Étude contrastive des documents vulgarisés et scientifiques de santé : sur la piste de la morphologie

Jolanta Chmielik[1,2], Natalia Grabar[1,3]

[1]INSERM, UMR_S 872, eq.20 Paris, F-75006 France; Université Paris Descartes, Paris, F-75006 France
[2]Commission européenne, Délégation à la recherche, Bruxelles, Belgique
[3]DIH, HEGP, AP-HP, Paris, France

Abstract

The amount of the health information on Internet is constantly growing although little is done to guide users towards documents which are the most appropriate to their expertise level. One of the distinctions is concerned with the specialization and technical level of documents: indeed documents differently specialized are co-existing while this difference is not explicitly indicated to users. The objective of this work is to propose clues for the automatic distinction between scientific and popularized medical content. Particularly, we propose to study the morphological level: morphological units, like bases and affixes, considered through their lexemes and morphological families. We apply the NLP tools, which provide access to the morphological content of documents. Our experiments and results highlight that this level may indeed provide clues for the distinction of the specialization of documents, and that they appear to be significant and stable across medical fields studied. Our work is done with French documents within three medical fields (cardiology, pneumology and diabetes).

Keywords

Medical Informatics, Access to Information, Language, Natural Language Processing, Documentation

1 Introduction

Une analyse récente de l'utilisation de l'Internet [1] montre qu'environ 80% de requêtes effectuées sont liées à la santé. Ce chiffre souligne la préoccupation que manifestent les citoyens vis-à-vis de leur santé. Il montre également que l'Internet propose des volumes d'information de santé extrêmement importantes. La qualité de ces documents n'est pourtant pas égale. Les initiatives comme CISMeF [2] ou HON [3] se concentrent sur la qualité et la transparence éthique des pages de santé : les documentalistes et experts du domaine médical effectuent une analyse des pages à indexer ou à certifier.

Mais à côté de ça, les documents de santé montrent aussi une différence quant à leur spécialisation et technicité. Ainsi, les documents scientifiques spécialisés (créés par des professionnels de santé à destination des professionnels de santé) et les documents vulgarisés (créés à destination des usagers non expert en médecine) co-existent. A ceci s'ajoutent aussi les documents médicaux de type didactique qui s'adressent aux étudiants en médecine. En fonction de la technicité des documents, qui se traduit souvent par les termes et mots employés et les contextes où ils apparaissent, les documents peuvent être plus ou moins difficiles à comprendre par des utilisateurs non expert en médecine. Cette hétérogénéité technique n'est pas transparente aux usagers, alors qu'il pourrait être utile de l'indiquer de manière explicite.

Il a été ainsi constaté que le niveau de technicité peut avoir un impact sur la compréhension des informations de santé par les non spécialistes et, par la suite, sur la communication des patients avec les professionnels de santé et la réussite des soins médicaux qui leur sont administrés [4,5]. Afin de guider les utilisateurs non expert vers des sources d'information qui leur sont plus appropriées, les moteurs de recherche pourraient faire la distinction entre les documents spécialisés et vulgarisés. Notons que les portails médicaux comme HON [3], CISMeF [2] ou le moteur de recherche général GoogleCoopsanté [6], proposent cette distinction mais elle est essentiellement basée sur une catégorisation manuelle de pages et sites Internet. Dans la situation actuelle, où le volume d'information augmente sans cesse sur l'Internet, il est souhaitable de chercher à faire la distinction entre les documents scientifiques et vulgarisés automatiquement. Pour cela il faut tout d'abord déterminer les critères sur lesquelles cette distinction peut être basée.

Nous faisons l'hypothèse qu'une description et analyse détaillées de différents niveaux linguistiques des documents scientifiques et vulgarisés peuvent fournir des critères complémentaires pour la détection des niveaux de spécialisation des documents. Dans ce travail, notre intérêt porte plus spécifiquement sur le niveau morphologique, qui de plus ne semble pas avoir été étudié dans la littérature. L'objectif que nous poursuivons concerne donc une étude des caractéristiques morphologiques des documents médicaux, vulgarisés et expert, et leur utilisation pour l'émergence et la sélection de critères morphologiques qui pourrait servir pour la distinction automatique de discours médicaux. Une telle étude se présente en effet comme une étape préalable à une classification automatique de discours. Nous utilisons les outils du traitement automatique de langues.

