Les utilisateurs de Doc'CISMeF peuvent-ils trouver ce qu'ils cherchent ? Une étude de l'adéquation du vocabulaire des requêtes des utilisateurs au MeSH

Natalia Grabar[a], **Pierre Zweigenbaum**[a], **Lina Soualmia**[b,c], **Stéfan Darmoni**[c]

[a] *DIAM — STIM/DSI, Assistance Publique – Hôpitaux de Paris & Département de Biomathématiques, Université Paris 6*
{ngr,pz}@biomath.jussieu.fr

[b] *Laboratoire Perception, Information et Systèmes, INSA, Rouen*

[c] *Département Informatique et Réseaux, CHU de Rouen*
{lina.soualmia,stefan.darmoni}@chu-rouen.fr

Abstract

*Context: Users of medical knowledge can write natural language queries to access more and more resources indexed using controlled vocabularies (*e.g.*, the MeSH): Medline and Internet catalogs such as CISMeF are well-known examples. This study examines an enabling condition of such natural language access: is the vocabulary of user queries comparable with that of the index terms (MeSH); can the observed gap between these vocabularies be reduced through the use of linguistic knowledge? Material: We matched the vocabulary found in the queries sent to CISMeF's Doc'CISMeF search tool to the target index terms: the French version of the MeSH, with the help of a morphological knowledge base for lemmatization and stemming. Methods: The two vocabularies were compared in their original form, then under incrementally normalized forms, using character-based normalizations (case and accents) then linguistic normalizations (lemmatization and stemming). Results: Only 16.7% of the user vocabulary, in its original form, is in the MeSH. Case normalization increases this proportion to 44.1%, then accent normalization to 50.6%, and morphological normalization to 56.8% ; a final spelling correction reaches 65.5%. Discussion: One third of the different words in user queries are uninterpretable by these methods. Morphological knowledge contributes 6 additional points to the score; but the coverage of our knowledge base is far from perfect, especially for stemming. Besides, if the frequency of occurrence of the words is taken into account, 89,3 % of user word occurrences can be matched to MeSH words. Conclusion: This shows the interest to take into account further matching methods between queries and index terms than those presented here; and to envisage, as a fallback, other access methods to target documents than the MeSH index.*

Keywords

MeSH; Information storage and retrieval; Subject headings; Controlled vocabulary; Natural language processing.

1 Introduction

Une masse croissante de connaissances médicales est mise à la disposition des professionnels de la santé et du grand public. D'une part, les bases de données bibliographiques comme Medline offrent un point d'entrée vers la littérature scientifique du domaine. D'autre part, des documents de diverses natures sont mis en ligne sur le Web par des acteurs variés : recommandations, sites thématiques, portails généralistes. Pour aider les « consommateurs » de connaissances à trouver ces ressources, une approche possible consiste à les *indexer* à l'aide d'un *vocabulaire contrôlé*. Ainsi, Medline (`www.ncbi.nlm.nih.gov/pubmed`) utilise des mots clés du thesaurus MeSH [1] pour caractériser le contenu informationnel des articles scientifiques ; et de même, des catalogues comme CISMeF (`www.chu-rouen.fr/cismef`) ou HON (`www.hon.ch`) indexent les sites web médicaux à l'aide du MeSH. L'un des intérêts majeurs de l'emploi d'un thesaurus comme le MeSH est sa hiérarchisation, qui permet d'interroger sur une notion générique (par exemple, *maladies cardiaques*) et de retrouver tous les « documents » indexés par des termes plus spécifiques (par exemple, *ischémies myocardiques*) : c'est le principe de l'*explosion* en sciences de l'information.

On trouve ainsi un avantage à passer par ces termes pivot pour accéder aux connaissances médicales. Cependant, on ne peut attendre de tous les utilisateurs de l'internet, et en particulier des « cybercitoyens », qu'ils connaissent les termes d'un vocabulaire contrôlé. Le MeSH compte quelque 20 000 termes, et malgré l'ajout de 9000 synonymes, un utilisateur non spécialiste a peu de chances d'employer d'emblée le terme canonique désignant ce qu'il cherche.

Les outils cités permettent donc l'interrogation sous forme de requêtes libres, et se chargent de rechercher la meilleure correspondance entre requête et termes d'indexation. Nous avons abordé cette problématique dans des travaux antérieurs [2], en examinant l'apport différentiel de connaissances linguistiques (flexion et dérivation) dans cette tâche. La question que nous soulevons ici se situe plus en amont, et concerne les conditions de faisabilité de cette mise en correspondance : dans quelle mesure les requêtes libres employées utilisent-elles un vocabulaire comparable à celui des termes cible ? Les écarts observés peuvent-ils être réduits par l'emploi de connaissances ? Des connaissances linguistiques suffisent-elles ?

Nous avons cherché à évaluer l'adéquation entre vocabulaire des utilisateurs et vocabulaire du MeSH à partir d'un échantillon de requêtes soumises à Doc'CISMeF (`doccismef.chu-rouen.fr`) [3], l'outil de recherche de CISMeF. Nous avons pour cela comparé ces vocabulaires, tout d'abord, sous leur forme initiale, puis sous une forme de plus en plus *normalisée*. Les normalisations appliquées concernent d'une part des opérations au niveau des caractères : casse (majuscules / minuscules) et accentuation. Elles concernent d'autre part des opérations fondées sur des connaissances morpholexicales : *lemmatisation* (réduction des pluriels et des féminins) et *racinisation* (réduction d'un mot dérivé à un mot initial dont il est dérivé, directement ou indirectement). Les « mots vides », comme c'est souvent le cas en recherche d'information, sont éliminés à l'étape initiale de la comparaison des deux vocabulaires.

