Apports de l'analyse automatique multilingue pour la veille épidémiologique

Gaël Lejeune, Romain Brixtel, Charlotte Lecluze, Antoine Doucet

Apports de l'analyse automatique multilingue pour la veille épidémiologique

Gaël Lejeune, Romain Brixtel, Charlotte Lecluze, Antoine Doucet

Université de Caen Basse-Normandie – prénom.nom@unicaen.fr

Abstract

The early detection of disease outbursts is an important objective of epidemic surveillance. The web news are one of the information bases for detecting epidemic events as soon as possible, but to analyze tens of thousands of articles published daily is costly. Recently, automatic systems have been devoted to epidemiological surveillance. The main issue for these systems is to process more languages at a reasonable cost. However, existing systems mainly process major languages (English, French, Russian, Spanish...). Thus, when the first news reporting a disease is written in a minor language, the timeliness of event detection is worsened. In this paper, we evaluate an automatic style-based method, designed to fill the gaps of existing automatic systems. It is parsimonious in resources and specially adapted for multilingual issues. The events detected by the human-moderated ProMED mail between November 2011 and January 2012 are used as a reference dataset and compared to events detected in 17 languages by the system DAnIEL from web articles of this time-window. We show how being able to process press articles in various languages allows quicker detection of epidemic events in some regions of the world.

Résumé

La détection précoce des épidémies de maladie est un objectif primordial pour les autorités sanitaires. La presse en ligne constitue l'une des principales bases d'information pour détecter dès que possible ces évènements épidémiologiques. L'analyse des dizaines de milliers d'articles publiés chaque jour est coûteuse. Différentes propositions d'approche automatique ont été formulées ces dernières années. Le principal problème pour ces systèmes est de traiter plus de langues à un coût limité. Cependant, les systèmes existants couvrent un éventail limité de langues (anglais, français, russe, espagnol,…). Ainsi, lorsque le premier article est rédigé dans une autre langue, la rapidité de la détection est moindre. Dans cet article, nous proposons de comparer un système automatique massivement multilingue avec le système manuel de référence ProMED. Nous montrons comment l'augmentation de la couverture en langues amène une amélioration du délai de détection des évènements épidémiologiques.

Mots-clés : veille, articles de presse, Extraction d'Information (EI), Recherche d'Information (RI), données textuelles, multilinguisme.

1. Introduction

Le but de la veille épidémiologique est de détecter le plus rapidement possible les épidémies de maladies à travers le monde. Ainsi, les articles de presses publiés en ligne constituent une source d'information pour les autorités sanitaires. L'augmentation dans différents pays du nombre de journaux en ligne permet d'envisager une détection accélérée des phénomènes épidémiologiques. Différents projets font état de l'usage de la presse en ligne pour accélérer la détection des épidémies. Les systèmes institutionnels ProMED[1] (Cowen *et al.* 2006) et

1 http://www.promedmail.org

GPHIN[2] (Mawudeku *et* al. 2006) se fondent sur des avis d'experts pour analyser les articles de presse publiés en ligne. D'autres systèmes se basent sur l'analyse automatique pour gérer ces masses de données : BioCaster (Son et al. 2008), DAnIEL (Lejeune *et al.* 2013), EpiSpider (Tolentino *et al.* 2007), HealthMap (Bronstein *et al.* 2009) ou encore PULS (Yangarber *et al.* 2008). Une des différences entre ces systèmes est le nombre de langues couvertes (*Tableau 1*).

Langues	Locuteurs	PULS	ProMED	GPHIN	HealthMap	BioCaster	DAnIEL
Anglais	1000	✓	✓	✓	✓	✓	✓
Russe	277	✓	✓	✓	✓	✓	✓
Espagnol	500		✓	✓	✓	✓	✓
Français	200		✓	✓	✓	✓	✓
Arabe	255			✓	✓	✓	✓
Portugais	240		✓		✓	✓	✓
Chinois	1151			✓	✓		✓
Allemand	166						✓
Vietnamien	86					✓	
Coréen	78					✓	
Turc	75						✓
Italien	62						✓
Thaï	60					✓	
Polonais	46						✓
Néerlandais	21						✓
Grec	13						✓
Tchèque	10						✓
Suédois	8						✓
Finnois	5						✓
Norvégien	5						✓

Tableau 1 - Couverture de différentes approches, nombre de locuteurs pour chaque langue (en millions).

