T.A.L., vol. 41, n° 3, pp. 727–757

STRATÉGIES D'IDENTIFICATION DES NOMS PROPRES À PARTIR DE NOMENCLATURES MÉDICALES PARALLÈLES

Olivier Bodenreider [*] **Pierre Zweigenbaum** [**]

Résumé - Abstract

Les domaines spécialisés comme la Médecine utilisent un grand nombre de noms propres dans leurs terminologies. L'objectif de ce travail est d'étudier les critères permettant d'identifier les noms propres dans les termes médicaux, débouchant sur une stratégie pour l'étude de nouvelles terminologies. Les critères étudiés portent sur la casse des caractères, la présence dans plusieurs langues et l'emploi de patrons. Des filtres complémentaires sont proposés pour améliorer la précision de ces méthodes. Plusieurs combinaisons de critères ont éte testées qui permettent d'améliorer aussi le rappel. Appliquées à deux terminologies médicales, cette méthode d'identification des noms propres permet d'atteindre des valeurs de rappel et de précision de l'ordre de 86 % pour une terminologie et de 98 % pour l'autre. Les implications stratégiques de ces résultats sont discutées.

Some specialized domains, such as medicine, make extensive use of proper names in their terminology. The present work studies criteria which can help to classify words as proper names in biomedical terminologies. Useful characteristics include word capitalization, the presence of words in several translations and patterns in terms. Since no individual criterion achieves both high precision and high recall, we propose a combination of five criteria that allowed us to reach 86% precision and recall on one terminology and 98% on another one. Strategic implications are discussed.

Mots Clefs - Keywords

Terminologies médicales, noms propres, corpus parallèles, mots invariants.

Medical terminologies, proper names, parallel corpora, invariant words.

*. U.S. National Library of Medicine, Bethesda, MD, USA.

**. DIAM - Service d'Informatique Médicale, DSI, AP-HP et Département de Biomathématiques, Université Paris 6, Paris, France.

INTRODUCTION

Les domaines spécialisés utilisent souvent une vaste terminologie technique. La Médecine fait partie de ces domaines. Les terminologies médicales ont été identifiées comme une source intéressante de vocabulaire (Baud R. *et al.* 1998). Alors que la plupart des termes médicaux ont la structure d'un syntagme nominal plus ou moins complexe (adjectifs, compléments déterminatifs), certains contiennent aussi un ou plusieurs noms propres. Dans ce travail, nous nous intéressons à l'identification des noms propres utilisés en Médecine à partir des terminologies médicales.

À notre connaissance, il n'existe aucun dictionnaire des noms propres utilisés en Médecine qui soit la fois complet, fiable et disponible sous forme électronique. Un index des noms propres, limité à quelques centaines d'entrées, figure généralement en appendice des ouvrages consacrés à la terminologie médicale (Chevallier J. 1995). Dans le *Unified Medical Language System* (NLM 1999) se trouve une table (LRPRP), dans laquelle sont recensés plus de 2000 noms propres, et une fonction du programme lvg qui permet de filtrer les noms propres. Les tests que nous avons effectués montrent que ni cette table, ni cette fonction ne permettent d'identifier de manière fiable les noms propres présents dans notre dictionnaire médical (absence de représentation des signes diacritiques, capitalisation parfois incorrecte, présence de noms communs ou d'expressions formées à partir de noms propres).

Notre but est donc de constituer une liste de noms susceptible de former le point de départ d'un dictionnaire des noms propres utilisés en Médecine. Ce travail complète une expérience précédente réalisée sur la Classification Internationale des Maladies (Bodenreider O. & Zweigenbaum P. 2000). Après avoir étudié les différentes caractéristiques des noms propres rencontrés dans plusieurs terminologies médicales, nous proposons différentes stratégies pour l'identification des noms propres adaptées aux caractéristiques de chaque terminologie médicale.

1. RECONNAÎTRE LES NOMS PROPRES DES TERMES MÉDICAUX

1.1. Les noms propres des termes médicaux

Dans le vocabulaire médical, un certain nombre de termes ont été forgés d'après un nom propre. Il s'agit le plus souvent du nom d'une personnalité scientifique (*« maladie de Parkinson »*), mais on trouve aussi des noms de lieux (*« fièvre de Lassa »*) ou des références historiques (*« syndrome de Münchhausen »*) ou mythologiques (*« tendon d'Achille »*). Les noms propres sont utilisés pour désigner des maladies dont la nature précise n'apparaît pas dans le nom (*« maladie de Parkinson »*), aussi bien que pour distinguer diverses formes d'une maladie particulière (*« aphasie de Wernicke »*, *« aphasie de Broca »*).

En dehors des maladies, la référence à un nom propre est souvent utilisée pour nommer les signes et symptômes (*« réflexe de Bainbridge »*, *« signe*

de Romberg »), les tests diagnostiques (*« test de Coombs »*, *« coloration de Gram »*), les éléments d'instrumentation médicale (*« valve de Starr-Edwards »*, *« broche de Kirschner »*), les actes médicaux diagnostiques ou thérapeutiques (*« opération de Billroth »*, *« manœuvre de Heimlich »*, *« indice d'Apgar »*) ou les micro-organismes (*« bacille de Hansen »*, *« virus d'Epstein-Barr »*). Finalement, c'est probablement dans le domaine anatomique que les noms propres ont été le plus largement utilisés (*« tubercule de Lisfranc »*, *« trompe de Fallope »*, *« ganglion de Gasser »*, *« cul de sac de Douglas »*, *« bronche de Nelson »*, *« polygone artériel de Willis »*), même si la nouvelle nomenclature anatomique (Nomina Anatomica) tend maintenant à en limiter l'usage (la *« trompe auditive »* remplace la *« trompe d'Eustache »*).

Les noms propres entrant dans la composition de termes médicaux sont le plus souvent isolés. Toutefois, certains termes associent plusieurs noms propres, généralement reliés par un trait d'union (*« sonde de Swan-Ganz »*, *« amyotrophie de Charcot-Marie-Tooth »*), plus rarement par la conjonction *« et »*, parfois utilisée pour mettre en évidence un nom propre composé au sein d'un éponyme (*« syndrome de Fitz-Hugh et Curtis »*, *« maladie de Pierre Marie et Sainton »*). Dans de rares cas, le prénom fait aussi partie du terme (*« signe pupillaire de Marcus Gunn »*). Enfin, certains noms propres comportent une particule, généralement séparée du nom principal par un espace ou une apostrophe (*« neurofibromatose de von Recklinghausen »*, *« syndrome de Van der Hoeve »*, *« thyroïdite de de Quervain »*), plus rarement accolée (*« point de McBurney »*). Une forme ultime d'intégration des noms propres à la terminologie médicale est représentée par leur transformation adjectivale (*« bloc suprahissien »*, relativement au nœud de His), la création d'un nom commun dérivé (*« pasteurellose »*, d'après Pasteur) ou encore leur latinisation dans le cas des noms de micro-organismes (*« Haemophilus ducreyi »*, d'après Ducrey).

1.2. Reconnaissance de noms propres

La reconnaissance des noms propres et, plus généralement, des entités nommées a fait l'objet de nombreux travaux depuis qu'elle a été introduite comme tâche dans les compétitions MUC (DARPA 1993; DARPA 1996). Cependant, les caractéristiques propres aux terminologies médicales rendent difficile l'utilisation de méthodes traditionnellement proposées pour identifier, catégoriser ou désambiguïser les noms propres.

(McDonald D. D. 1993) répartit en deux classes les indices utiles pour réaliser cette tâche. Les *indices internes* concernent les mots qui composent un nom propre ; ce sont par exemple des propriétés morphologiques (emploi de majuscules, suffixes spécifiques) ou la présence de classifieurs comme *« Monsieur »* ou *« Ltd. »*. Les *indices externes* proviennent du contexte dans lequel le nom propre apparaît ; ce sont par exemple les relations sémantiques que ces noms peuvent avoir avec d'autres éléments du texte.

Parmi les critères morphologiques, la *casse* (majuscules ou minuscules) est l'indice le plus évident pour déterminer si un mot est un nom propre. Elle

est utilisée dans toutes les approches. Cependant, deux facteurs réduisent sa prédictivité. D'une part, le premier mot d'une phrase commence normalement par une majuscule. Dans certaines terminologies, le premier mot de chaque terme commence effectivement par une majuscule, qu'il soit un nom propre ou pas. Une heuristique consiste à supposer qu'un mot qui possède au moins une occurrence en minuscules dans le corpus n'est pas un nom propre (Thielen C. 1995). (Mikheev A. 1999) pousse plus loin ce principe et recherche les occurrences non ambiguës de mots débutant par une majuscule. D'autre part, d'autres mots que les noms propres peuvent commencer par une majuscule ; c'est en particulier le cas des substantifs en allemand (Thielen C. 1995). Nous verrons que les noms de micro-organismes sont également dans ce cas.

Autre critère morphologique, la présence de *suffixes* particuliers peut aider à catégoriser certains types de noms propres. Ainsi, en anglais, les suffixes *« -corp »* ou *« -tec »* sont caractéristiques de noms d'entreprises (Gallippi A. F. 1996), et en allemand des suffixes du type *« -dorf »* ou *« -ingen »* sont révélateurs de noms de lieux (villes, etc.) (Thielen C. 1995). Cependant, les noms propres utilisés dans les termes médicaux sont le plus souvent des noms de personnes, présentant une importante variabilité morphologique.