Nous rappelons pour commencer quelques travaux sur l'appariement entre requêtes libres et termes d'une terminologie contrôlée. Nous précisons les requêtes et la terminologie cible sur lesquelles nous avons travaillé. Nous expliquons la méthode suivie pour comparer et normaliser vocabulaire des requêtes et vocabulaire cible. Nous détaillons ensuite les résultats de ces comparaisons, et discutons leurs implications. Nous proposons pour conclure les recommandations qui en découlent.

2 Travaux antérieurs

De nombreux travaux ont été menés sur la « variation terminologique » : l'identification d'expressions différentes de notions identiques ou proches. La question peut être vue comme la recherche de méthodes permettant d'*apparier* deux expressions. On peut pour celà travailler au niveau des lettres [4] (fautes de frappe, majuscules, accents), des mots et de leurs variantes morphologiques [5,6,7], de la syntaxe des expressions [7] ou en s'aidant de synonymes généraux [8] ou spécifiques au domaine médical [9]. Enfin, l'étude de la similarité des distributions statistiques de mots en corpus peut également aider à identifier des mots de sens proche (p. ex., [10] pour des familles morphologiques). Notons également les méthodes de correction phonémiques, qui visent à corriger un mot en conservant sa prononciation.

Par ailleurs, l'apport de traitements morphologiques en recherche d'information en langue française a été montré dans plusieurs travaux récents [11,12,13]. L'observation générale est que la lemmatisation apporte une amélioration statistiquement significative et que la racinisation apporte une contribution supplémentaire, mais statistiquement non significative.

3 Matériel

Nous présentons successivement les différents matériels qui entrent dans cette étude : un échantillon de requêtes libres envoyées à Doc'CISMeF, les termes cible de ces requêtes (essentiellement, le thesaurus MeSH), et les connaissances linguistiques utilisées pour aider leur mise en correspondance.

Doc'CISMeF recevait, début 2002, 5000 requêtes par jour à travers ses interfaces de recherche : simple (75 %), avancée (15 %), booléenne (7 %) et pas à pas (3 %). Les requêtes étudiées sont les requêtes envoyées à l'interface de recherche simple de Doc'CISMeF lors de la période de septembre 2000 à janvier 2001. Nous avons éliminé du journal des requêtes (le « log » du serveur) les requêtes vides et celles envoyées par l'équipe CISMeF (sur la base de l'adresse de la machine source). Cela constitue un ensemble de 108 660 requêtes (29 092 requêtes différentes). Nous avons cherché à tenir compte du fait que certaines requêtes sont plus fréquentes que d'autres. Pour cela, nous avons compté pour chaque requête le nombre de machines différentes d'où elle a été envoyée. Cela donne un nombre d'occurrences de *requêtes.machines* de 76 341 sur cette période.

Les termes cible sont ceux qui servent à indexer CISMeF : la version française du MeSH [14] (19971 termes, leurs 9151 synonymes et 83 qualificatifs), augmentée de 38 « métatermes » et 101 « types de ressources »[1] [3]. Notons que sur les 19971 termes du MeSH, de l'ordre de 5500 sont actuellement employés dans CISMeF pour indexer des sites Web, soit 23% du thesaurus. Au total, notre cible est constituée de 29 035 termes différents. Les termes du MeSH sont en majuscules non accentuées. Néanmoins, les termes employés dans CISMeF ont été mis en casse mixte (minuscules avec emploi « normal » des majuscules) et réaccentués.

Les connaissances linguistiques sont celles présentées dans [2], complétées comme indiqué ci-dessous. Elles servent pour *lemmatiser* et *raciniser* les mots. La lemmatisation réduit une forme fléchie à sa forme canonique (*abdominaux* - *abdominal*). La racinisation met en correspondance un mot dérivé avec le mot dont il dérive (*abdominal* - *abdomen*). Ces connais-

1. Début 2002, 800 termes supplémentaires ont été ajoutés dans le MeSH ; les métatermes de CISMeF sont au nombre de 51, et les types de ressources au nombre de 115.

sances prennent la forme de couples de mots comme ceux des exemples mentionnés. Nous disposons actuellement de 308 847 couples pour la lemmatisation (dont beaucoup représentent les flexions verbales) et de 1 041 couples pour la racinisation. Les ressources pour la lemmatisation ont été compilées à partir de dictionnaires généraux (lexique de l'ABU, `abu.cnam.fr/DICO`) et de différents corpus médicaux étiquetés (corpus MENELAS, parties du corpus CLEF). Les ressources pour la racinisation ont été générées à partir de terminologies [2,15]. De plus, des règles générales sont disponibles, par exemple pour passer d'un pluriel régulier en *-s* à un singulier.

Certains mots, dits *mots vides*, sont peu porteurs de sens dans une tâche de recherche d'information, et peuvent généralement être ignorés. Il s'agit généralement des mots extrêmement fréquents dans les documents cible. Une liste de mots vides a été établie dans une optique de recherche d'information dans le domaine médical [2]. Elle comprend des mots grammaticaux (articles, prépositions, conjonctions, pronoms), des numéraux, certains adverbes et adjectifs (mais aucun adjectif relationnel) et quelques formes verbales, pour un total de 199 formes. Une liste de 217 mots vides, mise au point pour la recherche d'information [13], est disponible à l'adresse `www.unine.ch/info/clef`. Nous l'avons fusionnée à notre liste, ce qui donne au total 344 mots vides.

