Groupement de termes basé sur des régularités linguistiques et sémantiques dans un contexte cross-langue

Marie Dupuch[1,2] Thierry Hamon[3] Natalia Grabar[1]
(1) CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France
mdupuch@objetdirect.com, natalia.grabar@univ-lille3.fr
(2) Viseo-Objet Direct, 4, avenue Doyen Louis Weil, 38000 Grenoble
(3) LIM&BIO (EA3969), Université Paris 13, Sorbonne Paris Cité,
74, rue Marcel Cachin, 93017 Bobigny Cedex France
thierry.hamon@univ-paris13.fr

RÉSUMÉ

Nous proposons d'exploiter des méthodes du Traitement Automatique de Langues dédiées à la structuration de terminologie indépendamment dans deux langues (anglais et français) et de fusionner ensuite les résultats obtenus dans chaque langue. Les termes sont groupés en clusters grâce aux relations générées. L'évaluation de ces relations est effectuée au travers de la comparaison des clusters avec des données de référence et la baseline, tandis que la complémentarité des relations est analysée au travers de leur implication dans la création de clusters de termes. Les résultats obtenus indiquent que : chaque langue contribue de manière équilibrée aux résultats, le nombre de relations hiérarchiques communes est plus grand que le nombre de relations synonymiques communes. Globalement, les résultats montrent que, dans un contexte cross-langue, chaque langue permet de détecter des régularités linguistiques et sémantiques complémentaires. L'union des résultats obtenus dans les deux langues améliore la qualité globale des clusters.

ABSTRACT

Grouping of terms based on linguistic and semantic regularities in a cross-lingual context

We propose to exploit the Natural Language Processing methods dedicated to terminology structuring independently in two languages (English and French) and then to merge the results obtained in each language. The terms are grouped into clusters thanks to the generated relations. The evaluation of the relations is done via the comparison of the clusters with the reference data and the baseline, while the complementarity of the relations is analyzed through their involvement in the clusters of terms. Our results indicate that : each language contributes almost equally to the generated results ; the number of common hierarchical relations is greater than the number of common synonym relations. On the whole, the obtained results point out that in a cross-language context, each language brings additional linguistic and semantic regularities. The union of the results obtained in each language improves the overall quality of the clusters.

MOTS-CLÉS : Relations sémantiques, termes, domaine de spécialité, médecine, contexte cross-langue.

KEYWORDS: Semantic relations, terms, specialized areas, medicine, cross-lingual context.

1 Introduction

Plusieurs travaux de recherche ont démontré qu'au travers des langues, il est possible de trouver des régularités linguistiques et sémantiques. De plus, ces régularités peuvent être renforcées dans un contexte cross-langue. Ce point peut être intéressant pour différentes applications du Traitement Automatique de Langues (TAL). L'analyse de travaux existants en TAL et en linguistique montre que le contexte cross-langue peut en effet être exploité de différentes manières :

– études comparatives, qui permettent de trouver des régularités et universaux interlangues. Ce type d'approche a été par exemple exploitée pour l'étude de la grammaticalisation (Willett, 1988), de la modalité (Diewald et Smirnova, 2010), des structures argumentatives (Li, 2011) ou stylistiques (Vinay et Darbelnet, 1958) ;
– études cross-langues contrastives, qui visent à faire des analyses comparatives entre les langues afin de relever des constantes aux langues comparées et des différences propres à chaque langue (Cartoni et Namer, 2012; Lefer et Grabar, 2013) ;
– transposition et adaptation de méthodes et ressources d'une langue vers une autre, qui visent à faire profiter une langue grâce aux travaux, méthodes et ressources déjà réalisés et éprouvés dans une autre langue (Farreres *et al.*, 1998; Huang *et al.*, 2002; Rodrigues *et al.*, 2006) ;
– collaboration entre les langues, qui vise à appliquer des méthodes ou ressources dans des langues différentes pour ensuite combiner les résultats. Ce type d'approches a été par exemple exploitée pour la désambiguïsation sémantique (Ceusters *et al.*, 2003; Banea *et al.*, 2011), l'indexation et recherche d'information (Schulz et Hahn, 2000; Malaisé *et al.*, 2007; Steinberger, 2011), et l'extraction d'information (Collier, 2011). Par ailleurs, la combinaison des résultats obtenus dans les langues différentes peut prendre différentes formes : un enrichissement mutuel afin d'obtenir des résultats plus exhaustifs, un système de vote ou de validation mutuelle afin d'obtenir des résultats plus précis, etc.

Nous proposons de travailler en mode collaboratif entre les langues et visons essentiellement l'amélioration de la complétude des résultats. L'hypothèse de notre travail est la suivante : le traitement du même matériel avec les mêmes méthodes dans deux langues (anglais et français), peut fournir des résultats différents et complémentaires, tandis que la combinaison de ces résultats peut améliorer les performances globales du système automatique.

Nous détectons les termes qui sont liés sémantiquement et les clusterisons. Nous travaillons avec les termes médicaux qui décrivent les effets indésirables dus à la prise de médicaments. La tâche visée dans notre travail est difficile, car il s'agit souvent de termes qui n'ont pas de similarité lexicale entre eux, comme *leucémie* (pathologie) et *ponction de moelle osseuse anormale* (résultats d'examen qui permet de la détecter) (Fleischman, 2001). Cependant, l'établissement de relations est très utile pour plusieurs applications, comme (1) la recherche et l'extraction d'information (Baeza-Yates et Ribeiro-Neto, 1999; Hahn *et al.*, 2001; Alfonseca *et al.*, 2002; Anizi et Dichy, 2009), où il est très utile de pouvoir détecter des contenus similaires afin d'augmenter le rappel des systèmes automatiques, (2) l'alignement de terminologies (Fridman Noy et Musen, 2000; Marko *et al.*, 2006), nombreuses dans le domaine médical et dont l'interopérabilité sémantique constitue un objectif très prisé dans le contexte clinique, (3) la fouille de bases de données de pharmacovigilance (Fescharek *et al.*, 2004; Hauben et Bate, 2009) pour la surveillance des médicaments et la génération des alertes lorsqu'un médicament présente un danger statistiquement significatif pour la population. Notre travail concerne la surveillance des médicaments.

