Structuration de terminologies pour la création de groupements de termes en pharmacovigilance

Marie Dupuch[1,2], Laëtitia Dupuch[3],Amandine Périnet[4], Thierry Hamon[4], Natalia Grabar[2]

[1]CNRS LIFL/Mostrare UMR 8022, Université Lille 1&3, 59653 – Villeneuve d'Ascq– France
[2]CNRS STL UMR 8163 – Université Lille 3 – 59653 – Villeneuve d'Ascq– France
[3]Université Toulouse III Paul Sabatier – 31062 – Toulouse Cedex– France
[4]LIM&BIO (EA3969) – Université Paris 13 – 93017 – Bobigny Cedex– France

Abstract

Pharmacovigilance is the activity related to the collection, the analysis and the prevention of adverse drug reactions (ADR) induced by drugs. Several statistical methods and more recently the groupings of terms are used to detect and to improve the signal in pharmacovigilance. Standardized MedDRA Queries (SMQs) are the gold standard groupings of terms in the pharmacovigilance area. Currently the SMQs are created by the experts and several medical conditions are not covered by the SMQ. The objective of this work is to propose an automatic method for the generation of groupings of terms and of SMQs for non described safety topics. We propose to use the methods designed for the structuring of terms (detection of synonymous and hierarchical relations) for the generation of these groupings. Then we evaluate and discuss in details the generated results.

Résumé

La pharmacovigilance est l'ensemble des activités liées au recueil, à l'analyse et à la prévention des effets indésirables (EI) susceptibles d'être dus à un médicament. Des méthodes statistiques et plus récemment des groupements de termes sont utilisés pour détecter et améliorer le signal en pharmacovigilance. Standardized MedDRA Queries (SMQ) sont des groupements de références dans le domaine de la pharmacovigilance. Actuellement ils sont réalisés manuellement par des experts et plusieurs conditions médicales ne sont pas couvertes par les SMQ. L'objectif de ce travail est de proposer une méthode automatique pour la génération de groupements de termes et de SMQ pour des problèmes médicaux non encore décrits. Nous proposons d'utiliser les méthodes de structuration de termes (détection de relations de synonymes et hiérarchiques) pour la génération de ces groupements. Enfin nous proposons une évaluation et discutons en détail les résultats générés.

Mots-clés : Structuration de termes, groupements de termes, pharmacovigilance, détection du signal.

1. Introduction

La détection de termes synonymes est cruciale pour plusieurs applications, comme la recherche et l'extraction d'information, la structuration de terminologies ou l'annotation sémantique de documents. Par exemple, en recherche d'information, il est important de retrouver les documents répondant au mot-clé *muscle ache* lorsque l'utilisateur donne le mot-clé *muscle pain*. Dans les

domaines de spécialité cette information est souvent calculée avec des méthodes de structuration de termes et de détection de variantes de termes (Jacquemin et al, 1997 ; Hamon et al, 1998 ; Verspoor *et al.*, 2003). De plus, l'utilisation de termes équivalents qui sont sémantiquement proches, sans pour autant être des synonymes, apparaît aussi nécessaire pour augmenter la couverture et la sensibilité d'une application.

Dans le contexte de la pharmacovigilance (recueil, à l'analyse et à la prévention des effets indésirables susceptibles d'être dus à un médicament), la détection et l'exploitation de termes sémantiquement proches présente un enjeu important. En effet, lorsque cette information est disponible, elle permet de grouper ensemble les effets indésirables qui sont codés avec des termes différents mais qui sont pourtant très proches, comme dans ces exemples : {*asystolic* ; *asystole*}, {*hematoma muscle* ; *hemorrhage muscle*}, {*muscle ache* ; *muscle pain*}, {*localized muscle* ; *localized muscle weakness*}. Par conséquent, cette information permet d'agglomérer les cas de pharmacovigilance et ainsi d'intensifier le signal (ou l'alerte) de pharmacovigilance : les effets indésirables de médicaments peuvent ainsi être identifiés plus efficacement et rapidement. Pour la même raison, la sécurité des produits de santé est augmentée car le signal peut émerger plus rapidement.

Dans la situation actuelle, il existe des banques de données de pharmacovigilance à différents niveaux : régional, national (AFSSAPS), européen (EMA) et mondial (OMS). Dans ces banques, les effets indésirables de médicaments sont codés avec la terminologie MedDRA (Brown *et al.*, 1999). Les méthodes statistiques sont actuellement les plus utilisées pour la détection de signaux de pharmacovigilance (Meyboom et al, 2002). Par contre, comme il a été observé que pour augmenter l'intensité ou détecter plus tôt un signal, le regroupement de cas similaires de pharmacovigilance est nécessaire (Hauben et Bate, 2009), il existe une initiative qui consiste à créer des SMQ (Standardized MedDRA Queries) regroupant des termes médicaux proches. Les SMQ contiennent les termes préférés (*Prefered Terms* ou *PT*) de MedDRA, car ces termes sont utilisés pour le codage des effets indésirables. La création des SMQ est effectuée manuellement par les experts. Dans ce processus, les experts s'appuient sur la structure hiérarchique de MedDRA. Cette hiérarchie est composée de cinq niveaux hiérarchiques, dont le plus haut correspond au *SOC* (*System Class Organ*) et subsume des termes relatifs à une localisation anatomique (système nerveux, système digestif, ...) ou à des procédures et des analyses de laboratoire. Les experts exploitent également la littérature scientifique. Les SMQ regroupent les termes associés à une condition médicale donnée, comme par exemple *Agranulocytosis*, *Acute renal failure* ou *Embolic and thrombotic events, arterial* ...