La présence de noms propres dans une *liste* spécifiée est certes une façon pratique de les détecter : listes de noms d'entreprises, de noms courants de personnes (*« Smith »*, *« Michael »*), de lieux (Gallippi A. F. 1996; Wakao T. *et al.* 1996; Wacholder N. *et al.* 1997). C'est précisément une telle liste que nous cherchons à obtenir pour le domaine médical. À l'inverse, une liste « négative » permet d'exclure des mots capitalisés qui ne sont le plus souvent pas des noms propres (Wacholder N. *et al.* 1997).

Certains marqueurs lexicaux constituent des indices précieux de la présence d'un nom propre. Ce sont en général des *classifieurs* (quelquefois appelés « désignateurs ») qui précisent le type de ces noms propres ; par exemple, *« Ltd. »*, *« G.m.b.H. »* ou *« S.A. »* pour des entreprises, *« M. »* ou *« Mme »* pour des personnes, ainsi que l'emploi de titres comme *« Prof. »* ou *« Dr. »*. Ces marqueurs sont utilisés dans tous les travaux que nous avons cités. Cependant, même si des marqueurs comme *« maladie »*, *« syndrome »*, *« tumeur »* ou *« cellule »* sont parmi les plus fréquents, les exemples mentionnés précédemment laissent présager qu'ils ne permettront pas d'identifier l'ensemble des noms propres. En effet, plus de 150 marqueurs différents ont été identifiés dans la seule Classification Internationale des Maladies (OMS 1993), concernant essentiellement des termes diagnostiques (tableau 1). À l'inverse, le même nom propre peut être utilisé dans plusieurs termes différents (Wacholder N. *et al.* 1997) : *« fièvre de Lassa »*, *« virus de Lassa »*.

Des *patrons* (Gallippi A. F. 1996), ou grammaires de noms propres (Wakao T. *et al.* 1996), fournissent une méthode complémentaire pour définir ces noms propres de façon interne, en s'appuyant éventuellement sur des listes de noms propres élémentaires ou des marqueurs lexicaux. Un patron souvent employé décrit une séquence de noms propres reliés par les caractères *« - »*, *« , »* et *« & »* (ou en français *« et »*) (Gallippi A. F. 1996; Wakao T. *et al.* 1996).

Marqueur lexical	Fréquence	Marqueur lexical	Fréquence
maladie	505	ulcère	6
syndrome	352	anneau	5
tumeur	14	atrophie	5
paralysie	13	ostéochondrite	5
anémie	12	cellules	4
ostéochondrose	11	chancre	3
hernie	8	cirrhose	3
anomalie	7	complexe	3
dysenterie	7	dystrophie	3
fièvre	7	kyste	3
fracture	6	...	

Tableau 1 : Liste des marqueurs lexicaux les plus fréquemment rencontrés dans le contexte d'un nom propre dans l'index alphabétique de la Classification Internationale des Maladies.

Nous chercherons à mettre au point des patrons appropriés à la forme des termes médicaux.

Parmi les indices externes, le fait qu'un nom participe à un type d'événement particulier peut être révélateur du type de nom propre dont il s'agit. Par exemple, une occurrence dans le contexte *« – entered a venture »* est révélatrice d'un nom de société (Gallippi A. F. 1996), ainsi que *« – stocks »* (Wakao T. *et al.* 1996) ; plus simplement, une préposition locative comme *« bei »* (Thielen C. 1995) peut introduire un nom de lieu. (Wakao T. *et al.* 1996) montre que l'ajout de ces indices aux indices internes apporte une amélioration significative des performances. Ce type d'indice est hélas rare dans un terme, syntagme nominal comportant souvent seulement quelques mots.

Plus généralement, le co-texte (phrases adjacentes) peut être précieux pour la catégorisation des noms propres. Les liens de coréférence permettent par exemple de propager une catégorisation connue à un nom propre non encore catégorisé (Wakao T. *et al.* 1996; Wacholder N. *et al.* 1997). Pour ce qui nous concerne, l'environnement du nom propre étant limité au seul terme, aucun indice externe provenant de phrases adjacentes n'est utilisable.

La récurrence d'un mot ou d'une expression dans un corpus permet d'accumuler les informations sur ses propriétés (Thielen C. 1995). Notre source étant une terminologie plutôt qu'un corpus de textes, la plupart des noms propres sont des hapax (60 % pour la Classification Internationale des Maladies, 75 % dans le Répertoire d'anatomopathologie de la nomenclature SNOMED). Les méthodes fondées sur la récurrence des mots sont donc difficilement applicables.

La disponibilité d'une version parallèle du corpus dans une autre langue apporte un autre type d'indice que l'on peut ranger dans les indices externes. Ainsi, (Fung P. 1995) repère des noms propres et leur traduction par alignement de corpus parallèles anglais-chinois. L'étiquetage préalable des noms

propres de la partie anglaise du corpus permet de détecter des noms propres chinois qui leur correspondent. Certaines terminologies médicales sont effectivement disponibles en plusieurs langues (MeSH, CIM, SNOMED). Chaque concept a un identifiant unique commun aux différentes traductions, offrant ainsi une possibilité très simple d'alignement. Des stratégies exploitant ces corpus parallèles ont été définies pour constituer un dictionnaire multilingue (Baud R. *et al.* 1998). Comme les noms propres conservent souvent leur orthographe dans ces différentes traductions (ce n'est bien sûr pas une règle absolue ; nous y revenons plus bas), les noms propres figurant dans une traduction d'une terminologie ont de bonnes chances d'appartenir à la liste des mots communs à plusieurs traductions de la même terminologie : nous les appellerons *invariants*. Cette propriété devient alors un élément supplémentaire pouvant contribuer à l'identification des noms propres à partir de terminologies médicales disponibles en plusieurs langues.

2. MÉTHODES

Nous avons étudié différents critères permettant d'identifier les noms propres dans les terminologies médicales : casse des caractères, présence dans plusieurs langues, emploi de patrons. Comme aucun de ces critères ne permet à lui seul d'obtenir à la fois un bon rappel et une bonne précision, nous avons défini plusieurs combinaisons de critères permettant d'atteindre ce but. Afin d'évaluer la performance de chaque critère ainsi que celle des combinaisons, les mots d'une terminologie médicale identifiés comme noms propres par le critère ont été comparés à une liste de noms propres de référence pour cette terminologie. Nous avons utilisé pour cette évaluation les mesures standard de rappel et de précision.

2.1. Critères individuels

Nous avons étudié les critères suivants pour leur capacité à identifier un mot d'une terminologie médicale comme étant un nom propre [1].

2.1.1. Invariants (INV)

Les mots communs à plusieurs traductions d'une terminologie médicale, ou invariants, sont généralement soit des noms propres, soit des termes étrangers (en particulier, des mots latins). L'hypothèse à l'origine de ce critère est, d'une part, que la majorité des noms propres vont être invariants (c'est-à-dire que leur orthographe va rester la même dans les différentes traductions) et, d'autre part, que les invariants vont contenir une majorité de noms propres. Les libellés provenant des différentes traductions d'une terminologie médicale ont été segmentés par simple découpage sur les espaces, les apostrophes et la ponctuation. La liste des invariants est définie comme l'intersection des

1. Nous associons à chaque critère une abréviation, par exemple INV pour *invariants*, qui sera notée en petites majuscules.

mots provenant des différentes traductions d'une terminologie médicale.

Comme nous l'avons montré, les noms propres utilisés dans les termes médicaux peuvent être composés, notamment reliés par un trait d'union, et l'ordre de composition des noms propres pour un même terme peut varier d'une langue à l'autre (*« maladie de Legg-Perthes-Calvé »* ; *« Legg-Perthes-Calvé disease »* ; *« Perthes-Legg-Calvé-Krankheit »*). Pour cette raison, un découpage additionnel sur les traits d'union a été effectué avant de procéder à l'intersection des lexiques (l'intersection des trois termes cités ci-dessus contient les trois noms propres *« Calvé »*, *« Legg »*, et *« Perthes »*). Dans certaines langues comme l'allemand, le trait d'union est utilisé pour créer les éponymes, y compris à partir de noms propres simples (*« Parkinson-Syndrom »*), rendant le découpage sur le trait d'union nécessaire, indépendamment du problème de la composition des noms propres. Le trait d'union étant utilisé dans le vocabulaire médical (*« contre-indication »*, *« méningo-encéphalocèle »*), le découpage sur le trait d'union crée quelques fragments de mots qui n'ont pas d'autonomie propre (*« méningo »*). Ceci s'est avéré sans conséquence majeure sur la suite du traitement.

Selon la façon dont on le considère, ce critère peut être rangé dans les critères internes ou externes. Dans la mesure où il tient à une relation particulière entre le terme dans lequel il se trouve et un autre terme qui en est une traduction, on peut le considérer comme un critère externe. Mais comme on se fonde sur la forme du mot examiné (pour l'identifier à la forme d'un mot dans la terminologie traduite), sans réellement mettre en jeu de relation sémantique ou discursive, on peut aussi le voir comme un critère interne.

