Évaluation de l'indexation des comptes rendus médicaux à l'aide d'un outil états-unien adapté pour le français

Saoussen Sakji[1], Peter Elkin[2], Stéfan Darmoni[1]

[1]CISMeF, Centre Hospitalier Universitaire de Rouen ;
[2]Center for Biomedical Informatics, Mount Sinai School of Medicine, One Gustave L. Levy Place, Box 1023, New York, NY 10029, United States.

Abstract

Objective: To evaluate the ability of the bilingual (French-English) language vocabulary server to identify the same main concepts of ICD-10 (International Classification of Disease v.10) in the same set of French and English Electronic Health Records. Material and methods: The Intelligent Natural Language Processor (iNLP) was adapted to parse French as well as English clinical content. Six clinical notes were translated from French to English. The English notes were processed with the English version of ICD-10 and the French notes were processed with the French version. The resultant codes were compared. Results: The English Version contained 36 concepts of which 26 were in common with the French (72.2%). Failure analysis showed that errors were mostly due to missing synonymy in the French ICD-10 terminology server and to problems associated with translating the records from French to English. Discussion: Multi-lingual iNLP is feasible and has the potential to enable querying across national boundaries and multi-lingual communities.

Keywords

Abstracting and indexing as topic; International Classification Diseases; Electronic Health Records; SNOMED CT; Intelligent Natural Language Processor

1 Introduction

Le dossier médical informatisé [1] est l'une des composantes essentielles du système d'information de santé. En France, il est précisé dans la loi du 4 mars 2002 relative aux droits des malades que le dossier médical concerne l'élaboration des suivis de diagnostic, les traitements, mais aussi plus généralement, tous les échanges écrits entre les professionnels de santé. Le dossier médical informatisé est, donc, constitué d'informations administratives et médicales nominatives qui forment une base de données dans le sens où il s'agit d'un « recueil d'œuvres, de données, ou d'autres éléments indépendants, disposés de manière systématique ou méthodique et individuellement accessibles par des moyens électroniques ou par tout autre moyen » (loi du 1 juillet 1998).

En effet, l'informatisation des dossiers patients permet, entre autres, de :

- faciliter l'accès aux données des patients par la fourniture d'outils de classification permettant de retrouver les informations rapidement selon plusieurs critères : par

nature des données (cliniques, biologiques, imagerie), par ordre chronologique, par nom, par âge, par lieu de domiciliation, par type d'affection ;

- offrir au patient la possibilité d'accéder à son dossier à n'importe quel endroit du monde et en plusieurs langues. De plus, il permet de favoriser la prise de conscience et la prise en charge par le patient lui-même de sa santé grâce par exemple à l'implémentation de messages d'alerte automatique (rappel de vaccinations obligatoires, de consultations annuelles ou d'examens complémentaires à effectuer.

Pour ce faire, la représentation des informations spécifiques au patient sous forme de texte libre peut être utile dans le cas d'une lecture simple des rapports, cependant, ce format est difficilement utilisable pour les systèmes d'aide à la décision ou lors de la réalisation des études cliniques démographiques [2].

L'avancée des recherches sur le traitement automatique du langage naturel (TAL) et l'extraction et l'indexation automatique des documents ont fait évoluer le traitement de l'information et des connaissances. Dans ce contexte, plusieurs terminologies médicales permettent de représenter et de décrire le contenu des dossiers patients informatisés, notamment la SNOMED (Systematized Nomenclature of Medicine). Dans la même perspective, l'United Medical Langage System (UMLS) [3], développé par la NLM, se propose d'établir un lien conceptuel entre le besoin d'une information exprimé par un utilisateur et différentes sources d'informations informatisées comme des bases de données sur la littérature médicale, les dossiers médicaux ou les bases de connaissances [4] [5]. Plusieurs travaux se sont intéressés à cette piste de recherche et nous pouvons citer, à titre d'exmple, l'étude réalisée par *La Mayo Clinic* qui a développé un système d'indexation basé sur la SNOMED et sur un lexique propre à la Mayo. Ensuite, un test portant sur l'utilisation d'UMLS dans le processus de recherche de documents s'est révélé positif, ouvrant ainsi la porte à une interface entre les données propres aux patients et aux autres sources de connaissances biomédicales [6]. *H. Warner* décrit de même un système de dossier médical basé sur UMLS, capable de partager la représentation de patients entre plusieurs systèmes de connaissance, tels que Iliad (développé par l'Université d'Utah) et QMR (Quick Medical Reference de Pittsburgh). Le but du système est de communiquer des données de patients entre plusieurs institutions, afin de comparer des paramètres tels que prévalence de maladie, sensibilités et spécificités des signes dans les maladies, mesure d'évolution clinique ("outcome") [7].

Plusieurs travaux ont été réalisés pour extraire et manipuler, à partir du texte libre, une information structurée. Le système d'extraction d'information médicale (MedIE) [8] a été mis au point pour repérer les diagnostics des patients à partir des comptes rendus cliniques en format texte libre. F-MTI, un outil d'indexation automatique a été utilisé pour indexer les documents et les dossiers de santé électroniques, avec plusieurs terminologies en langue française (CCAM, TUV) ou traduites en français (MeSH, CIM-10 et SNOMED) [9]. eQuality, un outil d'indexation fondé sur les concepts de la nomenclature SNOMED-CT, a été employé pour indexer les rapports des examens [10]. Les résultats obtenus sont prometteurs dans la mesure où le système était aussi performant qu'un être humain, selon une grille d'indicateurs de qualité [10].