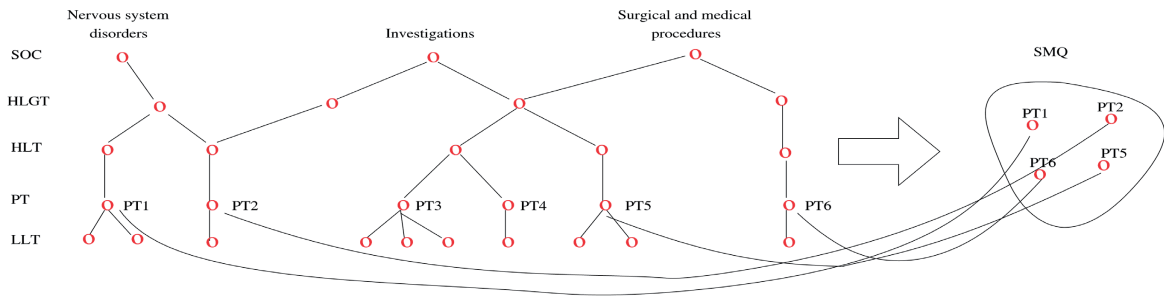


Figure 1 : Recrutement de termes de MedDRA pour la création des SMQ.

Le schéma de la figure 1 illustre le résultat de recrutement des termes PT pour les SMQ. Nous pouvons voir sur ce schéma que, d'une part les termes présents dans les SMQ peuvent appartenir à différents SOC de la terminologie MedDRA. Dans l'exemple de la figure 1, les termes provenant des SOC *Nervous System Disorders*, *Investigations* et *Surgical and medical procedures* sont ainsi recrutés. Au sein des 84 SMQ existants, la variété des SOC varie entre 4 et 25 (le nombre total de SOC étant 32), alors que la moyenne est de 8,26 SOC par SMQ. D'autre part, un même PT peut appartenir à plus d'un SMQ. Par exemple, les termes PT *Kidney failure* et *Acute renal insufficiency* appartiennent respectivement à 9 et 8 SMQ. En effet, ces effets indésirables peuvent apparaître en relation avec différentes conditions médicales. Pour le terme *Kidney failure* il s'agit par exemple des SMQ suivants : *Rhabdomyolysis/myopathy ; Acute renal failure ; Retroperitoneal fibrosis ; Torsade de pointes, shock-associated conditions ; Hypovolaemic shock conditions ; Toxic-septic shock conditions ; Shock-associated circulatory or cardiac conditions (excl torsade de pointes) ; Anaphylactic/anaphylactoid shock conditions ; Hypoglycaemic and neurogenic shock conditions*. Ces observations montrent que le recrutement des termes PT pour les SMQ suit une logique médicale bien précise et surtout que la structure actuelle de MedDRA ne peut pas permettre de bien respecter cette logique.

Comme le processus de création des SMQ est long et méticuleux, notre objectif est de proposer une méthode automatique pour assister la création de ces groupements de termes. Il existe actuellement quelques travaux qui ont abordé cette problématique. Par exemple, directement dans MedDRA, les experts en pharmacovigilance peuvent exploiter la structure hiérarchique (niveaux intermédiaires *HLT* et *HLGT* du schéma 1), mais comme le montre l'état de l'art (Mozzicato, 2007 ; Pearson *et al.*, 2009) et nos observations, la subsomption hiérarchique de MedDRA ne peut pas satisfaire les différents aspects pertinents pour décrire une condition médicale. D'autres travaux cherchent à exploiter une ressource spécifique, appelée ontoEIM (Alecu *et al.*, 2008), qui propose une structuration plus évoluée des termes MedDRA. Mais comme cette ressource ne couvre pas tous les termes PT de MedDRA, les groupements qui sont réalisés sont aussi lacunaires (Alecu *et al.*, 2008 ; Jaulent *et al.*, 2009 ; Dupuch *et al.*, 2011).

Dans le présent travail, nous proposons d'utiliser les méthodes du Traitement Automatique de Langues (TAL) pour la structuration de termes (détection de termes synonymes ou reliés hiérarchiquement). L'avantage de ces méthodes est qu'elles ne dépendent pas de ressources existantes (MedDRA et ontoEIM) et permettent de traiter l'ensemble de termes PT de MedDRA. Pour valider les résultats obtenus, nous évaluons les groupements calculés automatiquement par rapport aux SMQ. Une évaluation supplémentaire est effectuée avec un expert. Dans la suite de ce travail, nous présentons d'abord le matériel exploité (sec. 2) et la méthode proposée (sec. 3). Nous présentons et discutons les résultats (sec. 4) et concluons avec quelques perspectives (sec. 5).

2. Matériel

Nous exploitons trois types de matériel en langue anglaise : termes PT de MedDRA que nous cherchons à regrouper, ressources lexicales utilisées par les méthodes de TAL et groupements de référence SMQ par rapport auxquels nous évaluons nos groupements.

2.1. Termes PT de MedDRA

La terminologie MedDRA a été spécifiquement conçue pour le codage des effets indésirables. MedDRA contient un large spectre de termes (signes et symptômes, diagnostics, examens de laboratoire, procédures médicales et chirurgicales, antécédents), organisés en cinq niveaux hiérarchiques. Nous exploitons les 18 209 termes PT.

2.2. Ressources lexicales

Les ressources lexicales exploitées sont des paires de termes ou de mots synonymes. Nous utilisons trois ensembles de synonymes :

1. Synonymes médicaux extraits directement d'UMLS (n=228 542) et nettoyés (n=73 093);
2. Synonymes médicaux acquis à partir de trois terminologies biomédicales grâce à l'exploitation de leur compositionnalité (Grabar et Hamon, 2010) (n=28 691);
3. Synonymes de la langue générale fournis par WordNet (Fellbaum, 1998) (n=45 782).

Parmis les paires de mots fournies par ces ressources, nous retrouvons par exemple {*accord*, *concordance*}, {*adrenaline*, *epinephrine*} ou {gastrointestinal *bleeding*, gastrointestinal *hemorrhage*}. Les deux dernières paires sont fournies par les ressources de la langue générale et médicale, tandis que la première paire {*accord*, *concordance*} est fournie uniquement par les ressources médicales. Dans WordNet, *accord* et *concordance* appartiennent à des synsets distincts.

