Médicaments qui soignent, médicaments qui rendent malades : étude des relations causales pour identifier les effets secondaires

François Morlane-Hondère, Cyril Grouin, Véronique Moriceau, Pierre Zweigenbaum

Médicaments qui soignent, médicaments qui rendent malades : étude des relations causales pour identifier les effets secondaires

François Morlane-Hondère[1] Cyril Grouin[1] Véronique Moriceau[1,2] Pierre Zweigenbaum[1]
(1) LIMSI–CNRS, UPR 3251, rue John von Neumann, 91400 Orsay
(2) Université Paris-Sud, Campus universitaire d'Orsay, 91400 Orsay
{prenom.nom}@limsi.fr

Résumé. Dans cet article, nous nous intéressons à la manière dont sont exprimés les liens qui existent entre un traitement médical et un effet secondaire. Parce que les patients se tournent en priorité vers internet, nous fondons cette étude sur un corpus annoté de messages issus de forums de santé en français. L'objectif de ce travail consiste à mettre en évidence des éléments linguistiques (connecteurs logiques et expressions temporelles) qui pourraient être utiles pour des systèmes automatiques de repérage des effets secondaires. Nous observons que les modalités d'écriture sur les forums ne permettent pas de se fonder sur les expressions temporelles. En revanche, les connecteurs logiques semblent utiles pour identifier les effets secondaires.

Abstract.

Drugs that cure, drugs that make you sick : study of causal links to identify drug side effects
In this paper, we study the textual manifestations of the relation between drugs and side effects in online health forums. Our goal is to find relevant linguistic cues in order to improve the automatic identification of side effects by leveraging the ambiguity between actual side effects and indications (the reason for drug use). We find that the use of discourse markers can be relevant for the identification of indications – a third of indication mentions follow markers like 'pour' ('for') or 'dans le but de' ('with the aim of') – while temporal informations are not as discriminating.

Mots-clés : Pharmacovigilance, forums de santé, relations causales.

Keywords: Pharmacovigilance, Health Forums, Causal Links.

1 Introduction

Selon l'Organisation Mondiale de la Santé [1] (OMS), un effet secondaire est une réaction inattendue due à un traitement médical. Bien que des tests cliniques soient réalisés en laboratoire avant la commercialisation des médicaments, il est difficile de prévoir l'ensemble des effets secondaires d'un traitement pendant cette phase de tests, et ce, pour diverses raisons : durée limitée, différences entre patients, modifications des spécifications d'un traitement après les tests cliniques (Megahed, 2014). Il est alors nécessaire de réaliser une veille pharmacologique après l'autorisation de mise sur le marché d'un traitement. Parce qu'il est inattendu, un effet secondaire est généralement négatif (« effet indésirable »), mais peut se révéler positif (le Baclofène, initialement autorisé pour le traitement de troubles musculaires, permet le traitement de l'alcoolisme).

L'identification automatique d'effets indésirables de médicaments est une problématique récente. Alors que seuls 4 à 5% des effets indésirables sont rapportés de façon spontanée [2] auprès des centres de pharmacovigilance, il est nécessaire d'exploiter d'autres sources d'information, telles que les réseaux sociaux, vers lesquels les patients se tournent désormais pour obtenir des informations. De nombreuses études ont ainsi porté sur l'utilisation des réseaux sociaux pour réaliser une veille épidémiologique (Velardi *et al.*, 2014), suivre les conséquences d'un problème environnemental (Cha & Stow, 2015), ou à des fins de pharmacovigilance (Sampathkumar *et al.*, 2014). Les données issues des réseaux sociaux présentent de nombreux avantages en raison de leur accessibilité et de leur caractère massif. Le fait que ces données sont produites en continu constitue également un intérêt pour le processus de pharmacovigilance, qui nécessite une grande réactivité.