Pour la détection de relations sémantiques, nous proposons d'exploiter des méthodes de structuration de termes indépendamment sur deux langues (français et anglais), puis de regrouper les résultats afin de consolider l'ensemble et en augmenter la qualité. Nous visons la détection de trois types de relations : variantes morpho-syntaxiques {*sténose de l'aorte*, *sténose aortique*}, synonymie {*tumeur gastrique*, *cancer gastrique*} et relations de subsomption hiérarchique {*défaillance rénale*, *défaillance rénale post-opératoire*}. Nous présentons d'abord le matériel (section 2) et décrivons la méthodologie (section 3). Nous présentons et discutons les résultats obtenus (section 4) et concluons avec des perspectives (section 5).

2 Matériel

Type de matériel	**anglais**	**français**
1. Termes médicaux	18 209	18 786
2. Données de référence	84	84
3. Ressources linguistiques		
UMLS : Synonymes d'UMLS	227 887	126 892
3t : Synonymes biomédicaux acquis	28 691	1 314
Gen : Synonymes de la langue générale	50 970	115 720

TABLE 1 – Matériel traité et exploité dans les deux langues (anglais et français).

Nous exploitons trois types de matériel (table 1). Chaque matériel existe en anglais et en français : il s'agit de ressources qui ont des contenus comparables dans les deux langues.

2.1 Termes de pharmacovigilance

Les termes exploités proviennent de la terminologie MedDRA (*Medical Dictionary for Regulatory Activities*) (Brown *et al.*, 1999), créée pour l'indexation, l'analyse et la surveillance des effets indésirables de médicaments. C'est une terminologie internationale créée et maintenue en anglais, et traduite en français et espagnol. Nous exploitons les termes préférés PT de cette terminologie en anglais et en français, 18 209 et 18 786 respectivement. Il s'agit donc globalement du même ensemble de termes, mais dont les libellés sont en langues différentes (anglais et français) : *leukaemia* et *leucémie*, *B-cell type acute leukaemia* et *Leucémie B aiguë*, *atypical depressive disorder* et *trouble dépressif atypique*, etc. Chaque terme reçoit un identifiant unique, qui reste le même quelle que soit la langue de la terminologie. Les termes de MedDRA sont structurés en cinq niveaux hiérarchiques. Au-dessus des termes PT, que nous exploitons, les termes de niveau HLT (*High Level Terms*) subsument hiérarchiquement les termes PT.

2.2 Données de référence

Les données de référence se présentent sous forme de clusters de termes liés à une condition médicale donnée. Ces données sont indépendantes de notre travail et elles ont été constituées

manuellement par des groupes d'experts. Il existe actuellement 84 clusters (CIOMS, 2004). Les conditions médicales sont par exemple : *Affections hépatiques, Rhabdomyolyse/Myopathie, Infarctus myocardique, Convulsions*. Il s'agit des conditions médicales graves qui peuvent causer des atteintes de santé et des hospitalisations, voire un décès. Les données de référence contiennent des relations sémantiques implicites entre les termes : on sait que tous les termes au sein des clusters de référence sont liés à une condition médicale mais la logique ou bien la nature de relations entre les termes ne sont pas connues.

2.3 Ressources linguistiques

Les ressources linguistiques externes que nous utilisons apportent des connaissances linguistiques et sémantiques sur les mots des termes. Ces ressources sont aussi en deux langues, français et anglais. Typiquement, ces ressources contiennent des relations de synonymie entre les mots ou les termes, comme par exemple {*accord*, *concordance*}, {*aceperone*, *acetabutone*} ou {*bleeding*, *hemorrhage*}. Trois types de synonymes sont utilisés :

UMLS : synonymes de la langue médicale extraits directement de la ressource terminologique UMLS (*Unified Medical Language System*) (NLM, 2011). Ces synonymes correspondent aux termes qui appartiennent à un même concept d'UMLS ;

3t : ressources de synonymie construites lors des travaux précédents pour le français (Grabar *et al.*, 2009) et l'anglais (Grabar et Hamon, 2010). Elles sont également adaptées à la langue médicale car elles ont été acquises à partir de trois terminologies biomédicales grâce à l'exploitation du principe de compositionnalité ;

Gen : synonymes de la langue générale fournis par le WordNet (Fellbaum, 1998) en anglais et par le Petit Robert (Robert, 1990) en français.

3 Méthodologie pour la détection de relations sémantiques

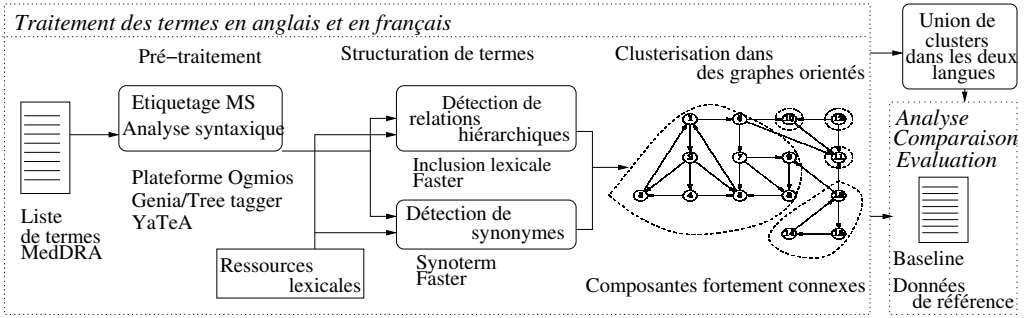


FIGURE 1 – Schéma général de la méthode.