Seules deux langues (anglais et russe) sont couvertes par tous les systèmes sur les vingt langues présentées. Ces vingt langues représentent 4 milliards de locuteurs ; environ 40% de la population mondiale reste donc hors du domaine de validité de ces systèmes. Dès lors il est difficile de savoir si tous les évènements épidémiologiques seront un jour ou l'autre

[2] http://www.phac.aspc.gc.ca/gphin/

mentionnés dans une des langues couvertes. D'autre part, si l'évènement est finalement décrit dans une langue couverte, nous pouvons nous interroger sur le temps écoulé depuis le premier rapport en langue locale, par exemple si cet article doit être traduit avant d'être traité.

Quelques études ont été menées pour mesurer l'impact de la couverture en nombre de langues sur la qualité de l'information extraite. Piskorski et al. (2011) ont montré une corrélation significative entre augmentation de la couverture et qualité des informations extraites pour différentes applications. Lyon et al. (2012) sont arrivés à la même conclusion dans le domaine particulier de la veille épidémiologique en ne considérant que cinq langues, ce qui est peu au regard des trente langues disponibles sur *Google News*.

Nous proposons ici une étude globale de la plus-value apportée par des systèmes automatiques analysant de nombreuses langues. À cette fin, nous comparons ProMED, le système manuel de référence et DAnIEL, le système automatique offrant la plus grande couverture. Il s'agit de mesurer si un système multilingue permet de détecter plus rapidement certains évènements et ainsi réduire le temps écoulé entre les premiers cas d'une maladie et le moment où les autorités sanitaires sont informées. Un cas typique d'incidence importante du délai de détection est l'épidémie de SRAS de 2002-2003 en Chine dont les principales étapes sont représentées dans la Figure 1.

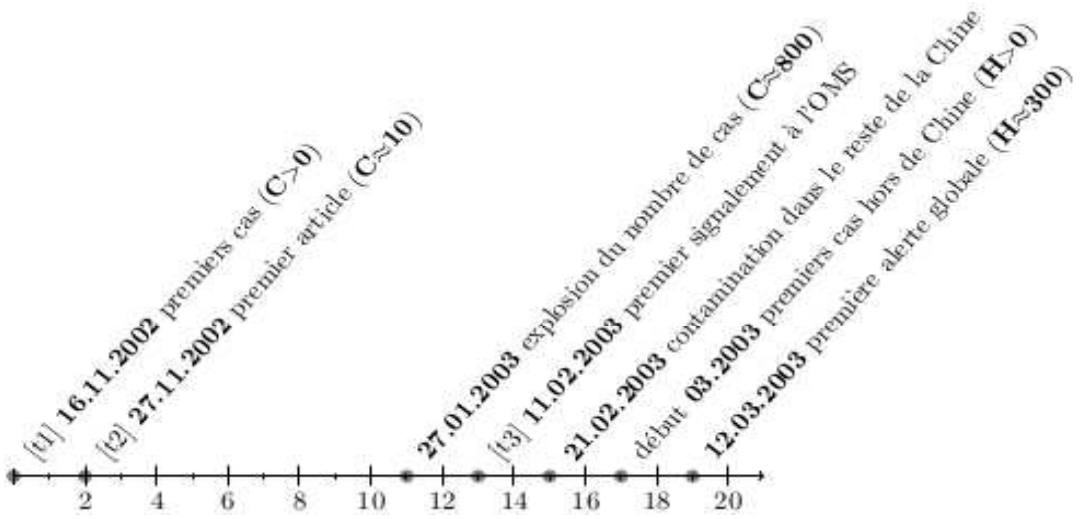


Figure 1 - L'épidémie de SRAS de 2002-2003 en Chine[3]*, principales étapes de propagation et de signalement, en semaines, avec* ***C*** *le nombre de cas en Chine et* ***H*** *le nombre de cas hors de Chine.*

Entre les premiers cas attestés (**t1**) et le signalement par l'OMS (**t3**) d'un « problème majeur » il s'est écoulé 13 semaines. Le délai de connaissance officielle est donc particulièrement long. Le stade **t2**, où l'information est relayée pour la première fois dans la presse, est pourtant atteint en moins de 2 semaines. Le délai de publication était relativement court mais le délai de signalement a été bien plus important (11 semaines). Nous constatons que pendant le délai de signalement, le nombre de cas a fortement augmenté. Ce nombre (**M** dans la figure) est ainsi passé d'une dizaine en **t2**, à plusieurs centaines peu avant **t3**. Le retard dans le

[3] *China Raises Tally of Cases and Deaths in Mystery Illness* : http://www.nytimes.com/2003/03/27/world/ china-raises-tally-of-cases-and-deaths-in-mystery-illness.html

signalement a eu un impact sur l'explosion du nombre de cas hors de Chine. Quand les autorités internationales ont disposé des données leur permettant de réagir, le nombre de cas atteint rendait déjà la situation difficilement contrôlable. Le stade épidémique était déjà dépassé et c'est à un risque aigu de pandémie que les autorités ont dû faire face.