1. http://www.who.int/fr/
2. http://www.acadpharm.org/dos_public/GTNotif_Patients_Rap_VF__2015.01.22.pdf

2 État de l'art

Les systèmes développés s'appuient – au moins en partie – sur des lexiques d'effets indésirables (Wang *et al.*, 2009; Sarker *et al.*, 2015). Ces lexiques sont soit projetés directement sur les textes, soit utilisés comme des traits dans un système d'apprentissage automatique. La première approche montre toutefois ses limites face à la variabilité orthographique et stylistique qui caractérise les textes – par nature non contrôlés – issus de forums de discussion en ligne ou de réseaux sociaux. L'entraînement d'un système d'apprentissage automatique à partir d'un corpus annoté manuellement est une autre technique qui consiste à construire un modèle dans lequel les entités à extraire sont caractérisées par un ensemble de traits. Ces derniers portent aussi bien sur l'entité elle-même (présence de majuscules, de chiffres, de certains types de suffixes…) que sur son contexte d'apparition dans le corpus (n-grammes de mots, de parties du discours…).

L'un des problèmes rencontrés lors du repérage automatique d'effets secondaires [3] est que ces effets, ou *événements*, se confondent avec un autre type d'entité, les *indications* (Nikfarjam *et al.*, 2015). Alors que les événements sont des symptômes ressentis après la prise d'un médicament (exemple 1), les indications sont les raisons pour lesquelles le médicament a été pris (exemple 2).

(1) J'ai avalé le médoc a 12h, j'ai mangé et a 12h30 grosse crampes dans le ventre et brusque gonflement suivi d'un urticaire géant...

(2) En faite c mon endo qui me la prescrit je n'arrivais pas a perdre de poids suite a des pb endocri.

Le fait qu'un symptôme peut souvent aussi bien constituer une indication qu'un événement complique la tâche d'identification des événements et génère du bruit.

Megahed (2014) montre que, du fait de cette ambiguïté, les entités appartenant aux classes "événement" et "indication" sont moins correctement classifiées que d'autres entités non ambiguës comme les noms de traitements. Il ressort également de cette étude que l'importance du phénomène varie en fonction du corpus étudié : le recouvrement entre les entités appartenant à la classe des événements et à celle des indications est plus important dans un corpus composé de messages portant sur la thématique des anti-dépresseurs que sur celle de la migraine. Sarker *et al.* (2015) ont également mis en lumière l'importance de ce phénomène. Ils montrent que le problème de l'ambiguïté entre indication et événement intervient dans 60 % des faux positifs produits par leur système et suggèrent d'analyser la polarité du contexte (les événements auront plus de chances d'être associés à des contextes qui portent une polarité négative que les indications).

L'étude que nous présentons ici se situe en aval d'une annotation préalable : nous partons d'un corpus annoté en entités pertinentes pour le domaine médical et nous cherchons à identifier celles qui relèvent de la cause ou de la conséquence de la prise d'un traitement médicamenteux. Pour ce faire, nous nous proposons d'étudier la pertinence pour ce type de tâche des indices que sont les connecteurs logiques et les marques temporelles. Ces indices ont été utilisés dans de précédentes études. Segura-Bedmar *et al.* (2011) mobilisent ainsi les connecteurs linguistiques pour identifier les interactions entre médicaments tandis que Sun *et al.* (2013) ont souligné l'utilité des éléments temporels pour typer des relations entre concepts médicaux.

3 Corpus

3.1 Constitution

Nous avons limité la thématique abordée dans le corpus au Médiator, pour tester et valider notre méthode d'identification des effets secondaires à un premier traitement. Le Médiator est un hypoglycémiant qui permet de lutter contre les glycémies excessives chez les diabétiques [4], dont l'usage a été détourné pour permettre la perte de poids chez les personnes non diabétiques. Le choix du Médiator s'appuie sur le fait que les effets secondaires de ce traitement sont connus et documentés (les plus courants concernent des troubles digestifs, de la fatigue et des vertiges), et qu'il est possible d'étudier la manière dont sont exprimés les problèmes rencontrés avant et après la date de retrait du marché en novembre 2009.