2.1.2. Capitalisation (CAP)

Il s'agit bien sûr d'un critère interne, de type morphologique. Les noms propres commencent généralement par une majuscule (à l'exception de ceux qui comportent une particule). Ceci fait de la casse un critère utile, au moins dans les terminologies médicales qui la respectent et l'utilisent avec constance. Toutefois, certains mots autres que des noms propres sont aussi capitalisés par convention. Certaines terminologies capitalisent aussi le premier mot de chaque terme, rendant inutilisable ce critère, au moins de manière isolée.

On peut, à l'inverse, noter que certains noms propres sont homographes de noms communs ou de verbes, différant seulement par la capitalisation (en français : *« pompe »* et *« maladie de Pompe »* ; en anglais : *« burns »* et *« Burns disease »*).

Les deux critères suivants ont trait à des formes particulières de capitalisation.

2.1.3. Symboles (N-S)

Bien que capitalisés, les acronymes (composés entièrement de majuscules, séparées ou non par des points) et les symboles (contenant des chiffres en plus des lettres) peuvent être aisément distingués des noms propres en rai-

son de leurs caractéristiques morphologiques spécifiques. On isole ainsi des noms propres des mots comme *« SIDA »*, *« [vitamine] B12 »* ou encore *« [caryotype] 46,XX »*.

De même, les chiffres romains, parfois suivis d'un indice, sont fréquemment utilisés pour désigner diverses formes d'une maladie (*« glycogénose de type VIII »*) ou les stades d'extension des cancers (*« lymphome stade IIe »*) et sont facilement détectables.

En outre, tout mot de deux caractères ou moins est considéré comme n'étant pas un nom propre. N-S est clairement un critère interne.

2.1.4. Micro-organismes (N-M)

Par convention, en biologie, les noms de famille et de genre sont capitalisés. Par exemple, la bactérie responsable de la fièvre typhoïde est *« Salmonella typhi »*, espèce (S. typhi) du genre Salmonella, appartenant lui-même à la famille des Enterobacteriaceæ.

Les micro-organismes peuvent toutefois être facilement identifiés en utilisant une liste de leurs noms. Il s'agit encore ici d'un critère interne. Une telle liste est parfois incluse dans la terminologie elle-même comme c'est le cas dans SNOMED (axe L des êtres vivants).

Une liste des noms de micro-organismes peut également être extraite de l'UMLS en prenant les concepts dont le type sémantique correspond à une classe d'organisme vivant (Bactéries, Champignons, Virus, Archéobactéries, Rickettsies/Chlamydiæ). À ces types sémantiques correspondent environ 9000 concepts comportant quelque 6000 mots différents.

Certains noms de micro-organismes contiennent un nom propre, le plus souvent de lieu (*« virus de Lassa »*, *« virus Ross River »*), parfois de personne (*« agent d'Eaton »* [mycoplasme], *« bacille de Pfeiffer »* [Haemophilus influenzæ]). Il n'est pas rare de retrouver au sein de noms de maladies des noms propres faisant partie du vocabulaire des micro-organismes (*« fièvre de Lassa »*, *« syndrome de Lambert-Eaton »*, *« maladie de Pfeiffer »*). Le choix d'une liste de référence de noms de micro-organismes adaptée à la terminologie à étudier doit minimiser ce recouvrement partiel pour éviter d'éliminer à tort des noms propres en filtrant les mots présents dans les noms de micro-organismes.

2.1.5. Patrons (PAT)

L'observation des contextes gauches des noms propres montre leur extrême diversité. Nous avons donc utilisé un patron très lâche qui se fonde uniquement sur l'emploi comme marqueur lexical de la préposition *« de »*. Il repère la première unité atomique suivant la forme *« de »* (*« système veineux de Batson »*) ou *« d' »* (*« tumeur d'Abrikossof maligne »*). Notons que dans le patron utilisé ici, nous n'imposons pas de contrainte sur le fait que cette unité débute par une majuscule. Cette contrainte est gérée indépendamment (voir les paragraphes 2.1.2, 2.1.3 et 2.1.4), le principe de ces expériences étant

d'étudier l'effet spécifique de chaque propriété avant de les combiner. Dans la mesure où nous nous focalisons sur les noms propres, plutôt que sur les expressions entières qui les contiennent, on peut considérer que nous employons ici un critère externe : le fait pour un nom propre d'être relié à ce qui précède par la préposition *« de »*.

Pour exploiter pleinement les constructions coordonnées (*« cathéter de Swan-Ganz »*, *« carcinome intra-épidermique de Borst et Jadassohn »*, *« syndrome de Rotor, Manahan et Florentin »*), si la forme trouvée est suivie d'un tiret, d'un *« et »* ou d'une virgule, on repère également l'unité suivant ce marqueur supplémentaire. Ce principe se retrouve dans la reconnaissance de groupes de noms propres dans les noms d'entreprises (Gallippi A. F. 1996; Wakao T. *et al.* 1996). Le patron résultant est le suivant [2] :

(1) `(de |d')<atome>(( et |-|, |, et )<atome>)*`
avec :
`<atome> := [a-zA-Z]+` *(séquence de lettres éventuellement accentuées)*

Il repère, à la suite de la préposition *« de »*, une unité atomique ou un groupe coordonné d'unités atomiques.

Il intègre maintenant également des critères que l'on peut considérer comme internes à un groupe de noms propres. Dans un tel groupe, chaque unité atomique `<atome>` est un candidat nom propre (pour les exemples de termes ci-dessus, on obtient la liste *« Batson »*, *« Abrikossof »*, *« Swan »*, *« Ganz »*, *« Borst »*, *« Jadassohn »*, *« Rotor »*, *« Manahan »*, *« Florentin »*).

Lorsqu'un nom propre comprend deux unités atomiques (*« maladie de von Gierke »*, *« naevus bleu de Max Tiesch »*), ce patron ne repère en revanche que la première unité. Une version étendue de ce patron (PAT2), à défaut de groupe coordonné, collecte les deux unités qui suivent le marqueur *« de »* (ou une seule en cas de fin de terme) :

(2) `(de |d')<atome>((( et |-|, |, et )<atome>)+| <atome>)?`

Elle permet ainsi de traiter correctement les cas ci-dessus, au risque d'un bruit très important.

Un patron plus sophistiqué peut être envisagé pour décrire de manière plus précise les noms propres composés formés d'atomes multiples, afin de limiter le bruit identifié avec le patron précédent :

(3) `(de |d')<npc>((, <npc>)* et <npc>)?`
avec :
`<npc> := <npr>(-<npr>)*`
`<npr> := ((van der|de|von|van|di|da) )?<atome>( <atome>)?`

2. Ces patrons sont exprimés par des expressions régulières, que nous présentons en suivant des conventions habituelles : la disjonction A ou B est notée par $(A \mid B)$; la répétition de $0-n$ occurrences de A est notée par A^* (symbole de Kleene) ; l'optionalité (répétition $0-1$) par $A?$; le groupement de plusieurs symboles par une paire de parenthèses, comme dans la séquence optionnelle $(ABC)?$; enfi n, pour faciliter la lecture, nous avons noté entre chevrons des sous-expressions (par exemple, $<atome>$) qui sont défi nies plus bas.

Ce patron est destiné à être appliqué après suppression des parenthèses, sans segmentation préalable des termes en unités atomiques. Un élément de nom propre (<npr>) est constitué d'un ou deux atomes (prenant en compte les cas où le prénom est mentionné), éventuellement précédé d'une particule. Dans un nom propre composé (<npc>), les éléments sont reliés par un trait d'union. Finalement, les noms propres composés peuvent être coordonnés par une virgule, remplacée par *« et »* pour la dernière occurrence. Le principal intérêt de ce patron est de caractériser plus précisément la forme « interne » d'un nom propre, avec possibilité de particule et prénom, et ce pour chaque nom d'un groupe coordonné. Il permet ainsi d'extraire des groupes comme *« Claude Bernard-Horner »*, *« Bruck-de Lange »* ou *« Cécile et Oscar Vogt »*[3].

2.2. Combinaisons de critères

Seule une combinaison des critères présentés plus haut permet d'obtenir à la fois une précision et un rappel suffisants. La combinaison optimale de ces critères binaires permettant d'identifier un mot quelconque comme étant un nom propre peut être déterminée à l'aide de techniques statistiques. Néanmoins, nous allons montrer que la connaissance du domaine permet déjà d'arriver à des résultats satisfaisants.

Les trois principaux critères sont la capitalisation, les invariants et les patrons :

– La capitalisation est un critère insuffisamment précis pour analyser les terminologies capitalisant le premier mot de chaque terme, par convention typographique.

– Les invariants aussi ont une précision nécessairement insuffisante, puisqu'ils contiennent des emprunts en plus des noms propres. On trouve en particulier dans les invariants de nombreux mots latins ainsi que les noms de micro-organismes.

– Les patrons ont un rappel variable selon l'homogénéité de la terminologie étudiée. Leur précision est généralement médiocre.

Afin d'améliorer la précision des deux derniers critères, nous allons leur associer des critères internes vus plus haut, basés sur la capitalisation (CAP), les caractéristiques morphologiques des acronymes et des symboles (N-S), ou encore sur des ressources supplémentaires telles qu'une liste de noms de micro-organismes (N-M). Les critères N-S et N-M sont définis comme des filtres permettant d'éliminer les mots qui présentent ces caractéristiques.

