Détection de la spécialisation scientifique et technique des documents biomédicaux grâce aux informations morphologiques

Jolanta Chmielik* **– Natalia Grabar****

** Commission européenne, DG RTD, recherche et innovation, Bruxelles*
jolachmielik@hotmail.com
*** UMR 8163 STL, CNRS & U de Lille 3 et Lille 1, Villeneuve d'Ascq*
natalia.grabar@univ-lille3.fr

RÉSUMÉ. *La distinction du degré de spécialisation des documents de santé en ligne est une indication importante, surtout lorsque ces documents sont consultés par des utilisateurs non experts, comme le sont les patients. Effectivement, une très grande technicité des documents empêche les patients de bien comprendre le contenu et peut avoir des conséquences négatives sur la gestion de leur maladie et la communication avec les médecins. Lorsque les portails de santé proposent cette distinction, elle est effectuée manuellement. Nous effectuons une catégorisation automatique des pages de la Toile en fonction de leur spécialisation. Nous exploitons l'information morphologique obtenue grâce à l'analyse morphologique des lexèmes. L'évaluation montre que la précision, le rappel et la f-mesure sont souvent supérieurs à 90 %.*

ABSTRACT. *Distinction of the specialization level of the health documents on Internet is an important indication, especially when documents are read by non expert users such as patients. Indeed, a high technicity of documents impedes the patients to understand the content and may have a negative consequence on their health care process and on their communication with medical doctors. When medical portals propose such a distinction, it is obtained further to a human categorisation. We propose an automatic categorization of health documents according to their specialization. We exploit morphological information obtained thanks to the morphological analysis of lexems. The evaluation shows that precision, recall and f-measure are often higher than 90%.*

MOTS-CLÉS : *documents médicaux, spécialisation, apprentissage supervisé, morphologie constructionnelle, sémantique.*

KEYWORDS: *medical documents, specialization, supervised machine learning, contructional morphology, semantics.*

TAL. Volume 52 – n° 2/2011, pages 151 à 179

1. Contexte

Les analyses régulières de l'utilisation de la Toile montrent qu'environ 80 % des internautes s'intéressent aux questions liées au domaine biomédical (Fox, 2006 ; Fox, 2011). Ce chiffre souligne la préoccupation que manifestent les citoyens vis-à-vis de leur santé. Il montre également que la Toile propose d'énormes volumes d'informations consacrées à la santé. La qualité de ces documents n'est pourtant pas égale. Du point de vue des utilisateurs non experts, comme le sont typiquement les patients, différents aspects liés à la qualité peuvent être distingués, parmi lesquels citons par exemple :

– la qualité scientifique qui correspond à la correction médicale des informations ;

– l'aspect éthique lorsqu'il est lié, par exemple, à la publicité des traitements non prouvés ou à des conseils de soins délivrés par une personne non spécialiste en médecine ;

– l'opacité ou l'accessibilité technique, dont dépend très souvent la compréhension du contenu par les patients ;

– la présentation visuelle des informations, comme par exemple la taille ou le contraste des caractères pour les malvoyants.

Ces aspects sont différemment pris en charge par des initiatives qui existent dans ce domaine (Risk et Dzenowagis, 2001). Par exemple, deux portails de santé, CISMeF[1] et HON[2], qui sont les deux initiatives les plus présentes en France, se concentrent essentiellement sur la qualité scientifique et la transparence éthique des pages de santé : les documentalistes et les experts du domaine biomédical effectuent une analyse des pages à indexer ou à certifier. Notons qu'une assistance automatique à la certification peut être proposée (Gaudinat *et al.*, 2007 ; Névéol *et al.*, 2007).

Dans notre travail, nous nous intéressons à l'aspect lié à l'étude et à la détection automatique de l'accessibilité technique des documents de santé et du niveau de leur spécialisation. Cet aspect n'est pas dépendant des initiatives qui assurent la démarche qualité des pages de santé : il est propre à tout document produit dans le domaine biomédical. En effet, les documents appartenant à différents registres de spécialisation coexistent sur la Toile. Par exemple, quatre types de situations communicatives peuvent être distingués : expert/expert, expert/initié, expert/non-initié et enseignant/élève (Pearson, 1998). En relation avec ces situations communicatives, trois types de documents sont très présents sur la Toile dans le domaine biomédical. Ces trois types de documents sont : des documents scientifiques spécialisés (créés par des professionnels de santé à destination des professionnels de santé ou des initiés), des documents vulgarisés (créés à destination des usagers non experts en médecine), ainsi que des documents biomédicaux de type didactique qui s'adressent aux étudiants en médecine. Nous pouvons en effet regrouper les deux premières situations, expert/expert et expert/initié, où les utilisateurs initiés se rapprochent des experts au

1. Catalogue et index des sites médicaux de langue française : www.chu-rouen.fr/cismef
2. Health on the Net : www.hon.ch

moins dans certains domaines. Cette hétérogénéité technique peut avoir des effets néfastes et élitistes, surtout lorsqu'elle n'est pas signalée aux usagers. Il a été ainsi constaté que si le degré de technicité est élevé, cela peut avoir un impact négatif sur la compréhension des informations par les patients et par la suite déteriorer la communication des patients avec les médecins, de même que la réussite des soins médicaux qui sont administrés aux patients (AMA, 1999 ; McCray, 2005 ; Thi Tran *et al.*, 2009). 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). La spécialisation et l'hétérogénéité technique des documents ne sont pas évidentes pour les internautes, alors qu'il pourrait être utile de les indiquer de manière explicite. Notons que les portails médicaux comme HON, CISMeF ou le moteur de recherche généraliste GoogleCoopsanté[3] proposent cette distinction mais exploitent essentiellement une catégorisation manuelle des pages et des sites de la Toile.

2. Objectifs scientifiques

L'objectif de ce travail consiste à proposer des méthodes automatiques pour arriver à effectuer la distinction de la spécialisation et de la technicité des documents que l'on rencontre sur la Toile. Cette catégorisation, lorsqu'elle est indiquée de manière explicite, peut être exploitée pour guider les utilisateurs non experts vers des sources d'informations qui leur sont plus appropriées. Quant à l'exploitation de méthodes automatiques, elle trouve sa justification dans le fait que le volume d'informations augmente sans cesse sur la Toile rendant impossible la gestion manuelle du fonds documentaire.

Le contenu et le style des documents de santé sont marqués par le contexte de leur production et de leur usage (Biber, 1994 ; Zweigenbaum *et al.*, 2001). Par exemple, le destinataire et le producteur du document, de même que les objectifs poursuivis lorsque le document a été créé, y laissent leurs traces. Nous proposons de nous centrer sur le contenu des documents pour faire émerger les descripteurs qui permettraient de discriminer leur degré de spécialisation. Notre intérêt porte plus spécifiquement sur la morphologie de la langue biomédicale. Nous proposons donc d'effectuer une étude des caractéristiques morphologiques de trois registres liés à la spécialisation des documents biomédicaux (vulgarisés, didactiques et experts) dans trois domaines de spécialités biomédicales (cardiologie, hématologie et pneumologie) et d'exploiter des critères morphologiques pour permettre la distinction automatique de ces registres. Ce travail repose sur l'hypothèse qu'il existe un lien entre la spécialisation des textes et l'emploi des procédés morphologiques. Plus particulièrement, nous supposons que ce lien est fort au niveau des procédés constructionnels tels que la composition (*pneumo-*

3. www.google.com/coop : accès sur abonnement.

coni-ose, *cardio-mégalie*) 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 (expert, didactique et patient). Une étude parallèle de plusieurs domaines biomédicaux va renforcer ou modérer nos résultats. Nous n'étudions pas la morphologie flexionnelle (*angioblastiques*, *cardiomégalies*) dans ce travail. Dans un travail précédent (Chmielik et Grabar, 2009a ; Chmielik et Grabar, 2009b), nous avions effectué une étude de faisabilité manuelle sur un échantillon des données biomédicales (46 bases), en étudiant plus particulièrement les fréquences des bases en corpus. Dans le travail actuel, les corpus sont différents et une automatisation complète de la chaîne de traitement permet de traiter des données plus volumineuses. De plus, l'aspect lié à la catégorisation automatique n'était pas abordé dans le travail précédent.

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

3. État de l'art

Deux principaux types de travaux existent autour de l'étude de la spécialisation des documents biomédicaux et de leur accessibilité par les patients : (1) l'établissement de lexiques monolingues[4] ou bilingues (Messai *et al.*, 2006 ; Deléger et Zweigenbaum, 2008) mettant en contrast le vocabulaire experts et patients, et (2) la comparaison et la distinction des degrés de spécialisation des documents à destination des experts et des patients. C'est le deuxième aspect qui nous intéresse ici et nous présentons son état de l'art plus en détail.