3. Nous reprenons la terminologie utilisée par les centres de pharmacovigilance avec lesquels nous interagissons dans le cadre du projet qui soutient ce travail. Pour les centres de pharmacovigilance, un effet secondaire est un événement (quelque chose qui se produit au niveau clinique) qui peut se révéler aussi bien positif que négatif. Dans ce dernier cas, on parlera d'effets indésirables.

4. http://www.eurekasante.fr/medicaments/vidal-famille/medicament-dmedia01-MEDIATOR.html

Le corpus se compose de dix fils complets de discussions autour du Médiator et de son principe actif (benfluorex), publiés à deux périodes différentes (avant et après la date de retrait du marché de ce traitement) et issus de deux forums de santé [5] en français. Les fils de discussion ont été découpés en messages individuels [6], pour un total de 157 messages. Une désidentification [7] manuelle a été réalisée pour masquer quatre types d'informations identifiantes (nom, prénom, pseudonyme, âge) que nous avons remplacées par une balise typante (e.g. <prenom/>) et nous avons changé les dates [8] présentes dans les documents lorsqu'elles se rapportent aux patients. Aucun autre traitement n'a été appliqué au corpus : ni tokénisation, ni correction orthographique ou syntaxique.

3.2 Annotation

Pour aborder la problématique de la détection des effets secondaires résultants d'une prise médicamenteuses, nous avons défini un schéma d'annotation composé de 16 catégories sémantiques [9], inspirées des types sémantiques de l'UMLS (Lindberg *et al.*, 1993). Ces catégories nous permettent de couvrir l'ensemble des informations nécessaires pour détecter la cause d'un effet secondaire (le traitement médical) et l'effet secondaire en lui-même (un symptôme, une maladie, une fonction biologique dégradée, etc.). Le principe d'annotation qui a été retenu consiste à annoter les têtes de syntagme afin (i) de se focaliser sur les entités porteuses de sens et (ii) de limiter l'impact des erreurs de reconnaissances des frontières des annotations. Puisque certaines catégories peuvent être aussi bien la cause que la conséquence d'une prise médicamenteuse, nous avons défini un attribut "rôle" qui permet de spécifier, lorsque le cas s'y prête, si l'annotation renvoie à l'*indication* (cause) ou à l'*événement* (conséquence). La figure 1 donne un exemple d'annotation du corpus.

```
Suite à quelques <SOSY role="indication">malaises</SOSY> avec <SOSY role="indication">perte</SOSY> de <FUNC
    role="indication">connaissance</FUNC>, mon <JOB>endocrinologue</JOB> m'a <PROC>prescris</PROC> du
    <CHEM>Médiator<CHEM>.
Au début c'est vrai j'ai eu le phénomène <PROC role="evenement">perte</PROC> de <FUNC
    role="evenement">poids</FUNC> (<WEIGHT role="evenement">6kg</WEIGHT>).
Au fil des années, <DISO role="evenement">migraines</DISO>, <DISO role="evenement">diarhées</DISO>, <DISO
    role="evenement">crampes</DISO>, une <SOSY role="evenement">fatigue</SOSY> de plus en plus grande, une <SOSY
    role="evenement">hyper-émotivité</SOSY>, toujours sur les <ANAT role="evenement">nerfs</ANAT>.
```

FIGURE 1 – Exemple d'annotations issues du corpus (SOSY = Sign or Symptom, FUNC = Biological Process or Function, PROC = Medical Procedure, CHEM = Chemical or drugs, DISO = Disorders, ANAT = Anatomy)

Le tableau 1 présente la répartition des annotations par catégorie en fonction de la valeur prise par l'attribut "rôle" (indication/événement) et lorsqu'aucune de ces valeurs n'est pertinente. Sur 16 catégories, seule la moitié est sous-spécifiée avec une valeur d'attribut. Pour ces huit catégories, les annotations en *Anatomy* et *Sign or Symptom* sont majoritairement sous-spécifiées comme "événement" (53,8% pour *Anatomy* et 66,9% pour *Sign or Symptom*). Cette prépondérance correspond également à la manière dont le corpus est annoté, avec des annotations connexes entre ces deux catégories (la portion *"mal de tête"* sera annotée avec la catégorie *Sign or Symptom* sur *"mal"* et la catégorie *Anatomy* sur *"tête"*).