À la figure 1, nous présentons le schéma général de notre approche. Si la détection de relations sémantiques est l'étape principale, notre approche comporte quatre autres étapes : le pré-traitement des termes, la clusterisation des relations générées, l'union des clusters générés

dans chaque langue, et l'évaluation des clusters (dans chaque langue et après leur union). La clusterisation est nécessaire pour effectuer l'évaluation. Comme nous l'avons indiqué (section 2.2), les données de référence sont des clusters de termes relatifs à une condition médicale donnée. Au sein de ces clusters, les termes ont des relations sémantiques (et médicales) entre eux, mais ces relations ne sont pas explicites. L'évaluation, que nous pouvons effectuer avec ces données de référence, porte donc sur l'appartenance des termes à un cluster de termes indépendamment des types de relations qui ont permis de relier ces termes.

La méthodologie que nous proposons est guidée par les données langagières et leurs propriétés telles que détectées dans le corpus de termes exploités dans les deux langues. Notre méthodologie ne requiert donc pas une étape d'apprentissage, elle ne requiert pas non plus de ressources sémantiques spécifiques. Dans la suite de cette section, nous présentons les cinq étapes de l'approche : (1) opérations effectuées pour le pré-traitement des termes (section 3.1), (2) étape de détection de relations sémantiques entre les termes (section 3.2), (3) clusterisation des termes grâce aux relations générées (section 3.3), (4) union de clusters générés dans chacune des deux langues (section 3.4), et (5) évaluation (section 3.5). Il est important de souligner que, de la même manière que le matériel, nous effectuons les mêmes traitements dans les deux langues : les méthodes et ressources exploitées sont en effet adaptées aux deux langues traitées.

3.1 Pré-traitement des termes

Le pré-traitement est effectué avec la plate-forme Ogmios (Hamon et Nazarenko, 2008). Après la segmentation en mots, les termes sont étiquetés morpho-syntaxiquement avec l'étiqueteur Genia (Tsuruoka *et al.*, 2005) en anglais et TreeTagger (Schmid, 1994) en français. Comme les termes sont des structures syntaxiques particulières (souvent des groupes nominaux et non pas des phrases bien formées) et pour garantir une meilleure qualité de l'étiquetage, nous transformons les termes en pseudo-phrases bien formées. Par exemple, le terme *fibrome du sein* est transformé en *C'est un fibrome du sein*. Par la suite, seuls les mots originaux des termes (*fibrome du sein*) sont considérés. Les termes sont aussi traités par l'analyseur syntaxique YTEA (Aubin et Hamon, 2006), qui permet de détecter les dépendances syntaxiques au sein des termes.

3.2 Méthodes de structuration de termes

Nous appliquons trois méthodes de l'état de l'art pour l'acquisition de relations sémantiques entre termes. Les relations visées sont la synonymie, la variation morpho-syntaxique et les relations hiérarchiques. Nous nous attendons à ce que ces relations relient des termes qui sont des équivalents sémantiques dans le contexte de notre travail. L'originalité et l'apport de notre travail concerne : (1) l'application de ces méthodes dans le contexte biomédical et aux données en deux langues, et (2) l'exploitation du contexte cross-langue pour exploiter les régularités sémantiques des termes dans deux langues et pour obtenir ainsi des résultats plus performants.

Variantes morpho-syntaxiques. L'identification de variantes morpho-syntaxique est effectuée avec Faster (Jacquemin, 1996). Trois règles de transformation sont appliquées : insertion (*cardiac disease*/*cardiac valve disease*), dérivation morphologique (*artery restenosis*/*arterial restenosis*) et permutation (*aorta coarctation*/*coarctation of the aorta*). Nous établissons une correspondance entre ces règles et les types de relations sémantiques :

– l'insertion introduit les relations hiérarchiques : *cardiac valve disease* est plus spécifique que *cardiac disease*,
– la permutation introduit les relations de synonymie : *aorta coarctation* et *coarctation of the aorta* sont en effet très proches sémantiquement,
– la dérivation morphologique introduit aussi des relations de synonymie : *artery restenosis*/*arterial restenosis*,
– par contre, lorsque plusieurs règles sont impliquées et lorsqu'il s'agit de règles correspondant aux relations hiérarchiques et de synonymie, ce sont les relations hiérarchiques qui prévalent (parce qu'elles sont plus spécifiques). Ainsi, pour la paire *gland abscess* et *abscess of salivary gland*, qui montre une insertion et une permutation, nous retenons la relation hiérarchique.

Compositionnalité et synonymie. Les relations de synonymie sont acquises de deux manières :
– la relation de synonymie est établie entre deux termes simples si cette relation existe dans les ressources linguistiques ;
– la relation de synonymie est établie entre deux termes complexes si la compositionnalité sémantique (Partee, 1984) est vérifiée pour ces termes. Ainsi, deux termes complexes sont considérés comme synonymes si au moins un de leurs composants à une position syntaxique donnée est synonyme et l'autre composant est identique (Hamon et Nazarenko, 2001). Par exemple, étant donné la relation de synonymie entre deux mots, *tumeur* et *cancer*, les termes *tumeur gastrique* et *cancer gastrique* sont identifiés comme synonymes.