2.3. Groupements de référence SMQ

Les SMQ (Standardised MedDRA Queries) sont des groupements de termes MedDRA liés à une condition médicale (ou diagnostic), comme par exemple *Acute renal failure*, *Agranulocytosis* ou *Embolic and thrombotic events, arterial*. Les SMQ sont créés pour apporter une aide dans la recherche de cas pertinents. Nous utilisons les SMQ comme nos données de référence. Actuellement il existe 84 SMQ, parmi ces SMQ nous pouvons aussi distinguer 20 SMQ qui sont structurés hiérarchiquement. Les SMQ hiérarchiques peuvent être structurés en de nombreux niveaux hiérarchiques. Bien que la majorité des SMQ contiennent en moyenne deux niveaux hiérarchiques, le nombre de niveaux peut varier entre deux et six. Par exemple, le SMQ hiérarchique *Cerebrovascular disorders* possède trois niveaux hiérarchiques et cinq sous-SMQ (entre parenthèses nous indiquons le nombre de PT présents dans chaque niveau) :

- *Cerebrovascular disorders* (198)
  - -- *Central nervous system haemorrages and cerebrovascular conditions* (30)
    - --- *Ischaemic cerebrovascular conditions* (67)
    - --- *Haemorrhagic cerebrovascular conditions* (35)
    - --- *Conditions associated with central nervous system haemorrhages and cerebrovascular accidents* (30)
  - -- *Cerebrovascular disorders, not specified as haemorrhagic or ischaemic* (18)

Nous utilisons trois ensembles de données de référence : 84 SMQ, 20 SMQ hiérarchiques et 92 sous-SMQ.

3. Méthodes

La figure 2 illustre l'approche que nous proposons d'appliquer. Cette approche s'appuie d'une part sur l'analyse de la structure interne des termes (identification de relations synonymiques et hiérarchiques) mais aussi sur le partitionnement du réseau de termes pour l'identification de sous-graphes. La méthode est appliquée à tous les termes PT de MedDRA (n=18 209).

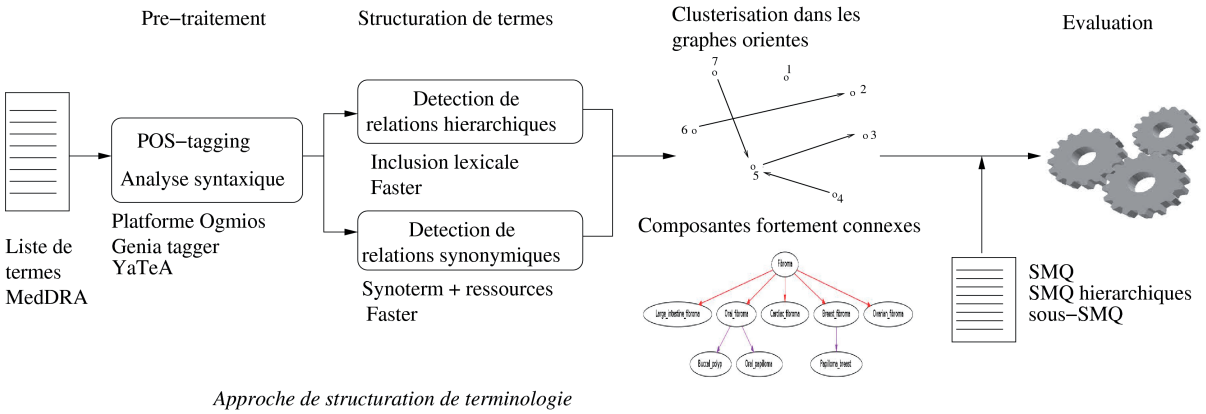


Figure 2 : Schéma général de la méthode composée de 4 étapes : (1) pré-traitement des données, (2) identification des relations hiérarchiques et synonymiques, (3) génération des groupements de termes et (4) évaluation des groupements de termes.

Comme le montre le schéma 2, notre approche est composée de plusieurs étapes. À partir de la liste à plat de termes PT de MedDRA, nous effectuons : (1) pré-traitement des données, (2) acquisition des relations sémantiques (hiérarchiques et synonymiques), (3) génération des groupements de termes et (4) évaluation des groupements de termes.

3.1. Pré-traitement des termes PT de MedDRA

Les termes de MedDRA sont analysés linguistiquement à travers la plateforme de TAL Ogmios (Hamon et Nazarenko, 2008). Leur étiquetage morpho-syntaxique est réalisé à l'aide de GeniaTagger (Tsuruoka *et al.*, 2005), et leur analyse syntaxique en tête/modifieur avec l'extracteur de termes YATEA (Aubin et Hamon, 2006). Un exemple de deux termes analysés syntaxiquement *Infection of navel cord* et *Infection of umbilical stump* est présenté sur la figure 3. Les deux termes sont analysés selon leurs dépendances syntaxiques : têtes et expansions syntaxiques. La tête syntaxique des deux termes est *infection*, leurs expansions sont respectivement *navel cord* et *umbilical cord*.

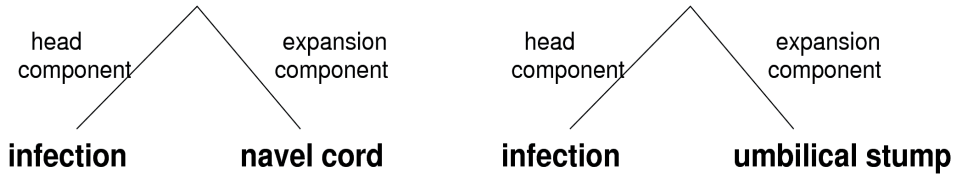


Figure 3 : Exemple de deux termes (Infection of navel cord et Infection of umbilical stump) analysés syntaxiquement. Les termes sont décomposés syntaxiquement en têtes et expansions.