5. Nous avons extrait quatre fils de discussion du site atoute.org (période 2004/2006, avant le retrait du marché) et six fils de discussion du site doctissimo.fr (période 2013/2014, après le retrait du marché).

6. Soit 75 messages du site atoute.org et 82 messages du site doctissimo.fr

7. La désidentification consiste à masquer ou modifier toutes informations relevant de catégories prédéfinies *(nom, prénom, adresse, téléphone, date, etc.)* permettant d'identifier l'auteur d'un message ou une personne mentionnée dans un message. S'il peut paraître inutile de désidentifier des documents récupérés sur internet, dans la mesure où il est possible de les retrouver par une simple recherche, la désidentification permet néanmoins de respecter la vie privée des utilisateurs des forums, notamment dans le cas où un utilisateur demanderait à ce que ses messages soient retirés du forum.

8. Nous avons réalisé une antédatation aléatoire manuelle des dates en retranchant quelques jours et en conservant le format d'origine.

9. *Anatomy* : parties du corps, y compris fluides et tissus (cerveau, peau, sang) — *Biological Process or Function* : processus ou état qui se produit naturellement, ou résultant d'une activité (respirer) — *Disorders* : maladies (cancer) — *Sign or Symptom* : manifestation observable d'une maladie, condition fondée sur un jugement clinique (fatigue, douleurs, ballonement) — *Chemical or Drugs* : médicament, principe actif, classe pharmacologique (Médiator, benfluorex) — *Genes Proteins* : protéines, lipides, acides nucléiques, gènes (insuline, lipase, triglycérides) — *Medical procedure* : activité médicale ou chirurgicale, liée au soin des patients, y compris diagnostiques, procédures et méthodes de traitement (radiothérapie) — *Weight* : poids total ou partiel du patient (82 kg, -5 kgs) — *Job* : activité professionnelle (médecin, gygy). — Des informations posologiques sur le traitement : *Concentration, Dosage, Mode* et des informations temporelles liées au traitement : *Date, Duration, Frequency, Time*.

	Anatomy	Disorders	Duration	Function	Gene	Procedure	Sign or Symptom	Weight
Indication	5	34	0	12	24	30	8	1
Événement	42	17	8	14	1	9	91	20
Aucun	31	52	79	43	13	149	37	25
Total	78	103	87	69	38	188	136	46

TABLE 1 – Répartition des annotations par catégorie selon la valeur de l'attribut "rôle" (indication/événement) et en l'absence de valeur associée à cet attribut (aucun)

4 Études distributionnelles

4.1 Connecteurs logiques

Lorsqu'elles se manifestent dans les textes, les relations discursives peuvent s'accompagner de marqueurs comme des connecteurs logiques, qui explicitent une relation entre deux phrases ou deux segments de phrases (*"parce que"* introduit une explication, *"plutôt que"* une alternative). Nous faisons l'hypothèse que la différence entre indication et événement peut être envisagée comme relevant des relations de discours. Dans ce cas, la présence de connecteurs logiques dans le texte constitue un indice pertinent pour la désambiguïsation des indications et des événements.

Nous avons utilisé Lexconn REF (Roze *et al.*, 2012; Roze, 2013), un lexique contenant 231 connecteurs logiques associés à une ou plusieurs des 20 relations discursives issues de la SDRT (Asher & Lascarides, 2003). Les connecteurs ont été cherchés dans dans une fenêtre de 10 mots précédant une entité à désambiguïser. Après une première projection des connecteurs, nous avons jugé nécessaire d'apporter trois modifications à Lexconn : (i) les connecteurs que nous avons jugés trop polysémiques (*à, en, et, si*) n'ont pas été pris en compte, (ii) nous avons regroupé certaines relations comme les relations d'opposition et de contraste, et (iii) nous avons ajouté les connecteurs *"dans le cadre de"* et *"pour cause de"*, associés à la relation *explication*.

