TALN 2003, Batz-sur-Mer, 11–14 juin 2003

Apprentissage de relations morphologiques en corpus

Pierre Zweigenbaum, Fadila Hadouche, Natalia Grabar
Mission de recherche en Sciences et Technologies de l'Information Médicale,
STIM/DPA/DSI, Assistance Publique – Hôpitaux de Paris
& ERM 202, INSERM
{pz,fha,ngr}@biomath.jussieu.fr http://www.biomath.jussieu.fr/

Résumé - Abstract

Nous proposons une méthode pour apprendre des relations morphologiques dérivationnelles en corpus. Elle se fonde sur la cooccurrence en corpus de mots formellement proches et un filtrage complémentaire sur la forme des mots dérivés. Elle est mise en œuvre et expérimentée sur un corpus médical. Les relations obtenues avant filtrage ont une précision moyenne de 75,6 % au 5000è rang (fenêtre de 150 mots). L'examen détaillé des dérivés adjectivaux d'un échantillon de 633 noms du champ de l'anatomie montre une bonne précision de 85–91 % et un rappel modéré de 32–34 %. Nous discutons ces résultats et proposons des pistes pour les compléter.

We propose a method to learn derivational morphological relations from a corpus. It relies on corpus cooccurrence of formally similar words, with additional filtering on the form of derived words. It is implemented and tested on a medical corpus. The relations obtained before filtering have an average precision of 75.6% at rank 5000 (150-word window). A detailed examination of derived adjectives for a sample of 633 anatomy nouns shows a good precision of 85–91% and a moderate recall of 32–34%. We discuss these results and propose directions for improvement.

Mots-clefs – Keywords

Morphologie, apprentissage, corpus, langue de spécialité, médecine
Morphology, learning, corpus, specialized language, medicine

1 Introduction

Pour palier le manque pour le français de bases de connaissances morphologiques telles que CELEX (Burnage, 1990), une base morphologique est en cours de construction pour le « français général » dans le cadre du projet MorTAL (Hathout *et al.*, 2002). Des bases complémentaires seront utiles pour les langues de spécialité ; par exemple, le « Specialist Lexicon » de l'UMLS fournit des données et des outils pour la morphologie flexionnelle et dérivationnelle de l'anglais médical (McCray *et al.*, 1994). L'équivalent n'est pas encore disponible pour le français médi-

cal ; le projet UMLF[1] a donc pour objectif la constitution d'un lexique médical francophone incluant des connaissances flexionnelles et dérivationnelles (Zweigenbaum *et al.*, 2003).

Pour faciliter ce travail, de la même façon que dans le projet MorTAL (Hathout *et al.*, 2002), il est utile d'assister automatiquement le recensement de ces liens morphologiques. Dans des travaux antérieurs, nous avons abordé l'acquisition automatique de ressources morphologiques à partir de terminologies structurées (Grabar & Zweigenbaum, 1999; Grabar & Zweigenbaum, 2000). Cependant, la taille nécessairement limitée des terminologies disponibles limite le vocabulaire et la variation morphologique qui y sont effectivement présents ; et le caractère normatif des termes qui y sont consignés peut masquer des usages effectifs différents et donc d'autres relations morphologiques. En complément des terminologies, nous avons donc cherché à explorer l'apprentissage de ce même type de ressources morphologiques à partir d'une autre source : des corpus. C'est maintenant un lieu commun d'indiquer qu'il est de plus en plus aisé, essentiellement grâce au Web, de se procurer des corpus d'une taille qui va grandissante. Si l'on peut contrôler suffisamment les conditions de sélection des documents, on peut à la fois rester dans un domaine spécifique (ici le domaine médical) et représenter des types de documents divers, favorisant la richesse du vocabulaire impliqué. Il reste à concevoir une méthode permettant de mettre au jour des relations morphologiques entre les mots du corpus construit.

De nombreux travaux ont été menés ces dernières années sur l'apprentissage de relations morphologiques. Citons entre autres (Jacquemin, 1997; Xu & Croft, 1998; Grabar & Zweigenbaum, 1999; Gaussier, 1999; Dal *et al.*, 1999; Daille, 1999; Hathout *et al.*, 2002; Tanguy & Hathout, 2002; Namer, 2002). Parmi les méthodes fondées sur corpus, certaines se situent à la charnière entre traitement automatique des langues et recherche d'information. Nous nous attarderons sur la méthode de (Xu & Croft, 1998), qui fonctionne en corpus sans terminologie ni règles a priori. Les mots du corpus sont réduits par un raciniseur algorithmique « agressif » (Porter, 1980). Deux mots racinisés à la même forme réduite et qui cooccurrent significativement plus qu'ils ne le feraient s'ils étaient indépendants (variante de l'information mutuelle) sont considérés comme faisant partie de la même classe d'équivalence morphologique.

