Utilisation de corpus de spécialité pour le filtrage de synonymes de la langue générale*

Natalia Grabar[1, 2], Pierre Zweigenbaum[1, 2]
(1) INSERM, U729, 75006 Paris ;
(2) INALCO, CRIM, 75343 Paris Cedex 07 ;
(3) Assistance Publique - Hôpitaux de Paris, STIM/DSI, 75674 Paris Cedex 14
{ngr,pz}@biomath.jussieu.fr

Mots-clefs : Langue de spécialité, langue générale, structuration de terminologies, synonymes, portabilité, filtrage

Keywords: Specialized language, general language, terminology structuring, synonyms, portability, filtering

Résumé Les ressources linguistiques les plus facilement disponibles en TAL ressortissent généralement au registre général d'une langue. Lorsqu'elles doivent être utilisées sur des textes de spécialité il peut être utile de les adapter à ces textes. Cet article est consacré à l'adaptation de ressources synonymiques générales à la langue médicale. L'adaptation est obtenue suite à une série de filtrages sur un corpus du domaine. Les synonymes originaux et les synonymes filtrés sont ensuite utilisés comme une des ressources pour la normalisation de variantes de termes dans une tâche de structuration de terminologie. Leurs apports respectifs sont évalués par rapport à la structure terminologique de référence. Cette évaluation montre que les résultats sont globalement encourageants après les filtrages, pour une tâche comme la structuration de terminologies : une amélioration de la précision contre une légère diminution du rappel.

Abstract General language resources are often more easily available for NLP applications. When using them to process specialized texts it might be useful to adapt them to these texts. This paper describes experiments in adapting general language synonymous resources to the medical domain. A set of filtering methods through a domain corpora is applied. Original and filtered synonyms are then used for normalizing term variation in a terminology structuring task. Their relative contributions are evaluated in comparison with the original structure of the reference terminology. This evaluation shows that the overall results are encouraging, as for the terminology structuring task : improvement of precision while recall is slightly decreased.

*Nos expériences en structuration de terminologies ont été présentées dans (Grabar & Zweigenbaum, 2004). Cet article est plus spécifiquement consacré à l'adaptation de ressources linguistiques générales aux textes de spécialité et à l'influence de cette adaptation sur les résultats.

1 Introduction

Les ressources linguistiques les plus facilement disponibles en TAL ressortissent généralement au registre général d'une langue : lexiques flexionnels, dictionnaires généraux, synonymes, etc. Par définition, la contrainte de spécialisation domaniale est absente de ces ressources. Pour obtenir de meilleurs résultats lors de leur utilisation dans un domaine de spécialité, une adaptation à ce domaine peut être utile. Le but de ce travail consiste ainsi à adapter un ensemble de synonymes généraux au domaine médical grâce aux filtrages effectués sur un corpus médical. Nous utilisons les synonymes originaux et filtrés, à côté d'autres ressources et traitements, pour la normalisation de variantes de termes médicaux. Ces normalisations s'avèrent importantes dans de nombreuses applications (indexation et recherche d'information, questions-réponses, codage dans des terminologies contrôlées, acquisition terminologique, traduction, etc.). Avec des objectifs similaires, (Jacquemin, 1999) applique des règles de transformation morpho-syntaxique pour apparier les termes d'un corpus avec une terminologie contrôlée. Dans le domaine médical, (McCray *et al.*, 1994) exploitent en plus d'autres niveaux de normalisation pour mettre en correspondance des termes provenant de différentes terminologies. En structuration de terminologies, (Hamon *et al.*, 1998) utilisent des synonymes simples pour la détection de liens de synonymie entre termes complexes, grâce à la compositionnalité sémantique. Dans les expériences présentées ici, nous effectuons différents types de normalisation de termes (traitements au niveau de caractères et de l'ordre de mots, ressources morphologiques et synonymiques) : des termes potentiellement proches par leur sens peuvent ainsi être appariés. Nous appliquons ces différentes normalisations en structuration de terminologies, à travers l'hypothèse d'inclusion lexicale qui nous permet d'induire des relations hyperonymiques entre termes.

Dans la section 2, nous présentons les synonymes généraux provenant du Petit Robert ; dans la section 3, les méthodes pour le filtrage de ces synonymes et pour leur évaluation. L'évaluation est faite au sein d'une tâche de structuration de terminologie, où la structure induite est comparée avec la structure originale de ces mêmes termes. Dans la section 4, nous présentons et analysons les résultats. Nous terminons avec une conclusion et des perspectives (sec. 5).

