SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100305

Morphologie contrastive au travers des domaines de spécialité et des registres

Chmielik, Jolanta[1] et Grabar, Natalia[2]

[1] Commission européenne, DG RTD, recherche et innovation, Bruxelles

[2] UMR 8163 STL, CNRS & Université de Lille 3 et Lille 1, Villeneuve d'Ascq

Avec la démocratisation des savoirs, dont l'évolution s'accentue grâce à l'utilisation de la Toile, un nombre croissant de documents spécialisés peut être trouvé en ligne. Cette situation est typique de différents domaines de spécialité (biomédical, juridique, énergétique, télécommunication ...). Si de très nombreux documents peuvent ainsi être trouvés, leur contenu n'est pas pour autant accessible à tous les utilisateurs. En effet, les documents vulgarisés et spécialisés co-existent. Or, ces derniers peuvent contenir des informations très techniques et difficiles à comprendre par les non experts. Dès alors, l'indication sur leur degré de spécialisation est une information importante, car elle permet justement de guider les utilisateurs vers des contenus qui leur sont plus appropriés. Dans ce travail, nous montrons qu'il est possible d'exploiter les informations morphologiques des documents de santé afin d'effectuer une distinction automatique efficace (avec des performances souvent supérieures à 90 %) de leur degré de spécialisation. Nous effectuons une analyse quantitative et qualitative du matériel morphologique. Nous effectuons également une comparaison entre l'exploitation du matériel morphologiques et le matériel lexical. Pour ce faire, nous exploitons plusieurs outils du Traitement Automatique des Langues qui nous permettent ainsi d'accéder au niveau morphologique des mots des documents et de nous concentrer de cette manière sur leur structure morphologique et sémantique.

1 Contexte et positionnement des objectifs scientifiques

La démocratisation des savoirs, devenue possible surtout grâce à l'évolution de la Toile, ouvre des possibilités importantes dans la diffusion de ces savoirs et surtout dans leur accès instantané et très souvent libre. Toutefois, si tout utilisateur de la Toile peut accéder physiquement aux documents mis en ligne, leur contenu risque de rester sémantiquement opaque si le niveau d'explicitation n'est pas suffisant ou bien s'il n'est pas adapté à la compétence de l'utilisateur. Notre travail est ciblé sur les documents biomédicaux, car environ 80 % des internautes s'intéressent aux questions liées à la santé (Fox, 2006 ; Fox, 2011). Comme ce chiffre souligne aussi la préoccupation que manifestent les citoyens vis-à-vis de leur santé, il devient intéressant d'analyser (1) si la technicité des documents du domaine biomédical est élevée, (2) si cette technicité est actuellement différenciée en ligne, mais aussi (3) si la distinction de la technicité est observable directement dans les documents. C'est ce que nous analysons dans la suite de cette section.

1.1 Les documents en ligne ont-ils une technicité importante ?

Depuis plusieurs années déjà, il a été constaté que la technicité élevée des documents de santé peut avoir des effets néfastes sur leurs utilisateurs (AMA, 1999). Les conséquences sont d'autant plus graves que les documents sont consultés par les patients. Par exemple, plusieurs études montrent que le degré de technicité élevé conduit en effet à un impact négatif sur la compréhension des informations par les patients. De plus, cela peut avoir pour effet une déterioration de la communication des patients avec les médecins, mais surtout causer l'incompréhension et l'échec des soins médicaux administrés aux patients (McCray, 2005 ; Tran et al., 2009). Si cet état de fait a été observé, le changement des pratiques ne va pas de soi. En effet, il a été aussi constaté que les notions véhiculées dans plusieurs sites et pages, de même que leur présentation, restent très complexes. Ainsi, une analyse de 25 sites rédigés en anglais et en espagnol a montré que tous les sites en anglais et 86 % de sites en espagnol exigent d'avoir un niveau d'études universitaires pour que leur contenu puisse être correctement compris (Berland et al., 2001). Les

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20120100305

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100305

informations véhiculées par les sites de la Toile montrent ainsi une technicité importante et risquent donc d'être inaccessibles pour la plupart des patients.