Les traitements informatiques ont été effectués avec des scripts utilisant les filtres Unix standard `tr`, `sed`, `grep`, `awk` ainsi que `perl5`.

4 Méthodes

La méthode mise en œuvre consiste à segmenter en mots les termes source (requêtes) et cible (MeSH), et à comparer les vocabulaires résultants après des normalisations successives.

La segmentation en mots se fait en coupant la chaîne initiale aux espaces, ponctuations et autres caractères non alphanumériques. Le résultat de cette étape est une liste S de mots source, et une liste C de mots cible. Ces listes permettent une première évaluation du recouvrement des vocabulaires : on recense l'ensemble $S_{.c} = S \cap C$ des mots source qui se trouvent dans la cible (les mots communs aux deux listes) et son complémentaire $S_{.i} = S - S_{.c}$ (mots « inconnus »). Les traitements ultérieurs se feront sur les mots inconnus $S_{.i}$.

Les deux vocabulaires ($S_{.i}$ et C) sont soumis à une série de normalisations réalisées au niveau des caractères (section 4.1) ou qui demandent un apport de connaissances linguistiques (section 4.2). Les mots vides V (voir la section 3) sont rapidement supprimés des deux vocabulaires (ci-dessous, opération $-_v$).

4.1 Normalisations au niveau des caractères

Les normalisations au niveau des caractères ont été conditionnées par la nature du MeSH (mots en majuscules non accentuées, hormis les termes en minuscules ajoutés par l'équipe CISMeF), de même que par la nature non prévisible des mots employés dans les requêtes par les utilisateurs de CISMeF (mots en minuscules ou en majuscules, accentués ou non). Nous effectuons deux types de traitement à ce niveau : la minusculisation et la désaccentuation. Ce type de normalisation est élémentaire, mais important du fait de l'observation précédente.

Une première source potentielle de différences entre mots source et mots cible est donc la *casse* : la différence entre majuscules et minuscules. Pour effacer cette différence, nous normalisons les deux listes de mots en les passant en minuscules, soit $S_{.i.m}$ et C_m. Ce traitement

n'a pas d'incidence sur la signification des mots traités, sauf dans le cas de certains noms propres (*Pierre vs pierre*), qui ne se produit pas ici. Ici encore, on recense $S_{.^{i}.m} \cap C_m$, soit $S_{.^{i}.m^{c}}$, et son complémentaire $S_{.^{i}.m^{i}}$ (les « mots minuscules inconnus »).

Une deuxième normalisation des caractères est ensuite appliquée : il s'agit de la suppression des accents ; par exemple, chaque occurrence des lettres *éèêë* est convertie en *e*. En effet, les mots du MeSH sont non accentués. De plus, les requêtes ne sont pas toujours accentuées, ou peuvent être accentuées incorrectement (par exemple, *athlètisme*). Cette transformation peut effacer les différences de signification entre certains mots. Par exemple, les mots *sténose* et *sténosé* sont réduits à une seule forme : *stenose*. Nous avons remarqué que l'absence ou bien la différence d'accentuation bloquaient dans beaucoup de cas l'application des connaissances linguistiques. Nous avons donc testé son application en début de chaîne, et avons de même désaccentué les connaissances linguistiques. On s'attend alors à un *rappel* plus grand (davantage d'appariements entres mots des requêtes et mots cible), au prix éventuel d'une diminution de la précision (davantage d'appariements erronés). On recense ici $S_{.^{i}.m^{i}.d} \cap C_{m.d}$, soit $S_{.^{i}.m^{i}.d^{c}}$, et son complémentaire $S_{.^{i}.m^{i}.d^{i}}$.

4.2 Normalisations morpholexicales

Les normalisations présentées maintenant ont recours à des connaissances linguistiques. Ces connaissances ont une complétude variable, et doivent être considérées comme une approximation de traitements linguistiques plus complets.

C'est à ce moment que les mots vides sont supprimés des deux vocabulaires, donnant les ensembles $S_{.^{i}.m^{i}.d^{i}.\text{-}v}$ et $C_{m.d.\text{-}v}$, les mots connus $S_{.^{i}.m^{i}.d^{i}.\text{-}v^{c}} = S_{.^{i}.m^{i}.d^{i}.\text{-}v} \cap C_{m.d.\text{-}v}$ et leur complémentaire $S_{.^{i}.m^{i}.d^{i}.\text{-}v^{i}}$.

La *lemmatisation* réduit les formes fléchies d'un mot à sa forme canonique. Nous avons testé deux formes de lemmatisation.

lm une liste de 308 812 couples {*lemme*, *forme*} présentée à la section 3 ;

-s une heuristique de suppression de la marque du pluriel qui est utilisée en recherche d'information avec de bons résultats [13] : elle consiste dans les grandes lignes à supprimer les finales en *-s*, à réduire les *-aux* en *-al* et à supprimer les *-x* ;

lm-s la combinaison des deux.

Les mots qui peuvent être lemmatisés le sont. On obtient deux nouveaux ensembles de mots $S_{.^{i}.m^{i}.d^{i}.\text{-}v^{i}.lm}$ et $C_{m.d.\text{-}v.lm}$, et on calcule leur intersection $S_{.^{i}.m^{i}.d^{i}.\text{-}v^{i}.lm^{c}}$ et son complémentaire $S_{.^{i}.m^{i}.d^{i}.\text{-}v^{i}.lm^{i}}$: les mots lemmatisés inconnus. Les mêmes calculs sont effectués sur les autres méthodes de lemmatisation, soit $S_{.^{i}.m^{i}.d^{i}.\text{-}v^{i}.\text{-}s^{i}}$ et $S_{.^{i}.m^{i}.d^{i}.\text{-}v^{i}.lm\text{-}s^{i}}$.

