Veille épidémiologique multilingue : une approche parcimonieuse au grain caractère fondée sur le genre textuel

Gaël Lejeune

Université de Caen Basse-Normandie

École doctorale SIMEM

Thèse de doctorat

présentée et soutenue le : *16 octobre 2013*

par

***Gaël* Lejeune**

pour obtenir le

Doctorat de l'Université de Caen Basse-Normandie

Spécialité : Informatique et applications

Veille épidémiologique multilingue : une approche parcimonieuse au grain caractère fondée sur le genre textuel

Directeur de thèse : *Nadine Lucas*

M. Luigi Lancieri	Professeur	Université Lille I	(rapporteur)
M. Jose Gabriel Pereira Lopes	Professeur	Université N^{lle} de Lisbonne	(rapporteur)
M^{me} Florence Sèdes	Professeure	Université Toulouse III	(rapporteur)
M. Gaël Dias	Professeur	Université de Caen	
M^{me} Natalia Grabar	Chargée de Recherche	STL–CNRS	
M. Ludovic Tanguy	Maître de Conférences HDR	Université Toulouse II	
M. Antoine Doucet	Maître de Conférences HDR	Université de Caen	(co-encadrant)
M^{me} Nadine Lucas	Chargée de Recherche HDR	GREYC–CNRS	(directeur)

Mis en page avec la classe thloria.

Mercis

Dans les méandres du contrat doctoral, il est difficile de progresser sans assistance extérieure. C'est à mes personnes ressources, mes adjuvants, que s'adresse ce préambule.

Je remercie tout d'abord les deux personnes qui ont permis que débute ce travail de thèse et qui en ont assuré l'encadrement de bout en bout : Antoine Doucet et Nadine Lucas.

Romain Brixtel aura été un adjuvant de première classe en venant à mon secours aux moments où j'en avais le plus besoin ... ainsi que dans les autres. Des remerciements particuliers également à Charlotte Lecluze pour son soutien et sa collaboration à de nombreux projets. Je me dois de citer également Emmanuel Giguet pour les nombreuses et riches discussions que nous avons eu pendant cette thèse. Grand merci à Harold Houivet et Ludovic Thérèze pour avoir témoigné autant d'enthousiasme à l'exposé de mes travaux, votre curiosité scientifique m'a ouvert de nouvelles portes.

Le document ici présent doit beaucoup à mes relecteurs patients et désintéressés, un *großes Dankeschön* aux D$^{r\cdot}$ Chatelier, Lécluze (*again* mais avé l'accent) et Moulintraffort. Un grand nombre des expériences menées pendant cette thèse n'auraient pu voir le jour sans ma troupe d'annotateurs. Je n'ai à offrir qu'un *grandan dankon* à mes principaux contributeurs Alex, Andreï, Britta, Christophe, Dorota, Kristina, Lichao, Markus, Marta, Pavel, Vikki, Wigdan... merci aussi à tous les autres.

Toujours dans la partie recherche (mais pas que), un merci à Pierre Beust, Gaël Dias, Stéphane Ferrari, Julien Gosme, Yves Lepage, Yann Mathet, Fabrice Maurel, Jacques Vergne, Antoine Widlöcher et mes autres coéquipers d'ISLand/DLU/HULTech pour leur précieuses lumières. Merci également à Bruno Crémilleux et Bruno Zanuttini pour leur influence positive sur mes travaux.

Il m'a été extrêmement agréable pendant ces 36,48 mois de thèse de côtoyer mes excellents collègues au GREYC. Je remercie les différents wagons et locomotives du petit train du midi d'avoir fait œuvre de sustentation continue et assidue, merci à Cyril Bazin, Grégory Bonnet, Mathieu Fontaine, Jean-Philippe Métivier, Alexandre Niveau, Thibaut Vallée et consorts. La vie au labo n'aurait pas la même saveur sans des personnages tels que Jerzy Karzmarczuk, François Rioult ou Agnès Zannier : merci d'avoir partagé avec moi tant de discussions impromptues. Je n'oublie pas nos chers sysadmins sans qui rien ne serait possible (c'est cadeau), ni notre équipe administrative pour son savoir-faire et son sourire.

Je remercie également mes étudiants pour le plaisir que j'ai eu à enseigner, mention particulière à mes étudiants de projet : Benoit Samson, Poulard Charles, Nathan Didier, Anne-Lise cahu, Émile Dufournier et Igor Davy.

Merci également à l'*International Master* Maxence Godard pour m'avoir tenu en forme physique pendant toutes ces années par son habileté écœurante au squash. Une grosse pen-

sée au club d'échecs Caen Alekhine et à tous ses membres auxquels je dois beaucoup. Enfin, des remerciements collectifs à tous ceux que j'ai oublié ici et qui m'ont déjà pardonné ;-)

Il n'est pas évident d'être tributaire des contraintes inhérentes à 3 ans de thèse. Je remercie ma femme, mes enfants et mes parents d'avoir une patience dont je ne fais pas toujours preuve.

Table des matières

•••••

Introduction

Notre travail de recherche porte sur la définition et l'application de méthodes de Traitement Automatique des Langues (TAL) adaptées au traitement de corpus multilingues. L'application principale recherchée est la collecte en temps réel d'informations sur un domaine spécifique : les maladies infectieuses. Il s'agit de mettre en place un système de veille adapté au domaine de l'épidémiologie. La veille est définie comme une activité de collecte et de traitement de l'information. Elle peut concerner différents domaines, parmi lesquels on peut citer la veille technologique, la veille concurrentielle ou encore la veille juridique.

Daniel Rouach ([Rouach-2010]) définit la veille technologique comme :

> [...] *l'art de* ***repérer, collecter, traiter, stocker des informations*** *et des signaux pertinents (faibles, forts) qui vont irriguer l'entreprise à tous les niveaux de rentabilité, permettre d'orienter le futur (technologique, commercial...) et également de protéger le présent et l'avenir face aux attaques de la concurrence.*

Si une partie de cette définition est très spécifique à la veille technologique, certains éléments sont néanmoins généralisables à d'autres types de veille. Nous en retiendrons que la veille n'est pas simplement le traitement de l'information, mais que c'est aussi en amont la collecte et en aval le stockage de l'information.

L'activité de veille peut être définie comme un processus partant d'un **besoin d'information** pour lequel il sera nécessaire de **collecter** et d'**analyser** des données avant d'en extraire des **informations pertinentes** pour une tâche donnée. Du point de vue du traitement informatique, c'est l'analyse des données et l'extraction des informations pertinentes qui constituent les aspects les plus intéressants. Nous n'oublions pas toutefois qu'il existe une étroite relation entre, d'une part, l'analyse et l'extraction et, d'autre part, la définition des besoins et la collecte des données de base.

Le contexte de la veille épidémiologique

En veille épidémiologique, l'objectif est de communiquer aux autorités sanitaires les informations les plus pertinentes sur la propagation de maladies infectieuses. Ceci est d'autant plus important que la propagation peut être rapide. Lorsque la phase épidémique

est atteinte, un temps précieux pour la mise en œuvre de contre-mesures a peut être déjà été perdu.

La première question est de savoir où trouver l'information, où sont les sources pertinentes pour la veille épidémiologique.

Les sources ouvertes concentrent une partie importante des travaux dans le domaine de la veille en général et de la veille épidémiologique en particulier. La gratuité et la large accessibilité de ces données en font un élément central des systèmes de collecte et de traitement des données. La presse dite « en ligne » représente la part la plus visible de ces sources ouvertes. Les articles de journaux publiés en ligne constituent un formidable moyen d'information, avec des documents disponibles dans un nombre de langues de plus en plus important. L'augmentation du nombre de langues et du nombre d'articles disponibles permet une accélération de la vitesse de transmission de l'information. Les évènements [1] du monde « réel » sont plus rapidement retranscrits sous forme électronique, autorisant une détection plus précoce des épidémies aux quatre coins du globe.

Les informations relatives à l'actualité dans différents domaines sont *a priori* plus facilement disponibles. Cela peut concerner différents sujets. Les plus souvent étudiés restent sans doute la sécurité (terrorisme, santé...) et le monde des affaires (informations boursières notamment). Le spécialiste du domaine espère profiter de la grande disponibilité de sources journalistiques en ligne pour dénicher plus rapidement les informations stratégiques que son commanditaire recherche.

La seconde problématique est la modalité de traitement des documents. En effet, la quantité de données à traiter est de l'ordre de plusieurs dizaines de milliers d'articles de presse par jour. La forme électronique des documents invite naturellement à opter pour un traitement automatique de ces données. La puissance apparente des outils informatiques est alors utilisée pour compléter ou suppléer l'analyse humaine « traditionnelle ». Or, si la collecte de documents de manière automatique semble peu problématique, le dépouillement automatique des textes, l'extraction automatique d'informations précises et structurées restent des défis importants. L'objectif d'un système automatique est d'offrir une qualité approchant au maximum celle obtenue par un veilleur spécialiste du domaine. Les recherches sur la veille automatique de la presse visent ainsi le meilleur compromis entre la vitesse du traitement automatique et cette optimalité, réelle ou supposée, de la qualité du traitement manuel. La perte de qualité peut être acceptable si le gain en temps est significatif pour le spécialiste.

Selon Roman Yangarber ([Yangarber-2008]), les phases d'analyse puis d'extraction peuvent être rattachées à deux pans de la recherche en TAL : la Recherche d'Information (RI) et l'Extraction d'Information (EI). Le regroupement des évènements détectés par les techniques d'EI permet ensuite de limiter la redondance pour l'utilisateur final.

La veille nécessite, dans un premier temps, un filtrage des données (ici des documents).

1. L'orthographe classique « événement » est toujours admise mais nous avons choisi dans ce manuscrit de nous conformer à l'orthographe « évènement » recommandée par le Conseil Supérieur de la Langue Française depuis 1990.

Cette classification fait usuellement appel à des techniques de RI. Elle vise à donner pour chaque document une étiquette « pertinent » ou « non-pertinent » pour une tâche donnée. Les techniques d'EI permettent ensuite dans les documents pertinents, d'extraire des informations plus précises et plus structurées que la simple appartenance d'un document à une classe. Par exemple, si l'on cherche à collecter des informations sur les feux de forêts, les précisions pertinentes pourront être le lieu exact du feu de forêt, le nombre d'hectares qui ont été touchés ou encore le fait qu'y ait eu des victimes. Cette sortie, structurée selon des règles bien définies, sera destinée à faciliter l'accès à l'information, l'émission d'alertes et le remplissage de bases de données.

S'agissant de textes disponibles sur le web, un grand nombre de langues sont utilisées. Ce qui soulève une question : l'objectif recherché requiert-il une approche monolingue ou une approche multilingue ? Une autre façon de voir le problème est de savoir s'il est **possible** de traiter toute langue disponible ou seulement certaines langues bien déterminées. De nombreux systèmes, dans différents domaines de la veille, ont été conçus spécifiquement pour l'anglais. Or, l'importance de la langue anglaise sur le web est dans une phase de régression, l'offre de contenus s'élargit et la part de la langue anglaise diminue. Des systèmes limités à l'anglais souffrent d'une couverture très partielle : toutes les données disponibles ne sont pas analysées. Parmi les documents diffusés par l'agrégateur de flux *Google News*, l'anglais reste majoritaire. Mais cette majorité est toute relative : 26% des documents que nous avons collecté sur la catégorie santé de « Google News » entre septembre 2011 et septembre 2012 sont écrits en anglais. L'anglais offre la couverture monolingue la plus grande et c'est tout naturellement la première langue que l'on est tenté de traiter pour obtenir une veille efficace. Mais cette couverture est en réalité très partielle à l'échelle de flux de presse comportant des dizaines de langues. Ce qui est disponible sur *Google News* n'est que la partie la plus visible de l'offre de contenus.

Dans le domaine de la veille épidémiologique, l'exigence de multilinguisme est sans doute plus importante qu'ailleurs. Il serait en effet particulièrement dommageable de devoir attendre qu'un article en anglais (ou dans une autre langue de grande diffusion) signale une épidémie avant de pouvoir réagir. L'émission d'une alerte dès le premier article publié en langue vernaculaire serait donc une avancée considérable.

Les systèmes spécialisés dans la veille épidémiologique ne sont pas à l'heure actuelle véritablement multilingues. Ceux qui affichent le traitement de plusieurs langues sont constitués par parallélisation de systèmes monolingues. La chaîne de traitement employée dans ces systèmes est fortement dépendante de la langue traitée. En effet, les modules d'analyse locale (lemmatiseurs, étiqueteurs...) sont spécifiques à une langue ([Steinberger-2011]). La factorisation des procédures est très limitée. Pour certaines langues, de nombreux analyseurs de qualité sont disponibles. A l'opposé pour d'autres langues, ces analyseurs restent à créer. La construction de tels analyseurs n'est par ailleurs pas triviale comme en témoigne l'existence d'ateliers de travail dédiés à ce problème. Une véritable communauté de recherche existe par exemple autour de la question des langues faiblement dotées en

ressources [2] ou spécifiquement sur les langues à morphologie riche [3]. Traiter une langue supplémentaire devient alors chaque fois un peu plus coûteux, en temps de développement ou en ressources financières investies. L'élargissement pas à pas de la couverture fait que l'on traite des langues de moins en moins dotées en ressources : le **coût marginal de traitement d'une nouvelle langue est élevé**.

Vers une analyse parcimonieuse

A contrario de l'approche classique où la couverture multilingue s'obtient par parallélisation de systèmes monolingues , nous mettons en avant ici un système multilingue « par essence ».

C'est par une approche textuelle, fortement basée sur des invariants du genre, que nous proposons d'aborder la problématique du traitement multilingue. Parmi ces invariants nous examinons particulièrement la structuration interne des documents et les stratégies d'écriture utilisées par les journalistes, leur « style collectif » tel que décrit dans les travaux de Nadine Lucas ([Lucas-2000, Lucas-2012]).

De plus, nous proposons de tenir compte des stratégies de communication à l'œuvre dans ces textes, stratégies validant le rôle décisif tenu par le lecteur de l'article en tant que destinataire du message ([Coursil-2000]). Le sens que nous cherchons à extraire est donc déduit des textes eux-mêmes plutôt que de la somme des phrases qui les composent. Nous souhaitons ainsi inscrire nos travaux dans la lignée de l'approche textuelle définie par François Rastier (voir par exemple [Rastier-2008]). Nous exploitons au maximum les traces laissées par l'émetteur du texte, traces qui ne sont pas réductibles à des aspects purement lexicaux ou grammaticaux. Notre méthode s'appuie sur la **répétition** d'éléments à des **positions** remarquables dans les textes.

Nous avons implanté un système d'analyse automatique de la presse qui découle de cette approche. Nous l'avons baptisé *DAnIEL* pour *Data Analysis for Information Extraction in any Language*. Ce système est fondé sur un noyau central d'analyse indépendant de la langue traitée. DAnIEL utilise des règles du genre journalistique, robustes sur le plan multilingue. Le cœur de l'analyse est indépendant de la langue traitée. Par ailleurs, le système s'appuie sur une description locale minimale de la langue. Aucune analyse morphologique ou syntaxique n'est requise, seule une liste de termes généraux du domaine (noms usuels de maladies et de lieux) est utilisée. L'analyse au grain « mot », prépondérante en TAL, est laissée de côté au profit d'une analyse en chaînes de caractères (mots ou non-mots). Cette analyse évite par exemple d'avoir à retrouver des mots graphiques dans des langues où ils n'existent pas. Le coût marginal de traitement d'une nouvelle langue se limite à la fourniture de quelques dizaines de noms de maladie, les plus usités. Par

2. Voir par exemple la série des *Workshop on Spoken Languages Technologies for Under-resourced languages.*

3. *ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages.*

ailleurs, l'approche n'amène pas de contrainte d'organisation de la ressource, sous forme d'ontologie par exemple. Ceci facilite la constitution et la maintenance de la ressource. L'observation du vocabulaire employé par les journalistes permet en effet de ne retenir que les termes nécessaires et suffisants à l'analyse. Les noms « scientifiques » des maladies sont utilisés comme supplétifs à l'échelle du texte. Les journalistes utilisent essentiellement des termes du langage commun de manière à s'assurer que l'information est bien décodée par le lecteur. La description complète du vocabulaire spécialisé s'avère alors inutile.

DAnIEL analyse des articles de journaux sous forme électronique dans un grand nombre de langues afin de détecter des évènements épidémiologiques : l'émergence de maladies dans un ou plusieurs lieux donnés. Les résultats obtenus par DAnIEL sont très prometteurs dans la mesure où il est efficace dans les trois sous-tâches de la veille épidémiologique :

- le **filtrage** des documents, aboutissant à une classification : pertinents VS non-pertinents ;
- le **typage**, à des fins d'alerte, des évènements épidémiologiques décrits dans ces documents ;
- le **regroupement** de ces évènements, afin de limiter la redondance pour l'utilisateur final.

DAnIEL dépasse la couverture (en nombre de langues et en nombre de locuteurs) des systèmes actuels en permettant de traiter à moindre coût des langues communément considérées comme difficiles. Ce sont par exemple des langues avec une morphologie riche [4] (grec, finnois, polonais...) ou dotées d'un système d'écriture différent (chinois, arabe, hindi...). Ces langues sont généralement peu dotées en ressources, l'approche parcimonieuse que nous prônons est donc particulièrement pertinente. Notre système est aussi en mesure de rivaliser avec des systèmes de l'état de l'art sur des langues plus communément traitées par les systèmes d'EI. Bien que le domaine d'étude soit relativement spécialisé, nous démontrons que les propriétés textuelles et stylistiques ont un réel intérêt en traitement des langues. Ceci est d'autant plus vrai dans un contexte multilingue. Les invariants de genre que nous décrivons sont une réelle alternative à l'analyse descriptive habituellement à l'œuvre en EI. La collecte, la formalisation et le stockage en mémoire de formes lexicales, de règles décrivant les phénomènes langagiers sont donc ici évités.