3.2. Acquisition de relations sémantiques

Nous exploitons plusieurs méthodes pour acquérir des relations sémantiques entre les termes MedDRA. Ces méthodes fonctionnent hors contexte et cherchent à détecter des indices sur les relations sémantiques au sein même des termes. Les méthodes exploitées sont : inclusion lexicale, variation morpho-syntaxique et acquisition de relations de synonymie par propagation de la compositionnalité.

3.2.1. Inclusion lexicale

Pour identifier des relations hiérarchiques, nous nous appuyons sur l'hypothèse d'inclusion lexicale, qui stipule que lorsqu'un terme est inclus lexicalement dans un autre, une relation hiérarchique peut généralement être établie entre le terme court (le père hiérarchique) et le terme long (son fils). Nous calculons deux types d'inclusions lexicales qui exploitent la décomposition syntaxique des termes en tête et expansion :

1. l'inclusion avec la tête minimale utilise la décomposition syntaxique des termes et calcule la plus petite forme lexicale du terme. Par exemple, pour le terme *kaolin cephalin clotting time,* la tête minimale du terme est réduite à *time,*
2. l'inclusion avec la tête maximale utilise la décomposition syntaxique des termes et calcule la forme la plus complète de la tête syntaxique. Dans l'exemple précédent, la tête maximale est alors *cephalin clotting time.*

3.2.2. Identification de variantes morpho-syntaxiques

Pour identifier des variantes morpho-syntaxiques parmi des termes PT MedDRA, nous exploitons l'outil Faster (Jacquemin et al, 1997). Faster effectue plusieurs opérations de transformation, comme par exemple l'insertion d'un modifieur (*cardiac disease / cardiac valve disease),* dérivation *(artery restenosis / arterial restenosis)* et permutation *(aorta coarctation / coarctation of the aorta).* À chacune des opérations de transformation morpho-syntaxique nous avons associé un des deux types de relations sémantiques (synonymie et hyperonymie). Ainsi, l'insertion d'un modifieur permet d'identifier des relations hiérarchiques : la relation de variation entre les termes *cardiac valve disease* et *cardiac disease* est interprétée comme une relation d'hyperonymie*,* le premier terme étant plus spécifique que le second. Quant à la transformation par permutation, comme dans {*aorta coarctation ; coarctation of the aorta*}, elle correspond à la relation de synonymie.

Cependant, certaines règles de transformation sont ambiguës et ne correspondent pas à un type précis de relation. C'est le cas notamment lorsque plusieurs opérations de transformation sont

impliquées. Pour désambiguïser ces règles, nous avons procédé à un tri manuel des relations identifiées. Par exemple, la relation de variation entre *gland abscess* et *abscess of salivary gland* est issue à la fois d'une opération de permutation et d'insertion. Alors que la première opération correspond généralement à la synonymie, la seconde permet d'identifier des relations d'hyperonymie. Dans ce type de situations, nous considérons que la relation d'hyperonymie prévaut.

3.2.3. Identification des relations synonymiques

Les relations de synonymie entre termes sont acquises de deux manières. D'une part, nous mettons en relation deux termes simples MedDRA lorsque ceux-ci apparaissent en relation de synonymie dans les ressources lexicales exploitées. D'autre part, l'identification de relations de synonymie entre les termes complexes s'appuie sur le principe de compositionnalité (Partee, 1984). Selon ce principe, la synonymie est préservée à travers la compositionnalité. Ainsi, deux termes complexes sont considérés comme synonymes si au moins un de leurs composants dans la même position syntaxique sont synonymes. Par exemple, étant donnée la relation de synonymie entre les mots *infections* et *sepsis*, les termes *wound infection* (*infection de blessure*) et *wound sepsis* (*septicité de blessur*e) sont synonymes (Hamon et al, 1998). Les trois règles de transformation appliquées sont les suivantes :

> *Règle 1* : Deux termes sont synonymes si leurs structures syntaxiques sont identiques, si leurs têtes syntaxiques sont identiques et si une relation de synonymie existe entre leurs expansions syntaxiques. Par exemple, si nous avons la relation de synonymie {*navel cord* ; *umbilical stump*} alors nous pouvons inférer la relation de synonymie entre les termes complexes *Infection of navel cord* et *Infection of umbilical stump* où la tête syntaxique *infection* est identique (figure 3).
>
> *Règle 2* : Deux termes sont synonymes si leurs structures syntaxiques sont identiques, si leurs expansions syntaxiques sont identiques et si une relation de synonymie existe entre leurs têtes syntaxiques. Nous pouvons ainsi inférer une relation de synonymie entre les termes *Grippe aviaire* et *Influenza aviaire* si {*grippe ; influenza*} sont synonymes.
>
> *Règle 3* : Deux termes sont synonymes si leurs structures syntaxiques sont identiques et si leurs tête syntaxiques ou leurs expansions syntaxiques sont synonymes. Par exemple, la relation de synonymes entre les termes *Angine pustuleuse* et *Pharyngite vésiculeuse* est inférée car {*pustuleux* ; *vésiculeux*} et {*angine* ; *pharyngite*} sont des synonymes connus.

Pour calculer les relations de synonymie entre les termes MedDRA, nous effectuons plusieurs expériences. Chaque ressource de synonymes médicaux est d'abord utilisée individuellement et ensuite en combinaison avec WordNet.

3.3. Génération des groupements de termes MedDRA

Les ensembles de relations hiérarchiques générés par nos méthodes sont considérés comme des graphes orientés : les termes sont les nœuds des graphes tandis que les relations hiérarchiques sont les arcs orientés (figure 4).

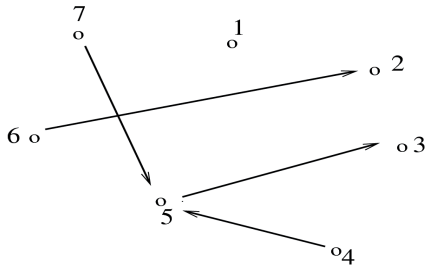


Figure 4 : Exemple de graphe orienté.