Dans la suite de ce papier, nous présentons d'abord l'état de l'art, ensuite le matériel et les méthodes appliquées. Nous présentons et discutons les résultats, et terminons avec des perspectives à ce travail.

2 Etat de l'art

Il existe plusieurs travaux en distinction automatique du niveau de spécialisation dans le domaine médical. Les formules linguistiques de lisibilité [7-9] prennent en compte les critères comme la longueur moyenne de mots ou de phrases. Ainsi, plus les mots sont longs, plus ils sont considérés comme savants. Ces formules peuvent être combinées avec le vocabulaire spécialisé (*ie*, MeSH [10]), ce qui permet de prendre en compte la dimension médicale

des documents du domaine, où les mots courts peuvent aussi être difficiles à comprendre [11]. Les approches qui appliquent les algorithmes d'apprentissage sont également très répandues et se basent sur différents types de descripteurs (n-grams de caractères [12], pondération manuelle [13] ou automatique [14] des termes MeSH, critères stylistiques [15] ou discursifs [16] des documents, leur niveau lexical [17]). Ce domaine d'activité est extrêmement actif actuellement et, de plus en plus, l'accent est mis sur la combinaison de différents descripteurs [18-20]. Par contre des études approfondies d'un type de descripteurs restent rare. A cet effet, citons par exemple l'alignement de vocabulaires scientifiques et vulgarisés [21, 22] ou l'étude assez détaillée du niveau syntaxique [23].

3 Matériel et méthodes

Dans les sections 3.1 à 3.3, nous présentons la manière dont nous avons constitué et sélectionné le matériel morphologique qui nous servira de base à ce travail. Nous décrivons ensuite (sec. 3.4) la méthode proposée pour l'étude contrastive des discours médicaux spécialisé et vulgarisé. Nous utilisons les outils du traitement automatique de langues et des outils et programmes Perl.

3.1 Corpus de documents scientifiques et vulgarisés

Notre travail est effectué sur les documents en français. Ces documents sont collectés au travers le portail CISMeF, qui propose des ressources satisfaisant les critères de qualité de l'information de santé. Nous exploitons deux types d'informations sur chaque ressource recensée par CISMeF : (1) Leur caractérisation selon les domaines médicaux effectuée grâce à l'indexation avec les mots-clés MeSH. Nous nous intéressons à trois spécialités médicales : cardiologie, pneumologie et diabète. (2) Leur caractérisation selon les types de discours et les types de pratiques sociales distinctes : grand public, professionnels de santé, et étudiants en médecine. Les ressources indexées sont proposées par CISMeF sous forme d'adresses URL téléchargeables. Les documents ont été téléchargés automatiquement à partir des listes d'URL, à l'aide d'un « aspirateur de Web » wget, paramétré pour télécharger les URL et les URL filles (par exemple, pour accéder aux documents fractionnés en plusieurs pages). Les documents ainsi téléchargés ont été ensuite filtrés pour sélectionner les documents au format HTML ou XML, que l'on peut convertir en texte plus facilement. Grâce à l'annotation CISMeF, ces documents ont été regroupés en des ensembles différenciés en fonction de leur spécialisation médicale et leur étiquetage comme expert, étudiant et vulgarisé. Au total, nous disposons des trois corpus : cardiologie, pneumologie et diabète, dont chacun se compose de trois parties : expert, étudiant et vulgarisé.

Tableau 1 : Taille des trois corpus étudiés

Domaine	Discours	Nombre de documents	Nombre d'occurrences
Cardiologie	expert	1 583	942 409
	étudiant	205	449 765
	vulgarisé	143	157 382

	total	1 931	1 549 556
Pneumologie	expert	724	600 524
	étudiant	127	213 379
	vulgarisé	134	96 559
	total	985	910 462
Diabète	expert	213	181 039
	étudiant	23	44 847
	vulgarisé	52	29 817
	total	236	255 703

La tableau 1 indique la taille de ces corpus. La colonne *Nombre de documents* indique le nombre de documents réunis dans chaque corpus (*Total* indique la somme des trois discours réunis). La colonne *Nombre d'occurrences* indique le nombre d'occurrences dans chaque ensemble. Nous pouvons voir que la taille des corpus varie d'une spécialité médicale à l'autre et d'un type de discours à l'autre. Le corpus cardiologie est ainsi le plus volumineux de tous, tandis que le corpus diabète est le plus petit. Par ailleurs, de manière générale, les textes expert sont les plus nombreux dans chaque spécialité. Les documents étudiant viennent en deuxième position et derniers en terme du volume arrivent les textes vulgarisés.