Ne disposant pas de l'équivalent du raciniseur de Porter pour le français, nous proposons ici une adaptation de cet algorithme à notre problème, avec de plus une focalisation sur l'identification de mots dérivés (section 2). Nous mettons en œuvre et évaluons cet algorithme sur un corpus médical (section 3), puis discutons les résultats obtenus (section 4).

2 Méthodes

Nous présentons le corpus constitué pour les expérimentations (section 2.1), la méthode proposée, mise en œuvre initialement dans (Hadouche, 2002) (section 2.2), le filtrage supplémentaire des mots dérivés (section 2.3) et le mode d'évaluation des résultats obtenus (section 2.4).

2.1 Constitution d'un corpus médical

Le corpus qui va nous servir de base de travail a été construit à partir du Web en s'adossant au catalogue CISMeF (Darmoni *et al.*, 2000). CISMeF recense des sites médicaux francophones (11 000 en 2002) satisfaisant des critères de qualité, et les indexe avec des mots clés du thesaurus

[1] ACI #02C0163, Ministère de la recherche, Réseau national des technologies pour la santé, 2002–2004.

MeSH, ce qui en fait un outil de choix pour constituer un corpus médical. Nous avons collecté toutes les pages cataloguées sous le mot clé *Signes et symptômes*. Pour nous affranchir des problèmes liés aux adresses pointant sur des cadres (« frames ») ou constituant de simples sommaires, nous avons également collecté les pages situées un lien plus loin au-dessous de chaque page indiquée. Nous les avons ensuite converties en texte brut (dans le présent travail, seuls les documents HTML ont été convertis), puis avons filtré les lignes non écrites en français en adaptant la méthode et en reprenant les données de (Grefenstette & Nioche, 2000). Nous avons ensuite étiqueté ce corpus avec TreeTagger (Schmid, 1994) couplé au lemmatiseur FLEMM (Namer, 2000). Le corpus résultant contient 4 627 documents et 5 204 901 occurrences de mots (180 000 formes différentes et 142 000 lemmes différents). Nous avons conservé ses « mots pleins » (nom, adjectif, verbe, adverbe), soit 2 055 419 occurrences. Beaucoup de ces « mots » sont bruités ; nous avons supprimé des espaces insécables qui étaient restés collés en début ou en fin de mot, puis éliminé les mots qui contenaient encore des caractères non-alphanumériques autres que le tiret. Il reste alors 2 041 627 occurrences (54 324 lemmes différents).

2.2 Apprentissage de relations morphologiques en corpus

Le principe fondateur de la méthode de (Grabar & Zweigenbaum, 1999) est de (i) repérer des mots proches par la graphie et qui (ii) possèdent des liens sémantiques. La méthode de (Xu & Croft, 1998) procède de même, en raffinant le premier critère grâce à un raciniseur existant (Porter, 1980). Dans une terminologie structurée, nous avons instancié ce principe en repérant (i) des mots qui partagent la même chaîne de caractères initiale et qui (ii) figurent dans des termes reliés par des liens sémantiques (Grabar & Zweigenbaum, 1999).

En corpus, les liens sémantiques vont s'appuyer sur la notion de continuité thématique : le fait que le thème du discours ne change pas à chaque phrase. Cette continuité se traduit généralement par des liens thématiques lexicaux (redondance lexicale) : les mots employés pour parler d'un thème donné ont souvent des liens (par exemple, *hôpital*, *médecin*, *opérer*). Ces liens thématiques peuvent être instanciés par des mots d'une même famille morphologique (*opérer*, *opération*). Par conséquent, parmi les mots employés à l'intérieur d'un segment de texte thématiquement homogène, on trouve souvent des mots d'une même famille morphologique.