Nous cherchons ensuite à partitionner ces graphes orientés afin de générer des groupements de termes. Pour cela, nous exploitons la notion de composantes fortement connexes. Il s'agit d'identifier dans un graphe orienté G, les sous-graphes maximaux H de G tel que pour tout couple {x, y} de sommets H, il existe un arc orienté de x vers y. Ainsi, le graphe de la figure 4 comporte six composantes fortement connexes : {1}, {2}, {3}, {2, 6}, {3, 5}, {3, 4, 5} et {3, 5, 7}. Dans notre étude, ces composantes fortement connexes sont considérées comme des groupements de termes : il existe un recouvrement entre les SMQ et c'est également le cas avec les composantes fortement connexes. Pour améliorer le rappel des groupements, nous ajoutons les synonymes générés : si un terme a une relation de synonymie avec un terme du groupement alors il est ajouté à ce groupement. Du point de la théorie des graphes, il s'agit d'augmenter le graphe initial par deux arcs orientés vers et à partir des termes synonymes.

3.4. Évaluation des groupements de termes

Tout d'abord, nous effectuons une évaluation des relations hiérarchiques générées. C'est une évaluation qualitative effectuée manuellement.

Nous effectuons également une évaluation des résultats par rapport à l'application que nous visons et grâce à une comparaison avec les 84 SMQ, les 20 SMQ hiérarchiques et les 92 sous-SMQ. Une évaluation quantitative est effectuée avec trois mesures classiques : précision P (pourcentage de termes pertinents retrouvés rapporté au nombre de termes total groupés), rappel R (pourcentage de termes pertinents retrouvés rapporté au nombre de termes dans un SMQ) et F-mesure F (la moyenne harmonique de P et R). L'association entre les SMQ et les groupements est guidée par la précision de ces groupements : le groupement qui a la précision la plus élevée par rapport à un SMQ est associé à ce SMQ. Le partitionnement du graphe en composantes fortement connexes favorise les groupements de petite taille, nous effectuons donc également *a posteriori* la fusion des *n* meilleurs groupements proposés par notre méthode. Nous réalisons la fusion de groupements en fonction de seuils de précision compris entre 10 et 90%, et permettant d'obtenir une F-mesure optimale. Nous présentons les résultats pour une précision supérieure à 10%. Ces mêmes résultats sont aussi évalués qualitativement avec un expert.

Par ailleurs, nous étudions le contenu des groupements afin d'identifier les indices permettant d'améliorer la qualité de notre méthode, et nous évaluons la contribution des différentes approches et des ressources de synonymes. Pour ceci, nous étudions l'apport quantitatif des ressources et des approches exploitées dans notre travail pour la génération des groupements. Nous calculons d'une part le nombre et la proportion de groupements auxquels les approches contribuent. D'autre part, nous évaluons la macro-précision de ces approches par rapport à la

génération des groupements : nous considérons alors que si une relation ajoute au moins un terme au groupement qui ne devrait pas y apparaître (car non présent dans le SMQ) celle-ci n'est pas correcte.

4. Description et discussion des relations et des groupements générés

Dans le tableau 1, nous présentons le nombre de relations générées pour les deux types de relations visés : hiérarchiques et synonymiques. Nous pouvons observer que les relations hiérarchiques sont plus nombreuses que les relations synonymiques. C'est surtout le cas de relations hiérarchiques générées avec l'analyse syntaxique des termes, tandis que la variation morpho-syntaxique fournit 743 paires hiérarchiques. Quant aux relations synonymiques, nous obtenons presque 2 000 paires de termes fournit par les trois terminologies biomédicales et seulement 190 avec les synonymes extraits d'UMLS. La variation morpho-syntaxique fournit le plus petit nombre de relations. L'influence des synonymes de la langue générale de WordNet est très faible (70 paires en moyenne).

Type de relations	Nombre de relations
Relations hiérarchiques	
Tête syntaxique maximale	3 3663 816
Tête syntaxique minimale	743
Variation morpho-syntaxique	
Synonymes médicaux	
3 terminologies biomédicales	1879
UMLS	190
UMLS nettoyé	190
Variation morpho-syntaxique	100
Synonymes médicaux et WordNet	
3 terminologies biomédicales	1939
UMLS	270
UMLS nettoyé	270

Table 1 : Types et nombres de relations sémantiques générées avec les méthodes de structuration et de variation de termes.

À partir de ces paires de termes générées, nous partitionnons les graphes orientés et obtenons les composantes fortement connexes (ou des groupements) de termes. Ainsi, nous obtenons 836 et 504 composantes fortement connexes avec les têtes syntaxiques maximales et minimales respectivement. Parmi les 18 209 termes MedDRA, 4 061 (22,3 %) termes sont mobilisés au sein des composantes avec les têtes maximales et 4 319 (23,7 %) dans les composantes avec les têtes minimales. Le nombre moyen de termes par groupement est de 5,3 avec les têtes maximales et de 8,6 avec les têtes minimales, sachant que les groupements de 2 termes sont très fréquents. L'enrichissement des groupements avec les synonymes augmente le nombre de termes mobilisés : avec les têtes minimales, nous avons 5 007 termes différents mobilisés, avec la moyenne de 9,9 termes par groupement, tandis qu'avec les têtes maximales, nous obtenons 4 661 termes différents, avec une moyenne de 6,2 termes par groupement.