Afin de mesurer la plus-value des systèmes multilingues dans le domaine de la veille épidémiologique, cet article est articulé de la manière suivante : dans la section 2, nous présentons le fonctionnement de ProMED et DAnIEL, puis dans la section 3, nous décrivons les jeux de données utilisées pour notre étude. La section 4 est consacrée à l'analyse des résultats. Les conclusions et perspectives de cette étude sont détaillées en section 5.

2. Veille épidémiologique multilingue avec ProMED and DAnIEL

Dans cette section, nous présentons le fonctionnement des deux systèmes et l'influence que celui-ci a sur le nombre de langues couvertes.

2.1. Le système ProMED

ProMED diffuse chaque jour des informations sur les épidémies survenant dans différentes parties du monde. Les modérateurs du système exploitent différentes sources d'information pour produire leurs rapports. Les principales sources utilisées sont des articles de presse, des rapports officiels ainsi que des informations transmises par des observateurs locaux.

ProMED diffuse des rapports en anglais depuis les débuts du projet en 1994. Des rapports sont aussi disponibles en français, portugais, russe et espagnol. Les rapports sont classés par paire Maladie-Lieu. La qualité des alertes émises dépend de la finesse de l'analyse effectuée par les experts humains. Or, cette analyse est longue et coûteuse. Toutes les langues ne sont donc pas analysables et toutes les sources ne peuvent pas être examinées.

Quel est l'impact de la chaîne de traitement de ProMED sur le délai de détection ? Pour répondre à cette question, nous comparons ses résultats à ceux d'un système automatique multilingue plus léger en ressources et offrant une couverture plus large.

2.2. Le système DAnIEL

Ce système propose une large couverture multilingue en se fondant sur une approche parcimonieuse destinée à limiter le coût marginal de traitement d'une nouvelle langue. La détection de répétitions de chaînes de caractères à des positions spécifiques des textes constitue la base de son fonctionnement, s'appuyant sur des spécificités du style d'écriture journalistique. Pour traiter une nouvelle langue le système[4] ne nécessite que des ressources lexicales de petite taille, en l'occurrence les noms de maladies tels qu'ils figurent dans Wikipedia, l'approche est détaillée dans (Brixtel et al. 2013).

DAnIEL exploite les articles de presse diffusés sur des agrégateurs tels que *Google News* ou *European Media Monitor* pour détecter des évènements épidémiologiques. Le système est entièrement automatique et couvre 43 langues. Il permet de traiter 2 000 documents par minute (Lejeune, 2013) ce qui est compatible avec la grande quantité de documents diffusés par ces agrégateurs. Par exemple, *European Media Monitor* diffuse 50 000 articles par jour et *Google News* plus de 100 000 (soit environ 100 articles par minute).

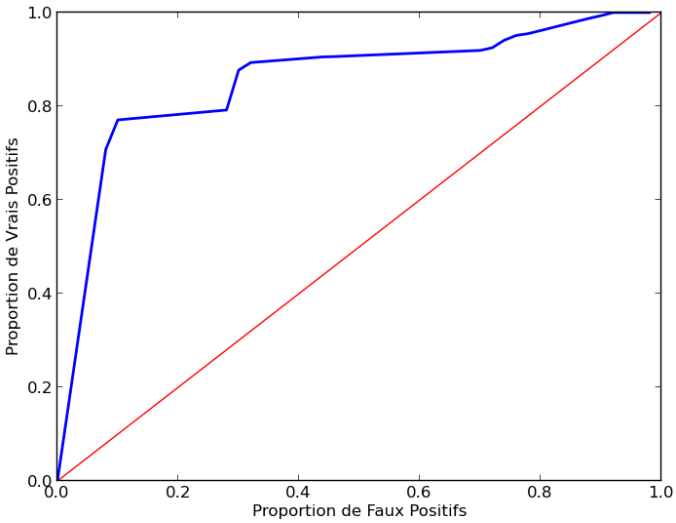
[4] Le système est disponible en ligne à l'adresse suivante :https:daniel.greyc.fr

DAnIEL filtre les documents en deux classes : pertinents ou non-pertinents pour la veille épidémiologique. Pour les documents classés dans la première catégorie, DAnIEL les étiquette plus finement sous la forme de Paires Maladie-Lieu (PML). Extraire une paire X-Y signifie pour le système l'existence d'une épidémie de la maladie X dans le pays Y. Nous pouvons donc confronter ces extractions avec les résultats de référence issus de ProMED.