La méthode de Xu et Croft approxime cette notion de continuité thématique à l'aide d'une fenêtre glissante de N mots. La proximité morphologique entre deux mots leur est donnée par le raciniseur de Porter. Nous reprenons cette méthode en remplaçant ce dernier par un « raciniseur » encore plus agressif : la réduction aux c premiers caractères du mot ($c = 4$ dans nos expérimentations). En résumé, nous recensons les mots qui partagent la même chaîne de caractères initiale de longueur supérieure ou égale à c et qui se trouvent « souvent » dans une même fenêtre de M mots. Ce dernier critère sera mis en œuvre par une mesure statistique d'association qui évalue dans quelle mesure cette cooccurrence est plus fréquente que ce que donnerait le hasard. Nous avons choisi le rapport de vraisemblance (« likelihood ratio ») (Manning & Schütze, 1999) : rapport $\lambda = \frac{L(H_1)}{L(H_2)}$ entre la probabilité d'observer le nombre de cooccurrences du mot m_2 avec le mot m_1 dans l'hypothèse H_1 où les mots sont indépendants et la probabilité d'observer leur nombre de cooccurrences dans l'hypothèse H_2 où les mots sont dépendants (on calcule $-2 \log \lambda$).

Les données pour calculer ce rapport sont les suivantes. Probabilité de l'observation selon H_1 (indépendance) : $L(H_1) = b(c_{12}, c_1, p)b(c_2 - c_1, N - c_1, p)$; probabilité de l'observation selon H_2 (dépendance) : $L(H_2) = b(c_{12}, c_1, p_1)b(c_2 - c_1, N - c_1, p_2)$; loi binômiale (probabilité d'une

séquence de k succès parmi n tirages) : $b(k,n,p) = C_k^n p^k (1-p)^{n-k}$; probabilités élémentaires : $p = \frac{c_{12}}{N}$; $p_1 = \frac{c_{12}}{c_1}$; $p_2 = \frac{c_2 - c_{12}}{N - c_1}$; c_1 est le nombre d'occurrences du mot m_1, c_2 est le nombre de fenêtres où apparaît le mot m_2, c_{12} est le nombre de fenêtres où cooccurrent les mot m_1 et m_2, N est la taille du corpus.

(Xu & Croft, 1998) appliquent un seuil d'association en-dessous duquel les couples cooccurrents sont éliminés. Nous considérons cependant que ce critère d'association doit être pris comme un facteur parmi d'autres pour le *classement* des couples potentiellement en relation morphologique plutôt que pour leur élimination directe. Notons que cette mesure d'association est asymétrique car elle dépend différemment de la fréquence propre de chaque mot. Ainsi, on a plus de chances d'observer le nom *canal* (481 occurrences) dans le voisinage de l'adjectif *canalaire* (65 occurrences) que l'inverse ; notons que l'adjectif peut également être le plus fréquent des deux (63 occurrences pour *laryngé* contre 48 pour *larynx*). Nous conservons le score d'association le plus fort des deux directions.

2.3 Critères de filtrage des mots dérivés

Pour le projet UMLF, nous avons besoin plus particulièrement de connaissances dérivationnelles[2] ; nous allons commencer ici par les couples nom – adjectif dérivé par suffixation. Pour les repérer, nous avons ajouté à la méthode précédente des tests reflétant des caractéristiques supplémentaires des dérivés suffixaux :

1. Pas de dérivation régressive : la dérivation ajoute un suffixe[3]. Le test correspondant est la différence de longueur entre les deux mots concernés ; pour conserver un peu de souplesse, nous acceptons un dérivé s'il contient jusqu'à un caractère de moins que le mot de base (par exemple, *sacrum* / *sacré*).

2. Dérivation plutôt que composition : la composition savante assemble des morphèmes, d'origine généralement grecque ou latine, qui sont en moyenne plus longs que les suffixes de la dérivation. Le système considère comme suspect (et élimine) un « dérivé » qui dépasse de plus de 5 lettres le mot de base (par exemple, *bronche* / *bronchopneumonique*).

3. Fréquence de la « règle » : le même opérateur morphologique s'applique généralement à plus d'un mot. Le test correspondant est le nombre de radicaux différents (chaînes initiales communes maximales) sur lesquels l'opérateur est observé. En approximant un opérateur par un couple de terminaisons substituées (par exemple, *-e* / *-ique* dans *kyste* / *kystique*), nous ne tenons pas compte des éventuelles adaptations morpho-phonologiques et séparons artificiellement ses diverses réalisations. Par exemple, la règle *-e* / *-ique* est observée 138 fois, et une règle distincte *-e* / *-que* (*urémie* / *urémique*) 242 fois.