4.1. Évaluation de la méthode

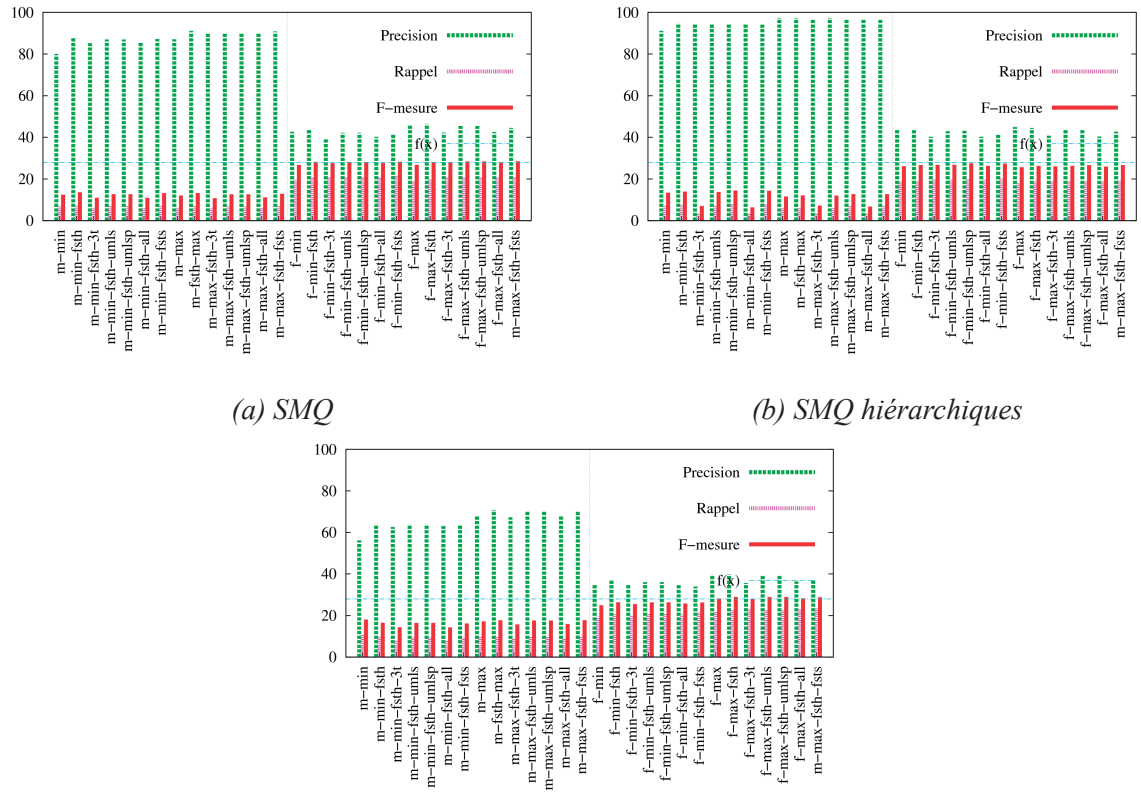
Une analyse manuelle des paires de termes reliées avec les relations hiérarchiques a montré que ces relations induites sont toujours correctes. Par contre, il peut y avoir des ambiguïtés de dépendance et de portée syntaxique au sein des termes. Ces ambiguïtés apparaissent surtout lors du calcul des têtes syntaxiques maximales. Par exemple, notre méthode calcule une relation hiérarchique entre *anticonvulsant drug level* et *drug level*, où le terme MedDRA *drug level* est le père hiérarchique. Il nous semble que dans cette paire, nous avons une ambiguïté sur le rattachement de *drug.* Ainsi, la tête syntaxique maximale peut être soit *drug level*, comme proposé par la méthode, soit *level*. L'expansion syntaxique est alors soit *anticonvulsant,* soit *anticonvulsant drug*. D'autres paires sont dans cette situation, comme *blood smear test* et *smear test*, *cranial nerve injury* et *nerve injury*, *eye movement disorder* et *movement disorder*, *central nervous system neoplasm* et *nervous system neoplasm*. Au total, avec les têtes syntaxiques maximales, nous avons 144 paires avec ce type d'ambiguïté, ce qui correspond à environ 5 %. Mais même s'il existe cette ambiguïté syntaxique, les relations sémantiques induites restent correctes. Avec les têtes minimales, nous avons juste deux paires avec des ambiguïtés de rattachement. Dans la majorité de cas, le calcul de dépendances syntaxiques est correct, comme dans les exemples suivants : *clear cell sarcoma of the kidney* et *sarcoma, blood incompatibility haemolytic anaemia of newborn* et *anaemia*, *radiotherapy to soft tissue* et *radiotherapy* avec les têtes minimales, *peripheral neuroepithelioma of bone* et *peripheral neuroepithelioma*, *electron radiation therapy to soft tissue* et *electron radiation therapy*, *neonatal respiratory distress syndrome prophylaxis* et *prophylaxis*, *malignant neoplasm of orbit* et *neoplasm of orbit* avec les têtes maximales. Dans la section suivante, nous évaluons ces mêmes groupements, mais dans le cadre de la validité des relations induites pour l'application visée (création des SMQ).

4.2. Évaluation par rapport à l'application

Au sein de l'ensemble de groupements générés avec les têtes maximales, l'étude de l'apport quantitatif des ressources et des approches exploitées montre que ce sont les relations hiérarchiques, quelles que soient les combinaisons avec les synonymes effectuées, qui sont impliquées dans la majorité des groupements (96 %) avec une précision de 69 %. Seuls 3 groupements (*Renovascular disorders ; Ovarian neoplasms, malignant and unspecified* et *Torsade de pointes/QT prolongation*) sont constitués uniquement avec des relations proposées par Faster ou des relations de synonymie. Les relations proposées par Faster interviennent dans seulement 50 % des groupements, mais avec une précision variant entre 75 % et 85 %. Un tiers des groupements est constitué à partir de relations de synonymie acquises à partir de ressources. La précision varie alors entre 55 % et 69 %. Les relations acquises à partir de l'UMLS contribuent le moins à la génération des groupements : environ 14 %, avec une précision de 38 % seulement. Ces observations montrent d'une part que notre méthode a permis de générer de très nombreuses relations hiérarchiques, alors que les termes traités proviennent tous du même niveau hiérarchique de MedDRA : niveau de termes préférés PT. Cela veut dire que la structure hiérarchique de la terminologie MedDRA pourrait être améliorée si des niveaux hiérarchiques intermédiaires y était ajoutés. D'autre part, nous observons une très forte participation de ces relations hiérarchiques dans la création de groupements et en plus avec une précision assez élevée (69 %). Par contre, les relations de synonymie sont plus rares et sont moins impliquées dans la génération des groupements. Cela veut dire que les termes PT de MedDRA sont assez