Plan de la Thèse

La première partie de ce manuscrit est consacrée à la présentation de l'enjeu central de notre travail, la question de la couverture multilingue dans la veille épidémiologique. Le chapitre 1 propose une analyse des besoins des autorités sanitaires en terme de veille. Ces besoins s'articulent autour de deux exigences : la couverture et la rapidité. La rapi-

4. La richesse de la morphologie, mesurable par la quantité de morphèmes par mots, complexifie la tâche d'analyse locale de la langue et met en lumière les limites des approches au grain mot.

dité est un critère temporel, il s'agit de détecter les évènements aussi vite que possible. La couverture est un critère géographique, il s'agit de recevoir des informations sur des évènements quel que soit le lieu dans le monde où ils se produisent. Sans doute, un évènement épidémiologique finit toujours par être connus des autorités sanitaires tôt ou tard. Mais, ce délai varie selon le lieu concerné et la langue dans laquelle les premières retranscriptions sont disponibles. Cette double contrainte de couverture et de rapidité montre la priorité que constitue la couverture multilingue. Le chapitre 2 présente les différentes approches visant à combler le besoin d'information de ces autorités, et la façon l'exigence de multilinguisme est abordée. Nous montrons que les approches existantes, qu'elles soient manuelles ou automatiques, sont toutes freinées dans leur couverture. Nous défendons l'idée que les systèmes existants sont davantage multi-monolingues que véritablement multilingues. Leurs limites se situent dans leur coût d'extension, incompatible avec les moyens disponibles (ressources lexicales et modules d'analyse).

Nous proposons donc une approche véritablement multilingue dont nous exposons les principes dans notre deuxième partie. Les fondements opératoires de cette approche sont présentés dans le chapitre 3 : ce sont la factorisation des procédures et la parcimonie dans l'utilisation de ressources externes. Notre but est de favoriser une couverture multilingue efficace à faible coût. L'approche utilisée est **différentielle**, en ce sens qu'elle s'appuie sur les relations entre unités. Elle est **non-compositionnelle** puisque le sens d'un grain n'est pas reconstruit à partir de grains plus petits. Enfin, l'approche est **endogène**, les données de base utilisées sont limitées car nous considérons que les indices pertinents sont déjà visibles dans les textes.

Dans un objectif de parcimonie dans l'utilisation des ressources, nous proposons dans le chapitre 4 un modèle d'analyse faiblement descriptif. Ce modèle utilise des propriétés stylistiques et rhétoriques spécifiques d'un genre textuel. Ces propriétés sont observables au grain texte, au contraire des propriétés grammaticales spécifiques à chaque langue qui sont habituellement examinées au grain phrase. La dépendance au genre textuel se substitue ainsi à la dépendance à la langue. Nous proposons une analyse au grain caractère, indépendante de toute description morpho-syntaxique de la langue, basée sur un modèle de document adapté au genre textuel à traiter. Des expériences menées sur des articles scientifiques nous permettent de proposer une mise en œuvre comparative. Cette comparaison a pour but de décrire l'importance de la variation en genre pour une approche textuelle. Nous montrons ensuite comment cette approche est applicable dans le cadre d'une activité de veille. Nous examinons les manifestations des propriétés du genre journalistique et du style collectif des journalistes. Nous étudions en particulier la répétition de segments d'intérêt, pertinents pour la recherche d'information, à certaines positions. Ceci nous permet de dessiner l'architecture d'un système de veille épidémiologique parcimonieux et multilingue.

Dans la troisième partie, nous présentons le corpus de référence que nous avons construit et les résultats obtenus sur ce corpus avec notre méthode fondée sur le genre

textuel. Nous proposons ensuite des expérimentations sur l'application de la même approche à d'autres genres textuels et sur sa robustesse hors « conditions de laboratoire ». Le chapitre 5 est consacré à *DAnIEL*, le système de veille épidémiologique que nous avons conçu. DAnIEL se fonde sur un modèle de document, dépendant du genre textuel, et une analyse au grain caractère, indépendante de la langue. Après avoir décrit précisément son architecture, nous proposons une évaluation de ses performances et de la plus-value qu'il peut apporter par rapport à des systèmes classiques d'analyse. Nous exposons en détail le corpus multilingue annoté que nous avons constitué et mis à disposition de la communauté. Nous proposons également une réflexion sur les modalités d'évaluation les mieux à même de mesurer l'impact d'une approche multilingue. Nous utilisons des modalités classiques d'évaluation dont nous proposons plusieurs variations de manière à mieux coller aux attentes de l'utilisateur final. Ensuite, dans le chapitre 6, nous proposons plusieurs exemples d'application de la combinaison entre un modèle de document fondé sur le genre textuel et une analyse au grain caractère. Nous montrons que lors du traitement d'un nouveau genre textuel, seule l'adaptation du modèle de document est nécessaire. Avec notre méthode, les ajustements du système sont donc dépendants du genre analysé plutôt que de la langue. Enfin, dans le chapitre 7, nous examinons dans quelle mesure notre approche est dépendante de la qualité des documents fournis en entrée. En effet, la structure des documents disponibles dans des flux de presse, par exemple, est loin d'être normalisée. En particulier, la présence de publicités et de traces du squelette du site Web impose un pré-traitement des documents. Différents outils existent pour ce pré-traitement, nous examinons comment ces outils ont une influence sur les résultats obtenus par DAnIEL et comment ce dernier peut en retour contribuer à leur évaluation.

Première partie

État de l'art dans le domaine de la veille épidémiologique

Introduction

Pour les grandes institutions publiques ou privées, la capacité à traiter la grande quantité d'informations diffusées par les organes de presse sur l'Internet est un enjeu stratégique. Même à l'heure de *Twitter* et de *Facebook*, la presse reste un organe de diffusion capital, tout à la fois rapide et fiable.

Dans le chapitre 1, nous proposons une analyse des besoins spécifiques des acteurs de la politique sanitaire en matière de veille. Nous montrons pourquoi l'analyse des articles de presse demeure la principale manière d'obtenir rapidement des informations à coût raisonnable. Nous exposons les raisons qui motivent la couverture multilingue. Une telle couverture est, dans le domaine épidémiologique, plus encore que dans tout autre, capitale. En particulier, nous examinons la plus-value que peuvent attendre ces acteurs d'un traitement automatisé de la presse qui soit résolument multilingue.

Dans le chapitre 2, nous proposons une vue d'ensemble des systèmes de veille épidémiologique actuels. Nous présentons les modalités de combinaison des techniques manuelles de veille, des techniques classiques de Recherche d'Information et des techniques plus récentes d'Extraction d'Information. Nous opérons un distingo entre les systèmes de veille multi-monolingues et les systèmes véritablement multilingues. Les premiers ne sont en fait que la concaténation de systèmes monolingues alors que les seconds proposent une architecture et un fonctionnement compatible avec la couverture multilingue. Nous montrons comment les approches actuelles en veille traitent la question de l'extension multilingue, c'est-à-dire comment ces approches permettent d'assurer le traitement de nouvelles langues. L'analyse de leurs atouts et des contraintes rencontrées nous permet de définir les aspects sur lesquels une approche multilingue de la veille épidémiologique doit s'appuyer.

Chapitre 1

La veille épidémiologique : Pourquoi et Comment ?

Sommaire

Contexte

Les autorités sanitaires telles que l'Organisation Mondiale de la Santé (OMS [5]) ou le Centre Européen pour la Prévention et le Contrôle des maladies (ECDC [6]) s'intéressent depuis de nombreuses années à la manière d'accélérer et d'améliorer la détection et le suivi des foyers d'épidémie. Certaines initiatives dédiées à des maladies particulières offrent une information efficace et structurée générée par des praticiens. En France, les Groupes Régionaux d'Observation de la Grippe (GROG) en sont un exemple.

Se contenter de sources d'informations spécialisées est toutefois insuffisant. Le renseignement par les praticiens est certes d'excellente qualité, mais son coût est élevé puisque les praticiens doivent eux-mêmes consacrer du temps à produire l'information. Ce type de

5. En anglais, WHO pour *World Health Organisation*.
6. *European center for Disease Control*.

ressource est difficilement généralisable au niveau mondial. D'autres sources d'information sont nécessaires pour améliorer la qualité du renseignement. À ce titre, remarquons que le traitement de toute source possible est un objectif central de la lettre d'information épidémiologique de référence *ProMED-mail*[7] émise par le « Programme de Surveillance des Maladies Émergentes » (ProMED[8]).

Le Renseignement d'Origine Source Ouverte (ROSO[9]) constitue ainsi une piste de recherche importante pour les projets de veille. Il s'agit de profiter autant que possible des documents librement disponibles, notamment sur le Web. C'est un changement de paradigme vis-à-vis des problématiques traditionnelles de renseignement. Il s'agit moins de découvrir une information « cachée » que de trouver les informations pertinentes dans le vaste ensemble que constituent les corpus de documents librement disponibles. Les progrès réalisés dans l'analyse automatique de texte rendent cet objectif atteignable. Ils ont permis de mettre en lumière l'intérêt de ces sources pour le domaine épidémiologique dans une vision automatique. Le travail de collecte d'articles de presse effectué par les agrégateurs permet un accès aisé à un grand nombre de sources. C'est la raison pour laquelle l'analyse des articles de presse disponibles en ligne est devenue le thème de nombreuses recherches sur la veille épidémiologique ([Linge-2009]).

Un agrégateur tel que l'*European Media Monitor*[10] collecte près de 150.000 documents chaque jour. Un tel volume de documents rend le dépouillement manuel des sources disponibles coûteux, sinon impossible, à réaliser. Nous cherchons donc dans ce chapitre à cerner au mieux les besoins exprimés par les autorités sanitaires (Section 1.1) de manière à pouvoir offrir une réponse adaptée aux problèmes identifiés (Section 1.2).

1.1 Le besoin d'information des autorités sanitaires

Les autorités sanitaires cherchent à optimiser leur réaction face aux épidémies. Leur objectif premier est de collecter toutes les informations possibles sur les maladies infectieuses et leur dissémination à travers le monde. Elles souhaitent restreindre le nombre d'informations qui pourrait leur échapper, autrement dit limiter le **silence**. Dans le même temps, ne pas avoir à traiter des documents non pertinents est capital pour ces organismes. Les fausses alertes doivent être rares, le **bruit** doit être restreint.

En premier lieu, les foyers d'épidémie doivent être détectés afin d'évaluer le risque de propagation (Section 1.1.1) ; et ceci le plus tôt possible (Section 1.1.2). Diversifier les sources d'information est capital pour réaliser cet objectif (Section 1.1.3).

7. `http://www.promedmail.org/aboutus/`
8. *Program for Monitoring Emerging Diseases.*
9. OSINT en anglais pour *Open Source INTelligence.*
10. `emm.newsexplorer.eu/NewsExplorer/`

1.1.1 Définition de l'évènement épidémiologique

En matière de maladies infectieuses, la collaboration entre différents états est requise pour contrôler l'expansion des agents pathogènes (virus, bactéries, parasites ...). Pour cela, l'OMS donne des consignes précises de signalement des évènements épidémiologiques à ses états membres ([OMS-2005]). Surveiller les maladies dans leur zone endémique permet de suivre l'évolution des phénomènes infectieux et de mettre à jour les instructions sanitaires données aux voyageurs se rendant dans ces zones. Il est également nécessaire de repérer et de maîtriser les épidémies émergentes. La contamination de zones limitrophes et la vitesse d'expansion de l'épidémie sont des caractéristiques fondamentales et donc particulièrement examinées. L'objectif est d'éviter une phase pandémique où l'étendue géographique du phénomène serait telle que le confinement deviendrait impossible. L'OMS insiste particulièrement d'une part sur l'importance de l'identification et de la localisation de la menace, et d'autre part sur la rapidité du signalement. Cette même préoccupation figure dans d'autres études effectuées pour le compte d'autres autorités sanitaires ([Smolinski-2003, Wilson-2009]). Pour le Système Mondial d'Information sur la Santé Animale (WAHIS pour *World Animal Health Information System*), organisme chargé au niveau mondial de la surveillance des épidémies affectant spécifiquement les animaux (épizooties), le critère géographique et le critère temporel revêtent également un caractère central [11].

Nous en déduisons la définition suivante : l'évènement épidémiologique est constitué par l'existence de cas problématiques d'infection par un agent pathogène donné. Chacun de ces cas désigne la contamination d'un ou plusieurs sujets par une **maladie** donnée, dans un certain **lieu** à un moment dans le **temps**. Une part importante du travail de veille consiste à déterminer si les cas de contamination sont suffisamment importants pour constituer un évènement épidémiologique à part entière. Un certain nombre d'indications supplémentaires peuvent être ajoutées comme le nombre de cas signalés ou encore la source d'information ayant permis le signalement.

Dans le cadre de cette thèse, nous ne décrivons pas les mesures que peuvent prendre les autorités sanitaires à l'issue du signalement d'un évènement épidémiologique. Notre étude se limite à la **sélection** des documents pertinents pour la veille épidémiologique et à l'**extraction** automatique d'évènements épidémiologiques à partir de ces documents. C'est, de notre point de vue, sur ces deux aspects que le besoin d'information des autorités sanitaires peut être le mieux satisfait. Il serait évidemment intéressant d'examiner comment pouvoir de façon automatique conseiller la réponse adéquate face à la survenance d'une épidémie. Il nous semble que cette partie incombe au spécialiste et que tenter d'interpréter l'évènement signalé à sa place relève d'une autre étude.

Notre objectif est de créer un système permettant de détecter avec efficacité un maxi-

11. Voir par exemple la description du format des rapports de signalement d'épidémies demandés aux états membres à l'adresse suivante : `http://web.oie.int/fr/info/fr_info.htm` (consulté le 28 juin 2013).

mum d'évènements le plus précocement possible. Il est nécessaire, d'une part, de diminuer de façon notable le nombre de documents que l'épidémiologiste devra consulter et, d'autre part, de s'assurer que les évènements lui sont signalés aussi tôt que possible. Nous cherchons donc à traiter un maximum de sources de façon à minimiser le risque d'ignorer un document potentiellement pertinent. De cette façon, nous favorisons la détection rapide d'épidémies potentielles.

1.1.2 Vers une réaction au plus près de la première alerte

En matière de signalement d'un évènement épidémiologique, il existe pour les autorités sanitaires trois étapes principales.

La première est l'apparition du cas d'infection lui-même : une personne est contaminée par un agent pathogène. Nous nommons $t1$ la date où les premiers cas d'infection surviennent.

La seconde étape a lieu lorsque l'information est disponible sur un canal de diffusion quelconque, cela correspond au moment où est publiée la première information sur les cas survenus. Ce canal peut être un canal officiel de diffusion d'information sur le domaine épidémiologique, par exemple un cas repéré par un médecin qui propose une rediffusion de l'information qu'il a recueillie lors d'une consultation. Ce canal peut aussi être une source ouverte (article de presse, flash d'information radiodiffusé ou télédiffusé...). Nous dénommons $t2$ la date de cette publication.

La troisième étape est atteinte lorsque les autorités sanitaires ont connaissance de ce cas d'infection. À partir de ce moment, elles peuvent mettre en place les mesures nécessaires à la limitation de la propagation de l'infection. Nous appelons $t3$, le moment où une autorité sanitaire compétente décrit officiellement une série de cas d'infection comme faisant partie d'un évènement épidémiologique.

Nous résumons donc de la façon suivante :

- $t1$: date des premiers cas de contamination ;
- $t2$: date de la première publication se rapportant à ces cas ;
- $t3$: date où l'information parvient à l'autorité sanitaire qui juge si la maladie est en phase épidémique.

Nous dénommons :

délai de publication Le temps écoulé entre $t1$ et $t2$;

délai de signalement Le temps écoulé entre $t2$ et $t3$;

délai de connaissance officielle Le temps écoulé entre $t1$ et $t3$.

Le but principal de notre étude est de réduire le délai de signalement. Plus ce temps est court, plus la réponse pourra être efficace. Si le canal intervenant à la phase $t2$ est lié de près au domaine épidémiologique, le délai de signalement sera quasi-nul. L'informa-

tion sera en effet transmise plus rapidement aux autorités sanitaires et dans un format directement exploitable.

Pour illustrer l'importance de la question temporelle, nous prenons l'exemple de l'épidémie de Pneumonie Atypique. Cette maladie est plus connue sous son acronyme anglais *SARS*[12] ([Heymann-2004]) ou français *SRAS*[13]. Cette épidémie a réactualisé la problématique de la surveillance épidémiologique à grande échelle ([Arnold-2013]). Elle a été une des causes de la modification des règles de coopération internationale édictées par l'OMS ([Baker-2007, Wilson-2009]). Dans la Figure 1.1 sont représentés les délais de publication, de signalement et de connaissance officielle pour cette épidémie.

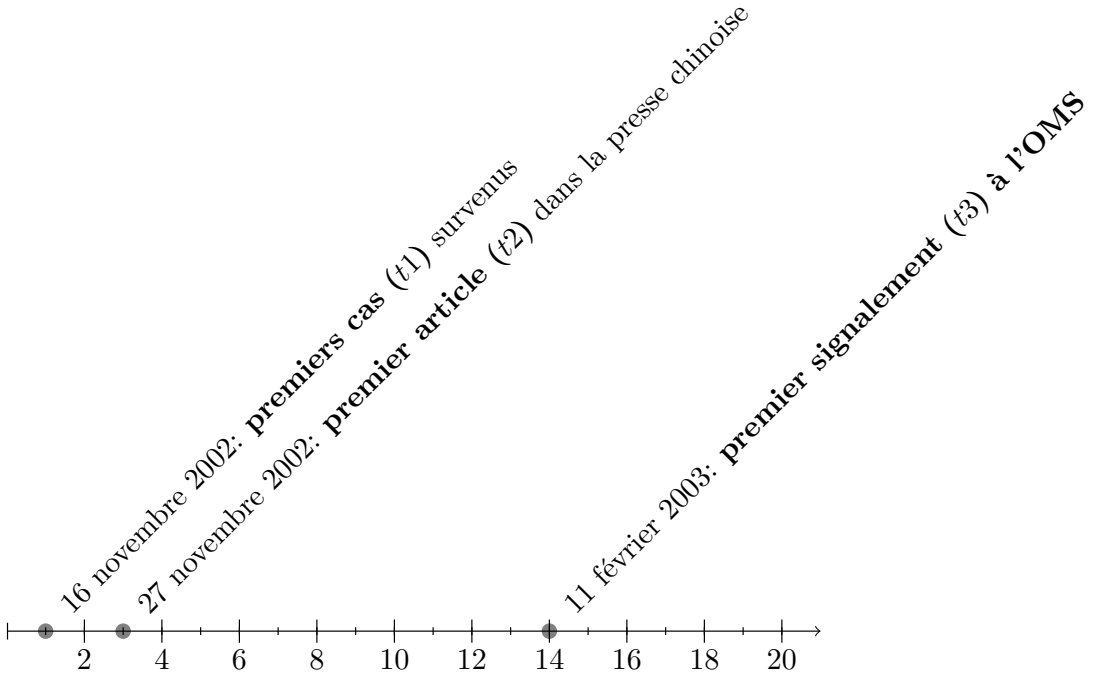


FIGURE 1.1 – L'épidémie de SRAS de 2002-2003 en Chine : dates des premiers cas de contamination ($t1$), de la première publication ($t2$) et date de connaissance officielle par l'autorité sanitaire ($t3$)

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, est pourtant atteint en moins de 2 semaines. Le **délai de publication** était relativement court mais le **délai de signalement** a été particulièrement long : 11 semaines.