Les critères sont appliqués dans l'ordre indiqué : les critères de forme (1 et 2) éliminent certains couples candidats. Le critère de fréquence de la règle (3) s'applique ensuite : si plusieurs adjectifs sont proposés pour le même nom, celui correspondant à la règle d'application la plus

[2]Nous utilisons le terme *dérivation* dans son sens plus restreint d'*affixation*, et nous nous focalisons sur la suffixation. Rappelons que nous approximons cette suffixation par une proximité formelle ; parmi les couples répertoriés, nous trouverons de « vrais » dérivés, mais également des mots d'une même famille morphologique (*articulation* / *articulaire*). Dans les deux cas, nous parlons de couples de mots dérivés.

[3]Nous ne nous intéressons pas aux *conversions*, où la forme reste inchangée ; elles sont faciles à repérer, et sont éliminées par notre méthode.

fréquente est conservé. La force d'association est ensuite prise en considération pour départager deux couples produits par des règles de fréquence identique : *tendon* / *tendineux* (fréq = 1, association = 86) plutôt que *tendon* / *tendinite* (fréq = 1, association = 11 — erreur d'étiquetage).

Enfin, les couples produits par une règle de très faible fréquence ne sont conservés que si leur force d'association est suffisante. Nous avons fixé expérimentalement un seuil d'association de 50 pour les couples issus de règles hapaxiques (une seule application) : cela élimine *colonne* / *colobomateux* (association = 10,94), mais conserve *cortex* / *cortical* (association = 173,07).

2.4 Expérimentation et évaluation

Nous avons procédé à deux évaluations. La première concerne la précision de l'ensemble des couples cooccurrents relevés (section 2.2), avant filtrage supplémentaire, en fonction de leur score d'association. Pour cela, parmi les couples classés par ordre décroissant d'association, nous avons examiné la proportion cumulée de couples corrects en fonction du nombre de couples examinés, ainsi que le pourcentage local de ces cooccurrents satisfaisants (calculé sur des tranches de 200 couples). Le résultat est une courbe *précision cumulée(rang)* et une courbe *précision locale(rang)*.

La seconde évaluation, plus spécifique, mesure la précision et le rappel de la recherche des dérivés adjectivaux (section 2.3) d'un échantillon de noms du champ sémantique de l'anatomie. Cet échantillon a été construit en partant des termes d'anatomie (axe T, *topographie*) du Répertoire d'anatomopathologie de la nomenclature SNOMED Internationale (Côté, 1996). Nous avons étiqueté les termes de cette terminologie (Grabar & Zweigenbaum, 2000) et pris le premier nom de chaque terme. Une liste de 633 noms a ainsi été obtenue, dont ceux qui commencent par la lettre *a* suivent : *abdomen, acinus, acromion, acétabulum, adventice, adénohypophyse, aile, aine, aisselle, alvéole, amas, amnios, ampoule, amygdale, anastomose, angle, anneau, annexe, anse, anthélix, antre, anus, aorte, apex, aponévrose, apophyse, appareil, appendice, aqueduc, arachnoïde, arcade, arc, arrière, articulation, artère, artériole, aréole, arête, astragale, astrocyte, atlas, auricule avorton, axis, axone*.

Les méthodes exposées ci-dessus ont été appliquées au corpus CISMeF (section 2.1). La détection des cooccurrents morphologiques a été testée avec des tailles de (demi-)fenêtre de 10, 40, 100 et 150 mots (une demi-fenêtre de M mots correspond à une distance maximale de M mots entre le mot pivot m_1 centre de la fenêtre et un cooccurrent m_2). La seconde évaluation s'est faite avec les couples obtenus avec la fenêtre de taille 100.

Les traitements ont été effectués à l'aide de scripts Perl. Étant donné leur nombre, le décompte des cooccurrences ne peut se faire en mémoire, et des « hachages liés » sont employés. La gestion des couples candidats, des règles associées et des mots d'anatomie a employé, par commodité, une base de données relationnelle (PostgreSQL).