Un autre apport d'information linguistique consiste à « raciniser » les mots. On applique la racinisation sur les mots du vocabulaire cible et sur les listes de mots qui sont toujours « inconnus » à la sortie des trois méthodes de lemmatisation. On obtient ainsi $C_{m.d.\text{-}v.lm.ra}$ et la version racinisée des trois listes de mots « inconnus » : $S_{.^{i}.m^{i}.d^{i}.\text{-}v^{i}.lm^{i}.ra}$, $S_{.^{i}.m^{i}.d^{i}.\text{-}v^{i}.\text{-}s^{i}.ra}$ et $S_{.^{i}.m^{i}.d^{i}.\text{-}v^{i}.lm\text{-}s^{i}.ra}$. De la même manière, les intersections et leurs complémentaires du côté des mots des requêtes sont calculés.

4.3 Analyse du reliquat

Les mots source qui n'ont été reconnus à aucune de ces étapes comme un mot cible ont été examinés et classés pour déterminer si des connaissances supplémentaires permettraient de les mettre en correspondance avec des mots du MeSH. Parmi ces méthodes, une correction orthographique a été tentée.

4.4 Correction orthographique : comparaison approximative de mots

Les mots qui restent « inconnus » après les normalisations précédentes sont peut-être des mots du MeSH mal orthographiés. Nous avons donc tenté une *correction orthographique* sur ces mots. Nous avons pour cela employé l'outil `ispell` d'Unix, avec le vocabulaire cible comme dictionnaire de référence. Cependant, comme une correction orthographique est trop aléatoire sur les mots courts, nous avons exclu de cette correction les mots de longueur inférieure à cinq lettres. Enfin, nous avons également exclu les propositions de correction multiples pour un même mot (correction ambiguë).

On obtient au final $C_{m.d.\text{-}v.lm.ra.co}$ et la version racinisée des trois listes de mots de l'étape précédente : $S_{.^i.m^i.\cdots.lm^i.ra^i.co}$, $S_{.^i.m^i.\cdots.\text{-}s^i.ra^i.co}$ et $S_{.^i.m^i.\cdots.lm\text{-}s^i.ra^i.co}$, avec les mots encore inconnus et leur complémentaire.

4.5 Types et occurrences

Chaque mot des requêtes peut apparaître dans une ou plusieurs requêtes ; et comme indiqué dans la section Matériel, chaque requête peut avoir été envoyée par plusieurs utilisateurs (machines) différents. Nous avons maintenu tout au long de ces traitements le décompte des mots différents (*types*), mais aussi celui du nombre d'*occurrences* de chaque mot. Ainsi, si dans les requêtes le mot *anorexie* est apparu N fois en majuscules et M fois en minuscules, une fois mis en minuscules (dans $S_{.^i.m}$), sa forme canonique minusculisée représentera $N + M$ occurrences. Le nombre total d'occurrences de mots reste donc constant tout au long des normalisations successives, sauf lors de la suppression des mots vides ($S_{.^i.m^i.d^i.\text{-}v}$). En revanche, le nombre total de types de mots diminue à mesure que certains types sont remplacés par une forme normalisée.

Le suivi différencié des types et des occurrences permet de tenir compte de l'importance relative des mots dans l'évaluation réalisée : il est a priori important que les mots qui concernent un plus grand nombre d'utilisateurs soient bien couverts.

5 Résultats

L'étude de la faisabilité d'appariements entre les mots des requêtes et du vocabulaire contrôlé consiste donc en une application progressive des différents traitements sur les mots qui n'ont pas été reconnus à l'étape précédente. Le tableau 1 détaille les résultats de ces appariements pour l'ensemble de la période allant de septembre 2000 à janvier 2001. Chaque rangée correspond à une étape. Les colonnes extérieures indiquent le nombre de types (mots différents) et d'occurrences (*occ.* : nombre total des emplois de ces mots) des mots de la *source* S (le vocabulaire des requêtes) et de la *cible* C (le vocabulaire du MeSH). Les colonnes intérieures montrent le résultat de la comparaison entre source et cible : les mots *inconnus* et les mots *communs*. Pour chaque étape, en plus des nombres absolus, on indique également

TAB. 1 – *Résultats de l'appariement (période totale de septembre 2000 à janvier 2001).*