12. *Severe Acute Respiratory Syndrome*.
13. Syndrome Respiratoire Aigu Sévère.

La Figure 1.2 montre une version enrichie de la figure précédente : nous ajoutons les principales étapes de propagation. Nous constatons que pendant le délai de signalement, le nombre de cas a fortement augmenté. Ce nombre (m dans la figure) est ainsi passé de l'ordre de la dizaine en $t2$, à plusieurs centaines peu avant $t3$. Le retard dans le 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.

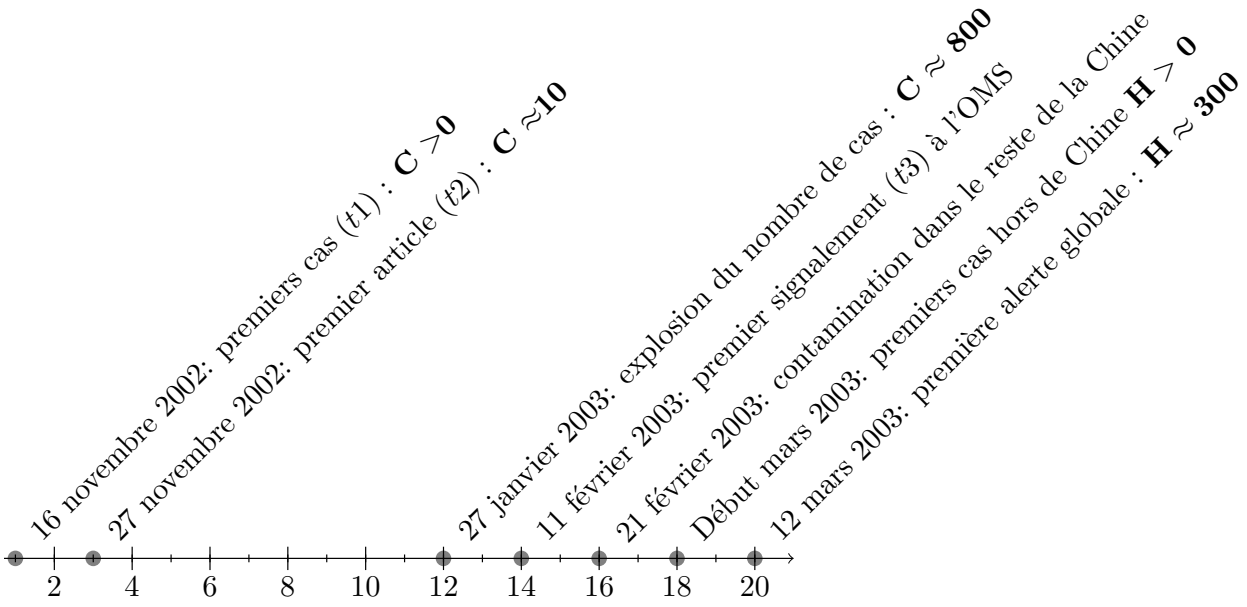


FIGURE 1.2 – L'épidémie de SRAS de 2002-2003 en Chine : principales étapes de propagation et de signalement sur les 20 premières semaines de l'épidémie avec C le nombre de cas en Chine et H le nombre de cas hors de Chine

Plusieurs articles de la presse chinoise sur le sujet ont été publiés dans les premières semaines de l'épidémie. Généralement, la connaissance de ce type d'évènement se propage rapidement à travers de nouveaux articles qui reprennent et développent les premières informations diffusées. Lorsqu'un nombre grandissant de sources relatent l'évènement, il est de plus en plus probable que les autorités compétentes soient mises au courant.

La politique des autorités sanitaires chinoises visant à « dissuader » les journaux de parler de la propagation de la mystérieuse maladie [14] a au contraire limité la propagation de l'information et occasionné une augmentation du délai de signalement. Malgré cette

14. *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`, article consulté le 28 juin 2013.

censure, l'un ou l'autre des premiers articles diffusés aurait pu permettre une détection anticipée de l'évènement si les autorités sanitaires avaient eu les outils pour les analyser automatiquement dans les premiers temps suivant leur publication.

L'épisode du SRAS de 2002-2003 en Chine a constitué un tournant pour la surveillance de la progression des épidémies au niveau mondial. Les délais de connaissance officielle ont depuis lors été réduits, tant grâce aux sources officielles que grâce aux sources non officielles ([Mondor-2012]). Cette épidémie nous montre qu'il est important de tenir compte de toutes les sources possibles. Ceci doit être un objectif majeur pour réduire les délais de signalement en profitant notamment de la complémentarité des sources gouvernementales et non gouvernementales mise en lumière à l'issue de l'épidémie de SRAS en 2002.

1.1.3 Détection des évènements épidémiologiques à partir de n'importe quelle source

Il serait sans doute idéal de n'avoir à analyser qu'un nombre limité de sources pour combler efficacement le besoin d'information des autorités sanitaires. Il suffirait de choisir d'emblée les « bonnes sources » pour pouvoir extraire à moindre coût les informations recherchées. Néanmoins, il est difficile d'attester que cette sélection des sources ne se fait pas au détriment du rappel, surtout en ce qui concerne les sources ouvertes. Des recherches récentes ont par ailleurs montré que la diversification des sources pouvait apporter une réelle plus-value à la détection d'évènements épidémiologiques ([Cataldi-2010]).

Des systèmes institutionnels assurant la collecte et la dissémination d'information sur le domaine épidémiologique tels que les GROG proposent une information de qualité mais avec un coût important : le temps passé par les médecins acteurs du réseau pour collecter et mettre en forme les informations. Par ailleurs, cela suppose que ces acteurs aient connaissance de cas de grippe par le biais de leurs consultations.

Les sources non institutionnelles signalent régulièrement des évènements plus rapidement que des sources officielles ([Mondor-2012]). La multiplicité d'observateurs qui s'expriment au travers des canaux non officiels (articles de presse ou plus récemment réseaux sociaux par exemple) offre une large couverture des évènements. La complémentarité de ces deux types de sources est donc importante. Les sources ouvertes impliquent toutefois des traitements particuliers. En effet, l'information intéressant l'autorité sanitaire n'est pas d'emblée disponible sous une forme exploitable.

Par opposition, les informations émanant des sources officielles suivent, en général, un schéma bien défini destiné à faciliter le travail du destinataire de l'information. Seul ce qui est important est transmis et, pour favoriser le traitement, les informations sélectionnées sont transmises à l'aide d'un formalisme bien particulier. C'est ce qui est appelé en informatique une **sortie structurée**.

Par exemple, les données transmises par les acteurs (ou vigies) du réseau des GROG suivent des contraintes particulières d'organisation.

Le rapport doit contenir les éléments suivants ([Grog-2012]) :

- la localisation de la vigie ;
- le décompte du nombre de patients ;
- le diagnostic des tests rapides de surveillance ;
- la description détaillée des cas confirmés.

Obtenir une telle structuration de l'information de façon automatique nous paraît difficilement envisageable pour ce qui concerne les sources non officielles. Nous pensons donc que ce travail relève des compétences du spécialiste du domaine. C'est ici où son expertise est la mieux valorisée alors que c'est en amont que se situe la plus-value des systèmes automatiques. Au-delà d'un certain seuil, enrichir les informations extraites se ferait au détriment de l'augmentation de la couverture. Plus l'information à extraire est complexe, détaillée plus le traitement d'une nouvelle source ou d'une nouvelle langue est coûteux. La couverture et la complexité sont deux objectifs difficiles à mener de front ; il convient donc à un certain stade d'effectuer un choix. Préparer le travail du spécialiste du domaine en filtrant et classifiant les documents disponibles offre à notre sens la meilleure plus-value. C'est de cette manière que nous concevons notre projet d'instrumenter la veille en facilitant un traitement efficace de l'information à un coût raisonnable.

Le but est donc de disposer d'un système de signalement des faits épidémiologiques, y compris quand les signaux sont faibles (peu de canaux ou de documents sur un fait particulier). Pour cela il faut traiter un maximum de sources, pour optimiser la couverture (favoriser le rappel), tout en limitant le nombre de fausses alertes émises (optimiser la précision). L'ambition de multilinguisme est particulièrement pertinente dans les domaines qui ont trait à la sécurité ([Atkinson-2013, Tulechki-2013]).

Dans le domaine de la recherche documentaire, le **rappel** est la proportion de documents pertinents sélectionnés par un système en fonction du nombre de documents pertinents connus pour une requête donnée. Les documents pertinents non retournés par le système constituent le **silence**. Lorsque l'on traite autre chose que des données de référence, lorsque le corpus n'est pas fermé, le rappel est difficile à évaluer : il faut savoir que des informations manquent. C'est l'objectif primordial de la couverture multilingue que nous recherchons : limiter au maximum la proportion de documents qui ne peuvent être traités par le système.

La **précision** est la proportion de documents qui sont pertinents parmi tous les documents qui ont été sélectionnés par le système. Les documents qui ont été sélectionnés mais ne sont pas pertinents constituent le **bruit**. La précision est généralement plus facile à mesurer puisqu'il semble envisageable d'analyser l'ensemble des documents retournés par un système.

De façon très classique, nous avons pour objectif de trouver le meilleur compromis entre rappel et précision. Toutefois, du fait du domaine traité et de notre objectif multilingue, le rappel revêt un intérêt tout particulier. Les coûts de production et de fonctionnement du système d'alerte sont également des éléments importants dès lors qu'ils affectent la

robustesse et la pérennité du système.

1.2 Instrumentation et outillage de la veille

Nous détaillons dans cette section l'objectif opératoire de la veille épidémiologique multilingue. Nous montrons que le besoin d'information exprimé par les autorités sanitaires doit être satisfait par l'augmentation de la couverture des sources disponibles (Section 1.2.1). La grande quantité de données ainsi obtenue impose de filtrer les documents pertinents pour la veille épidémiologique et d'établir une classification efficace des évènements rapportés dans ces documents (Section 1.2.2).

1.2.1 Maximisation de la couverture par le traitement de nouvelles langues

L'exemple de l'épidémie de SRAS en Chine (cf. Figure 1.2 page 24) a montré l'importance du traitement multilingue. En effet, il est important de détecter un évènement épidémiologique quel que soit la langue dans laquelle il est signalé en premier.

Cette langue est communément la langue vernaculaire, la langue qui est utilisée en priorité dans la zone géographique concernée. Ne pas être en mesure de traiter cette langue implique mécaniquement une augmentation du délai de détection de l'évènement épidémiologique. Il faudrait dans ce cas attendre que l'évènement soit retranscrit dans une langue de plus grande diffusion. Traiter un grand nombre de langues permet au contraire de réduire le délai de détection de l'évènement en réduisant le délai de signalement. L'analyse automatique des rares documents en chinois ayant relayé les premiers cas de SRAS aurait sans doute permis un gain de temps précieux à l'OMS.

À l'inverse, certains évènements peuvent même être « indétectables », si l'on ne dispose pas des outils permettant de collecter et de traiter les documents diffusés.

1.2.2 Filtrage des documents disponibles et extraction des alertes épidémiologiques

Le filtrage des documents et leur catégorisation sont deux étapes intimement liées qui n'offrent cependant pas le même intérêt pour le destinataire de la veille. La plus-value n'est pas du même ordre. Le filtrage offre un gain principalement quantitatif, en éliminant les documents non-pertinents le nombre de documents que le spécialiste doit consulter est réduit. La catégorisation, dans le cas de la veille sous forme d'un triplet maladie-lieu-date, offre un gain qualitatif : les documents sont regroupés par évènement décrit pour éliminer la redondance.

Filtrer les documents permet donc de réduire la quantité de données qui reste à analyser pour le spécialiste du domaine. Cela permet d'éviter que l'épidémiologiste ait à consul-

ter de grandes quantités de documents lorsqu'en réalité une faible proportion d'entre eux l'intéressent véritablement. Dans notre travail, le premier objectif est de déterminer si un document est pertinent ou non pour la surveillance des épidémies. Plus précisément, nous cherchons à déterminer si un document apporte ou non des renseignements importants sur un évènement épidémiologique. La fonction de filtrage du système de veille vise à constituer *a minima* deux classes de documents :

- celle des documents pertinents, susceptibles d'intéresser un épidémiologiste ou une autorité sanitaire ;
- celle des documents non pertinents, qui relatent des faits secondaires ou sans lien avec le domaine épidémiologique.

Le but est de réduire massivement le nombre de documents que le spécialiste du domaine aura à consulter pour obtenir l'information qui l'intéresse.

Catégoriser les documents aboutira à une classification plus fine. L'objectif est, dans la classe des documents pertinents, de réduire la redondance. Un système utile pour le spécialiste doit proposer de regrouper les documents dans des classes cohérentes selon les caractéristiques des évènements qui y sont rapportés. Avec notre définition de l'évènement épidémiologique, il est possible de déterminer plusieurs types de classification :

- par maladie concernée ;
- par lieu concerné ;
- par paire maladie-lieu ;
- par date.

D'autres classifications utilisant d'autres traits ou d'autres combinaisons de traits sont évidemment envisageables. Il est possible d'opérer des regroupements par familles de maladies, par type de germe, par mode de transmission, par fonction vitale affectée ou encore par morbidité [15].

Quel que soit le degré de précision de la classification, l'objectif reste, à notre sens, de limiter la redondance. Il est souhaitable de regrouper les documents très proches par leur contenu pour que le spécialiste du domaine puisse travailler efficacement. Il faut donc regrouper entre eux les documents apportant pas ou peu de plus-value informationnelle les uns par rapport aux autres. La phase de catégorisation des évènements vise donc à enrichir le filtrage.

Synthèse

La profusion de sources disponibles en ligne représente une formidable opportunité pour la détection de foyers d'épidémie. L'existence de sources dispersées sur l'ensemble du globe permet aux autorités sanitaires d'envisager une meilleure réactivité face à la

15. La morbidité mesure la vitesse de propagation d'une maladie donnée. Par exemple, le nombre de personnes qui risquent d'être infectées en un mois.

survenance de nouvelles épidémies. Il convient tout de même de filtrer, sélectionner et classer les dizaines de milliers de documents qui sont publiés chaque jour. Le prochain chapitre est donc consacré aux systèmes qui tentent à l'heure actuelle de combler les attentes des autorités sanitaires, et aux approches qui sous-tendent ces systèmes.

Chapitre 2

Veille épidémiologique, perpectives multilingues

Sommaire

Dans ce chapitre, nous examinons différentes approches dédiées à la veille épidémiologique, en prenant comme fil conducteur le « degré d'automatisation » qu'elles impliquent. Ce qui relève des principes de la veille, au niveau monolingue, est ici traité séparément de ce qui relève de la veille à visée multilingue. Après une présentation de l'éventail de systèmes existants, nous décrivons ce qui les relie en matière d'objectifs et en matière d'approche du traitement automatique des langues.

En mettant en avant, dans ce chapitre comme tout au long de ce manuscrit, la dimension multilingue, nous choisissons d'étudier séparément la dimension « multi-genre » (au sens de genre textuel). Ce choix n'implique pas un désintérêt pour les publications plus spécifiques au monde de l'Internet tels que les blogs ou les réseaux sociaux, publications qui ont une importance grandissante dans le domaine de la veille épidémiologique ([Denecke-2012]). Nous choisissons simplement de nous concentrer sur les invariants de genre, c'est à dire les propriétés du genre qui restent valables quelle que soit la langue du texte.

Considérons deux axes pour le traitement des textes en langue naturelle : le genre de texte et la langue utilisée. L'hypothèse est que l'axe du genre offre plus de variabilité que l'axe de la langue. Autrement dit, traiter n textes du même genre écrit dans n langues différentes est plus facile que traiter n textes dans la même langue appartenant à n genres différents. Dans notre approche fondée sur le genre, nous différencions fortement un *Tweet* en français traitant de la grippe d'un article de presse en français traitant de la grippe. Au contraire, nous postulons qu'il existe nombre de points communs entre des articles de presse traitant de la grippe même s'ils sont écrits dans dix langues différentes. L'état de l'art présenté ici est donc consacré exclusivement, ou presque, à la veille sur le genre journalistique.

La part prise par le traitement automatique et la taille de la couverture en nombre de langues sont deux dimensions particulièrement pertinentes dans le cadre de la veille. Dans la section 2.1 différents degrés d'automatisation de la veille sont étudiés : du traitement manuel des données au traitement entièrement automatisé en passant par le traitement assisté par ordinateur. Puis, dans la section 2.2 nous examinons de quelle manière les systèmes multilingues actuels se sont globalement construits par accumulation d'architectures monolingues existantes. Les réponses à apporter aux besoins de veille multilingue définies dans le chapitre précédent se déclinent en deux aspects : vitesse de réaction et couverture du plus grand nombre de sources.

2.1 Principes généraux de la veille monolingue

Pour répondre à un besoin d'information dans n'importe quel domaine, il semble naturel de procéder par étapes et de s'intéresser en premier lieu au traitement d'une seule langue. Dans cette vision, le traitement de plusieurs langues amène à la construction « stratifiée » d'un système multilingue cumulatif, langue après langue.