Une autre manière de l'évaluer est d'utiliser une courbe ROC[5]. Celle ci permet de représenter les performances d'un classifieur binaire pour différents jeux de paramètres.

La courbe ROC (Figure 2) a été calculée sur un échantillon de 2089 documents en cinq langues (anglais, chinois, grec, polonais et russe) du corpus de 26091 documents ayant servi à comparer DAnIEL et ProMED (Tableau 4, page 7). En ordonnée figure le pourcentage de vrais positifs (le rappel), en abscisse la proportion de faux positifs (ou bruit). Plus un point est proche de 1 en ordonnée, plus le rappel est élevé.

Plus l'abscisse d'un point est proche de 1, plus la précision est faible. Un excellent classifieur atteint un rappel très élevé sans que sa précision chute trop vite. La courbe ROC idéale a donc une pente très forte dès les faibles valeurs de bruit (à gauche de la courbe). La diagonale qui part du point (0,0) et qui va vers le point (1,1) représente un classifieur aléatoire.



La courbe ROC ci-contre présente deux paliers. Le premier se situe aux coordonnées (0.09,0.77). DAnIEL obtient une sensibilité de 0,77 avec un bruit assez faible.Le second palier se situe aux coordonnées (0.31,0.91). À partir de ce point, tout gain en sensibilité devient prohibitif au regard du bruit engendré. Ainsi, un gain de 0,04 en sensibilité occasionne une augmentation du bruit de 0,41. Avec une aire sous la courbe de 0,86 DAnIEL s'avère performant vis-a-vis de l'échantillon annoté sur lequel il a été évalué (Brixtel et al. 2013).

Figure 2 - Courbe ROC sur un échantillon de 2089 documents (cn, el, en, pl, ru)

3. Jeu de données pour ProMED et DAnIEL

3.1. Jeu de données issu des rapports ProMED

Nous avons collecté 2558 rapports diffusés sur ProMED entre octobre 2011 et février 2012. Nous disposons ainsi de 2558 comptes-rendus structurés Ceux-ci sont principalement disponibles en cinq langues (anglais, espagnol, français , portugais et russe). Quelques comptes-rendus en thaï et en vietnamien ont également été produits dans les premiers jours de cette période. Chacun des rapports collectés expose des faits relatifs à un ou plusieurs évènements épidémiologiques, sous la forme d'une PML

5 *Receiver Operating Characteristics*

Pour cette expérience nous nous intéressons spécifiquement au premier signalement. Pour chaque PML de la période, nous ne nous intéressons donc qu'au premier rapport émis par ProMED. Nous disposons par ailleurs d'informations sur la source qui a permis l'émission du rapport. Ceci nous permet d'établir quelle est la langue couverte qui a permis la détection de l'évènement.

Le corpus ainsi créé est présenté dans le *Tableau 2*. Les rapports en anglais représentent plus de la moitié des rapports émis. De la même façon, la majorité des sources à l'origine des rapports sont elles-mêmes en anglais. L'importance de l'anglais dans l'émission des rapports ProMED s'explique par le fait que de très nombreuses sources sont disponibles dans cette langue. L'anglais offre ainsi la meilleure couverture par rapport aux autres langues.

	anglais	français	portugais	russe	espagnol	thaï	vietnamien	Ensemble
# Rapports	819	148	129	127	220	25	78	1546
Nov. 2011	285	3	26	49	68	25	78	534
Déc. 2011	291	33	15	28	78	0	0	445
Jan. 2012	193	62	48	37	37	0	0	377
Fév. 2012	54	50	40	33	38	0	0	215

Tableau 2 - Répartition des rapports ProMED pour chaque langue et chaque mois de la période d'étude.

