Dr. João Alves Falcato. 21

Tardelli, A. (2007). DeCS/MeSH Description, Uses, Services, Updating. In *Proceedings of the Global Health Library Workshop*. 1, 17

Tardelli, A. (2009). DeCS/MeSH – Descritores em Ciências da Saúde – Descrição, Usos, Serviços e Atualização. 4a Reunião dos Comitês Consultivo e Executivo da Biblioteca Virtual em Saúde Temática em Integralidade, FIOCRUZ/ICICT. 7

Uppsala Monitoring Centre (2011). Welcome to WHO-ART - WHO Adverse Reaction Terminology. `http://www.umc-products.com/DynPage.aspx?id=73589&mn1=1107&mn2=1664`. 16

WHO Collaborating Centre for International Drug Monitoring (2005). The WHO Adverse Reaction Terminology – WHO-ART. Guide, WHO Collaborating Centre for International Drug Monitoring. 16

Widdows, D., Peters, S., Cederberg, S., Chan, C.K., Steffen, D. & Buitelaar, P. (2003). Unsupervised monolingual and bilingual word-sense disambiguation of medical documents using UMLS. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, 9–16, Association for Computational Linguistics, Morristown, NJ, USA. 1

Wikipédia (2010). Robot d'indexation — Wikipédia, l'encyclopédie libre. [En ligne ; Page disponible le 25-juin-2010]. 26

WONCA International Classification Committee (1998). *ICPC-2 : International Classification of Primary care, second edition*. ISBN : 978-0192628022, U.K. : Oxford University Press. 14

World Health Organization (2005). International Statistical Classification of Diseases and Related Health Problems (ICD-10) – Volume 2. Instruction manual – 2nd edition, Tenth Revision. 10

Zweigenbaum, P. (2004). L'UMLS entre langue et ontologie : une approche pragmatique dans le domaine médical. *Revue d'Intelligence Artificielle*, **18**, 111–137. 7

Zweigenbaum, P. (2005). Fusion de terminologies : le projet UMLS. *IFIP/IEEE International Symposium on Integrated Network Management (IX IM 2005). Nice, France*. xvii, 9

Zweigenbaum, P., Bachimont, B., Bouaud, J., Charlet, J. & Boisvieux, J.F. (1996). Le rôle du lexique sémantique et de l'ontologie dans le traitement automatique de la langue médicale. In P.L. Beux & A. Burgun, eds., *Actes du Colloque CRISTAL'S*, 9–16. 11

Zweigenbaum, P., Baud, R., Burgun, A., Namer, F., Jarrousse, E., Grabar, N., Ruch, P., Duff, F.L., Thirion, B. & Darmoni, S. (2003). UMLF : a Unified Medical Lexicon for French. *AMIA Annual Symposium Proceedings*, 1062. xi, 1, 10

Annexe A

Problèmes types trouvés à l'UBI par les professionnels de la santé

Liste de termes créées par le Professeur José Martinez de Oliveira le 02 Mai 2007

Sigles des Écoles utilisés : Br pour Brésilienne ; C pour Coimbra ; L pour Lisbone ; P pour Porto.

Problèmes de genre

O enzima ou A enzima
O grama ou A grama

Problèmes d'accentuation

Amílase (P) ou Amilase (C)
Anémia (L) ou Anemia (C,P)
Atresia ou Atrésia
Carácter ou Caracter ou Caractere => plural Caracteres
Clítoris ou Clitóris ou Clitóride
Gâmeta (P) ou Gameta (Br)
Leucémia (L) ou Leucemia (C,P)
Pólipo (P) ou Polipo (L)
Tacrolimus ou Tacrólimus
Triptófano ou Triptofano
Trofozóito (P) ou Trofozoíto (L)

Problèmes de préférence

Ansas ou Alsas (Br)
Atrofogénico ou Atrofiante
Biopsar ou Biopsiar
Bolbo ou Bulbo
Canais ou Ductos
Centrossomo ou Centrosoma ou Centrossoma
Cisto ou Quisto
Citocina ou Citoquina
Climatérico ou Climático (de climatério e de clima)
Clitorectomia (clitor/o + ectomia) ou Clitoridectomia (clitorid/o+ectomia)
Cólica ou Colónica (de cólon)
Consulentes, Utentes, Usuárias, Clientes

Cromossomo ou Cromosomo ou Cromossoma
Encefalina ou enquefalina
Enervação ou Inervação (distribution par nerfs)
Equipe ou Eequipa
Espinocelular ou espinhocelular (*células do estrato espinhoso de Malpighi*)
Exsudado, Exsudato, Exudado ou Exudato
Fenda ou Rima (Br)
Fímbrias ou Franjas (tubáricas)
Fórnix ou Fórnice (B) ou Fundo-de-saco (P) ou Betesga (C)
Freio ou Frénulo ou Frenulum
Glicose ou Glucose
Hematometra, Hematométrio ou hematometria
Hidrometra, Hidrométrio ou hidrometria
Leucina ou Leuquina
Lípido ou Lipídeo
Maduro ou maturo
Mamologia, Mastologia ou Senologia
Ocitocina ou Oxitocina
Oligoâmnio ou Oligohidrâmnio ou Oligohidrâmnios (exagération de liquide amniotique)
Oócito ou Ovócito
Oogénese ou Ovogénese
Oportunistas ou Oportunísticas (maladies)
Ostium ou Óstio ou Orificio (C) ou Buraco (P) ou Foramen
Pavimento (C) ou Soalho (P) ou Assoalho (Br)
Prolapsado ou Prolabado
Sacros ou Sacrados ou Sagrados ou Sacrais
A Sindroma, A Síndrome, O Sindromo
Tabagismo ou Tabaquismo
Tiróide ou Tireóide
Toma ou Tomada (de medicamentos)
Trompas ou Tubas (Br)
Tubar, Tubário ou Tubárico (referente a trompa)
Tuboclasia ou Laqueação de trompas ou Laqueadura (B) das trompas
Verrugosa ou Verrucosa (en forme de verrues)