L'ordre de construction d'un système peut être influencé par la volonté de commencer par les langues les plus aisées à traiter ou par les langues plus pertinentes pour la tâche. La première langue traitée peut être une langue stratégique vis-à-vis de l'objectif recherché : par exemple le russe lorsque l'on s'intéressait aux renseignements sur la situation de l'URSS dans les années 1970. Ce peut être aussi une langue supposée plus simple à traiter, constituant par conséquent un point de départ réaliste sur le plan opératoire. Dans différents domaines, il est souvent considéré que le traitement de l'anglais constitue un pré-requis indispensable à la crédibilité d'un système de veille.

En matière de veille épidémiologique, comme dans d'autres domaines, le traitement de l'anglais est généralisé. Ceci est rendu possible grâce aux nombreuses ressources humaines et informatiques disponibles. Nous verrons ici comment se sont construits les systèmes monolingues et l'influence que cette construction a eu sur les approches dédiées au multilinguisme.

Les principes et techniques de la veille manuelle constituent une référence (Section 2.1.1).

La veille manuelle peut toutefois bénéficier pour certaines tâches spécifiques, telles que le filtrage des documents, de l'apport de techniques automatiques (Section 2.1.2). La phase ultérieure de l'automatisation est classiquement de simuler plus avant l'analyse humaine en extrayant des informations spécifiques à l'aide de techniques d'extraction d'information (Section 2.1.3).

2.1.1 La veille manuelle : mètre-étalon du domaine ?

La veille manuelle monolingue peut constituer une référence de base pour déterminer les étapes du processus de veille. Le but du processus est d'établir, à partir d'un corpus de documents, une photographie de l'information disponible sur un ou plusieurs domaines d'intérêt. Le veilleur devra disposer d'une sorte de cahier des charges, lui permettant de définir ce qu'il doit extraire des documents. Son travail consiste schématiquement à classer des documents en fonction de catégories bien déterminées et d'autre part à compiler dans des fiches les informations détectées et leurs occurrences dans les documents.

La détermination du cahier des charges est une tâche essentielle, puisque cela doit permettre de discriminer ce qui est important de ce qui ne l'est pas ([CDC-2004, ECDC-2006]). Si le cahier des charges est trop libre, la structuration de l'information risque d'être fluctuante et donc peu fiable sur la durée : par exemple, la même information extraite par deux personnes pourrait figurer dans des catégories différentes. Qu'il soit au contraire trop strict et des informations d'un type nouveau ou imprévu pourraient être difficiles à intégrer. De plus, une trop grande précision dans la description implique un coût de traitement supérieur : un nombre plus important de descripteurs sont à rechercher, un plus grand nombre de champs sont à renseigner pour chaque document.

La veille manuelle est fortement affectée par le passage à l'échelle : le coût est proportionnel à la couverture et à la vitesse recherchée. À grande échelle, le coût peut devenir prohibitif. Les ressources importantes à mettre en œuvre pour réaliser la veille manuelle impliquent alors des choix. Mais, il peut être difficile de se contenter de traiter simplement 10 ou 20% des documents disponibles. D'autre part, il faut éventuellement accepter de traiter les données avec un certain délai d'avoir un retard continuel sur le traitement du flux d'information ([Chan-2010]).

Il existe toutefois la possibilité de préparer en amont le travail du veilleur en opérant un pré-traitement automatique pour limiter le travail de l'humain. Ceci a pour but de limiter l'intervention de l'humain aux opérations pour lesquelles il est réellement compétitif face à l'ordinateur. Ces techniques d'assistance permettent alors d'aboutir à la veille semi-automatisée.

2.1.2 La veille semi-automatisée : traitement de sources pré-filtrées

La « veille semi-automatisée » est définie comme un processus où une partie décisive des différentes étapes de traitement de l'information est faite automatiquement. L'étape

où la plus-value offerte par le passage au traitement automatique reste la plus évidente est la collecte des données. C'est en effet une tâche pour laquelle les techniques automatiques offrent un très bon compromis coût-efficacité. Ce rôle est dévolu aux agrégateurs de flux *RSS* qui permettent de regrouper en un même canal un nombre important de sources d'information.

Après la collecte, le bénéfice le plus important offert par le du traitement automatique concerne le pré-filtrage des sources. Pouvoir exclure les sources manifestement non-pertinentes pour le domaine étudié permet de désengorger la chaîne de traitement en aval, là où le processus est géré par l'humain. Par exemple, il est fort peu probable de découvrir une information pertinente pour la surveillance de la fièvre hémorragique Ebola au Mali dans *Sports Illustrated*[16]. Dès lors, la plus-value du traitement exhaustif de cette source est très faible par rapport au coût supplémentaire de traitement impliqué. Il est alors intéressant d'exclure cette source afin de limiter la quantité de données à traiter avec un impact minimal sur le rappel.

Plus le domaine est spécialisé, plus ce pré-filtrage est utile. Il est plus aisé de définir les sources les plus susceptibles de procurer des informations pertinentes pour un domaine précis. Nous avons étudié les documents issus des catégories « santé » de Google News et de différents journaux. Cette étude révèle que dans la catégorie « santé », la proportion d'articles non pertinents pour l'activité de veille épidémiologique est d'environ 93%. Sur un flux non-filtré, c'est à dire en examinant l'ensemble des catégories de *Google News*, cette proportion est supérieure à 99,5%.

Par ailleurs, la partie étudiée du corpus non-filtré ne comportait pas de documents pertinents absents de la catégorie « santé ». Le gain d'efficacité obtenu est donc très important et la perte d'information faible. Quand l'humain est impliqué dans la boucle, la coûteuse et inutile analyse exhaustive de l'ensemble des documents est évitée ([Yangarber-2008]). Les premières publications de la *newsletter* ProMED-Mail concernaient ainsi l'extraction par des humains d'évènements épidémiologiques contenus dans des sources en anglais préalablement sélectionnées par des experts du domaine. La perte d'information, lorsqu'un document pertinent est exclu par erreur en amont, a en effet été jugée dérisoire par rapport au gain en productivité.

2.1.3 La veille automatique : l'extraction d'information

Pour aller plus loin dans le processus d'automatisation, l'objectif est cette fois d'automatiser le processus d'analyse. Bien plus que le filtrage des documents, cette phase fait appel à des notions de compréhension fine des textes. Le classement des documents en « potentiellement pertinents » ou « non-pertinents » ne justifie pas un examen approfondi et minutieux de chacun des textes. Ce traitement en profondeur est réservé à des documents dûment sélectionnés.

16. Magazine sportif publié aux États-Unis.

C'est à ce stade que se situe le champ de l'Extraction d'Information (EI). Son objet est de déléguer à la machine la tâche d'identification des informations pertinentes, par opposition au filtrage qui se borne à la sélection des documents pertinents. C'est la phase de remplissage des champs, prévue par le cahier des charges. Après une définition de l'EI, une description synthétique des processus qu'elle implique est proposée.

Définition de l'extraction d'information

Prenons comme définition globale de l'EI celle proposée par Rik de Busser ([DeBusser-2006a]) :

> *Information extraction is the identification, and consequent or concurrent classification and structuring into semantic classes, of specific information found in unstructured data sources, such as natural language text, making the information more suitable for information processing tasks.*

Notre traduction :

> L'extraction d'information consiste en l'identification, la classification (concurrente ou postérieure) et la structuration en classes sémantiques d'informations spécifiques, extraites de données non-structurées telles que les textes en langue naturelle, rendant ainsi l'information plus adaptée à un traitement automatique.

L'automatisation de la phase d'analyse profonde des textes amène à ne laisser à l'humain que la responsabilité de l'action à mener en fonction des informations qui lui sont transmises. C'est à la machine qu'incombe la phase de *dépouillement des textes*, auparavant confiée à l'humain. La forme structurée de la sortie proposée par la machine facilite la manipulation des données, la navigation dans les documents et la découverte de tendances ([Yangarber-2011b]). Le travail décisionnel de l'humain est donc grandement facilité. C'est la principale plus-value attendue d'un système d'EI. Il est maintenant important de définir ce qui va permettre à la machine de classer et typer l'information avec l'efficacité attendue. C'est ce que l'on nomme dans le domaine de l'EI les patrons ou *patterns*. Il s'agit usuellement de phrases types qui permettent d'identifier dans les documents les informations pertinentes pour une tâche donnée.

Les patrons au centre du processus

L'architecture des systèmes d'EI a été décrite par Hobbs ([Hobbs-1993]) dans le cadre historique des *Message Understanding Conference* ([MUC-1991, MUC-1992, MUC-1993]). Si les techniques ont bien entendu évolué dans l'intervalle, les grandes lignes du processus ont été peu modifiées ([DeBusser-2006b]). Le cœur du problème est l'énumération puis la reconnaissance des patrons permettant d'identifier et de typer les segments d'information pertinents ([Cowie-1996]).

La définition donnée par Riloff ([Riloff-1999]) est la suivante :

> *« IE systems extract domain-specific information from natural language texts. [..] Domain-specific patterns (or something similar) are used to identify relevant information. »*

Notre traduction :

> « Les systèmes d'EI extraient de l'information spécifique à un domaine à partir de textes en langue naturelle. [...] Des patrons spécifiques au domaine sont utilisés pour extraire l'information pertinente. »

C'est au niveau de la phrase ou de la proposition que les patrons sont recherchés. Ces patrons doivent représenter l'expression la plus abstraite possible d'un certain nombre d'énoncés pertinents connus.

Une définition plus récente ([Hobbs-2010]) apporte certaines précisions sur les éléments recherchés :

> *« Information Extraction (IE) techniques aim to extract the names of entities and objects from text and to identify the roles that they play in event descriptions. »*

Notre traduction :

> « Les techniques d'extraction d'information (IE) visent à extraire des entités et des objets dans des textes, et à identifier le rôle qu'ils jouent dans la description d'évènements. »

Il convient donc de définir la notion d'évènement, spécifique à un domaine, au travers des entités qui le composent et des relations que ces entités entretiennent entre elles. Pour extraire l'ensemble des cas de contamination recensés dans un flux de presse, il faudra se baser sur un nombre significatif d'énoncés connus décrivant des évènements épidémiologiques.

L'évènement peut être réduit à une Paire $Maladie - Lieu$ comme dans les travaux de Collier ([Collier-2006]). De cette façon, le regroupement des documents au sein des sujets d'intérêt, au sens de *topic* dans *Topic Detection and Tracking*, est rendu possible. Cette technique permettra pour l'utilisateur final une vision synoptique des évènements épidémiologiques et une limitation des redondances ([Linge-2009]). Pour faciliter les comparaisons, et donc l'évaluation des systèmes, un évènement épidémiologique est défini comme l'existence de cas d'une **maladie** dans un **lieu** donné sur une certaine période de **temps**. Un évènement épidémiologique, c'est donc une Paire Maladie-Lieu que l'on peut situer dans le temps ([Chan-2010]).

Certains travaux soulignent *a contrario* que la plus-value du système automatique devrait résider également dans un raffinement de l'information ([Yangarber-2008, Breton-2010]). Cette plus-value peut être, par exemple, le classement des évènements selon leur gravité (simples contaminations ou décès) ou le degré de nouveauté (nouvel évènement,

suivi d'une épidémie en cours...). La plus-value offerte par l'analyse automatique peut aussi consister en la détection d'indices annonçant un risque de pandémie ([Reilly-2008]).

La définition des patrons phrastiques ou sous phrastiques dépend donc d'énoncés types. Ces énoncés peuvent être exploités de deux façons différentes. Dans une approche descriptive, un jeu de motifs récurrents est déduit de ces énoncés. Dans une approche orientée « apprentissage », ces énoncés fourniront également des indices permettant de découvrir de nouveaux motifs automatiquement. Il est également possible de recourir à un expert du domaine pour créer ou valider les motifs qui serviront à l'extraction ([Breton-2012]). Dans les deux cas il s'agira de trouver le meilleur compromis rappel-précision pour le but recherché, afin de maximiser la plus-value offerte par le traitement automatique ([Chanlekha-2010]).

Déterminer les patrons

Selon le niveau d'abstraction offert par les patrons (patrons de mots, de lemmes, d'étiquettes morpho-syntaxiques...), un pré-traitement plus ou moins important est nécessaire pour les identifier dans des textes. Le but des motifs n'est d'ailleurs pas de recenser de manière exhaustive l'ensemble des cas possibles mais de les réduire à un nombre limité de formes canoniques. Une tâche importante est la reconnaissance des entités nommées (en anglais NER pour *Named-Entity Recognition*). Le but est de détecter des termes appartenant à des catégories prédéfinies. Les classes d'entités sont dépendantes du domaine concerné et de la tâche.

Soient les énoncés :

- Paris : 2 cas de gale signalés ;
- VIIIe arrondissement : deux nouveaux cas de gale ;
- Un couple de personnes contaminé à Saint-Lazare ;
- Une famille espagnole a contracté la scabiose l'année dernière.

Il est tout d'abord nécessaire de reconnaître les entités nommées concernées par notre recherche, ici ce sont des maladies et des lieux, et les stocker dans une ressource lexicale. Cette ressource peut être constituée manuellement ou automatiquement. Elle n'est pas nécessairement gravée dans le marbre, elle peut être complétée au fur et à mesure du temps. La détection des entités nommées de type maladie permet de se fonder sur des patrons d'extractions indépendants de la maladie concernée. Les énoncés précédents permettent d'extraire au moins deux patrons[17] :

Adjectif numéral + « CAS DE » + *Nom_Maladie*

Animé + *Contaminé ou Contracté* + *Nom_Maladie*

Opérer des regroupements de termes permet de factoriser la description des phénomènes linguistiques. En créant une classe de verbes du domaine *Verbes_Transmission_Maladie*,

17. Il ne s'agit ici que d'une illustration.

on obtient un second patron avec une forme plus généralisée :

Animé + Verbes_Transmission_Maladie + Nom_Maladie.

Le patron présenté ici est assez fréquent mais il n'offre sans doute pas une grande couverture. Pour élargir la couverture, des motifs plus complexes sont alors nécessaires. Ces motifs imposent un certain nombre de pré-traitements permettant de repérer la forme canonique des mots, leur fonction dans la phrase.

On retrouve ces pré-traitements dans le schéma décrit par Hobbs dans le cadre des conférences MUC ([Hobbs-1993]), ils mettent en jeu des analyseurs spécialisés dans différentes « couches » linguistiques :

- lemmatisation ;
- analyse morphologique, impliquant un lexique ;
- reconnaissance des entités nommées ;
- analyse syntaxique, impliquant une base de motifs ;
- analyse sémantique, impliquant des ontologies ;
- analyse discursive, impliquant des règles d'inférence.

Dans cette approche, la phase d'extraction d'information sur une langue nécessite donc de disposer d'un analyseur pour tout ou partie de ces phases [18]. Un certain nombre de ressources externes sont également impliquées. L'EI comporte donc une série de modules d'analyse, que l'on appelle chaîne de traitement (en anglais *pipeline*). Ceci constitue le noyau dur de l'analyse (*core analysis*). Ce *pipeline* est couplé avec des connaissances externes ou *knowledge bases*. Il est fortement dépendant de la langue traitée, tandis que les connaissances externes sont elles dépendantes du domaine étudié.

2.2 Vers une veille massivement multilingue ?

Cette section présente les caractéristiques de chaque degré d'automatisation pour évaluer les différentes approches utilisées pour la conception de systèmes de veille multilingue. La démarche considérée comme la plus naturelle consiste à dupliquer un système de veille prévu pour une $langue_1$ pour l'adapter à une $langue_2$. Il s'agit alors « simplement » de remplacer tout ce qui dans le processus est dépendant de la langue traitée. Nous examinons dans quelle mesure cette approche est adaptée à la dimension multilingue.

Cette section ne vise pas l'exhaustivité, un très grand nombre de projets ayant été menés dans le domaine de la veille épidémiologique comme dans tant d'autres. Les systèmes décrits ici sont des systèmes marquants, soit par leur importance actuelle, soit par l'originalité de l'approche utilisée. Les caractéristiques principales de chaque système sont décrites, les systèmes étudiés sont regroupés selon le niveau d'intervention humaine.

18. L'analyse discursive n'est par exemple pas systématiquement pratiquée.

En premier lieu, le cas où l'intervention humaine est centrale est examiné (ProMED : Section 2.2.1). Puis, nous présentons des systèmes où l'humain opère une validation *a posteriori* (*GPHIN*, *Argus* et *HealthMAP* : Section 2.2.2). Enfin, dans la Section 2.2.3 sont exposés les systèmes dont la motivation de totale automaticité se rapproche le plus des objectifs de cette thèse (*EpiSPIDER*, *PULS* et *BioCaster*).

2.2.1 Veille manuelle de ProMED : l'humain comme émetteur et récepteur du signalement

Dans le cadre d'une veille manuelle, l'extension du système nécessite *a minima* le « recrutement » d'un locuteur de la langue concernée. L'extension de la couverture de ce type de système de veille épidémiologique ne se fait donc pas à moyens constants.

L'humain remplit ici le rôle de décideur de la validité du signalement effectué par un système automatique. Bien souvent, il a aussi la charge du typage du signalement (paire maladie-lieu par exemple).

L'évolution du système d'alerte « *ProMED-mail* » l'a amené vers une organisation où les sources textuelles sont analysées par des experts humains. Plusieurs rapports quotidiens sont rédigés à partir de sources pré-sélectionnées, dont la liste est mise à jour régulièrement. En plus des sources journalistiques, les analystes de *ProMED-mail* utilisent des rapports émis par les autorités sanitaires nationales, ainsi que par d'autres agrégateurs de contenus. *ProMED-mail* diffuse actuellement des rapports dans cinq langues : anglais, français, espagnol, portugais et russe. Le site de *ProMED-mail* [19] annonce également la diffusion de rapports en thaï, en vietnamien et en chinois, mais aucun rapport n'est actuellement diffusé pour ces langues [20]. En règle générale, nous y reviendrons dans le chapitre 5, quand la source est une source ouverte la langue d'émission du rapport est la langue de la source utilisée. Au vu de ce qui est diffusé par *ProMED-mail*, l'utilisation de données en anglais, français, espagnol, portugais ou russe est la règle ; le traitement d'autres langues est l'exception.