Parmi de nombreux travaux qui étudient la comparaison et la distinction des degrés de spécialisation, les 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. Ainsi, si les mots et les phrases sont longs, 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

4. www.consumerhealthvocab.org

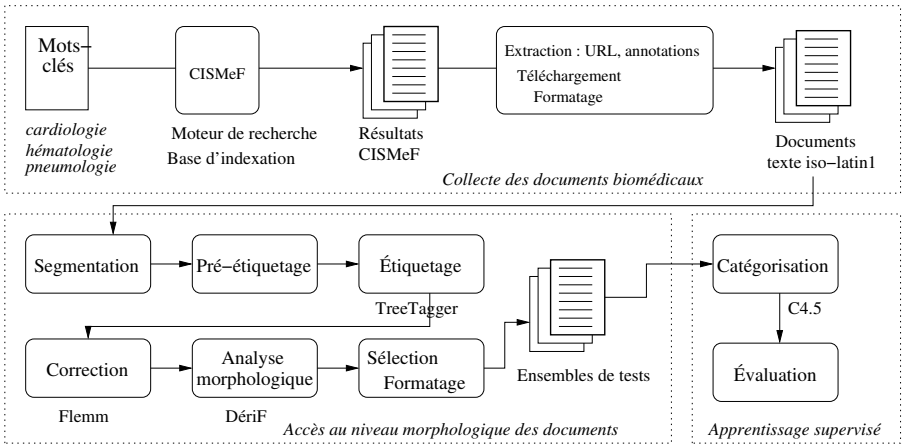

Figure 1. *Schéma général de la méthode*

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). Néanmoins, 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). Dans notre travail, nous proposons d'étudier les descripteurs du niveau morphologique et de nous concentrer également sur trois registres : patient (*vul*), didactique (*étu*) et expert (*pro*), tandis que dans l'état de l'art la tradition est d'opposer deux registres (expert et patient).

4. Collecte et préparation du matériel

Nous présentons ici le matériel exploité dans notre expérience. Les étapes de ce processus sont présentées dans la figure 1. Les corpus sont d'abord collectés (section 4.1) et une série de traitements avec des outils de TAL est effectuée afin d'accéder au niveau morphologique (section 4.2). Ensuite, nous décrivons la réalisation de l'étude constrastive des documents biomédicaux experts, didactiques et vulgarisés (section 4.3). Nous utilisons ainsi plusieurs outils de TAL, des modules et scripts Perl et une implémentation de l'algorithme d'apprentissage supervisé `C4.5`.

4.1. *Collecte de documents biomédicaux et constitution de corpus*

4.1.1. *Portail CISMeF*

Nous exploitons le portail CISMeF[5], qui propose des ressources biomédicales de qualité, pour collecter les documents en français pour notre étude. Les documentalistes de CISMeF travaillent sur le recensement et l'annotation des documents à indexer. Plusieurs annotations sont associées aux documents, parmi lesquelles nous exploitons :

1) les adresses URL des ressources indexées ;

2) la caractérisation des documents selon leurs domaines biomédicaux, définie grâce à leur indexation avec les mots-clés MeSH (NLM, 2001). Le thésaurus MeSH est conçu spécifiquement pour l'indexation et la recherche des documents biomédicaux. Ce thésaurus, créé à l'origine en langue anglaise, est maintenant traduit en plusieurs autres langues, dont le français. Ceci permet de l'exploiter dans des bases de données et les portails francophones. Au sein de CISMeF, l'annotation des mots-clés MeSH est proposée de manière automatique (Névéol *et al.*, 2007). Elle est ensuite validée par les documentalistes. Grâce à cette annotation, nous pouvons aisément extraire les documents relatifs aux trois spécialités médicales étudiées : cardiologie (étude du cœur et de ses maladies), pneumologie (étude des poumons et de leurs maladies) et hématologie (étude du sang et de ses maladies) ;

3) la caractérisation des documents selon leur spécialisation et les types de pratiques sociales : grand public (documents vulgarisés), professionnels de santé (documents spécialisés ou experts) et étudiants en médecine (documents didactiques). 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 *pro*, les documents annotés comme *examen national classant* ou *questions à choix multiple* sont considérés comme des documents *étu*, et les documents annotés comme *information patient et grand public* ou *brochure pédagogique pour les patients* sont associés à la catégorie *vul*.

4.1.2. *Téléchargement des documents et homogénéisation de leur format*

Une fois les URL et les annotations associées collectées, les documents sont téléchargés. Nous utilisons l'*aspirateur de Web* `wget`, paramétré pour télécharger uniquement les URL de la liste, et non leurs filles, ce qui permet de mieux contrôler la corrélation entre l'annotation des documents proposée par CISMeF et leur contenu. Les documents ainsi téléchargés sont ensuite testés par l'utilitaire Linux `file`, qui permet de détecter le format des documents. Les documents au format HTML ou XML, que l'on peut convertir en texte plus facilement, sont ainsi détectés et sélectionnés. Grâce à l'annotation CISMeF, ces documents sont regroupés en des ensembles différenciés en fonction de leur spécialité médicale et leur annotation comme *expert*, *étudiant* et

5. Catalogue et index des sites médicaux de langue française : www.chu-rouen.fr/cismef

	Cardio		Pneumo		Hémato	
	doc.	occ.	doc.	occ.	doc.	occ.
pro	2 922	1 285 665	1 823	1 265 726	1 580	1 512 064
étu	582	384 550	304	229 639	293	198 264
vul	404	253 402	317	189 205	203	100 126
Total	3 908	1 923 617	2 444	1 684 570	2 076	1 810 454

Tableau 1. *Nombre de documents et d'occurrences de mots dans chaque corpus et sous-corpus, à l'étape de téléchargement et d'homogénéisation de leur format*

	Cardio		Pneumo		Hémato	
	taille	langue	taille	langue	taille	langue
pro	1 064	874	674	582	768	613
étu	326	163	109	109	92	91
vul	504	249	195	191	103	102

Tableau 2. *Nombre de documents dans les sous-corpus à l'étape des filtrages : par rapport à la taille et ensuite par rapport à la langue. Ces données sont obtenues suite aux filtrages appliqués aux données du tableau 1.*

vulgarisé. Au total, nous disposons des trois corpus, cardiologie, pneumologie et hématologie, dont chacun se compose de trois sous-corpus, *pro*, *étu* et *vul*. Au sein de chaque sous-corpus, les documents téléchargés sont convertis en texte et leur encodage est homogénéisé vers `ISO-8859-1`, pour rendre possible l'application des outils de TAL (section 4.2). La détection du codage courant des documents est également effectuée avec l'utilitaire Linux `file`, qui en plus du format permet aussi de détecter plusieurs encodages possibles, comme par exemple *ASCII*, *ISO-8859-**, *UTF-8*, ou *extended-ASCII*. Le recodage des documents vers *iso-latin1* est effectué avec un autre utilitaire Linux `recode`. Le tableau 1 indique la taille des sous-corpus au moment de téléchargement et d'homogénéisation de leur format (en nombre de documents et d'occurrences de mots). À la dernière ligne du tableau, sont fournies les tailles totales des trois corpus. À cette étape, nous pouvons observer deux déséquilibres dans les données : (1) les sous-corpus *pro* sont plus grands que les deux autres sous-corpus ; (2) le corpus *cardiologie* est plus volumineux que les deux autres corpus.

4.1.3. *Filtrage et sélection des documents*

Nous effectuons ensuite des filtrages liés à la taille des documents et à leur langue. Le filtrage lié à la taille des documents permet de ne pas prendre en compte les documents vides (il s'agit souvent des URL qui n'existent plus) ou des documents trop courts. Dans ce dernier cas, les documents peuvent contenir essentiellement le péritexte (Genette, 1987 ; Baroni *et al.*, 2008) qui sert à la navigation ou à l'organisation

d'un site mais qui ne correspond pas au contenu du document à proprement parler. Suite à une analyse de la taille des documents, la taille minimale en nombre de mots est fixée à 250. Notons que dans cette tâche, notre objectif n'est pas de séparer le péritexte du contenu informationnel d'une page, comme c'est le cas dans les campagnes d'évaluation WAC *CleanEval*[6] décrites dans (Baroni *et al.*, 2008), mais plus simplement de ne pas prendre en compte les pages où seul le péritexte est présent.

Un filtrage supplémentaire concerne la langue des documents. CISMeF est dédié à l'indexation des documents francophones et nous nous attendons à ce que tous les documents collectés au travers de CISMeF soient en français. Or, nous nous sommes rendu compte de la présence de documents en anglais. Nous effectuons donc un filtre sur la langue en exploitant les mots grammaticaux les plus fréquents et les plus spécifiques (Grefenstette et Nioche, 2000) pour trois langues (français, anglais et allemand). Par exemple, nous utilisons *that*, *you*, *have* pour l'anglais et *dans*, *une*, *qui* pour le français. Sur la base de ces marqueurs, nous calculons le pourcentage de textes de chacune de ces langues dans un document. Nous retenons un document s'il contient au moins 70 % de textes en français.

Le tableau 2 présente le nombre de documents dans chaque sous-corpus suite à ces deux filtrages. La colonne *langue*, qui intervient après le filtre *taille*, indique le nombre de documents disponibles après les filtres. Là encore, nous pouvons observer des déséquilibres dans les données : (1) les sous-corpus *pro* restent plus grands que les deux autres ; (2) le corpus *cardiologie* reste aussi plus volumineux, même si la différence est moins importante qu'auparavant (tableau 1).

4.2. *Accès au niveau morphologique des documents textuels*

Afin d'accéder aux données du niveau morphologique des documents, 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).

4.2.1. *Prétraitement des documents*

Afin d'améliorer la qualité de l'étiquetage syntaxique présenté dans le paragraphe suivant (section 4.2.2), 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.*) même si ces dernières ne sont pas exploitées) et un pré-étiquetage avec un lexique du français. La segmentation est effectuée avec un script Perl. 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.