Problèmes de sens

Aborto (L) ou Abortamento (P-onde aborto é o produto de abortamento)
Cólica de cólon ou de dor ou ambos
Constipação ou Obstipação
Fisiometria ou Fisiometra : fetidez do liquido amniótico
Iterativa
Lobos e Lóbulos
Polipóide (ant : séssil) => pólipo séssil aceitável?
Semiologia (do grego semeion, sinal, e logos, tratado) : parte da Medicina que estuda os sintomas e os sinais das doenças.
Semiótica (do grego semeiotike) : embora tendo o mesmo significado etimológico semiologia

aplica-se mais especificamente à metodologia da colheita e ao processo de sistematização dos sintomas e sinais clínicos.
Significado, Significação, Significância
Tumefacção e Tumoração
Ulceração e Úlcera

Problèmes de neologismos

Cosmecêutica : cosméticos com uso terapêutico
Nutracêuticos : nutrientes com uso terapêutico

Annexe B

Les DTDs

B.1 DTD correspondant aux sept XMLs extraits

```
1.  <!ELEMENT esculapio (dico,entry+)>

2.  <!ELEMENT dico (#PCDATA)>

3.  <!ELEMENT entry (word,source,trusted_source*,url,etymology?,domain?,paths?,category?,
                     number?,gender?,plural?,definition?,image?,categorias?,synonyms?,
                     antonym?,related_adj?, related_nouns?,related_verb?,related_word?,
                     abbreviations?,symbols?,compound?,translation*)>
4.  <!ATTLIST entry id CDATA #REQUIRED>

5.  <!ELEMENT word (#PCDATA)>
6.  <!ATTLIST word id CDATA #IMPLIED>
7.  <!ATTLIST word type CDATA #IMPLIED>

8.  <!ELEMENT source (#PCDATA)>

9.  <!ELEMENT trusted_source (#PCDATA)>

10. <!ELEMENT url (#PCDATA)>
11. <!ATTLIST url doc_date CDATA #IMPLIED>
12. <!ATTLIST url search_date CDATA #IMPLIED>
13. <!ATTLIST url type CDATA #IMPLIED>

14. <!ELEMENT etymology (#PCDATA)>

15. <!ELEMENT domain (word+)>

16. <!ELEMENT paths (path+)>
17. <!ELEMENT path (#PCDATA)>

18. <!ELEMENT category (#PCDATA)>

19. <!ELEMENT number (#PCDATA)>

20. <!ELEMENT gender (#PCDATA)>

21. <!ELEMENT plural (#PCDATA)>
```

```
22.  < !ELEMENT definition (#PCDATA)>

23.  < !ELEMENT image (legend*,url)>

24.  < !ELEMENT legend (#PCDATA)>

25.  < !ELEMENT categorias (#PCDATA)>

26.  < !ELEMENT synonyms (synonym+)>
27.  < !ELEMENT synonym (word,etymology*,category*,number*,gender*,usage*,abbreviation*)>
28.  < !ATTLIST synonym id CDATA #IMPLIED>

29.  < !ELEMENT antonym (word+)>

30.  < !ELEMENT related_adj (word+,usage*,gender*,translation*)>

31.  < !ELEMENT related_nouns (related_noun+)>
32.  < !ELEMENT related_noun (word,gender*,number*,usage*)>
33.  < !ATTLIST related_noun id CDATA #IMPLIED>

34.  < !ELEMENT related_verb (word+)>

35.  < !ELEMENT related_word (word+,usage*)>

36.  < !ELEMENT usage (#PCDATA)>

37.  < !ELEMENT abbreviations (abbreviation+)>
38.  < !ELEMENT abbreviation (word+,usage*)>
39.  < !ATTLIST abbreviation id CDATA #IMPLIED>

40.  < !ELEMENT symbols (symbol+)>
41.  < !ELEMENT symbol (#PCDATA)>

42.  < !ELEMENT compound (word+)>

43.  < !ATTLIST translation lang (en|us|fr|sp) #REQUIRED>
44.  < !ELEMENT translation (word+)>
```