On définit ainsi trois associations de critères, présentées sur les trois premières rangées du tableau 2. L'association C-CAP basée sur la capitalisation représente une sorte de référence. Les autres associations C-INV et C-PAT

3. Il se trouve néanmoins que les deux premiers sont mal orthographiés dans la Classifi cation Internationale des Maladies (*« Claude-Bernard-Horner »*, *« Bruck-de-Lange »*) et que le troisième en est absent.

(C-CAP$_i$)	CAP$_i$ et N-S et N-M
(C-INV$_j$)	INV$_j$ et CAP et N-S et N-M
(C-PAT$_k$)	PAT$_k$ et CAP et N-S et N-M
(COMBI$_n$)	(INV ou PAT) et CAP et N-S et N-M

Tableau 2 : Combinaison des critères individuels.

utilisent C-CAP pour améliorer la précision de INV et de PAT. Nous étudierons différentes variantes CAP$_i$, INV$_j$ et PAT$_k$ qui donneront respectivement les associations C-CAP$_i$, C-INV$_j$ et C-PAT$_k$.

Néanmoins, si les associations de critères par un ET logique permettent d'améliorer la précision, une combinaison de plusieurs de ces associations par un OU logique est nécessaire pour améliorer le rappel. La combinaison que nous avons établie (COMBI) peut être définie comme la réunion des ensembles de noms propres identifiés par deux méthodes, l'une basée sur les invariants (C-INV) et l'autre basée sur les patrons (C-PAT). COMBI réunit les noms propres identifiés à partir des invariants ou à partir du patron, et en élimine les mots non capitalisés, les acronymes et les symboles, ainsi que les noms de micro-organismes. Ici encore, plusieurs variantes COMBI$_n$ seront examinées.

3. CLASSIFICATION INTERNATIONALE DES MALADIES

3.1. Matériel

La Classification Internationale des Maladies (CIM) a été créée au XIX^{e} siècle par Jacques Bertillon dans l'objectif de répertorier les causes de décès de manière uniforme au niveau international. La première révision comptait 161 rubriques. L'Organisation Mondiale de la Santé en assure la maintenance depuis 1948. La plus récente révision (CIM-10) a été publiée en 1992 pour la version anglaise. Une vingtaine de traductions sont actuellement disponibles. La CIM-10 est utilisée dans un grand nombre de pays pour établir les statistiques sanitaires (mortalité, morbidité) et, plus généralement, pour répertorier les maladies, les traumatismes et les problèmes de santé.

La CIM-10 est constituée de plusieurs volumes et comprend en particulier un volume analytique classant environ 17 000 maladies, complété par un index alphabétique comportant 42 000 entrées dans la version française (OMS 1993). Les libellés du volume analytique sont généralement des syntagmes nominaux d'une complexité et d'une longueur variables (*« Céphalée »*, *« Sarcome de Kaposi de la peau »*, *« Intoxication par médicaments agissant essentiellement sur le système nerveux autonome »*). La structure syntaxique des entrées de l'index alphabétique est par contre moins bien définie. Par rapport au contexte mis en place dans un terme mentionné plus haut dans l'index, seuls les modificateurs du terme fils sont mentionnés, les autres mots hérités du terme père étant matérialisés par des tirets qui représentent la profondeur hiérarchique du terme. Par exemple, le terme *« – chronique »* signifie en fait *« gastrite chronique »* lorsqu'il apparaît en dessous du terme *« Gastrite »*.

L'index alphabétique représente la principale source de noms propres dans la CIM-10. L'index reflète l'usage des termes médicaux dans une langue donnée, et particulièrement des éponymes, tant sur le plan linguistique que sur le plan culturel. À l'inverse, le volume analytique est constitué d'une liste de termes dont la structure est assez similaire d'une langue à l'autre (*« Sarcome de Kaposi de la peau »*, *« Kaposi's sarcoma of skin »*, *« Kaposi-Sarkom der Haut »*). Pour cette raison, les noms propres présents dans l'index alphabétique de la CIM-10 dans une traduction ne sont pas nécessairement retrouvés dans les autres traductions. Finalement, les noms propres composés apparaissant dans plusieurs traductions présentent parfois des variations non seulement dans l'ordre, mais aussi dans le nombre de noms propres utilisés (la panniculite récidivante est appelée *« maladie de Weber-Christian »* en français, *« Weber-Christian disease »* en anglais, mais *« Pfeifer-Weber-Christian-Krankheit »* en allemand[4]).

3.2. Étalon

Pour établir la liste de référence des noms propres que nous avons utilisée comme étalon, nous avons extrait le vocabulaire de la Classification Internationale des Maladies (dixième révision, volume 1 et index alphabétique). Ce lexique a été comparé à un dictionnaire général du français, le DELAF (80 000 entrées), enrichi de termes du vocabulaire médical provenant d'autres terminologies médicales. Les mots inconnus du dictionnaire ont été examinés manuellement et étiquetés. 1222 noms propres ont ainsi été identifiés, constituant notre étalon pour les travaux sur la Classification Internationale des Maladies.

3.3. Précisions sur les critères employés

Nous précisions ici la définition employée pour les critères dans leur application à la Classification Internationale des Maladies, lorsqu'elle est différente de la définition présentée au paragraphe 2.

3.3.1. Invariants (INV)

La méthode INV effectue l'intersection des lexiques (unités atomiques hors ponctuation et traits d'union) provenant des versions anglaise (EN), française (FR) et allemande (GE) de la Classification Internationale des Maladies (volume 1 et index alphabétique), représentant respectivement 17 739, 16 049 et 23 394 formes distinctes.

Comme seuls les termes provenant du volume 1 peuvent faire l'objet d'un alignement un à un sur le code, nous avons choisi de réaliser l'intersection des lexiques sans alignement.

Pour tenir compte des différences d'orthographe de certains noms propres entre les différentes traductions, nous avons défini deux types d'invariants. Le premier type (INV3) inclut un mot dans les invariants seulement s'il

4. Soit *« maladie de Pfeifer-Weber-Christian »*.

est présent dans chacune des traductions (EN et FR et GE). Le deuxième type (INV2) est plus permissif et requiert seulement que le mot soit présent dans au moins deux traductions ([EN et FR] ou [FR et GE]). INV2 devrait donc avoir un meilleur rappel et INV3 une meilleure précision.

3.3.2. Capitalisation (CAP)

Par convention, le premier mot de chaque terme de la Classification Internationale des Maladies commence par une majuscule, ce qui donne au critère capitalisation défini au paragraphe *Méthodes* une précision médiocre.

Nous définissons donc un critère de capitalisation plus précis CAP2 à la façon de (Thielen C. 1995). Ce critère est vrai exclusivement pour les mots qui sont *toujours* capitalisés dans le corpus. Ainsi, bien que le mot *« choléra »* soit souvent orthographié avec un *« C »* majuscule (en début de terme), il existe des cas où il est entièrement en minuscules (exemple (4)). Le critère CAP2 est donc négatif pour ce mot, ce qui est correct. On peut le considérer comme une version simplifiée de la méthode proposée par (Mikheev A. 1999).

(4) *Choléra*
Choléra à Vibrio cholerae 01, biovar cholerae
Choléra à Vibrio cholerae 01, biovar El Tor
Choléra, sans précision
[...]
Nécessité d'une vaccination contre le choléra seul
Nécessité d'une vaccination contre le choléra et la typhoïde-paratyphoïde (choléra+TAB)

Cette variante CAP2 va certes faire perdre quelques homographes. En effet, l'exemple (5) montre que coexistent dans la CIM-10 à la fois un nom propre *« Pompe »* et le nom commun *« pompe »*.

(5) *Maladie de Pompe*
Ajustement et entretien d'une pompe à perfusion

Elle va néanmoins permettre d'éliminer de la liste des candidats un grand nombre de mots dont la capitalisation n'est liée qu'à leur position en début de terme[5].

5. Une autre possibilité, suggérée par l'un des relecteurs, serait de ne pas tenir compte du premier mot de chaque terme. Si l'on utilise le patron ci-dessous (paragraphe 3.3.4, (6)), cette méthode est correcte pour le volume analytique. En revanche, elle n'est pas applicable à l'index alphabétique, dans lequel les noms propres sont souvent placés en premier. De plus, cela introduirait dans CAP un critère positionnel comparable à un patron. Nous préférons ici pouvoir tester individuellement chaque type de critère, quitte à les combiner ensuite. C'est d'ailleurs aussi la stratégie adoptée par (Mikheev A. 1999).

3.3.3. Micro-organismes (N-M)

La liste de référence pour les noms de micro-organismes comporte 12 979 formes différentes et provient de la réunion de plusieurs listes :

– Les noms de Bactéries, Champignons, Virus, Archéobactéries, Rickettsies/Chlamydiæ mentionnées dans l'UMLS, extraits en faisant une recherche sur leur type sémantique ;

– Diverses classifications de micro-organismes et listes de cultures disponibles dans le domaine public.