6. http ://cleaneval.sigwac.org.uk/

4.2.2. *Étiquetage morphosyntaxique*

Nous utilisons l'étiqueteur `TreeTagger` (Schmid, 1994), qui assigne à chaque mot d'un document une étiquette morphosyntaxique 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.

4.2.3. *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.

4.2.4. *Analyse morphologique*

L'analyseur morphologique `DériF` (Namer, 2009) effectue ensuite l'analyse des lemmes en fonction de leur structure morphologique. Nous utilisons la version distribuée en 2007 de cet analyseur. `DériF` génère quatre types d'informations, 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 à leur catégorie syntaxique (*ADJ*). Lorsqu'il s'agit d'une base supplétive, qui n'existe pas 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).

4.2.5. *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. Cette tâche est effectuée avec un script Perl.

4.3. *Catégorisation automatique des documents*

La catégorisation automatique des documents selon leurs degrés de spécialisation est effectuée avec une approche par apprentissage supervisé et selon les étapes et les points méthodologiques décrits dans la suite de cette section. Ces étapes sont effectuées au sein de chaque corpus (*cardio*, *hémato*, *pneumo*) mais aussi sur la totalité de ces corpus (*total*).

4.3.1. *Algorithmes d'apprentissage supervisé*

Nous exploitons un algorithme d'apprentissage supervisé pour vérifier nos hypothèses et pour observer s'il existe effectivement une corrélation entre les procédés morphologiques et les catégories relatives à la spécialisation des documents biomédicaux. Si cette corrélation existe, le degré de spécialisation des documents pourra être détecté assez aisément et leur catégorisation automatique pourra être effectuée avec des résultats performants.

En apprentissage supervisé, les systèmes ont besoin d'exemples en entrée ou d'un corpus d'apprentissage, pour pouvoir construire un modèle de classification. 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. Dans notre expérience, nous poursuivons l'objectif de détecter trois catégories de documents, selon leurs registres (section 4.3.2). L'efficacite du modèle dépend notamment beaucoup des descripteurs que l'on fournit (section 4.3.3). Nous utilisons l'algorithme d'apprentissage supervisé qui applique les arbres de décision `C4.5` (Quinlan, 1993), tel qu'il est implémenté dans la plate-forme `Weka` (Witten et Frank, 2005), et dont nous avons gardé le paramétrage par défaut.

4.3.2. *Catégories à reconnaître*

Nous cherchons à distinguer automatiquement trois catégories de documents :

– *pro* : documents hautement spécialisés à destination des experts ;

– *étu* : documents didactiques moyennement spécialisés pour les étudiants ;

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

Nous effectuons une catégorisation bicatégorie en testant les trois couples possibles de catégories (la troisième catégorie n'est alors pas prise en compte) :

– *pv* : catégorisation des documents *pro* à destination des experts et des documents *vul* à destination des patients ;

– *pe* : catégorisation des documents *pro* à destination des experts et des documents *étu* à destination des étudiants ;

– *ve* : catégorisation des documents *vul* à destination des patients et des documents *étu* à destination des étudiants.

Nous effectuons aussi un test multicatégorie, où les trois catégories de documents sont exploitées en même temps :

– *tri* : catégorisation des trois types de documents *pro*, *étu* et *vul*.

Nous avons vu dans le tableau 2 que les données ne sont pas équilibrées entre les différents sous-corpus. Afin de ne pas introduire un biais lors de la catégorisation, nous effectuons un nivellement des données. Pour chacune des combinaisons présentée ci-dessus, le nombre de documents au sein de chaque sous-corpus est réduit à la taille du plus petit sous-corpus. Par exemple, lors de l'expérience multicatégorie pour le corpus *cardio*, après le filtrage sur la taille et la langue des documents, nous obtenons 874 documents *pro*, 163 documents *étu* et 249 documents *vul*. Le nombre de documents à traiter avec l'apprentissage supervisé est donc limité à 163 pour chaque catégorie. Au sein de la plus petite catégorie, tous les documents sont retenus, quant aux autres catégories, le choix des documents est fait aléatoirement.

4.3.3. *Choix des descripteurs*

Les descripteurs que nous exploitons proviennent du niveau morphologique de la langue. Deux ensembles de descripteurs sont testés dans notre étude, exemplifiés ici pour *angioblastique/ADJ* :

– *b* : bases (ou composants) des lexèmes analysés : *angi* et *blast* ;

– *ba* : bases et affixes des lexèmes analysés : *angi*, *blast* et *-ique*.

Dans les deux cas, l'accès à ces descripteurs est effectué grâce aux outils de TAL et plus particulièrement grâce à l'analyseur morphologique `DériF`.

4.3.4. *Pondération des descripteurs*

Les descripteurs sont pondérés de plusieurs manières. Tout d'abord, nous faisons la distinction entre les types et les occurrences de lexèmes où les procédés morphologiques étudiés apparaissent :

– *occ* : chaque occurrence d'un lexème dans un document est comptabilisée. Par exemple, si *activité/NOM* apparaît trois fois dans un document, il compte pour trois

occurrences ;

– *type* : chaque lexème est compté une fois, indépendamment du fait qu'il apparaisse une ou plusieurs fois dans un document. Pour l'exemple précédent, même si *activité/NOM* apparaît trois fois, il compte pour un seul type.

Quant aux valeurs assignées aux descripteurs, elles prennent en compte le fait que les types ou les occurrences des lexèmes sont considérés. Les valeurs elles-mêmes sont calculées de trois manières :

– *freq* correspond à la fréquence du descripteur. Il s'agit soit du nombre de types où le descripteur apparaît (la base *acte* apparaît dans *action* et *activité* et reçoit la valeur de 2), soit de la somme des occurrences de ces types (la base *acte* reçoit alors la valeur de 5, car *action* apparaît 2 fois et *activité* 3 fois dans le document). Dans le premier cas, nous mesurons la variété lexicale ou la taille des familles morphologiques autour d'un procédé, tandis que dans le deuxième cas nous mesurons la fréquence des unités morphologiques. Les deux valeurs suivantes associées aux descripteurs dérivent de ces notions de fréquence ;

– *norm* correspond à la normalisation de la fréquence par la longueur du document. Le poids du descripteur dépend alors de sa fréquence et de la longueur du document. Le document de notre exemple contient 1 068 occurrences. Avec ce calcul, nous obtenons les valeurs suivantes de descripteurs, selon que les types ou les occurrences de lemmes sont pris en compte : 2 / 1068 = 0,0019, 5 / 1068 = 0,0047 ;

– *tfidf* correspond à la pondération dite *term frequency*inverse document frequency* (Salton, 1991 ; Singhal *et al.*, 1996). Cette mesure permet d'évaluer l'importance du descripteur par rapport au corpus. Le poids augmente proportionnellement à la fréquence du descripteur dans le document. Il varie également en fonction de la fréquence de ce descripteur dans le corpus. Nous appliquons la version suivante de cette formule : $freq * \log(tot/nbdoc)$, où $freq$ est la fréquence du descripteur, tot le nombre de documents dans le corpus et $nbdoc$ le nombre de documents où ce descripteur apparaît. De la même manière que pour le poids *norm*, *tfidf* dépend du nombre de types ou d'occurrences de lexèmes où le descripteur apparaît.

4.3.5. *Sélection des descripteurs*

La sélection des descripteurs est étudiée depuis plusieurs années de façon intensive (Koller et Sahami, 1996). Elle poursuit un double objectif :

– réduire le nombre de descripteurs, en en sélectionnant une partie ;

– obtenir de meilleurs résultats en essayant d'éliminer les descripteurs, qui produisent le plus de bruit, et en ne gardant que les descripteurs les plus discriminants.

Dans cette étude, nous nous limitons à réaliser une sélection de descripteurs à partir du critère de leur distribution dans les documents, ce qui favorise les descripteurs qui sont distribués dans le plus grand nombre de documents. L'avantage d'une telle sélection est d'offrir de bons résultats pour une réduction de plus de 80 % du coût de calcul (Yang et Liu, 1999). Nous testons plusieurs filtres de sélection en fixant le

nombre minimal de documents où les descripteurs apparaissent, sur une échelle allant de 1 à 10. Lorsque ce filtre est fixé à 1, il suffit que les descripteurs apparaissent au moins dans un document de chaque sous-corpus : un maximum de descripteurs est alors exploité. Lorsque le seuil est fixé à 10, les descripteurs doivent apparaître dans 10 documents au moins au sein de chaque sous-corpus : cela correspond au filtrage des descripteurs le plus contraint.

Nous exploitons aussi un ensemble réduit de descripteurs. Au sein des neuf sous-corpus, nous prenons les 100 descripteurs les plus fréquents en termes d'occurrences. Une fois la liste dédoublonnée, nous obtenons un ensemble de 256 descripteurs. Nous appliquons le même filtre de sélection relatif au nombre de documents où ces descripteurs apparaissent (allant de 1 à 10).