3 Résultats

3.1 Évaluation de l'ensemble des cooccurrents trouvés

Avec une fenêtre de taille 150, 48 002 couples cooccurrents différents (un seul ordre étant conservé) sont relevés. La figure 1 montre les courbes *précision cumulée(rang)* et *précision*

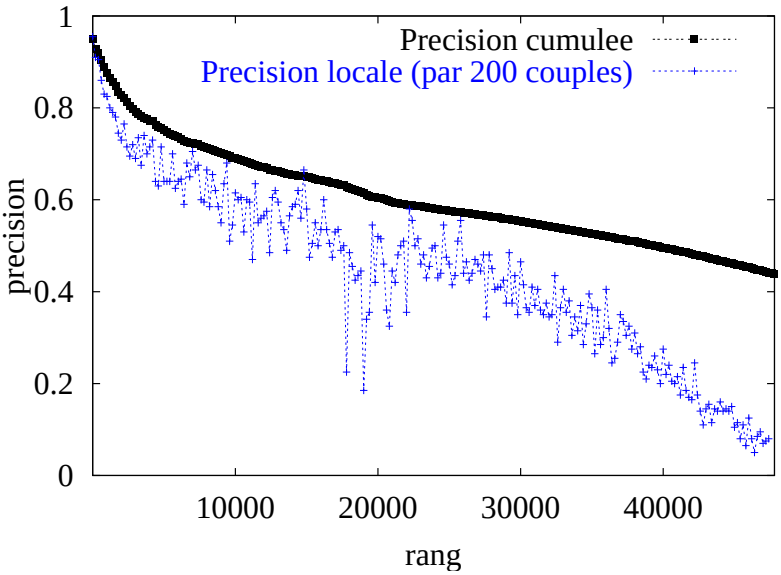

Figure 1: Précision cumulée et locale des couples cooccurrents trouvés avant filtrage, en fonction de leur rang par ordre décroissant de score d'association.

locale(rang). Par exemple, sur les 5 000 premiers cooccurrents trouvés avec une fenêtre de 150 mots, 3778 étaient corrects (précision cumulée = 75,6 %). Localement, les 200 couples trouvés du rang 4801 au rang 5000 avaient une précision de 71,5 %, les 200 précédents une précision de 63,0 %. Les courbes montrent que la précision décroît avec le rang, et que les derniers rangs ont une précision locale très basse (moins de 20 % pour les 6000 derniers couples). Cela confirme la pertinence de la mesure d'association. Néanmoins, on continue à trouver un nombre non négligeable de couples corrects même aux rangs les plus élevés (728 parmi les 6000 derniers couples). Il serait dommage d'éliminer ces couples : le filtrage mis en place dans la suite des traitements s'efforce de repérer des couples pertinents parmi ceux dont l'association est faible.

Parmi les erreurs, on trouve des omissions d'accents (*hypoglycémie* / *hypoglycemie*), des erreurs d'orthographe (*travaille* / *travalle*), des erreurs de segmentation (mots collés : *maladie* / *maladiede*), des préfixes (très nombreux : *trans*, *télé*, *hyper*, *hypo*, *iso*, *méso*, etc.), des composés avec des tirets (*chien* / *chien-guide*, *aldostérone* / *aldostérone-synthase*), et des mots de différentes langues (anglais, espagnol, allemand) qui n'ont pas été correctement filtrés. On relève d'ailleurs de belles paires espagnoles (*nuevo* / *nueva*, *infeccione* / *infectada*) ou anglaises (*child* / *children*), qui illustrent le fait que la recherche d'associations est indépendante de la langue... Sur le même registre, la lemmatisation ne traite souvent pas bien les mots latins.

3.2 Évaluation sur un échantillon de noms d'anatomie

Le tableau 1 montre, à titre d'exemple, les propositions d'adjectifs dérivés pour les noms d'anatomie commençant par la lettre *a* : sur 45 noms, 17 se voient proposer un adjectif dérivé, dont 13 sont considérés comme corrects. Le tableau 2 donne les résultats de l'évaluation de la précision et du rappel de la recherche de dérivés adjectivaux pour tous les noms d'anatomie (le paragraphe suivant définit les « dérivés attendus »). En bon accord avec l'exemple des noms en *a*, on observe une précision de l'ordre de 85–91 % et un rappel de 32–34 %.