B.2 DTD finale correspondant au XML unifié

```
1.  <!ELEMENT umlp (entry+)>

2.  <!ELEMENT entry (word,usage*,origine_id?,origine_syn?,origine_RA?,origine_RN?,
                     origine_RV?,origine_RW?,origine_abbr?)>

3.  <!ELEMENT word (#PCDATA)>

4.  <!ELEMENT usage (#PCDATA)>

5.  <!ELEMENT origine_id (prib*,medpt*,glo*,chcb*,wikip*,wikt*,decs*)>

6.  <!ELEMENT origine_syn (prib*,medpt*,glo*,chcb*,wikip*,wikt*,decs*)>

7.  <!ELEMENT origine_RA (prib*,medpt*,glo*,chcb*,wikip*,wikt*,decs*)>

8.  <!ELEMENT origine_RN (prib*,medpt*,glo*,chcb*,wikip*,wikt*,decs*)>

9.  <!ELEMENT origine_RV (prib*,medpt*,glo*,chcb*,wikip*,wikt*,decs*)>

10. <!ELEMENT origine_RW (prib*,medpt*,glo*,chcb*,wikip*,wikt*,decs*)>

11. <!ELEMENT origine_abbr (prib*,medpt*,glo*,chcb*,wikip*,wikt*,decs*)>

12. <!ELEMENT prib EMPTY>
13. <!ATTLIST prib id CDATA #REQUIRED>

14. <!ELEMENT medpt EMPTY>
15. <!ATTLIST medpt id CDATA #REQUIRED>

16. <!ELEMENT glo EMPTY>
17. <!ATTLIST glo id CDATA #REQUIRED>

18. <!ELEMENT chcb EMPTY>
19. <!ATTLIST chcb id CDATA #REQUIRED>

20. <!ELEMENT wikip EMPTY>
21. <!ATTLIST wikip id CDATA #REQUIRED>

22. <!ELEMENT wikt EMPTY>
23. <!ATTLIST wikt id CDATA #REQUIRED>

24. <!ELEMENT decs EMPTY>
25. <!ATTLIST decs id CDATA #REQUIRED>
```

Annexe C

Quelques résultats trouvés avec la Distance de Levenshtein

C.1 Problèmes types trouvés grâce à l'application de la Distance de Levenshtein

Cefaleia	MdePt id="2180"	glo id="311"	Wpdia id="3006"	Wtnry id="620a"
Cefaléia	Decs id="7204"			
bicos de papagaio	Wpdia syn id="2247"			
bicos-de-papagaio	WPri id="461"			
Ultra-sonografia	MdePt id="10771"	Wpdia syn id="10450"	Wtnry id="81"	
Ultrassonografia	Wpdia id="10450"	Wtnry id="1928"	Decs id="21002"	
Tónus muscular	MdePt id="11544"	Wpdia id="12213"		
Tônus muscular	Wpdia syn id="12213"	Decs syn id="22197"		
contração	Wtnry id="239"	Wtnry id="240"		
Contracção	MdePt id="3061"	glo id="407"	Wtnry rW id="268"	

C.2 Exemples de La Distance de Levenshtein appliquée aux entrées directes du DPLP

prião ID = "2511"	**príon** ID = "2515" ; pulsão ID = "2586" ; purina ID = "2593" ; pícrico ID = "2406" ;pínico ID = "2413" ; pólipo ID = "2461" ; pólipo ID = "2461a" ; recão ID = "2641" ;região ID = "2665" ; ritmo ID = "2724" ; seio ID = "2794" ; tifo ID = "3050" ; timo ID = "3052" ; tipo ID = "3057" ; tipo ID = "3057a" ; torção ID ="3085" ; trago ID = "3103" ; trismo ID = "3143" ; ustão ID = "3215" ; ustão ID = "3215a" ;vergão ID = "3274" ; vergão ID = "3274a" ; viro ID = "3293" ; visão ID = "3296" ; ácido ID = "69" ; ácido ID = "69a" ; ácido ID = "69b" ; ácido ID = "69c" ; ácido ID = "69d" ; ácino ID = "74" ; ático ID = "376" ; ázigo ID = "406" ; ílio ID = "1728" ; íris ID = "1806" ; íris ID = "1806a" ; ítrio ID = "1827" ; óxido ID = "2255" ; úraco ID = "3187"
acatastático ID = "33"	**acatástico** ID = "34" ; catagmático ID = "594"
	cocus ID = "685" ; cofose ID = "689" ; coifa ID = "690" ; coma ID = "726" ; conca ID = "747" ; cone ID = "758" ; corda ID = "795" ; corpo ID = "805" ; cálcio ID = "525" ; cérico ID = "626" ; cíato ID = "635" ; cínico ID = "644" ; cólico ID = "708" ; cólon ID = "714" ; cúneo ID = "849" ;