Cette situation est, par exemple, visible sur les agrégateurs multilingues. L'anglais occupe une place centrale dans la catégorie santé de *Google News* dont les statistiques sont exposées dans le *Tableau 4*. Toutefois, la part relative de l'anglais est significativement supérieure sur les rapports émis par ProMED. La dépendance de ProMED par rapport à l'anglais est probablement disproportionnée par rapport à la réalité des corpus disponibles. Les rapports émis en anglais concernent une grande variété de maladies et de lieux même si 40% des évènements extraits se concentrent sur trois pays : États-Unis, Australie et Royaume-Uni (*Tableau 3*).

	anglais	français	portugais	russe	espagnol	thaï	vietnamien
# Rapports	819	148	129	127	220	25	78
Maladies	183	33	34	47	58	10	31
Lieux	151	37	23	15	46	8	26
PML	366	63	40	55	46	12	26

Tableau 3 - Détails sur les rapports ProMED : répartition par maladie, lieux et Paire Maladie-Lieu (PML).

Dans ce jeu de données, très peu de rapports sont tirés d'une source émise dans une langue non couverte par ProMED. Nous avons étudié un sous-corpus de 200 rapports montrant la même répartition par langue que celle présentée dans le *Tableau 3*. Seuls deux de ces rapports provenaient d'une source rédigée dans une langue autre que celles traitées par ProMED. Pour faciliter la comparaison, nous considérons que la langue du rapport ProMED correspond à la langue de la source qui a permis l'émission du rapport.

3.2. Corpus pour DAnIEL

Nous avons constitué un corpus à partir de la catégorie santé de *Google News* sauf pour le grec, le finnois et le polonais, langues pour lesquelles cette catégorie n'existe pas et pour lesquelles les documents ont été collectés à partir de fils RSS relatifs à la santé et dans la catégorie santé de journaux de diffusion nationale. Ces trois langues ont été choisies pour augmenter le nombre de langues à morphologie riche dans le corpus, ces langues étant réputées difficiles à traiter pour les systèmes automatiques. Le corpus contient des documents publiés dans la période s'étendant du premier octobre 2011 au 31 janvier 2012. La composition du corpus par langue et par période est détaillée dans le *Tableau 4*. 40% de ces documents sont rédigés dans des langues non couvertes par ProMED.

Langues	Total	Oct. 2011	Nov. 2011	Déc. 2011	Jan. 2012
Anglais (en)	4742	1301	1181	1082	1178
Espagnol (es)	4389	952	1020	1517	900
Arabe (ar)	3093	780	819	735	759
Allemand (de)	2509	631	809	712	357
Français (fr)	2132	412	506	832	382
Russe (ru)	1896	240	312	487	857
Grec (el)	1380	220	289	400	471
Portugais (pt)	1362	343	205	485	329
Chinois (zh)	1122	243	174	303	402
Néerlandais (nl)	876	197	172	253	254
Polonais (pl)	801	182	199	122	298
Italien (it)	703	173	100	224	206
Norvégien (no)	311	52	61	111	87
Turc (tr)	239	74	79	52	34
Tchèque (cs)	208	42	99	37	30
Suédois (sv)	196	41	72	37	46
Finnois (fi)	132	23	37	32	40
Ensemble	26091	5906	6134	7421	6630

Tableau 4 - Nombre d'articles par langue et par mois. Les documents composant ce corpus sont en html brut.

À partir de ce corpus, DAnIEL a extrait 1571 évènements épidémiologiques (*Tableau 5* et *Tableau 6*). Parmi ces signalements, 32% ont été extraits à partir de documents dans des langues non couvertes par ProMED. L'arabe présente un nombre limité de signalements eu égard au nombre de documents analysés dans cette langue. Moins de 1% des documents de cette langue ont contribué à la production d'un signalement. Aucun signalement n'a été émis à

partir des documents en turc. Le nombre peu élevé de documents disponibles sur cette période constitue une explication partielle.

Afin de mieux appréhender le contenu du corpus, nous avons fait annoter 100 documents en turc par des locuteurs natifs, aucun d'entre eux n'a été étiqueté comme pertinent pour la veille épidémiologique. Le choix de collecter des documents sur la catégorie santé de *Google News* n'est pas pertinent pour le turc : le filtre est trop strict.