Nom	Adjectif	# cooc	loglike	c.i.c.m.	suf1	suf2	f
abdomen	abdominal	101	584.2161	abdom	en	inal	2
acinus	acini*	1	16.2631	acin	us	i	3
alvéole	alvéolaire	24	213.3956	alvéol	e	aire	72
amygdale	amygdalien	8	100.2468	amygdal	e	ien	24
anastomose	anastomotique	20	222.8728	anastomo	se	tique	29
angle	anglais*	9	26.0384	angl	e	ais	2
annexe	annexes*	1	10.4252	annexe		s	46
antre	antral	13	170.0993	antr	e	al	42
aorte	aortique	170	1314.7484	aort	e	ique	131
apophyse	apophysaire	3	39.6601	apophys	e	aire	72
appareil	apparent*	16	52.1557	appare	il	nt	1
appendice	appendiculaire	19	225.241	appendic	e	ulaire	5
articulation	articulaire	216	1406.3482	articula	tion	ire	13
artériole	artériolaire	15	99.9927	artériol	e	aire	72
aréole	aréolaire	2	27.5564	aréol	e	aire	72
astrocyte	astrocytaire	2	28.6077	astrocyt	e	aire	72
axone	axonal	8	93.2184	axon	e	al	42

Table 1: Proposition d'adjectifs dénominaux pour des noms du champ sémantique de l'anatomie. *# cooc* = nombre de fenêtres où les deux mots sont cooccurrents ; *c.i.c.m.* = chaîne initiale commune maximale ; *suf1* = chaîne finale du nom ; *suf2* = chaîne finale de l'adjectif ; *f* = fréquence de la règle *-suf1/-suf2*. L'astérisque indique les couples « incorrects ».

étalon	# noms	# adj proposés	# corrects	précision	rappel
dérivés attendus	633	235	200	85 %	32 %
tous les dérivés	633	235	213	91 %	34 %

Table 2: Précision et rappel de la recherche d'adjectifs dénominaux (anatomie).

3.3 Étude des erreurs et du silence

Parmi les adjectifs dérivés proposés, on trouve 13 adjectifs que l'on peut considérer comme dérivés du nom correspondant, mais qui ne constituent pas l'adjectif relationnel attendu. Par exemple, *média* / *médiatique* (*médial*), *sang* / *sanglant* (*sanguin*), *système* / *systématique* (*systémique*), *embryon* / *embryonné* (*embryonnaire*). Certains des adjectifs attendus sont formés sur une base supplétive, que notre méthode ne permet pas de repérer : *poil* / *poilu* (*pileux*), *figure* / *figuré* (*facial*), *pointe* / *pointu* (*apical*). Selon que l'on compte comme « corrects » ou pas ces dérivés, on obtiendra deux mesures de précision (tableau 2) : l'une où *tous les dérivés* sont acceptés, et l'autre où seuls les *dérivés attendus* le sont.

Dans le reste, 2 dérivations sont étymologiquement correctes, mais non pertinentes dans notre contexte : *partie* / *partiel*, *passage* / *passager*. 4 composés savants ont passé notre filtre heuristique : *monocyte* / *monocytogene*, *iléon* / *iléorectal*, *myélocyte* / *myéloïde*, *érythroblaste* / *érythrocytaire*. 13 couples erronés proviennent d'un mauvais étiquetage de mots en tant qu'adjectifs : mots erronés (*côlonb*), anglais (*origin*), ou français (*pochite*). Enfin, 3 couples constituent des erreurs pures et simples : *appareil* / *apparent*, *angle* / *anglais*, *trou* / *trouble*.

Nous avons examiné systématiquement les raisons du silence pour les 45 noms d'anatomie commençant par la lettre *a* (tableau 3). Un nombre important de ces noms ne figuraient pas dans le corpus (ou n'ont pas été étiquetés en tant que tels) ; au total, parmi les 633 noms exam-

Diagnostic	Nombre	%	Exemples
Trouvé automatiquement	13	29	*abdomen* / *abdominal*
Nom absent du corpus	4	9	*adénohypophyse*, *amnios*
Adjectif absent ou inconnu	15	33	*aile*, *aisselle*, *amas*, *anse*
Chaîne initiale commune < 4	6	13	*artère* / *artériel*, *apex* / *apical*
Non cooccurrents	4	9	*aponévrose* / *aponévrotique*
Base supplétive	1	2	*aine* / *inguinal*
Conversion	1	2	*arrière*
Cooccurrents mais mal classés	1	2	*annexe* / *annexiel*
Total	45	99	

Table 3: Causes de silence pour les noms d'anatomie commençant par *a*. Les cas pris en compte dans une rangée ne le sont plus dans les rangées suivantes.

inés, 89 (soit 14 %) ne se trouvaient pas dans le corpus, et ne pouvaient donc se voir attribuer d'adjectif dérivé par notre méthode. Un nombre encore plus grand (près d'un tiers) des noms, pourtant présents dans le corpus, n'ont pas d'adjectif dérivé identifiable dans le corpus. La contrainte des quatre caractères initiaux élimine des couples qui auraient été fortement associés (par exemple, *artère* / *artériel*, *ampoule* / *ampullaire*). Enfin, certains noms et adjectifs appariés étaient présents dans le corpus, mais pas ensemble dans une fenêtre de 100 mots pleins.