2 Synonymes de la langue générale : Le Robert

L'innovation du dictionnaire Le Robert a été le dépassement de l'organisation alphabétique des lexèmes grâce à l'ajout des rapports analogiques établis sur la base des étymologies, définitions, enchaînements syntaxiques, liens de synonymie et d'antonymie (Robert, 1967). L'ensemble de ces rapports permet aux usagers de saisir plus aisément le sens des lexèmes. Il a fourni également les séries de synonymes que nous utilisons : plus de 140 000 paires de synonymes simples, comme par exemple {*culot*, *fond*}. Elles semblent correspondre à une édition des années 70 de ce dictionnaire[1]. Plusieurs questions peuvent se poser lors de l'utilisation de ressources synonymiques en TAL. Nous discutons ici de leur symétrie, transitivité et spécialisation.

La synonymie, reliant des lexèmes ou expressions contextuellement interchangeables (Cruse, 1986, p. 88), est souvent considérée comme une relation symétrique. Les ressources synonymiques du Robert se présentent sous forme {*entrée dictionnairique*, *famille de synonymes*} ou {*entrée dictionnairique*, *synonyme*}. Nous considérons alors l'*entrée dictionnairique* comme

[1]Nous remercions l'INaLF et Didier Bourigault d'avoir rendu ces ressources disponibles, Thierry Hamon de nous les avoir fournies nettoyées et formatées et Jean-Luc Manguin de nous avoir communiqué leur datation.

canon vers lequel sa famille ou ses synonymes peuvent être « normalisés ». Par contre, nous n'autorisons pas la normalisation dans le sens opposé, ni donc la symétrie. Cela semble raisonnable, comme le montre l'exemple de la famille *boulimie* :

> *boulimie* : *cynorexie*, *hyperorexie*, *hyperphagie*, *sitiomanie*, *faimcalle*, *appétit*, *avidité*,

qui reçoit, si nous considérons la synonymie comme une relation symétrique, *faim*, *fringale* et *frénésie*. De la même manière, *ventre* passe de 11 synonymes à 19 et *trouble* de 22 à 64.

Lors de l'utilisation de synonymes, pour ne pas générer trop de bruit, nous ne calculons pas de fermeture transitive complète. Par contre, nous autorisons une transitivité « locale » : au sein d'une famille, deux synonymes (comme *cynorexie* et *hyperorexie*) peuvent être normalisés vers leur entrée (*boulimie*), et donc considérés eux-mêmes comme synonymes.

Et enfin, la spécialisation de ces ressources. D'une part, les sens spécifiques peuvent manquer, ce qui peut être cause de silence. Par exemple, la famille *culot* :

> *culot* : *fond*, *dépôt*, *résidu*, *benjamin*, *aplomb*, *assurance*, *audace*, *effronterie*, *toupet*, *estomac*.

ne contient pas l'acception médicale (« *Amas d'érythrocytes tassés au fond du récipient de conservation après centrifugation du plasma sanguin* ») proche de *transfusion*. D'autre part, le mélange de registres et donc l'instabilité sémantique des lexèmes, surtout dans un contexte de spécialité, peuvent être cause de bruit. Rappelons que les rapports analogiques du Robert peuvent correspondre en réalité à plusieurs relations, généralement non distinguées (synonymes, analogies, thèmes d'expressions, variantes orthographiques, sous-entrées, hyperonymes, etc.) et mettent ensemble des données fondamentalement hétérogènes (Marcus, 2003). Pour toutes ces raisons, il peut être utile de filtrer les synonymes de la langue générale, surtout lorsqu'ils doivent être utilisés dans des traitements automatiques appliqués à un domaine de spécialité.

3 Méthodes

3.1 Méthode d'adaptation des synonymes généraux

Pour l'adaptation de synonymes généraux aux textes de spécialité, nous utilisons un corpus médical d'environ 8,5 millions d'occurrences. Ce corpus contient des documents hospitaliers (lettres, comptes rendus hospitaliers) et des documents Web collectés à travers le portail médical CISMeF (Darmoni *et al.*, 2001)[2]. Nous supposons ainsi que les synonymes les plus pertinents doivent appartenir au même registre. Les tests utilisés se basent sur le fait que ces synonymes apparaissent dans les mêmes textes et « pas trop loin » (*cf.* Recherche d'associations), et plus spécifiquement qu'il existe des constructions syntaxiques qui sont des indices forts de synonymie (*cf.* Repérage d'indices de synonymie).