Nous rapportons dans le tableau 2 la fréquence des marqueurs extraits pour trois relations (*explication, but* et *opposition*) et, entre parenthèses, la proportion qu'elle représente par rapport au nombre total d'indications ou d'événements.

	explication	but	opposition
indication	11 (9,6 %)	38 (33,3 %)	6 (5,3 %)
événement	9 (4,5 %)	8 (4 %)	8 (4 %)

TABLE 2 – Distribution des connecteurs logiques pour les relations *explication, but* et *opposition*

Nous n'avons fait apparaître que les relations pour lesquelles au moins dix marqueurs ont été extraits, ce qui ne concerne que trois relations sur les vingt contenues dans Lexconn. Le fait que si peu de relations soient représentées – et la faiblesse relative du nombre de marqueurs extraits en général – peut s'expliquer de plusieurs façons. Une première explication est que Lexconn a été construit en prenant la base Frantext, qui est constituée de textes littéraires. On peut donc prédire un décalage entre la nature des connecteurs utilisés dans Frantext et dans notre corpus. On peut également faire l'hypothèse que le caractère non contrôlé des textes de forums incite les scripteurs à utiliser moins d'indices explicites de la structure textuelle (exemple 3) ou à utiliser des connecteurs atypiques (exemple 4).

(3) Au fil des années, migraines, diarrhées, crampes, une fatigue de plus en plus grande, une hyper-émotivité, toujours sur les nerfs.

(4) j'ai pris médiator pendant un certain temps = problèmes intestinaux, diarrhée

Le résultat le plus intéressant que nous fournit le tableau 2 est qu'un tiers des indications sont introduites par un marqueur de but (alors que ce n'est le cas que pour 4 % des événements) comme *"pour"* – principalement –, *"afin de"* ou *"dans le but de"*. Cette relation exprime un lien entre la prise d'un médicament et le but de cette prise, à savoir résoudre un problème médical, analysé ici comme une indication (exemple 5).

(5) j'ai moi-même en 1998, pris ce médicament pour maigrir

Les marqueurs d'explication que sont *"car"*, *"dans le cadre de"* et *"pour cause de"* semblent jouer un rôle similaire, mais leur différence d'emploi pour introduire une indication ou un événement est moins flagrante (exemple 6).

(6) ce médicament m'a été prescrit dans le cadre d'un problème d'hyperinsulinisme

La relation d'opposition apparaît potentiellement intéressante en cela qu'elle peut être l'indice d'un phénomène inattendu (comme elle l'est à l'aide du marqueur *"or"* dans l'exemple 7).

(7) mon endocrinologue me prescrit du LEVOTHYROX 75mg par jour + 3MEDIATOR. or je prends de plus en plus de poids"

La distribution des marqueurs d'opposition n'apparaît toutefois pas potentiellement discriminante.

4.2 Expressions temporelles

Intuitivement, on peut penser que les problèmes de type *indication* se produisent temporellement avant les problèmes de type *événement*. Pour vérifier cette hypothèse, nous nous sommes intéressés aux expressions temporelles associées à ces deux types de problème. Pour cela, nous avons utilisé les annotations manuelles des expressions temporelles du corpus pour les types DATE (*depuis mai 2005*), TIME (*à 12h30, au coucher*), DURATION (*pendant de longs mois*) et FREQUENCY (*par jour, régulièrement*).