Langues	# A	# S	100*S/A	Oct. 2011	Nov. 2011	Déc. 2011	Jan. 2012
Anglais (en)	4742	285	6,01%	63	75	67	80
Espagnol (es)	4389	230	5,24%	42	62	71	55
Arabe (ar)	3093	30	0,97%	3	5	12	10
Allemand (de)	2509	63	2,51%	7	13	24	19
Français (fr)	2132	142	6,66%	17	50	48	27
Russe (ru)	1896	296	15,61%	49	84	54	109
Grec (el)	1380	83	6,01%	17	25	18	23
Portugais (pt)	1362	92	6,75%	30	22	25	15
Chinois (zh)	1122	73	6,51%	12	25	14	22
Néerlandais (nl)	876	24	2,74%	2	4	12	6
Polonais (pl)	801	140	17,47%	15	37	36	52
Italien (it)	703	54	7,68%	12	19	15	8
Norvégien (no)	311	11	3,53	0	4	4	3
Turc (tr)	239	0	0	0	0	0	0
Tchèque (cs)	208	15	7,21%	2	7	3	3
Suédois (sv)	196	26	13,26%	2	10	9	5
Finnois (fi)	132	7	5,3%	2	0	3	2
Ensemble	26091	1571	6,02%	275	442	415	439

Tableau 5 - Nombres d'articles analysés (A) et de signalements (S) émis par DAnIEL et proportion de signalements en fonction du nombre de documents disponibles par langue.

Le *Tableau 6* (page 9) présente le nombre de maladies et de lieux différents concernés par les signalements émis par DAnIEL. Les langues de grande diffusion rapportent plus facilement des évènements qui se produisent en dehors de leur zone d'influence. Au contraire, pour des langues plus rares comme le suédois ou le finnois les signalements se concentrent sur un nombre limité de lieux.

Ces données invitent à s'interroger sur la relation entre la couverture d'une langue et la qualité de la surveillance de telle ou telle zone du globe. Une des motivations de la couverture

multilingue est de pouvoir traiter le premier article relatant un fait épidémiologique indépendamment de la langue dans laquelle il est rédigé.

Langues	Rapports	Maladies	Lieux	PML
Russe (ru)	296	21	70	141
Anglais (en)	285	33	55	161
Espagnol (es)	230	29	35	115
Français (fr)	142	32	39	85
Polonais (pl)	140	19	45	83
Portugais (pt)	92	23	14	50
Allemand (de)	63	12	19	32
Grec (el)	83	13	7	25
Chinois (zh)	73	16	6	23
Italien (it)	54	22	9	28
Arabe (ar)	30	7	3	12
Suédois (sv)	26	7	2	10
Néerlandais (nl)	24	9	7	11
Tchèque (cs)	15	6	2	9
Norvégien (no)	11	6	1	6
Finnois (fi)	7	6	2	4
Turc (tr)	0	0	0	0

Tableau 6 - Nombre de maladies, de lieux et de PML impliqués dans les signalements produits par DAnIEL.

4. Évaluation

Nous étudions dans cette section le bénéfice que peut offrir un système de veille épidémiologique automatique tel que DanIEL. Nous évaluons l'apport d'une couverture accrue en nombre de langues dans la tâche d'émission des signalements.

Nous supposons qu'un évènement ayant lieu dans un pays donné sera tout d'abord décrit dans une publication dans une langue officielle de ce pays. C'est donc sur les langues non traitées par ProMED et les zones géographiques qu'elles recouvrent que l'on s'attend à ce que le système automatique soit en avance. À l'opposé, à données égales, l'analyse humaine est sans doute plus efficace. Lorsque l'humain et la machine ont accès aux mêmes documents, la machine réagira au mieux aussi vite. Nous ne tenons pas compte ici du temps consacré à l'analyse mais simplement de la première source qui déclenche le signalement. On peut considérer que cette mesure désavantage le système automatique DAnIEL.

Parmi les PML extraites, 167 l'ont été par les deux systèmes. Afin de mesurer les différences en termes de délai de détection nous avons comparé pour chaque PML la date du plus ancien rapport issu de chaque système.

PML		ProMED		DAnIEL		Retard
Maladie	Lieu	Lg.	Date	Lg.	Date	(jours)
Grippe	Canada	**en**	2011-11-04	**en**	2011-12-01	27
Pneumonie	Russie	**ru**	2011-11-12	**ru**	2011-12-09	27
Grippe	Italie	en	2011-11-05	***it***	2011-12-01	26
Gale	Espagne	en	2011-12-25	**es**	2012-01-12	18
Hépatite	Russie	en	2011-11-22	**ru**	2011-12-06	14
Rougeole	Ukraine	**ru**	2011-12-28	**ru**	2012-01-06	9
Syphilis	Espagne	**es**	2011-11-29	**es**	2011-12-01	2

Tableau 7 - Exemples de PML pour lesquelles le premier signalement vient ProMED. Pour chacune, nous indiquons la langue et la date ainsi que le retard de DAnIEL. En gras, les langues officielles du pays concerné, en italique les langues non-couvertes par ProMED.