Dans le même contexte, le serveur terminologique de traitement de langage naturel iNLP a été mis en place. Il s'agit d'un ensemble d'outils développés pour faciliter l'indexation des dossiers patients informatisés (DPI) en utilisant un vocabulaire fondé sur des concepts. Le système est dépendant des terminologies médicales et, souvent, utilise la SNOMED CT ; la terminologie de plus en plus utilisée aux États-Unis pour la description des DPI.

L'identification des concepts qui sont explicitement affirmés négativement (par exemple « aucune évidence de la pneumonie ») et les distinguer des affirmations positives devient un souci de plus en plus important si nous souhaitons analyser et comprendre les implications du texte médical [11]. La négation linguistique dans les dossiers médicaux devient le problème et le défi à soulever pour les professionnels de santé. Dans cet exemple « *une femme âgée de 62 ans qui se présente, avec un érythème au-dessus du dos du pied gauche et une douleur exquise au-dessus d'une blessure située à mi pied. Après une étude clinique, il s'est avéré qu'elle avait une cellulite du pied gauche sans signes de propagation lymphangitique de son infection* », il est important de déterminer que la patiente n'a pas de « Lymphangitie » liée à sa « cellulite, pied gauche » par opposition à un cas distinct où le diagnostic de « Lymphangitie » était présent. Pour une recherche épidémiologique et dans le cas où on étudierait la « Lymphangitis », il serait important d'exclure le cas de cette patiente de l'étude réalisée et de l'analyse des données.

À cette fin, dans la version anglaise du serveur terminologique, les DPI sont analysés pour identifier les expressions négatives. Le système identifie des affirmations positives, négatives et incertaines et un ensemble d'assertions logiques (vrai, plus probablement vrai que faux, même probabilité vrai que faux, plus probablement faux que vrai, faux…) pour exprimer le degré de certitude.

En 2005, le système a été évalué pour déterminer la précision à l'identification des cas de présence de la pneumonie dans un corpus de rapports radiologiques. Des assertions (vrais, possibles et absence) ont été désignées pour séparer les différents cas possibles [11]. Les rapports radiologiques ont été analysés et indexés par la terminologie de référence, la SNOMED CT. Chaque section du rapport a été identifiée (par exemple historique, résultats…). Dans chaque section, chaque concept a été codé et étiqueté comme étant une affirmation positive, négative ou incertaine.

À travers cet article, nous exposons notre étude à évaluer la capacité d'un serveur terminologique de traitement de langage naturel, développé pour la langue anglaise à s'adapter à d'autres langues, en l'occurrence le français. Notre but est de déterminer si le serveur terminologique bilingue, ainsi mis en place, permet d'identifier les mêmes concepts à partir de texte libre des comptes rendus médicaux.

2 Matériel et méthodes

Hypothèse :

La validation d'un analyseur multilingue implique que le même dossier de santé, disponible dans des langues différentes (par exemple dans ce cas, anglais et français), soit indexé par les mêmes codes d'une terminologie donnée.

En comparant les résultats de l'indexation du même ensemble de comptes rendus médicaux français (et traduits en anglais), par le serveur terminologique bilingue, nous nous attendons à obtenir les mêmes codes CIM-10.

Conception de l'étude :

Etant donné que la SNOMED CT n'est toujours pas disponible en français, nous avons créé une version anglaise et française du serveur terminologique pour la CIM-10. Par ailleurs, en France, comme dans beaucoup d'autres pays, les praticiens emploient la CIM-10 pour coder les dossiers patients informatisés pour la tarification à l'activité (T2A).

La CIM-10 permet le codage des maladies, des traumas et des raisons du recours aux services de santé. La classification a été conçue pour « permettre l'analyse systématique, l'interprétation et la comparaison des données de la mortalité et de la morbidité recueillies dans différents pays ou régions et à des époques différentes».

Pour évaluer l'étude, dix dossiers patients informatisés ont été fournis et pré anonymisés par le CHU de Rouen et indexés par la CIM-10. Ensuite, ces derniers ont été traduits en anglais et indexés par la version anglaise de la CIM-10. Enfin, une comparaison des codes restitués par l'indexation permet de mettre en relief les correspondances exactes entre les deux versions du serveur. (Voir la figure1 ci-dessous).

