Étude des verbes introducteurs de noms de médicaments dans les forums de santé

François Morlane-Hondère[1] Cyril Grouin[1] Pierre Zweigenbaum[1]
(1) LIMSI–CNRS, UPR 3251, rue John von Neumann, 91400 Orsay
{prenom.nom}@limsi.fr

Résumé. Dans cet article, nous combinons annotations manuelle et automatique pour identifier les verbes utilisés pour introduire un médicament dans les messages sur les forums de santé. Cette information est notamment utile pour identifier la relation entre un médicament et un effet secondaire. La mention d'un médicament dans un message ne garantit pas que l'utilisateur a pris ce traitement mais qu'il effectue un retour. Nous montrons ensuite que ces verbes peuvent servir pour extraire automatiquement des variantes de noms de médicaments. Nous estimons que l'analyse de ces variantes pourrait permettre de modéliser les erreurs faites par les usagers des forums lorsqu'ils écrivent les noms de médicaments, et améliorer en conséquence les systèmes de recherche d'information.

Abstract.

Study of Drug-Introducing Verbs on Health Forums

In this paper, we combine manual/automatic annotation to identify the verbs used by the users of a health forum to say that they are taking a drug. This information is important in many aspects, one of them being the identification of the relation between drugs and side effects : the mere mention of a drug in a message is not enough to assess that the user is taking this drug, and is thus likely to provide a feedback on it. In a second part of the study, we show how the set of verbs that we identified can be used to automatically extract variants of drug names. We assume that the analysis of the variants could shed light on patterns of mistakes that users make when spelling drug names and thus, improve medical information retrieval systems.

Mots-clés : contenu généré par l'utilisateur, forum, verbes, noms de médicaments.

Keywords: user-generated content, forum, verbs, drug names.

1 Introduction et état de l'art

Cette étude s'inscrit dans le cadre d'un projet visant l'identification d'effets secondaires de médicaments dans des textes produits par des utilisateurs sur des forums de discussion. Il s'agit d'une problématique récente née du besoin de surveiller les médicaments après leur mise sur le marché et du développement des forums de santé en ligne.

L'autorisation de mise sur le marché d'un médicament est soumise à une batterie d'essais cliniques qui permettent d'évaluer le rapport entre effets bénéfiques et les éventuels effets néfastes. Si ce rapport est jugé acceptable, le médicament est commercialisé et les effets indésirables sont documentés dans sa notice. Toutefois, certains effets indésirables échappent aux tests cliniques et, dans les cas les plus graves, peuvent conduire à des crises sanitaires comme l'*affaire du Mediator* en France, qui n'était pas qu'une affaire politique.

En 2007, l'étude EMIR (Effets indésirables des Médicaments : Incidence et Risque) menée par le réseau des centres régionaux de pharmacovigilance a estimé que 3,6 % des admissions dans des hôpitaux français étaient dues à des effets indésirables de médicaments, soit 1 480 885 journées d'hospitalisation par an [1]. Ces chiffres témoignent des enjeux humains et financiers considérables liés au processus de pharmacovigilance, qui consiste à détecter, identifier et gérer les effets indésirables de médicaments après leur commercialisation.

Les dispositifs mis en œuvre par les organismes de pharmacovigilance et les laboratoires pour permettre aux praticiens et aux patients de rapporter les effets indésirables se composent principalement de centres d'appel et de formulaires en

1. http://www.sante.gouv.fr/IMG/pdf/EMIR.pdf

ligne [2]. Un rapport de 2014 de l'Académie nationale de Pharmacie montre toutefois que seuls 4 à 5 % des effets indésirables sont signalés de façon spontanée [3]. Cela est principalement dû à la méconnaissance des dispositifs de signalement de la part des utilisateurs de médicaments et professionnels de santé (Alshakka *et al.*, 2013).