Recherche d'associations. La recherche de mots associés (cooccurrences, collocations, etc., voir (Manning & Schütze, 1999)) est une méthode courante en TAL « à base de corpus ». Nous utilisons la mesure du « rapport de vraisemblance » (*log likelihood ratio*) comme dans (Zweigenbaum *et al.*, 2003), dont nous avons repris et adapté les programmes. Nous recherchons des

[2] `http://www.chu-rouen.fr/cismef/`

paires de synonymes qui cooccurrent plus souvent que le hasard dans une fenêtre de 2*150 mots *pleins*. Cela permet de confirmer des paires de synonymes comme :

{*abcès*, *phlegmon*}, {*biopsie*, *ponction*} et {*dernier*, *culot*}, {*signal*, *appel*}.

Avec cette approche de filtrage, une première sélection est faite à travers le corpus : les synonymes doivent apparaître dans la fenêtre de mots fixée. Une deuxième sélection est faite grâce au classement : les associations de synonymes les plus stables ont un meilleur classement.

Repérage d'indices de synonymie en corpus. Les patrons lexico-syntaxiques (Séguéla & Aussenac-Gilles, 1999) et les marqueurs de coordination (Lame, 2002, sec.6.1) sont utilisés par les auteurs pour la structuration de termes. Ils sont projetés sur les corpus et permettent de mettre au jour des relations sémantiques recherchées entre les termes `X` et `Y` :

« `X appelé Y` », « `X est défini comme 1-MOT Y` »,
« `X est confondu avec Y` », « `X n'est autre que 1-MOT Y` »,
« `X ou Y` », « `X ni Y` », « `X et Y` », « `pas de X, pas de Y` ».

La projection de patrons de synonymie et marqueurs de coordination sur notre corpus permet de valider des paires de synonymes comme :

- {*gonflement*, *œdème*} : « *L'œdème* `est défini comme` *un gonflement palpable produit par l'expansion du volume interstitiel liquidien.* »
- {*rhinopharynx*, *cavum*} : « *Le rhinopharynx* `appelé` *cavum est situé sous la base du crâne, en arrière des fosses nasales, au-dessus de l'oropharynx et en avant des 2 premières vertébres cervicales.* »
- {*syndrome*, *affection*} : « *Cette affection* `est appelée` *le syndrome hépato-rénal qui est défini comme une augmentation progressive de la créatinine plasmatique, sans cause évidente autre chez un patient atteint de maladie hépatique avancée.* »
- {*repos*, *sommeil*} : « *Trop souvent, repos* `est confondu avec` *sommeil et activité avec éveil.* »
- {*bruit*, *souffle*} : « *Examen cardiaque : bruits bien frappés aux 4 foyers sans souffle* `ni` *bruit surajoutés.* »
- {*orthopnée*, *dyspnée*} : « *Examen cardio-vasculaire :* `pas` *de dyspnée,* `pas` *d'orthopnée, présence d'œdèmes des membres inférieurs avec un godet positif.* »

3.2 Méthodes d'évaluation du filtrage des synonymes

L'évaluation des synonymes originaux et des synonymes filtrés est faite à travers une tâche de structuration de termes (Grabar & Zweigenbaum, 2004). Les synonymes, à côté d'autres traitements et ressources, sont utilisés pour la normalisation de la variation des termes du thesaurus MeSH (NLM, 2001)[3]. Nous recourons à ces normalisations en effectuant des calculs d'inclusions lexicales pour la détection de relations hiérarchiques. Nous utilisons ensuite la structure originale du MeSH comme référence pour évaluer le rappel et la précision des relations induites.

Détection d'inclusions lexicales. L'inclusion lexicale est naturellement utilisée dans la formation de termes (Kleiber & Tamba, 1990) :

acides gras / *acides gras indispensables*.

Nous exploitons ce fait pour la détection de relations hyperonymiques entre termes. Pour ceci, nous vérifions si tous les mots d'un terme sont inclus dans un autre terme. Si c'est le cas, le terme inclus est supposé être le père hiérarchique P, et le terme incluant, son fils F. Nous effectuons les tests sur les termes segmentés en mots, d'abord sur leurs formes brutes et ensuite avec une série de normalisations :

[3] `http://www.nlm.nih.gov/mesh/meshhome.html`

- normalisation de base : conversion en caractères minuscules, suppression des accents, de la ponctuation, des nombres et des mots « vides » ;
- normalisations avec des ressources morphologiques flexionnelles de la langue médicale et générale, des ressources dérivationnelles et allomorphiques ;
- normalisations avec des synonymes qui sont ceux de la langue médicale, soit ceux de la langue générale, soit les synonymes de la langue générale filtrés sur les corpus médicaux.

Les ressources linguistiques utilisées se présentent sous forme de paires de mots {*canon*,*forme*}. Lors des traitements, si un mot d'un terme correspond à une *forme*, il est normalisé en son *canon* correspondant.