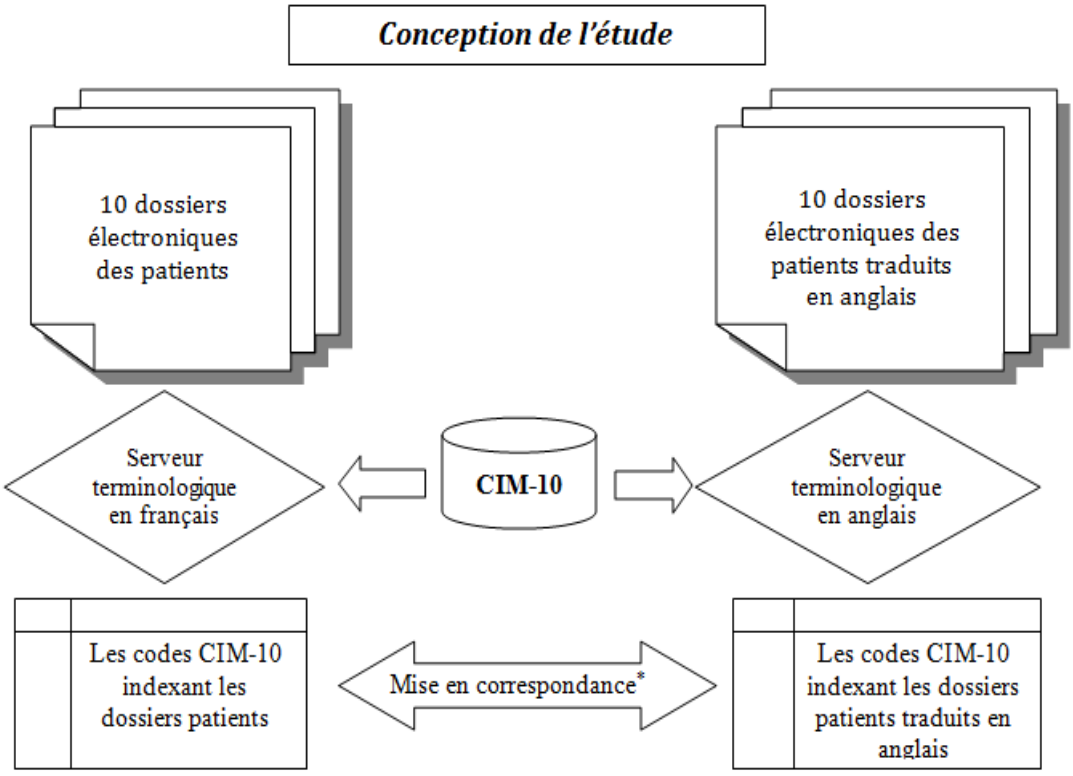


Figure 1 : Conception de l'approche de comparaison des résultats du serveur terminologique

Les dossiers patients informatisés ont été choisis en provenance de services différents : le service de pneumologie, le service de cardiologie, le service de chirurgie thoracique et cardio-vasculaire, le service d'endocrinologie, diabète et maladies métaboliques et le service de physiologie digestive, urinaire et respiratoire…

Pour comparer les résultats de l'indexation entre les deux versions du serveur, chaque compte rendu a été entièrement parcouru en utilisant l'analyseur de traitement de langage naturel et indexé par la CIM-10 (voir les figure2 et figure3). Chaque section du compte rendu a été identifiée (par exemple antécédents, résultats…). Dans chaque section, chaque concept a été codé et étiqueté comme étant une affirmation positive, négative ou incertaine. S'ajoute à cela, l'identification des opérateurs booléens (et, ou, pas) entre les différents concepts permettant de mettre en relief une conjonction ou une disjonction entre des traitements ou des diagnostics.

Pour la version française du traitement, nous avons dû mettre au point quelques processus et modèles du système existant. L'indépendance du langage était le but majeur des efforts entrepris.

Les étapes principales de l'adaptation du serveur terminologique sont décrites ci-dessous :

1. *Création du modèle de langue français* : à partir d'un ensemble représentatif de dossiers électroniques des patients français, livrés par l'hôpital de Rouen, nous avons étudié les différentes formes d'assertions négatives, positives et incertaines. La différence entre les deux modèles de langue anglaise et française est due à la manière et la façon dont les médecins expriment leurs affirmations et la connaissance au sujet des patients.

   Pour cela, nous avons dû identifier et extraire toutes les différentes formes d'expressions négatives françaises (principalement les opérateurs négatifs tels que *aucun*, *sans*) et les distinguer des affirmations positives (par exemple *détecté*, *persistance*) et les incertaines (tels que *probablement*, *suspect*). En outre, nous avons identifié les « andOperators » pour exprimer la conjonction des affirmations, les « orOperators » pour identifier une distinction entre deux traitements ou deux idées et finalement les opérateurs temporels (par exemple *depuis*, *ensuite*) pour établir la chronologie des événements.

2. *Adaptation de l'outil de désuffixation pour la langue française* : le principe de base de l'indexation est de générer les principaux concepts représentatifs d'un document. Pour cela, combiner les variantes d'un mot dans la même racine s'avère une bonne pratique pour identifier le maximum de concepts. Par exemple, les mots « pensé », « penseurs » ou « pensait » peuvent être représentés par la même racine « pense». Ce principe a pour but, en finalité, d'améliorer la recherche d'information dans la mesure où les mots ayant la même racine se rapportent à la même idée ou concept et doivent, donc, être indexés sous la même forme.

   La première étape pour l'algorithme de désuffixation est d'enlever les suffixes flexionnels ou, pour la langue anglaise, de combiner les formes singulières et plurielles des mots et d'enlever la terminaison du participe passé « - ed » et la terminaison du gérondif ou du participe présent « ing ». La plupart des algorithmes de désuffixation doivent respecter certaines contraintes quantitatives (par exemple, une longueur minimale de la chaine de caractère de la racine doit être respectée) et des contraintes qualitatives (par exemple, la fin de la chaine de caractère doit être sous certaine forme ; enlever les doublons au cas de besoin). En conclusion, un ensemble de règles peut être appliqué afin de changer des racines (par exemple, « hopping » sans « ing » donne le « hop » et pas « hopp »).