Avec la mise au format électronique des données de santé, l'utilisation de méthodes d'extraction d'information apparaît alors comme une approche intéressante pour identifier les effets indésirables de médicaments par l'analyse de dossiers de patients (Trifirò *et al.*, 2009; Wang *et al.*, 2009; Gurulingappa *et al.*, 2012). Les résultats encourageants fournis par ces méthodes ont conduit à les appliquer sur les données produites par les utilisateurs de médicaments eux-mêmes, principalement sur les réseaux sociaux et forums de discussion (voir Sarker *et al.* (2015) pour un état de l'art). Ces données présentent plusieurs avantages par rapport à celles produites par des professionnels de santé dans le milieu médical : (i) elles sont facilement accessibles, (ii) elles sont massives, (iii) elles sont produites en continu et peuvent être soumises à un système de pharmacovigilance dès leur mise en ligne, ce qui permet une réactivité inédite (Egberts *et al.* (1996) rapportent un délai de 229 jours entre le signalement par téléphone et la production d'un rapport sur les effets indésirables sur la paroxétine par la Pharmacovigilance Foundation), et (iv) elles sont volontairement mises à disposition par les utilisateurs à destination de la communauté, ce qui facilite la question de leur confidentialité. Le principal inconvénient est lié au caractère non contrôlé des données, ce qui entraîne une variabilité orthographique et stylistique. Ceci complique l'utilisation d'outils d'analyse morpho-syntaxique traditionnels et la projection de lexiques d'entités (Nikfarjam & Gonzalez, 2011). Nous relevons par exemple dans nos données plus de neuf orthographes différentes pour le mot *anxiolytique* ainsi que de nombreuses classes d'équivalences comme *être agité/survolé/comme une pile électrique*...

Cette étude prospective, fondée sur une analyse manuelle, est envisagée comme une première étape dans l'identification automatique d'effets indésirables. Elle vise à donner des éléments de description linguistique de la façon dont les utilisateurs de forums de santé expriment le fait qu'ils prennent un médicament par l'analyse des verbes employés. Comme mis en évidence par Leaman *et al.* (2010), les verbes constituent des indices forts pour le repérage d'informations médicales (indications, effects bénéfiques ou indésirables...). Les verbes qui ont pour rôle d'introduire des noms de médicaments présentent donc un intérêt pour identifier des effets indésirables dans les textes : en plus d'indiquer la prise d'un médicament, ils donnent des indices sur la nature de ce médicament. Nous montrons également que ces verbes peuvent servir à identifier des variantes de noms de médicaments.

2 Méthode et données

La mention de l'utilisation d'un médicament peut être modélisée comme une construction composée de trois éléments : (i) un nom de médicament, (ii) un verbe introducteur de médicament (VIM) ; ce type de verbe est indispensable pour distinguer les cas où un utilisateur témoigne de son expérience avec un médicament ou demande des renseignements en vue de le prendre (Wu *et al.*, 2013), et (iii) un pronom de première personne singulier. Bien que seuls le verbe et l'objet soient suffisants pour identifier la mention d'une prise de médicament, nous restreignons le sujet au scripteur lui-même dans le but d'écarter les cas de témoignages indirects. Des modalités comme le dosage ou le rythme de prise du médicament peuvent se greffer à cette construction mais ne font pas partie de ses éléments essentiels. Puisqu'il existe des listes des noms de médicaments et que les pronoms de première personne singuliers se limitent à *je* et *me*, les VIMs sont les seuls éléments inconnus de cette construction. Cette section présente les approches manuelle puis automatique que nous avons adoptées pour les identifier, puis une expérience visant à montrer leur pertinence pour identifier les variantes de noms de médicaments.

2.1 Annotation manuelle des VIMs

Dans une approche préliminaire, nous avons manuellement annoté les VIMs dans un corpus de 11 735 mots constitué de messages extraits du forum de santé francophone Doctissimo [4]. Nous avons identifié douze verbes, dont dix ont une fréquence inférieure à 5. Afin d'estimer la productivité des verbes identifiés sur un plus gros volume de données, nous avons utilisé le moteur de recherche Google pour accéder à l'ensemble des données du forum Doctissimo. Nous avons construit 91 requêtes constituées d'un sous-ensemble de sept verbes parmi les douze identifiés précédemment et de treize noms de médicaments choisis parmi les plus prescrits en France. Ces requêtes sont construites selon les modalités suivantes :