1.2 La technicité des documents en ligne est-elle différenciée ?

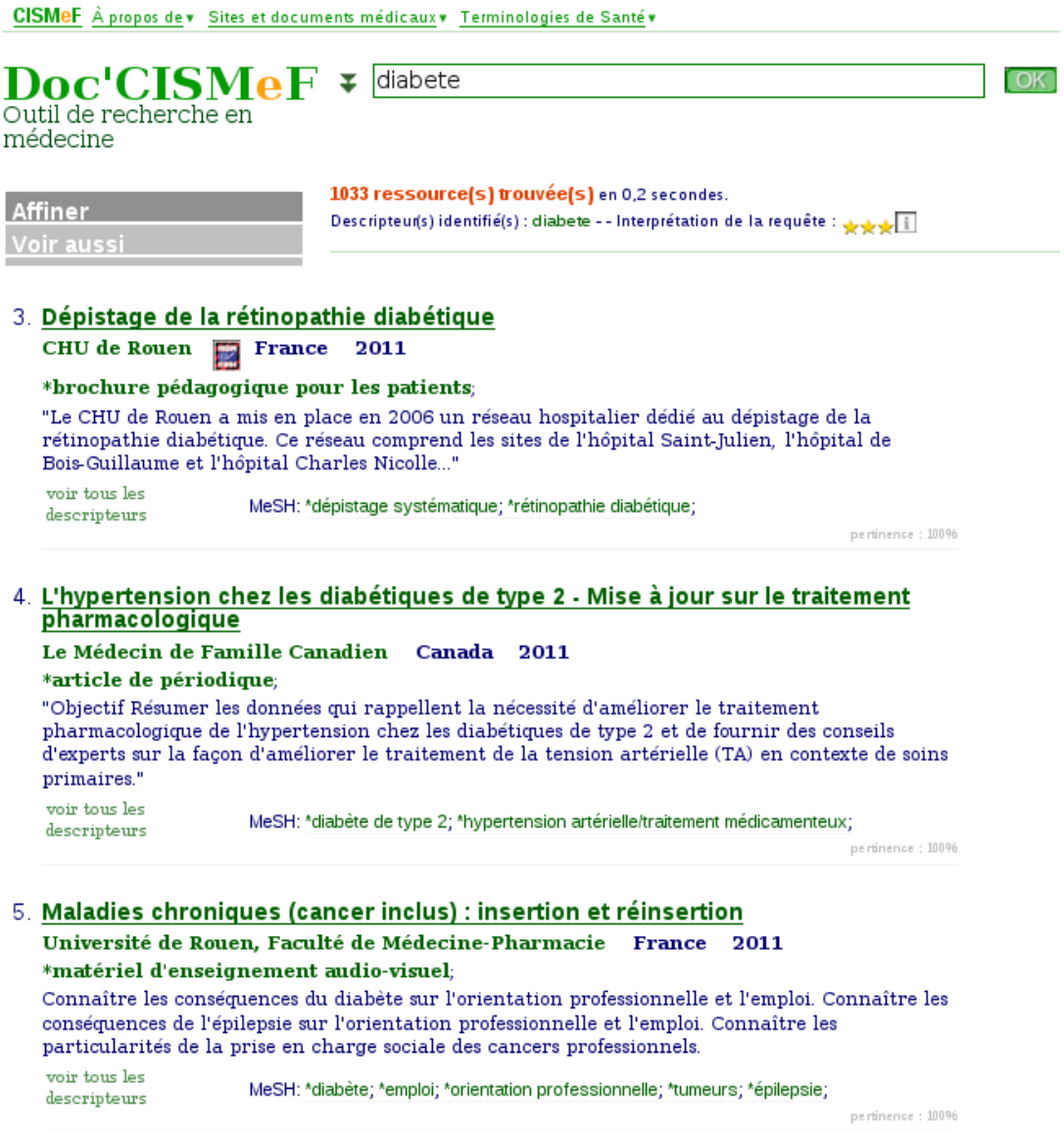

Figure 1 : Un extrait de la recherche effectuée sur le portail CISMeF avec le mot-clé « diabète ».

La technicité et l'hétérogénéité technique des documents restent ainsi importantes sur la Toile. Par contre, il est raisonnable de supposer que, si cette hétérogénéité est clairement indiquée aux utilisateurs, cela peut leur permettre de consulter les documents dont la spécialisation est plus appropriée à leur niveau. Ceci peut donc les aider dans la compréhension des documents.

Il est possible d'effectuer cette analyse grâce à une visite sur les portails de santé. Les plus connus et utilisés en France sont HON (Health on the Net) et CISMeF (Catalogue et Index des Sites MEdicaux de

langue Française). Notons qu'il existe aussi un moteur de recherche généraliste GoogleCoopsanté (COOP).

Le portail HON ne propose pas une annotation indiquant le degré de spécialisation des documents, bien que des travaux de catégorisation automatique des documents selon deux catégories (documents spécialisés et documents non spécialisés) soient menés dans l'équipe HON (Borst et al, 2008). Quant au portail CISMeF, comme le montre la figure 1, les documents sont annotés avec les mots-clés thématiques (*diabète de type 2, dépistage systématique...*) et avec une typologie de documents (*article de périodique, brochure pédagogique pour les patients, matériel d'enseignement...*), cette dernière étant évocatrice des destinataires. Par exemple, une brochure pédagogique pour les patients est créée à destination des patients, le matériel d'enseignement est destiné aux étudiants en médecine, tandis qu'un article de périodique est essentiellement destiné aux professionnels du domaine biomédical. Dans le portail CISMeF, cette annotation est effectuée manuellement par les documentalistes. Finalement, le moteur de recherche GoogleCoopsanté propose également cette distinction. Il faut cependant noter que ce moteur ne produit pas ces annotations, mais exploite celles qui lui sont fournies par les partenaires.

Globalement, il est très rare que la technicité des documents soit indiquée de manière explicite en ligne. Les documents appartenant à différents registres de spécialisation co-existent donc sur la Toile sans que ce soit explicite pour les utilisateurs.

1.3 La technicité est-elle observable directement dans les documents ?

Une analyse de l'état de l'art actuel montre que, moyennant les méthodes et outils adaptés, la technicité des documents peut être observée directement en leur sein. Deux principaux types de travaux existent à cet effet.

D'une part, il s'agit d'établir des lexiques biomédicaux monolingues (CHV) ou bilingues (Messai et al., 2006 ; Deléger et Zweigenbaum, 2008). Dans ces lexiques, les termes techniques employés par les professionnels de santé sont alignés avec les expressions utilisées par les patients. Par exemple, le terme *infarctus du myocarde* correspond à l'expression grand public *crise cardiaque*, alors que *rhagade* correspond à *crevasse*. Les lexiques de ce type peuvent aider les non professionnels du domaine à mieux comprendre et appréhender le contenu des documents biomédicaux. Malheureusement, ces lexiques sont très incomplets.

D'autre part, il s'agit d'effectuer une comparaison et une distinction des degrés de spécialisation des documents à destination des experts et des patients. Parmi de nombreux travaux existants, les formules dites *formules linguistiques de lisibilité* (Flesch, 1948 ; Gunning, 1973 ; Björnsson et Härd af Segerstad, 1979) sont largement utilisées, notamment parce qu'elles peuvent être intégrées dans les éditeurs de texte. La définition de ces formules repose sur les critères liés à la longueur moyenne des mots ou des phrases dans la langue considérée. Ainsi, si les mots et les phrases sont complexes et longs (en nombre de caractères pour les mots et en nombre de mots pour les phrases), ils sont considérés comme savants et difficiles à comprendre par un non expert. Ces formules peuvent être combinées avec le vocabulaire spécialisé (Kokkinakis et Gronostaj, 2006), afin de prendre en compte la dimension médicale du lexique. Une autre approche productive consiste en l'application d'algorithmes d'apprentissage et permet d'étudier les documents experts et vulgarisés de manière contrastive. Différents types de descripteurs sont alors exploités, comme par exemple : les n-grammes de caractères (Poprat et al., 2006), la pondération manuelle (Zheng et al., 2002) ou automatique (Borst et al., 2008) des termes médicaux, les critères stylistiques (Grabar et al., 2007) ou discursifs (Goeuriot et al., 2007) des documents et leur niveau lexical (Miller et al., 2007). L'accent est mis parfois sur la combinaison de différents descripteurs (Wang, 2006 ; Zeng-Treiler et al., 2007 ; Goeuriot et al., 2007 ; Leroy et al., 2008). Notons néanmoins que des études approfondies d'un type donné de descripteurs restent assez rares. À cet effet, citons par exemple des études assez détaillées des niveaux syntaxique (Zeng et Tse, 2006) et discursif (Goeuriot et al., 2007).

Nous proposons de situer nos expériences dans ce deuxième type de travaux liés à la comparaison et la distinction des degrés de spécialisation des documents à destination des experts et des patients. Nous

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100305

proposons surtout d'étudier les descripteurs du niveau morphologique de la langue au travers de trois registres : vulgarisation à destination des patients (*pat*), didactique à destination des étudiants (*étu*) et expert à destination des professionnels (*pro*). L'étude du niveau morphologique de la langue n'a pas encore été explorée dans le contexte ciblé ici. Par ailleurs, dans l'état de l'art, deux registres (professionnel et patient), et non les trois étudiés ici, sont opposés.

2 Objectifs scientifiques

L'objectif de ce travail consiste à étudier de manière détaillée les documents biomédicaux de la Toile qui se distinguent entre eux sur le plan de leur spécialisation et technicité. Ce travail se distingue des travaux existant dans l'état de l'art par quatre dimensions :

- L'objet de cette étude concerne le niveau morphologique de la langue. Notre hypothèse principale est que la morphologie représente un vecteur de différenciation important entre les documents qui montrent de différents degrés de spécialisation. Nous pensons en effet que les documents spécialisés contiennent à la fois plus de lexèmes composés et plus de composants et affixes d'origine grecque et latine. De plus, comme le montre l'état de l'art réalisé, cet aspect n'a pas été étudié par d'autres chercheurs.