	dorso ID = "988" ; enojo ID = "1092" ; feto ID = "1345" ; feto ID = "1345a" ; figo ID = "1358" ; filo ID = "1360" ; fixo ID = "1375" ; foco ID = "1409" ; foice ID = "1410" ; gota ID = "1545" ; hilo ID = "1630" ; icto ID = "1720" ; iodo ID = "1801" ; ião ID = "1717" ; loba ID = "1894" ; lobo ID = "1896" ; local ID = "1900" ; meso ID = "1990" ; mola ID = "2048" ; motor ID = "2069" ; motor ID = "2069a" ; mucol ID = "2070" ; mucro ID = "2071" ; nocebo ID = "2124" ; noma ID = "2125" ; ose ID = "2217" ; ovo ID = "2249" ; pano ID = "2279" ; papo ID = "2283" ; porro ID = "2478" ; recto ID = "2656" ; recão ID = "2641" ; saco ID = "2746" ; sacro ID = "2749" ; seio ID = "2794" ; soda ID = "2897" ; soda ID = "2897a" ; soro ID = "2908" ; suco ID = "2933" ; tac ID = "2979" ; talo ID = "2983" ; timo ID = "3052" ; tipo ID = "3057" ; tipo ID = "3057a" ; tofo ID = "3067" ; tono ID = "3075" ; tronco ID = "3154" ; tronco ID = "3154a" ; varo ID = "3252" ; viro ID = "3293" ; volvo ID = "3304" ; zona ID = "3339" ; ácido ID = "69" ; ácido ID = "69b" ; ácido ID = "69c" ; ácido ID = "69d" ; ácino ID = "74" ; ático ID = "376" ; ácido ID = "69a" ; ílio ID = "1728" ; úraco ID = "3187" ; cru ID = "845" ; pica ID = "2402" ; tifo ID = "3050"
viro ID = "3293"	visão ID = "3296" ; vitelo ID = "3300" ; volvo ID = "3304" ; **vírus** ID = "3295" ; vórmio ID = "3307" ; zerbo ID = "3330" ; zirbo ID = "3334" ; ácido ID = "69" ; ácido ID = "69a" ; ácido ID = "69b" ; ácido ID = "69c" ; ácido ID = "69d" ; ácino ID = "74" ; ático ID = "376" ; ázigo ID = "406" ; ílio ID = "1728" ; ítrio ID = "1827" ; óvulo ID = "2250" ; óvulo ID = "2250a" ; óxido ID = "2255" ; útero ID = "3218"
acinésico ID = "72"	**acinético** ID = "73" ; adinâmico ID = "107" ; amiélico ID = "203" ; amiélico ID = "203a" ;cinâmico ID = "641"
anosteose ID = "298"	**anostose** ID = "299" ; anástase ID = "241" ; diostose ID = "948" ; exostose ID = "1273"
blastocele ID = "474"	**blastocélio** ID = "475" ; galactocele ID = "1474"
osido ID = "2220"	ossinho ID = "2222" ; **osídeo** ID = "2219" ; ovo ID = "2249" ; oxácido ID = "2251" ; oócito ID = "2196" ; prião ID = "2511" ; rosado ID = "2733" ; saco ID = "2746" ; seio ID = "2794" ; soda ID = "2897" ; soda ID = "2897a" ; soro ID = "2908" ; suco ID = "2933" ; sódio ID = "2898" ; tecido ID = "2999" ; tifo ID = "3050" ; timo ID = "3052" ; tipo ID = "3057" ; tipo ID = "3057a" ; ustão ID = "3215" ; ustão ID = "3215a" ; viro ID = "3293" ; ácido ID = "69" ; ácido ID = "69a" ; ácido ID = "69b" ; ácido ID = "69c" ; ácido ID = "69d" ; ácino ID = "74" ; ático ID = "376" ; ázigo ID = "406" ; ílio ID = "1728" ; óxido ID = "2255"
salter ID = "2762"	**saltério** ID = "2763" ; scanner ID = "2786" ; sortes ID = "2909" ;

	uréter ID = "3194" ; zóster ID = "3345" ; áster ID = "369" ; éster ID = "1221" ; éter ID = "1257"
zooparasita ID = "3341"	**zooparasito** ID = "3342"
alcalosa ID = "152"	**alcalose** ID = "153" ; calculose ID = "527" ; calculoso ID = "528"
aplotomia ID = "318"	artrotomia ID = "357" ; biotomia ID = "469" ; colostomia ID = "716" ; colotomia ID = "717" ; **haplotomia** ID = "1569" ; laparotomia ID = "1855" ; litotomia ID = "1890" ; rinotomia ID = "2722" ; rizotomia ID = "2725" ; tenotomia ID = "3021" ; tiflotomia ID = "3049"
acatesia ID = "35"	**acatisia** ID = "37" ; acidemia ID = "65" ; acrisia ID = "90" ; acrisia ID = "90a" ; alestesia ID = "165" ; analgesia ID = "237" ; anataxia ID = "244" ; astasia ID = "364" ; atelia ID = "374" ; atelia ID = "374a" ; atresia ID = "386" ; ectasia ID = "1015" ; estesia ID = "1231"
acetolado ID = "55"	acetolar ID = "56" ; **acetolato** ID = "57" ; acetulado ID = "60" ; enolado ID = "1094" ; oleolado ID = "2166"
formigueiro ID = "1426"	**formiguilho** ID = "1427"
geriatra ID = "1514"	geriatria ID = "1515" ; **geriatro** ID = "1517" ; pediatra ID = "2335" ; pediatria ID = "2336" ; pediatro ID = "2338" ; zoiatra ID = "3336"
pediatra ID = "2335"	pediatria ID = "2336" ; **pediatro** ID = "2338" ; pedra ID = "2342" ; pequiagra ID = "2355" ; zoiatra ID = "3336"
poliestireno ID = "2455"	polietileno ID = "2456" ; **polistireno** ID = "2465"
xénio ID = "3319"	**xénon** ID = "3320" ; ílio ID = "1728" ; ínion ID = "1766" ; ítrio ID = "1827"
torcicolo ID = "3086"	torcilhão ID = "3087" ; **torticolo** ID = "3093"
calaza ID = "522a"	calazar ID = "523" ; **calázio** ID = "524" ; cana ID = "533" ; canal ID = "534" ; cava ID = "602" ; célula ID = "610" ; célula ID = "610a" ; célula ID = "610b" ; galapo ID = "1480" ; malato ID = "1923" ; malha ID = "1925" ; nagana ID = "2083" ; palma ID = "2264" ; pelada ID = "2343" ; placa ID = "2426" ; tala ID = "2980" ; talha ID = "2982" ; valva ID = "3240"
entubação ID = "1103"	indução ID = "1751" ; **intubação** ID = "1795" ; maturação ID = "1940" ; nutação ID = "2136" ; obturação ID = "2150" ; saturação ID = "2784" ; titulação ID = "3063"
entubar ID = "1104"	entubação ID = "1103" ; escutar ID = "1159" ; **intubar** ID = "1796" ; rotular ID = "2737" ; rotular ID = "2737a" ; titular ID = "3064"
endozoário ID = "1083"	**entozoário** ID = "1101" ; fitozoário ID = "1370"
fosfato ID = "1432"	**fosfito** ID = "1433" ; fosfático ID = "1431" ; fósforo ID = "1438" ; oleato ID = "2164" ; opiato ID = "2202" ; opiato ID = "2202a" ; rosado ID = "2733" ; sulfato ID = "2942"
ácido ID = "69"	ácino ID = "74" ; ático ID = "376" ; áxis ID = "405" ; ázigo ID = "406" ; ílio ID = "1728" ; óxido ID = "2255"
ácido ID = "69a"	ácino ID = "74" ; ático ID = "376" ; áxis ID = "405" ;