Nous effectuons la mise en minuscules et la suppression d'accents parce que dans la version 2001 du MeSH les termes étaient encore écrits en majuscules non accentuées (*EPITHELIOMA SQUIRRHEUX*), tandis que nos ressources linguistiques sont en minuscules accentuées {*carcinome*, *épithélioma*}. À côté de cela, les termes du MeSH, qui correspondent à un langage d'indexation artificiel, peuvent comporter des virgules (*VIRUS A ADN, INFECTIONS*) ou faire omission des mots grammaticaux (*LESION REPERFUSION MYOCARDIQUE*), ce qui explique que nous appliquons une suppression automatique de la ponctuation, des mots « vides » et que nous ignorons l'ordre des mots. L'abstraction de l'ordre des mots et des mots « vides » est aussi utile lorsque les termes à apparier comportent des dérivations (*sténose de l'aorte* et *aorte sténosée*, *TRANSPLANTATION CARDIAQUE* et *TRANSPLANTATION COEUR-POUMON*). Notons également que nous n'effectuons pas d'analyse syntaxique. L'ensemble de nos normalisations peut ainsi conduire vers des appariements non pertinents (*AGE DENTAIRE / SOINS DENTAIRES SUJET AGE*). On peut supposer qu'une analyse syntaxique des dépendances dans les termes permettrait d'en éliminer un certain nombre.

Évaluation par rapport au référentiel existant. Pour évaluer les relations induites, nous les comparons avec les relations qui existent dans la structure originale du thesaurus MeSH. Nous cherchons ainsi à savoir si les relations hiérarchiques du MeSH pouvent être induites avec l'hypothèse d'inclusion lexicale. Nous calculons alors le rappel et la précision (r_{MeSH} est le nombre de relations dans le MeSH (95 815), r_e le nombre de relations induites existant dans le MeSH, et r_n le nombre de relations induites hors-MeSH) : $R = \frac{r_e}{r_{MeSH}}$; $P = \frac{r_e}{r_e + r_n}$

4 Apport du filtrage des synonymes : résultats et analyse

Nous avons mis en œuvre les méthodes de filtrage présentées en 3.1 aux synonymes du Robert (sec. 2) à l'aide du corpus médical. Les synonymes originaux et filtrés ont été utilisés dans la tâche de structuration de terminologies (sec. 3.2). Nous présentons ici les résultats du filtrage des synonymes et leur impact relatif sur la tâche de structuration : évolution quantitative des relations, utilisation effective des ressources linguistiques, évaluation de ces relations par rapport à la structure originale du MeSH et analyse de quelques relations.

Filtrage des synonymes. Le calcul d'associations permet de valider le nombre le plus élevé de synonymes : 15 589 paires en gardant 60 % des meilleures associations ; les marqueurs de coordination en valident 1 736 et les patrons lexico-syntaxiques 46 paires. Le filtrage complet (union) nous donne un ensemble de 16 154 paires de synonymes, soit une réduction de 88,5 % par rapport aux 140 141 paires d'origine. Il est intéressant de remarquer qu'aucune instanciation