2.2.2 Veille semi-automatisée : vers une collaboration efficace de l'humain et de la machine

Dans cette section sont exposées les caractéristiques de trois systèmes : *GPHIN*, *Argus* et *HealthMap*. Deux limites importantes de ces systèmes semi-automatiques dans une perspective multilingue sont mises en lumière : le pré-filtrage des sources et la validation des informations par l'humain.

19. `http://www.promedmail.org/`
20. Dernière consultation le 12 octobre 2013.

GPHIN

Le projet canadien de Réseau Global de Renseignement sur la Santé publique (*Global Public Health Intelligence Network*[21] est le seul système à accès réservé présenté dans cette section. Ce système est fondé sur la validation *a posteriori* par des analystes humains de prototypes d'évènements détectés par un système commercial de fouille de textes (*Text Mining*) nommé *TME* (*Text Mining Engine*)[22]. Six langues (anglais, arabe, chinois, espagnol, français et russe) sont actuellement traitées mais aucune information n'est donnée sur la cause de cette limitation du nombre de langues. Il est impossible de savoir s'il s'agit d'un choix assumé, d'une incapacité de *TME* à analyser d'autres langues ou encore d'un manque d'analystes pour la validation *a posteriori*.

Argus

Le projet *Argus* ([Wilson-2007]) avait pour ambition de contourner le problème du nombre d'analystes en cherchant à améliorer spécifiquement le pré-filtrage des documents au profit d'une équipe d'« analystes multilingues » agissant dans une approche proche de *ProMED-mail* et de *GPHIN*. L'équipe d'*Argus* était composée de 36 analystes chargés de couvrir 34 langues ([Wilson-2008]). Le pré-filtrage était effectué à partir de modèles probabilistes pour 11 langues. Pour 8 autres langues, les textes étaient traduits automatiquement avant de passer l'étape de pré-filtrage. Aucune indication n'est disponible sur les pré-traitements éventuellement appliqués pour les 17 autres langues. Le projet n'apparaît plus en ligne et n'a fait l'objet d'aucune publication depuis 2010 ([Chen-2010]).

Les limites du pré-filtrage des sources

GPHIN et *Argus* misent beaucoup sur le pré-filtrage pour garantir la fiabilité de leur analyse à un coût raisonnable. Or, il n'est pas évident que l'on puisse disposer d'un pré-filtrage compétitif dans toutes les langues. Nous avons donc, dans le cadre de campagnes d'annotation, mené une évaluation sur quelques langues pour évaluer la qualité globale du pré-filtrage « santé » proposé sur deux agrégateurs. *Google News* propose par exemple une classification d'articles de presse de diverses sources dans différentes langues. Parmi les catégories disponibles figure la catégorie « santé » dans laquelle environ 7% des documents s'avèrent pertinents pour la veille épidémiologique. Toutefois, seules 12 langues sur les 18 disponibles sur *Google News* bénéficient de cette rubrique « santé ».

Le pré-filtrage proposé par l'agrégateur *MediSys*[23] couvre 27 langues mais la qualité offerte est très variable selon les langues. La qualité du filtarge de *MediSys* se rapproche pour le français et l'anglais de ce que propose *Google News* puisque respectivement 8 et 9% des documents de *MediSys* concernent spécifiquement la veille épidémiologique. En

21. `http://www.phac-aspc.gc.ca/gphin/`
22. Nstein's Text Mining Engine : `http://www.nstein.com`
23. `medusa.jrc.it`

russe, au contraire, le pourcentage de documents pertinents pour la veille épidémiologique était inférieur à 1%. L'objet de la catégorie « santé » n'est certes pas de référencer des documents relatifs à la veille épidémiologique. Toutefois, la nette différence de résultat amène des questions sur la confiance que l'on peut leur accorder à ces filtres. Ce point est développé dans le chapitre 4 qui traite des spécificités des corpus.

Une autre motivation du pré-filtrage est plus technique : la relative lenteur des processus de pré-traitement impliqués. La vitesse de l'étiquetage en parties du discours (*POS tagging*) est d'ailleurs un aspect généralement ignoré dans l'évaluation des systèmes automatiques. Pour donner un ordre de grandeur, la vitesse d'étiquetage est au mieux égale à 10^5 mots par seconde ([Brants-1998, Poudat-2004, Wilkens-2008, Perkins-2010]). En prenant les statistiques de l'*EMM* ($1,5*10^5$ articles par jour) et en partant sur une hypothèse de base de 10^3 mots en moyenne par document, il y aurait donc $1,5 * 10^8$ mots à étiqueter. Une seule des phases de la chaîne de traitement prendrait donc environ 30 minutes. Cette contrainte a été formulée par Steinberger ([Steinberger-2008a]) qui considère qu'il n'est pas raisonnable d'appliquer le processus d'extraction d'information à l'intégralité des documents disponibles.

Le pré-filtrage est fortement dépendant des langues. Les techniques classiques de RI, utilisées par *MediSys* par exemple, n'offrent pas la même fiabilité dans toutes les langues. De plus, le ralentissement occasionné par le pré-filtrage peut être très important selon les techniques de pré-traitement impliquées. Le pré-filtrage ne permet pas aux systèmes semi-automatisés d'assurer une couverture multilingue très large.

HealthMap

Concernant les systèmes semi-automatiques, le dernier système décrit ici est *HealthMap* ([Freifeld-2008]). C'est un système massivement automatisé puisqu'il propose notamment une vision condensée de différentes sources disponibles. Certaines sources, *ProMED-mail* ou les bulletins de l'OMS par exemple, sont des sorties structurées produites manuellement.

Une modération humaine est pratiquée par les analystes de *HealthMap* sur les évènements ainsi récupérés. Son principal objet est de mesurer la pertinence des signalements effectués par le système. Sept langues sont traitées par *HealthMap* (anglais, arabe, chinois, espagnol, français, portugais et russe). Ceci correspond à l'ensemble des langues traitées en amont par ProMED et *GPHIN*.

Le coût de la validation par l'humain

Pour les approches où l'humain est décisif dans l'analyse, l'extension de la couverture est coûteuse. Il faut en effet pouvoir disposer au moins d'un locuteur de chaque langue à traiter. Si un même analyste peut être en mesure de traiter plusieurs langues, une double contrainte existe toutefois : celle de la charge de travail d'une part et celle de la vitesse de traitement d'autre part ([Linge-2009]). La question de la vitesse est importante car il faut

pouvoir traiter des documents qui n'arrivent pas en flux tendu mais par paquets (quand un retard est pris par exemple).

Le pré-filtrage offre un gain de productivité intéressant mais, comme nous l'avons montré plus haut, la qualité du filtre n'est pas garantie dans toutes les langues. Un filtre trop rigide pourrait ignorer des documents intéressants. Mais s'il est trop lâche, le filtrage perd de son intérêt. Nous pensons donc que la veille semi-automatisée est sévèrement limitée dans ses possibilités de couverture multilingue. Il est donc intéressant de se tourner vers des systèmes plus automatisés encore.

2.2.3 Veille automatique multilingue

Le coût impliqué par la veille humaine invite à se tourner vers les méthodes entièrement automatiques pour augmenter la couverture des systèmes de veille à un coût raisonnable. Nous présentons ici trois systèmes représentant trois approches de la veille épidémiologique multilingue : agrégation de contenus (*EpiSPIDER*), combinaison RI/EI (*MediSys-PULS*) et traduction automatique (*BioCaster*).

EpiSPIDER

L'approche d'agrégation de contenus d'autres systèmes, utilisée par HealthMAP, a également été utilisée dans le défunt projet EpiSPIDER ([Tolentino-2007]). La validation *a posteriori* n'a par contre pas été retenue dans ce projet. La principale innovation d'*EpiSPIDER* se situait au niveau de la valorisation et de la mise en perspectives de connaissances extraites par d'autres systèmes automatiques. L'objectif de ce système n'était donc pas l'extraction de nouveaux contenus ou de nouveaux évènements mais l'exploitation de contenus existants. Cela se traduisait notamment par l'usage accru de frises chronologiques (*timeline*) et par la mise en ligne d'alertes par le biais d'un compte *Twitter* associé. Ce projet n'est toutefois plus maintenu depuis décembre 2011 [24].

In fine, c'est donc une parallélisation de systèmes monolingues (certes hétérogènes) qui a été utilisée. La couverture en langues est donc entièrement dépendante de traitements dépendants de la langue effectués en amont.

MediSYS-*PULS*

PULS (Pattern-Based Understanding and Learning System) se base sur un filtrage opéré par le système européen d'information MediSys. MediSys est lui même une branche de l'*EMM*, il rassemble des articles relevant du domaine de la santé dans 45 langues différentes. Il propose une classification en sous-catégories pour 25 de ces langues. Ces catégories sont formées à l'aide d'un modèle vectoriel et de la similarité cosinus ([Steinberger-2008a]). La pertinence de ce filtrage pour le domaine épidémiologique a été brièvement discutée dans la Section 2.2.2.

24. `http://www.epispider.org/` et `https://twitter.com/epispider`, consultés le 12 octobre 2013.

Le système *PULS* est issu d'un système initialement destiné à traiter automatiquement les alertes provenant de ProMED ([Grishman-2002]). *PULS* propose une plus-value à la classification issue de MediSys sous la forme d'extraction de quadruplets maladie-lieu-cas-date. La chaîne de traitement utilisée est basée sur des techniques classiques d"EI avec une partie d'apprentissage. Les bons résultats du système sur l'anglais ont incité les promoteurs du projet à étendre le système à d'autres langues.

Une première tentative d'extension de la couverture a consisté à créer un système simple permettant de traiter le français à faible coût ([Lejeune-2009a, Lejeune-2009b]). Le traitement du français a été réalisé avec une version altérée de ce système jusqu'en avril 2010.

La seconde tentative d'extension a donné lieu à un choix exactement opposé. L'approche a consisté cette fois à développer pas à pas les analyseurs nécessaires au processus d'EI de manière à dupliquer le *PULS* anglais pour le russe ([Du-2011]).

Nous sommes donc ici dans le cas d'école où un grand nombre de composants doivent être recréés ou réutilisés pour permettre d'étendre la couverture du système. Il s'agit dès lors d'une duplication du système d'une $langue_1$ pour s'adapter à une $langue_2$ comme dans certains projets d'EI plus généralistes ([Efimenko-2004, Etzioni-2011]).

BioCaster

À l'opposé des approches précédentes, le projet BioCaster ([Doan-2008]) utilise son système d'EI pour l'anglais comme analyseur final. Chaque texte dans une autre langue est donc traduit automatiquement en anglais avant d'être analysé. La traduction est donc un pré-traitement nécessaire au système au même titre que peuvent l'être la lemmatisation ou l'étiquetage morpho-syntaxique pour d'autres systèmes. En plus de cette phase de traduction, ce sont les éléments de l'ontologie du domaine utilisée qui sont ici spécifiques à la langue traitée. Cette ontologie, qui couvre 10 langues (anglais, arabe, chinois, coréen, espagnol, français, portugais, russe, thaï et vietnamien) , est librement mise à disposition par les promoteurs du projet [25]. Le système BioCaster a bénéficié d'un très grand nombre d'évaluations sur différents aspects ([Collier-2011]). Parmi les systèmes existants, il est celui qui offre la meilleure couverture en nombre de langues.

L'utilisation de la traduction automatique est une alternative originale pour augmenter la couverture multilingue. Il reste à savoir si cela revient à déplacer le problème plus qu'à le régler. Il convient de disposer d'un système de traduction automatique fiable. En effet, un système d'EI est naturellement basé sur des propriétés phrastiques précises. Le texte traduit doit être conforme aux règles morphologiques et syntaxiques de la langue pour que le module d'EI puisse reconnaître les patrons. Or, les systèmes de traduction automatique ne sont pas réputés pour obtenir pour tous les couples de langues des phrases syntaxiquement correctes. En réduisant le problème à la traduction vers l'anglais, cette

25. `http://biocaster.nii.ac.jp/_dev/static/ontology`

approche n'offrira toujours pas une garantie de qualité dans toutes les langues. Par ailleurs, construire l'ontologie nécessaire pour chaque nouvelle langue à traiter reste coûteux.

Les systèmes entièrement automatiques cherchent à réduire le coût marginal de traitement d'une nouvelle langue. L'outillage nécessaire reste toutefois très important et la factorisation des procédures est limitée. Par ailleurs, plusieurs études ont montré que le taux de recouvrement de différents systèmes était très faible ([Lyon-2011, Lejeune-2013a]), illustrant la difficulté de garantir un rappel élevé.

2.2.4 Perspectives pour l'augmentation de la couverture

Dix langues différentes sont actuellement prises en compte par les systèmes de veille épidémiologique décrits précédemment. Nous proposons ici une analyse de la couverture ainsi offerte et des raisons qui freinent son extension.

La couverture actuelle

Le tableau 2.1 présente les langues traitées[26] par les systèmes de veille épidémiologique présentés dans la section précédente et encore fonctionnels à ce jour. Nous pouvons voir que, conformément à l'hypothèse formulée dans la section 2.1, l'anglais occupe une place centrale : il est traité par chacun des systèmes. La couverture en nombre de locuteurs permise par l'anglais et le nombre élevé de sources disponibles motivent cette place particulière. La prééminence de l'anglais dans les travaux de TAL et de fouille de textes est certainement aussi un facteur important de cette prééminence de l'anglais.

Langues (Code)	#Locuteurs (10^6)	Systèmes				
		ProMED	*GPHIN*	*HealthMap*	*PULS*	*BioCaster*
anglais (en)	1000	✓	✓	✓	✓	✓
arabe (ar)	255		✓	✓		✓
chinois (zh)	1151		✓	✓		✓
coréen (ko)	78					✓
espagnol (es)	500	✓	✓	✓		✓
français (fr)	200	✓	✓	✓		✓
portugais (pt)	240	✓		✓		✓
russe (ru)	277	✓	✓	✓	✓	✓
thaï (th)	60					✓
vietnamien (vi)	86					✓

Tableau 2.1 – Couverture des principaux systèmes existants et nombre de locuteurs pour chaque langue

26. Pour les abréviations nous utilisons la norme ISO 639-1, plus connue et plus synthétique que la norme 693-3.

De façon assez étonnante, le russe est également traité par tous les systèmes présentés ici. Il faut probablement y voir une volonté de couverture territorial plus qu'en terme de nombre de locuteurs. Le nombre de locuteurs du russe n'est pas négligeable, loin s'en faut, avec 277 millions. Toutefois, le russe compte nettement moins de locuteurs que le chinois-mandarin (1151 millions), l'espagnol (500 millions) ou l'hindi (490 millions). Ce chiffre reste dans le même ordre de grandeur que l'arabe ou le portugais (respectivement 255 et 240 millions). La position préférentielle du russe n'est donc pas justifiée par un critère purement démographique. Le russe offre par contre une couverture géographique très importante : les pays où c'est la langue officielle (Russie, Kazakhstan, Kirghizistan, Biélorussie) représentent une superficie de plus de 20 millions de kilomètres carrés et partagent une frontière avec 14 pays différents.

Cinq autres langues de très grande diffusion (200 millions de locuteurs et plus) sont également bien représentées dans le tableau 2.1 :

- l'espagnol (4 systèmes) ;
- le français (4 systèmes) ;
- l'arabe (3 systèmes) ;
- le chinois-mandarin (3 systèmes) ;
- le portugais (3 systèmes).

Le fait que le français soit mieux représenté que des langues de plus grande diffusion (arabe, chinois et portugais notamment) peut s'expliquer par deux phénomènes. Le premier est la recherche de la couverture des pays africains. Le français a un statut de langue officielle ([OIF-2007]) dans 21 états africains [27]. Il est par ailleurs une langue de première importance dans cinq autres états africains [28]. Le second phénomène est la grande disponibilité de ressources d'analyses. Derrière l'anglais, le français fait partie des langues qui bénéficient de nombreux modules de traitement de la langue fiables ainsi que de ressources lexicales. Par exemple, le MeSH [29] (*Medical Subject Headings*), thésaurus de référence dans le domaine médical n'existe qu'en français et en anglais. Le français est par ailleurs la troisième langue, après l'anglais et l'allemand, à avoir franchi la barre du million d'articles dans l'encyclopédie en ligne Wikipédia.

Nous avons enfin dans ce tableau trois langues d'Asie du sud-est (coréen, thaï et vietnamien) qui sont traitées uniquement par BioCaster. Ceci permet à ce système d'afficher à la fois la plus large couverture en nombre de langues et une couverture géographique consistante en Asie.

Nous donnons également à titre indicatif dans le tableau 2.1 le nombre de locuteurs estimés de chaque langue ([Katsiavriades-2007]). Ces chiffres sont à manier avec prudence

27. Bénin, Burkina Faso, Burundi, Cameroun, Comores, Côte d'Ivoire, Djibouti, Gabon, Guinée, Guinée équatoriale, Madagascar, Mali, Niger, République centrafricaine, République démocratique du Congo, République du Congo, Rwanda, Sénégal, Seychelles, Tchad, Togo.

28. Algérie, Maroc, Île Maurice, Mauritanie, Tunisie.

29. `http://www.ncbi.nlm.nih.gov/mesh`

dans la mesure où les langues secondes sont également comptabilisées [30]. Ainsi, un canadien francophone ayant l'anglais comme langue seconde apparaîtrait deux fois dans ces statistiques. Les chiffres ne sont donc pas suffisamment précis pour pouvoir évaluer exactement la population potentiellement couverte : un certain nombre de doublons seraient pris en compte. Nous pouvons par contre remarquer que des langues de diffusion très importante ne sont pas traitées par les systèmes de l'état de l'art : l'hindi (490 millions) le bengali (215 millions) ou encore l'indonésien (175 millions).

Le coût d'extension des systèmes

Le tableau 2.2 présente les pré-requis de chaque système pour traiter une nouvelle langue. Nous avons exclu *HealthMap* de ce tableau puisqu'il est basé sur l'agrégation de résultats provenant d'autres systèmes. L'analyse approfondie des textes, cœur de la veille, est traitée de façon différente par les quatre systèmes présentés dans ce tableau.