2. https://www.formulaires.modernisation.gouv.fr/gf/cerfa_15031.do
3. http://www.acadpharm.org/dos_public/GTNotif_Patients_Rap_VF__2015.01.22.pdf
4. http://forum.doctissimo.fr

– le verbe est à la première personne du singulier du présent de l'indicatif, sauf dans le cas de *prescrire* et *donner*, qui sont au participe passé (l'utilisateur est celui à qui l'on donne un médicament) ;
– le verbe et le nom de médicament sont séparés par un astérisque, qui remplace un ou plusieurs mots dans la syntaxe de requêtes de Google ;
– la requête est encadrée de guillemets pour que l'ordre des éléments se retrouve à l'identique dans les textes.

Ces 91 requêtes ont été soumises manuellement à Google avec le paramètre `site:forum.doctissimo.fr` pour restreindre la recherche à tous les forums de Doctissimo. Bien que cette expérience ait fourni des résultats intéressants, le protocole mis en place reste limité (i) par le nombre de contextes et de médicaments testés, (ii) par l'utilisation de requêtes Google, qui ne permettent pas de prendre en compte les formes fléchies des verbes et (iii) par l'utilisation du nombre de pages retournées comme un substitut de la fréquence.

2.2 Extraction semi-automatique des VIMs

Cette deuxième approche vise à étendre la liste des douze VIMs identifiés manuellement par la projection de patrons sur un corpus de 17,5 millions de mots constitué de messages extraits du forum médical Atoute.org [5]. Elle consiste à chercher les VIMs dans les contextes dont nous faisons l'hypothèse qu'ils sont les plus susceptibles d'apparaître, à savoir entre un pronom de première personne singulier et un nom de médicament.

Cette approche se divise en trois étapes :

1. une liste de noms de médicaments est utilisée pour annoter automatiquement les noms de médicaments qui apparaissent dans les messages. Cette liste est composée de 4 ressources distinctes. La construction de cette liste prend en compte les versions accentuées et désaccentuées et ne tient pas compte de la casse. Cette liste contient :
   – les 8691 entités en français de l'UMLS appartenant au type sémantique *Pharmacologic substance* ;
   – 9064 noms de médicaments génériques fournis par l'Agence nationale de sécurité du médicament et des produits de santé (ANSM) [6] ;
   – 10 870 noms de médicaments extraits du dictionnaire de médicaments en ligne EurekaSanté [7] ;
   – la liste des 100 médicaments les plus prescrits en France [8].
2. les séquences de n ($n \leq 6$) mots qui figurent entre un pronom de première personne singulier et un nom de médicament sont extraites ;
3. les verbes figurant dans ces séquences sont identifiés et lemmatisés par la projection de Glàff [9], un lexique généraliste du français. Bien que fruste, cette approche se justifie par la nature de nos données, qui complique l'utilisation d'un étiqueteur morpho-syntaxique. Cette méthode reste néanmoins très bruitée.

La pertinence des verbes candidats identifiés a été évaluée manuellement.

2.3 Utiliser les VIMs pour extraire les variantes de noms de médicaments

Parce que les textes produits sur les forums sont non contrôlés et que les noms de traitements sont complexes, il est fréquent que les noms de médicaments soient mal orthographiés ou abrégés (Pimpalkhute *et al.*, 2014). Ces orthographes alternatives ne figurent pas dans les lexiques et posent un problème pour les systèmes d'extraction d'information.

Dans cette expérience, nous illustrons une utilisation possible des VIMs pour extraire des variantes de noms de médicaments. Le protocole utilisé est une déclinaison de celui décrit en section 2.2 : au lieu de chercher les verbes situés entre un pronom et un nom de médicament, nous ciblons les mots inconnus – définis comme n'apparaissant ni dans notre lexique de médicaments ni dans Glàff – situés après une séquence pronom+VIM. Nous faisons l'hypothèse que les VIMs sont quasi-systématiquement suivis de noms de médicaments. En conséquence, un mot inconnu qui apparaît dans une séquence de 1 à 6 mots après un VIM a plus de chances d'être la variante d'un nom de médicament que les mots inconnus qui apparaissent ailleurs dans le texte.