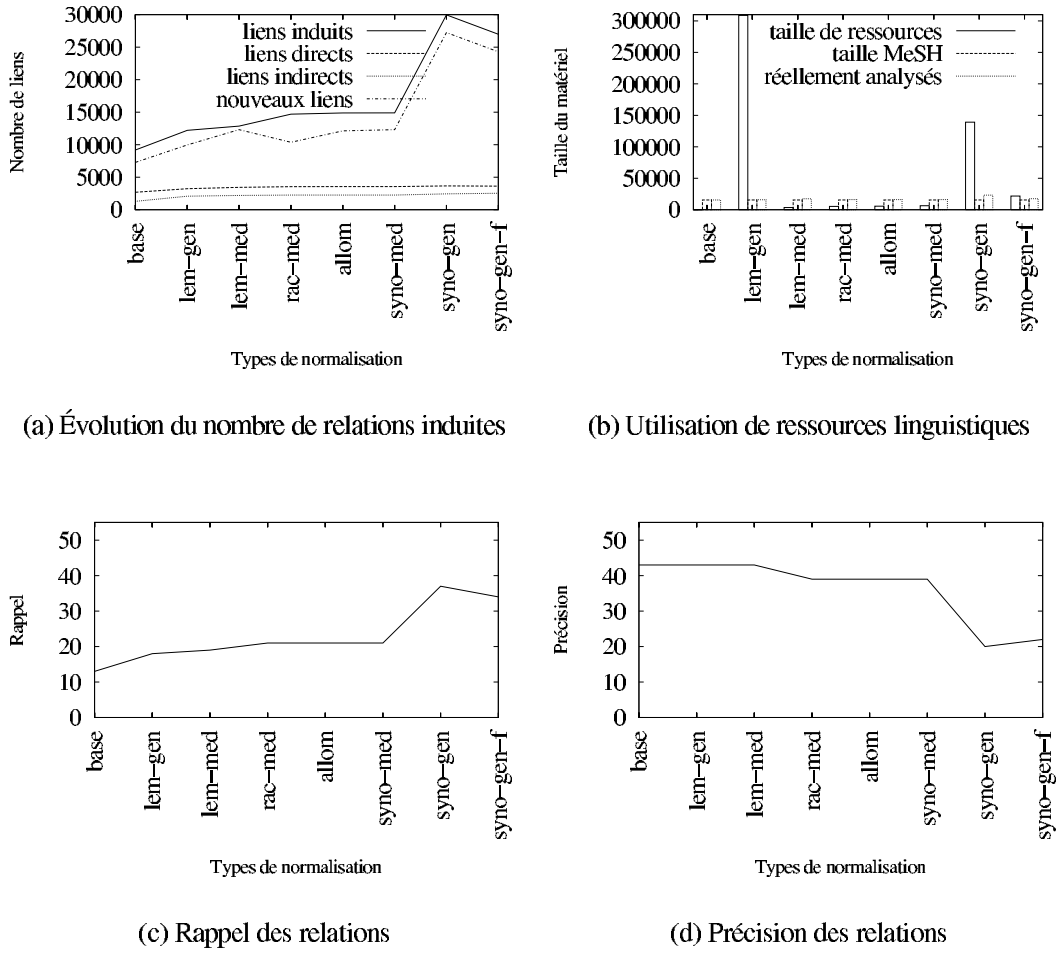
de patrons de (Séguéla & Aussenac-Gilles, 1999) n'a été relevée dans la partie *documents hospitaliers* du corpus. Cela semble raisonnable : ces documents s'adressent à des spécialistes avec, comme but principal, la transmission d'informations sur les patients. Par contre dans les documents du Web, destinés souvent à un public plutôt non averti, les reformulations et le recours à la synonymie sont fréquents. Il apparaît en outre que les marqueurs de coordination relient souvent non des synonymes mais des co-hyponymes ({*bruit*, *souffle*}, {*orthopnée*, *dyspnée*} dans les exemples de la sec. 3.1). Ce caractère des marqueurs déjà noté dans (Pearson, 1998) est accentué ici par la nature hétérogène des relations qui constituent les rapports analogiques dans Le Robert. Notons que l'ensemble de filtrages utilisé exploite les relations syntagmatiques entre les mots. Il serait intéressant d'étudier en plus la piste paradigmatique, par exemple les calculs distributionnels (Nazarenko *et al.*, 2001).

Évolution quantitative des relations. Le calcul d'inclusion lexicale a été appliqué à une liste « à plat » de 19 638 termes du MeSH (écrits à l'époque en majuscules non accentuées). La figure 1(a) montre le nombre de relations induites à chaque étape des normalisations : normalisation des caractères et suppression des mots vides (*base*), application des ressources flexionnelles de la langue générale (*lem-gen*) et de la langue médicale (*lem-med*), des ressources dérivationnelles (*rac-med*) et allomorphiques (*allom*), des synonymes de la langue médicale (*syno-med*), de la langue générale (*syno-gen*) et des synonymes de la langue générale filtrés (*syno-gen-f*). Chaque ressource synonymique est utilisée alternativement, en plus des normalisations antérieures (*base*, *lem-med*, *rac-med* et *allom*). Nous analysons ici les deux dernières étapes : *syno-gen* (synonymes de la langue générale) et *syno-gen-f* (synonymes de la langue générale filtrés). L'analyse de l'impact des synonymes médicaux *syno-med* constitue une perspective. La figure 1(a) décompose le nombre total des relations induites en relations correctes (directes et indirectes du MeSH[4]) et les nouvelles relations (hors-MeSH). Les ressources *syno-gen*, comportant un nombre très élevé de paires de synonymes, doublent le volume des relations induites à l'étape précédente : de 14 884 relations *allom* nous passons à 29 969 avec *syno-gen*, ce qui fait 15 085 relations en plus. Avec les synonymes généraux filtrés le volume est un peu moins important : 26 986 relations (12 102 de plus qu'avec *allom*).

Utilisation effective des ressources linguistiques. La figure 1(b) montre l'utilisation effective de ressources linguistiques. Pour chaque normalisation, nous indiquons le nombre de mots dans la ressource, le nombre de mots dans les termes du MeSH (15 446) et le nombre de mots réellement traités. Nous pouvons ainsi voir que la totalité des ressources est rarement utilisée, surtout avec les ressources de la langue générale (*lem-gen* et *syno-gen*). Cela ne ralentit pas les calculs car seules les ressources pertinentes sont chargées. Par contre, cela montre *a posteriori* que le recouvrement entre les deux ensembles (les données à traiter et les ressources) est faible. Il serait intéressant de pouvoir évaluer ce recouvrement avant les traitements et choisir les ressources qui conviennent le mieux, comme proposent de le faire (Ninova *et al.*, 2005).