Moyens / Systèmes	Locuteurs	Modules d'analyse	Ressources lexicales
ProMED-mail	veilleur		
GPHIN	validateur	*Text Mining*	inconnues
PULS		Extraction d'Information	ontologie
BioCaster		Traduction Automatique	ontologie

Tableau 2.2 – Moyens nécessaires à l'extension vers une nouvelle langue pour chacun des systèmes décrits

L'approche manuelle de *ProMED-mail* impose de disposer de nouveaux locuteurs pour chaque nouvelle langue à traiter. Cette démarche ne semble pas compatible avec tous les domaines de la veille car elle est très coûteuse. La veille épidémiologique est un domaine stratégique ([CDC-2004, ECDC-2006]) qui bénéficie de nombreuses sources de financement. Ceci n'a pas permis aux approches manuelles de dépasser une couverture de huit ou dix langues. Par exemple, le système *ProMED-mail* n'affiche qu'une couverture limitée en dépit de sa relative ancienneté et du budget important de l'OMS.

Le fonctionnement du système commercial *Nstein's Text Mining Engine* utilisé par *GPHIN* pose question. Il est décrit sur le site Web de la compagnie éditrice comme étant multilingue sans autre précision. Nous supposons que ce logiciel met en œuvre des techniques de pré-traitement usuelles dans le domaine de la fouille de textes. *A minima*, il doit nécessiter un lemmatiseur et un système de reconnaissance des entités nommées. Selon toute vraisemblance, le multilinguisme affiché par ce module est restreint. Là encore, le fait que le système se limite à six langues est un indice de la difficulté que le traitement de nouvelles langues doit représenter pour ce type d'approche.

30. `http://www.krysstal.com/spoken.html`, consulté le 12 octobre 2013.

Le système *PULS* est plus transparent sur son fonctionnement et sur les ressources que nécessite le traitement d'une nouvelle langue ([Du-2011]). Le traitement du russe a pu s'appuyer sur l'architecture existant pour l'anglais mais avec un certain nombre d'aménagements. Il a en effet fallu développer ou intégrer les modules suivants : lemmatiseur, étiqueteur, analyseur syntaxique. Par ailleurs, l'ontologie spécialisée en russe a dû être créée pour l'occasion. La taille des ressources requises par l'architecture prônée par les auteurs freine donc considérablement l'extension de la couverture de ce système.

Ces trois premiers exemples illustrent le fait que l'augmentation de la couverture au-delà de six ou sept langues constitue une difficulté importante. Un grand nombre d'étapes de traitement sont dépendantes de la langue. La plupart des acteurs (modules de traitement ou humains) impliqués dans chaque système doivent être secondés ou remplacés pour permettre le traitement d'une nouvelle langue. Cela rejoint l'étude de Kabadjov *et al.* ([Kabadjov-2013]). Pour ses auteurs, la couverture multilingue est difficilement atteignable par simple parallélisation de systèmes monolingues. Ce peut être une explication à la limitation de la couverture des trois systèmes précités (*ProMED-mail*, *GPHIN* et *PULS*) : la **factorisation des procédures est limitée**.

Concernant *BioCaster*, cette factorisation est recherchée par un recours à la traduction automatique. Plutôt que d'analyser le texte dans sa langue d'origine, il est traduit automatiquement en anglais puis soumis au système d'analyse spécialisé pour cette langue. Cela revient à opérer un pré-traitement du texte afin de le rendre analysable par un module central unique. Mais extraire une information fiable impose que la traduction automatique n'altère pas ou peu le contenu du texte. Aucune ambigüité ne doit en effet être ajoutée pendant cette phase.

Or, le problème de la qualité de la traduction automatique est clairement posé. La traduction automatique a pourtant pu apparaître comme une solution efficace à moindre coût ([Kanayama-2004]). Selon Linge ([Linge-2009]), la fouille de textes doit pouvoir bénéficier des avancées de la traduction automatique pour traiter de nouvelles sources. Au contraire, une étude de Piskorski ([Piskorski-2011]) a montré que cette technologie n'apportait pas de plus-value vis-à-vis d'une addition de systèmes monolingues de qualité hétérogène. La traduction automatique est par ailleurs mal adaptée aux textes et aux vocabulaires spécialisés. Cette contrainte est contournée dans *BioCaster* par l'utilisation d'une ontologie du domaine, mais cette ontologie est coûteuse à construire et pose à nouveau le problème de la duplication des ressources ([Steinberger-2008b]).

En résumé, les approches actuellement utilisées sont trop dépendantes de ressources propres à chaque langue pour permettre une couverture réellement multilingue. Ces ressources peuvent êtres :

- humaines : des veilleurs ou des évaluateurs ;
- modulaires : par exemple des analyseurs grammaticaux, des traducteurs automatiques ou des outils de fouille de texte ;
- lexicales : par exemple des dictionnaires ou des ontologies.

Ces ressources induisent des coûts importants qui se ressentent à différents niveaux. Le premier niveau est le **coût d'utilisation**, d'acquisition ou de création de chaque ressource impliquée. Ce peut être un frein important à la satisfaction du besoin d'information. Il faut également faire face à un coût de maintenance des systèmes qui est d'autant plus grand que ceux-ci sont complexes. En second lieu, il faut faire face à un **coût en temps** dans la mesure où l'analyse par l'humain est *a priori* plus lente que l'analyse automatique. De son côté, l'approche automatique impose principalement un temps de développement. Enfin, le plus important est sans doute le **coût en efficacité** lorsqu'il n'est pas possible de garantir une qualité de service comparable entre les langues.

Ces différentes modalités ne sont pas indépendantes, il est possible de compenser la lenteur de l'analyse humaine, au prix d'un coût financier important, en recrutant une armée de veilleurs. On peut sacrifier l'efficacité en se contentant de traiter un moins grand nombre de langues pour diminuer les coûts. Il nous semble que ceci traduit la limite des approches de l'état de l'art de la veille épidémiologique.

Synthèse

Les systèmes de veille épidémiologique existant ne sont pas multilingues mais multi-monolingues ([Steinberger-2008b]). C'est à dire qu'il ne s'agit pas de systèmes multilingues par essence ([Gey-2009]), mais de la mise en parallèle de systèmes monolingues. La couverture ne dépasse pas les dix langues et l'extension se fait véritablement pas à pas, langue après langue. La factorisation est limitée du fait de la dépendance à la langue de la majorité des étapes de traitement et du coût des ressources utilisées.

Le gain marginal est décroissant au fur et à mesure que l'on traite des langues plus rares : il est de plus en plus difficile de s'appuyer sur l'existant. Dans le même temps, le coût marginal grandit puisque les ressources spécifiques requises seront de moins en moins disponibles.

Si les bénéfices de l'automatisation du processus de veille au niveau monolingue sont avérés, il n'en est pas de même dans une perspective plus large. En effet, les difficultés rencontrées par les méthodes actuelles semblent limiter le cadre multilingue à une poignée de langues pour lesquelles des modules d'analyse locale (lemmatiseurs, étiqueteurs...) sont disponibles. Une approche véritablement efficace et à large couverture devrait être multilingue par essence et non pas multilingue par parallélisation de systèmes monolingues.

Conclusion de la première partie

Dans cette partie, nous avons présenté une évaluation des besoins d'information des autorités sanitaires vis-à-vis des articles diffusés dans la presse Internet. La plus-value que ces organismes attendent d'une veille automatique est liée à la problématique de la couverture : il faut pouvoir traiter autant que possible toutes les sources disponibles.

Cette couverture est principalement liée au nombre de langues que le système peut traiter. Nous avons examiné comment les approches actuelles cherchaient à traiter la dimension multilingue. Il s'avère que le coût marginal d'extension de ces systèmes est très important. De plus, il existe une forte disparité entre les langues. Certaines langues seraient très coûteuses, sinon impossibles à traiter avec les techniques actuelles.

Ceci ne concerne pas que des langues rares ou d'extension géographique limitée. La limitation de la couverture des systèmes actuels ne résulte d'ailleurs pas tant de choix de traiter telle langue plutôt qu'une autre mais véritablement de contraintes résultant d'approches trop peu génériques. De ce fait, la problématique de la couverture ne peut être résolue par des approches ne permettant que le traitement des 5 ou 10 langues les plus répandues.

Cette couverture limitée est d'autant plus problématique que les épidémies ne surgissent pas systématiquement dans les zones d'influence des langues de grande diffusion. La problématique du multilinguisme dans la veille épidémiologique mérite dès lors un traitement différent et une approche adaptée.

Deuxième partie

Vers un traitement de la langue adapté à la dimension multilingue

Introduction

Nous exposons dans cette partie des fondements scientifiques adaptés à un objectif de traitement des textes en langue naturelle avec une dimension multilingue. Nous présentons dans le chapitre 3 une approche différentielle, endogène et non-compositionnelle. Cette approche se fonde sur les relations existant entre les unités, elle vise à limiter les données de base nécessaires au traitement de ces relations. La méthode ne se base pas sur la recomposition d'unités atomiques telles que le mot graphique pour en déduire un sens. Au contraire, l'information recherchée est extraite à un grain donné sans composition à partir des grains plus petits.

Nous montrons dans le chapitre 4 comment exploiter cette approche. L'utilisation des propriétés du genre textuel offre une alternative crédible à l'analyse à fondement grammatical. Les propriétés qui forment la base de notre système sont choisies en fonction de leur aptitude à combiner efficacité, simplicité et généricité. Nous présentons des manifestations de ces propriétés ainsi que les opportunités qu'elles représentent pour la veille en général et pour la veille épidémiologique en particulier.

Chapitre 3

Fondements pour une veille multilingue parcimonieuse

Sommaire

Ce chapitre présente les fondements de notre approche ; ces fondements sont choisis pour leur conformité avec notre problématique multilingue. L'objectif est de promouvoir une **factorisation** massive des procédures mises en œuvre et d'utiliser des ressources externes avec une grande **parcimonie**. La factorisation implique que les procédures mises en œuvre soient strictement indépendantes de la langue traitée. Dans un objectif de parcimonie, les ressources utilisées doivent être aussi légères que possible, et constructibles, quelle que soit la langue à traiter.

L'approche présentée se fonde sur l'idée de parvenir à traiter « n » langues par des moyens raisonnés. En effet, il est difficile d'obtenir une grande couverture en développant un système pour la langue_1 puis en l'adaptant à la langue_2, à la langue_3... L'approche cumulative (une langue puis une autre) montre rapidement des limites en terme de couverture. Ces limites sont souvent imputées à la complexité de certaines langues ou à l'absence de ressources permettant de les traiter. Le développement de nouvelles ressources pour combler les manques est alors la seule solution envisagée. Cette solution est pourtant coûteuse, notre approche vise donc à offrir d'autres perspectives.

L'objectif de couverture multilingue peut et doit être poursuivi en cherchant d'emblée à bâtir un système aussi générique que possible. Cette généricité permet de factoriser au

maximum les traitements appliqués, indépendamment du nombre de langues à traiter. Ceci implique que la plus grande partie des opérations effectuées soit rigoureusement indépendante de la langue analysée. De cette façon, l'effort à fournir pour traiter une langue supplémentaire est réduit au minimum. L'objectif est d'atteindre un **coût marginal** minimal, idéalement ce coût serait nul.

3.1 Principes généraux de notre approche

De notre point de vue, le traitement multilingue ne revient pas à chercher absolument dans certaines langues des propriétés qu'elles ne connaissent pas. Un exemple de cette pratique est de chercher des mots graphiques dans des langues qui ne les connaissent pas comme le chinois. Les applications réellement multilingues ne peuvent pas se baser sur des extensions pas à pas de principes valables au niveau bilingue ou trilingue. Au contraire, de telles applications doivent se baser sur des concepts, des propriétés alingues ([Vergne-2004a, Roy-2007, Lardilleux-2010]). Ce constat est issu d'une longue tradition caennaise [31] du traitement des langues dans laquelle l'aspect multilingue est central. Pour Jacques Vergne ([Vergne-2004b]) les propriétés alingues sont des « propriétés très générales des langues » ou, *a minima*, communes à un groupe de langues.

Ayant fait le constat que la description des phénomènes linguistiques était très coûteuse, puisque fortement dépendante de chaque langue, nous définissons ici une approche :

différentielle : les relations entre les unités d'analyse importent plus que les unités elles-mêmes ;

non-compositionnelle : le sens d'un grain n'est pas reconstruit par la sommation du sens des grains plus petits ;

endogène : les documents sont à la fois objets d'analyse et ressources permettant d'améliorer cette analyse, le recours aux ressources externes est limité.

Ces trois aspects se recoupent fortement. Exploiter les différences et les relations entre les unités permet de limiter la description des phénomènes linguistiques étudiés. L'approche différentielle « pure » impliquerait sans doute l'absence totale de ressources externes. Nous n'en faisons néanmoins pas un dogme, c'est le meilleur rapport coût–efficacité qui est recherché ici. Dès lors, des ressources légères sont envisageables si leur coût de construction est réaliste par rapport à la tâche.

Éviter les phases consécutives de décomposition–recomposition des unités permet de ne pas dépendre de modules d'analyse locale. L'aspect endogène est ainsi grandement favorisé, les pré-traitements sont limités. Ces principes sont présents dans certains travaux de RI qui placent l'indépendance vis-à-vis des langues traitées au cœur de leur

31. À travers des travaux menés au sein du laboratoire GREYC dans l'équipe ISLanD et perpétués dans les équipes DLU puis HULTECH.

approche ([Doucet-2010]). Dans notre approche, la transformation des unités disponibles en unités interprétables par l'humain n'est pas requise. L'intérêt d'une telle approche a déjà été démontré dans le domaine de la détection de plagiat dans du code source ([Brixtel-2010]).

La limitation des pré-traitements renforce l'indépendance par rapport à la langue (ou au langage de programmation).

D'autre part, tirer parti au maximum des propriétés internes du matériau traité permet une moindre dépendance aux ressources externes. Les ressources externes sont les ressources lexicales stockées en mémoire ainsi que les outils traditionnels de traitement des données langagières (lemmatisation, étiquetage...). Nous postulons que certains indices pertinents sont directement accessibles dans le texte sans pré-traitement « linguistique ». Parmi les informations qui figurent dans un document, toutes ne sont pas lexicales. La structure et l'organisation du texte laissent des traces, des indices qui peuvent prendre différentes formes. Ces indices, pas nécessairement explicites, n'en sont pas pour autant inexploitables. Ils permettent de notre point de vue une meilleure abstraction des méthodes d'analyse. Cette abstraction autorise une généralisation efficace des hypothèses.

Nous proposons donc une **analyse parcimonieuse** fondée sur une **factorisation** massive des procédures. La parcimonie est définie ici comme la recherche de l'économie dans la description des phénomènes et de la simplicité dans la mise en œuvre des méthodes. La factorisation est le corolaire de la parcimonie. Mettre en avant la factorisation et la parcimonie constitue une réponse adaptée aux problématiques de coût de traitement. Un nombre minimal de composants nécessaires et suffisants sont définis, par la suite leur réutilisation doit être maximisée.

La factorisation est *a priori* consubstantielle à toute réalisation d'un projet en informatique. Cette vérité est partielle dans le domaine du TAL. De nombreux outils sont certes conçus pour être indépendants du domaine ou du genre traité. Mais, la dépendance à la langue de ces outils est, par contre, la règle. Ceci provoque un coût de traitement incompatible avec une couverture multilingue (Section 3.2). Notre approche vise donc la factorisation des procédures (Section 3.3) et la parcimonie dans l'utilisation des ressources externes (Section 3.4).

3.2 Le coût de traitement et ses conséquences

Au fur et à mesure des années, la communauté du TAL est parvenue à s'accorder sur une architecture de traitement « idéale », qui met en jeu des modules d'analyse au sein d'une **chaîne de traitement**. Elle implique un coût de construction (fabriquer ou collecter les maillons de la chaîne) et un coût d'utilisation (maintenance de la chaîne et temps de traitement impliqué). Ce coût est un frein très important pour bon nombre de tâches.

Cette chaîne de traitement vise à traiter successivement différents niveaux de la langue.

Les principaux maillons qui constituent cette chaîne ont été exposés dans la section 2.1.3. Il ne s'agit pas de discuter de l'opportunité ou de la pertinence de tel ou tel de ces maillons. Ce sont plutôt les principes qui sous-tendent cette chaîne de traitement qui nous intéressent.

Lorsque cette chaîne classique de traitement est utilisée pour s'attaquer à une problématique multilingue, nous remarquons que peu de maillons sont finalement réutilisables pour d'autres langues. Peu d'outils génériques existent. Soit parce que le nombre de règles spécifiques à chaque langue est important ; soit parce que les outils basés sur l'apprentissage automatique se heurtent à la faible disponibilité de données d'entraînement. C'est donc une réponse très partielle au problème du multilinguisme. Par ailleurs, de plus en plus de travaux en TAL s'appuient sur des ressources lexicales importantes ou sur des ontologies. Ces ressources sont d'autant plus coûteuses que leur taille est grande et leur degré de description est fin. Elles sont très peu réutilisables ; couvrir une nouvelle langue impose quasiment de repartir de zéro [32].

Le coût des outils se manifeste principalement dans l'énergie, le temps consacré à leur « fabrication ». C'est aussi, et c'est particulièrement vrai pour les ontologies, un coût en terme de maintenance. Un autre coût important, et souvent ignoré, est le **coût opératoire**. Celui-ci comporte deux aspects, un aspect calculatoire d'une part, un aspect d'efficience d'autre part. L'enchaînement des étapes de traitement implique une complexité importante en terme de calcul. En prenant l'exemple de l'étiquetage morpho-syntaxique, nous pouvons remarquer que la question de la vitesse de traitement a quasiment disparu des préoccupations de la communauté du TAL. Ce problème a été évoqué dans la section 2.1.3 (page 39) où le temps d'étiquetage d'une journée d'articles repris sur le portail *EMM* a été évalué à 30 minutes. Bien que ce chiffre puisse sembler raisonnable [33], il amène quelques remarques :

- Il ne s'agit que d'une seule étape du traitement ;
- La vitesse est supposée indépendante de la langue ;
- L'analyse d'archives courant sur de longues durées est coûteuse [34].

Cette question du temps de calcul peut sembler relever plutôt du domaine de l'ingénierie. Au contraire, ce coût est révélateur d'une complexification non–justifiée du processus d'analyse, incompatible avec le passage à l'échelle sur de grands corpus.