Étape	**Source**		**Inconnus**		**Communs**		**Cible**	
	occ.	types	occ.	types	occ.	types	occ.	types
original	S		$S_{.i}$		$S_{.c}$		C	
	131570	*21112*	*61968*	*17674*	69602	3438	58912	21475
δ_{etape}			47,1 %	83,7 %	52,9 %	16,3 %		
$\%_{original}$			*47,1 %*	*83,7 %*	52,9 %	16,3 %		
minusculisation	$S_{.i.m}$		$S_{.i.m^i}$		$S_{.i.m^c}$		C_m	
	61968	16333	23217	11806	38751	4527	58912	20251
δ_{etape}			37,5 %	72,3 %	62,5 %	27,7 %		
$\%_{original}$			17,6 %	55,9 %	29,5 %	21,4 %		
désaccentuation	$S_{.i.m^i.d}$		$S_{.i.m^i.d^i}$		$S_{.i.m^i.d^c}$		$C_{m.d}$	
	23217	11292	20004	10420	3213	872	58912	19351
δ_{etape}			86,2 %	92,3 %	13,8 %	7,7 %		
$\%_{original}$			15,2 %	49,4 %	2,4 %	4,1 %		
suppression vides	$S_{.i.m^i.d^i.\text{-}v}$		$S_{.i.m^i.d^i.\text{-}v^i}$		$S_{.i.m^i.d^i.\text{-}v^c}$		$C_{m.d.\text{-}v}$	
	19919	10387	19919	10387	–	–	57053	19279
$\%_{original}$			15,1 %	49,2 %	–	–		
lemmatisation	$S_{.i.m^i.d^i.\text{-}v^i.lm}$		$S_{.i.m^i.d^i.\text{-}v^i.lm^i}$		$S_{.i.m^i.d^i.\text{-}v^i.lm^c}$		$C_{m.d.\text{-}v.lm}$	
[par règles]	19919	10173	17631	9370	2288	803	57053	17923
δ_{etape}			88,5 %	92,1 %	11,5 %	7,9 %		
$\%_{original}$			13,4 %	44,4 %	1,7 %	3,8 %		
[par heuristiques]	19919	10155	17659	9328	2260	827	57053	17957
δ_{etape}			88,7 %	91,9 %	11,3 %	8,1 %		
$\%_{original}$			13,4 %	44,2 %	1,7 %	3,9 %		
[combinée]	19919	10088	17456	9176	2463	912	57053	17633
δ_{etape}			87,6 %	91,0 %	12,4 %	9,0 %		
$\%_{original}$			13,3 %	43,5 %	1,9 %	4,3 %		
racinisation	$S_{.i.m^i.d^i.\text{-}v^i.lm^i.ra}$		$S_{.i.m^i.d^i.\text{-}v^i.lm^i.ra^i}$		$S_{.i.m^i.d^i.\text{-}v^i.lm^i.ra^c}$		$C_{m.d.\text{-}v.lm.ra}$	
[par règles]	17631	9370	17485	9311	146	59	57053	17663
δ_{etape}			99,2 %	99,4 %	0,8 %	0,6 %		
$\%_{original}$			13,3 %	44,1 %	0,1 %	0,3 %		
[par heuristiques]	17659	9328	17516	9260	143	68	57053	17703
δ_{etape}			99,2 %	99,3 %	0,8 %	0,7 %		
$\%_{original}$			13,3 %	43,9 %	0,1 %	0,3 %		
[combinée]	17456	9176	17306	9114	150	62	57053	17370
δ_{etape}			99,1 %	99,3 %	0,9 %	0,7 %		
$\%_{original}$			13,2 %	43,2 %	0,1 %	0,3 %		
correction	$S_{.i.m^i.\cdots.lm^i.ra^i.co}$		$S_{.i.m^i.\cdots.lm^i.ra^i.co^i}$		$S_{.i.m^i.\cdots.lm^i.ra^i.co^c}$		$C_{m.d.\text{-}v.lm.ra.co}$	
[par règles]	17485	8855	14397	7521	3088	1334	57053	17379
δ_{etape}			82,3 %	84,9 %	17,7 %	15,1 %		
$\%_{original}$			10,9 %	35,6 %	2,3 %	6,3 %		
[par heuristiques]	17516	8782	14434	7465	3082	1317	57053	17547
δ_{etape}			82,4 %	85,0 %	17,6 %	15,0 %		
$\%_{original}$			11,0 %	35,4 %	2,3 %	6,2 %		
[combinée]	17306	8643	*14111*	*7277*	3195	1366	57053	17370
δ_{etape}			81,5 %	84,2 %	18,5 %	15,8 %		
$\%_{original}$			*10,7 %*	*34,5 %*	2,4 %	6,5 %		

pour les types et les occurrences la répartition en pourcentage des mots connus et inconnus (δ_{etape}), et le pourcentage que ces nombres représentent vis-à-vis du vocabulaire original ($\%_{original}$). Les mots restant inconnus à l'étape n servent de source à l'étape $n + 1$.

Les requêtes contiennent 29 092 termes différents, qui comportent 21 112 mots différents (types de mots) pour un total de 131 570 occurrences dans les requêtes.machines relevées (voir la section Matériel). Parmi ceux-ci, 3438 mots (69 602 occurrences) sont communs avec le vocabulaire du MeSH (qui compte 58 912 occurrences pour 21 475 mots différents) ; et les 17 674 restants (61 968 occurrences) sont inconnus du MeSH. La répartition entre mots connus et inconnus est de 16,3/83,7 % (52,9/47,1 % si l'on tient compte des occurrences). Comme aucune transformation n'a été effectuée à cette étape sur les deux vocabulaires, cet appariement direct a toutes les chances d'être correct ; seuls quelques cas de mots ambigus (par exemple, *voie* dans *pays voie de developpement* et *anesthesie voie rectale*) sont des sources potentielles d'erreurs et risquent de causer du bruit.

La mise en minuscules regroupe les 17 674 mots restants en 16 333 mots en minuscules. À elle seule, elle permet de reconnaître 62,5 % des occurrences restantes, soit au total 82,4 %, au prix éventuel d'une perte de spécificité de la recherche. 17,6 % des occurrences (23 217) restent alors inconnues, correspondant à 11 806 mots différents (55,9 %). La désaccentuation permet de traiter 13,8 % des occurrences de mots encore non reconnues. Il reste maintenant 20 004 occurrences de mots non reconnues (15,2 %) pour 10 420 types (49,4 % des types d'origine). Les 33 mots vides supprimés à ce stade ne comptent que 85 occurrences. En effet, la plupart des mots vides des requêtes sont également dans le MeSH et ont été reconnus aux étapes précédentes. Les résultats des trois méthodes de lemmatisation sont donnés côte à côte. Les traitements suivants s'appliquent à chacun de ces résultats séparément.