Ce matériel, constitué au travers le portail CISMeF, nous sert de gold standard dans notre étude. Les différences dans la taille des corpus correspondent sans doute à la réalité Internet, mais peuvent éventuellement introduire des biais dans les résultats que nous pourrons obtenir. Nous supposons par contre que la caractérisation des documents selon les types de ressources (*article de périodique, rapport technique, recommandation, brochure pédagogique pour les patients, information patient et grand public, cours, cas clinique, etc*) ou l'indexation avec les mots-clés MeSH, étant effectuée par des professionnels du domaine, est fiable.

3.2 Accès au niveau morphologique des documents textuels

Afin d'accéder au niveau morphologique de textes de nos corpus, nous appliquons des outils du traitement automatique de langues : (1) L'étiqueteur TreeTagger [24] assigne à chaque mot des documents une étiquette morpho-syntaxique et effectue une lemmatisation. Par exemple, *angioblastiques* est étiqueté comme adjectif (*ADJ*) et lemmatisé en singulier *angioblastique/ADJ*, *antiinflammatoire* est étiqueté comme pronom (*PRO*), etc. (2) Le lemmatiseur Flemm [25] reprend la lemmatisation proposée et la corrige ou bien ajoute des traits morphologiques supplémentaires. Ainsi, grâce à Flemm, l'étiquetage de *antiinflammatoire/PRO* est corrigé en adjectif. (3) Finalement, l'analyseur morphosémantique DériF [26] effectue l'analyse des termes en fonction de leur structure morphologique et sémantique. Il fonctionne en quatre étapes, que nous illustrons sur l'exemple de *angioblastique/ADJ* :

1. Calcul de l'arbre d'analyse d'un lemme étiqueté :

   [[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 :

   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)

La partie de la méthode qui suit exploite cette analyse morphosémantique pour le choix des bases qui seront analysées et pour la constitution de familles morphologiques.

3.3 Préparation de matériel morphologique

Nous avons ainsi considéré qu'un des critères de différentiation au niveau morphologique entre le discours des spécialistes et celui du grand public concerne l'usage qui est fait des bases : d'une part les lexèmes qu'elles permettent de former, d'autre part la systématicité et la fréquence avec laquelle ces unités sont utilisées dans les documents analysés.

Sélection de bases

Nous avons sélectionné, pour chaque type de discours et pour chaque spécialité médicale, des bases spécifiques à ces corpus et représentant les plus grandes familles morphologiques. Ainsi par exemple la base *cardio*, appartenant au sous domaine de la cardiologie, forme un nombre de lexèmes parmi les plus élevées dans le corpus cardio-étudiant : 57 lexèmes. Sur ce critère nous avons retenu *cardio* et comparé sa productivité dans les corpus cardio-expert et cardio-vulgarisé (respectivement, 26 et 20), de la même manière qu'au travers les corpus de pneumologie et diabète, qui nous servent de données témoins pour cette base.

Constitution de familles morphologiques

Une famille morphologique regroupe l'ensemble des lexèmes reliés par des relations formelles et sémantiques au travers des règles de formation de lexèmes. Dans notre cas, chaque famille morphologique est formée autour d'une base. Nous nous basons pour ceci sur l'analyse morphosémantique du lexique par Derif. L'analyse morphosémantique de chaque lexème est pris en compte, comme la notation entre les crochets dans ces exemples :

angioblastique/ADJ: *[[angi N*] [blast N*] ique ADJ]*

cardiogénique/NOM: [[[cardio N] [gène V*] ique ADJ] NOM]*

électrocardiographie/NOM: [[électr N] [[cardio N*] [graphie N*] NOM] NOM]*

echocardiographie/NOM: [[echo N] [[cardio N*] [graphie N*] NOM] NOM]*

cardiovasculaire/ADJ: [[cardio N] [vascul N*] aire ADJ]*

angiocardiographie/NOM: [[angi N] [[cardio N*] [graphie N*] NOM] NOM]*

Les lexèmes composés, comme *angioblastique* dans les exemples présentés, appartiennent à autant de familles que le nombre de bases qu'ils comportent. Ainsi, *angioblastique* enrichit deux familles (*angi, blast*), et *cardiogénique* enrichit aussi deux familles (*cardio, gène*), etc.