	ázigo ID = "406" ; ílio ID = "1728" ; óxido ID = "2255"
ácido ID = "69b"	ácino ID = "74" ; ático ID = "376" ; áxis ID = "405" ; ázigo ID = "406" ; ílio ID = "1728" ; óxido ID = "2255"
ácido ID = "69c"	ácino ID = "74" ; ático ID = "376" ; áxis ID = "405" ; ázigo ID = "406" ; ílio ID = "1728" ; óxido ID = "2255"
ácido ID = "69d"	ácino ID = "74" ; ático ID = "376" ; áxis ID = "405" ; ázigo ID = "406" ; ílio ID = "1728" ; óxido ID = "2255"
acme ID = "77"	**acne** ID = "78" ; anel ID = "262" ; aná ID = "225" ; argo ID = "343" ; aura ID = "391" ; base ID = "432" ; base ID = "432a" ; c ID = "513" ; coma ID = "726" ; cone ID = "758" ; cru ID = "845" ; e ID = "1002" ; facote ID = "1298" ; fase ID = "1317" ; fene ID = "1330" ; gema ID = "1505" ; gémeo ID = "1507" ; icto ID = "1720" ; m ID = "1910" ; má ID = "1911" ; noma ID = "2125" ; ose ID = "2217" ; palma ID = "2264" ; raque ID = "2631" ; rede ID = "2657" ; saco ID = "2746" ; sacro ID = "2749" ; sémen ID = "2800" ; tac ID = "2979" ; taxe ID = "2996" ; timo ID = "3052" ; vómer ID = "3305"
acracia ID = "85"	acrinia ID = "89" ; **acrisia** ID = "90" ; acrisia ID = "90a" ; acromia ID = "93" ; alalia ID = "145" ; alalia ID = "145a" ; anaraxia ID = "238" ; apraxia ID = "333" ; astasia ID = "364" ; atresia ID = "386" ; ectasia ID = "1015" ; sacralgia ID = "2747" ; uracrasia ID = "3188" ; xerasia ID = "3321"
aferente ID = "127"	aférese ID = "128" ; deferente ID = "863" ; **eferente** ID = "1022" ; frenite ID = "1450"
aferente ID = "127a"	aférese ID = "128" ; deferente ID = "863" ; **eferente** ID = "1022" ; frenite ID = "1450"
afusão ID = "135"	agrião ID = "139" ; agudo ID = "143" ; ambustão ID = "195" ; avulsão ID = "404" ; difusão ID = "936" ; difusão ID = "936a" ; efluxão ID = "1026" ; **efusão** ID = "1028" ; emulsão ID = "1065" ; emulsão ID = "1065a" ; frusto ID = "1453" ; função ID = "1463" ; oclusão ID = "2153" ; oclusão ID = "2153a" ; perfusão ID = "2365" ; pulsão ID = "2586" ; sufusão ID = "2937" ; tensão ID = "3022" ; ustão ID = "3215" ; ustão ID = "3215a" ; visão ID = "3296"
agnóia ID = "137"	agustia ID = "144" ; alalia ID = "145" ; alalia ID = "145a" ; algália ID = "168" ; amixia ID = "219" ; anemia ID = "263" ; aniria ID = "288" ; **anóia** ID = "292" ;atelia ID = "374" ; atelia ID = "374a" ; atimia ID = "377" ; sinóvia ID = "2870" ; tonia ID = "3071"
alelo ID = "161"	alilo ID = "174" ; amileno ID = "212" ; anel ID = "262" ; argo ID = "343" ; arilo ID = "345" ; astela ID = "366" ; atelia ID = "374" ; atelia ID = "374a" ; azoto ID = "409" ; bolo ID = "482" ; ceco ID = "606" ; dreno ID = "995" ; feto ID = "1345" ; feto ID = "1345a" ; filo ID = "1360" ; flagelo ID = "1378" ; flexor ID = "1392" ; flexão ID = "1391" ; fluxo ID = "1408" ; fluxo ID = "1408a" ; freio ID = "1446" ;