Le *Tableau 7* présente des PML signalées en premier lieu par ProMED, le *Tableau 8* présente celles où DAnIEL a donné le premier signalement. Nous pouvons remarquer que les cas où ProMED effectue le premier signalement sont massivement dus à des sources en anglais. C'est particulièrement vrai pour des pays où l'anglais constitue la langue officielle, même s'il existe des exceptions (dans notre exemple la détection de la PML Grippe-Italie). DAnIEL apporte une plus-value aux alertes ProMED dans des zones géographiques dont les langues ne sont pas couvertes par ProMED.

Parmi les PML extraites par les deux systèmes, dans 37% des cas DAnIEL a fourni le tout premier signalement (*Tableau 9*). Nous pouvons remarquer que DAnIEL obtient de meilleurs résultats dès lors qu'il dispose de documents en langue locale a priori non-directement traitables par les analystes ProMED. C'est particulièrement le cas pour l'Europe (hors France Royaume-Uni et péninsule ibérique). Ces zones correspondent sans surprise à la sphère d'influence des langues traitées par les analystes ProMED. Pour ces zones, le système automatique offre une grande complémentarité bien qu'il ne puisse pas à proprement parler remplacer le système manuel.

Paire		ProMED		DAnIEL		Gain
Maladie	Lieu	Lg.	Date	Lg.	Date	(jours)
Méningite	Russie	**ru**	2011-12-18	**ru**	2011-12-06	12
Rage	Russie	**ru**	2011-12-21	fr	2011-12-09	12
Fièvre jaune	Brésil	**pt**	2011-12-20	**pt**	2011-12-09	11
Botulisme	Finlande	en	2011-11-01	***fi***	2011-10-21	11
Dengue	Colombie	**es**	2012-02-03	**es**	2012-01-23	11
Salmonelle	Russie	**ru**	2012-01-14	**ru**	2012-01-06	8
Grippe	Espagne	**es**	2011-12-14	**es**	2011-12-09	5

Tableau 8 - Exemples de PML signalées en premier lieu par *DAnIEL. Pour chacune nous indiquons la langue et la date ainsi que le gain par rapport à ProMED. En gras, les langues officielles du pays concerné, en italique les langues non-couvertes par ProMED.*

Ce phénomène est plus contrasté en Afrique où DAnIEL tire avantage du traitement de l'arabe dans certaines régions. Mais DAnIEL est moins performant dès lors qu'il ne bénéficie pas de documents dans la langue locale. DAnIEL est par contre fréquemment plus rapide pour les évènements ayant lieu en Europe Centrale (République Tchèque par ex.), Europe du Nord (Finlande) ou Europe du Sud (Grèce).

La Russie et l'Ukraine sont deux contre-exemples. Le russe est couvert par ProMED, ce qui laisserait supposer une plus grande réactivité du système manuel. Toutefois, DAnIEL bénéficie d'articles en polonais relatant les évènements se produisant dans ces pays.

Nous présentons dans le *Tableau 10*, la comparaison entre les deux systèmes, en fonction cette fois de la langue de la source utilisée. Nous voyons ici de façon plus claire que DAnIEL offre une complémentarité intéressante dès lors que des documents sont disponibles dans d'autres langues que les langues traitées par ProMED. DAnIEL offre toutefois des résultats moins bons que ProMED dès lors que les sources permettant l'émission de l'alerte sont en anglais. C'est vrai dans une moindre mesure lorsque les sources disponibles sont en espagnol ou en portugais.