   Pour la langue française, il faut prendre en compte un certain nombre d'irrégularités. Bien que la langue anglaise contienne, déjà, des irrégularités morphologiques (box/boxes, mouse/mice, keep/kept), la langue française est beaucoup plus complexe. En effet, il faut considérer les suffixes flexionnels concernant les variations de genre (masculines vs. féminin) et les variations de nombre (singulier vs. pluriel) pour les noms et les adjectifs. Pour les verbes, la conjugaison se fait selon un nombre de traits grammaticaux tels que le temps et la personne et qui doivent être pris en compte. Une étude récente effectuée par [13] a pour objectif de comparer quelques algorithmes de désuffixation français, mettait en relief les avantages de l'algorithme de Lucene [14] qui s'inspire largement du l'algorithme anglais Porter [15]. L'évaluation réalisée a été effectuée avec trois algorithmes : celui de CISMeF, l'algorithme de Carry [16] et

l'algorithme de Lucene. Bien que, généralement, la désuffixation soit fondée sur le même principe, la différence observée entre les trois algorithmes est due aux règles appliquées. Le résultat de l'évaluation révèle une F-mesure de 77,9% pour l'algorithme de Lucene, 70,4% pour l'algorithme de CISMeF et finalement de 66,7% pour Carry.

Pour cette raison, dans notre étude nous avons choisi l'algorithme de Lucene pour avoir les meilleurs cas de désuffixation.

3. *Construction du serveur terminologique* : Afin d'adapter le serveur terminologique à la langue française, nous avons dû configurer la CIM-10 selon les informations structurées déjà existantes : concepts, termes, synonymes de concept et relation. Une fois la CIM-10 stockée dans la base de données, des termes complémentaires fondés sur les concepts originaux de la CIM-10 sont générés pour enrichir la base de données. Ensuite, en utilisant l'algorithme de désuffixation, les formes canoniques de l'ensemble des termes ont été créées. La troisième étape pour construire le serveur terminologique était d'indexer chaque mot de chaque terme (concept) dans le but d'améliorer le processus de la recherche d'information.

   La structure hiérarchique (chapitre, bloc, bloc secondaire, catégorie, catégorie secondaire et subdivision) de la CIM-10 permet d'avoir les relations « est un », « exclusion » et « inclusion » entre les concepts.

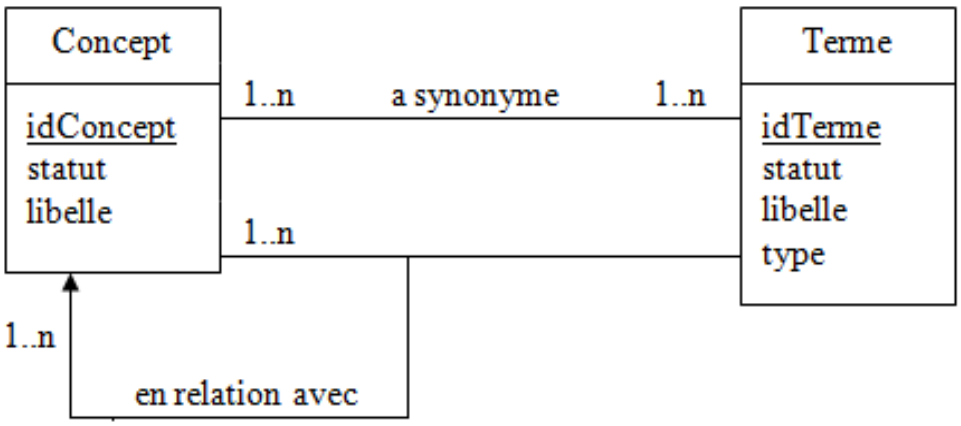


Figure 2 : Modélisation UML de la CIM-10 selon le modèle du langage

La base de données contient les concepts (termes préférées) des terminologies ainsi que les termes qui référencent les concepts (les synonymes).

Par exemple, le concept CIM-10 « angine de poitrine instable » admet comme concepts synonymes : « angine d'effort », « angine accélérée », « syndrome de préinfarctus »...

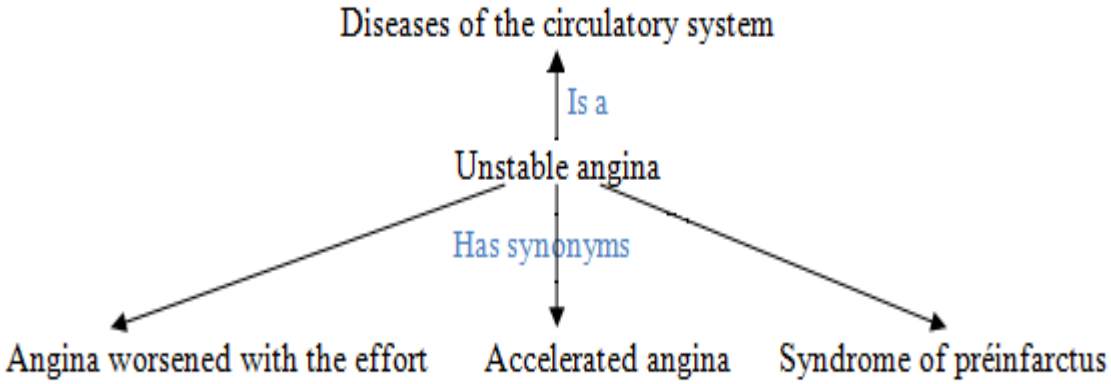


Figure 3 : Relations (hiérarchique et synonymie) de la CIM-10

3 Résultats

Nous récapitulons les résultats de l'indexation des six comptes rendus des dossiers patients informatisés dans le tableau 1. Nous remarquons, dans la plupart des cas, une différence entre le nombre de concepts d'indexation anglais et français.