3.4 Méthodes pour l'étude contrastive des discours

Afin de faire émerger des caractéristiques morphologiques saillantes, nous abordons l'analyse contrastive des discours médicaux de deux points de vue : au travers la taille de familles morphologiques et au travers les fréquences que les bases sélectionnées montrent en corpus. Nous pensons en effet que les critères, issus de ces deux types d'observations, peuvent être détectés automatiquement en corpus et montrer une spécificité suffisamment importante pour la distinction automatique de discours.

La taille de familles morphologiques est étudiée au travers du nombre de lexèmes que chaque base étudiée forme. Ainsi, plus une base est productive plus sa famille morphologique sera grande. Nous prenons en compte deux valeurs pour mesurer la taille de familles : brute et normalisée. La taille brute correspond exactement au nombre de lexèmes formés. La taille normalisée représente une valeur normalisée par la taille des corpus.

Les fréquences de lexèmes en corpus sont étudiées sur les données étiquetées et lemmatisées par Flemm (pour éviter le biais de la variation des lexèmes). De la même manière, les fréquences des lexèmes sont observées au travers des valeurs brutes et normalisées.

4 Résultats et Discussion

4.1 Matériel morphologique sélectionné

Les bases sélectionnées à partir des données de Derif appartiennent à deux catégories morphologiques : les bases dites 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 par exemple au travers *gastrique*) et les bases autonomes (p. ex : *bronches, bactérie*). Nous avons ainsi 38 bases supplétives (*pulmon, vascul(o), vas(o), muscul, séro, artério, angi(o), athér(o), endo, uro, myo, cardio, ite, patho, pharyngo, gastr(o), glyc(o), graphie, hépat(o), néphr(o), phléb(o), psych(o), lip(o), neur(o), thromb(o), osté(o), pneum(o), cyt(o), gène, rhin(o), hém(o), ose, ectomie, tomie, scopie, ome, oïde*) et 7 bases autonomes (*lipide, diabète, insuline, veine, bronche, bactérie, thérapie*). Ces bases sont différemment représentées dans les corpus (spécialités, discours) étudiés : certaines sont répandues dans tous les corpus d'autres restent spécifiques (en terme d'occurrence) à une spécialité ou un discours.

Lors de la sélection des bases, nous nous sommes rendus compte de quelques limites dans l'analyse morphosémantique de DériF, mise à part l'incomplétude de son lexique de référence (par rapport auquel l'existence des lexèmes et des bases est identifiée) et la couverture partielle des bases et affixes (comme *-eux)* du français analysés par Derif. Ainsi, Derif propose parfois une analyse

morphosémantique erronée, comme dans l'exemple du lexème *gymnase/NOM,* analysé comme l'*« enzyme du nu »* :

gymnase/NOM: [[gymno A] [ase N*] NOM]*

" *(Partie de -- Type particulier de) enzyme caractérisé par la propriété : nu* "

Constituants = /gymno/ase/

Type = produit chimique

Une autre limite, qui avait plus de répercussion sur notre travail, concerne la notation des bases reconnues. Dérif identifie et regroupe les bases selon leur forme (chaîne de caractères). Par exemple, quatre bases suivantes (*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] (*hém(o), héma*), [hémato] et [èm]. Un regroupement de familles morphologiques effectué sur le critère formel va inévitablement former trois familles morphologiques distinctes, alors que le sens véhiculé par ces bases est le même. D'ailleurs, la glose sémantique associée reste la même (*relatif au sang, liée au sang*). Cet état de choses suggère que le regroupement de familles morphologiques devrait être basée autant sur la segmentation de lexèmes effectuée par Derif que sur la glose sémantique qu'il en propose.