Des traitements morpholexicaux permettent de reconnaître de 11,3 à 12,4 % des occurrences restantes selon la méthode de lemmatisation, plus moins de 1 % pour la racinisation. Il reste alors au mieux 9176 mots non reconnus (43,5 %), soit à 13,3 % des occurrences des mots.

Enfin, la correction propose de corriger de l'ordre de 6,5 % des mots restants (18 % des occurrences restantes). Néanmoins, malgré les précautions prises, de nombreuses propositions sont erronées. Au final, 65,5 % des mots peuvent être mis en rapport avec des mots du MeSH, soit 89,3 % des occurrences.

Le tableau 2 trace l'évolution de ces mesures pour chaque mois de la période considérée en retenant les données de synthèse suivantes, chaque fois en nombre d'occurrences et de types : taille du vocabulaire *source*, *comparaison initiale* obtenue sans normalisation (première ligne du tableau 1) et *comparaison finale* obtenue après lemmatisation combinée, racinisation et correction orthographique (tableau 1, antépénultième ligne). Pour ces deux comparaisons, on a indiqué les chiffres en absolu et en pourcentage du vocabulaire source. Pour aider à repérer ces données, nous avons mis en italiques ces six chiffres dans le tableau 1.

Le nombre d'occurrences du vocabulaire source augmente au fil des mois (davantage de requêtes sont soumises), mais cela ne semble pas être le cas pour le nombre de types. Le pourcentage de mots non reconnus dans leur forme d'origine diminue au fil des mois, que ce soit en termes d'occurrences (nettement) ou de types (plus doucement). La même évolution s'observe sur le pourcentage de mots non reconnus après les normalisations. Nous avons également rappelé sur ce tableau la synthèse des données de l'ensemble de la période. On remarque que dans les comparaisons, les pourcentages d'occurrences non reconnues sont proches de la médiane des données mensuelles. En revanche, les pourcentages de types non reconnus sont nettement supérieurs à la valeur maximale mensuelle correspondante.

TAB. 2 – *Évolution mensuelle de septembre 2000 à janvier 2001.*

Données	**Source**		**Comparaison initiale**				**Comparaison finale (**$_{i}.m^{i}...lm\text{-}s^{i}.ra^{i}.co^{i}$**)**			
	occ.	types	occ.	types	% occ.	% types	occ.	types	% occ.	% types
09/00	16909	6519	9269	4902	54,8 %	75,2 %	2532	1732	15,0 %	26,6 %
10/00	22635	7510	12089	5716	53,4 %	76,1 %	3038	2061	13,4 %	27,4 %
11/00	28277	7826	13935	5902	49,3 %	75,4 %	3263	2106	11,5 %	26,9 %
12/00	28186	6875	12601	5081	44,7 %	73,9 %	2645	1796	9,4 %	26,1 %
01/01	36317	7167	14355	5203	39,5 %	72,6 %	2614	1787	7,2 %	24,9 %
09/00–01/01	131570	21112	61968	17674	47,1 %	83,7 %	14111	7277	10,7 %	34,5 %

6 Discussion

Lors de la comparaison directe des vocabulaires (S *vs* C), on constate qu'un peu plus de la moitié des mots (occurrences) du vocabulaire des requêtes fait partie du vocabulaire d'indexation de CISMeF (le MeSH) ; on en dénombre seulement un sixième si l'on considère les mots différents (types). Cette proportion est faible, car elle signifie que sans traitements supplémentaires, près de la moitié de ce que recherchent les utilisateurs obtiendrait des résultats de moins bonne qualité. En effet, dans Doc'CISMeF, la demande d'un terme hors du MeSH génère une recherche sur le contenu intégral des notices. On a vu qu'heureusement, les traitements étudiés améliorent de beaucoup ce premier chiffre.

Si l'on tient donc compte du fait que le MeSH est écrit en majuscules non accentuées et que les utilisateurs écrivent avec une casse et une accentuation variables, cette proportion peut être considérée comme élevée. Cela provient selon nous de plusieurs origines. D'une part, un grand nombre d'utilisateurs de Doc'CISMeF sont des documentalistes, qui connaissent le MeSH (... en anglais). D'autre part, un grand nombre de pathologies courantes sont des mots clés MeSH, qui de plus sont couverts par Doc'CISMeF et donc écrits en minuscules accentuées. Par exemple, on trouve parmi les vingt requêtes les plus fréquentes en janvier 2002 les mots *euthanasie, anorexie, alcoolisme, hepatite/hépatite, pneumonie, sida, varicelle, paludisme*. Enfin, certains sites proposent des liens consistant en des requêtes toutes prêtes à Doc'CISMeF : ces requêtes sont systématiquement correctes. L'augmentation de plusieurs de ces facteurs peut aider à expliquer la diminution progressive du pourcentage de requêtes non reconnues : l'accroissement du nombre de liens « précâblés » vers Doc'CISMeF, amplifié par une couverture croissante des pathologies courantes, avec leur mise en minuscules ; et une familiarité croissante des utilisateurs avec le contenu de CISMeF.