3.3.4. Patrons (PAT)

Le patron discuté au paragraphe 2 a été adapté aux conventions d'écriture particulières à la Classification Internationale des Maladies. Il est appliqué sans modification aux termes provenant du volume 1 (*« maladie de Parkinson »*) :

(6) `<contexte> (de |d')<npc>((, <npc>)* et <npc>)?`

Pour la plupart des termes contenant un nom propre, dans l'index, le nom propre figure au début du terme, séparé du contexte par une virgule (*« Cooley, anémie ou maladie de »*). La préposition *« de »* reste le marqueur lexical de référence, mais elle définit la borne droite du contexte. Une version modifiée du patron a été créée pour tenir compte de cette particularité :

(7) `<npc>((, <npc>)* et <npc>)?, <contexte> de`

Le patron est limité à une seule ligne et ne prend donc pas en compte les termes définis sur plusieurs lignes comme c'est souvent le cas dans l'index alphabétique. En pratique, les lignes commençant par un tiret sont exclues du traitement par le patron. Un nom propre dont l'occurrence unique figure sur une de ces lignes ne sera donc pas identifié par le patron [6] .

3.4. Résultats

Les valeurs de précision et de rappel ont été calculées pour les critères individuels puis combinés sur les 16 401 mots extraits de la Classification Internationale des Maladies, comparés à la liste de référence des noms propres.

Le tableau 3 résume les valeurs de précision et de rappel calculées pour chacun des critères individuels définis au paragraphe 2 et pour leurs variantes

6. On pourrait chercher à remplacer les tirets initiaux dans ces lignes « indentées » par le terme « racine », situé au niveau d'indentation précédent : ce terme fournit le contexte dans lequel interpréter ces lignes. Cela donne cependant des concaténations de morceaux de termes dont le traitement automatique est malaisé, sauf dans les cas les plus simples. De plus, le libellé « racine » est formulé de telle sorte qu'il puisse accommoder tous ses descendants, avec des parenthèses pour mentionner ce qui est optionnel.

Critère	INV2	INV3	CAP	CAP2	N-S	N-M	PAT
Précision	0,289	0,473	0,192	0,276	0,075	0,073	0,482
Rappel	0,874	0,713	0,999	0,993	1,000	0,950	0,837

Tableau 3 : CIM-10 : Performance des critères individuels.

du paragraphe 3.3. Le tableau 4 résume les valeurs de précision et de rappel calculées pour chacune des combinaisons de critères définies au paragraphe 2 et pour leurs variantes. COMBIn fait l'union de C-PAT et C-INVn. Les valeurs de précision et de rappel des principaux critères sont représentées sur la figure 1.

Critère	C-INV2	C-INV3	C-PAT	C-CAP2	COMBI2	COMBI3
Précision	0,663	0,827	0,991	0,280	0,683	0,859
Rappel	0,822	0,666	0,807	0,944	0,909	0,881

Tableau 4 : CIM-10 : Performance des combinaisons de critères.

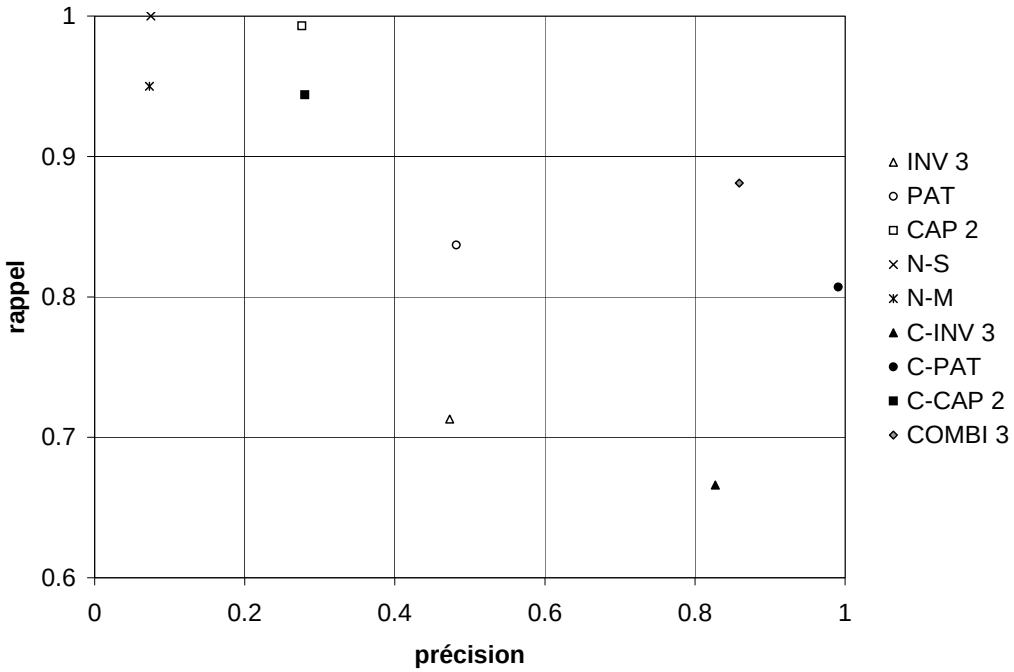

Figure 1 : Performance des principaux critères appliqués à la CIM-10.

4. RÉPERTOIRE D'ANATOMOPATHOLOGIE DE LA NOMENCLATURE SNOMED

4.1. Matériel

La Nomenclature systématique de la médecine humaine et vétérinaire, ou SNOMED Internationale (Côté R. A. *et al.* 1993), se divise en onze *modules* ou *axes sémantiques* : anatomie, atteintes morphologiques, fonctions, maladies et diagnostics, etc. La version originale de la SNOMED est en anglais, et contient plus de 150 000 termes. Des traductions sont en cours ou terminées dans plusieurs langues. Pour ce qui concerne le français, seul le Répertoire d'anatomopathologie, sous-ensemble spécialisé de la SNOMED, est actuellement disponible (Côté R. A. 1996). Il contient 12 555 termes répartis dans les onze modules.

Nous avons utilisé deux sources différentes pour la version anglaise. La première source est l'UMLS (NLM 1999), qui contient 128 855 termes de la SNOMED Internationale (version 3.5, 1998, en anglais), représentant 89 343 concepts différents. Cependant, 2221 des 9098 concepts différents présents dans le Répertoire d'anatomopathologie en français n'apparaissent pas dans cet ensemble de concepts.

La seconde source est la traduction russe, en cours, du Répertoire d'anatomopathologie anglais (Emelin I. V. *et al.* 1995). Elle se compose d'une table parallèle des termes anglais et russes du Répertoire d'anatomopathologie. Elle contient exactement 9098 concepts différents (pour 12 462 termes), qui recouvrent à 95 % ceux du Répertoire français. Le parallélisme entre Répertoires spécialisés d'anatomopathologie est donc meilleur que celui entre Répertoire d'anatomopathologie et SNOMED générale. Cependant, comme dans la CIM-10, les éponymes et leur orthographe peuvent varier d'une langue à l'autre. On trouve par exemple en français *« sarcome de Kaposi »* et le terme synonyme *« sarcome hémorragique multiple »*, et en anglais *« Kaposi's sarcoma »* et *« Multiple hemorrhagic sarcoma »* ; mais aussi *« syndrome de Sjögren »* et *« syndrome de Gougerot-Sjögren »*, et en anglais *« Sicca syndrome »*, *« Sjogren's disease »* et *« Sjogren's syndrome »*.

Les termes de la SNOMED sont de façon générale des syntagmes nominaux (sans déterminant initial). Les termes du Répertoire d'anatomopathologie français comptent de une à vingt unités atomiques (*tokens*, ponctuation comprise), avec une moyenne à 2,8 unités. Ces syntagmes nominaux sont en général bien construits. Hors ponctuations, le Répertoire d'anatomopathologie contient 7251 formes atomiques différentes.

4.2. Étalon

Dans un travail antérieur (Grabar N. & Zweigenbaum P. 2000), nous avons étiqueté les termes du Répertoire d'anatomopathologie [7]. Nous avions

7. Nous avons pour cela utilisé l'étiqueteur de Brill entraîné à l'INaLF, puis nous l'avons réentraîné sur des portions corrigées du thesaurus. Nous avons enfi n corrigé manuellement

ainsi à notre disposition un recensement des noms propres qui apparaissent dans ce thesaurus. Notons que les micro-organismes (voir section 2.1.4) y sont bien étiquetés comme des noms communs. Les 339 noms propres ainsi collectés constituent notre étalon pour les travaux sur le Répertoire d'anatomopathologie.

4.3. Précisions sur les critères employés

Nous précisons ici la définition employée pour les critères dans leur application au Répertoire d'anatomopathologie, lorsqu'elle est différente de la définition présentée au paragraphe 2.

4.3.1. Invariants (INV)

La méthode INV effectue l'intersection des vocabulaires (unités atomiques hors ponctuations) du Répertoire d'anatomopathologie français (7251 formes) et d'une version anglaise correspondante. Cette version anglaise est :

– soit la version extraite du Répertoire d'anatomopathologie russe (8860 formes [ren]),

– soit la SNOMED générale extraite de l'UMLS (63 856 formes [en]).