Tableau 1 : Résultats de l'indexation par les codes CIM-10 de la version anglaise et française de l'analyseur

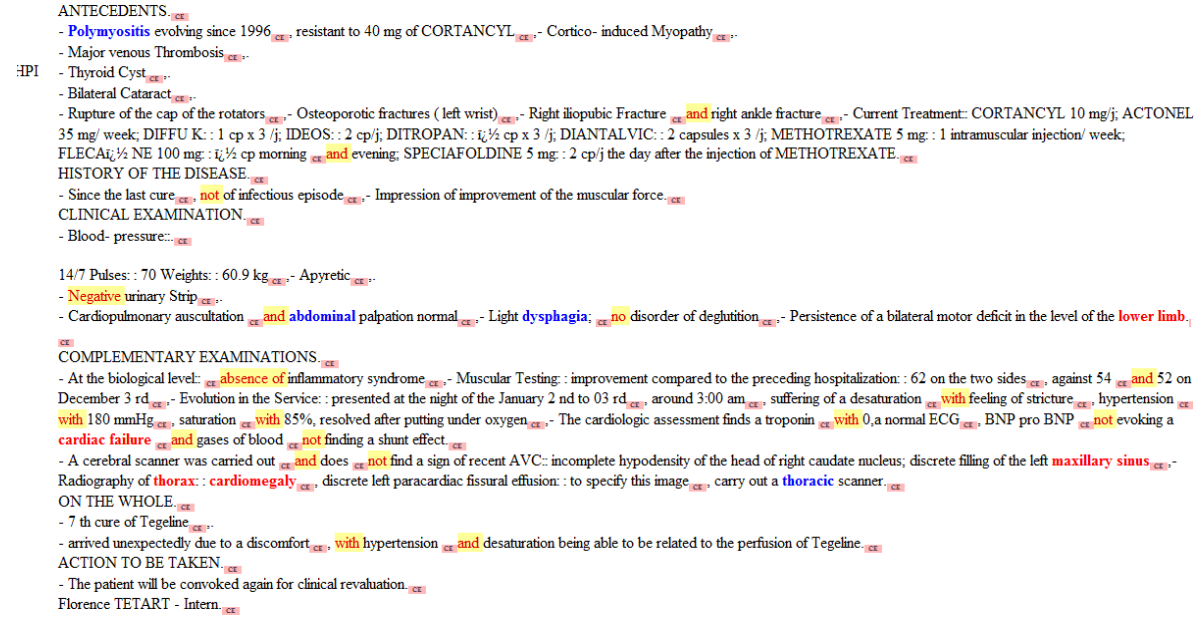
N° du DEP	**Nombre des concepts anglais détectés**	**Nombre des concepts français détectés**	**Nombre des concepts en correspondance**	**Nombre des concepts détectés seulement en anglais**	**Nombre des concepts détectés seulement en français**
1	4	3	3	1	0
2	3	4	3	0	1
3	1	0	0	1	0
4	9	12	7	2	5
5	10	9	7	3	2
6	9	10	6	3	4

La version anglaise de l'analyseur terminologique CIM-10 a permis de retrouver 36 codes CIM-10 tandis que la version française a permis de retrouver 38 codes ; 26 codes d'indexation étaient présents dans les deux versions du parseur.

Les figures 4 et 5 montrent un exemple de résultats de l'indexation d'un compte rendu (version française et anglaise) par la CIM-10. Les concepts restitués sont mis en relief par des couleurs différentes, afin d'améliorer son interprétation (les affirmations positives sont en bleu et les négatives en rouge).

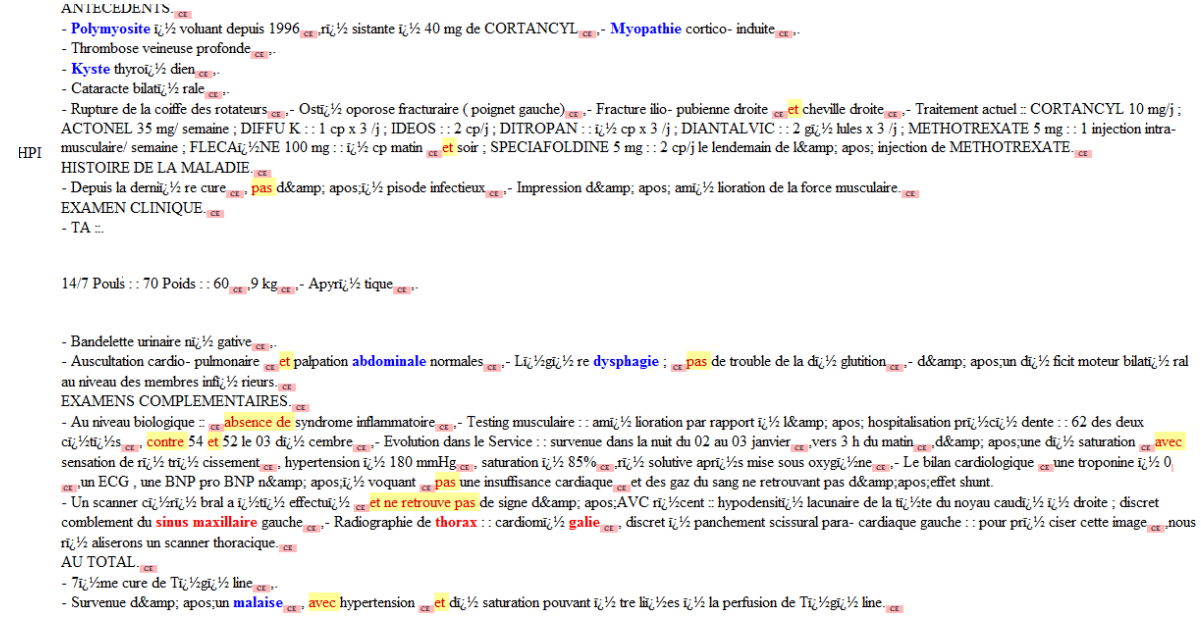
L'analyse de l'indexation de l'analyseur terminologique nous a permis de pointer les causes du non correspondance qui était due, en partie, à la traduction des comptes rendus de la version française à la version anglaise. S'ajoute à cela, un manque de synonymie, généralement, de la version française de la CIM-10 et à quelques problèmes dus au traitement du langage naturel lors du processus de mise en correspondance (voir la figure 6). Parmi les correspondances imprécises, nous enregistrons 16% due à la traduction des notes françaises à la version anglaise ; 32% au manque de synonymie et 52% à l'analyseur