L'utilisation de traitements supplémentaires, y compris une proposition de correction orthographique, amène à un taux final de reconnaissance de 89,3% des occurrences de mots (65,5 % des types). Quantativement, à peine un dixième de ce qu'écrivent les utilisateurs était donc ininterprétable dans cette période avec les méthodes examinées ; cette proportion monte cependant à un tiers si l'on considère le nombre de types de mots employés.

Un examen manuel des mots non reconnus avant correction a permis de recenser plusieurs cas. Nous avons noté des mots du MeSH mal orthographiés (*acetylsalycilique*, *altzheimer*, *alzeihmer*, *1lzheimer*) ; des variantes morphologiques de mots du MeSH pour lesquels nos connaissances linguistiques étaient incomplètes (*cicatriciel* pour MeSH *cicatrice*, *encephalique* pour MeSH *encephale*) ; des abréviations (*bpco*, *esb*), *biam* ; des mots anglais (*acalculous*, *allergy*, *bronchiolitis*) ; des notions qui ne se trouvent simplement pas dans le MeSH (*calcemie*, *adenomectomie*, *aggir*, *amphitheatre*), y compris des noms propres (*darmoni*,

beuscart, *bethune*, *brest*, *bichat*, *broussais*). Nous avons vu qu'un rapprochement entre mots restants et mots du MeSH par correction orthographique permet d'en traiter 15 % des types et 18 % des occurrences. Il faut cependant noter que cette correction n'est qu'une proposition, qui serait à valider par l'utilisateur. Il reste à affiner les conditions de cette correction, et à évaluer dans quelle proportion le mot proposé est valide.

La différence observée entre le pourcentage de types de mots non reconnus chaque mois et ce même pourcentage sur l'ensemble de la période montre que ces mots se renouvellent d'un mois sur l'autre. Ce n'est pas étonnant pour les noms propres, les abréviations, les mots anglais et les mots hors MeSH. Il en existe un réservoir énorme, et leur diminution demanderait des utilisateurs une meilleure connaissance du MeSH. Mais heureusement, les mots les plus souvent demandés sont en moyenne ceux qui se trouvent dans le MeSH.

Parmi les étapes examinées, la normalisation en minuscules apporte le plus gros gain (gain relatif de 27,7 % sur les types, 62,5 % sur les occurrences), et avec la désaccentuation, la proportion de mots connus monte à 51,6 % des types d'origine et 84,8 % des occurrences. Étant donné la simplicité des méthodes employées, ce chiffre devrait être considéré comme la base de ce qu'il est possible de faire. On constate que ce taux de recouvrement entre vocabulaire utilisateur et vocabulaire du MeSH est tout à fait honorable en termes d'occurrences.

La contribution totale des connaissances morphologiques est plus faible, comparable à celle de la désaccentuation (gain relatif de l'ordre de 9 % des types et 13 % des occurrences), alors qu'elle demande davantage de ressources initiales (la base de connaissances morphologiques). Elle permet d'atteindre 56,8 % des types et 86,8 % des occurrences du vocabulaire utilisateur. On entre probablement ici dans un schéma classique dans la résolution de problèmes : au-delà d'un certain point, les efforts nécessaires pour améliorer les résultats sont de plus en plus importants pour un gain qui va en diminuant. Il faut remarquer d'une part que la couverture des ressources linguistiques pour la lemmatisation et la racinisation n'est pas parfaite. C'est surtout vrai pour celles de racinisation, qui font encore défaut de façon générale pour le français [16]. L'étude du MeSH et des requêtes devrait permettre de les compléter. Il faut observer d'autre part que la contribution de ces méthodes en termes de pourcentage de types de mots reconnus, qui conditionne la variété de recherche possible pour les utilisateurs, est plus grand que celui en termes de pourcentage d'occurrences reconnues (lignes $\%_{original}$).

Notons aussi que toutes les normalisations effectuées peuvent rencontrer des ambiguïtés, avec des fréquences diverses. C'est rarement le cas pour la mise en minuscules, un peu plus pour la désaccentuation, la lemmatisation et la racinisation. Le gain de sensibilité obtenu par ces normalisations peut ainsi s'accompagner d'une perte de spécificité. Enfin, rappelons que les mesures effectuées concernent le recouvrement des mots individuels. La plupart des requêtes et des termes MeSH comprennent plusieurs mots, et leur appariement demande des conditions et des traitements supplémentaires (voir par exemple [2]).

Ces mesures portent sur une période de cinq mois du journal de Doc'CISMeF. Nous avons noté des tendances sur cette période, qui demandent à être vérifiées. Nous comptons donc appliquer les mêmes méthodes et les outils réalisés pour analyser les mois plus récents. Avec le temps, les utilisateurs qui ont visité plusieurs fois CISMeF peuvent devenir plus familiers avec les termes du MeSH et utiliser davantage son vocabulaire. À l'inverse, Doc'CISMeF recrute un nombre toujours croissant d'utilisateurs (2500 par jour ouvré début 2002), et conserve ainsi une proportion importante de novices. La baisse lente du taux de types de mots inconnus, telle que mesurée par les méthodes présentées ici, demande confirmation.

Si ces mots non reconnus ne sont pas appariables avec le vocabulaire du MeSH, ils pour-

raient cependant être présents en texte libre dans les notices de CISMeF, ou dans les pages sur lesquelles pointe CISMeF (les ressources primaires). Cette hypothèse reste à tester. Si elle se vérifie, le traitement des requêtes de recherche pourrait, pour les mots non reconnus, se replier sur un modèle classique de recherche d'information dans ces pages.