4.2 Familles morphologiques

Nous avons constitué un ensemble de 46 familles morphologiques. La plupart de ces familles apparaissent dans tous les corpus. De manière générale, ces familles sont les mieux représentées dans les corpus cardiologie et pneumologie, ce qui n'est pas surprenant car ces corpus sont beaucoup plus grands que le corpus diabète. Par ailleurs, certaines familles morphologiques occurrent dans les corpus d'une spécialités médicales parmi les trois qui sont étudiées et, par conséquence, peuvent être absentes d’autres corpus. Par exemple, la famille de la base *athér*, que l’on peut associer au domaine de cardiologie, est absente des corpus vulgarisés de diabète et de pneumologie, mais apparaît dans les discours expert et étudiant de ces spécialités. Notons que trois familles (*rhin*, *tomie* et *phléb*) sont totalement absentes dans le corpus diabète (ce qui pourrait être expliqué par la petite taille de ce corpus).

Le nombre total de lexèmes que regroupent les 46 familles est de 2 295. Mais tous ces lexèmes n'apparaissent pas dans tous les corpus. Par exemple, dans le corpus cardio-étu, la famille *angio* comporte 19 lexèmes (*macroangiopathie, angiocardiographie, angioneurotique, microangiopathie, angiocoronarographique, angiocardiographique, cinéangiographie, angiopneumographie, angiomatose, angiographique, angiopathie, angiome, angioscopie, télangiectasie, angioscanner, angiographie, angioplastie, angiographique, angioplasticien*), alors qu'elle en comporte 15 dans le corpus cardio-expert. Notons que *angiomateux* n'est pas inclus dans cette famille : le lexème n'est pas segmenté en bases et affixes ni analysé car l'affixe *-eux* n'est pas traité dans la version actuel de l'outil. Ceci correspond à une limite de notre matériel, mais comme la situation est la même pour toutes les bases, nous supposons que cela relativise cette limite.

4.3 Etude comparative de la taille des familles morphologiques

La taille de chaque famille morphologique correspond à la productivité de la base qui permet de la fonder. Plus une famille est grande, plus la base est utilisée dans le corpus (par les locuteurs qui ont produit les documents), plus de lexèmes elle permet de former et plus son lexique est diversifié.