- Nous proposons d'effectuer une étude quantitative et une étude qualitative.

- Nous proposons d'effectuer une comparaison entre les niveaux morphologique (grâce à l'accès aux composants et affixes) et lexical (en exploitant directement les mots des corpus).

- Finalement, nous proposons de travailler avec les documents de trois registres liés à la technicité des documents biomédicaux (patients, étudiants et professionnels) dans trois domaines de spécialités biomédicales (cardiologie, hématologie et pneumologie).

Parmi les procédés morphologiques, nous exploitons les procédés constructionnels tels que la composition (*pneumoconiose*, *cardiomégalie*) et l'affixation (*cardiaque*, *angioblastique)*. Il nous semble en effet que d'une part les procédés constructionnels sont très abondants dans les documents biomédicaux et que d'autre part leur emploi peut être corrélé aux registres étudiés (professionnel, étudiant et patient). Une étude parallèle de plusieurs domaines biomédicaux va renforcer ou modérer nos résultats. Dans nos travaux précédents, nous avons effectué une étude de faisabilité manuelle sur un échantillon (46 bases) des données biomédicales (Chmielik et Grabar, 2009a ; Chmielik et Grabar, 2009b), et une étude quantitative sur des données plus importantes (Chmielik et Grabar, 2011). Dans le travail actuel, nous effectuons une analyse qualitative, qui permet d'appréhender les résultats plus en détail.

Dans la suite de ce travail, nous décrivons comment nous préparons et étudions le matériel (section 3), ensuite nous présentons et discutons les résultats qualitatifs (section 4) et quantitatifs (section 5). Nous terminons avec des perspectives (sec. 6).

3 Collecte et préparation du corpus

Les corpus sont d'abord collectés et une série de traitements avec des outils de TAL (Traitement Automatique des Langues) est effectuée afin d'accéder au niveau morphologique. Ensuite, nous décrivons la réalisation de l'analyse quantitative pour une étude contrastive des documents biomédicaux professionnels, étudiants et patients, de même que leur analyse qualitative.

3.1 Collecte de documents biomédicaux et constitution de corpus

Nous exploitons la catégorisation selon les registres effectuée par les documentalistes du portail CISMeF. Cette caractérisation est effectuée selon la technicité de documents et leurs destinataires : grand public (documents patients), professionnels de santé (documents professionnels) et étudiants en médecine (documents étudiants). Plusieurs dizaines d'étiquettes sont utilisées actuellement par les documentalistes

de CISMeF pour effectuer cette annotation. Un travail spécifique a été ainsi effectué pour établir une correspondance entre ces étiquettes et les trois catégories considérées. Par exemple, les documents annotés comme *monographie*, *article de périodique* ou *avis de vigilance sanitaire* sont associés à la catégorie professionnelle ; les documents annotés comme *examen national classant* ou *questions à choix multiple* sont considérés comme des documents étudiants ; et les documents annotés comme *information patient et grand public* ou *brochure pédagogique pour les patients* sont associés à la catégorie patient. Deux filtrages sont effectués : sur la langue des documents (pour ne garder que les documents en français (Grefenstette et Nioche, 2000) car CISMeF propose également des documents en d'autres langues) et sur la taille des documents (pour éliminer les documents vides ou trop courts).

	Cardiologie (nombre de documents)		Pneumologie (nombre de documents)		Hématologie (nombre de documents)	
	Collecte	Filtrage	Collecte	Filtrage	Collecte	Filtrage
Professionnel (pro)	2 922	874	1 823	582	1 580	613
Étudiant (étu)	582	163	304	109	293	91
Patient (pat)	404	249	317	191	203	102

Tableau 1 : Taille des sous-corpus en fonction des spécialités biomédicales (cardiologie, pneumologie et hématologie) et de la technicité des documents (documents professionnels, étudiants et patients).

Dans le tableau 1, nous présentons la taille de nos données. Pour chaque domaine biomédical traité (cardiologie, pneumologie et hématologie) et pour chaque registre de documents (professionnel, étudiant et patient), nous indiquons les tailles de sous-corpus correspondants au moment de leur collecte et après les filtrages. À cette étape, nous pouvons faire plusieurs observations. D'une part, les corpus de type professionnel sont toujours les plus volumineux. Il en est de même pour les corpus du domaine de la cardiologie. D'autre part, le filtrage des données collectées réduit énormément la taille des corpus.

3.2 Accès au niveau morphologique des documents

Afin d'accéder aux données du niveau morphologique, nous exploitons plusieurs outils de TAL. Plus particulièrement, grâce à la lemmatisation et à l'analyse morphologique, nous pouvons nous concentrer sur les procédés de la morphologie constructionnelle. Une partie de cette chaîne de traitement a fait ses preuves dans un travail précédent (Fradin et al., 2008).

Pré-traitement des documents. Afin d'améliorer la qualité de l'étiquetage syntaxique, nous effectuons d'abord une segmentation adaptée aux documents biomédicaux en français (par exemple, avec la sauvegarde des composés de type *vertebro-médullaire* ou *anatomo-clinique* et des abréviations (*M.V.S.*)) et un pré-étiquetage avec un lexique du français. Le lexique utilisé pour le pré-étiquetage se compose de plus de 300 000 mots. Il est issu du projet UMLF (Zweigenbaum et al., 2005) et contient des mots de la langue générale et de la langue biomédicale.

Étiquetage morpho-syntaxique. Nous utilisons l'étiqueteur TreeTagger (Schmid, 1994), qui assigne à chaque mot d'un document une étiquette morpho-syntaxique et effectue la lemmatisation. Il s'agit d'un étiqueteur entraîné sur la langue générale et ses performances peuvent être moindres sur les documents d'une langue de spécialité, comme la biomédicine.

Lemmatisation et correction. Le lemmatiseur Flemm (Namer, 2000) reprend l'étiquetage et la lemmatisation proposés par TreeTagger et les corrige lorsque c'est possible. Dans tous les cas, Flemm ajoute des traits morphologiques supplémentaires. Par exemple, le mot *coronarien*, correctement étiqueté *ADJ* par TreeTagger, reçoit les traits complémentaires calculés par Flemm *ADJ:m:s* qui spécifient qu'il s'agit d'un adjectif masculin au singulier.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100305

3.3 Analyse des données

Nous proposons d'effectuer une analyse quantitative et une analyse qualitative en nous concentrant sur l'emploi des lemmes et des procédés morphologiques dans les corpus.