7 Conclusion

Les mesures effectuées dans ce travail montrent l'importance de traitements de « bas niveau » dans la mise en correspondance de requêtes en texte libre avec les termes d'un vocabulaire contrôlé.

Elles indiquent aussi l'intérêt mais également les limites des connaissances morphologiques pour aider cette mise en correspondance. Si l'apport de la lemmatisation est net, celui de la racinisation reste extrêmement faible dans cette étude. Il faut toutefois noter la différence quantitative des connaissances mises en jeu, deux cents fois plus importantes pour la lemmatisation que pour la racinisation. Des connaissances plus complètes pour cette dernière pourraient donc peut-être modifier ces résultats.

Une correction orthographique est en revanche un facteur d'amélioration important. Cependant, ses résultats sont moins fiables, et ses conditions d'utilisation restent donc à préciser. Avant d'effectuer cette correction, il sera utile d'effectuer une recherche dans la version originale du MeSH pour identifier d'éventuels mots anglais. Notons que Doc'CISMeF effectue déjà cette recherche. Il sera bon également de rechercher les mots non reconnus dans le MeSH dans un lexique français plus grand, pour ne pas chercher à « corriger » des mots existants dont le seul tort est de ne pas figurer dans le MeSH.

La proportion de mots non traités par les méthodes étudiées, qui reste importante, témoigne de la nécessité de prendre en compte d'autres méthodes d'appariement et d'accès. Parmi ces autres méthodes, rappelons l'emploi de synonymes [9] ou de similarités distributionnelles. Le repli sur un accès classique en texte intégral, mentionné plus haut, est une autre option. La combinaison de ces méthodes est le thème de recherche de l'un des auteurs (LS).

Enfin, les résultats de cette étude fournissent une mesure très utile de l'adéquation entre vocabulaire des utilisateurs et vocabulaire d'indexation ; elle devrait devenir l'une des métriques servant au suivi régulier de Doc'CISMeF.

Références

[1] Medical Subject Headings. page WWW `http://www.nlm.nih.gov/mesh/meshhome.html`, National Library of Medicine, Bethesda, Maryland, 2001.

[2] Zweigenbaum P, Grabar N, et Darmoni SJ. Projection de requêtes en langue naturelle sur les termes du MeSH : l'apport de connaissances morphologiques. In: Staccini P et Fieschi M, eds, Actes de IPM 2001, Nice. novembre 2001. *À paraître*.

[3] Darmoni SJ, Thirion B, Leroy JP, et al. A search tool based on 'encapsulated' MeSH thesaurus to retrieve quality health resources on the Internet. *Med Inform Internet Med* 2001;26(3):165–78.

[4] Lovis C et Baud R. Fast exact string pattern-matching algorithms adapted to the characteristics of the medical language. *J Am Med Inform Assoc* 2000;7(4):378–91.

[5] McCray AT, Srinivasan S, et Browne AC. Lexical methods for managing variation in biomedical terminologies. In: Proc Eighteenth Annu Symp Comput Appl Med Care, Washington. Mc Graw Hill, 1994; pp. 235–9.

[6] Lovis C, Michel PA, Baud R, et Scherrer JR. Word segmentation processing: a way to exponentially extend medical dictionaries. In: Greenes RA, Peterson HE, et Protti DJ, eds, Proc 8 th World Congress on Medical Informatics, 1995; pp. 28–32.

[7] Jacquemin C et Tzoukermann E. NLP for term variant extraction: A synergy of morphology, lexicon, and syntax. In: Strzalkowski T, ed, *Natural language information retrieval*, (vol7) of *Text, speech and language technology*. Kluwer Academic Publishers, Dordrecht & Boston, 1999; pp. 25–74.

[8] Hamon T, Nazarenko A, et Gros C. A step towards the detection of semantic variants of terms in technical documents. In: Boitet C, ed, Proceedings of the 17 th COLING, Montréal, Canada. 10–14 August 1998; pp. 498–504.

[9] Pouliquen B, Delamarre D, et Le Beux P. Indexation de textes médicaux par extraction de concepts, et ses utilisations. In: Actes de JADT, 2002. *À paraître*.

[10] Xu J et Croft BW. Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems* 1998;16(1):61–81.

[11] Gaussier E, Grefenstette G, Hull D, et Roux C. Recherche d'information en français et traitement automatique des langues. *Traitement automatique des langues* 2000;41(2):473–93.

[12] Zweigenbaum P, Darmoni SJ, et Grabar N. The contribution of morphological knowledge to French MeSH mapping for information retrieval. *J Am Med Inform Assoc* 2001;8(suppl):796–800.

[13] Savoy J. Morphologie et recherche d'information. Cahier de recherche en informatique CR-I-2002-01, Université de Neuchatel, Division économique et sociale, Faculté de Droit et des Sciences Économiques, 2002.

[14] Institut National de la Santé et de la Recherche Médicale, Paris. Thésaurus Biomédical Français/Anglais, 2000.

[15] Grabar N et Zweigenbaum P. A general method for sifting linguistic knowledge from structured terminologies. *J Am Med Inform Assoc* 2000;7(suppl):310–4.

[16] Hathout N, Namer F, et Dal G. An experimental constructional database: the MorTAL project. In: Boucher P, ed, *Morphology book*. Cascadilla Press, Cambridge, MA, 2001. *À paraître*.

Addresse de correspondance

Pierre Zweigenbaum
DIAM — STIM/DSI/AP-HP
91, boulevard de l'Hôpital
75634 Paris Cedex 13, France
e-mail: `pz@biomath.jussieu.fr`
url: `http://www.biomath.jussieu.fr/~pz/`