de TAL. Pour le dernier cas, parfois la non correspondance est due à l'orthographe des concepts (« surénale » au lieu de « surrénale »).


ANTECEDENTS.
- Polymyositis evolving since 1996, resistant to 40 mg of CORTANCYL,- Cortico- induced Myopathy,.
- Major venous Thrombosis,.
HPI - Thyroid Cyst,.
- Bilateral Cataract,.
- Rupture of the cap of the rotators,- Osteoporotic fractures (left wrist),- Right iliopubic Fracture and right ankle fracture,- Current Treatment:: CORTANCYL 10 mg/j; ACTONEL 35 mg/ week; DIFFU K: : 1 cp x 3 /j; IDEOS: : 2 cp/j; DITROPAN: : ï¿½ cp x 3 /j; DIANTALVIC: : 2 capsules x 3 /j; METHOTREXATE 5 mg: : 1 intramuscular injection/ week; FLECAï¿½ NE 100 mg: : ï¿½ cp morning and evening; SPECIAFOLDINE 5 mg: : 2 cp/j the day after the injection of METHOTREXATE.
HISTORY OF THE DISEASE.
- Since the last cure, not of infectious episode,- Impression of improvement of the muscular force.
CLINICAL EXAMINATION.
- Blood- pressure::.

14/7 Pulses: : 70 Weights: : 60.9 kg,- Apyretic,.
- Negative urinary Strip,.
- Cardiopulmonary auscultation and abdominal palpation normal,- Light dysphagia; no disorder of deglutition,- Persistence of a bilateral motor deficit in the level of the lower limb.
COMPLEMENTARY EXAMINATIONS.
- At the biological level:: absence of inflammatory syndrome,- Muscular Testing: : improvement compared to the preceding hospitalization: : 62 on the two sides, against 54 and 52 on December 3 rd,- Evolution in the Service: : presented at the night of the January 2 nd to 03 rd, around 3:00 am, suffering of a desaturation with feeling of stricture, hypertension with 180 mmHg, saturation with 85%, resolved after putting under oxygen,- The cardiologic assessment finds a troponin with 0,a normal ECG, BNP pro BNP not evoking a cardiac failure and gases of blood not finding a shunt effect.
- A cerebral scanner was carried out and does not find a sign of recent AVC:: incomplete hypodensity of the head of right caudate nucleus; discrete filling of the left maxillary sinus,- Radiography of thorax: : cardiomegaly, discrete left paracardiac fissural effusion: : to specify this image, carry out a thoracic scanner.
ON THE WHOLE.
- 7 th cure of Tegeline,.
- arrived unexpectedly due to a discomfort, with hypertension and desaturation being able to be related to the perfusion of Tegeline.
ACTION TO BE TAKEN.
- The patient will be convoked again for clinical revaluation.
Florence TETART - Intern.

Figure 4 : Résultats de la version anglaise de l'analyseur terminologique


ANTECEDENTS.
- Polymyosite ï¿½ voluant depuis 1996,rï¿½ sistante ï¿½ 40 mg de CORTANCYL,- Myopathie cortico- induite,.
- Thrombose veineuse profonde,.
- Kyste thyroï¿½ dien,.
- Cataracte bilatï¿½ rale,.
- Rupture de la coiffe des rotateurs,- Ostï¿½ oporose fracturaire (poignet gauche),- Fracture ilio- pubienne droite et cheville droite,- Traitement actuel :: CORTANCYL 10 mg/j ; ACTONEL 35 mg/ semaine ; DIFFU K : : 1 cp x 3 /j ; IDEOS : : 2 cp/j ; DITROPAN : : ï¿½ cp x 3 /j ; DIANTALVIC : : 2 gï¿½ lules x 3 /j ; METHOTREXATE 5 mg : : 1 injection intra-
HPI musculaire/ semaine ; FLECAï¿½NE 100 mg : : ï¿½ cp matin et soir ; SPECIAFOLDINE 5 mg : : 2 cp/j le lendemain de l& apos; injection de METHOTREXATE.
HISTOIRE DE LA MALADIE.
- Depuis la dernii¿½ re cure, pas d& apos;ï¿½ pisode infectieux,- Impression d& apos; amï¿½ lioration de la force musculaire.
EXAMEN CLINIQUE.
- TA ::.