Analyse morphologique. L'analyseur morphologique DériF (Namer, 2009) effectue l'analyse des lemmes en fonction de leur structure morphologique. Nous utilisons la version distribuée en 2007. DériF génère quatre types d'information, que nous illustrons sur l'exemple de *angioblastique/ADJ* :

1) Calcul de l'arbre d'analyse morphologique d'un lemme étiqueté. Les bases et affixes détectés sont associés aux catégories syntaxiques des lexèmes qu'ils forment (ici, *ADJ).* Lorsqu'il s'agit d'une base supplétive, à laquelle ne correspond pas de mot-forme dans la langue française moderne, DériF assigne la catégorie syntaxique probable (*N** pour un nom) :

[[angi N] [blast N*] ique ADJ]*

2) Reprise de l'arbre sous forme de famille ordonnant l'ensemble des bases successives reconnues par l'analyseur :

(angioblastique/ADJ, [angi,N]:blast/N*)*

3) Représentation en langage naturel de la relation sémantique entre le lemme et ses bases (glose sémantique) :

"Qui est en relation avec cellule embryonnaire et vaisseau"

4) Description d'autres traits sémantiques acquis automatiquement. D'une part, nous pouvons y trouver les constituants des lemmes composés (*angi, blast, ique*), mais aussi le type sémantique du lemme (ici, il s'agit d'un terme d'anatomie) et des relations sémantiques possibles avec d'autres composants : *eql* pour la relation d'équivalence (par exemple, *blast* et *angéio* sont équivalents), *isa* pour la relation hiérarchique (*angéio* et *cyt* ont une relation hiérarchique entre eux), et *see* pour la relation voir aussi (par exemple, la relation voir aussi existe entre *angéio* et *bléph) :*

Constituants = /angi/blast/ique
Type = anatomie
Relations possibles = (eql:ang/blast, eql:angé/blast, eql:angéio/blast, eql:vas/blast, eql:vascul/blast, isa:ang/cyt, isa:angi/cyt, isa:angé/cyt, isa:angéio/cyt, isa:vas/cyt, isa:vascul/cyt, see:ang/bréph, see:angi/bréph, see:angé/bréph, see:angéio/bréph, see:vas/bréph, see:vascul/bréph).

Collecte du matériel morphologique. L'exploitation de l'arbre morphologique, comme *([[angi N*] [blast N*] ique ADJ])* pour *angioblastique/ADJ*, permet d'accéder aux composants (*angi*, *blast)* et aux affixes (*-ique*) des lexèmes analysés. Cette analyse arborée est effectuée par DériF pour les lexèmes composés, comme dans cet exemple, mais aussi pour les lexèmes affixés. Pour pouvoir accéder systématiquement à ces éléments morphologiques, nous exploitons donc cette analyse arborée.

Analyse quantitative. L'objectif de l'analyse quantitative consiste à effectuer une catégorisation automatique des documents selon leurs technicité. Ceci permet de vérifier s'il existe une corrélation entre les procédés morphologiques et les registres des documents biomédicaux. Si cette corrélation existe, le registre pourra être détecté assez aisément et la catégorisation automatique des documents pourra être effectuée avec des résultats performants. Lors de la réalisation de cette étape, nous combinons l'utilisation des données linguistiques (les lemmes ou leur analyse morphologique telle que décrite ci-dessus) et une méthode d'apprentissage supervisé (arbres de décision C4.5 (Quinlan, 1993) implémentés dans la plateforme Weka (Witten et Frank, 2005)). En apprentissage supervisé, les systèmes ont besoin d'exemples en entrée ou d'un corpus d'apprentissage, pour pouvoir construire un modèle de classification. Les exemples en entrée de notre expérience proviennent du portail CISMeF. Ce modèle est ensuite appliqué aux données de tests, qui correspondent à des données nouvelles et non encore vues par le système. Le système fait alors des prédictions sur des données nouvelles et sur leur catégorisation dans

les catégories prévues par le modèle. Ayant pour objectif d'étudier la technicité des documents, nous visons la détection automatique de trois registres déjà signalés auparavant :

– pro : documents hautement spécialisés à destination des professionnels,

– étu : documents didactiques moyennement spécialisés pour les étudiants,

– pat : documents vulgarisés et peu spécialisés à destination des patients.

Comme nous l'avons indiqué, nous exploitons deux types d'indices pour effectuer cette analyse quantitative : les lemmes et les procédés morphologiques. De plus, nous considérons deux ensembles de procédés : bases seulement (par exemple, *angi* et *blast* pour *angioblastique*) mais aussi les bases et les affixes (*angi, blast* et *-ique* pour *angioblastique*). Dans les deux cas, l'accès à ces procédés est effectué grâce aux outils de TAL et plus particulièrement grâce à l'analyseur morphologique DériF. Notons aussi que nous considérons soit les types de ces indices ou bien leurs occurrences. Dans le premier cas, seule la présence d'un lemme ou des lemmes comportant un procédé morphologique sera exploitée, alors que dans le deuxième cas, chaque apparition d'un lemme ou des lemmes comportant un procédé morphologique sera comptée. Par exemple, dans un des documents, la base *acte* apparaît dans deux lemmes (*action* et *activité*). En terme de type, elle reçoit la valeur de 2. Mais comme *action* apparaît 2 fois et *activité* 3 fois, la base *acte* reçoit la valeur de 5 en termes d'occurrences. Dans le premier cas, nous mesurons la variété lexicale ou la taille des familles morphologiques autour des bases (ou des affixes), tandis que dans le deuxième cas nous mesurons la fréquence des bases et des affixes.

Avec cette analyse quantitative, nous pouvons effectuer une évaluation de la performance que montre la méthode par apprentissage supervisé. La catégorisation manuelle des documents proposée par CISMeF correspond aux données de référence. Les mesures d'évaluation sont standard pour ce type de tâche : précision (pourcentage de documents correctement catégorisés parmi tous les documents assignés à une catégorie), rappel (pourcentage de documents correctement catégorisés par rapport aux documents qui doivent être assignés à une catégorie) et f-mesure (moyenne harmonique de la précision et du rappel). L'évaluation est effectuée avec les lemmes et avec les procédés morphologiques.

Analyse qualitative. L'objectif de l'analyse qualitative consiste à effectuer une étude plus détaillée de nos données et des résultats obtenus automatiquement. Si l'analyse quantitative permet de traiter les corpus de manière massive et de montrer la saillance du matériel linguistique, l'analyse qualitative permet d'aller plus au fond des raisons de la réussite ou de l'échec de la méthode par apprentissage. Nous proposons ainsi d'analyser les unités linguistiques saillantes pour la catégorisation des documents (les plus fréquentes, les hapax ou bien celles retenues par les algorithmes de catégorisation). Pour ceci, nous examinons les arbres de décision qui ont permis de faire la distinction entre les registres. Nous proposons également une analyse des unités linguistiques (lemmes, bases, affixes) provenant de différents registres et de différents corpus.