Évaluation des relations par rapport à la structure du MeSH. La comparaison des relations induites avec la structure originale du MeSH permet d'évaluer les rappel et précision (fig. 1(c) et 1(d)). Nous pouvons ainsi voir que le rappel augmente de manière générale avec l'injection de connaissances linguistiques supplémentaires. De 13,7 % à l'étape *base* et 21,6 %

[4]Les relations directes existent entre un père et un fils hiérarchiques. Les relations indirectes correspondent à toute relation d'un terme vers l'un de ses ancêtres.



(a) Évolution du nombre de relations induites

(b) Utilisation de ressources linguistiques

(c) Rappel des relations

(d) Précision des relations

FIG. 1 – Illustration des différentes étapes du calcul des inclusions lexicales avec et sans le filtrage des synonymes de la langue générale.

à l'étape *allom* il atteint un sommet avec les synonymes de la langue générale (37,6 %), ces derniers présentant beaucoup de candidats synonymes. Suite aux filtrages, le rappel diminue : il perd 3 % et se fixe à 34,6 %. Face à ces valeurs de rappel très faibles, n'oublions pas que nous testons une seule approche de structuration de termes à travers les relations hyponymiques, basée sur un type d'indices lexicaux (inclusions lexicales). Sa combinaison avec d'autres approches devrait donner une structuration plus complète (Kavanagh, 1995). Comme c'est souvent le cas, l'évolution de la précision est opposée : l'injection de connaissances supplémentaires apporte plus de « risque » dans la génération de relations incorrectes. Par rapport à la configuration de *base* (43,3 %) et d'*allom* (39 %), la précision est de 20,4 % avec *syno-gen* et 22,9 % avec *syno-gen-f*, avec une amélioration de 2,5 % suite aux filtrages. L'apport relatif des ressources synonymiques adaptées ou non au domaine semble alors être similaire à celui des ressources spécialisées ou générales (voir par exemple (Hamon *et al.*, 1998)) : les ressources générales prises dans leur ensemble augmentent le rappel, tandis que leur adaptation augmente la précision. Le choix de filtrer ou non doit donc être fait selon qu'on privilégie le rappel ou la précision. Dans une tâche comme la structuration de terminologies, qui demande une intervention humaine lors de la validation des inductions automatiques, il peut ainsi être plus utile d'avoir une meilleure précision. Une autre piste pour l'amélioration de la précision est liée sans doute à

	Relation hyponymique	Canon	Synonyme(s)
Liens directs	{*adénocarcinome*, *épithélioma squirrheux*}	*carcinome*	*adénocarcinome*, *épithélioma*
	{*céramiques*, *porcelaine dentaire*}	*céramique*	*porcelaine*
	{*famille*, *filiation illégitime*}	*filiation*	*famille*
	{*gravitation*, *modification pesanteur*}	*attraction*	*gravitation*, *pesanteur*
	{*immunisation*, *rappel vaccination*}	*immunité*	*immunisation*, *vaccination*
	{*mouvement*, *activité motrice*}	*vie*	*mouvement*, *activité*
	{*rhinite*, *coryza spasmodique*}	*rhume*	*rhinite*, *coryza*
	{*thérapeutique*, *traitement combiné*}	*thérapeutique*	*traitement*
Liens indirects	{*anesthésie*, *hypnose dentisterie*}	*narcose*	*anesthésie*, *hypnose*
	{*assainissement*, *drainage sanitaire*}	*assèchement*	*assainissement*, *drainage*
	{*calculs*, *lithiase rénale*}	*calcul*	*lithiase*
	{*épithélioma*, *carcinome lobulaire*}	*carcinome*	*épithélioma*
	{*personnalité*, *concept soi*}	*personnalité*	*soi*
	{*thérapeutique*, *traitement par art*}	*thérapeutique*	*traitement*
	{*vascularite*, *artérite temporale*}	*angéite*	*vascularite*, *artérite*

TAB. 1 – Exemples de relations directes et indirectes du MeSH perdues suite aux filtrages.

notre décision d'autoriser la transitivité « locale » des synonymes au sein d'une famille, comme le montre l'exemple {*abcès*, *volume sanguin*} dans la suite de cette section. Ce fait est amplifié d'une part parce que les termes du MeSH ont pour vocation de couvrir tout le domaine médical. La cohésion sémantique de l'ensemble de ces termes est donc moins importante que ce qu'elle pourrait être dans un texte. D'autre part, nous ne vérifions pas l'existence de termes intermédiaires qui servent à l'appariement (*grosseur* et *grosseur sanguine* pour l'exemple ci-dessus). Notons que lors de la détection de relations de synonymie entre termes complexes, (Hamon *et al.*, 1998) bloquent la transitivité « locale ».