4.3.6. *Évaluation*

Nous effectuons plusieurs tests, en faisant varier les différents paramètres décrits dans cette section 4.3 afin de tester l'influence qu'ils peuvent avoir sur les résultats. La catégorisation manuelle des documents proposée par CISMeF correspond à nos données de référence. Pour chaque test, nous effectuons une validation croisée ou *cross-validation* (Mitchell, 1996, section 4.6.5) (Sebastiani, 2002, section 4.1). Le principe d'une telle évaluation est de permettre aux algorithmes d'apprentissage d'utiliser deux ensembles distincts de données pour les étapes d'entraînement et de validation. La validation croisée à n plis (*n fold*) est ainsi effectuée n fois sur des partitions de données différentes et le résultat global correspond à la moyenne des performances. Dans notre travail, nous effectuons une validation croisée à 10 plis. Le corpus est donc divisé en 10 partitions : à chaque fois, l'entraînement est effectué sur 9 partitions et la 10^e est utilisée pour la validation. À chaque itération, la partition d'évaluation est donc différente. Les mesures d'évaluation correspondent aux moyennes de toutes les itérations. Nous calculons les mesures d'évaluation standard dans la tâche de catégorisation automatique : 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).

La *baseline* correspond à l'exploitation des descripteurs de niveau lexical, qui sont des descripteurs plus facilement et directement accessibles dans les textes et qui sont traditionnellement exploités en apprentissage. Nous prenons en compte les lemmes obtenus suite à l'étiquetage `TreeTagger` et à leur correction par `Flemm`. Les tests réalisés avec ces descripteurs de baseline sont parallèles à ceux réalisés avec les descripteurs morphologiques.

5. Distinction automatique des catégories de documents

Nous présentons les résultats et discutons les aspects liés au matériel traité (section 5.1), aux données morphologiques collectées (section 5.2), et à la catégorisation automatique des documents selon leur degré de spécialisation (section 5.3).

5.1. *Collecte des documents*

Le matériel, collecté au travers de CISMeF, correspond à nos données de référence. Nous exploitons ces données pour la création de modèles représentant le contenu des documents biomédicaux selon leurs registres, mais aussi pour l'évaluation des modèles générés. Le portail CISMeF existe depuis 1995 (Darmoni *et al.*, 1999) et, grâce aux efforts de plusieurs documentalistes et de plusieurs projets de recherche, les documents référencés représentent bien le contenu de la Toile biomédicale francophone. La représentativité du matériel est d'autant plus importante que les documents indexés proviennent de différents sites créés et maintenus dans différents pays francophones (France, Belgique, Canada, Suisse, Luxembourg...).

Comme nous l'avons indiqué (section 4.1), chaque document est annoté par plusieurs types d'informations, comme par exemple les mots-clés biomédicaux et le destinataire des documents. Les mot-clés assignés à chaque document proviennent de la terminologie MeSH, dédiée à la recherche d'information. Cette terminologie est structurée hiérarchiquement, ce qui permet au moteur de recherche CISMeF d'effectuer l'explosion hiérarchique de la requête. Par exemple, une requête, qui contient le mot-clé *cardiologie*, est explosée en plusieurs termes qui sont subsumés par ce mot-clé (comme *électrophysiologie cardiaque*). Par ailleurs, chaque spécialité médicale peut être décrite de plusieurs points de vue, comme par exemple *économie*, *enseignement et éducation*, *instrumentation*, *législation et jurisprudence*, *méthodes*, *normes*. Cette approche d'indexation et de recherche documentaire au sein de CISMeF garantit la variété des documents retrouvés, qui vont nourrir le corpus correspondant. Notons aussi que CISMeF propose le taux de pertinence des mots-clés par rapport au document, mais nous n'exploitons pas cette information. En revanche, comme nous l'avons indiqué auparavant, lors de la constitution du corpus, nous téléchargeons uniquement le document correspondant à l'URL indexée, sans analyser les liens hypertextes éventuellement disponibles. De cette manière, nous pensons contrôler mieux la pertinence de l'indexation par rapport au contenu du document exploité.

Le deuxième type d'annotation que nous exploitons est relatif à la caractérisation du document par rapport à ses destinataires et à son degré de spécialisation, que nous avons converti en trois catégories (*pro*, *étu* et *vul*). Cette annotation est effectuée manuellement par les documentalistes. Elle n'est pas créée spécifiquement pour la tâche que nous traitons ni pour notre expérience. Ceci garantit une indépendance et objectivité de cette annotation. Nous pouvons exploiter ce matériel afin de tester nos hypothèses et réaliser nos expériences.

5.2. *Accès au niveau morphologique des documents textuels*

Dans cette section, nous nous intéressons aux différentes étapes de TAL qui permettent d'accéder aux données morphologiques. Le succès dans l'obtention de meilleures données dépend beaucoup de la qualité de l'étiquetage morphosyntaxique. Ainsi, plusieurs des étapes préliminaires effectuées poursuivent les objectifs suivants :

l'homogénéisation de l'encodage des documents vers *iso-latin1*, la reconnaissance de la langue, la segmentation et le pré-étiquetage des documents avec un lexique adapté et la correction de l'étiquetage avec `Flemm`. En ce qui concerne la reconnaissance de la langue, nous avons actuellement deux limites. D'une part, nous pouvons récupérer des pages multilingues (par exemple, avec des résumés d'articles scientifiques en anglais et en français). La contrainte que nous posons lors de cette étape consiste à avoir au moins 70 % de textes en français : par ce biais, nous pouvons trouver des mots issus d'autres langues (*body*, *alcoholic*, *hemochromatosis*, *hyperferritinemia*). Ces mots restent souvent non analysés à l'étape de `Dérif`. D'autre part, certaines pages en français ne sont pas détectées comme telles parce qu'elles ne contiennent pas les mots grammaticaux prévus. Une analyse de ces pages montre qu'elles ne contiennent pas de vraies phrases, mais semblent surtout contenir du péritexte.

Ces différents prétraitements n'empêchent toutefois pas que des erreurs persistent suite entre autres à la spécificité du lexique, des phrases ou des constructions grammaticales. Certaines erreurs de `TreeTagger` sont corrigées par `Flemm`. Par exemple, `TreeTagger` étiquette *antiinflammatoire* comme pronom (*PRO*) et `Flemm` corrige cet étiquetage : *antiinflammatoire/ADJ*. Parmi d'autres erreurs typiques de `TreeTagger`, mentionnons l'étiquetage d'un nom au début de la phrase comme un verbe ou encore l'assignation de mots lexicaux à des classes fermées (*DET, SENT, PRO, PRP...*). Par exemple, dans l'expression *Affections de longue durée*, le mot *affections* est étiqueté comme verbe à l'imparfait *VER:impf* et se trouve lemmatisé vers *affecter*, tandis que dans l'expression *Etablissement français du sang*, *etablissement* est étiqueté comme verbe au subjonctif présent *VER:subp* et se trouve lemmatisé vers *etablissemer*. Même si un tel étiquetage et parfois aussi la lemmatisation ne sont pas corrects, l'assignation des catégories comme *verbe*, *nom* ou *adjectif* permet aux mots correspondants d'être analysés par `DériF`. En revanche, l'étiquetage de *Labellisation* comme d'un séparateur de fin de phrase (*SENT*) ou de *Classantes* comme un nom propre (*NAM*) n'est pas corrigé. De plus, un tel étiquetage va empêcher une analyse morphologique de ces mots par `DériF`. Afin de seconder `Flemm`, il est possible de faire un traitement supplémentaire pour détecter et corriger les erreurs lorsque des classes fermées et manifestement erronées sont proposées par `TreeTagger`.

Quant au matériel morphologique à proprement parler fourni par `DériF`, nous avons pu bénéficier de l'existence de cet outil unique, qui arrive à analyser un grand nombre de lemmes. Les affixes peuvent être des préfixes ou des suffixes. En ce qui concerne les bases morphologiques détectées par `Dérif`, elles peuvent être de deux types : (1) supplétives ou savantes, qui n'apparaissent pas de manière isolée dans la langue mais toujours couplées avec d'autres éléments morphologiques (p. ex. : *gastr(o)* qui se réalise au travers de *gastrique* ou *hém* qui se réalise au travers de *hémorragique*) et (2) les bases autonomes (comme *bronches* ou *bactérie*).

Nous nous sommes rendu compte de quelques limites dans l'analyse morphologique. Ainsi, `DériF` peut générer une analyse morphologique erronée, comme pour le lexème *gymnase/NOM* analysé comme l'« *enzyme du nu* » ou *mention/NOM* formé sur

le verbe *mentir*, présentés dans la série d'exemples (1) à (3) qui suit. Notons toutefois que nous n'avons pas repéré beaucoup d'erreurs d'analyse.

(1) *gymnase/NOM: [[gymno A*] [ase N*] NOM]*
« *(Partie de – Type particulier de) enzyme caractérisé par la propriété : nu* »
Constituants = /gymno/ase/
Type = produit chimique

(2) *mention/NOM: [[mentir VERBE] ion NOM]*
(mention/NOM, mentir/VERBE)
« *(Action - résultat de l'action) de mentir* »

(3) *information/NOM: [[[informe ADJ] VERBE] ion NOM]*
(information/NOM, informer/VERBE, informe/ADJ)
« *(Action - résultat de l'action) de informer* »

Une autre limite, qui avait plus de répercussion sur notre travail, concerne la notation des bases reconnues et analysées. `DériF` identifie les bases selon leur forme (chaîne de caractères). Par exemple, quatre bases – *hém(o)*, *héma*, *hémat(o)* et *èm* – proviennent du même mot grec *haima*, signifiant *relatif au sang*. `DériF` les segmente correctement mais continue de les différencier formellement. Ainsi, trois bases sont distinguées dans l'analyse morphologique des lexèmes : *[hém]*, qui regroupe *hém(o)* et *héma* (exemples (4) et (5)), *[ém]* (exemples (6) et (7)) et *[hémato]* (exemple (8)).