	galapo ID = "1480" ; gel ID = "1504" ; glena ID = "1527" ; hilo ID = "1630" ; janela ID = "1829" ; julepo ID = "1832" ; lobo ID = "1896" ; malato ID = "1923" ; maléolo ID = "1924" ; martelo ID = "1936" ; meso ID = "1990" ; oleato ID = "2164" ; oleula ID = "2170" ; plexo ID = "2447" ; safeno ID = "2752" ; talo ID = "2983" ; valgo ID = "3239" ; vela ID = "3266" ; vitelo ID = "3300" ; ílio ID = "1728" ; óvulo ID = "2250" ; óvulo ID = "2250a" ; útero ID = "3218"
aloftalmia ID = "179"	**anoftalmia** ID = "291" ; buftalmia ID = "506" ; oftalgia ID = "2160" ; oftalmia ID = "2161" ; xeroftalmia ID = "3322"
aloquezia ID = "182"	**aloquiria ID = "183"**
amida ID = "201"	amielia ID = "202" ; amina ID = "216" ; amixia ID = "219" ; amonita ID = "222" ; anemia ID = "263" ; aniria ID = "288" ; anóia ID = "292" ; arcada ID = "339" ; arilo ID = "345" ; atimia ID = "377" ; aura ID = "391" ; aurina ID = "394" ; coifa ID = "690" ; corda ID = "795" ; imido ID = "1731" ; lambda ID = "1849" ; lira ID = "1882" ; manita ID = "1931" ; miva ID = "2044" ; mola ID = "2048" ; mula ID = "2073" ; oliva ID = "2183" ; oliva ID = "2183a" ; oliva ID = "2183b" ; osido ID = "2220" ; papila ID = "2282" ; pica ID = "2402" ; pira ID = "2416" ; raiva ID = "2626" ; reira ID = "2671" ; soda ID = "2897" ; soda ID = "2897a" ; tenda ID = "3018" ; vagina ID = "3233" ; vagina ID = "3233a" ; veia ID = "3264" ; via ID = "3286" ; ácido ID = "69" ; ácido ID = "69a" ; ácido ID = "69b" ; ácido ID = "69c" ; ácido ID = "69d" ; óxido ID = "2255"
anfigénio ID = "271"	**anfígeno** ID = "272" ; antigénio ID = "309" ; oxigénio ID = "2257"
anfólito ID = "273"	**anfórico** ID = "275" ; anfótero ID = "276" ; apósito ID = "330" ; fólico ID = "1411"
apofilaxia ID = "322"	**profilaxia** ID = "2523"
basificar ID = "436"	basificação ID = "435"
basificar ID = "436a"	basificação ID = "435"
basílico ID = "438"	basófilo ID = "441" ; benzílico ID = "448" ; berílio ID = "453" ; básico ID = "434"

Annexe D

Nombre de corrections effectuées au niveau des entrées directes et indirectes grâce à l'application de l'AO

	MP	DPLP	GM	Wikip	Wikc	DeCS	CHCB
ÉI	rien	rien	rien	rien	entry : 5 syn : 1 rW : 230	entry : 71 syn : 98 rW : 15	rien
ÓI	entry : 184 syn : 47 rA : 15 rW : 25 comp :1	entry : 44 syn : 2 rA : 3	entry : 9 syn : 2	syn : 2	entry : 7 syn : 2 rW : 45	entry : 263 syn : 343 rW : 66	entry : 2
CT	entry : 214 syn : 68 rA : 23 rN : 2 rV : 2 rW : 7 ant : 4 comp :4	entry : 26	entry : 40 syn : 14	entry : 8 syn : 4	entry : 5 syn : 4 rW : 23	entry : 5 syn : 19 rw : 5	entry : 8 syn : 1
CÇ	entry : 98 syn : 15 rW : 45 comp : 36	entry : 20	entry : 21 syn : 5	entry : 19 syn : 4	entry : 2 syn : 2 rW : 2	entry : 148 syn : 185 rw : 30	entry : 6 syn : 2
Ü	rien	rien	rien	entry : 5	syn : 3 syn : 1	entry : 24 syn : 85 rw : 7	rien
CC	entry : 9 syn : 5 rW : 2 rA : 5	entry : 3	entry : 1 syn : 1	entry : 5 syn : 2	entry : 1	entry : 21 syn : 46 rw : 11	rien
Hífen1	rien	rien	rien	entry :2	rW : 2	entry : 29 syn : 35 rw : 5	rien
Hífen2	entry :129 syn : 48 rA : 3 rN : 4 rW : 26 comp :2	entry : 12 syn : 3 rA : 1	entry : 15 syn : 8	entry : 25 syn : 24 rW : 2	entry : 18 syn : 26 rW : 150	entry : 208 syn : 219 rw : 53	rien
PÇ	entry : 4	rien	entry : 3	entry :10	rien	entry : 34	rien