Par ailleurs, nous avons testé une version alignée de la méthode d'intersection des vocabulaires (INV_a). Dans cette variante, l'intersection se fait concept à concept : pour chaque concept, l'intersection des formes apparaissant dans les termes français et anglais de ce concept est réalisée séparément. On suppose en effet que si un terme contient un nom propre et que ce nom propre est employé dans la langue cible, il devrait apparaître dans la traduction de ce terme. Cette variante identifie nécessairement moins d'invariants et pourrait améliorer la précision des résultats. Il ne faudrait cependant pas qu'elle diminue le rappel.

Au total, on distingue donc quatre variantes, selon que l'on emploie la version simple (INV) ou alignée (INV_a) sur la SNOMED générale (INV^{en}, INV_a^{en} : «en» pour anglais) ou sur le Répertoire anglais obtenu du russe (INV^{ren}, INV_a^{ren} : «ren» pour russe - anglais) ; le Répertoire français, lui, reste constant.

4.3.2. Micro-organismes (N-M)

La liste des micro-organismes employée ici a été dérivée de l'axe L (organismes vivants) du Répertoire d'anatomopathologie. Il s'agit des mots capitalisés des termes de cet axe, après correction manuelle : dix noms propres s'y trouvaient, et deux noms de micro-organismes employés dans le reste du Répertoire n'y figuraient pas (ils figuraient en revanche dans les termes de l'axe L de la SNOMED générale). La liste des micro-organismes contient 275 formes.

---

l'ensemble de cet étiquetage.

4.3.3. Patron (PAT)

Les résultats pour les patrons PAT (paragraphe 2.1.5, exemple (1)) et PAT2 (paragraphe 2.1.5, exemple (2)) sont présentés. Le patron de l'exemple (3) donne ici les mêmes résultats que PAT2.

4.4. Résultats

Les valeurs de précision et de rappel ont été calculées en appliquant aux 7251 formes du Répertoire d'anatomopathologie les critères individuels puis combinés et en comparant les résultats obtenus aux 339 noms propres de notre étalon.

Le tableau 5 résume les valeurs de précision et de rappel calculées pour chacun des critères individuels définis au paragraphe 2 et pour leurs variantes.

Critère	INV^{ren}	INV^{en}	INV$_a^{ren}$	CAP	N-S	N-M	PAT	PAT2
Précision	0,213	0,165	0,229	0,473	0,048	0,049	0,418	0,302
Rappel	0,737	0,732	0,735	1,000	1,000	1,000	0,917	0,950

Tableau 5 : SNOMED : Performance des critères individuels.

Le tableau 6 résume les valeurs de précision et de rappel calculées pour chacune des combinaisons de critères définies au paragraphe 2 et pour leurs variantes du paragraphe 4.3. Les valeurs de précision et de rappel des principaux critères sont représentées sur la figure 2.

Critère	C-INV^{ren}	C-INV^{en}	C-INV$_a^{ren}$	C-PAT	C-PAT2	C-CAP
Précision	0,969	0,980	0,969	0,997	0,997	0,971
Rappel	0,737	0,732	0,735	0,917	0,950	1,000

Critère	COMBI^{ren}	COMBI^{en}	COMBI$_a^{ren}$	COMBI^{ren}2	COMBI^{en}2	COMBI$_a^{ren}$2
Précision	0,976	0,985	0,976	0,976	0,985	0,976
Rappel	0,962	0,965	0,962	0,971	0,971	0,971

Tableau 6 : SNOMED : Performance des combinaisons de critères.

5. SYNTHÈSE ET DISCUSSION

5.1. Insuffisance de la casse seule

Les expériences menées montrent que l'application du critère de casse (CAP), s'il possède un rappel proche de 1, obtient une précision médiocre (moins de 0,2 pour la CIM, et moins de 0,5 pour la SNOMED). Cela tient, d'une part, au fait que les noms de micro-organismes sont capitalisés et, d'autre part, pour la CIM, au fait que le premier mot d'un terme est capitalisé. Les symboles et sigles, en proportion moindre, entrent pour une plus faible part dans cette médiocre précision.

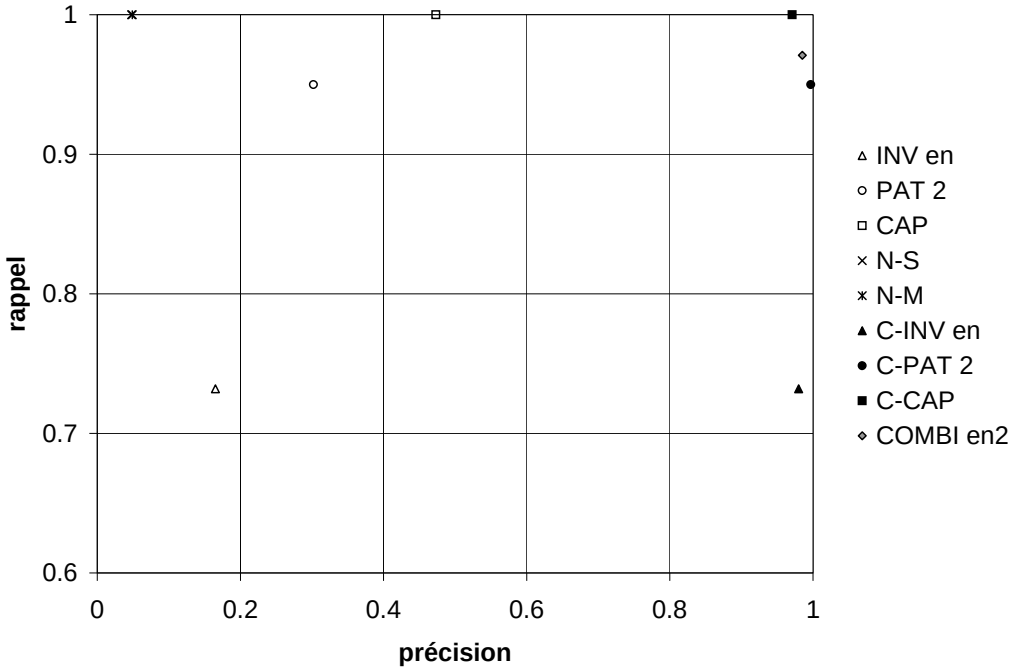

Figure 2 : Performance des principaux critères appliqués au Répertoire d'anatomopathologie de la nomenclature SNOMED.

Une première correction consiste à filtrer les « symboles » (N-S) et les noms de micro-organismes (N-M) : c'est la combinaison C-CAP.

Pour la SNOMED, comme ces deux filtres ont un rappel de 1, le rappel reste constant à 1, comme pour CAP ; la précision, elle, monte à 0,971. C-CAP est ainsi une excellente méthode d'extraction de noms propres pour le Répertoire d'anatomopathologie de la SNOMED, dans lequel on trouve peu de mots capitalisés en dehors des noms propres, des symboles et des micro-organismes.

Pour la CIM, du fait que N-M a un rappel (0,950) inférieur à celui de CAP, la combinaison C-CAP voit également le sien baisser (0,944) ; sa précision (0,280) s'améliore un peu, mais reste faible. Cette médiocre précision est due au fait que la proportion de mots capitalisés en raison de leur présence en début de terme est très importante (6368 formes sont capitalisées dont 4499 n'apparaissent que sous cette forme), en comparaison du nombre de noms propres (1222), symboles (73) et micro-organismes (557).

Cette combinaison de critères peut être considérée comme une méthode de base, fondée exclusivement sur des critères internes positifs (forme des mots) ou négatifs (exclusion de noms de micro-organismes).

Les performances de cette méthode dépendent d'une part des conventions suivies dans la capitalisation des termes. Elles dépendent d'autre part de la complétude de la liste de noms de micro-organismes utilisée.

5.2. Propriétés des critères complémentaires

Deux critères complémentaires sont ensuite étudiés : invariants et patrons. Sur la CIM comme sur la SNOMED, pour une précision (médiocre) comparable à celle des invariants (0,2 à 0,5), les patrons obtiennent un rappel bien meilleur (0,837 contre 0,713 pour la CIM, 0,950 contre 0,73 pour la SNOMED). Ce rappel reste en-deçà de celui de C-CAP, d'où l'intérêt de filtrer les résultats de ces méthodes par la combinaison C-CAP : ce sont les méthodes C-INV et C-PAT.

Ce filtrage augmente très nettement la précision de INV et PAT, pour la CIM comme pour la SNOMED, C-PAT obtenant une précision proche de 1. Leur rappel décroît très peu (CIM) ou reste constant (SNOMED) selon le rappel de C-CAP.

C-PAT obtient ainsi un rappel bon (CIM, 0,807) à excellent (SNOMED, 0,950). C-PAT se positionne donc comme une méthode de précision maximale pour un rappel honorable, face à la méthode de base C-CAP dont le rappel est très élevé mais la précision médiocre pour la CIM. C-INV, de son côté, reste en-deçà de C-PAT.

5.3. Propriétés des méthodes combinées

En utilisant en parallèle les résultats de C-INV et C-PAT, on propose un autre compromis entre rappel et précision (COMBI) : un rappel situé entre C-PAT et C-CAP et une précision elle aussi intermédiaire.

Le fait que le rappel de COMBI soit supérieur à celui des méthodes C-INV et C-PAT dont il réunit les résultats indique la complémentarité de ces deux méthodes :

– L'efficacité des patrons repose sur la régularité de formation et de présentation des termes ;

– Le rappel des invariants est sans doute corrélé à la proximité des langues et surtout de leur transcription des noms étrangers[8]. La précision, au contraire, peut être favorisée par la distance entre langues.