Inclusion lexicale et hiérarchie. Selon l'hypothèse de l'inclusion lexicale (Kleiber et Tamba, 1990), il existe une relation de subsomption hiérarchique entre deux termes lorsqu'un terme est lexicalement inclus, à une position syntaxique donnée, dans un autre terme. Par exemple, le terme court *cancer* est le père hiérarchique, tandis que le terme long *cancer gastrique* est le fils hiérarchique parce que *cancer* est la tête syntaxique de *cancer gastrique*.

3.3 Clusterisation de termes

Les termes et les relations hiérarchiques sont représentés sous forme de graphes orientés : les termes sont les noeuds et les liens hiérarchiques les arcs orientés. Ces graphes sont partitionnés en composantes fortement connexes : dans un graphe orienté G, nous identifions des sous-graphes maximaux H de G, où pour chaque paire $\{x, y\}$ de noeuds de H, il existe un chemin composé d'arcs orientés de x à y. Avec ce type de composantes, un terme peut appartenir à plus d'un cluster, ce qui est aussi le cas des données de référence. Les clusters peuvent correspondre aux ensembles ou aux sous-ensembles des données de référence. Pour améliorer la couverture, nous ajoutons les synonymes : si un terme a une relation de synonymie avec un terme du cluster, ce terme est ajouté au cluster. Le terme central d'un cluster lui donne son libellé.

3.4 Union de clusters générés dans les deux langues

L'union de clusters, générés dans chaque langue, repose sur le libellé de ces clusters. Comme nous l'avons indiqué dans la section 2, chaque terme MedDRA reçoit un identifiant unique, qui reste le même quelle que soit la langue de ses termes. Ces deux informations (les libellés des clusters et les identifiant de ces libellés) permettent d'établir le lien entre les clusters correspondants dans

chacune des langues. Ainsi, lorsqu'il existe des clusters avec les mêmes libellés dans les deux langues, ils sont fusionnés, sinon l'union ne peut pas avoir lieu.

3.5 Évaluation et analyse de la complémentarité

Pour pouvoir exploiter ces données de référence, nous considérons l'ensemble des termes au sein des clusters et non pas chaque relation prise individuellement.

Le lien entre les clusters générés et les clusters de référence est effectué grâce au nombre de termes qu'ils partagent : pour un cluster de référence donné, nous sélectionnons celui des clusters générés qui partage le plus de termes communs avec lui. Une fois que les clusters de référence sont associés avec les clusters générés, les relations sémantiques générées sont évaluées contre les données de référence avec trois mesures :
– précision P (nombre de termes pertinents au sein d'un cluster divisé par le nombre total de termes au sein de ce cluster),
– rappel R (nombre de termes pertinents au sein d'un cluster divisé par le nombre total de termes dans le cluster de référence correspondant),
– F-mesure F_1 (la moyenne harmonique de P et de R).

Pour l'analyse de la complémentarité, nous analysons par exemple les points suivants :
– existence de relations uniques et communes entre les deux langues,
– amélioration de la couverture et/ou de la précision des résultats à l'aide des informations issues de deux langues.

Pour la baseline, nous exploitons l'approche la plus communément utilisée pour ce type de tâche : les relations sémantiques qui correspondent aux relations hiérarchiques de MedDRA (Mozzicato, 2007; Pearson *et al.*, 2009; Yuen *et al.*, 2008). Plus particulièrement, il s'agit de l'exploitation de la subsomption hiérarchique des termes PT au travers de leurs termes HLT de MedDRA. Parmis les 1 688 HLT et 84 groupements de référence, 46 ont une correspondance directe (*Thrombocytopenias* et *Thrombocytopenia (HLT)*) ou une correspondance non ambiguë (*Renal failure and impairment* et *Acute renal failure (HLT)*). Nous utilisons ces 46 clusters de référence pour l'évaluation des résultats obtenus avec la baseline. Ces 46 clusters sont un sous-ensemble de toutes les données de référence (84 clusters).

4 Résultats et Discussion

4.1 Détection de relations sémantiques

Dans la table 2, nous indiquons le nombre de relations acquises dans les deux langues. Nous pouvons faire les observations suivantes :
– il existe plus de relations générées en anglais qu'en français,
– chaque ressource linguistique exploitée en anglais contribue à l'acquisition de relations entre les termes, tandis qu'en français les synonymes d'UMLS ne fournissent pas de résultats,
– l'ensemble de relations hiérarchiques induites avec la subsomption lexicale en français (3 980) est plus grand qu'en anglais (3 366).

Relations et méthodes	# relations	
	anglais	français
Hiérarchique (inclusion lexicale)	3 366	3 980
Hiérarchique (variantes morpho-synt.)	316	178
Synonymie (UMLS)	54	-
Synonymie (relations acquises)	1 110	31
Synonymie (termes simples)	214	-
Synonymie (langue générale)	28	142
Nombre total de synonymes	1 459	164

TABLE 2 – Relations générées dans chaque langue.

Une remarque intéressante peut aussi être faite sur l'apport de ressources linguistiques. Nous voyons par exemple que les synonymes d'UMLS, qui sont directement accessibles dans cette ressource, fournissent un apport faible en anglais et un apport nul en français. Nos résultats indiquent ainsi clairement l'intérêt de construire et d'utiliser d'autres ressources linguistiques.

4.2 Génération de clusters

	anglais	français	union
Nombre de clusters	965	1 133	1 571
Taille des clusters (intervalle)	[2 ; 257]	[2 ; 205]	[2 ; 301]
Taille des clusters (moyenne)	6,39	4,97	6

TABLE 3 – Clusters générés dans chaque langue (anglais, français) et avec leur union.