	MP	DPLP	GM	Wikip	Wikc	DeCS	CHCB
			syn : 2			syn : 15 rw : 14	
GD	entry : 9 syn : 3	entry : 5 syn : 1	syn : 1	entry :1 syn : 3	syn : 1	entry : 1 syn : 13	rien
PC	rW : 3	rien	rien	rien	entry : 4	entry : 18 syn : 4 rw : 3	rien
PT	entry :40 syn : 10 rA : 4 rW : 18 comp :1	entry : 5	entry : 6 syn : 1	entry : 30 syn : 6	entry : 4 rW : 11	entry : 547 syn : 928 rw : 22	entry : 4
BD	rien	rien	rien	rien	rien	rien	rien
BT	rien	rien	rien	rien	rien	rien	rien
MN	rien	rien	rien	rien	rien	rien	rien
TM	rien	rien	rien	rien	rien	rien	rien
ÊN	entry : 1	rien	rien	entry : 9 syn : 6 rA : 1	entry : 1 rW : 16	entry : 365 syn : 595 rw : 164	rien
ÊM	rien	rien	rien	entry : 3 syn : 1	entry : 5 syn :3 rW : 22	entry : 62 syn : 119 rw : 24	rien
ÔM	rien	rien	rien	rien	entry : 1 rW :30	entry : 152 syn : 136 rw : 66	rien
ÔN	rien	rien	rien	entry : 7 syn : 2	entry : 1 rw : 6	entry : 231 syn : 356 rw : 52	rien
ÔO	rien	rien	rien	rien	syn : 1	entry : 3 syn : 7 rw : 1	rien
-Ê	entry : 1	rien	rien	rien	entry : 3	entry : 5 syn : 1 rw : 3	rien
-Ô	rien	rien	rien	rien	rien	rien	rien
Totaux	1117	125	129	176	634	6132	23

Annexe E

Liste de termes uniques

E.1 Liste de termes uniques après l'unification intra-terminologique

Termes uniques	Terminologies correspondantes			
escutar	Pri id="1159"			
escápula	MedPt syn id="8633"	Pri id="1143"	Wpdia id="957"	Decs id="178"
escândio	Pri id="1142"	Decs id="9345"		
escócia	Wtnry rW id="2263"	Wtnry rW id="2264"		
Escólex	MedPt id="4900"	Wpdia id="3171"		
Eserina	Wpdia syn id="11189"			
Eserinum	Decs id="24822"			
esfacelado	Pri id="1160"			
esfacelo	glo syn id="773"	glo syn id="1202"		
Esfagno	Decs syn id="4942"			
Esfagnópsidas	Decs syn id="4942"			
Esfenisciforme	Decs syn id="1682"			
Esfeniscídeos	Decs syn id="1682"			
esfenoide	Pri id="1161"	Wtnry rW id="2466"	Wtnry rW id="2466b"	
Esferas de Látex	Decs syn id="18897"			
esferocitose hereditária	MedPt syn id="10590"	Wpdia id="2480"	Decs id="8011"	
Esferoides Celulares	Decs id="1113"			
Esferoplastos	Decs id="1350"			
esferócito	MedPt syn id="10435"	Wpdia id="5138"	Wtnry rW id="1534"	Decs id="1044"
esfigmo-	MedPt rW id="9982"			
esfigmomanómetro	MedPt syn id="11273"	Pri id="1163"	Wpdia id="3776"	Decs id="18940"

E.2 Liste de termes uniques avant l'unification inter-terminologique

desamínase	Desassimilação
Desamino Arginina Vasopressina	desassossego
Desaminoarginina Vasopressina	Desastre Geológico
Desamparo Adquirido	Desastre Hidrológico
Desamparo Aprendido	Desastre Humano
desarranjo	Desastre Meteorológico
desarranjo mental	Desastre Provocado pelo Homem
Desarticulação	Desastres
desarticulado	Desastres Antropogénicos
desarticular	Desastres Naturais

Desastres Tecnológicos
desatinado
desatinar
desazotar
Desbloqueamento
Desbridamento
descabelado
descabelar
descalçamento dentário
Descalcificação
Descalcificação Patológica
descalcificado
Descalcificante
descalvado
Descamação
descamativo
descansar
Descanulação
Descapsulação
descarbonatado
descarbonatar
Descarboxicistina
Descarboxilação
descarboxílase
Descarboxilase Pirúvica
Descarboxilases
Descarboxilases de Aminoácido Aromático
Descarboxilases de Aminoácido-L-Aromático
descarga
Descarga dos Membros Posteriores
Descarga Esquelética
Descarga Vaginal
descartar
Descendência Adulta
descenso
Descentralização
Descerebração
Descida a Rappel
Descoloração
Descloretação
descloretado
Descloretante
descoberta
Descoberta de Drogas
Descobertas Incidentais
Descobrimentos Incidentais
descolado
Descolagem Dentária
Descolamento
Descolamento da retina
Descolamento de retina
Descolamento do Epitélio Pigmentado da Retina
Descolamento do Epitélio Pigmentar da Retina
Descolamento do Vítreo
Descolamento do Vítreo Posterior
Descolamento Prematuro da Placenta
Descolamento Retiniano
Descoloração de Dente
Descompensação
Descompensação Cardíaca
Descompensado
Descompressão
Descompressão Abdominal
Descompressão Cirúrgica
Descompressão Explosiva
Desconcentração
descondicionado
Descondicionamento
Descondicionamento Cardiovascular
descondicionar
Desconexão
desconforto psíquico
Descongestionante
Descongestionante nasal
Descontaminação
Descontaminação Radioativa
descontaminado
Descontração
descontracturante
descontraído
descontraturante
descorticado
Irídio
Irmãos
irmãos siameses
IRMAs
Irmãs
IRMf
IRPPP
Irradiação
Irradiação a Laser de Baixa Intensidade
Irradiação a Laser de Baixa Potência
Irradiação Corporal Total
Irradiação Craniana