Le graphique comparatif entre CIM et SNOMED montre une constance du phénomène pour ces deux terminologies (Figure 3). Concernant la différence de rappel entre la combinaison et la meilleure des deux méthodes, le bénéfice semble plus important pour la CIM.

5.4. Variantes de la méthode des invariants

5.4.1. INV3 et INV2

Pour tenir compte des différences d'orthographe de certains noms propres entre les différentes traductions de la CIM-10, nous avons défini deux

8. Il dépend toutefois aussi des modalités de représentation des caractères dans la version électronique des terminologies (par exemple, de la présence ou non des signes diacritiques).

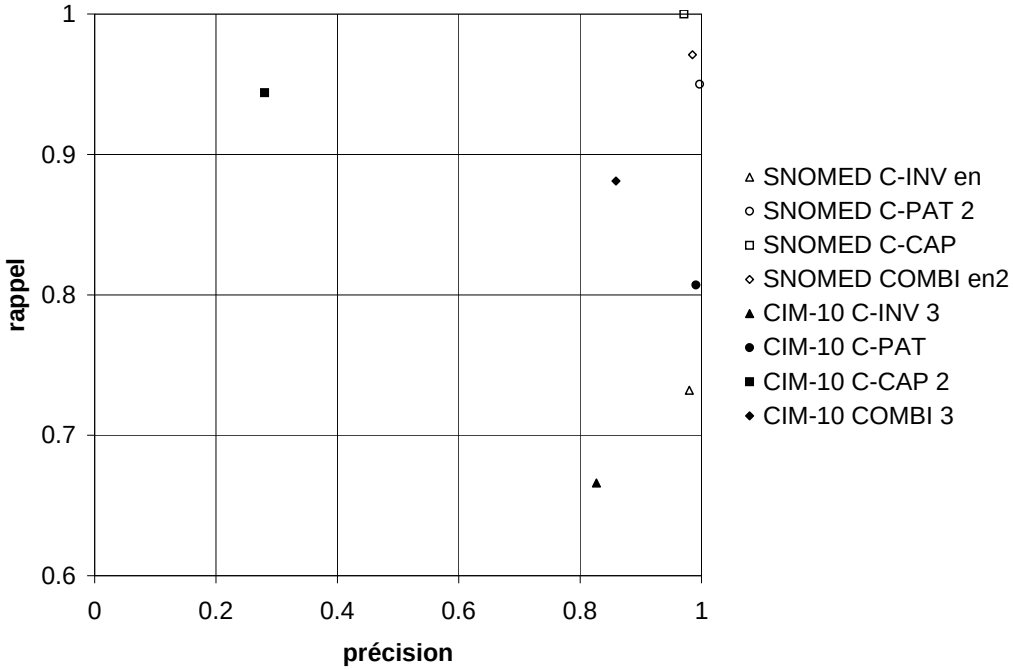

Figure 3 : Comparaison de la performance des principaux critères appliqués au Répertoire d'anatomopathologie de la nomenclature SNOMED et à la CIM-10.

types d'invariants : INV3 inclut un mot dans les invariants seulement s'il est présent dans chacune des traductions (EN et FR et GE), alors que INV2 requiert seulement que le mot soit présent dans au moins deux traductions ([EN et FR] ou [FR et GE]). INV2 retrouve environ deux fois plus d'invariants que INV3, mais seulement 15 % de noms propres en plus. On pourra donc utiliser INV2 pour privilégier le rappel et INV3 pour privilégier la précision.

5.4.2. Alignement des invariants

La variante des invariants réalisant un alignement sur les codes (INV_a^{ren}) ne modifie pas significativement les performances. Ceci est probablement lié à la proximité des traductions (le rappel n'est pas diminué) et au fait que les invariants de SNOMED contiennent une importante proportion de mots qui ne sont pas des noms propres (le gain de précision est minime).

5.5. Typologie des terminologies

Parmi les critères influençant la possibilité d'identifier les noms propres au sein des termes médicaux, on peut retenir les caractéristiques suivantes des terminologies médicales :

– Proportion de formes capitalisées (liée, en pratique, à la capitalisation du premier mot de chaque terme) ;

– Homogénéité (régularité de formation et de présentation des termes) ;

– Disponibilité en plusieurs langues ;

– Complétude des différentes traductions (conditionnant la possibilité d'un alignement sur les codes) ;

– Proximité des différentes traductions (conditionnant la possibilité d'un alignement sur les mots) ;

– Disponibilité des signes diacritiques dans les différentes traductions ;

– Disponibilité de ressources complémentaires intégrées à la terminologie (comme une liste de micro-organismes).

Le tableau 7 présente un classement en fonction de ces caractéristiques du Répertoire d'anatomopathologie de la SNOMED, de la CIM-10 et du MeSH (Medical Subject Headings), un ensemble de 20 000 termes utilisé par la Bibliothèque Nationale de Médecine des États-Unis pour indexer les articles de la base de donnée bibliographique MEDLINE.

Caractéristiques	SNOMED	CIM-10	MeSH
Proportion de formes capitalisées	Faible	Importante	Toutes
Homogénéité	Bonne	Moyenne	Faible
Disponibilité en plusieurs langues	Oui	Oui	Oui
Complétude des différentes traductions	Bonne	Moyenne	Bonne
Proximité des différentes traductions	Bonne	Moyenne	Bonne
Disponibilité des signes diacritiques	Oui	Oui	Non
Liste intégrée de micro-organismes	Oui	Non	Oui

Tableau 7 : Comparaison des caractéristiques du Répertoire d'anatomopathologie de la SNOMED, de la CIM-10 et du MeSH.

5.6. Considérations stratégiques

En fonction de l'objectif poursuivi, différentes méthodes pourront être utilisées. Dans les activités de repérage des noms propres, c'est le rappel qui est privilégié et les méthodes de choix sont celles basées sur la casse (C-CAP). Pour le filtrage, à l'inverse, on privilégie la précision en choisissant de préférence une association basée sur les invariants (C-INV) ou sur les patrons (C-PAT), selon les caractéristiques de la terminologie. Enfin, la combinaison des différentes méthodes (COMBI) représente un compromis entre rappel et précision.

5.7. Application à une nouvelle terminologie

5.7.1. Le cas du MeSH (Medical Subject Headings)

Les méthodes d'identification des noms propres que nous avons proposées pourraient être appliquées à d'autres terminologies médicales disponibles sous forme électronique, par exemple le MeSH.

Une dizaine de traductions du MeSH sont intégrées dans l'UMLS (NLM 1999). Toutefois, comme pour les autres terminologies présentes dans l'UMLS, les signes diacritiques ne sont disponibles dans aucune des traductions. De plus, la traduction française du MeSH, au contraire des autres traductions, comporte uniquement des majuscules non accentuées. La quasi totalité des termes est présente dans la plupart des traductions du MeSH, qui sont par ailleurs proches les unes des autres. En outre, chaque terme est doté d'un identifiant unique commun aux différentes traductions.

Dans le MeSH, les termes apparaissent fréquemment sous forme inversée (*« PARKINSON, MALADIE »*), comme dans l'index alphabétique de la CIM, et la préposition *« de »* est en général absente. Le MeSH est organisé sous forme de plusieurs hiérarchies de termes, permettant d'en extraire une liste de noms de micro-organismes. Ces caractéristiques sont résumées dans le tableau 7.

Les patrons risquent donc d'être difficiles à mettre en œuvre (les noms propres sont cependant en général en première position). L'absence d'accentuation risque d'induire un bruit supplémentaire pour la méthode des invariants, qui semble cependant être mieux placée pour le MeSH que les deux précédentes terminologies. La précision pourrait être augmentée en réalisant l'intersection de 3 traductions ou plus, incluant, par exemple l'espagnol, le portugais, etc. Une approche différenciée selon les chapitres pourrait être indiquée. Par exemple, les produits chimiques sont sans doute de mauvais candidats pour les invariants. Enfin, même avec le MeSH français, un alignement sur les codes devrait permettre d'utiliser une version des invariants ne tenant pas compte de la casse (ni des diacritiques s'ils sont disponibles dans une des langues), permettant au passage une recapitalisation (et éventuellement une réaccentuation).

5.7.2. Hors du domaine médical

Plus généralement, la méthode proposée est a priori applicable à des terminologies hors du domaine médical. Comme pour le MeSH, l'examen des caractéristiques de la terminologie doit permettre de déterminer quels critères seront les plus efficaces. Dans la plupart des cas, la partie externe du patron (ici, la préposition *« de »*), devra néanmoins être adaptée.

Considérons par exemple le thesaurus de l'astronomie (Shobbrook R. R. & Shobbrook R. R. 1994) dont des termes traduction l'un de l'autre figurent dans l'exemple (8). Les termes de ce thesaurus sont entièrement en majuscules non accentuées, ce qui rend le critère de casse inapplicable. Cependant, ce sont des syntagmes nominaux bien formés, et de plus le thesaurus est tra-

duit dans plusieurs langues (anglais, français, allemand, italien, espagnol). Si l'on perd en casse, on peut donc potentiellement gagner sur les critères fondés sur des patrons (ici, *« de X »* semble un bon point de départ) et sur les invariants. On voit ainsi sur l'exemple (8) comment le nom propre *« KREUTZ »* pourrait être identifié.