5. http://www.atoute.org/n/forum/
6. http://ansm.sante.fr/var/ansm_site/storage/original/text/97b3c42da571c69da1e837f759076675.txt
7. http://www.eurekasante.fr/medicaments/alphabetique.html
8. http://www.doctissimo.fr/asp/medicaments/les-medicaments-les-plus-prescrits.htm
9. http://redac.univ-tlse2.fr/lexiques/glaff.html

3 Résultats

L'étape d'annotation manuelle nous a permis d'identifier les douze VIMs suivants (la fréquence du verbe apparaît entre parenthèses) : *prendre* (37), *prescrire* (19), *être sous* (4), *passer* (4), *donner* (3), *commencer* (3), *avaler* (1), *absorber* (1), *entamer* (1), *suivre* (1), *tester* (1) et *utiliser* (1). Le nombre de pages Web indiqué par Google après la soumission des requêtes contenant 7 de ces VIMs et 13 noms de médicaments est rapporté au tableau 1. La proportion de pages rapportées pour chaque VIM et chaque contexte est fournie dans le tableau 2

	Doliprane	Levothyrox	Kardégic	Spasfon	Tahor	Voltarène	Forlax	Subutex	Gaviscon	Lexomil	Lysanxia	Atarax	Xanax	total
prendre	6350	1550	71	35700	55	36	319	56	4570	806	576	365	1340	51 794
prescrire	2070	334	27	7270	23	64	510	35	5040	750	426	824	1570	18 943
donner	3730	86	17	8570	12	34	619	8	1560	414	74	1190	673	16 987
être sous	80	1340	202	6390	24	30	47	75	61	536	177	93	811	9866
avaler	1787	0	0	35	0	1	0	0	12	9	2	5	20	1871
commencer	8	89	5	28	1	0	18	10	63	12	6	4	26	270
passer	35	21	5	33	4	3	15	6	37	1	7	1	18	186
total	16 600	5080	327	58 026	119	168	1528	190	11 343	2528	1268	2482	4458	

TABLE 1 – Nombre de pages rapportées par Google pour chaque requête.

	Doliprane	Levothyrox	Kardégic	Spasfon	Tahor	Voltarène	Forlax	Subutex	Gaviscon	Lexomil	Lysanxia	Atarax	Xanax	moyenne
prendre	38.3	30.5	20.7	58.4	40.4	20.1	20.6	22.3	40.1	31.1	43	14.6	25.5	31.2
prescrire	12.5	6.6	7.9	11.9	16.9	35.8	32.9	13.9	44.3	28.9	31.8	32.9	29.9	23.6
être sous	0.5	26.4	58.9	10.4	17.6	16.8	3	29.9	0.5	20.7	13.2	3.7	15.4	16.7
donner	22.5	1.7	5	14	8.8	19	40	3.2	13.7	16	5.5	47.5	12.8	16.1
avaler	10.8	0	0	0.1	0	0.6	0	0	0.1	0.3	0.1	0.2	0.4	1.0
passer	0.2	0.4	1.5	0.1	2.9	1.7	1	2.4	0.3	0	0.5	0	0.3	0.9
commencer	0	1.8	1.5	0	0.7	0	1.2	4	0.6	0.5	0.4	0.2	0.5	0.9

TABLE 2 – Proportion de pages rapportées pour chaque VIM et chaque nom de médicament.

La méthode d'extraction de VIMs (cf. section 2.2) nous permet d'identifier 28 934 occurrences de 1053 verbes candidats parmi lesquels 44 ont été manuellement identifiés comme des VIMs. Le tableau 3 présente ces verbes et leur fréquence.

freq.	verb	freq.	verb	freq.	verb	freq.	verb
2953	*prendre*	37	*passer*	13	*manger*	5	*repasser*
1570	*prescrire*	33	*continuer*	12	*avaler*	4	*vacciner*
842	*être sous*	31	*appliquer*	11	*(se) soigner*	4	*refiler*
764	*donner*	25	*tester*	10	*consommer*	4	*recommencer*
400	*avoir*	24	*injecter*	9	*rajouter*	2	*sniffer*
296	*mettre*	20	*remettre*	9	*administrer*	2	*(se) droguer*
181	*commencer*	16	*tenter*	8	*(se) gaver*	2	*effectuer*
140	*essayer*	15	*represcrire*	7	*(se) badigeonner*	2	*absorber*
133	*utiliser*	15	*filer*	7	*baisser*	1	*bouffer*
122	*reprendre*	15	*diminuer*	6	*boire*	1	*bénéficier*
115	*suivre*	15	*augmenter*	5	*sucer*	1	*abuser*

TABLE 3 – Fréquence des 44 VIMs identifiés dans le corpus Atoute.org.