Irradiação de Alimentos
Irradiação Hemicorpórea
Irradiação Hemicorpórea Sequencial
Irradiação Hemicorpórea Sistémica
Irradiação Hipofisária
Irradiação Linfática
Irradiação Linfoide
Irradiado
Irredutível
irregular
irregularidade
irreversível
irrigação
Irrigação Gástrica
Irrigação Peritoneal
Irrigantes do Canal Radicular
irrigar
Irritabilidade
Irritação
Irritante
irritativo
irritável
IRS
IRV
Isatina
Isatis
Isavirus
Ischnocera
iscn(o)-
Iscnóceros
iscnofonia
iscnofónico
ISCOMs
Iscúria
Isertia
Isetionato de Pentamidina
ISFJ
ISFP
ISL
ISN
Iso-
Isoaglutinação
Isoamilase
Isoanticorpos
Isoantígenos
Isobamato
Isobutanos
Isobutilamidas Poli-insaturadas
Isobutilcianoacrilato
Isobutilteofilina
Isobutrazina
Isocarboxazida
Isocianatos
Isocianetos
Isocitrase
Isocitratase
Isocitrato Desidrogenase
Isocitrato Hidroliase
Isocitrato Liase
Isocitratos
Isocoria
Isocoros
isocromático
Isocromossomos
isocronismo
Isocumarinas
Isodesmosina
isodinamia
Isodissomia Uniparental
Isodon
Isoefedrina
Isoelétrico
Isoenxerto
Isoenzima
Isoenzima MB
Isoenzimas
isoenzimático
Iso-Eptanos
Isoetarina
Isoetretina
Isoexanos
Isofenilefrina
Isoflavona
Isoflavonas
Isoflurano
Isoflurofato
Iso-Hemaglutininas
Iso-Heptanos
Iso-octanos
Iso-OMPA
Prótons
prússico
PSA
Psacalium

Psalliota bispora
Psammomys
PSAS
Pselismo
Pseud-
Pseudacacia odorata
Pseudallescheria
Pseudallescheria boydii
Pseudallescheriose
Pseudalopex
Pseudartrose
Pseudo-
Pseudo Inhame
Pseudoacatisia
Pseudoafacia
Pseudoafaquia
Pseudoalteromonas
Pseudoanemia
Pseudoaneurisma
Pseudoaneurisma Carotídeo
Pseudoangioma
Pseudoarnica
Pseudoartrose
Pseudociese
Pseudocirrose de Pick
Pseudocisto
Pseudocisto pancreático
Pseudocoartação da aorta
Pseudococcus cacti
Pseudocolinesterase
Pseudocoma
Pseudodemência
Pseudodistonia
Pseudoefedrina
Pseudoepiléptico
Pseudoesclerose Cerebral
Pseudofacia
Pseudofaquia
Pseudofolliculitis barbae
Pseudogenes
Pseudoglobulinas
Pseudogonococia enterítica
Pseudogota
Pseudogravidez
Pseudo-hermafrodita
pseudo-hermafroditismo
Pseudo-Hipoaldosteronismo
Pseudo-Hipoparatireoidismo
pseudo-hipoparatiroidismo
Pseudoinsulinorresistência
Pseudolinfoma
Pseudomelia
Pseudomembrana
pseudomembranoso
Pseudomeningite
Pseudomiotonia
Pseudomixoma Peritoneal
Pseudomonadaceae
Pseudomonas
Pseudomonas acidovorans
Pseudomonas aeruginosa
Pseudomonas alcaligenes
Pseudomonas cepacia
Pseudomonas fluorescens
Pseudomonas fragi
Pseudomonas mallei
Pseudomonas mendocina
Pseudomonas oleovorans
Pseudomonas pseudoalcaligenes
Pseudomonas pseudomallei
Pseudomonas putida
Pseudomonas pyocyanea
Pseudomonas stutzeri
Pseudomonas syringae
Pseudomonas testosteroni
Pseudomonas-Fago 7s
Pseudomonas-Fago PP7
pseudomorfose
Pseudo-Obstrução Colónica
Pseudo-Obstrução Intestinal
Pseudoparalisia
Pseudopelada
Pseudopleuronectes
Pseudopódios
Pseudopoliartrite rizomélica
Pseudo-Pseudo-Hipoparatireoidismo
Pseudopsicose
pseudoptose
Pseudoquisto
Pseudorraiva
Pseudorrotaxanos
Pseudosclerose
Pseudotabes alcoólica
Pseudotsuga

pseudotuberculose pseudotumor	Pseudotuberculose por Pasteurella