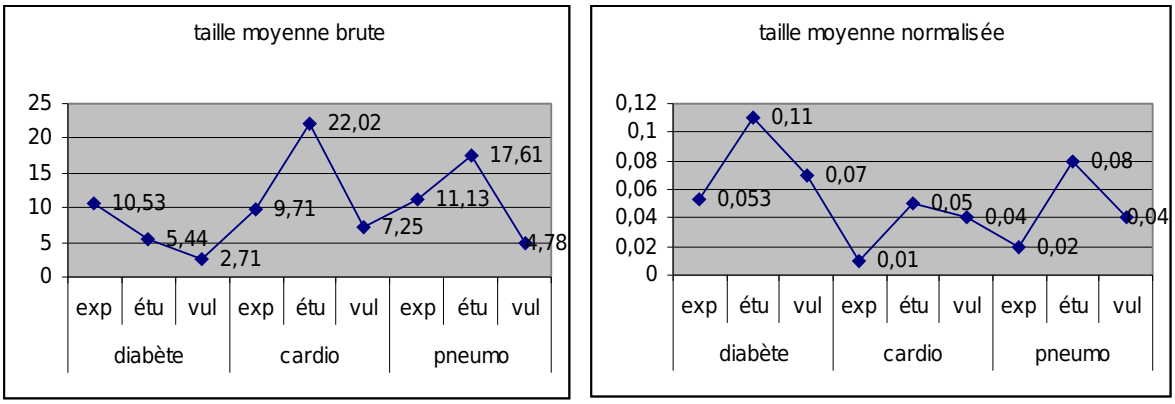

Figure 1: Taille moyenne des familles morphologiques : brute et normalisée

La figure 1 indique la taille de familles morphologiques : brute (fig.1 gauche) et normalisée (fig.1 droite). Nous pouvons en effet observer une variabilité selon les spécialités et les discours. Nous n'allons pas nous arrêter sur les résultats observés sur le corpus diabète, car ce corpus est relativement petit, mais surtout sur les deux autre corpus. Nous pouvons ainsi remarquer que les documents étudiants montrent toujours des familles morphologiques plus grandes que les deux autres discours, que ce soit avec les résultats bruts ou normalisés. En effet, les corpus étudiant réunissent des documents didactiques pour l'enseignement de la médecine. Compte tenu de leur destination, ces textes sont généralement plus précis et traitent des sujets avec plus de détails, ce qui explique la diversité du vocabulaire médical, et donc la taille de familles morphologiques plus grande que dans les autres discours.

Nous avons ainsi constaté que, dans toutes les trois spécialités, la taille normalisée de familles en corpus expert est inférieure à celle de textes vulgarisés (fig. 1 droite), alors que les valeurs brutes indiquent le contraire (fig. 1 gauche). L'observation d'une diversité plus grande du vocabulaire médical dans les corpus expert a été faite par un autre travail [11], où les auteurs étudient le langage médical suédois dans le domaine de maladies cardiovasculaires. Comme dans notre travail, une comparaison est effectuée entre les documents provenant des portails professionnels et les sites web grand public. Si nos deux travaux arrivent à des conclusions similaires, la différence quant au matériel utilisé persiste : nous positionnons notre travail au niveau de la morphologie (et unités lexicales correspondantes), alors que [11] étudient l'usage (et reconnaissance) des termes MeSH. Par ailleurs, lorsque nous analysons les valeurs normalisées, nous constatons que la taille de familles morphologiques normalisée est plus grande dans les textes vulgarisés que dans les textes expert. Ce cas de figure n'a pas été envisagé dans le travail cité. Justement, ce que les valeurs normalisées montrent c'est que, parmi toutes les occurrences des corpus, la proportion des bases étudiées, et du

lexique correspondant, est plus grande dans les corpus vulgarisés. Une des raisons de cette situation est que la taille de corpus expert est la plus importante (Tableau 1) : la normalisation de la taille des familles par la taille des corpus va inévitablement faire baisser cette valeur plus qu'elle ne le fait pour les corpus vulgarisés. En effet, les corpus vulgarisés sont en moyenne 6 fois moins volumineux que les corpus expert. Il est possible qu'une analyse de la taille de familles morphologiques dans des corpus de tailles comparables aurait aboutit à des résultats différents.

4.4 Etude comparative de l'utilisation de familles morphologiques en corpus

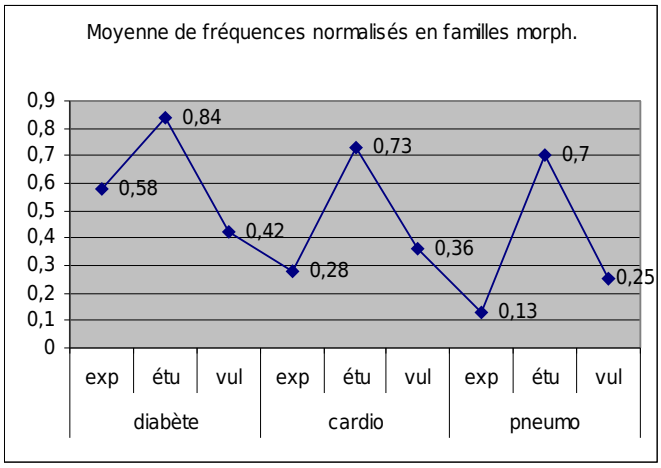

Figure 2: Fréquence normalisée des familles morphologiques en corpus

La figure 2 présente les fréquences normalisées d'apparition des familles morphologiques (au travers leurs lexèmes) dans les corpus. Dans toutes les trois spécialités médicales, cette utilisation est la plus élevée dans les corpus étudiant (0,84 en diabète, 0,73 en cardio, 0,70 en pneumo). Dans la section précédente, nous avons vu que dans le discours étudiant les familles étaient les plus fournies. Maintenant, nous pouvons voir de plus que dans ce discours la répétition est également importante. En deuxième position viennent les corpus vulgarisés (0,36 en cardio, 0,25 en pneumo). Enfin, l'utilisation et répétition, une fois normalisées, sont les plus basses dans les corpus expert (0,28 en cardio, 0,13 pneumo). Les documents expert sont certainement consacrés à des questions ciblées et précises et, par ailleurs, la grande taille des corpus expert semble pénaliser les valeurs normalisées.