4 Discussion et conclusion

La méthode proposée repère un grand nombre de couples en relations morphologique, dont une part importante est correcte (avec une fenêtre de 150 mots, 70 % au 9000è rang, 43,9 % soit 21 000 pour l'ensemble des 48 000 couples). Une proportion importante des erreurs peut de plus être filtrée en prenant en compte des critères supplémentaires (section 2.3) sur les couples de mots dérivés. Ces critères sont mis en œuvre sur un échantillon de noms d'anatomie, dans une tâche de recherche d'adjectifs dérivés qui obtient une précision de 85 à 91 %. La proportion de noms pour lesquels un adjectif dérivé est trouvé reste cependant modeste (32 à 34 %). Ceci étant, la bonne précision obtenue permet d'envisager la combinaison de cette méthode avec celle, de précision encore meilleure, appliquée aux terminologies (Grabar & Zweigenbaum, 2000), afin d'augmenter leur rappel total.

La méthode proposée ici possède plusieurs limitations ; indiquons-en quelques unes. L'examen manuel de l'ensemble des 48 000 couples est une tâche fastidieuse, qui a pris plusieurs jours, au cours de laquelle des erreurs d'appréciation ont pu se produire (dans un sens comme dans l'autre)[4]. Nous pensons néanmoins que leur influence reste marginale dans les chiffres produits.

Si l'on extrapole l'étude effectuée sur les noms en *a*, de l'ordre de 42 % (9+33) des couples dérivés recherchés n'étaient pas présents dans le corpus. Il faut toutefois tempérer cela par le fait que certains des noms considérés n'ont peut-être pas d'adjectif dérivé (*amas*, *avorton* ?). En tout état de cause, une première mesure sera d'élargir et de diversifier le corpus : d'une part encore sur le Web, d'autre part en puisant dans des genres qui n'y sont pas représentés, comme les comptes rendus hospitaliers. La recherche d'une complémentarité avec les mots présents

[4]Les personnes intéressées par ces données peuvent contacter les auteurs.

dans les terminologies médicales constituera la seconde mesure. Notons que si l'on met de côté ces 42 % hors corpus, le rappel monte à 50 % pour les mots qui sont effectivement représentés dans le corpus : la moitié des couples présents sont trouvés.

La contrainte des quatre caractères initiaux élimine elle aussi des couples qui auraient été fortement associés. Il faudra examiner la quantité de bruit apportée par une descente à trois caractères, mais aussi rechercher des méthodes prenant plus largement en compte les autres lettres des couples impliqués (par exemple, *poumon* / *pulmonaire*). Les méthodes employées pour l'identification de cognats, par exemple (Kraif, 2001), pourraient trouver ici une application supplémentaires. Pour compléter les principales causes de silence, près de 10 % des couples recherchés (mots en *a*) étaient présents dans le corpus, mais pas à une distance inférieure ou égale à 100 mots pleins. Ceux-ci pourraient être détectés par application des règles acquises sur le corpus ou sur des terminologies (c'est le cas des quatre couples en *a*).

Certains noms donnent naissance à plusieurs adjectifs relationnels concurrents (*pharynx* / *pharyngien*, *pharyngé*). En cherchant à isoler le meilleur candidat, notre méthode passe ces cas sous silence. Il faudrait pour l'éviter mettre en place un procédé permettant de montrer à la personne qui va valider ces dérivations les autres propositions du système.

En parallèle avec l'amélioration de la méthode, il reste à tester son rappel sur d'autres ensembles de noms : autres axes de la SNOMED, noms tirés d'autres terminologies médicales, et plus généralement les noms présents dans les corpus étudiés ; et bien sûr sur d'autres types de dérivations. D'autre part, sa pertinence sur des corpus non spécialisés (par exemple, les habituels corpus journalistiques) sera aussi à examiner.

Enfin, le travail présenté ici n'explore qu'une partie de l'arsenal possible pour l'acquisition morphologique. La combinaison à d'autres méthodes parmi celles mentionnées en début d'article devrait contribuer à un recensement accéléré des dérivés de la langue médicale (Zweigenbaum *et al.*, 2003), et pourquoi pas de ceux de la langue générale (Hathout *et al.*, 2002).