Le processus d'analyse est difficile à évaluer puisqu'il implique le chaînage de plusieurs étapes dépendantes les unes des autres. L'accumulation d'erreurs au fur et à mesure du processus amène également des interrogations. Il faut pouvoir évaluer dans quelle mesure ces erreurs se multiplient ou se compensent. L'impact des erreurs d'un module

32. *Wordnet* existe ainsi pour différentes langues mais la qualité et la taille de ces différentes versions sont très disparates. Le coût du passage à d'autres langues n'est donc pas un problème anodin.

33. Après tout, il « suffirait » d'avoir plusieurs serveurs pour calculer plus rapidement puis de stocker les résultats de chaque étape de traitement pour une réutilisation ultérieure...

34. Ce qui est un problème réel que nous avons eu à traiter lors de notre travail au sein du projet PULS à Helsinki.

sur les modules d'analyse suivants a par exemple été étudié par différents chercheurs ([Debili-2006, McCallum-2006, Poulard-2011]). Les modules sont le plus souvent évalués indépendamment les uns des autres. Il est rarement tenu compte de la qualité de l'entrée fournie comme des besoins réels en sortie. Par exemple, les performances d'un étiqueteur morpho-syntaxique dépendent fortement de la qualité de la lemmatisation produite en amont de la chaîne.

Notre objectif n'est pas de condamner la chaîne de traitement classique en tant que telle. Nous voulons souligner qu'elle est inadaptée dans certains cas. La couverture multilingue est un de ces cas. Il n'est pas réaliste de créer une chaîne de traitement classique pour chaque langue, il faut donc s'adapter aux différentes langues en jeu. Il est possible que localement, sur des langues bien dotées en ressources, la chaîne de traitement classique puisse être pertinente. Mais elle n'est pas compatible avec une perspective multilingue. Nous rejoignons ici la vision proposée par Thibault Roy dans sa thèse de doctorat ([Roy-2007]) :

> « Une méthode sans ressource pourra être moins efficace, mais sera applicable sur un très grand nombre de langues, alors qu'une méthode avec ressources et règles pourra être plus efficace sur une langue donnée, mais ne sera utilisable que sur cette langue. »

C'est un constat qui a été également fait par l'équipe du Centre de Recherche de la communauté européenne (JRC pour *Joint Research Center*) : la couverture multilingue impose de penser différemment le traitement de la langue ([Steinberger-2008a]). Il convient de mettre beaucoup plus en avant la factorisation et la parcimonie.

3.3 Le choix d'un noyau de traitement adapté à la dimension multilingue

Un noyau de traitement adapté à la dimension multilingue ne peut être obtenu qu'à partir de la définition de propriétés communes à toutes les langues. C'est ce que nous appelons ici des **propriétés alingues**. Nous pourrions aussi parler d'**invariants** ou d'**universaux**. Ce dernier terme pourrait toutefois créer une confusion avec l'utilisation qui en est faite en grammaire générative.

La dépendance vis-à-vis de la chaîne de traitement classique, mais complexe, comprenant lemmatiseurs, étiqueteurs, analyseurs syntaxiques et lexiques ou ontologies est problématique pour analyser de nouvelles langues. Les techniques d'apprentissage pourraient constituer la solution idoine. Il suffirait en effet à partir de données d'entraînement de construire le meilleur jeu de règles pour traiter de nouvelles données. Les techniques d'apprentissage automatique sont supposées totalement indépendantes de la langue traitée. Mais leur utilisation suppose tout de même que l'on dispose de données d'entraînement avec une taille et une qualité qui soient suffisantes. De notre point de vue, ces données

constituent également des ressources externes, ce qui implique un coût. Par ailleurs, les données d'entraînement sont souvent elles-mêmes produites à partir de traitements dépendants de la langue. Par exemple, l'étiquetage morpho-syntaxique qui est utilisé dans les travaux de Charnois pour préparer l'extraction de patrons ([Charnois-2009]). Ces ressources en nécessitent d'autres qui ont elles même une dépendance forte à la langue. La qualité des données d'apprentissage est fortement dépendante de ces pré-traitements. Il semble dès lors difficile d'affirmer que les techniques d'apprentissage peuvent résoudre le problème de la disponibilité des ressources. Les techniques d'apprentissage ne sont donc pas à même de solutionner les problématiques multilingues.

La factorisation est ici envisagée de façon plus large. C'est l'opération de traitement dans son ensemble qui est reconsidérée. Nous entendons par là qu'aucun maillon de la chaîne n'est considéré comme indispensable. Aucun des modules d'analyse impliqué ni aucune des ressources mises en œuvre ne constituent à notre sens des pré-requis indispensable au traitement des textes. C'est l'ensemble du processus qui est revu de manière à obtenir une chaîne de traitement adaptée à la couverture multilingue. Nous ne cherchons pas à segmenter le texte en unités interprétables, au contraire c'est le document lui même qui est placé au centre du processus d'analyse. Plus précisément c'est l'appartenance du document à un genre textuel qui est exploitée. Ce point est développé dans la Section 4.1.1.

3.4 L'exploitation de ressources accessibles et de taille raisonnable

Usuellement, des tâches telles que la veille épidémiologique requièrent des ressources de taille importante. L'ontologie utilisée dans le système *BioCaster* représente ainsi plus de 10 Mo [35] au format RDF. Certains des termes décrits sont disponibles dans dix langues. Le coût de constitution de cette ontologie est difficile à mesurer faute d'indication des auteurs sur le sujet. La limitation de l'ontologie à dix langues, malgré une volonté affichée de couverture maximale, en donne une première indication. Les quatre années qui se sont écoulées entre le dévoilement de la version 2 de cette ontologie ([Collier-2006]) et celui de la version 3 ([Collier-2010]) en fournissent une seconde.

D'autres projets de ressources multilingues dans le domaine médical existent. L'*UMLS* [36] (*Unified Medical Language System*) est une initiative de constitution de ressource lexicale multilingue qui vise notamment à combler les lacunes de l'existant. Cette ressource a servi de base à un certain nombre de travaux pour améliorer des terminologies multilingues ([Hazem-2013, Mougin-2013]). Toutefois, les possibilités d'utilisation à grande échelle restent limités du fait du déséquilibre de la ressource. La couverture affichée est de 21 langues

35. Elle est téléchargeable à l'adresse suivante : `http://code.google.com/p/biocaster-ontology/downloads/list`

36. `http://www.nlm.nih.gov/research/umls/`

mais l'anglais et l'espagnol représentent à eux seuls 85% des termes recensés [37].

Nous en déduisons que le coût de constitution et de maintenance de ce genre de ressources reste très élevé et qu'offrir un statut égal à toutes les langues n'est pas envisageable.

Or, nous avons montré que la taille et la richesse des ressources engagées étaient un frein considérable à la couverture multilingue. En l'absence de ressources lexicales multilingues structurées et de grande taille, il est nécessaire d'envisager de solutionner les problèmes de couverture de façon parcimonieuse. L'extraction automatique de lexiques multilingues à partir de corpus comparables ([Gaussier-2004]) reste difficile du fait de la faible couverture en langues de ces corpus.

Le fait de mettre le document et le genre textuel auquel il appartient au centre de nos préoccupations permet de limiter la taille des ressources lexicales stockées en mémoire. La parcimonie dans les ressources est donc à la fois un objectif, lié a la volonté de large couverture en langues, et une conséquence de l'utilisation centrale de la notion de genre textuel. Au grain phrase, une grande quantité de vocabulaire est nécessaire pour analyser les énoncés. Au contraire, au grain texte il est beaucoup moins probable que le vocabulaire utilisé soit entièrement inconnu.

3.5 Synthèse

Nous avons abordé dans ce chapitre les problèmes posés par l'architecture classique de TAL lorsqu'il s'agit de traiter un grand nombre de langues. L'idée défendue est de **minimiser le coût marginal** de traitement de nouvelles langues afin de permettre une couverture large en nombre de langues à moindre coût. Nous avons montré que ceci n'était pas envisageable par le biais d'une architecture classique. Nous avons proposé de contourner ce problème en articulant notre approche autour d'un noyau d'analyse commun à toutes les langues, favorisant la factorisation des procédures. L'objectif est de limiter le nombre de ressources impliquées dans le traitement, puisque ce sont ces ressources qui sont dépendantes des langues. Elles constituent un frein à l'extension multilingue, nous cherchons donc la parcimonie dans l'utilisation de ressources.

37. `http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html`

Chapitre 4

Une approche textuelle fondée sur le grain caractère

Sommaire

Pour atteindre notre double objectif de factorisation et de parcimonie, nous nous fondons sur le genre textuel d'une part et sur une analyse au grain caractère d'autre part. Les **propriétés du genre journalistique**, les universaux de style, constituent la base de l'analyse. L'indépendance des règles du genre par rapport aux langues constitue un atout considérable pour la veille multilingue. Les propriétés du genre permettent d'envisager des traitements massivement factorisés. Le texte est exploité en tant qu'unité plutôt que comme la simple somme des phrases successives qui le composent. Ainsi, la variable n'est plus la langue mais le genre. L'analyse pratiquée est donc dépendante du genre

étudié plutôt que des langues analysées. Le **grain caractère** permet de s'affranchir d'un grand nombre de caractéristiques spécifiques à chaque langue, les mots par exemple. Les techniques d'algorithmique du texte et la standardisation de l'encodage des caractères fournissent à l'analyse au grain caractère un support solide. La description linguistique requise par le système est ainsi limitée.

La factorisation et la parcimonie sont nos principes méthodologiques, les propriétés du genre et l'analyse au grain caractère permettent de les mettre en application.

Nous devons donc modifier les termes de l'équation habituelle du traitement automatique de la langue. À l'utilisation de la place des mots dans la phrase, nous substituons l'exploitation de la position de chaînes de caractères dans les textes. La description, à vocation exhaustive, d'éléments du lexique est remplacée par la définition du vocabulaire nécessaire et suffisant pour une analyse basée sur le genre textuel.

C'est le grain texte et son ancrage dans un genre déterminé qui servent de pivot à notre étude (Section 4.1). Les propriétés du genre textuel étudié assurent la parcimonie dans l'utilisation de lexiques (Section 4.2). L'inadaptation de la notion de mot à la problématique multilingue amène à analyser les textes au grain caractère (Section 4.3).

4.1 Le document comme unité minimale d'interprétation

Le document est dans notre approche l'unité principale par ce qu'il possède des propriétés d'organisation qui sont indépendantes des langues. L'hypothèse est que les caractéristiques détectables au grain document offrent une grande robustesse à l'échelle multilingue.

Au contraire, la phrase est difficile à modéliser dans une perspective multilingue. La syntaxe est souvent difficilement comparable entre différentes langues. L'analyse syntaxique dépend également de pré–traitements, eux-mêmes dépendants de la langue. Le traitement correct des phrases d'une langue donnée nécessite donc un grand nombre de ressources, disponibles ou à construire, qui ne sont que partiellement réutilisables pour d'autres langues.

Par ailleurs, la phrase n'est pas un cadre d'interprétation totalement satisfaisant. Le contexte dans lequel elle apparaît est important. Analyser une phrase mot par mot puis un texte phrase par phrase revient finalement à déployer de nombreux efforts pour découper les textes en unités de plus en plus petites (phrases et mots) puis à s'efforcer de réagencer ces unités dans un second temps.

Nous proposons au contraire d'utiliser le texte comme unité d'analyse minimale au travers de sa relation avec le genre dont il est issu. L'article de presse fait ainsi partie d'un genre qui possède des règles précises. Ces règles, de plus haut niveau que les règles grammaticales, sont très proches dans des langues différentes. Elles régissent les relations entre l'auteur et son public, dans notre cas entre le journaliste et ses lecteurs. Ces règles sont

spécifiques au genre ; elles définissent la structure de l'article de presse et le vocabulaire qui y est utilisé (Section 4.1.1). Elles ont des visées communicationnelles et sont connues de la source comme de la cible des documents (Section 4.1.2). À partir de la connaissance de ces règles sont définies des positions remarquables qui sont indépendantes des langues (Section 4.1.3).

4.1.1 L'importance de la dimension textuelle

La dimension textuelle des données à traiter en TAL est souvent ignorée ou reléguée au rang de paramètre secondaire (voir par exemple [Yangarber-2011a] sur l'utilisation de la position des unités dans les textes). Pourtant, l'importance de la dimension textuelle est reconnue par les linguistes. D'après Rastier, le grain texte est fondamental dans l'étude de la langue : « le texte est l'unité d'analyse minimale » [38] ou en plus polémique ([Rastier-2002]) :

> « Le texte est, pour une linguistique évoluée, l'unité minimale »

Il nous semble important de ne plus considérer un document comme non–structuré simplement par ce qu'il n'est pas directement disponible dans un format XML canonique. La recherche absolue du « texte brut » délibérément dépossédé de nombreuses caractéristiques essentielles (mise en forme matérielle notamment) n'est donc pas une solution de notre point de vue. La topologie des énoncés, les indices positionnels ou plus généralement les contraintes d'un genre textuel peuvent en effet être extrêmement utiles dans des langues où les ressources sont parcellaires, voire inexistantes. Un aspect capital de ces indices est qu'ils sont inhérents au genre et peu coûteux à repérer. Les règles du genre laissent des marques fortes dans les textes, leur détection est donc peu coûteuse. De plus, elles sont robustes au changement de langue. Elles peuvent être articulées dans un jeu de règles d'analyse indépendant des langues.

Prenons par exemple la phrase suivante tirée d'un article du *Daily Mail* [39] :

> *« Gauguin took himself off to Tahiti where he entertained under-age mistresses, consumed vast quantities of absinthe and morphine and* ***died of syphilis*** *in 1903. »*

La séquence en gras est un déclencheur typique d'alerte dans un système d'Extraction d'Information. Cette séquence pourrait être extraite à partir du motif déclencheur décrit dans la Section 2.1.3 (page 37). Effectivement, l'article en question avait été sélectionné par le système *PULS*. Un *cluster* « Syphilis en France en mai 2009 » avait ainsi été constitué. Il s'agit pourtant, pour le lecteur humain, d'un **faux positif** évident. Il est clair en effet que cette information n'est pas susceptible d'intéresser la veille épidémiologique. C'est

38. Citation extraite de l'émission de France Culture « Tire ta Langue » du 18 février 2003.
39. `http://www.dailymail.co.uk/debate/article-1178163/`, consulté le 12 octobre 2013

TITRE : One was an arrogant bully. The other was a nervous wreck. So what is the truth about the war of Van Gogh's ear ?

Paragraphe 1/38 The iconic story about Dutch-born painter Vincent Van Gogh cutting off his own ear and presenting it as a gift to his favourite prostitute may not be true after all.
Paragraphe 2/38 Or so say some German art historians, who now claim the famous ear was cut off in a fight with rival artist Paul Gauguin.
...
Paragraphe 29/38 Gauguin took himself off to Tahiti where he entertained under-age mistresses, consumed vast quantities of absinthe and morphine and **died of syphilis** in 1903.
...
Paragraphe 37/38 Even this did not put an end to his torture. Van Gogh staggered back to the inn where he was lodging and lingered for two days before dying.
Paragraphe 38/38 His poignant last words, according to Theo, the distraught brother who had rushed to his side, were : 'The sadness will go on for ever.'

FIGURE 4.1 – L'importance de la position dans le genre journalistique

au contraire un exemple manifeste de bruit documentaire. Il est intéressant d'imaginer comment éviter qu'un système automatique produise ce type d'erreur.

Une réponse, au niveau phrastique, serait de dire que la classification aurait été meilleure si le système avait tenu compte de l'information temporelle contenue dans le segment « *in 1903* ». Une approche textuelle répondrait en utilisant une autre information : la position. La phrase en question se situe dans le corps d'un article relativement long (Figure 4.1). À cette échelle, l'information apparaît comme secondaire. Ce constat ne suppose pas de connaissance préalable sur la langue anglaise. La connaissance de la marque de paragraphe suffit à ce diagnostic [40]. Il est très peu probable qu'une information importante soit nichée dans le 29ème paragraphe, sur 38, d'un article de presse et reprise nulle part ailleurs.

Le journaliste expose des informations dans le texte et le placement de ces informations donne des renseignements sur leur importance. La mise en lumière n'est pas nécessairement verbalisée. Il est rare que les journalistes utilisent des moyens lexicaux pour le faire, écrire « ceci est important » ou « ceci ne l'est pas » n'est pas typique du le style journalistique. Ces tournures appartiennent plutôt au domaine de l'oral. A l'opposé, la position dans le texte fournit des informations sur la pertinence de tel ou tel segment. Pour exploiter les positions nous nous inspirons des travaux sur les invariants de genre menés par Nadine Lucas ([Lucas-2004, Lucas-2009a]).

40. D'autres mesures de la position pourraient être imaginées : position des phrases ou des caractères...

Les différentes positions dans le texte sont ici définies de la manière suivante :

Début de texte : composé idéalement du titre de l'article

Début de corps : contenant les deux premiers paragraphes

Fin de corps (pied) : comprenant les deux derniers paragraphes

Reste du corps (tronc) : constitué du reste des éléments textuels (paragraphes et intertitres éventuels)

Les débuts, de texte et de corps, et la fin du corps constituent les positions remarquables du texte. Le début de texte (titre) et le début de corps forment la **tête** de l'article. Le début de corps correspond idéalement [41] au chapeau (ou chapô) de l'article. En effet, les lois du genre journalistique accordent à ce segment de texte, lorsqu'il est présent, une importance capitale. Le chapeau fait partie des outils que le journaliste utilise pour accrocher le lecteur. Il permet d'éclairer sur les composantes principales de l'évènement décrit dans l'article. Après avoir lu le chapeau, le lecteur doit savoir de quoi l'on parle et si la suite de l'article va l'intéresser. Le chapeau est un indice visuel important pour celui qui interprète le texte. Le choix effectué par le journaliste de mettre en lumière certains éléments et pas d'autres n'est pas anodin.