Nous avons recensé les expressions temporelles qui se trouvent dans la même phrase qu'une des 8 catégories acceptant un rôle "indication"/"événement". Lorsqu'il existe plusieurs expressions temporelles dans une phrase, nous n'avons considéré que celle la plus proche (en nombre de mots) de l'indication ou l'événement. Nous avons aussi noté la position de l'expression temporelle par rapport à l'indication ou l'événement : par exemple, la date est avant l'indication dans :

```
Je prend <CHEM>Médiator</CHEM> depuis le mois de <DATE>mars</DATE> pour <PROC role="indication">controler</PROC>
 mon <DISO role="indication">cholesthérol</DISO>
```

ou la durée est après l'événement dans :

```
J'ai perdu <WEIGHT role="evenement">7 kgs</WEIGHT> en <DURATION>6mois</DURATION>.
```

Le tableau 3 montre la distribution des expressions temporelles dans le corpus par rapport à leur position vis-à-vis d'une indication ou d'un événement.

		DATE		TIME		DURATION		FREQUENCY		TOTAL
		avant	après	avant	après	avant	après	avant	après	
GENE	INDIC					1	1		2	4
	EVENT									0
DISORDER	INDIC		2			1		2		5
	EVENT	2		2						4
SIGN OR SYMPTOM	INDIC				1	1		2		4
	EVENT	1	1	3	4	8	5	3	1	26
PROCEDURE	INDIC	4	1	1		2	1			9
	EVENT					1				1
FUNCTION	INDIC									0
	EVENT			1						1
WEIGHT	INDIC						1			1
	EVENT	2				7	7	1		17
TOTAL	INDIC	4	2	1	1	5	3	4	2	22
	EVENT	5	1	6	4	16	12	4	1	51

TABLE 3 – Distribution des expressions temporelles

On remarque que les expressions temporelles, quel que soit leur type, sont très majoritairement associées aux catégories *Sign or symptom* et *Weight* et principalement de rôle *événement*. On note également que les expressions temporelles de type TIME et DURATION sont principalement associées à des rôles *événement* alors que pour les autres types d'expression temporelle, la distribution est assez uniforme. La catégorie *Procedure* est majoritairement de rôle *indication* quelle que soit l'expression temporelle associée. Enfin, la catégorie *Disorders* est de type *indication* si une date est positionnée après alors qu'elle est de type *événement* si la date est avant.

Ces observations sont des indices qui peuvent aider à catégoriser une annotation issue de notre schéma en *indication* ou *événement* même si le petit nombre d'occurrences ne permet pas de tirer de réelle conclusion. De plus, plutôt que la position des expressions temporelles, il faudrait connaître leurs relations de dépendance syntaxique : dans l'exemple suivant, la première date est associée à la catégorie *Procedure* dans la première proposition alors que la seconde date est associée à la catégorie *Sign or symptom* dans la seconde proposition :

```
Ma jeune cousine est sous <CHEM>médiator</CHEM> pour "<PROC role="indication">s'affiner</PROC>"
depuis <DATE>vendredi 12 mars 2004</DATE> et depuis <DATE>lundi 15</DATE> se plaint de
<SOSY role="evenement">douleurs</SOSY> au <ANATOMY role="evenement">ventre</ANATOMY>
```

On le voit également dans cet exemple, les valeurs normalisées des dates peuvent permettre de typer les catégories : ainsi, la date *vendredi 12 mars 2004* (2004-03-12) précède temporellement la date *lundi 15* (2004-03-15) et confirme qu'une *indication* se produit avant un *événement*. Pour obtenir ces informations, nous avons adapté les règles pour le français de l'outil libre HeidelTime (Moriceau & Tannier, 2014) afin d'extraire et normaliser les expressions temporelles exprimées dans un style propre aux forums en ligne (par exemple, *pdt 3 jrs* pour *pendant 3 jours*).

5 Conclusion

Dans cet article, nous avons présenté les premières études que nous avons menées pour identifier des indices linguistiques permettant de distinguer les entités cliniques qui sont des *indications* (ce pour quoi le traitement est administré) de celles qui sont des *événements* (ce qui est causé par le traitement). De cette étude, il ressort que ces indices permettent plus facilement de repérer les *indications* que les *événements*. Nous prévoyons toutefois de tester les indices que sont les verbes de causation (*provoquer*, *engendrer*, *entraîner*…), dont nous faisons l'hypothèse qu'ils sont pertinents pour le repérage des événements.