4 Analyse quantitative : étude contrastive automatique des documents

Les graphiques de la figure 2 montrent les performances des méthodes automatiques dans la distinction des registres des documents. Le graphique de gauche présente les résultats obtenus lorsque les types sont utilisés et le graphique de droite présente les résultats obtenus avec exactement la même configuration, si ce n'est que ce sont les occurrences qui sont utilisées. Sur chacun de ces deux graphiques, les points positionnent les performances pour les trois types de données linguistiques : les lemmes (points bleus), les bases (points rouges), et la combinaison des bases et des affixes (points verts). Pour observer la corrélation entre le matériel linguistique exploité et la performance de la détection automatique des registres, il est important d'examiner la position des points qui correspondent aux performances produites selon le type de matériel. Plus les points sont proches du coin droit supérieur meilleures sont les performances et plus saillant est le matériel linguistique. Globalement, nous pouvons voir que les performances sont assez respectables (pour les points bleus de lemmes), voire très performantes (pour les points verts et rouges qui correspondent au matériel morphologique). Ces observations amènent une discussion.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100305

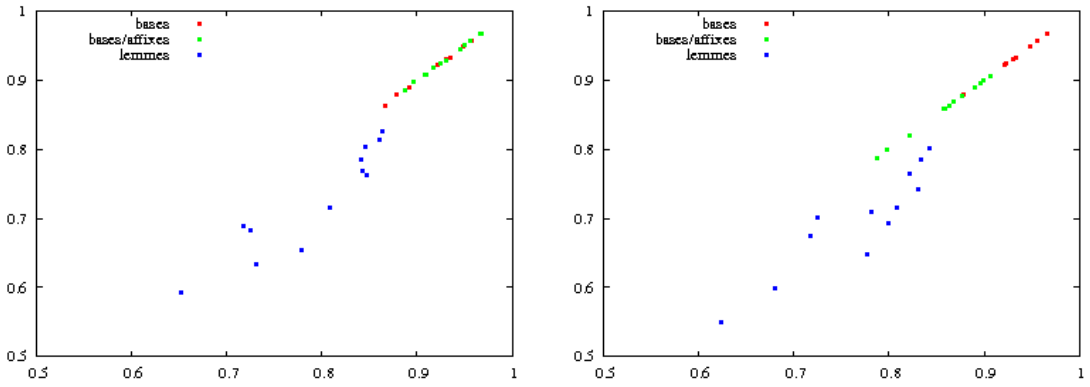

Figure 1 : Performance des méthodes automatiques par apprentissage supervisé dans la distinction des types de documents selon leur technicité.

- D'une part, les deux types d'unités morphologiques (bases seules ou combinaison des bases et des affixes) sont assez équivalents entre eux. Nous pouvons voir en effet que les poins correspondants à leur performances se superposent sur les graphiques.

- D'autre part, les unités lexicales (lemmes) produisent des résultats beaucoup moins performants que les procédés morphologiques. Nous pouvons donc voir que le matériel morphologique est plus saillant que le matériel lexical. Cette différence importante montre que les procédés morphologiques sont appropriés pour la tâche ciblée dans ce travail et qu'ils permettent de factoriser la représentation du contenu des documents. Grâce aux procédés morphologiques, nous pouvons ainsi appréhender le contenu des documents non à travers leurs formes de surface mais sur leur structure morphologique et sémantique interne.

- Finalement, nous observons une petite différence entre les types et les occurrences du matériel linguistique exploité. Les résultats sont légèrement meilleurs avec les types. Lorsque nous considérons les types, nous prenons en compte en réalité la taille des familles morphologiques et donc la variété lexicale. Il s'avère que la variété lexicale est en effet discriminante pour la distinction automatique de registres. Ainsi, non seulement les occurrences des mots et des unités morphologiques ont une importance mais également l'usage et la variété du vocabulaire des locuteurs.

Concernant ce dernier point, nous avons pu observer que les documents étudiants montrent la variété lexicale la plus grande (le plus grand nombre de types). L'objectif de ces documents consiste en effet à fournir des informations biomédicales précises et détaillées. En général, la vocation de ce type de documents étudiants est de présenter un grand nombre de notions médicales. En ce qui concerne les documents professionnels et patents, les premiers présentent plus d'occurrences de lexèmes, tandis que les deuxièmes montrent plus de variété.

Les trois figures qui suivent (figures 3, 4 et 5) illustrent les arbres de décision qui ont permis de faire la distinction avec les documents professionnels et les documents à destination des étudiants. Dans les trois cas, il s'agit de la même configuration et seules les unités linguistiques exploitées changent : lemmes (figure 3), bases (figure 4) et la combinaison des bases et des affixes (figure 5). Une toute première observation est que les algorithmes de classification choisissent les unités qui sont les plus saillantes statistiquement et non pas sur la base de leur sémantique. De ce point de vue, le fonctionnement est différent de l'approche manuelle, où la saillance sémantique serait privilégiée.

En ce qui concerne le contenu de ces arbres de décision, nous pouvons observer la différence entre leur complexité : l'arbre est très simple lorsque les unités lexicales sont exploitées (figure 3), avec seulement quatre lemmes (*abonnement*, *accréditation*, *accueil* et *accident*). Avec les unités morphologiques, les arbres sont beaucoup plus complets. Notons aussi que les performances se détériorent d'environ 0,10 % lorsque les lemmes sont exploités. Nous analysons ces arbres plus en détail dans la section suivante.

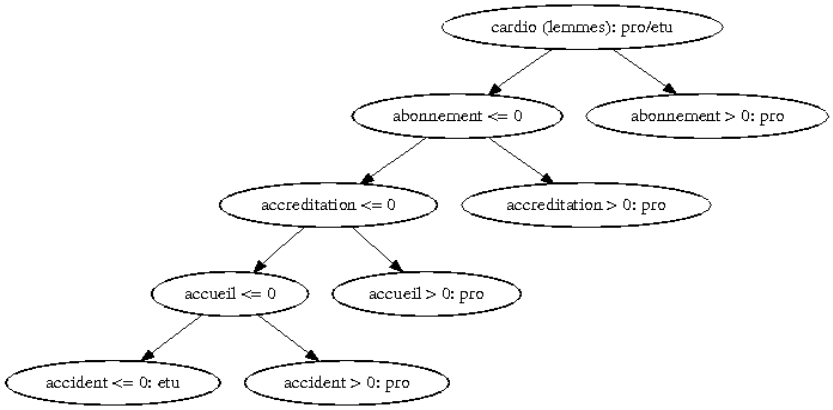

Figure 2 : Arbre de décision pour la distinction entre les documents professionnels et étudiants. La distinction automatique exploite les informations lexicales : occurrences des lemmes.