(4) *hémolytique/ADJ: [[hém N*] [lyt V*] ique ADJ]*
« *Qui dissoudre ou désintégrer le(s) sang* »

(5) *hémorragique/ADJ: [[[hém N*] [rragie N*] NOM] ique ADJ]*
« *En rapport avec le(s) hémorragie* »

(6) *urémique/ADJ: [[[uro N*] [ém N*] ie NOM] ique ADJ]*
« *En rapport avec le(s) urémie* »

(7) *ferritinémie/NOM: [[ferritine NOM] [ém N*] ie NOM]*
« *Affection liée au(x) sang en rapport avec le(s) ferritine* »

(8) *hématologique/ADJ: [[[hémato N*] [logie N*] NOM] ique ADJ]*
« *En rapport avec le(s) hématologie* »

D'autres séries de bases sont dans ce cas, comme par exemple *abdomen*/*abdomin* et *card*/*carde*/*cardio*. L'existence de plusieurs bases qui véhiculent le même sens peut introduire un biais dans nos données car les types et les occurrences de lexèmes correspondants sont alors distribués entre les différentes formes de ces bases. Dans le travail précédent, où nous avions effectué la sélection des bases manuellement, nous avions fusionné les bases équivalentes. En revanche, dans le travail actuel, où toute la

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

Tableau 3. *Bases et affixes les plus fréquents selon les trois registres étudiés*

Corpus	Total	Uniq.	Exemples
cardio	4 372	1 889	$jéjuno_e$ $plausible_e$ $flavone_p$ $ischio_p$ $méiose_p$ $normat_v$
pneumo	4 260	1 899	$abort_e$ $alopécie_e$ $amiodarone_p$ $monét_p$ $spéc_v$ $fécal_v$
hémato	4 097	1 877	$abdomen_e$ $naevo_e$ $cocaïne_p$ $phrén_p$ $angél_v$ $règle_v$
pro	4 887	1 786	$adipos_c$ $brachy_c$ $ankylo_h$ $carotène_h$ $agrégat_p$ $aphte_p$
étu	3 926	1 285	$abduct_c$ $crano_c$ $mnés_c$ $coccygien_h$ $exérèse_p$ $ptéryg_p$
vul	2 645	1 119	$adip_c$ $graphe_c$ $pexie_c$ $amnio_h$ $angél_h$ $cili_p$ $gnath_p$

Tableau 4. *Selon les corpus et les registres : nombre total de bases et d'affixes, nombre de bases uniques, et quelques exemples de bases uniques*

chaîne de traitement est effectuée automatiquement, une approche manuelle qui assurerait la fusion n'est plus possible. Ces bases restent donc isolées. Pour leur groupement, il est possible d'exploiter la glose sémantique associée aux lexèmes analysés. Si nous regardons les gloses des exemples (4) à (8), la notion de *en rapport avec le sang*, *liée au sang* peut y apparaître, comme dans (4) et (7), mais pas systématiquement. Il est possible qu'une ressource externe soit nécessaire pour le groupement de bases équivalentes ou bien des méthodes pour le calcul de la distance des chaînes d'édition.

Notre chaîne de traitement a permis de distinguer 5 968 bases et affixes différents. Le tableau 3 présente ceux qui se trouvent parmi les plus fréquents selon les registres. Nous pouvons voir que les procédés d'affixation (*ion*, *al* et *ique* dans ces exemples, mais ils sont bien plus nombreux) sont distribués dans tous les corpus. *is*, qui apparaît dans les noms et adjectifs d'origine latine, est fréquent dans les sous-corpus *pro*. Nous pouvons observer quelques verbes fréquents, comme *traiter*, *guider* ou *mentir*, ce dernier provenant de l'analyse du nom *mention* (exemple (2)).

Dans le tableau 4, nous présentons une autre vue de la distribution de ces unités morphologiques. Pour chaque domaine biomédical (*cardio*, *pneumo* et *hémato*) et pour chaque registre (*pro*, *étu* et *vul*), nous indiquons le nombre total de bases et affixes ainsi que le nombre de bases uniques dans un corpus. Notons que les affixes ne sont hapaxiques dans aucun sous-corpus. Dans la dernière colonne, nous indiquons aussi quelques exemples en mettant en indice le sous-corpus où l'unité morphologique en question est unique. Par exemple, dans le corpus *cardio*, *jéjuno* et *plausible* apparaissent uniquement dans le sous-corpus *étu*, *flavone*, *ischio* et *méiose* apparaissent dans le sous-corpus *pro*, et *normat* dans le sous-corpus *vul*. Lorsque la distribution

du matériel morphologique est indiquée par registre (*pro*, *étu* et *vul*), ce sont les domaines biomédicaux qui se trouvent en indice. Nous pouvons observer la présence de noms de médicaments ou de substances chimiques (*carotène*, *flavone*, *amiodarone* ou *cocaïne*) qui apparaissent uniquement dans les sous-corpus *pro*. Si nous faisons l'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 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*...), mais il nous est difficile d'en juger. Dans d'autres cas, il s'agit de bases grecques, comme *sphygm*, signifiant *pulsation*, qui semblent correspondre à une notion assez centrale en cardiologie. Mais comme ce sont les bases d'origine latine qui sont actuellement privilégiées en terminologie biomédicale, cela peut mener à un sous-emploi de bases grecques. Dans le registre *étu* du tableau 4, notons aussi l'adjectif *coccygien* construit morphologiquement, mais qui reste non analysé. D'autres lemmes sont dans ce cas (comme *zostérien*, *zygomatique*, *vomitif*, *bulleux*, *vitréen* ou *membraneux*). La raison de cette non-analyse est que ces bases apparaissent au sein de lemmes construits et encore plus complexes, comme dans ces quelques exemples :

(9) *sacro-coccygien/ADJ: [[sacro N*] [coccygien ADJ] ADJ]*
(sacro-coccygien/ADJ, [sacro,N]:coccygien/ADJ)*
« *Qui est coccygien par rapport au sacrum* »

(10) *vésiculo-bulleux/ADJ: [[vésicule NOM] [bulleux ADJ] ADJ]*
(vésiculo-bulleux/ADJ, [vésicule,NOM]:bulleux/ADJ)
« *Qui est bulleux par rapport au vésicule* »

(11) *musculo-membraneux/ADJ: [[muscul N*] [membraneux ADJ] ADJ]*
(musculo-membraneux/ADJ, [muscul,N]:membraneux/ADJ)*
« *Qui est membraneux par rapport au muscle* »

Dans ces cas, `Dérif` analyse la dernière opération morphologique. Elle correspond ici à la composition. Plus particulièrement, c'est cette opération qui permet de définir la glose sémantique du lemme et d'expliciter sa sémantique en relation avec ses bases. Comme nous l'avons discuté auparavant pour *hém(o)/héma/hémat(o)/èm*, ces cas peuvent aussi présenter une limitation pour la tâche que nous poursuivons : les unités morphologiques équivalentes ne sont en effet pas regroupées. Mais en dehors de cette limitation, cela indique aussi que les unités morphologiques ont des valeurs différentes selon les contextes où elles apparaissent et que c'est l'opération qui permet d'expliciter la sémantique du lemme qui est privilégiée par `DériF`.

5.3. *Catégorisation automatique des documents*

Le tableau 5 montre les résultats de la catégorisation automatique des documents

	Cardio			Pneumo			Hémato		
	P	R	F	P	R	F	P	R	F
pe	0,930	0,929	0,929	0,917	0,917	0,917	0,908	0,907	0,907
ve	0,966	0,966	0,966	0,950	0,950	0,950	0,909	0,907	0,906
pv	0,968	0,968	0,968	0,945	0,945	0,945	0,956	0,956	0,956
tri	0,887	0,885	0,886	0,925	0,924	0,923	0,897	0,897	0,897

Tableau 5. *Catégorisation automatique des documents. Validation croisée à 10 plis. Descripteurs* ba*, présents dans 10 documents au minimum, pondération* tfidf*, prise en compte des* types *des lemmes. Évaluation de la précision P, du rappel R et de la f-mesure F.*

obtenus avec la validation croisée à 10 plis. Les résultats sont donnés selon les corpus et les sous-corpus en termes de précision, rappel et f-mesure. Comme nous l'avons annoncé dans la section 4.3, nous avons fait varier et avons testé plusieurs paramètres. Dans le tableau 5, nous présentons un des meilleurs paramétrages, qui est le suivant :

– *type* : prise en compte des types des lemmes : c'est donc la variété lexicale et morphologique qui est mise en avant ;

– *ba* : les descripteurs couvrent les bases et les affixes ;

– *10* : chaque descripteur doit être présent au moins dans 10 documents : il s'agit du filtrage le plus contraint ;

– *tfidf* : les descripteurs sont pondérés au moyen de *tfidf*.