14/7 Pouls : : 70 Poids : : 60,9 kg,- Apyrï¿½ tique,.

- Bandelette urinaire nï¿½ gative,.
- Auscultation cardio- pulmonaire et palpation abdominale normales,- Lï¿½gï¿½ re dysphagie ; pas de trouble de la dï¿½ glutition,- d& apos;un dï¿½ ficit moteur bilatï¿½ ral au niveau des membres infï¿½ rieurs.
EXAMENS COMPLEMENTAIRES.
- Au niveau biologique :: absence de syndrome inflammatoire,- Testing musculaire : : amï¿½ lioration par rapport ï¿½ l& apos; hospitalisation prï¿½cï¿½ dente : : 62 des deux cï¿½tï¿½s, contre 54 et 52 le 03 dï¿½ cembre,- Evolution dans le Service : : survenue dans la nuit du 02 au 03 janvier,vers 3 h du matin,d& apos;une dï¿½ saturation avec sensation de rï¿½ trï¿½ cissement, hypertension ï¿½ 180 mmHg, saturation ï¿½ 85%,rï¿½ solutive aprï¿½s mise sous oxygï¿½ne,- Le bilan cardiologique une troponine ï¿½ 0, un ECG , une BNP pro BNP n& apos;ï¿½ voquant pas une insuffisance cardiaque et des gaz du sang ne retrouvant pas d'effet shunt.
- Un scanner cï¿½rï¿½ bral a ï¿½tï¿½ effectuï¿½ et ne retrouve pas de signe d& apos;AVC rï¿½cent :: hypodensitï¿½ lacunaire de la tï¿½te du noyau caudï¿½ ï¿½ droite ; discret comblement du sinus maxillaire gauche,- Radiographie de thorax : : cardiomï¿½ galie, discret ï¿½ panchement scissural para- cardiaque gauche : : pour prï¿½ ciser cette image,nous rï¿½ aliserons un scanner thoracique.
AU TOTAL.
- 7ï¿½me cure de Tï¿½gï¿½ line,.
- Survenue d& apos;un malaise, avec hypertension et dï¿½ saturation pouvant ï¿½ tre lié¿½es ï¿½ la perfusion de Tï¿½gï¿½ line.

Figure 5 : Résultats de la version française de l'analyseur terminologique

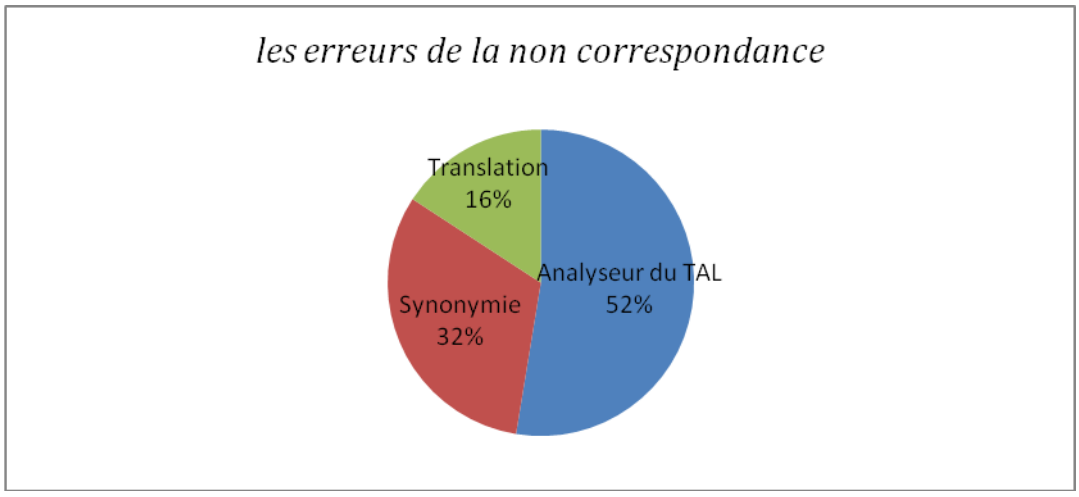


Figure 6 : Classement des erreurs de la non correspondance entre les deux versions (anglaise-française) de l'analyseur terminologique

4 Discussion

L'indexation des comptes rendus, avec la version anglaise de l'analyseur, a été réalisée avec 36 concepts dont 26 étaient en commun avec la version française; soit 72,2%. L'analyse de la non correspondance entre les deux versions (française et anglaise) a montré que le pourcentage le plus élevé des erreurs concerne l'analyseur TAL pour identifier les différentes expressions françaises pour exprimer un concept de la CIM-10. Par exemple, dans la version anglaise, nous avons pu restituer et indexer le compte rendu par l'assertion positive « *ischemic mixed heart disease* », alors que la version française a manqué l'identification de l'assertion « *cardiopathie mixte ischémique* ». S'ajoute à cela, le manque de synonymie dans le serveur terminologique français (i.e. *foie* synonyme *inter-hépatique*) et les problèmes liés à la traduction (français-anglais) des comptes rendus (i.e. mauvaise traduction du terme *Myopathie*).