La table 3 contient les données sur les clusters générés. Nous pouvons observer que le nombre de clusters, aussi bien que leurs tailles, sont plus grands lorsque les données des deux langues sont considérées. Ainsi, les deux langues sont complémentaires de différents points de vue : au niveau des relations et au niveau des clusters.

Relations et méthodes	% dans les clusters		
	anglais	français	union
Hiérarchique (inclusion lexicale)	79,57	96,66	89,2
Hiérarchique (variantes morpho-synt.)	8,62	2,56	4,36
Synonymes	11,81	0,78	6,44

TABLE 4 – Participation des relations dans la création de clusters.

Dans la table 4, nous indiquons à quelle hauteur les relations acquises participent dans la population des clusters. Ces valeurs sont indiquées en pourcentage par type de relations (hiérarchique, synonymie). Nous pouvons voir que les relations acquises par la subsomption hiérarchique apportent la majorité de termes dans les clusters (79,57 % en anglais et jusqu'à 96,66 % en

français), tandis que les relations de synonymie montrent seulement un impact très faible sur les clusters (moins de 1 % en français, mais jusqu'à 11,81 % en anglais).

4.3 Complémentarité des deux langues

Relations	anglais	français	intersection
Hiérarchique	1 919	2 395	1 763
Synonymie	1 332	137	27
Total	3 251	2 532	1 790

TABLE 5 – Nombre de relations spécifiques et communes (intersection) aux langues.

La table 5 indique la complémentarité entre les deux langues pour chaque type de relations : seulement 27 relations de synonymie, mais jusqu'à 1 763 relations hiérarchiques sont communes aux deux langues. Nous pouvons ainsi observer que la génération de relations hiérarchiques permet de détecter plus de régularités communes dans les deux langues. Mais plusieurs relations sont uniques à une langue (*i.e.*, {*abdominal rebound tenderness*, *abdominal tenderness*} en anglais, {*fibrome du sein*, *tumeur du sein*} en français). De manière générale, l'union indique que les langues contribuent de manière quasiment égale : 39,69 % de termes uniques en anglais, 34,03 % uniques en français, et 26,27 % de termes communs aux deux langues.

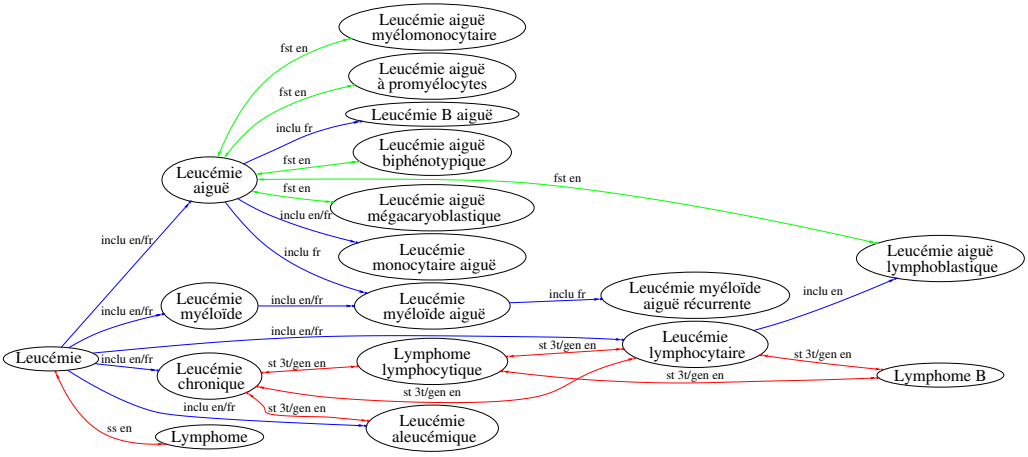
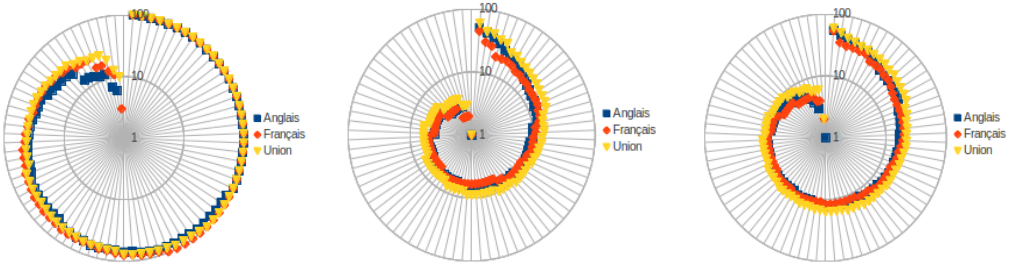


FIGURE 2 – Exemple de graphe (extrait du cluster *leucémie*).

À la figure 2, nous présentons un extrait de graphe sur l'exemple du cluster *leucémie*. Différentes couleurs de flèches correspondent aux différentes méthodes qui relient les termes (inclusion lexicale en bleu, Faster en vert, synonymie en rouge). Les flèches unidirectionnelles sont des relations hiérarchiques, les flèches bidirectionnelles sont des relations de synonymie. Nous indiquons également la et les langues où une relation donnée a été détectée. Souvent les inclusions lexicales sont détectées dans les deux langues. Mais il arrive aussi que le libellé d'un

terme ne permet de détecter une relation que dans une seule langue : les termes *acute leukaemia* et *b-cell type acute leukaemia* n'ont pas pu être reliés en anglais car les libellés sont trop éloignés lexicalement, par contre cette relation a été établie en français entre les termes correspondants *leucémie aiguë* et *leucémie B aiguë*. Nous pouvons aussi voir que la synonymie est surtout détectée en anglais (ressources de synonymie plus complètes). La synonymie directe entre les termes simples relie une seule paire de termes (*leukaemia* et *lymphoma*) dans cet exemple. Rappelons que les termes traités proviennent du même niveau hiérarchique dans MedDRA, alors que nous voyons que plusieurs niveaux de relations hiérarchiques peuvent être détectés : la structure actuelle de la terminologie MedDRA pourrait être affinée.