(8) *« KREUTZ SUNGRAZERS » (En)*
« COMETE DE KREUTZ FROLANT LE SOLEIL » (Fr)
« DIE SONNE STREIFENDE KOMETEN » (De)
« SUNGRAZERS TIPO KREUTZ » (It)
« RASANTES DEL SOL TIPO KREUTZ » (Es)

Un autre exemple est le thesaurus PASCAL de l'INIST, dont des termes figurent en (9). Dans la version dont nous disposons, les noms propres sont en majuscules, mais le premier mot de chaque terme aussi. Les caractères accentués sont présents. En revanche, les termes contiennent essentiellement les « mots pleins » (nom, adjectif, etc.), avec très peu de prépositions et de déterminants, ce qui rend difficile l'emploi de patrons [9]. Nous n'avons par ailleurs pas eu connaissance de traduction dans une autre langue. De plus, certaines expressions (emprunts, etc.) sont considérées comme des noms propres et capitalisées (*« Card Sorting Test Wisconsin »*). Il semble donc difficile d'en extraire et catégoriser les noms propres.

(9) *« Brique silicocalcaire », « Produit silicocalcaire » ;*
« Brise vent », « Abat vent », « Coupe vent » ;
« Bradydactylie héréditaire Mac Kinder », « Mac Kinder maladie » ;
« Brocoli », « Chou Broccoli » ;
« Canal artériel », « Canal Botal perméable » ;
« Champ proche », « Champ Fresnel » ;
« Children's Adaptive Behavior Scale », « Adaptive Behavior Inventory for Children » ;
« Wisconsin Card Sorting Test », « Card Sorting Test Wisconsin ».

Il serait utile d'automatiser, au moins partiellement, la mise au point de la meilleure stratégie pour chaque nouvelle terminologie abordée. (Gallippi A. F. 1996) propose une méthode dans laquelle on apprend, sur un échantillon d'un corpus, un arbre de décision servant à catégoriser les noms propres (personnes, entreprises, etc.) sur la base de leurs différents types de propriétés. Son travail pourrait servir d'inspiration pour aller dans ce sens.

CONCLUSION

Parmi les noms propres rencontrés dans les documents médicaux (noms de médecins et de patients, auteurs d'articles scientifiques, éponymes, noms

9. Contrairement à d'autres terminologies, les thesaurus destinés à la recherche documentaire (MeSH, PASCAL) présentent souvent des termes syntaxiquement « mal formés », probablement liés à des raisons historiques dans la constitution de ces terminologies.

de lieux), les éponymes (servant par exemple à désigner des maladies) constituent une classe relativement fermée. Un dictionnaire des éponymes médicaux serait une ressource utile pour typer ces noms propres. Or aucun dictionnaire de ce type n'est actuellement disponible sous forme électronique.

Les éponymes médicaux sont présents dans les terminologies médicales comme la CIM-10 ou le Répertoire d'anatomopathologie de la SNOMED, où ils représentent 5 à 7 % des différentes formes rencontrées. Il semble donc intéressant d'essayer de les recenser à partir des termes médicaux. Nous avons testé plusieurs méthodes pour identifier des noms propres dans deux terminologies présentant des caractéristiques différentes, avec des résultats satisfaisants en termes de précision et de rappel.

Nous avons également cherché à lier la performance de ces méthodes aux caractéristiques des terminologies, l'objectif étant de définir une stratégie pour aborder d'autres terminologies. Il existe en effet de nombreuses autres terminologies médicales, mono ou multilingues, qui permettront de contribuer à la constitution de ce dictionnaire de noms propres médicaux : nous avons vu comment les caractéristiques d'autres terminologies, médicales ou non, permettent d'envisager les méthodes utilisables ou d'anticiper leurs limites.

6. REMERCIEMENTS

Les auteurs tiennent à remercier le Docteur Roger Côté (Université de Sherbrooke, Québec) pour avoir mis à leur disposition une version pré-commerciale du Répertoire d'anatomopathologie de la SNOMED en français, et Yvan Emelin pour la version anglo-russe du Répertoire d'anatomopathologie.

Nous remercions également le LADL (Université Paris 7) pour nous avoir permis d'utiliser le dictionnaire DELAS, qui a servi à filtrer le vocabulaire de la CIM-10 ; l'INaLF (Nancy) pour l'étiqueteur WinBrill entraîné pour le français, qui a servi à amorcer l'étiquetage de la SNOMED ; l'IMS (Stuttgart) pour leur corpus WorkBench, qui a facilité les premières explorations de patrons sur la SNOMED ; Jean Royauté, à l'INIST (Nancy), pour le thesaurus PASCAL ; et les relecteurs de la revue, pour leurs conseils judicieux.

RÉFÉRENCES

BAUD, Robert ; Lovis, Christian ; Rassinoux, Anne-Marie ; Michel, Pierre-André ; Scherrer, Jean-Raoul (1998) : "Automatic extraction of linguistic knowledge from an international classification", *in Proceedings of the 9 th World Congress on Medical Informatics,* B. Cesnik ; C. Safran ; P. Degoulet (eds.), pp. 581–585, Seoul.

BODENREIDER, Olivier ; Zweigenbaum, Pierre (2000) : "Indentifying proper names in parallel medical terminologies", *in Medical Infobahn for Europe — Proceedings of MIE2000 and GMDS2000,* A. Hasman ; B. Blobel ; J.

Dudeck ; R. Engelbrecht ; G. Gell ; H.-U. Prokosh (eds.), pp. 443–447, Amsterdam.

CHEVALLIER, J. (1995) : *Précis de terminologie médicale*, Paris, Maloine, 6e édition.

CÔTÉ, Roger A. ; Rothwell, D. J. ; Palotay, J. L. ; Beckett, R. S. ; Brochu, Louise (eds.) (1993) : *The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International*, Northfield, College of American Pathologists.

CÔTÉ, Roger A. (1996) : *Répertoire d'anatomopathologie de la SNOMED internationale, v3.4*, Université de Sherbrooke, Sherbrooke, Québec.

DARPA (1993) : *Fifth Message Understanding Conference (MUC-5)*, San Francisco, Ca, Morgan Kaufmann, Defense Advanced Research Projects Agency.

DARPA (1996) : *MUC-6: Proceedings of the Sixth Message Understanding Conference*, Columbia, Maryland, Morgan Kaufmann, Defense Advanced Research Projects Agency, novembre 1995.

EMELIN, Ivan V. ; Levenson, R. ; Perov, Y. L. ; Rykov, V. V. (1995) : "A Russian version of SNOMED-International", *in Proceedings of the 8 th World Congress on Medical Informatics,* R. A. Greenes ; H. E. Peterson ; D. J. Protti (eds.), pp. 173–173, Vancouver.

FUNG, Pascale (1995) : "A pattern matching method for finding noun and proper noun translations from noisy parallel corpora", *in Proceedings of the 33rd ACL*, pp. 236–233, Boston, Mass.

GALLIPPI, Anthony F. (1996) : "Learning to recognize names across languages", *in Proceedings of the 16 th COLING,* J.-I. Tsujii (ed.), pp. 424–429, Copenhagen, Denmark.

GRABAR, Natalia ; Zweigenbaum, Pierre (2000) : "Automatic acquisition of domain-specific morphological resources from thesauri", *in Actes de RIAO 2000 : Accès à l'Information Multimédia par le Contenu*, C.I.D., pp. 765–784, Paris, France.

MCDONALD, David D. (1993) : "Internal and external evidence in the identification and semantic categorization of proper names", *in Corpus Processing for Lexical Acquisition,* B. Boguraev ; J. Pustejovsky (eds.), Cambridge (Mass.), MIT Press, pp. 61–76.

MIKHEEV, Andrei (1999) : "A knowledge-free method for capitalized word disambiguation", *in Proceedings of the 37th ACL*, College Park, Maryland.

NLM, Bethesda, Maryland (1999) : *UMLS Knowledge Sources Manual*, National Library of Medicine.

OMS, Genève (1993) : *Classification statistique internationale des maladies et des problèmes de santé connexes — Dixième révision*, Organisation mondiale de la Santé.

SHOBBROOK, Robert R. ; Shobbrook, Robyn R. (1994) : *The Astronomy Thesaurus*, The Anglo Australian Observatory, Epping, New South Wales, Australie, Disponible à `http://msowww.anu.edu.au/library/thesaurus/`.

THIELEN, Christine (1995) : “An approach to proper name tagging for German”, *in From Texts to Tags: Issues in Multilingual Language Analysis – Proceedings of the ACL SIGDAT Workshop*, pp. 35–40, Dublin.

WACHOLDER, N. ; Ravin, Y. ; Choi, M. (1997) : “Disambiguating proper names in text”, *in ANLP97*, pp. 202–208, Washington, DC.

WAKAO, Takahiro ; Gaizauskas, Robert ; Wilks, Yorick (1996) : “Evaluation of an algorithm for the recognition and classification of proper names”, *in Proceedings of the 16 th COLING,* J.-I. Tsujii (ed.), pp. 418–423, Copenhagen, Denmark.