La traduction automatique est un problème majeur de la linguistique informatisée. En plus, la différence des modèles de langue constitue, en soit même, un défi d'une traduction précise et fidèle au sens des concepts. Un modèle de langue permet de détecter les régularités linguistiques d'une langue. Or, les structures initiales qui étaient conçues pour le traitement optimal de l'anglais ne sont plus très adaptées pour coder les ressources linguistiques des autres langues, notamment le français dans notre cas. Chaque développement d'une nouvelle langue apporte sa portion d'ajustements. C'est pour cela, le passage automatique vers la version française du modèle de langue nous a paru impossible et d'où la nécessité de son adaptation d'une manière manuelle et intelligente, en respectant les contraintes de l'existant.

L'enrichissement du serveur terminologique, par l'ajout des synonymes des concepts CIM-10, a pour effet d'améliorer les résultats de la correspondance entre les deux versions d'indexation, d'une part, et le résultat global de reconnaissance de concepts, d'autre part. En effet, la robustesse des systèmes d'information dépend de la richesse du vocabulaire de référence, entre autres au niveau des synonymes.

Dans le futur proche, nous souhaitons refaire l'étude avec la terminologie SNOMED CT pour évaluer les résultats d'un vocabulaire plus détaillé et complet médicalement que la CIM-10. Notre but est de réaliser des échanges de données des dossiers médicaux d'une langue à une autre et d'un service ou d'un pays à un autre.

Cette étude démontre que le traitement de langage naturel multilingue est réalisable et a le potentiel de permettre l'échange de données et de recherches à travers des frontières nationales et des communautés multilingues.

Des langues telles que l'espagnol et le chinois sont programmées pour enrichir le serveur terminologique de traitement de langage naturel multilingue.

5 Conclusion

Cette étude nous permet de conclure que le traitement du langage naturel multilingue est faisable et révèle des résultats prometteurs. Les améliorations de la synonymie et de la traduction des DPI peuvent améliorer considérablement les résultats actuels.

Remerciements

Les auteurs remercient tous les membres de l'équipe états-unienne du Professeur Peter Elkin qui ont participé à la réalisation de cette étude.

Références

1. Degoulet P, Fieschi M. Informatique Médicale, 3ème édition, Eds Masson, 1998
2. Ceusters W, Elkin P, Smith B. Negative findings in electronic health records and biomedical ontologies: a realist approach. *Int J Med Inform*, 2007; pp. 326-333.
3. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. *Methods Inf Med* 1993; 32(4): 281-91
4. Lindberg DA, Humphreys BL. The UMLS Knowledge Sources: Tools for Building Better User Interfaces. In: *Proc. 14th SCAMC*. Washington, DC: IEEE. 1990; pp.121–125
5. McCray AT. The UMLS Semantic Network. In: *Proc. 12th SCAMC*. Washington, DC: IEEE. 1998; pp.503–507.
6. Chute CG, Y Yang, Tuttle MS, Sherertz DD, Olson NE, Erlbaum MS. A Preliminary Evaluation of the UMLS Metathesaurus for Patient Record Classification. In: Miller RA (ed). *Proceedings of the 14th annual SCAMC*. Washington, D.C. IEEE Computer Society Press, 1990; pp.161-165.
7. Fu LS, Bouhaddou O, Huff SM, Sorenson DK, Warner HR. Toward A Public Domain UMLS Patient Database. In: Miller RA (ed). *Proceedings of the 14th annual SCAMC*. Washington, D.C. IEEE Computer Society Press, 1990; pp.170-174.
8. Zhou X, Han H, Chankai I, Prestrud A, Brooks A. Approaches to text mining for clinical medical records. *Proceedings of the 2006 ACM symposium on applied computing*. Dijon, France, 2006; pp. 235-239.
9. Pereira S, Massari P, Buemi A, Dahamna B, Serrot E, Joubert M, Darmoni SJ. F-MTI : outil d'indexation multi-terminologique : application à l'indexation automatique de la SNOMED. *JFIM*. Nice, France, 2009 ; pp. 57-67.
10. Brown SH, Elkin PL, Rosenbloom ST, Fielstein E, Speroff T. eQuality for all: Extending automated quality measurement of free text clinical narratives. *AMIA Annu Symp Proc*. 2008 ; 6 : 71-5

11. Elkin PL, Brown SH, Bauer BA, Husser CS, Carruth W, Bergstrom LR, Wahner-Roedler DL. A controlled trial of automated classification of negation from clinical notes. *BMC Med Inform Decis Mak.* 2005 ; pp. 5-13.
12. Elkin PL, Froehling D, Wahner-Roedler D, Trusko B, Welsh G, Ma H, Asatryan AX, Tokars JI, Rosenbloom ST, Brown SH. NLP-based identification of pneumonia cases from free-text radiological reports. *AMIA Annu Symp Proc*. 2008; pp.172-6.
13. Pereira, S. *Indexation Multi-Terminologique de Concepts en Santé.* Mémoire de thèse, Rouen: Université de Rouen, 2008.
14. Hatcher E, Gospodnetic O. Lucene in Action. *Manning Publications*, 2004.
15. Porter MF. "An algorithm for suffix stripping." *Program*, 1980 ; pp. 130-137.
16. Paternostre M, Francq P, Lamoral J, Wartel D, Saerens M. Carry, un algorithme de désuffixation pour le français. Rapport Technique, Université libre de Bruxelles, 2002, http://beams.ulb.ac.be/beams/documents/carryfinal.pdf.

Adresse de correspondance

Saoussen SAKJI

Equipe CISMeF, CHU de Rouen

Laboratoire LITIS EA 4108

1 rue de Germont

76031, Rouen

E-mail: saoussen.sakji@chu-rouen.fr