Nous pouvons voir qu'avec ce paramétrage, les trois mesures d'évaluation montrent des performances supérieures à 0,90 pour les trois domaines et pour les tests bicatégories. Les tests multicatégories *tri* ont des résultats plus faibles dans les corpus *cardio* et *hémato*. Dans l'expérience *total*, lorsque la totalité des documents de tous les sous-corpus est prise en compte, les performances sont meilleures. Par exemple, avec le même paramétrage que dans le tableau 5, nous obtenons la f-mesure de 0,955, 0,972, 0,986 et 0,954 pour les tests *pe, ve, pv* et *tri* respectivement. Ceci montre qu'il existe des descripteurs stables et saillants au travers des domaines biomédicaux. Comme ces descripteurs ressortent grâce à l'analyse morphologique et à l'augmentation de la taille des corpus, il n'est pas nécessaire d'effectuer au préalable la distinction des domaines biomédicaux : la distinction de la spécialisation peut être effectuée directement.

5.3.1. *Comparaison avec la baseline*

Dans le tableau 6, nous présentons les résultats obtenus avec la baseline (descripteurs lexicaux lemmatisés) et avec le même paramétrage que dans le tableau 5. Nous pouvons observer que les performances sont beaucoup moins bonnes, avec une perte d'au moins 0,10 points et pouvant aller jusqu'à 0,30 points pour la f-mesure. La différence s'accentue avec la diminution de la taille des données traitées (sous-corpus *pneumo* et *hémato*). Une telle différence montre que les descripteurs positionnés au

	Cardio			Pneumo			Hémato		
	P	R	F	P	R	F	P	R	F
pe	0,860	0,813	0,807	0,717	0,688	0,677	0,842	0,769	0,756
ve	0,841	0,785	0,776	0,725	0,683	0,668	0,779	0,654	0,610
pv	0,864	0,825	0,821	0,846	0,804	0,797	0,809	0,716	0,693
tri	0,847	0,763	0,766	0,731	0,633	0,631	0,652	0,593	0,597

Tableau 6. *Performances obtenues avec la baseline (descripteurs lexicaux lemmatisés). Même paramétrage que dans le tableau 5.*

	Cardio			Pneumo			Hémato		
	P	R	F	P	R	F	P	R	F
pe	0,930	0,929	0,929	0,917	0,917	0,917	0,908	0,907	0,907
pro	0,938	0,920	0,929	0,917	0,917	0,917	0,885	0,934	0,909
étu	0,922	0,939	0,930	0,917	0,917	0,917	0,930	0,879	0,904
pv	0,968	0,968	0,968	0,945	0,945	0,945	0,956	0,956	0,956
pro	0,957	0,980	0,968	0,947	0,942	0,945	0,943	0,971	0,957
vul	0,979	0,956	0,967	0,943	0,948	0,945	0,970	0,941	0,955
tri	0,887	0,885	0,886	0,925	0,924	0,923	0,897	0,897	0,897
pro	0,840	0,902	0,870	0,931	0,872	0,900	0,880	0,890	0,885
étu	0,909	0,859	0,883	0,882	0,963	0,921	0,916	0,956	0,935
vul	0,913	0,896	0,904	0,962	0,936	0,949	0,895	0,846	0,870

Tableau 7. *Détail par catégorie sur la précision P, le rappel R et la f-mesure F pour les tests* pe*,* pv *et* tri*. Même paramétrage que dans le tableau 5.*

niveau morphologique de la langue sont approriés pour la tâche ciblée dans ce travail et qu'ils permettent de factoriser la représentation du contenu des documents.

Nous allons maintenant analyser nos résultats plus en détail, en mettant l'accent sur les différents paramètres liés au matériel : les catégories (section 5.3.2), les descripteurs (section 5.3.3), la pondération (section 5.3.4) et la sélection (section 5.3.5) des descripteurs. Nous faisons aussi une discussion générale de cette expérience en la mettant en relation avec d'autres travaux et en réfléchissant à la place de la morphologie dans cette tâche et dans la langue biomédicale (section 5.3.6).

5.3.2. *Catégories*

De manière générale, la reconnaissance des registres de spécialisation au sein des domaines est assez performante. Avec le paramétrage du tableau 5 et avec les expériences bicatégories, la précision et le rappel sont supérieurs à 0,90 %. Ils sont supérieurs à 0,88 % pour les expériences multicatégories *tri*. Parmi les expériences

	Cardio			Pneumo			Hémato		
	P	R	F	P	R	F	P	R	F
ba	0,968	0,968	0,968	0,945	0,945	0,945	0,956	0,956	0,956
b	0,949	0,948	0,948	0,966	0,966	0,966	0,957	0,956	0,956
réduit	0,952	0,952	0,952	0,966	0,966	0,966	0,961	0,961	0,961

Tableau 8. *Différence de la précision, du rappel et de la f-mesure au sein du test* pv *en fonction des descripteurs : bases et affixes (*ba*), bases seules (*b*) et une liste des bases et affixes les plus fréquents (*réduit*). Même paramétrage que dans le tableau 5.*

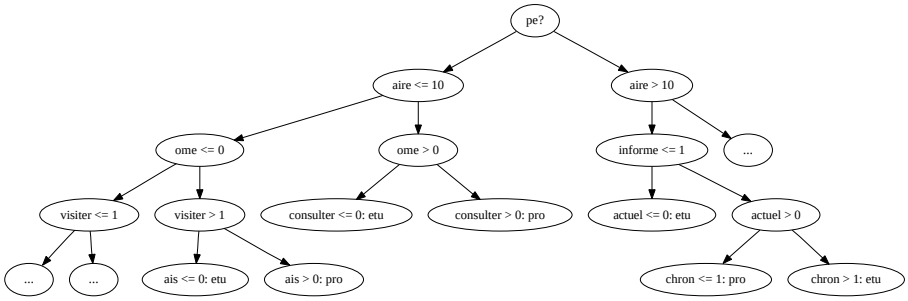

Figure 2. *Un extrait de l'arbre de décision* `C4.5` *pour le test* pe *sur le corpus* cardio. *Le paramétrage est le suivant : descripteurs* ba*, pondération* freq*, nombre minimal de documents 5,* occurrences *des lemmes.*

bicatégories, le test le plus difficile concerne la distinction binaire entre les documents *étu* et *pro*, tandis que le test le plus performant concerne la distinction entre les documents *vul* et *pro*. Dans le tableau 7, nous présentons le détail des résultats pour chaque catégorie. Selon les registres et les corpus, la f-mesure reste assez stable tandis que la précision et le rappel varient. Les documents en langue vulgarisée semblent présenter plus de différences par rapport aux deux autres registres. Dans le test *pv* au sein des corpus *cardio* et *hémato*, le registre *vul* est reconnu avec une grande précision. Le détail sur les catégories au sein du test *tri* semble confirmer aussi cette observation : au travers des domaines, les documents *vul* sont reconnus le plus souvent avec la f-mesure la plus élevée.

5.3.3. *Descripteurs*

Le tableau 8 présente les résultats obtenus pour le test *pv* avec différents ensembles de descripteurs (bases et affixes *ba*, bases seules *b* et la liste de bases et affixes les plus fréquents *réduits*) tout autre paramétrage restant identique. Sauf dans le corpus *cardio*, la liste *réduite* permet d'obtenir les meilleurs résultats.

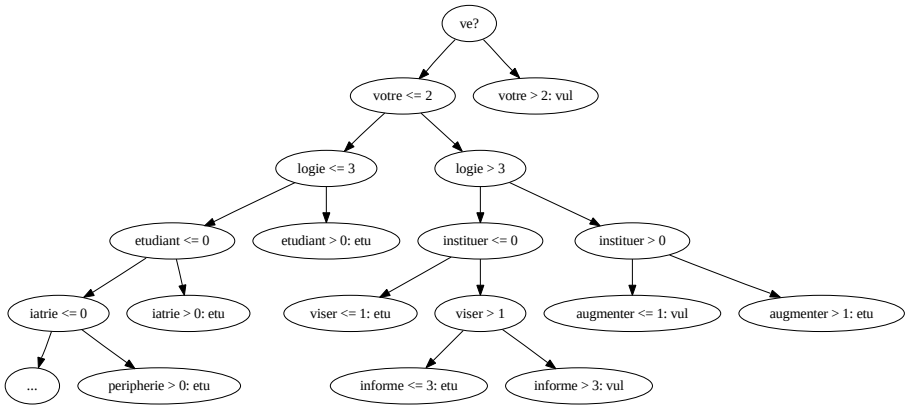

Figure 3. *Un extrait de l'arbre de décision* `C4.5` *pour le test* ve *sur le corpus* cardio*. Le paramétrage est le suivant : descripteurs* b*, pondération* freq*, nombre minimal de documents 10,* occurrences *des lemmes.*

Pour les expériences réalisées, la taille des arbres varie entre 7 et 52 feuilles décisionnelles. Nous présentons les extraits des arbres de décision pour deux tests *pe* (figure 2) et *ve* (figure 3). Les descripteurs exploités pour le test *ve* sont les bases seulement, alors que pour le test *pe* les bases et les affixes sont exploités. Dans le cas du test *pe*, nous remarquons que : (1) les affixes (*aire* et *ais*) ont un rôle assez discriminant pour différencier les documents *étu* et *pro* ; (2) l'élément de composition *ome* est placé assez haut dans l'arbre de décision ; (3) les bases verbales, *visiter, consulter* et *informe* pour *informer* (voir exemple (3)), ont aussi une place importante.