4 Discussion

Emploi des verbes La première approche consistant à annoter manuellement les VIMs dans un petit corpus a montré que 61 % des mentions de prise d'un médicament se font en employant les verbes *prendre* et *prescrire* (bien que ce verbe n'exprime pas la prise d'un médicament, les données montrent que cette dernière est très souvent sous-entendue). Les résultats fournis par l'approche basée sur la construction de requêtes Google montrent que les verbes *prendre*, *prescrire*, *donner* et *être sous* prévalent aussi bien en terme de nombre de pages rapportées que de proportion : sur l'ensemble des pages rapportées par le moteur de recherche, 98 % l'ont été par une requête contenant l'un de ces quatre VIMs. On constate également que le type de verbe employé pour évoquer la prise d'un médicament varie en fonction du médicament :

– Lexomil, Lysanxia, Atarax et Xanax, qui appartiennent à la classe des benzodiazépines, sont souvent employés avec le verbe *prescrire*. Ce n'est pas le cas de Levothyrox et Kardégic, ce qui peut s'expliquer par le fait que ce sont des traitements à vie : on peut alors supposer que l'action de prescrire ce type de traitement est moins fréquente que pour d'autres médicaments, et que cela se répercute dans les messages ;
– Doliprane, Gaviscon et Forlax ne s'emploient que très rarement avec *être sous*. On peut supposer que cela est dû au fait que ces médicaments s'utilisent de façon ponctuelle pour traiter des affections passagères. Le fait que Levothyrox et – surtout – Kardégic sont souvent introduits par cette locution verbale va dans le sens de cette hypothèse ;
– le verbe *avaler* s'emploie quasi-exclusivement avec Doliprane. L'emploi de ce verbe relativement familier est peut-être à mettre en relation avec le fait que le Doliprane est un médicament très répandu et considéré comme inoffensif, quand bien même aucun médicament n'est inoffensif. Ce VIM est également particulier en cela qu'il contient une information sur la façon dont le médicament est pris, donc sur sa forme galénique : il paraît improbable qu'un médicament sous forme de crème soit introduit par ce VIM.

Ces résultats mettent ainsi en lumière certaines restrictions distributionnelles imposées par les VIMs, qui sélectionnent certains types de médicaments en fonction de propriétés liées à leur fréquence de prise ou leur forme galénique. Comme le montre le tableau 3, la prévalence des 4 VIMs identifiés manuellement est confirmée par les données obtenues sur le corpus Atoute.org. De nouveaux VIMs particulièrement fréquents ont pu être identifiés, comme *mettre*, qui s'emploie soit avec un nom de médicament – principalement une lotion ou un crème – en COD, soit avec la préposition *sous*. La haute fréquence du verbe *avoir* est une erreur due à la méthode d'identification des verbes employée. Du fait de son caractère hautement polysémique, il apparaît dans une grande variété de contextes, mais n'est que rarement un VIM.

Informations complémentaires Une des caractéristiques des VIMs les plus fréquents est qu'ils sont relativement *neutres* : le fait qu'ils imposent peu de restrictions sélectionnelles sur le type de médicament qu'ils prennent comme objets entraine leur utilisation pour l'introduction d'une grande variété de médicaments, ce qui explique leur fréquence. À l'inverse, d'autres VIMs véhiculent une connotation qui réduit le spectre des médicaments avec lesquels ils peuvent s'employer, et donc leur fréquence générale dans les textes. On peut distinguer 4 groupes parmi ces VIMs : (i) *commencer*, *essayer*, *passer*, *continuer*, *tester*, *tenter*, *repasser*, *recommencer* fournissent des informations temporelles sur le traitement ; (ii) *diminuer*, *augmenter*, *baisser* indiquent l'évolution du dosage ; (iii) *manger*, *(se) gaver*, *(se) droguer*, *bouffer*, *abuser* apportent des précisions sur la quantité et la fréquence de prise du médicament. Ils renseignent également du point de vue du scripteur qui porte un jugement négatif sur la quantité de médicaments prise, jugée excessive ; (iv) *injecter*, *avaler*, *(se) badigeonner*, *boire*, *sucer*, *sniffer* informent sur la forme galénique du médicament. En plus d'indiquer la prise d'un médicament, ces VIMs apportent des informations complémentaires utiles pour identifier des effets indésirables.