4.4 Évaluation



(a) Précision (moyenne) : anglais 64,7%, français 70%, union 69,5%

(b) Rappel (moyenne) : anglais 9,5%, français 8,5%, union 12%

(c) F-mesure (moyenne) : anglais 16,5, français 15, union 20,4

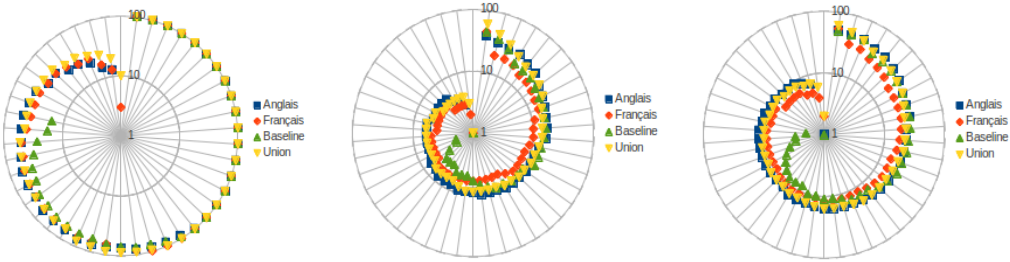
FIGURE 3 – Évaluation des clusters générés par rapport aux données de référence (84 clusters).

Les résultats de l'évaluation par rapport aux données de référence sont indiqués à la figure 3. Pour leur présentation, nous ne les projetons pas sur les axes x et y, mais sur un plan circulaire. Chaque rayon correspond à un cluster de référence (un total de 84 clusters de référence). L'échelle du rayon, réglée en mode logarithmique, va de 0 à 100 et permet de positionner les valeurs d'évaluation (précision, rappel et F-mesure). Dans cette présentation, plus une ligne (et une méthode) est proche du bord extérieur, meilleurs sont les résultats correspondants. Nous pouvons faire plusieurs observations sur les résultats obtenus :

– Très souvent, la précision est élevée tandis que le rappel est faible. La raison générale est que les clusters générés sont plus petits que les clusters de référence et peuvent de ce fait montrer leurs différents aspects. Du point de vue de la méthodologie, cela veut dire que les approches exploitées ne permettent pas de détecter toutes les relations qui seraient nécessaires pour grouper les termes des clusters de référence de manière automatique. Par exemple, nous ne détectons pas de relations entre les termes *ponction de moelle osseuse anormale* (*aspiration bone marrow abnormal*), *anémie réfractaire* (*anemia refractory*) et *leucémie* (*leukaemia*) qui sont pourtant tous liés aux anomalies du sang : *ponction de moelle osseuse anormale* est le résultat d'examen médical qui permet de dépister de telles anomalies, *anémie réfractaire* est une des conséquences possibles des leucémies.
– L'union des deux langues montre un effet positif sur le rappel et la F-mesure surtout. Notons que, contrairement à notre attente, la précision ne souffre pas beaucoup de l'union : elle est

améliorée par rapport à la valeur obtenue sur l'anglais (+4,8 %), mais elle perd 0,5 % par rapport à la valeur obtenue sur le français.

– Il existe une variabilité importante entre les performances de différents clusters générés. Cela est observable sur la forme des tracés : par exemple, pour la précision, presque la moitié des clusters générés montre des valeurs maximales (100 %) ou proches, mais pour les clusters restants, ces valeurs diminuent jusqu'à 10 %.



(a) Précision (moyenne) : anglais 71,1 %, français 70,5 %, union 74 %, baseline 69,3 %

(b) Rappel (moyenne) : anglais 11,8 %, français 8,4 %, union 12,1 %, baseline 9,2 %

(c) F-mesure (moyenne) : anglais 18,9, français 14, union 18,3, baseline 14,6

FIGURE 4 – Évaluation des clusters générés par rapport à la baseline et aux données de référence.

L'évaluation par rapport à la baseline (figure 4) indique que les résultats obtenus sur l'anglais sont meilleurs que la baseline, les résultats obtenus sur le français sont comparables et parfois moins bons que la baseline. Nous pouvons voir aussi que la qualité des clusters de la baseline décroît plus rapidement que dans les autres expériences, et ceci pour les trois mesures d'évaluation. En ce qui concerne les résultats de l'union, ils sont à la fois meilleurs que la baseline et meilleurs que chaque langue prise séparément (sauf une perte de 0,6 % pour la F-mesure).

Les performances d'autres méthodes automatiques exploitées pour cette tâche (similarité sémantique (Iavindrasana *et al.*, 2006; Dupuch *et al.*, 2012), requêtes OWL (Declerck *et al.*, 2012) ou subsomption hiérarchique (Jaulent et Alecu, 2009)) conduisent aux mêmes observations, lorsque une évaluation est effectuée : une des mesures d'évaluation est privilégiée. Notons que ces méthodes exploitent souvent des ressources dédiées à leur fonctionnement et qu'il est alors nécessaire d'y encoder les connaissances nécessaires à leur fonctionnement. Les avantages de notre approche sont : (1) elle ne requiert pas la création d'une ressource sémantique (terminologie ou ontologie) dédiée ; (2) elle n'est pas consommatrice en temps et effort pour constituer cette ressource, car elle fonctionne dans un contexte sémantique relativement pauvre ; (3) elle peut être exploitée avec une simple liste de termes et des ressources linguistiques disponibles ; (4) elle n'est pas spécifique aux données et à la tâche traitée ici et peut être exploitée dans d'autres contextes (acquisition de relations terminologiques, terminologies à facettes...).