En ce qui concerne le test *ve*, la première feuille décisionnelle est occupée par *votre*. Notons aussi l'importance des éléments de composition *logie* et *iatrie*, de la base *étudiant* et également de plusieurs bases verbales. L'apparition de *votre* dans le matériel morphologique est due au fait que ce mot a été parfois étiqueté comme nom, ce qui est erroné, et analysé ensuite comme une conversion par `Dérif` :

(12) *votre/NOM: [[votre ADJ] NOM]*
(votre/NOM, votre/ADJ)
« *Entité dont la propriété vue comme saillante est d'être votre* »

Même s'il s'agit d'une erreur d'étiquetage, elle fait apparaître l'importance des pronoms possessifs et des traces des locuteurs dans les documents. Nous avons déjà pu observer le caractère saillant des pronoms possessifs auparavant (Grabar *et al.*, 2007). Nous avons effectué le même test mais sans le descripteur *votre*, pour éliminer le biais éventuel dans les descripteurs. Les performances diminuent mais très légèrement : elles passent de 0,902 à 0,899 pour les trois mesures d'évaluation.

	Cardio			Pneumo			Hémato		
	P	R	F	P	R	F	P	R	F
$freq_t$	0,851	0,851	0,851	0,783	0,783	0,783	0,789	0,789	0,789
$norm_t$	0,842	0,841	0,841	0,775	0,775	0,775	0,800	0,799	0,799
$tfidf_t$	0,968	0,968	0,968	0,945	0,945	0,945	0,956	0,956	0,956
$freq_o$	0,837	0,837	0,837	0,804	0,804	0,804	0,840	0,838	0,838
$norm_o$	0,858	0,857	0,857	0,804	0,804	0,804	0,780	0,779	0,779
$tfidf_o$	0,906	0,906	0,906	0,859	0,859	0,859	0,869	0,868	0,868

Tableau 9. *Différence de la précision P, du rappel R et de la f-mesure F au sein du test* pv *en fonction de différentes pondérations des descripteurs (*freq, norm *et* tfidf*) et selon que les lemmes sont représentés par leurs* types *ou leurs* occurrences*.*

5.3.4. *Pondération des descripteurs*

Dans le tableau 9, nous présentons les résultats sur la variation des performances en fonction de la pondération (*freq*, *norm* ou *tfidf*) des descripteurs au sein du test *pv*. Lorsque les types de lemme sont pris en compte, les pondérations ont l'indice t. Lorsque les occurrences des lemmes sont prises en compte, les pondérations ont l'indice o. Nous pouvons voir que les pondérations des descripteurs ont une influence importante sur les résultats. Ainsi, avec *tfidf* le système fournit les résultats qui sont largement meilleurs que ceux obtenus avec les deux autres pondérations. La pondération *norm* est souvent la plus faible. Notons que dans les tests *total* la pondération *tfidf* montre encore plus de différence avec les deux autres pondérations.

Lorsque nous considérons les *types* des lemmes, nous prenons en compte en réalité la taille des familles morphologiques et donc la variété lexicale. Au vu des résultats du tableau 9, 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. Une étude complémentaire a permis d'observer que ce sont les documents *étu* qui montrent la variété lexicale la plus grande. En effet, ces documents fournissent des informations biomédicales précises et détaillées et en général présentent un grand nombre de notions médicales. En ce qui concerne les documents *pro* et *vul*, les premiers présentent plus d'occurrences de termes savants, tandis que les deuxièmes montrent plus de variété.

5.3.5. *Sélection des descripteurs*

Dans le tableau 10, nous indiquons le nombre de descripteurs sélectionnés avec trois seuils différents (1, 5 et 10). Cette sélection est effectuée au sein des tests *tri*, contrastant les trois catégories. Nous pouvons voir que le nombre de descripteurs est toujours plus élevé pour les ensembles *ba*, ce qui est attendu car cet ensemble contient l'ensemble *b*. Nous pouvons aussi voir que la différence entre ces deux ensembles

	Cardio		Pneumo		Hémato	
Seuil	b	ba	b	ba	b	ba
1	2 229	2 276	1 961	2 014	1 862	1 965
5	1 119	1 175	932	965	851	927
10	766	807	539	584	515	568

Tableau 10. *Nombre de descripteurs dans les tests* tri*, pour les deux types de descripteurs (*b *et* ba*) et pour trois des seuils appliqués (1, 5 et 10)*

ba et *b* est assez faible : ce sont les bases qui constituent le principal matériel morphologique et les affixes les complètent. Par ailleurs, avec l'augmentation du seuil, le nombre de descripteurs diminue, ce qui est aussi une évolution normale. En revanche, nous n'avons pas observé d'influence au niveau des résultats de `C4.5` car cet algorithme semble éliminer de lui-même les descripteurs trop rares.

5.3.6. *Au-delà de la morphologie*

Nos expériences indiquent la pertinence des descripteurs morphologiques de la langue biomédicale pour la catégorisation automatique des documents selon leur spécialisation. Les descripteurs morphologiques s'avèrent en effet plus efficaces que les descripteurs lexicaux. Ces descripteurs devraient toutefois être examinés dans leurs contextes et au sein des documents. Nous avons effectué une analyse supplémentaire des contextes de 30 lexèmes, avec l'objectif de voir si ces lexèmes apparaissent dans des contextes définitoires. Il s'agit par exemple de *bronchite*, *phlébectomie*, *hématome*, *myocarde*, *corticothérapie*, *pulmonaire* ou *bronchospasme*. Nous avons extrait tous les contextes phrastiques de ces lexèmes (n = 486) et les avons examinés. Nous donnons deux exemples de contextes définitoires : pour *bronchite chronique* en (13) observé dans le sous-corpus *vul* et pour *hématome pulmonaire* en (14) observé dans le sous-corpus *étu*.

(13) *La bronchite chronique est une maladie chronique caractérisée par une inflammation et un rétrécissement des bronches déclenchés par des substances irritantes.*

(14) *L'hématome pulmonaire se définit comme une hémorragie collectée au sein d'une cavité néoformée par dilacération du parenchyme pulmonaire.*

Suite à l'analyse des contextes définitoires, nous avons fait deux observations principales. D'une part, il existe une différence nette entre les définitions provenant de différents registres : dans les documents *vul*, les définitions sont plus facilement compréhensibles. D'autre part, la majorité de contextes analysés sont définitoires dans les sous-corpus *vul* (36 %), tandis que dans les sous-corpus *étu* et *pro*, ce pourcentage est de 18 % et 13 % respectivement. Nous pouvons donc voir que même si les docu-

	Algo.	Descripteurs	Taille	*pro*	*vul*
(Zheng *et al.*, 2002)	*ad, nb*	pondération manuelle des termes	288 doc*2	0,77	
(Wang, 2006)	*svm* *ad, nb*	unigrammes lexique	sites Web	0,81-0,84	
(Goeuriot *et al.*, 2007)	*svm, ad*	discursifs	variable	P 0,13-1,00 R 0,28-1,00	
(Poprat *et al.*, 2008)	*textcat*	n-grammes	sites Web	0,88-0,94	0,90-0,98
notre travail	*ad*	morphologie	100-250*2	0,94-0,97	0,94-0,97

Tableau 11. *Un tableau comparatif de quelques travaux en catégorisation automatique des documents comme expert et patient. Les performances en termes d'accuracy (précision) sont dans la dernière colonne. Les autres colonnes présentent les paramétrages et les descripteurs des méthodes.*

ments à destination des patients font un emploi conséquent de termes biomédicaux, le destinataire non spécialiste peut être accompagné par des informations qui pourraient l'aider à comprendre quand même ces termes. Nous n'avons pas étudié spécifiquement si c'est la première occurrence d'un terme qui est définie, mais ceci semble être le cas. Un autre type d'information qui pourrait aider les patients à mieux appréhender le contenu spécialisé consiste en l'utilisation d'images. Dans un travail précédent, nous avons ainsi pu observer que la présence des images est un descripteur très fiable pour la catégorisation automatique des documents selon les registres patient et expert (Grabar *et al.*, 2007). En effet, dans les documents à destination des patients, l'utilisation des images est fréquente. En revanche, nous n'avons pas étudié s'il existe un lien entre les termes biomédicaux et les images et quel est le rôle exact de ces images.

De manière générale, il nous semble que la morphologie occupe une place importante dans la langue. À travers les différents procédés qu'elle met en œuvre, elle est liée à la syntaxe, au lexique, à la sémantique, au style, etc. Nous avons comparé les descripteurs morphologiques avec les descripteurs lexicaux et avons ainsi montré la supériorité de la morphologie. Nous pensons que cela montre l'intérêt de combiner ou de comparer la morphologie avec d'autres types de descripteurs. Il serait par exemple aussi intéressant d'observer si les performances obtenues avec les descripteurs morphologiques et les n-grammes sont comparables et si la morphologie est cosubstantielle avec les n-grammes, également souvent utilisés en catégorisation automatique.