Remerciements

Nous remercions le Dr. Roger A. Côté de nous avoir gracieusement prêté une copie pré-commerciale de la version française du Répertoire d'anatomopathologie de la nomenclature SNOMED Internationale, et l'équipe CISMeF du CHU de Rouen pour son Catalogue et Index des Sites Médicaux Francophones, ressource précieuse pour le traitement automatique des langues.

Références

BURNAGE G. (1990). *CELEX - A Guide for Users*. Nijmegen: Centre for Lexical Information, University of Nijmegen.

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

DAILLE B. (1999). Identification des adjectifs relationnels en corpus. In P. AMSILI, Ed., *Actes de TALN 1999 (Traitement automatique des langues naturelles)*, p. 105–114, Cargèse: ATALA.

DAL G., NAMER F. & HATHOUT N. (1999). Construire un lexique dérivationnel : théorie et réalisations. In P. AMSILI, Ed., *Actes de TALN 1999 (Traitement automatique des langues naturelles)*, Cargèse:

ATALA.

DARMONI S. J., LEROY J.-P., THIRION B., BAUDIC F., DOUYERE M. & PIOT J. (2000). CISMeF: a structured health resource guide. *Methods of Information in Medicine*, **39**(1), 30–35.

GAUSSIER E. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. In A. KEHLER & A. STOLCKE, Eds., *ACL workshop on Unsupervised Methods in Natural Language Learning*, College Park, Md.

GRABAR N. & ZWEIGENBAUM P. (1999). Acquisition automatique de connaissances morphologiques sur le vocabulaire médical. In P. AMSILI, Ed., *Actes de TALN 1999 (Traitement automatique des langues naturelles)*, p. 175–184, Cargèse: ATALA.

GRABAR N. & ZWEIGENBAUM P. (2000). Automatic acquisition of domain-specific morphological resources from thesauri. In *Proceedings of RIAO 2000: Content-Based Multimedia Information Access*, p. 765–784, Paris, France: C.I.D.

GREFENSTETTE G. & NIOCHE J. (2000). Estimation of English and non-English language use on the WWW. In *Proceedings of RIAO 2000: Content-Based Multimedia Information Access*, p. 237–246, Paris, France: C.I.D.

HADOUCHE F. (2002). *Acquisition de resources morphologiques à partir de corpus*. DESS d'ingénierie multilingue, Institut National des Langues et Civilisations Orientales, Paris.

HATHOUT N., NAMER F. & DAL G. (2002). An experimental constructional database: the MorTAL project. In P. BOUCHER, Ed., *Many morphologies*, p. 178–209. Somerville, MA: Cascadilla Press.

JACQUEMIN C. (1997). Guessing morphology from terms and corpora. In *Proc 20th ACM SIGIR*, p. 156–167, Philadelphia, PA.

KRAIF O. (2001). Exploitation des cognats pour l'alignement : architecture et évaluation. *Traitement automatique des langues*, **42**(3).

MANNING C. D. & SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

MCCRAY A. T., SRINIVASAN S. & BROWNE A. C. (1994). Lexical methods for managing variation in biomedical terminologies. In *Proc 18th Annu Symp Comput Appl Med Care*, p. 235–239, Washington: Mc Graw Hill.

NAMER F. (2000). FLEMM : un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues*, **41**(2), 523–547.

NAMER F. (2002). Acquisition automatique de sens à partir d'opérations morphologiques en français : études de cas. In J.-M. PIERREL, Ed., *Actes de TALN 2002 (Traitement automatique des langues naturelles)*, Nancy: ATALA ATILF.

PORTER M. F. (1980). An algorithm for suffix stripping. *Program*, **14**, 130–137.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, p. 44–49, Manchester, UK.

TANGUY L. & HATHOUT N. (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web. In J.-M. PIERREL, Ed., *Actes de TALN 2002 (Traitement automatique des langues naturelles)*, p. 245–254, Nancy: ATALA ATILF.

XU J. & CROFT B. W. (1998). Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems*, **16**(1), 61–81.

ZWEIGENBAUM P., BAUD R., BURGUN A., NAMER F., JARROUSSE E., GRABAR N., RUCH P., LE DUFF F., THIRION B. & DARMONI S. (2003). Towards a unified medical lexicon for French. In R. BAUD, M. FIESCHI, P. LE BEUX & P. RUCH, Eds., *Actes Medical Informatics Europe*, p. 415–420, Amsterdam: IOS Press.