bien différenciés entre eux sur le plan conceptuel : il existe assez peu de doublons. Notons aussi que les relations de synonymie proposées par Faster montrent une meilleure précision que les relations de synonymie générées grâce au principe de compositionnalité. Il est possible que les ressources linguistiques exploitées comportent des ambiguïtés. Les relations induites avec WordNet ne sont pas impliquées dans la génération des groupements.

La figure 5 présente l'évaluation quantitative des groupements générés. Chaque graphique est construit selon différentes données de référence : 84 SMQ (figure 5(a)), 20 SMQ hiérarchiques (figure 5(b)) et 92 sous-SMQ (figure 5(c)). Chacun des graphiques est divisé en deux parties (par un impulse) : la partie droite correspond aux résultats générés avec le meilleur groupement (*m*) et la partie gauche aux résultats générés avec les *n* meilleurs groupements (*f*). Dans chaque partie nous présentons 14 expériences : tête minimale syntaxique (*min),* tête minimale syntaxique avec les relations hiérarchiques de Faster (*min-fsth*), tête minimale syntaxique avec les relations hiérarchiques de Faster et la combinaison des différents tests de synonymes à partir de : 3 terminologies biomédicales (*min-fsth-3t*), UMLS (*min-fsth-umls*), UMLS nettoyé (*min-fsth-umlsp*), fusion des différents tests de synonymes (*min-fsth-all*) et des synonymes de Faster (*min-fsth-fsts*). Ces expériences sont réitérées en utilisant la tête maximale syntaxique (*max*). Chaque graphique montre les résultats en termes de précision (en vert), de rappel (en rose) et de f-mesure (en rouge).



(a) SMQ *(b) SMQ hiérarchiques*

(c) sous-SMQ

Figure 5 : Résultats (précision, rappel et f-mesure) obtenus avec les différentes approches par rapport au matériel de référence (SMQ, SMQ hiérarchiques et sous-SMQ).

Globalement, nous pouvons observer que les meilleurs résultats en termes de f-mesure sont obtenus avec la fusion des n meilleurs groupements, alors que la meilleure précision est observée dans la première partie des graphiques (sans la fusion de *n* meilleurs groupements). L'apport des relations de synonymie montre une influence positive mais très faible.

Nous avons analysé en détail les groupements relatifs aux SMQ *Agranulocytosis* et *Angioedema* (avec fusion des *n* meilleurs groupements). Nous présentons les résultats de ces analyses pour l'approche qui fournit les meilleurs résultats : combinaison des relations obtenues avec les têtes syntaxiques maximales et Faster. Le SMQ *Agranulocytosis* contient 74 termes et le SMQ *Angioedema* contient 81 termes. La table 2 indique le nombre de termes dans le SMQ (*SMQ*) et dans le groupement correspondant (*gr*), ainsi que le nombre de termes communs (*com*) et les performances obtenues (précision *P*, rappel *R* et f-mesure *F*) avant (*Référence*) et après l'expertise réalisée par l'expert (*Après expertise*) pour les groupements relatifs à deux SMQ analysés. Nous avons ainsi analysé le bruit présent dans les groupements relatifs aux SMQ *Agranulocytosis* et *Angioedema*.

	Nombre de termes			Référence			Après expertise		
SMQ	SMQ	gr	com	P	R	F	P	R	F
Agranulocytosis	74	58	38	65	51	58	77	60	67
Angioedema	*81*	*89*	*52*	58	64	61	70	76	73

Table 2: Résultats (précision, rappel et f-mesure) obtenus avec la méthode de la tête syntaxique maximale combinés à Faster pour les deux SMQ: Agranulocytosis et Angioedema.

Le SMQ *Agranulocytosis* contient 74 termes. Les termes du SMQ Agranulocytosis concernent l'agranulocytose mais aussi les conséquences de l'agranulocytose. Les relations générées avec la tête syntaxique maximale combinée avec Faster fournissent 20 termes qui n'appartiennent pas au SMQ. Certains de ces termes devraient être inclus dans le SMQ parce qu'ils correspondent à d'autres conséquences d'agranulocytose (*Herpes sepsis, Candida sepsis, Fungal sepsis* et *Anthrax sepsis)* ou à des termes très proches des termes déjà inclus dans le SMQ (*Pseudomembranous colitis)*. D'autres termes ne sont pas des effets indésirables (*Idiopathic neutropenia, Chronic tonsillitis, Autoimmune neutropenia, Cyclic neutropenia* et *Autoimmune pancytopenia*) et sont de vrais faux positifs. Par exemple, dans le terme *Idiopathic neutropenia*, *idiopathic* signifie que la cause n'est pas médicamenteuse, et dans les termes *Cyclic neutropenia* et *Chronic tonsillitis* il s'agit de maladies cycliques ou chroniques or l'agranulocytose est un phénomène. Le filtrage de ces termes peut d'ailleurs être basé sur des marqueurs lexicaux comme *idiopathic, chronic, autoimmune* ou *cyclic*. Quelques autres termes (*Colitis, Phlebitis, Injection site phlebitis* et *Neutropenia*) sont des termes trop larges pour être inclus dans le SMQ. Finalement, quelques termes (*Eosinophilic colitis*, *Amoebic colitis* et *Allergic colitis)* ne sont pas liés à l'agranulocytose.