Identification de variantes Un des problèmes liés aux données générées par des utilisateurs est la fréquence d'entités nommées mal orthographiées. Il est possible d'identifier ces variantes en rapprochant leur différentes occurences sur la base de leur structure phonétique ou à l'aide de la distance d'édition. De manière similaire aux techniques de clustering non supervisées qui se fondent sur l'analyse du contexte, nous proposons ici d'utiliser les VIMs en les projetant sur le corpus Atoute.org dans le but de recueillir des variantes de noms de médicaments. Cette méthode nous a permis de recueillir 5 638 occurrences de 2 769 candidats. Ces derniers peuvent être classés dans cinq catégories :

– nom de médicament
  1. orthographe officielle : bien que correctement orthographiés, ces noms n'apparaissent pas dans notre liste. Nos lexiques ne contiennent aucun nom de produits paramédicaux (tels que les compléments alimentaires) et se limitent aux médicaments prescrits en France (certains médicaments sont évoqués sous leur nom commercial canadien par des utilisateurs québécois) ;
– orthographe non officielle
  2. variante intentionnelle : nous considérons comme des variantes intentionnelles les mots dérivés des noms offi-

ciels par des procédés morphologiques tels que l'apocope (*lévo* pour *Lévothyrox*), la réduplication (*dudu* pour *Duphaston*) ou encore la siglaison (*pdl* pour *pilule du lendemain*) ;

3. variante fautive : nous considérons comme des fautes les variantes de noms de médicaments qui ne semblent pas résulter d'un procédé de formation morphologique volontaire comme ceux évoqués précédemment. Nous avons distingué ces variantes selon les quatre opérations utilisées pour mesurer la distance de Damerau-Levenshtein (cf. tableau 4). Les erreurs d'orthographe n'ont pas été distinguées des éventuelles fautes de frappe ;

– autre type de nom

4. domaine médical : mots du lexique médical mal orthographiés ou n'apparaissant pas dans nos lexiques (variantes *amniosynthese* et *amnio* pour *amniocentèse*) ;
5. domaine non médical : mots du vocabulaire général mal orthographiés ou n'apparaissant pas dans Glàff.

effacement	substitution	insertion	inversion	combinaison
alocardyl (Avlocardyl)	allupirinol (Allopurinol)	corguard (Corgard)	pantesa (Pentasa)	calements (calmants)
lévotyrox (Lévothyrox)	anxiolitique (anxiolytique)	cortisonne (cortisone)	procolaran (Procoralan)	cylbatant (Cymbalta ?)
luthenyl (Lutényl)	anxyolitique (anxiolytique)	dacktarin (Daktarin)	steridil (Stediril)	rhénomicine (Rinomicine)
stomectol (Stromectol)	celccept (Cellcept)	duphastion (Duphaston)		doxycilline (Doxycycline)
utrogestant (Utrogestan)	cortencil (Cortancyl)	endoxant (Endoxan)		methojet (Metoject)

TABLE 4 – Les différentes catégories de variantes fautives (les noms officiels sont entre parenthèses). La colonne `combinaison` contient des variations produites par la combinaisons de plusieurs opérations.