De manière générale, nous constatons que ce sont les synonymes généraux qui présentent une rupture avec l'évolution des courbes du rappel et de la précision par rapport aux étapes précédentes. Il serait ainsi intéressant d'analyser de plus près ce que nous obtenons avec l'application de la transitivité « locale » et de compléter ces comparaisons en analysant également l'apport relatif des synonymes généraux (filtrés ou originaux) et des synonymes spécifiques au domaine.

Analyse de relations « filtrées ». Nous examinons ici des relations que nous n'induisons plus suite aux filtrages des synonymes. Elles correspondent aux relations correctes du MeSH et à des relations hors-MeSH, donc potentiellement incorrectes.

Dans le tableau 1, nous présentons des relations correctes du MeSH, directes et indirectes. La première colonne contient ces relations, la deuxième contient le *canon* et la troisième les synonymes normalisés vers le *canon* lors des traitements. Nous induisons 69 relations directes du MeSH et 106 indirectes en moins. Ainsi dans le premier exemple, la famille de *carcinome*, qui avait deux éléments à l'origine (*adénocarcinome*, *épithélioma*), n'en garde plus que le premier. Ce qui empêche d'induire la relation {*adénocarcinome*, *épithélioma squirrheux*}, pourtant correcte, après les filtrages. Dans le deuxième exemple {*céramiques*, *porcelaine dentaire*}, les termes sont appariés à travers une lemmatisation {*céramiques*, *céramique*} et une relation de synonymie {*porcelaine*, *céramique*}.

Nous avons analysé environ 5 % des 3 000 relations hors-MeSH que nous n'induisons plus. Sur cet ensemble, les filtrages semblent montrer pour la plupart leur efficacité. Par exemple, la famille *grosseur*, de 19 éléments à l'origine, parmi lesquels *abcès*, *volume* et *obésité*, permet d'induire les paires comme {*abcès*, *volume sanguin*} et {*obésité*, *volume sanguin*}, manifestement erronées. Cette famille étant réduite lors des filtrages à trois éléments, ces paires ne sont plus générées. De la même manière, les paires comme {*absorption*, *sidération myocarde*}, {*acétylène*, *infusion goudron*}, {*autoritarisme*, *gouvernement États Unis*}, {*crime*, *faute professionnelle*} ou {*émeute*, *lutte contre moustique*} n'apparaissent plus suite à la réduction des familles *anéantissement* (46 éléments à l'origine), *combustible* (27), *gouvernement* (38), *crime* (12) et *révolte* (16), respectivement.

Dans (Grabar & Zweigenbaum, 2004), nous soulignons que des relations hors-MeSH peuvent souvent être pertinentes, par exemple la paire hors-MeSH {*adénocarcinome*, *épithélioma mixte*} perdue suite à la réduction de la famille *carcinome*. Bien qu'étant hors-MeSH, cette paire est probablement correcte car elle suit le même modèle que la paire directe du MeSH {*adénocarcinome*, *épithélioma squirrheux*} (tab. 1). D'autres paires, comme {*agraphie*, *alexie pure*} ou {*agraphie*, *aphasie anomique*} (famille *aphasie* : *agraphie*, *alexie*), sont aussi des relations potentiellement pertinentes, mais non hiérarchiques.

Parmi les relations hors-MeSH analysées ici nous avons beaucoup d'erreurs, mais aussi des paires potentiellement pertinentes, qu'elles correspondent aux relations hyperonymiques ou non. La décision de les encoder dans une terminologie relève d'une décision humaine, que les approches automatiques ne peuvent pas induire.

5 Conclusion et perspectives

Les expériences et analyses présentées avaient pour but de montrer que lorsque des ressources spécifiques à un domaine de spécialité ne sont pas disponibles il est possible de recourir aux ressources de la langue générale. Pour que leur utilisation soit plus bénéfique, il peut être utile de les adapter à ce domaine. Nous avons ainsi testé les ressources synonymiques de la langue générale (Le Petit Robert) sur des données provenant du domaine médical. Ces ressources ont été utilisées en structuration de terminologie à côté d'autres ressources et traitements pour la normalisation des variantes de termes. L'intérêt de l'utilisation des synonymes lors des normalisations est dû au fait qu'ils permettent d'accéder à des variations qui ne sont pas accessibles avec d'autres ressources. L'adaptation est effectuée à travers des filtrages sur des corpus médicaux. Trois approches sont utilisées : recherche d'associations, marqueurs de coordination et patrons lexico-syntaxiques. L'ensemble des filtrages permet de réduire les synonymes originaux d'environ 90 %. La comparaison de l'apport de ressources synonymiques filtrées et originales montre une amélioration de la précision de 2,5 % et une perte du rappel de 3 %, ce qui nous semble bénéfique pour une tâche comme la structuration de terminologie où le bruit est important.