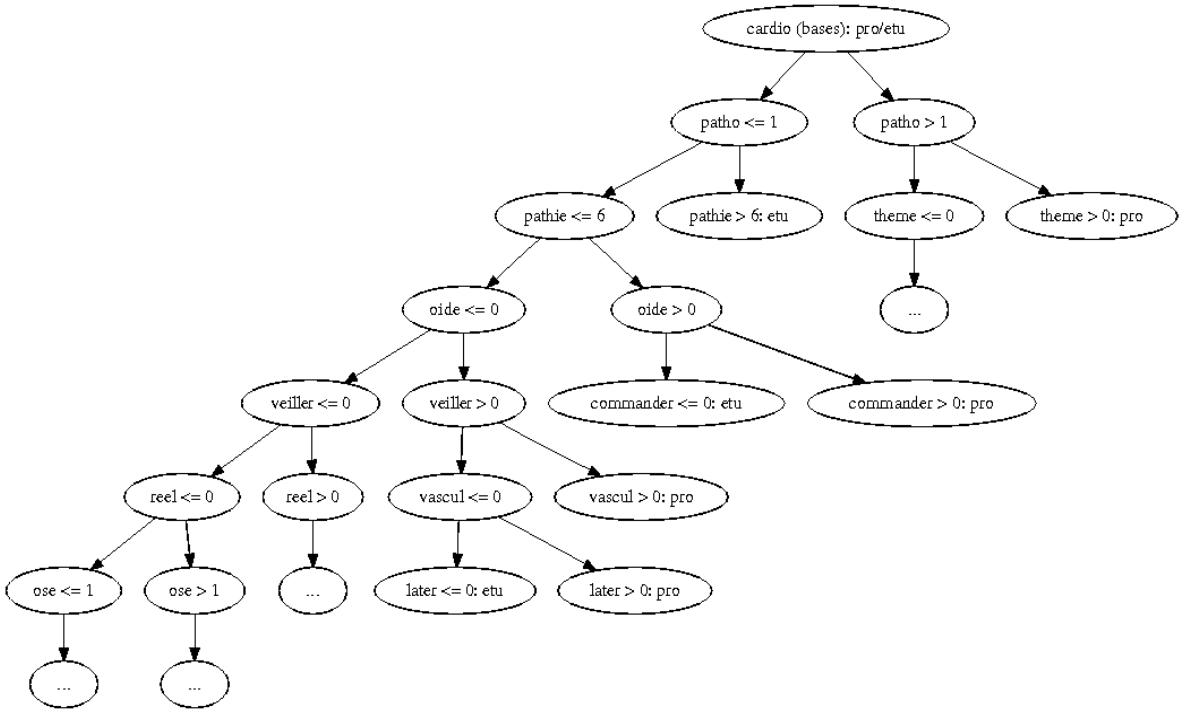

Figure 3 : Extrait de l'arbre de décision pour la distinction entre les documents professionnels et étudiants. La distinction automatique exploite les informations morphologiques : occurrences des bases.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100305

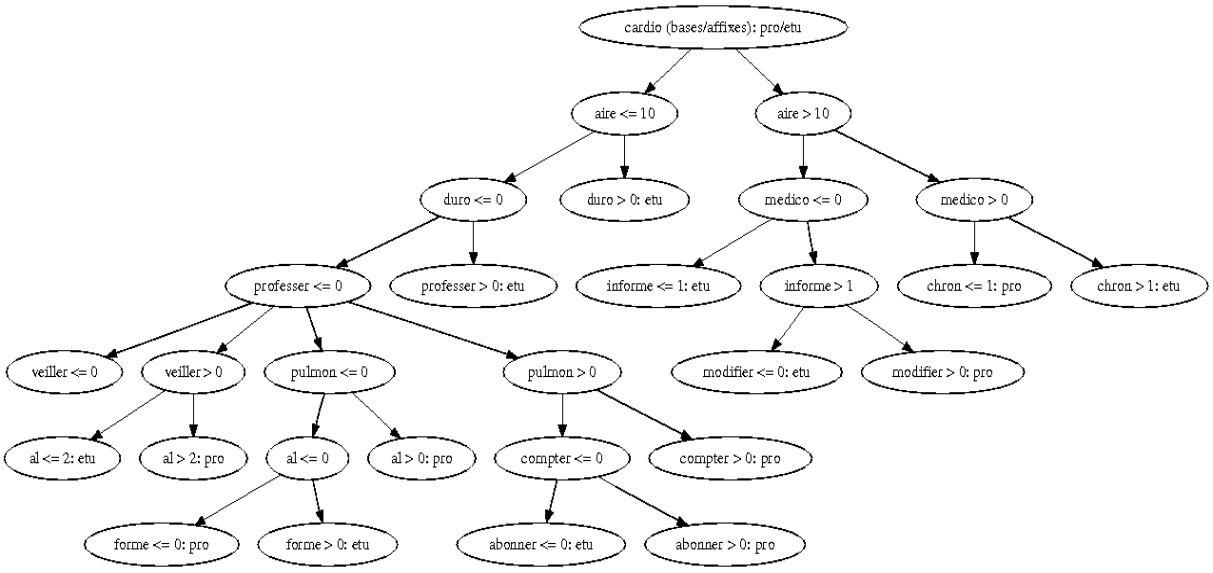

Figure 4 : Arbre de décision pour la distinction entre les documents professionnels et étudiants. La distinction automatique exploite les informations morphologiques : occurrences des bases et des affixes.

5 Analyse qualitative

Nous allons détailler et discuter les arbres de décision des figures 3 à 5. Avec les unités lexicales de la figure 3, nous avons déjà remarqué la pauvreté de la représentation des documents : seulement quatre unités lexicales sont retenues. Il s'agit de *abonnement*, *accréditation*, *accueil* et *accident*. De plus, nous pouvons aussi voir que le nombre d'occurrences de ces unités dans les corpus est bas, alors que le corpus consacré à la cardiologie est le plus volumineux. Effectivement, avec les valeurs égales à 0 ou supérieures à 0, il s'agit plutôt de l'absence ou de la présence des unités lexicales dans les documents. Avec la prise en compte des procédés morphologiques, nous constatons d'une part une plus grande variété des unités saillantes et d'autre part des nombres d'occurrences plus élevés pour plusieurs d'entre elles. Sur les arbres présentés, nous avons par exemple, *-aire* et *-pathie* dont les occurrences saillantes sont 10 et 6 respectivement. À côté de cela, nous remarquons aussi que :

1) les affixes (*-aire*, *-oïde* et *-al*) ont un rôle assez discriminant pour différencier les documents étudiants et professionnels ;

2) les éléments de composition (comme *-ose, vascul-, -pathie, -patho, médico-, pulmon-, chron-*) sont très présents dans les arbres de décision;

3) par ailleurs, les bases verbales (*compter, informe* (pour *informer*), *modifier, abonner, commander, veiller, professer*) ont aussi une place très importante. D'ailleurs la base verbale *abonner* semble être est le seul élément en commun avec l'arbre de décision lexical, qui comporte *abonnement.*

Les bases morphologiques détectées par Dérif peuvent être de deux types :

1) bases supplétives ou savantes, qui n'apparaissent pas de manière isolée dans la langue mais sont toujours couplées avec d'autres éléments morphologiques (p. ex : *-pathe, vascul-* ou *pulmon-*),

2) bases autonomes, qui peuvent se réaliser de manière isolée dans la langue (comme *actuel, thème* ou *forme* dans les trois arbres de décision présentés).

professionnel	*ion al is logie able techno ité alerte économie acte organe nation mentir*
étudiant	*ion patho traiter ose graphie génèr cardia thérapie isch thromb pulmon*
patient	*ion ique ité cardia prévent utile traiter infect guider informe allergie post*

Tableau 2 : Bases et affixes les plus fréquents selon les trois registres étudiés.