Notre travail n'est pas directement comparable avec les travaux de l'état de l'art, où les corpus et les données de référence sont différents. Nous allons néanmoins essayer de faire un parallèle entre ces travaux et notre expérience. Dans le tableau 11, nous présentons les paramètres et les résultats de quelques autres travaux. Nous indiquons les algorithmes de classification (*ad* arbres de décision, *nb* classifieur bayésien naïf, *svm* et une adaptation de `textcat`), les descripteurs, la taille des corpus et les performances obtenues pour les registres *pro* et *vul*. Ces performances sont essentiel-

lement exprimées en termes d'*accuracy*, qui peut être rapproché avec la précision. Pour présenter nos expériences, nous avons pris les résultats du test *pv* du tableau 5. Les descripteurs utilisés dans ces autres travaux sont assez typiques des tâches de catégorisation automatique (n-grammes et descripteurs lexicaux). Lorsque les données de la Toile sont utilisées, les corpus sont très volumineux (plusieurs dizaines ou même centaines de mégaoctets de texte), mais aussi moins contrôlés car la catégorisation est réalisée au niveau des sites et non au niveau des pages. Au vu des résultats que montrent les différentes expériences présentées dans ce tableau, il apparaît que les résultats obtenus avec les descripteurs morphologiques sont bons et même meilleurs que ceux obtenus dans d'autres travaux. De plus, nos résultats présentent une variation plus faible et en général le rappel comme la précision restent élevés. Par ailleurs, dans les travaux cités, les catégories habituellement distinguées concernent les documents à destination des experts et des patients, alors que dans notre travail nous avons aussi traité les documents à destination des étudiants.

6. Conclusion et perspectives

Dans le travail présenté, l'objectif poursuivi concerne la distinction automatique des documents vulgarisés, experts et étudiants dans le domaine biomédical. Plus particulièrement, nous effectuons une étude des caractéristiques du niveau morphologique de ces documents. L'étude est effectuée avec une approche par apprentissage supervisé, qui exploite le matériel fourni par les outils de TAL. Les documents des corpus sont collectés au travers du portail CISMeF, leur annotation par les documentalistes de ce portail correspond donc aux données de référence pour cette étude.

Les mesures d'évaluation (précision, rappel et f-mesure) indiquent une bonne performance (souvent au-dessus de 0,90 %) et sont stables au travers des expériences. Nous avons noté plusieurs facteurs qui influencent les résultats, comme par exemple : (1) la pondération des descripteurs (*tfidf* fournit les résultats les plus performants) ; (2) le registre *vul* est assez bien contrasté par rapport aux deux autres registres ; (3) la prise en compte de l'ensemble des documents tous registres confondus améliore encore les performances ; (4) la taille des familles morphologiques, et pas seulement les fréquences de lemmes et des unités morphologiques, joue aussi un rôle important dans la catégorisation ; (5) 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). Grâce à des études supplémentaires, nous avons aussi observé que dans les documents à destination des patients, les termes médicaux complexes apparaissent souvent dans des contextes définitoires et que ces documents contiennent souvent des images. Dans l'avenir, nous allons aussi étudier l'interaction entre la morphologie et d'autres types de descripteurs. Par exemple, nous pouvons exploiter et combiner les descripteurs provenant de niveaux lexical, syntaxique, stylistique, ou encore ceux exploités dans l'état de l'art (comme les n-grammes). La perspective prin-

cipale de notre travail concerne l'utilisation des critères morphologiques, et d'autres descripteurs saillants, 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 descripteurs du niveau morphologique de la langue pourra fournir des résultats assez fiables et assister les documentalistes qui effectuent cette annotation.

Remerciements

Nous remercions Fiammetta Namer, Nabil Hathout et les relecteurs anonymes pour nous avoir aidés dans l'amélioration du contenu de notre travail et de son étendu.

7. Bibliographie

AMA, « 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*, vol. 281, n° 6, p. 552-7, 1999.

Baayen H., « Quantitative aspects of morphological productivity », *Yearbook of Morphology*p. 109-149, 1991.

Baayen H., « Derivational productivity and text typology », *Journal of quantitative linguistics*, vol. 1, n° 1, p. 16-34, 1994.

Baroni M., Chantree F., Kilgarriff A., Sharoff S., « Cleaneval: a Competition for Cleaning Web Pages », *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008. http://www.lrec-conf.org/proceedings/lrec2008/.

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., « Health information on the Internet. Accessibility, quality, and readability in English ans Spanish », *JAMA*, vol. 285, n° 20, p. 2612-2621, 2001.

Biber D., « Representativeness in Corpus Design », *Linguistica Computazionale*, vol. IX-X, p. 377-408, 1994. Current Issues in Computational Linguistics: in honor of Don Walker.

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

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

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

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

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

Deléger L., Zweigenbaum P., « Paraphrase acquisition from comparable medical corpora of specialized and lay texts », *AMIA 2008*, p. 146-50, 2008.

Flesch R., « A new readability yardstick », *Journal of Applied Psychology*, vol. 23, p. 221-233, 1948.

Fox S., 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, 2006.

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

Fradin B., Dal G., Grabar N., Namer F., Lignon S., Tribout D., Zweigenbaum P., « Remarques sur l'usage des corpus en morphologie », *Langages*, vol. 171, n° 3, p. 34-59, 2008.

Gaudinat A., Grabar N., Boyer C., « Combination of heterogeneous criteria for the automatic detection of ethical principles on health web sites », *AMIA 2007*, p. 264-8, 2007.

Genette G., *Seuils*, Seuil, Paris, 1987.

Goeuriot L., Grabar N., Daille B., « Caractérisation des discours scientifique et vulgarisé en français, japonais et russe », *TALN*, p. 93-102, 2007.

Grabar N., Krivine S., Jaulent M.-C., « Classification of Health Webpages as Expert and Non Expert with a Reduced Set of Cross-language Features », *AMIA 2007*, Chicago, USA, p. 284-8, 2007.

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

Gunning R., *The art of clear writing*, McGraw Hill, New York, NY, 1973.

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

Koller D., Sahami M., « Toward Optimal Feature Selection », *International Conference on Machine Learning*, p. 284-292, 1996.

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

McCray A., « Promoting Health Literacy », *Journal of American Medical Informatics Association*, vol. 12, p. 152-163, 2005.

Messai R., Zeng Q., Mousseau M., Simonet M., « Building a Bilingual French-English Patient-Oriented Terminology for Breast Cancer », *MedNet*, 2006.

Miller T., Leroy G., Chatterjee S., Fan J., Thoms B., « A Classifier to Evaluate Language Specificity of Medical Documents. », *HICSS*, p. 134-140, 2007.

Mitchell T. M., *Machine learning*, McGraw-Hill, 1996.

Namer F., « FLEMM : un analyseur flexionnel du français à base de règles », *Traitement automatique des langues (TAL)*, vol. 41, n° 2, p. 523-547, 2000.

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

Névéol A., Pereira S., Kerdelhué G., andM Joubert B. D., Darmoni S., « Evaluation of a simple method for the automatic assignment of MeSH descriptors to health resources in a French online catalogue », *Medinfo*, p. 407-11, 2007.

NLM, *Medical Subject Headings*. 2001, *www.nlm.nih.gov/mesh/meshhome.html*.

Pearson J., *Terms in Context*, vol. 1 of *Studies in Corpus Linguistics*, John Benjamins, Amsterdam/Philadelphia, 1998.

Poprat M., Beisswanger E., Hahn U., « Building a BioWordNet Using WordNet Data Structures and WordNet's Software Infrastructure - A Failure Story », *ACL 2008 workshop "Software Engineering, Testing, and Quality Assurance for Natural Language Processing"*, p. 31-9, 2008.

Poprat M., Markó K., Hahn U., « A Language Classifier that Automatically Divides Medical Documents for Experts and Health Care Consumers », *MIE 2006 - Proceedings of the XX International Congress of the European Federation for Medical Informatics*, Maastricht, p. 503-508, 2006.

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

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

Salton G., « Developments in automatic text retrieval », *Science*, vol. 253, p. 974-979, 1991.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, p. 44-49, 1994.

Sebastiani F., « Machine learning in automated text categorization », *ACM Computing Surveys*, vol. 34, n° 1, p. 1-47, 2002.

Singhal A., Salton G., Mitra M., Buckley C., « Document Length Normalization », *Information Processing & Management*, vol. 32, n° 5, p. 619-633, 1996.

Thi Tran M., Chekroud H., Thiery P., Julienne A., « Internet et soins : un tiers invisible dans la relation médecine/patient ? », *Ethica Clinica*, vol. 53, p. 34-43, 2009.

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

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

Yang Y., Liu X., « Re-examination of text categorisation methods », *Proc of 22nd Annual International SIGIR*, Berkley, p. 42-49, 1999.

Zeng Q., Tse T., « Exploring and developing Consumer Health Vocabularies », *JAMIA*, vol. 13, p. 24-29, 2006.

Zeng-Treiler Q., Kim H., Goryachev S., Keselman A., Slaugther L., Smith C. A., « Text characteristics of clinical reports and their implications for the readability of personal health records », *in* K. Kuhn, A. McGray (eds), *MEDINFO*, IOS Press, Brisbane, Australia, p. 1117-1121, 2007.

Zheng W., Milios E., Watters C., « Filtering for medical news items using a machine learning approach », *AMIA*, p. 949-53, 2002.

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

Zweigenbaum P., Jacquemart P., Grabar N., Habert B., « Building a Text Corpus for Representing the Variety of Medical Language », *MEDINFO*, p. 290-294, 2001.