La plus grande différence, quant aux fréquences et répétitions, se creuse entre les textes étudiant et expert, ensuite entre les textes étudiant et vulgarisés. Voilà quelques exemples de familles morphologiques qui montrent des fréquences élevées. La famille *diabète* est très fréquente en corpus diabète (expert et vulgarisé) ; la famille *hém* en corpus diabète étudiant ; la famille *pulmon* en corpus pneumologie. Le corpus cardiologie se distingue par trois bases *vascul* dans les documents expert, *cardio* dans les documents étudiant, et *graphie* dans les documents vulgarisés. En effet, ces familles morphologiques sont étroitement liées aux trois spécialités médicales : diabétologie (*diabète, hém*), cardiologie (*vascul, cardio, graphie*) et pneumologie (*pulmon*).

5 Conclusion et Perspectives

Dans le travail présenté, l'objectif poursuivi concerne la distinction automatique des discours vulgarisé, expert et étudiant dans le domaine médical. Plus particulièrement, nous effectuons une étude détaillée des caractéristiques du niveau morphologique de ces documents. L'étude est faite de manière contrastive : nous analysons les corpus expert, étudiant et vulgarisés en parallèle et ensuite nous effectuons une comparaison de leurs spécificités. Les documents du corpus sont collectés au travers le portail CISMeF, leur annotation par les documentalistes de ce portail sert donc de gold standard pour cette étude. Ce travail mise sur l'utilisation des outils du traitement automatique de langues qui permettent d'accéder au niveau morphologique de la langue.

Nous avons abordé les objectifs de deux points de vue. Tout d'abord, nous avons étudié la productivité de 46 bases morphologiques. La productivité des bases correspond au nombre de lexèmes qu'ils forment. L'étude a été faite séparément dans chaque discours. Ensuite, nous avons observé les fréquences d'utilisation des ces bases en corpus, également séparément dans chaque discours.

Ces deux analyses montrent qu'il existe en effet des caractéristiques morphologiques spécifiques à chacun des trois discours : scientifique (expert), didactique (étudiant) et vulgarisé. Les résultats obtenus corroborent dans une certaine mesure. Ainsi, nous avons constaté que les textes didactiques destinés à l'enseignement de la médecine se caractérisent par la plus grande redondance de termes médicaux, ceci dans les trois spécialités étudiées (cardiologie, pneumologie, diabète). En effet, la répétition fait partie intégrante du processus de l'apprentissage. Si l'on ajoute à cela la taille de familles morphologiques plus élevée que dans d'autres discours, nous voyons que les textes étudiant se distinguent à la fois par la diversité (taille de familles morphologiques) et la densité (fréquences) du vocabulaire médical. De part de ces deux caractéristiques, les textes didactiques se distinguent largement des textes destinés au grand public ou aux professionnels. Concernant les discours vulgarisé et expert, notre étude montre que la redondance du vocabulaire est plus élevée en textes vulgarisés qu'en textes expert, tandis que sa diversité est plus grande dans les textes expert lorsque l'étude est faite avec des données brutes. Une fois normalisées, les données montrent que les textes expert sont proportionnellement moins redondants que les textes vulgarisés.

La perspective principale de notre travail concerne l'utilisation des critères morphologiques pour la distinction automatique de discours. En effet, les résultats obtenus dans le présent travail suggèrent que la distinction automatique pourra fournir des résultats assez fiables. Dans cette étude, nous nous sommes concentrés sur 46 familles morphologiques, alors que l'ensemble de familles morphologiques pourrait être pris en compte. Toutefois, le succès de cette piste dépend de l'efficacité des outils du traitement automatique de langues et nous avons vu que l'analyse morphosémantique, bien que efficace pour de très nombreux lexèmes, montre encore quelques limites (*ie*, analyse erronée des lexèmes, manque de normalisation des bases, analyse incomplète). Au travers la taille des familles morphologiques, nous avons abordé la notion de productivité des bases. Il serait intéressant d'explorer plus en détail la

productivité morphologique de nos corpus en appliquant des approches proposées à cette effet [27]. Notons que la productivité des bases et affixes semblent être en effet spécifique des discours et genres [28]. Cette étude pourrait être étendue à des documents provenant d'autres sources (*ie.,* Santé Canada) et à d'autres spécialités médicales.