Dans notre travail, nous avons distingué 5 968 bases et affixes différents. Le tableau 2 présente ceux qui se trouvent parmi les plus fréquents selon les registres. Nous pouvons voir que les procédés d'affixation (-*al,* -*ion* et -*ique* dans ces exemples, mais ils sont bien plus nombreux) sont distribués dans tous les corpus. L'affixe -*is*, qui apparaît dans les noms et adjectifs d'origine latine, est fréquent dans les sous-corpus professionnel. Nous pouvons observer quelques verbes fréquents, comme *traiter* ou *guider*.

corpus	total	unique	exemples		
			étudiant	professionnel	patient
cardio	4 372	1 889	*jéjuno plausible*	*flavone ischio méiose*	*normat*
pneumo	4 260	1 899	*abort alopécie*	*amiodarone monét*	*spécv fécalv*
hémato	4 097	1 877	*abdomen naevo*	*cocaïne phrén*	*angél règle*
			cardio	hémato	pneumo
professionnel	4 887	1 786	*adipos brachy*	*ankylo carotène*	*agrégat aphte*
étudiant	3 926	1 285	*abduct crano mnés*	*coccygien*	*exérèse ptéryg*
patient	2 645	1 119	*adip graphe pexie*	*amnio angél*	*cili gnath*

Tableau 3 : Nombre total de bases et d'affixes, nombre de bases uniques et quelques exemples de bases uniques indiqués selon les corpus et les registres.

Dans le tableau 3, nous présentons une autre vue de la distribution des unités morphologiques dans les différents corpus et registres. Pour chaque domaine biomédical (cardio, pneumo et hémato) et pour chaque registre (professionnel, étudiant et patient), nous indiquons le nombre total de bases et d'affixes ainsi que le nombre de bases uniques dans un corpus. Notons que les affixes ne sont hapaxiques dans aucun des sous-corpus. Dans la dernière colonne, nous indiquons aussi quelques exemples de bases et affixes selon les sous-corpus où le procédé morphologique en question est unique. Par exemple, dans le corpus cardio : *jéjuno* et *plausible* apparaissent uniquement dans le sous-corpus étudiant; *flavone, ischio* et *méiose* dans le sous-corpus professionnel et *normat* dans le sous-corpus patient. Lorsque la distribution du matériel morphologique est indiquée par registre (professionnel, étudiant et patient), nous pouvons voir la présence de noms de médicaments ou de substances chimiques (*carotène, flavone, amiodarone* ou *cocaïne*) qui apparaissent uniquement dans les sous-corpus professionnels. Si nous faisons abstraction des domaines biomédicaux, nous avons 2 553 bases uniques aux différents registres : 1 517 pro, 803 étu et 233 vul. C'est donc le registre *pro* qui propose le nombre le plus grand de bases uniques. De manière

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100305

générale, c'est ce registre qui fournit le matériel morphologique le plus conséquent. Certaines des bases hapaxiques semblent sortir du champ sémantique des trois domaines étudiés ici (*alopécie, abort, vitréo, uréthro* ...). Dans d'autres cas, il s'agit de bases grecques, comme *sphygm*, signifiant « pulsation », qui semblent correspondre à une notion assez centrale en cardiologie.

Les hapax sont beaucoup plus nombreux au sein des unités lexicales : 14 120 hapax dans le registre étudiant, 38 318 dans le registre professionnel et 5 583 dans le registre patient. Parmi les hapax ou les unités lexicales les plus fréquentes, toute catégorie morpho-syntaxique est présente. Cet aspect souligne encore la capacité des procédés morphologiques à décomposer et factoriser le contenu des documents biomédicaux.

6 Conclusion et Perspectives

Nous avons effectué une étude qualitative et quantitative des documents biomédicaux de trois registres (patient, professionnel et étudiants). Nous effectuons une étude comparative et contrastée des caractéristiques des niveaux morphologique et lexical de ces documents. L'étude est effectuée grâce à une analyse manuelle des documents, mais aussi grâce à leur annotation et traitement avec des outils de TAL et les algorithmes d'apprentissage supervisé. Les mesures d'évaluation (précision, rappel et f-mesure) indiquent une bonne performance lorsque les unités morphologiques sont utilisées (souvent au-dessus de 0,90 %), ce qui signifie qu'il existe une très bonne corrélation entre les procédés morphologiques et les registres de documents. Les performances obtenues avec les unités lexicales sont beaucoup mois élevées (inférieures de 0,10 à 0,20 %), ce qui indique que les lemmes représentent moins bien le contenu des documents et des registres différenciés par la technicité des documents.

Les données et analyses produites dans notre travail peuvent être exploitées dans l'avenir pour aligner les séquences (essentiellement, des syntagmes nominaux ou verbaux) avec des sens proches dans deux ou même trois registres. Ce travail s'apparente au traitement et à l'alignement de corpus ou de phrases comparables, où la dimension commune correspond au domaine de spécialité des documents (cardiologie, hématologie ...) et la dimension de divergence correspond au registre de ces documents.

Nous avons noté plusieurs facteurs qui influencent les résultats, dans ce travail nous en avons souligné deux essentiellement : (1) la taille des familles morphologiques, et pas seulement les fréquences de lemmes et des unités morphologiques, joue un rôle important dans la distinction automatique ; (2) parmi les descripteurs morphologiques, les bases comme les affixes participent à la distinction entre les registres. En relation avec cette dernière observation, nous proposons d'étudier la productivité morphologique (Baayen, 1991), qui semble être spécifique des discours et genres dans d'autres domaines (Baayen, 1994).

Dans le domaine biomédical, les unités morphologiques semblent ainsi jouer un rôle assez central et leur considération permet d'améliorer la représentation des documents et leur typage. Nous prévoyons de vérifier ces observations dans d'autres domaines de spécialité, par exemple dans le domaine juridique (Fernandez-Barrera, 2011). Il sera intéressant de voir si dans ce domaine la morphologie occupe un rôle aussi important ou bien si c'est aux niveaux lexical et terminologique que s'opère la distinction entre les registres.

Dans l'avenir, nous allons aussi étudier l'interaction entre la morphologie et d'autres types d'unités et informations linguistiques. Par exemple, nous pouvons exploiter et combiner les unités provenant de niveaux lexical, syntaxique, stylistique.