La distinction des variantes extraites permet d'observer que les variantes intentionnelles relèvent de procédés de création réguliers, contrairement aux variantes fautives. Parmi les variantes fautives, nous observons des tendances telles que les substitutions récurrentes *i/y* ou *s/z*, le rajout d'un *t* aux noms de médicaments terminés par la séquence *-an*, ou encore des confusions dans les noms contenant des doubles lettres (*Xyzaal* pour *Xyzall*). Bien qu'une analyse phonétique permettrait d'identifier la majorité de ces variantes, il reste des cas pour lesquels une analyse plus poussée demeure nécessaire (*steridil* vs *Stediril*). En conséquence, nous estimons que l'analyse des procédés de formation des variantes intentionnelles et fautives pourrait permettre de prédire les erreurs faites par les utilisateurs et ainsi d'améliorer les performances de systèmes d'identification des effets indésirables de médicaments.

5 Conclusion

Dans cet article, nous avons présenté l'étude réalisée pour identifier un ensemble de verbes introduisant des noms de médicaments dans un corpus de messages postés sur un forum de santé. Nous avons mis en évidence que les verbes sont utilisés différemment selon le médicament qu'ils introduisent. Nous avons montré que cet ensemble de verbes peut être utilisé dans des règles pour extraire automatiquement des variantes de noms de médicaments. Enfin, nous estimons que ces propriétés sont utiles pour identifier certaines catégories de médicaments, ou pour extraire des relations entre médicaments et effets secondaires.

Remerciements

Ce travail a été réalisé dans le cadre du projet Vigi4MED (ANSM-2013-S-060), financé par l'ANSM (Agence Nationale de Sécurité du Médicament).

Références

ALSHAKKA M. A., IBRAHIM M. I. M. & HASSALI M. A. A. (2013). Do health professionals have positive perception towards consumer reporting of adverse drug reactions ? *J Clin Diagn Res*, **7**(10), 2181–5.

EGBERTS T. C., SMULDERS M., DE KONING F. H., MEYBOOM R. H. & LEUFLENS H. G. (1996). Can adverse drug reactions be detected earlier ? a comparison of reports by patients and professionals. *Br Med J*, **313**(7056), 530–1.

GURULINGAPPA H., MATEEN-RAJPUT A. & TOLDO L. (2012). Extraction of potential adverse drug events from medical case reports. *Journal of Biomedical Semantics*, **3**(1).

LEAMAN R., WOJTULEWICZ L., SULLIVAN R., SKARIAH A., YANG J. & GONZALEZ G. (2010). Towards internet-age pharmacovigilance : Extracting adverse drug reactions from user posts to health-related social networks. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, BioNLP '10, p. 117–125, Stroudsburg, PA, USA : Association for Computational Linguistics.

NIKFARJAM A. & GONZALEZ G. H. (2011). Pattern mining for extraction of mentions of adverse drug reactions from user comments. *AMIA Annual Symposium Proceedings*, **2011**, 1019.

PIMPALKHUTE P., PATKI A., NIKFARJAM A. & GONZALEZ G. (2014). Phonetic spelling filter for keyword selection in drug mention mining from social media. *AMIA Summits on Translational Science Proceedings*, **2014**, 90.

SARKER A., NIKFARJAM A., O'CONNOR K., GINN R., GONZALEZ G., UPADHAYA T., JAYARAMAN S. & SMITH K. (2015). Utilizing social media data for pharmacovigilance : A review. *Journal of Biomedical Informatics*, (0), –.

TRIFIRÒ G., PARIENTE A., COLOMA P. M., KORS J. A., POLIMENI G., MIREMONT-SALAMÉ G., CATANIA M. A. A., SALVO F., DAVID A., MOORE N., CAPUTI A. P. P., STURKENBOOM M., MOLOKHIA M., HIPPISLEY-COX J., ACEDO C. D. D., VAN DER LEI J., FOURRIER-REGLAT A. & EU-ADR GROUP (2009). Data mining on electronic health record databases for signal detection in pharmacovigilance : which events to monitor ? *Pharmacoepidemiology and drug safety*, **18**(12), 1176–1184.

WANG X., HRIPCSAK G., MARKATOU M. & FRIEDMAN C. (2009). Research paper : Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records : A feasibility study. *JAMIA*, **16**(3), 328–337.

WU H., FANG H. & STANHOPE S. J. (2013). Exploiting online discussions to discover unrecognized drug side effects. *Methods Inf Med*, **52**(2), 152–9.