5 Conclusion et perspectives

Nous avons proposé et réalisé des expériences consistant à exploiter des données linguistiques provenant de deux langues, anglais et français, pour obtenir des résultats plus complets et performants en détection de relations sémantiques entre les termes médicaux. L'objectif principal a consisté en vérification de la complémentarité entre les langues. En effet, lorsque les mêmes

méthodes sont appliquées aux mêmes ensembles de termes dans des langues différentes, la différence dans les résultats provient essentiellement du fait que les libellés de ces termes sont différents et qu'ils permettent de détecter des régularités linguistiques et sémantiques différentes et complémentaires. De manière générale, cette approche a permis d'améliorer les résultats obtenus en termes de précision, de rappel et de la F-mesure. Nous avons par exemple observé que chaque langue contribue de manière quasiment équivalente, même si l'apport par types de relations n'est pas comparable (les relations hiérarchiques sont beaucoup plus nombreuses que les relations de synonymie). La tâche visée dans notre travail (détection de relations sémantiques et médicales entre termes) est difficile, car il s'agit souvent de termes qui n'ont pas de similarité lexicale entre eux. La difficulté de la tâche est due aussi au fait que la surveillance des effets indésirables est effectuée dans un contexte réglementaire très strict. Cependant, avec la méthode proposée nous obtenons de meilleurs résultats que ceux fournis par la baseline. De plus, notre méthode fonctionne avec des méthodes et ressources assez "pauvres", qui ne requièrent pas de ressources terminologiques ou ontologiques dédiées.

Nous avons plusieurs perspectives à ce travail : (1) enrichissement des ressources linguistiques avec des ressources de type associatif qui pourraient être acquises avec des méthodes distributionnelles à partir de corpus et/ou de terminologies ; (2) exploitation de la méthode compositionnelle non seulement pour construire des ressources linguistiques de synonymie mais aussi pour acquérir des relations hiérarchiques ou associatives ; (3) exploration de corpus et application d'autres méthodes pour la détection automatique de relations sémantiques entre les termes ; (4) combinaison des résultats présentées dans ce travail avec ceux obtenus avec d'autres méthodes automatiques (Dupuch *et al.*, 2012) ; (5) clusterisation des termes avec d'autres approches, mais aussi au sein d'un graphe commun de toutes les relations générées, tandis qu'actuellement les clusters sont générés dans chaque langue séparément et fusionnés par la suite ; (6) affinement de la structure hiérarchique actuelle de la terminologie MedDRA grâce à la détection de niveaux hiérarchiques intermédiaires ; (7) adaptation des méthodes et ressources exploitées à la langue espagnole pour améliorer encore les performances des résultats.

Remerciements. Ce travail a été en partie soutenu par l'Agence Nationale de la Recherche (ANR) et la DGA, sous le numéro Tecsan ANR-11-TECS-012.

Références

Alfonseca, E., de Boni, M., Jara-Valencia, H.-L. et Manandhar, S. (2002). A prototype question answering system using syntactic and semantic information for answer retrieval. *In TREC 10*.

Anizi, M. et Dichy, J. (2009). Assessing word-form based search for information retrieval in Arabic : towards a new type of lexical resource. *In MEDAR*, pages 12–19.

Aubin, S. et Hamon, T. (2006). Improving term extraction with terminological resources. *In FinTAL 2006*, numéro 4139 de LNAI, pages 380–387. Springer.

Baeza-Yates, R. et Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley, New York.

Banea, C., Mihalcea, R. et Wiebe, J. (2011). Multilingual sentiment and subjectivity. *In Multilingual Natural Language Processing*, Prentice Hall.

Brown, E., Wood, L. et Wood, S. (1999). The medical dictionary for regulatory activities (MedDRA). *Drug Saf.*, 20(2):109–117.

CARTONI, B. et NAMER, F. (2012). Linguistique contrastive et morphologie : les noms en -iste dans une approche onomasiologique. *In CMLF*, pages 1245–1259.

CEUSTERS, W., DESIMPEL, I., SMITH, B. et SCHULZ, S. (2003). Using cross-lingual information to cope with underspecification in formal ontologies. *In MIE*, pages 391–396.

CIOMS (2004). Development and rational use of standardised MedDRA queries (SMQs) : Retrieving adverse drug reactions with MedDRA. Rapport technique, CIOMS.

COLLIER, N. (2011). Towards cross-lingual alerting for bursty epidemic events. *J Biomed Semantics*, 2(5):S10.

DECLERCK, G., BOUSQUET, C. et JAULENT, M. (2012). Automatic generation of MedDRA terms groupings using an ontology. *In MIE*, pages 73–77.

DIEWALD, G. et SMIRNOVA, E. (2010). *Evidentiality in European languages : the lexical-grammatical distinction*, chapitre Introduction, pages 1–14. Walter de Gruyter Mouton.

DUPUCH, M., BOUSQUET, C. et GRABAR, N. (2012). Automatic creation and refinement of the clusters of pharmacovigilance terms. *In ACM IHI*, pages 181–190.

FARRERES, X., RIGAU, G. et RODRIGUEZ, H. (1998). Using WordNet for building WordNets. *In COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*.

FELLBAUM, C. (1998). A semantic network of english : the mother of all WordNets. *Computers and Humanities. EuroWordNet : a multilingual database with lexical semantic network*, 32(2-3):209–220.