La perspective principale de notre travail concerne l'utilisation des procédés morphologiques, et d'autres unités linguistiques saillantes, pour la distinction automatique de la spécialisation des documents dans un contexte applicatif, comme par exemple au sein d'un portail biomédical. Les résultats exploratoires obtenus dans le présent travail suggèrent en effet que l'exploitation des procédés morphologiques pourront fournir des résultats assez fiables et assister les documentalistes qui effectuent cette annotation.

Références

AMA (1999). Health literacy: report of the Council on Scientific Affairs. Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs, American Medical Association, *JAMA*, 281(6) : 552-7.

Baayen H. (1999). Quantitative aspects of morphological productivity, *Yearbook of Morphology* : 109-149.

Baayen H. (1994). Derivational productivity and text typology, *Journal of quantitative linguistics*, 1(1) : 16-34.

Berland G., Elliott M., Morales L., Algazy J., Kravitz R., Broder M., Kanouse D., Munoz J., Puyol J., Lara M., Watkins K., Yang H., McGlynn E. (2001). Health information on the Internet. Accessibility, quality, and readability in English ans Spanish, *JAMA* 285(20) : 2612-2621.

Björnsson H., Härd af Segerstad B. (1979). Lix pä franska och tio andra språk, Stockholm: Pedagogiskt centrum, Stockholms skolförvaltning.

Borst A., Gaudinat A., Boyer C., Grabar N. (2008). Lexically based distinction of readability levels of health documents, MIE 2008. Poster.

Chmielik J., Grabar N. (2009a). Comparative study between expert and non-expert biomedical writings: their morphology and semantics, *Stud Health Technol Inform*. 150 : 359-63.

Chmielik J., Grabar N. (2009b). Étude contrastive des documents vulgarisés et scientifiques de santé: sur la piste de la morphologie, JFIM, Informatique et Santé, Springer-Verlag France, chapter XVII.

Chmielik J., Grabar N. (2011). Détection de la spécialisation scientifique et technique des documents biomédicaux grâce aux informations morphologiques, *Traitement Automatique des Langues*:52(2): 151-179.

CHV. Consumer Health Vocabulary : www.consumerhealthvocab.org

CISMEF. Catalogue et Index des Sites Médicaux de langue Française : www.chu-rouen.fr/cismef

COOP. www.google.com/coop : Accès sur abonnement

Darmoni S., Leroy J., Baudic F., Douyère M., Piot J., Thirion B. (1999). CISMeF: cataloque and index of French speaking health resources, *Stud Health Technol Inform* : 493-6.

Deléger L., Zweigenbaum P. (2008). Paraphrase acquisition from comparable medical corpora of specialized and lay texts, AMIA 2008 : 146-50.

Fernandez-Barrera M. (2011). Exploring the terminological nature of citizens' queries in the domain od consumer justice, TIA 2011 : 52-58.

Flesch R. (1948). A new readability yardstick, *Journal of Applied Psychology* 23 : 221-233.

Fox S. (2006). Online Health Search 2006. Most Internet users start at a search engine when looking for health information online. Very few check the source and date of the information they find, Technical report, Pew Internet & American Life Project, Washington DC.

Fox S. (2011). Health topics. 80% of internet users look for health information online, Technical report, Pew Internet & American Life Project, Washington DC.

Fradin B., Dal G., Grabar N., Namer F., Lignon S., Tribout D., Zweigenbaum P. (2008). Remarques sur l'usage des corpus en morphologie, *Langages* 171(3) : 34-59.

Goeuriot L., Grabar N., Daille B. (2007). Caractérisation des discours scientifique et vulgarisé en français, japonais et russe. TALN : 93-102.

Grabar N., Krivine S., Jaulent M.-C. (2007). Classification of Health Webpages as Expert and Non Expert with a Reduced Set of Cross-language Features, AMIA 2007 : 284-8.

Grefenstette G., Nioche J. (2000). Estimation of English and non-English language use on the WWW, Recherche d'Information Assistée par Ordinateur (RIAO) : 237-246.

Gunning R. (1973). The art of clear writing, New York, NY :McGraw Hill.

HON. Health on the Net : www.hon.ch

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100305

Kokkinakis D., Gronostaj M. T. (2006). Comparing Lay and Professional Language in Cardiovascular Disorders Corpora, in A. Pham T., James Cook University (ed.), WSEAS Transactions on BIOLOGY and BIOMEDICINE : 429-437.

Leroy G., Helmreich S., Cowie J., Miller T., Zheng W. (2008). Evaluating Online Health Information: Beyond Readability Formulas, AMIA 2008 : 394-8.

McCray A. (2005). Promoting Health Literacy, *Journal of American Medical Informatics Association* 12 : 152-163.

Messai R., Zeng Q., Mousseau M., Simonet M. (2006). Building a Bilingual French-English PatientOriented Terminology for Breast Cancer, MedNet.

Miller T., Leroy G., Chatterjee S., Fan J., Thoms B. (2007). A Classifier to Evaluate Language Specificity of Medical Documents, HICSS : 134-140.

Namer F. (2000). FLEMM : un analyseur flexionnel du français à base de règles. Traitement automatique des langues 41(2) :523-547

Namer F. (2009). Morphologie, Lexique et TAL : l'analyseur DériF: TIC et Sciences cognitives, London : Hermes Sciences Publishing.

Poprat M., Markó K., Hahn U. (2006). A Language Classifier that Automatically Divides Medical Documents for Experts and Health Care Consumers, MIE 2006 : 503-508.

Quinlan J. (1993). C4.5 Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA.

Risk A., Dzenowagis J. (2001). Review of Internet information quality initiatives, *Journal of Medical Internet Research*.

Schmid H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees, Proceedings of the International Conference on New Methods in Language Processing : 4449.

Tran MT., Chekroud H., Thiery P., Julienne A. (2009). Internet et soins : un tiers invisible dans la relation médecine/patient ?, *Ethica Clinica* 53 : 34-43.

Wang Y. (2006). Automatic recognition of text difficulty from consumers health information, in IEEE (ed.), Computer-Based Medical Systems : 131-136.

Witten I., Frank E. (2005). Data mining: Practical machine learning tools and techniques. San Francisco : Morgan Kaufmann.

Yang Y., Liu X. (1999). Re-examination of text categorisation methods, Proc of 22nd Annual International SIGIR : 42-49.

Zeng Q., Tse T. (2006). Exploring and developing Consumer Health Vocabularies, *JAMIA* 13 : 24-29.

Zeng-Treiler Q., Kim H., Goryachev S., Keselman A., Slaugther L., Smith C. A. (2006). Text characteristics of clinical reports and their implications for the readability of personal health records, MEDINFO, IOS Press : 1117-1121.

Zheng W., Milios E., Watters C. (2002). Filtering for medical news items using a machine learning approach, AMIA : 949-53.

Zweigenbaum P., Baud R., Burgun A., Namer F., Jarrousse E., Grabar N., Ruch P., Le Duff F., Forget J., Douyère M., Darmoni S. (2005). UMLF: a unified medical lexicon for French, *Int J Med Inform* 74(2-4) : 119-24.