	ProMED		DAnIEL	
Zones des évènements	Langues	PS	Langues	PS
France, Portugal, Espagne, Royaume-Uni	en, es, fr, pt	31 (72%)	en, es, fr, nl, pt	12 (28%)
Reste de l'Europe	en, fr	7 (37%)	cs, de, el, fi, fr, it, sv	12 (63%)
Russie, Ukraine	en, ru	4 (40%)	pl, ru	6 (60%)
Afrique du Nord	en, fr	5 (62%)	ar, fr	3 (38%)
Reste de l'Afrique	en, fr, pt	10 (77%)	fr	3 (23%)
Chine/Inde	en	5 (62%)	en, zh	3 (38%)
Reste de l'Asie	en	6 (34%)	ru, zh	9 (66%)
Amérique du Nord	en, es	22 (85%)	en, es	4 (15%)
Amérique Centrale et du Sud	en, es, pt	16 (64%)	en, es, pt	9 (36%)
Ensemble	5	106 (63%)	15	61 (37%)

Tableau 9 - Localisation des Premiers Signalements (PS) de chacun des systèmes.

	ar	cs	de	el	en	es	fi	fr	it	nl	no	pl	pt	ru	sv	tr	zh	Total
ProMED	-	-	-	-	53	27	-	5	-	-	-	-	15	6	-	-	-	106
DAnIEL	1	2	3	4	8	8	2	8	3	3	0	2	4	9	1	0	3	61

Tableau 10 - Répartition par langue des premiers signalements de ProMED et DAnIEL. "-" signale une langue non couverte.

5. Conclusion

La comparaison des sorties des deux systèmes met en lumière les bénéfices que l'on pouvait attendre d'un traitement multilingue. Le système automatique permet de combler les manques du système manuel pour les évènements survenant dans des zones non couvertes par les langues de ProMED. Si le système automatique ne semble pas pouvoir remplacer le système manuel, il apporte une plus-value en terme de délai de détection dans un cas sur trois. Nous avons donc ici une bonne illustration d'une interaction réussie entre systèmes de traitements manuels et systèmes automatiques.

Néanmoins, certaines questions restent ouvertes. Il est difficile de savoir pourquoi la plupart des signalements n'ont pas pu être comparés entre les deux systèmes. Une étude sur une période de temps plus large permettrait sans doute d'éclaircir ce point.

Références

Brixtel, R., Lejeune, G., Doucet, A., Lucas, N. (2013). Any Language Early Detection of Epidemic Diseases from Web news Streams. *International Conference on Healthcare Informatics (ICHI)* : pp.159–168.

Bronstein, J.S., Freifeld, C.C. (2009). Digital disease detection - harnessing the web for public health surveillance. *New England Journal of Medicine 360(21)* : pp.2153–2157.

Cowen, P., Garland, T., Hugh-Jones, M.E., Shimshony, A., Handysides, S., Kaye, D., Madoff, L.C., Pollack, M.P., Woodall, J. (2006). ProMED-mail as an electronic early warning system for emerging animal diseases: 1996 to 2004. *JAVMA 229(7)*, pp.1090–1099.

Katsiavriades, K., Qureshi, T. (2007). The 30 most spoken languages of the world. http://www.krysstal.com/spoken.html

Lejeune G., Brixtel, R., Lecluze C., Doucet, A., Lucas, N. (2013). DAnIEL : Veille Épidémiologique Multilingue Parcimonieuse. *Traitement Automatique des Langues Naturelles (TALN)* : pp.77–78.

Lyon, A., Nunn, M., Grossel, G., Burgman, M. (2011). Comparison of Web-Based Biosecurity Intelligence Systems: BioCaster, EpiSPIDER and HealthMap. *Transboundary and Emerging Diseases* : pp.223-232.

Mawudeku, A., Blench, M. (2006). Global Public Health Intelligence Network (GPHIN). *7th Conference of the Association for Machine Translation in the Americas (AMTA)* .

Piskorski, J., Belyaeva, J., Atkinson, M. (2011). Exploring the usefulness of cross-lingual information fusion for refining real-time news event extraction: A preliminary study. *Proceedings of Recent Advances in Natural Language Processing* : pp.210–217.

Son, D., Quoc, H.N., Ai, K., Collier, N. (2008). Global Health Monitor - a Web-based system for detecting and mapping infectious diseases. *International Joint Conference on Natural Language Processing (IJCNLP) :* pp.951–956.

Steinberger, R. (2011). A survey of methods to ease the development of highly multilingual text mining applications. *Language Resources and Evaluation* : pp.1–22.

Tolentino, H., Kamadjeu, R., Fontelo, P., Liu, F., Matters, M., Pollack, M.P., Madoff, L. (2007). Scanning the Emerging Infectious Diseases Horizon - Visualizing ProMED Emails Using EpiSPIDER. *Advances in disease surveillance* : p.169.

Yangarber, R., von Etter, P., Steinberger, R. (2008). Content collection and analysis in the domain of epidemiology. *Proceedings of DrMED-2008: International Workshop on Describing Medical Web Resources*.