Parmi les perspectives de ce travail, notons essentiellement l'exploration de la piste paradigmatique, par exemple distributionnelle (Nazarenko *et al.*, 2001), pour le filtrage de synonymes ; une analyse plus poussée de l'impact de la transitivité « locale » ; la comparaison de l'apport relatif des synonymes généraux et spécialisés ; et l'application d'une analyse syntaxique des dépendances dans les termes. Par ailleurs, il serait intéressant d'observer l'influence de l'adaptation de synonymes, et de ressources linguistiques en général, dans d'autres contextes applicatifs.

Références

CRUSE D. A. (1986). *Lexical Semantics*. Cambridge : Cambridge University Press.

DARMONI S. J., THIRION B., LEROY J.-P., DOUYÈRE M., LACOSTE B., GODARD G., RIGOLLE I., BRISOU M., VIDEAU S., GOUPY E., PIOT J., QUÉRÉ M., OUAZIR S. & ABDULRAB H. (2001). A search tool based on 'encapsulated' MeSH thesaurus to retrieve quality health resources on the Internet. *MIIM*, **26**(3), 165–178.

GRABAR N. & ZWEIGENBAUM P. (2004). Lexically-based terminology structuring. In *Terminology*, volume 10, p. 23–54.

HAMON T., NAZARENKO A. & GROS C. (1998). A step towards the detection of semantic variants of terms in technical documents. In *International Conference on Computational Linguistics (COLING-ACL'98)*, p. 498–504, Université de Montréal, Montréal, Quebec, Canada.

JACQUEMIN C. (1999). Syntagmatic and paradigmatic representations of term variation. In *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, p. 341–348, University of Maryland.

KAVANAGH J. (1995). *The Text Analyser : A Tool for Extracting Knowledge From Text*. Master of computer science thesis, University of Ottawa, Ottawa, Canada.

KLEIBER G. & TAMBA I. (1990). L'hyperonymie revisitée : inclusion et hiérarchie. *Langages*, **98**, 7–32. L'hyponymie et l'hyperonymie (dir. Marie-Françoise Mortureux).

LAME G. (2002). *Construction d'ontologies à partir de textes. Une ontologie du droit dédié à la recherche d'information sur le Web*. Thèse de doctorat en informatique temps réel, robotique et automatique, École des Mines de Paris, Paris.

MANNING C. D. & SCHÜTZE H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA : MIT Press.

MARCUS A. (2003). *Dictionnaires électroniques et hypertextualité. Analyse critique des renvois doubles du Grand Robert*. Mémoire de DESS, CRIM/INaLCO. Sous la direction de David Piotrovsky.

MCCRAY A. T., SRINIVASAN S. & BROWNE A. C. (1994). Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the Annual SCAMC*, p. 235–239.

NAZARENKO A., ZWEIGENBAUM P., HABERT B. & BOUAUD J. (2001). Corpus-based extension of a terminological semantic lexicon. In *Recent Advances in Computational Terminology*, p. 327–351. John Benjamins.

NINOVA G., NAZARENKO A. & HAMON T. (2005). Comment mesurer la couverture d'une ressource terminologique pour un corpus ? In *TALN 2005*. À paraître.

NLM (2001). *Medical Subject Headings*. National Library of Medicine, Bethesda, Maryland. `http://www.nlm.nih.gov/mesh/meshhome.html`.

PEARSON J. (1998). *Terms in Context*, volume 1 of *Studies in Corpus Linguistics*. Amsterdam/Philadelphia : John Benjamins.

ROBERT P. (1967). *Préface du Petit Robert*. Paris : Dictionnaires Le Robert. Première édition.

SÉGUÉLA P. & AUSSENAC-GILLES N. (1999). Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In *Actes d'Ingénierie des Connaissances (IC)*, p. 79–88, Palaiseau, France.

ZWEIGENBAUM P., HADOUCHE F. & GRABAR N. (2003). Apprentissage de relations morphologiques en corpus. In B. DAILLE, Ed., *Actes de TALN 2003 (Traitement automatique des langues naturelles)*, p. 285–294, Batz-sur-mer : ATALA IRIN.