Références

[1] Fox S. On line 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, 2006; Washington DC

[2] www.chu-rouen.fr/cismef

[3] www.hon.ch

[4] 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, 1999; 281(6), pp. 552–7

[5] McCray A. Promoting health literacy. Journal of American Medical Informatics Association, 2005; 12, pp. 152–163

[6] www.google.com/coop

[7] Flesch R. A new readability yardstick. Journal of Applied Psychology, 1948; 23, pp. 221–233

[8] Gunning R. The art of clear writing. New York, NY : McGraw Hill, 1973

[9] Björnsson & Härd af Segerstad, Segezstad B. Lixpä franska och tio andra språk. Stockholm:Pedagogiskt centrum, Stockholms skolförvaltning. 1979

[10] Medical Subject Headings. National Library of Medicine, Bethesda, Maryland. http://www.nlm.nih.gov/mesh/meshhome.html.

[11] Kokkinakis D, Gronostaj MT. Comparing lay and professional language in cardiovascular disorders corpora. In A. Pham T., James Cook University, Ed., WSEAS Tran sactions on BIOLOGY and BIOMEDICINE, 2006; pp. 429–437

[12] Poprat M., Marko K. & Hahn U. A language classifier that automatically divides medical documents for experts and health care consumers. In MIE 2006; pp. 503–508

[13] Zheng W, Millos E., Watters C. Filtering for medical news items using a machine learning approach. In AMIA 2002; pp. 949–53

[14] Borst A, Gaudinat A, Boyer C, Grabar N. Lexically based distinction of readability levels of health documents. MIE 2008

[15] Grabar N, Krivine S, Jaulent MC. Classification of health web pages as expert and non expert with a reduced set of cross-language features. In AMIA 2007; pp. 284–8

[16] Goeuriot L, Grabar N, Daille B. Caractérisation des discours scientifique et vulgarisé en français, japonais et russe. TALN 2007; pp. 93-102

[17] Miller T, Leroy G, Chatterjee S., Fan J., Thoms B. A classifier to evaluate

language specificity of medical documents. In HICSS, 2007; pp. 134–140

[18]Wang Y. Automatic recognition of text difficulty from consumers health information. In IEEE, Ed., Computer-Based Medical Systems, 2006; pp. 131–136

[19]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 MEDINFO 2007; pp. 1117–1121

[20]G. Leroy; S. Helmreich; J. Cowie; T. Miller; W. Zheng. Evaluating Online Health Information: Beyond Readability Formulas. In AMIA 2008; pp. 394-8

[21]www.consumerhealthvocab.org

[22]Deléger L, Zweigenbaum P. Paraphrase acquisition from comparable medical corpora of specialized and lay texts. In AMIA 2008; pp. 146-50

[23]Zeng Q., Tse T. Exploring and developing consumer health vocabularies. JAMIA 2006; 13, 24–29

[24]Schmid H. Probabilistic part-of-speech tagging using decision trees. In Proceedings of the International Conference on New Methods in Language Processing, 1994; pp. 44–49

[25]Namer F. Flemm: Un analyseur Flexionnel de Français à base de règles. Traitement automatique des Langues, 2000; 41(2), pp.523-47

[26]Namer F. Automatiser l’analyse morphosémantique non affixale: le système DériF. Cahiers de Grammaire, 2003; 28, 31–48.

[27]Bayen H. Quantitative aspects of morphological productivity. Yearbook of Morphology. 1991; pp. 109-149

[28]Baayen H. Derivational productivity and text typology. Journal of quantitative linguistics, 1994; 1(1), pp. 16-34

Adresse de correspondance

Natalia Grabar

SPIM, UMRS Inserm 872 éq. 20

Centre de Recherche des Cordeliers

15 rue de l'Ecole de Médecine

75006 Paris

natalia.grabar@spim.jussieu.fr