FESCHAREK, R., KÜBLER, J., ELSASSER, U., FRANK, M. et GÜTHLEIN, P. (2004). Medical dictionary for regulatory activities (MedDRA) : Data retrieval and presentation. *Int J Pharm Med*, 18(5):259–269.

FLEISCHMAN, S. (2001). *Language and Medicine*, pages 470–502. Blackwell.

FRIDMAN NOY, N. et MUSEN, M. (2000). Prompt : Algorithm and tool for automated ontology merging and alignment. *In AAAI*, pages 450–455.

GRABAR, N. et HAMON, T. (2010). Exploitation of linguistic indicators for automatic weighting of synonyms induced within three biomedical terminologies. *In MEDINFO 2010*, pages 1015–9.

GRABAR, N., VAROUTAS, P., RIZAND, P., LIVARTOWSKI, A. et HAMON, T. (2009). Automatic acquisition of synonym ressources and assessment of their impact on the enhanced search in EHRs. *Methods of Information in Medicine*, 48(2):149–154. PMID 19283312.

HAHN, U., HONECK, M., PIOTROWSKY, M. et SCHULZ, S. (2001). Subword segmentation - leveling out morphological variations for medical document retrieval. *In AMIA*.

HAMON, T. et NAZARENKO, A. (2001). Detection of synonymy links between terms : experiment and results. *In Recent Advances in Computational Terminology*, pages 185–208. John Benjamins.

HAMON, T. et NAZARENKO, A. (2008). Le développement d'une plate-forme pour l'annotation spécialisée de documents web : retour d'expérience. *TAL*, 49(2):127–154.

HAUBEN, M. et BATE, A. (2009). Decision support methods for the detection of adverse events in post-marketing data. *Drug Discov Today*, 14(7-8):343–357.

HUANG, C.-R., TSENG, I. J. E. et TSAI, D. (2002). Translating lexical semantic relations : The first step towards multilingual WordNets. *In COLING Workshop SemaNet'02*.

IAVINDRASANA, J., BOUSQUET, C., DEGOULET, P. et JAULENT, M. (2006). Clustering WHO-ART terms using semantic distance and machine algorithms. *In AMIA*, pages 369–373.

JACQUEMIN, C. (1996). A symbolic and surgical acquisition of terms through variation. *In* WERMTER, S., RILOFF, E. et SCHELER, G., éditeurs : *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 425–438, Springer.

JAULENT, M. et ALECU, I. (2009). Evaluation of an ontological resource for pharmacovigilance. *In MIE*, pages 522–526.

KLEIBER, G. et TAMBA, I. (1990). L'hyperonymie revisitée : inclusion et hiérarchie. *Langages*, 98:7–32.

LEFER, M. et GRABAR, N. (2013). French evaluative prefixes in translation : from automatic alignment to semantic categorization. *In CIL TACMO workshop*. To appear.

LI, A. (2011). A comparative study of argument structure and lexicon. *In International Conference on Bilingualism and Comparative Linguistics*.

MALAISÉ, V., ISAAC, A., GAZENDAM, L. et BRUGMAN, H. (2007). Anchoring Dutch cultural heritage thesauri to WordNet : two case studies. *In Workshop on Language Technology for Cultural Heritage Data (LaTeCH*, pages 57–64.

MARKO, K., BAUD, R., ZWEIGENBAUM, P., BORIN, L., MERKEL, M. et SCHULZ, S. (2006). Towards a multilingual medical lexicon. *In AMIA*, pages 534–538.

MOZZICATO, P. (2007). Standardised MedDRA queries : their role in signal detection. *Drug Saf*, 30(7):617–619.

NLM (2011). *UMLS Knowledge Sources Manual*. National Library of Medicine, Bethesda, Maryland. www.nlm.nih.gov/research/umls/.

PARTEE, B. (1984). *Compositionality*. F Landman and F Veltman.

PEARSON, R., HAUBEN, M., GOLDSMITH, D., GOULD, A., MADIGAN, D., O'HARA, D., REISINGER, S. et HOCHBERG, A. (2009). Influence of the MedDRA hierarchy on pharmacovigilance data mining results. *Int J Med Inform*, 78(12):97–103.

ROBERT, L. (1990). *Le petit Robert*. Le Robert, Paris.

RODRIGUES, J., RECTOR, A., ZANSTRA, P., BAUD, R., INNES, K., ROGERS, J., RASSINOUX, A., SCHULZ, S., PAVIOT, B. T., TEN NAPEL, H., CLAVEL, L., VAN DER HARING, E. et MATEUS, C. (2006). An ontology driven collaborative development for biomedical terminologies : from the french CCAM to the australian ICHI coding system. *In MIE*, pages 863–868.

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In ICNMLP*, pages 44–49, Manchester, UK.

SCHULZ, S. et HAHN, U. (2000). Morpheme-based, cross-lingual indexing for medical document retrieval. *Int J Med Inform*, 58-59:87–99.

STEINBERGER, R. (2011). Cross-lingual keyword assignment. *Procesamiento del Lenguaje Natural*, 27:273–280.

TSURUOKA, Y., TATEISHI, Y., KIM, J., OHTA, T., MCNAUGHT, J., ANANIADOU, S. et TSUJII, J. (2005). Developing a robust part-of-speech tagger for biomedical text. *LNCS*, 3746:382–392.

VINAY, J. et DARBELNET, J. (1958). *Stylistique Comparée du Français et de l'Anglais*. Didier-Harrap.

WILLETT, T. (1988). A cross-linguistic survey of the grammaticalization of evidentiality. *Studies in Language*, 12:51–97.

YUEN, N., FRAM, D., VANDERWALL, D. et ALMENOFF, J. (2008). Do standardized MedDRA queries add value to safety data mining ? *In ICPE 2008*, pages 1–2.