Nous estimons que ces indices devraient permettre d'identifier automatiquement les effets secondaires dans les messages parus sur les forums de santé, à des fins de pharmacovigilance. A cet effet, nous envisageons de poursuivre ces travaux en utilisant les indices que nous avons mis en évidence pour identifier automatiquement les entités cliniques qui relèvent d'une indication ou d'un événement.

Le corpus que nous avons utilisé est limité à un seul traitement (Médiator) et se compose d'un nombre réduit de messages (157 messages). Nous avons envisagé la création de ce corpus uniquement dans la perspective d'identifier des indices pour l'extraction d'entités cliniques d'une part, et d'en vérifier leur utilité réelle d'autre part. Si la question de la mise à disposition du corpus et des annotations réalisées est pertinente, la problématique de la redistribution de contenus issus d'internet ne nous permet pas de donner accès à ce corpus. Si le contenu des messages n'est pas redistribuable pour des questions de droit, il est néanmoins possible de donner accès à la liste des URL correspondant aux discussions qui nous ont permis de constituer le corpus. C'est notamment l'approche qui a été suivie dans l'atelier DEFT 2015 (`https://deft.limsi.fr/2015/`) : les organisateurs ont fourni aux participants la liste des tweets sur lesquels ils devaient développer et appliquer leurs méthodes pour répondre à la problématique posée. La méthode présentée dans cet article reste cependant applicable sur n'importe quel corpus issus de forums de santé, ce qui conduit malgré tout à la possibilité de dupliquer la méthode présentée et de confirmer ou de réfuter les conclusions que nous avons présentées.

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

ASHER N. & LASCARIDES A. (2003). *Logics of Conversation*. Cambridge University Press.

CHA Y. & STOW C. (2015). Mining web-based data to assess public response to environmental events. *Environ Pollut*, **198**, 97–9.

LINDBERG D. A., HUMPHREYS B. L. & MCRAY A. T. (1993). The Unified Medical Language System. *Methods Inf Med*, **32**(4), 281–91.

MEGAHED D. (2014). Etude des forums de santé pour la détection d'événements secondaires. Master's thesis, INaLCO.

MORICEAU V. & TANNIER X. (2014). French resources for extraction and normalization of temporal expressions with heideltime. In *Proc of LREC*, p. 3239–43, Reykjavik, Iceland.

NIKFARJAM A., SARKER A., O'CONNOR K., GINN R. & GONZALES G. (2015). Pharmacovigilance from social media : mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J Am Med Inform Assoc*.

ROZE C. (2013). *Vers une algèbre des relations de discours*. PhD thesis, Université Paris-Diderot - Paris VII, Paris, France.

ROZE C., DANLOS L. & MULLER P. (2012). LEXCONN: a French lexicon of discourse connectives. *Discours*, **10**.

SAMPATHKUMAR H., CHEN X. & LUO B. (2014). Mining adverse drug reactions from online healthcare forums using hidden markov model. *BMC Med Infor Decis Mak*, **91**(14).

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. *J Biomed Inform*. Epub ahead of print.

SEGURA-BEDMAR I., MARTÍNEZ P. & DE PABLO-SÁNCHEZ C. (2011). A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. *BMC Bioinformatics*, **12(Suppl 2)**(S1).

SUN W., RUMSHISKY A. & UZUNER O. (2013). Evaluating temporal relations in clinical text : 2012 i2b2 challenge. *J Am Med Inform Assoc*, **20**(5), 806–13.

VELARDI P., STILO G., TOZZI A. & GESUALDO F. (2014). Twitter mining for fine-grained syndromic surveillance. *Artif Intell Med*, **61**(3), 153–63.

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.