Le SMQ *Angioedema* contient 81 termes. Les termes de ce SMQ correspondent aux manifestations et symptômes de l'angiœdème. Notre méthode fournit 37 termes qui n'appartiennent pas au SMQ. Dix de ces termes (*Testicular swelling, Testicular oedema, Injection site joint swelling, Injection site urticaria, Injection site oedema, Cervix oedema, Retroperitoneal oedema, Bronchial oedema, Injection site hypersensitivity, Injection site swelling*) ont pour étiologie le

médicament et sont des symptômes de l'angiœdème et devraient donc être inclus dans le SMQ. Seize autre termes ne sont pas des allergies ou ne sont pas des conséquences médicamenteuses or l'angiœdème est une manifestation allergique : ce sont de vrais faux positifs. Quelques autres termes ne sont pas liés à l'angiœdème ou ont une signification trop large : ce sont aussi de vrais faux positifs.

Cette analyse détaillée montre que le raisonnement médical, qui se trouve à la base des SMQ, est assez complexe. Nous pensons que différents types d'approches devraient être utilisés et combinés pour reproduire au mieux ce raisonnement.

5. Conclusion et Perspectives

Le travail présenté exploite les méthodes de structuration de termes et permet de générer des groupements de termes MedDRA qui montrent une très bonne précision, ce qui correspond aux attentes de pharmacovigilants. Une analyse détaillée des résultats indique que certains des termes absents des SMQ pourraient y être également inclus. Avec une précision élevée, ces groupements peuvent être utilisés pour constituer des composantes qui alimentent la création des SMQ mais aussi pour affiner la structure des termes au sein des SMQ. Une première analyse effectuée montre que les résultats obtenus dans cette expérience sont complémentaires aux méthodes qui exploitent la distance sémantique (Dupuch et al, 2011). Nous prévoyons ainsi de combiner ces deux types de méthodes pour optimiser les groupements. Les groupements seront ensuite évalués au travers l'exploration des bases de données de pharmacovigilance. Les premiers résultats semblent indiquer que les groupements de termes PT obtenus avec nos méthodes sont plus efficaces pour la détection du signal que les SMQ ou les niveaux HLT de MedDRA utilisés plus traditionnellement. Nous allons également analyser les types de relations sémantiques qui existent entre les termes au sein des SMQ pour ouvrir des pistes vers d'autres méthodes de groupements de termes. Par exemple, nous prévoyons d'exploiter des corpus pour détecter des relations sémantiques transversales entre termes, comme par exemple les causes d'une pathologie ou les résultats d'examens biologiques anormaux et relevant d'une pathologie.

Références

I. Alecu, C. Bousquet, M.C. Jaulent. (2008) A case report: using snomed ct for grouping adverse drug reactions terms. *BMC Med Inform Decis Mak*, 8(S1):4.

S. Aubin, T. Hamon. (2006) Improving term extraction with terminological resources. In *FinTAL 2006*, number 4139 in LNAI, pages 380-387. Springer.

C. Bousquet, C. Henegar, A. Lillo-Le Louët, P. Degoulet, M.C. Jaulent. (2005) Implementation of automated signal generation in pharmacovigilance using a knowledge-based approach. *Int J Med Inform*, 74(7-8):563--71.

E.G. Brown, L. Wood, S. Wood. (1999) The medical dictionary for regulatory activities (MedDRA). *Drug Saf.*, 20(2):109—17.

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

M. Dupuch, M. Lerch, A. Jamet, M.C. Jaulent, R. Fescharek, N. Grabar. (2011) Grouping pharmacovigilance terms with semantic distance. In *MIE*.

C. Fellbaum. (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.

N. Grabar, T. Hamon. (2010) Exploitation of linguistic indicators for automatic weighting of synonyms induced within three biomedical terminologies. In *MEDINFO 2010*, pages 1015--9.

N. Grabar, P. Zweigenbaum. (2000) A general method for sifting linguistic knowledge from structured terminologies. *JAMIASUP,* pages 310-314.

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

T. Hamon, A. Nazarenko, C. Gros. (1998) A step towards the detection of semantic variants of terms in technical documents. In *COLING-ACL'98*, pages 498-504.

M. Hauben, A. Bate. (2009). Decision support methods for the detection of adverse events in post-marketing data. *Drug Discov Today*, 14(7-8), 343-57.

J. Iavindrasana, C. Bousquet, P. Degoulet, M.C. Jaulent. (2006) Clustering who-art terms using semantic distance and machine algorithms. In *AMIA Annu Symp Proc*, pages 369-73.

C. Jacquemin, J.L. Klavans, E. Tzoukerman. (1997) Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *ACL/EACL 97*, pages 24-31, Barcelona, Spain.

M.C. Jaulent, I. Alecu. (2009) Evaluation of an ontological resource for pharmacovigilance. In *Stud Health Technol Inform*, pages 522—6.

R. Meyboom, M. Lindquist, A. Egberts, I. Edwards. (2002). Signal selection and follow-up in pharmacovigilance. *Drug Saf*, 25(6), 459-65.

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

B.H. Partee (1984) *Compositionality*. F Landman and F Veltman.

R.K. Pearson, M. Hauben, D.I. Goldsmith, A.L. Gould, D. Madigan, D.J. O'Hara, S.J. Reisinger, A.M. Hochberg. (2009) Influence of the meddra hierarchy on pharmacovigilance data mining results. *Int J Med Inform*, 78(12):97-103.

Y. Tsuruoka, Y. Tateishi, J.D. Kim, T. Ohta, J. McNaught, S. Ananiadou, J. Tsujii. (2005) Developing a robust part-of-speech tagger for biomedical text. *LNCS,* 3746:382-392.

C.M. Verspoor, C. Joslyn, G.J. Papcun. (2003) The Gene Ontology as a source of lexical semantic knowledge for a biological natural language processing application. In *SIGIR workshop on Text Analysis and Search for Bioinformatics*, pages 51-56.