Garder les éléments visuellement saillants est une propriété qui a par exemple été utilisée par Lehtonen ([Lehtonen-2007]) dans le cadre de campagnes d'évaluation en RI sur des documents *XML*. La structure même de ces documents a été exploitée pour des tâches de classification de documents ([Doucet-2006a, Elhadj-2012]). Ces indices dépassent le cadre purement littéral. Ceci permet de s'affranchir d'une vision uniquement lexicale du traitement des textes.

Le tableau 4.1 (page 68) propose une mise en saillance de certains éléments de l'exemple du tableau précédent : noms des protagonistes en bleu, information principale (*cut* et *ear*) en vert et le nom de maladie en rouge. Les éléments bleus et verts sont répétés à différentes positions dans les segments de textes sélectionnés pour l'exemple mais il serait utile d'avoir une vision complète du texte. Le formalisme précédemment décrit permet de produire le tableau 4.2 (page 68) qui représente la distribution des mêmes éléments dans le texte intégral. Le segment où le nom de maladie est placé ne figure pas à une position remarquable. Par ailleurs, il n'est pas répété. Or, la répétition est un effet de style important dans le genre journalistique.

Les éléments importants d'information sont répétés à des positions dans le texte qui sont facilement identifiables. Ces éléments figurent ordinairement à au moins deux de ces positions. Nous pouvons voir que les termes « Gauguin » et « Van Gogh » ont une distribution riche (en bleu). Il en est de même pour les termes relatifs à l'oreille coupée de Van Gogh (en vert). La position et la répétition permettent donc ici de hiérarchiser l'information sans avoir recours à une analyse locale.

41. Idéalement, car il n'est pas toujours repérable automatiquement.

TITRE : One was an arrogant bully. The other was a nervous wreck. So what is the truth about the war of **Van Gogh**'s **ear** ?

Paragraphe 1/38 The iconic story about Dutch-born painter Vincent **Van Gogh** **cutting** off his own **ear** and presenting it as a gift to his favourite prostitute may not be true after all.
Paragraphe 2/38 Or so say some German art historians, who now claim the famous **ear** was **cut** off in a fight with rival artist Paul **Gauguin**.
...
Paragraphe 29/38 **Gauguin** took himself off to Tahiti where he entertained under-age mistresses, consumed vast quantities of absinthe and morphine and **died of** **syphilis** in 1903.
...
Paragraphe 37/38 Even this did not put an end to his torture. **Van Gogh** staggered back to the inn where he was lodging and lingered for two days before dying.
Paragraphe 38/38 His poignant last words, according to Theo, the distraught brother who had rushed to his side, were : 'The sadness will go on for ever.'

Tableau 4.1 – Représentation des occurrences de différents termes dans notre exemple en anglais. En **rouge** le nom de maladie ayant entrainé l'erreur de classification. En **bleu** les noms des deux peintres dont il est question. Les constituants de l'évènement principalement décrit dans l'article apparaissent en **vert**.

Segment	Début de segment	Fin de segment
Début de texte		**V E**
Début de corps	**V C E**	**E C G**
Tronc	VGVVCEGCEVGGVVG	GGVGGGVCEGVGVVVGVCEVVGVGG**S**VEVV
Pied	**V**	

Tableau 4.2 – Représentation des occurrences de différents termes dans un article en anglais. En **rouge** le nom de maladie ayant entrainé l'erreur de classification. En **bleu** les noms des deux peintres dont il est question (G pour Gauguin et V pour Van Gogh). Les constituants de l'évènement principalement décrit dans l'article (E pour *ear* et C pour *cut* et ses synonymes) apparaissent en **vert**.

Cette représentation de la position des informations importantes permet d'inférer deux propriétés. Ces informations sont répétées et se trouvent régulièrement à des positions spécifiques : les débuts et les fins de segments.

Nous allons pratiquer l'opération inverse, c'est-à-dire extraire ce que l'on trouve à des positions d'intérêt. Ces positions se combinent pour former des figures. Une figure est définie comme un couple de positions dans le texte. Ce couple comporte au moins une position remarquable. Les positions remarquables dans un article de presse sont les

suivantes (entre parenthèses l'abréviation utilisée dans les tableaux) :

- début de texte (D) ;
- premier paragraphe (P) ;
- second paragraphe (S) ;
- avant-dernier paragraphe (A) ;
- paragraphe de fin (F).

Deux autres positions existent mais ne sont pas considérées comme remarquables : le début (H pour haut) et la fin (B pour bas) du tronc. Ces positions ne peuvent appartenir à une figure qu'en étant associée à une position remarquable telle que définie plus haut. Dans notre modèle de document le couple HB n'est donc pas une figure, il est ignoré en tant que couple. Par contre, les couples DH et DB sont bien des figures puisqu'elles comportent une position remarquable.

L'ordre linéaire du texte peut être représenté de la façon suivante : **D P S** H B **A F**. Dans le tableau 4.3, cet ordre linéaire disparaît au profit d'une représentation tabulaire, en deux dimensions, mettant en valeur les oppositions entre les débuts et les fins de segments.

début de texte (**D**)	
premier paragraphe (**P**)	(**S**)second paragraphe
début de tronc (H)	(B) fin de tronc
avant-dernier paragraphe (**A**)	(**F**) paragraphe de fin

Tableau 4.3 – Représentation schématique des oppositions début-fin de segments

La sélection par la position permet d'extraire des termes pertinents sous forme de n-grammes de mots avec n non fixé. Le tableau 4.4 (page 70) présente les 20 plus longs n-grammes de mots répétés et maximaux figurant à des positions remarquables dans le texte. Ces n-grammes sont classés par taille en caractères décroissante (espaces compris). L'indice des paragraphes où ils apparaissent est présenté ainsi que les figures impliquées. Les n-grammes présentés sont maximaux : c'est-à-dire qu'ils ne sont pas des sous-motifs de n-grammes plus grands et de même effectif. Les unigrammes « Paul » et « Gauguin » sont ainsi maximaux car ils n'apparaissent pas systématiquement dans le bigramme « Paul Gauguin ». Une définition plus complète de la maximalité figure dans la section 4.3.3.

Quelques exemples biens choisis ne peuvent bien sûr pas constituer à eux seuls la validation d'une approche. Toutefois, cela permet d'illustrer l'intérêt de la structure du document dans l'interprétation. Nous faisons l'hypothèse que l'interprétation de cette structure est **indépendante de la langue** car elle est du domaine de la rhétorique. Ce point est un fondement capital de notre méthode d'analyse d'articles de presse. L'intérêt que nous portons à la structure du document est liée à son lien fort avec le genre textuel.

Selon le genre textuel, la structuration des documents est différente. Cette structuration est en quelque sorte une convention entre l'émetteur et le récepteur du message. Elle correspond à la meilleure articulation possible entre les stratégies d'écriture et les

N-grammes répétés	Figures	Numéro des paragraphes
Paul Gauguin	SB	[**2**,22]
his own ear	PB	[**1**,21]
prostitute	PH-PB-HB	[**1**,18,30]
the famous	SH-SA-HA	[**2**,16,**37**]
favourite	SH	[**1**,4]
now claim	SH	[**2**,12]
Van Gogh	DP-DH-DB-DA-PH-PB-PA-HA-BA	[**0**,**1**,3,4,6,14,16,19,21-28,30,33,**37**]
cut off	SH	[**2**,13]
Gauguin	SH-SB	[**2**,4,13-23,25,26,27,28,29]
off his	PH	[**1**,11,13]
painter	PH-PB	[**1**,6,9,14,34]
artist	SH-SB	[**2**,13,14,19,31]
famous	SH-SB	[**2**,5,16,17,33,**37**]
to his	PH-PB	[**1**,8,25,26,27]
about	DP-DB-PB	[**0**,**1**,28]
after	PH-PB	[**1**,14,16,23,30]
bully	DH	[**0**,19]
Dutch	PB	[**1**,20]
other	DH-DB	[**0**,3,14,27,28]
Paul	SB	[**2**,22,34]

Tableau 4.4 – Les 20 plus longs (en caractères) n-grammes maximaux de mots figurant à des positions remarquables (indices des paragraphes d'apparition, 0 étant le titre). En gras, les paragraphes inclus dans des positions remarquables.

stratégies de lecture, entre les techniques d'encodage et les techniques de décodage de l'information ([Lejeune-2012a]). Il s'agit pour l'émetteur de s'assurer que son message soit le moins bruité possible. Pour ce faire, il facilite la tâche du lecteur, il crée la meilleure relation de communication possible, de façon à transmettre efficacement l'information.

Ces stratégies ont été étudiées dans le domaine de l'oral, de la conversation. Pour Grice ([Grice-1975]), une bonne communication résulte d'une coopération entre l'émetteur et le récepteur du message. Plus précisément, il définit dans ses maximes conversationnelles quatre critères qui régissent une communication efficace :

Qualité : l'énoncé ne doit rien contenir de faux ou de non–démontrable

Quantité : l'énoncé doit être aussi informatif que nécessaire

Pertinence : les informations transmises doivent susciter l'intérêt

Manière : la brièveté et l'absence d'ambiguïté sont requises

Les maximes de Grice ont été longuement étudiées. Baylon ([Baylon-2005]) en propose une vision condensée tandis que Foudon ([Foudon-2008]) offre une analyse plus récente et plus complète. Nous postulons que ces maximes sont également applicables au genre article de presse dans la mesure où il s'agit de textes de communication, idée qui figure également chez Maingueneau ([Maingueneau-2005]).

4.1.2 La relation de communication entre l'auteur et le lecteur du document

Les articles de presse sont des textes de communication de masse. La mission du journaliste est d'assurer que son intention de communication soit correctement interprétée par son lecteur. Or, plutôt qu'un lecteur destinataire unique de l'article, il s'adresse à une myriade de lecteurs avec des compétences cognitives diverses et des connaissances variées. Et c'est à cet ensemble hétéroclite qu'il souhaite transmettre son message.

Sperber et Wilson ([Sperber-1998]) proposent une unification des principes conversationnels de Grice sous l'égide **du principe de pertinence**. Pour les auteurs, une communication efficace impose de **minimiser les coûts de traitement** et de **maximiser les effets cognitifs**. Appliquons ces principes aux articles de presse : le journaliste doit s'assurer que le lecteur ait le moins possible d'inférences à faire pour retrouver l'intention de communication de l'auteur. Les inférences sont en effet cognitivement coûteuses pour le lecteur. Par ailleurs, plus le lecteur doit faire d'inférences, plus il est probable qu'il commette des erreurs d'interprétation. Les aspects centraux des évènements décrits doivent donc être facilement détectés par les lecteurs, il faut donc suivre un schéma adapté d'exposition des informations.

Le journaliste s'adresse en fait à une abstraction de public, à la représentation qu'il se fait de son lectorat. Il a donc besoin d'un lecteur modèle, un « lector in fabula » au sens d'Umberto Eco ([Eco-1985]). Comme dans les textes narratifs, cœur de l'étude d'Umberto

Eco, les articles de presse sont écrits en fonction d'une stratégie textuelle où le lecteur final est central. Le journaliste s'assure donc que le lecteur puisse effectivement extraire l'information qu'il a voulu transmettre. Il veille à ce que la construction de l'article facilite l'interprétation par le lecteur.

La formation des journalistes va d'ailleurs dans ce sens. Le principe des 5W, voir par exemple les travaux de Itule et Anderson ([Itule-2006]), compile les bonnes pratiques à mettre en œuvre dans l'écriture d'un article de presse. Les 5W sont les 5 informations majeures qui doivent figurer dans l'article : Qui (*Who*), Quoi (*What*), Où (*Where*), Quand (*When*), Pourquoi (*Why*). Les réponses à ces questions doivent être facilement repérées par le lecteur.

Ce principe est du domaine de la rhétorique [42]. Les réponses aux cinq questions sont placées à des positions clés de l'article : le titre et le chapeau (*headline* en anglais). Dans la suite de l'article, ces informations sont développées, explicitées. Cette stratégie d'écriture globale est connue sous le nom de pyramide inversée, la règle des 5W est un principe central de cette stratégie

Des explications technologiques et économiques ont été avancées pour expliquer l'apparition de cette stratégie. La faible vitesse et le coût élevé du télégraphe auraient poussé les journalistes à écrire des textes plus concis ([Wendt-1979]). La justification communicationnelle possède aussi ses défenseurs. La popularisation de cette stratégie d'écriture serait due à une excellente adéquation aux besoins éprouvés par les lecteurs de mettre en œuvre des stratégies de déchiffrage économes ([Pottker-2003]).

Ces stratégies parcimonieuses des lecteurs offrent des pistes intéressantes pour décoder informatiquement les textes. C'est d'ailleurs l'approche prônée par Jacques Coursil ([Coursil-2000]) lorsqu'il décrit la fonction muette du langage. L'interprétation devient alors le point de départ de l'activité d'analyse de la langue. Celle qu'effectue celui qui lit un journal est un excellent socle pour bâtir une méthodologie d'analyse automatique plus parcimonieuse que ce qui est habituellement pratiqué. Nous ne défendons pas pour autant l'idée que l'intelligence artificielle soit nécessairement une simulation de l'activité humaine. Néanmoins, l'idée de s'inspirer des stratégies humaines nous paraît justifié dès lors que ces stratégies sont efficaces et transposables.

Il convient donc de modéliser le plus simplement possible l'activité de déchiffrage des articles de presse menée par les lecteurs humains. Nous supposons que le moindre coût cognitif est lié à un moindre coût computationnel. Ce n'est pas la connaissance exhaustive des structures phrastiques ou du vocabulaire qui permettent à un locuteur de la langue de déchiffrer efficacement un article de presse. Ce sont les structures textuelles et l'adaptation du vocabulaire qui assurent que le canal source-cible soit le moins bruité possible. Ces stratégies tiennent plus de la communication que de la langue. C'est ce qui les rend adaptées d'un point de vue informatique à un objectif de couverture multilingue.

Les universaux, les principes alingues qui régissent le genre textuel sont les indices

42. Le rhéteur grec Hermagoras retient lui deux éléments supplémentaires : le moyen et la manière.

qui guident notre approche. Plutôt que la structuration au niveau phrastique, fortement dépendante de la langue, c'est la structuration au niveau textuel, supposée indépendante des langues, qui est utilisée.

4.1.3 Synthèse sur l'utilisation du grain document

Pour respecter le principe de factorisation précédemment exposé, un noyau d'analyse bâti autour des « lois du genre » est défini. Ces lois régissent les pratiques de lecture et d'écriture en vigueur dans un genre textuel donné, indépendamment de la langue traitée. La structuration de l'information dans le texte permet de remettre en cause la méthode *informatique* de recherche de l'information. Plutôt que de rechercher quelque chose de défini (stocké en mémoire) partout dans le texte, c'est ce qui est présent à une **position définie** qui constitue la marque pertinente. Les positions remarquables dans les textes sont exploitées car elles sont dépendantes du genre plutôt que de la langue. Nous disposons ainsi d'un grain d'interprétation compatible avec notre volonté de couverture multilingue. D'ailleurs, le grain phrase, habituellement utilisé en TAL, paraît bien petit au vu de la taille des corpus à analyser. Le décodage des phrases une à une n'est pas adapté au traitement de corpus qui s'enrichissent quotidiennement de dizaine de milliers de documents.

4.2 Termes médicaux et vocabulaire journalistique

Nous avons exposé précédemment pourquoi des ressources de grande taille étaient un frein à la couverture multilingue : les coûts de construction et de traitement qu'elles impliquent reflètent une **inadaptation à la tâche**. Nous considérons par ailleurs que la constitution de ressources très descriptives (de type ontologies) n'est pas envisageable pour la plupart des langues. Nous montrons même qu'elle n'est pas nécessaire.

L'utilisation du lexique n'est pas la même selon le genre textuel auquel on s'intéresse. L'auteur d'un article de presse s'adresse à un public fondamentalement divers, par ses attentes en terme de contenu et par ses connaissances du domaine traité. Les choix de vocabulaire effectués se basent sur ce constat : avant d'introduire des termes nouveaux ou scientifiques, il faut s'assurer que l'essence du message est transmise par des termes courants. C'est une stratégie qui offre des possibilité intéressantes d'exploitation d'un point de vue informatique. La description exhaustive du vocabulaire d'un domaine est superflue, le vocabulaire en usage dans le genre textuel est suffisant : du point de vue de l'humain pour comprendre les textes, du point de vue de la machine pour les traiter efficacement.

4.2.1 Considérations sur le vocabulaire journalistique dans le domaine des maladies infectieuses

Dans le domaine des maladies infectieuses, le vocabulaire recouvrera deux catégories principales. En premier lieu nous avons la catégorie des vecteurs de transmission de la maladie. Pour les cas les plus fréquents, le vecteur est d'origine bactérienne ou virale. Nous devons y ajouter les champignons (mycoses) et les parasites (parasitoses). La seconde catégorie contient les noms de maladies, qui définissent les différentes situations d'altération de la santé d'un individu. La partie visible et descriptible des maladies sera constituée par les symptômes.

L'examen des corpus montre que le nom de la maladie infectieuse et de son vecteur sont régulièrement confondus. Par exemple, le terme « Ebola » dans les articles de presse désigne à la fois le virus et la maladie : la fièvre hémorragique. Dans le genre qui nous intéresse, la différenciation entre le virus et la maladie n'est pas systématiquement pertinente. Dans l'usage des journaux non-spécialisés, les classes vecteurs, maladies et symptômes sont très peu distinguées. En effet, la distinction entre ces classes de termes n'est pas pertinente pour la grande majorité des lecteurs. La richesse offerte par une ontologie se justifie sans doute du point de vue du spécialiste, elle est par contre superflue pour le grand public.

Dans la suite de cette étude nous utiliserons, par abus de langage assumé, le terme « maladie » pour désigner la maladie infectieuse concernée ou son vecteur. En corpus le terme utilisé par le journaliste a fonction de déclencheur. Le choix du terme ne s'opère donc pas sur des critères sémantiques mais sur des critères purement communicationnels : créer une réaction chez le lecteur. Le but n'est pas de décrire de façon rigoureuse le processus de contamination mais d'alerter le lecteur. Il est donc possible de se passer d'une description exhaustive du vocabulaire lié aux maladies pour analyser les articles de presse sans craindre d'affecter le rappel.

4.2.2 Illustrations de l'usage du vocabulaire médical dans la presse

La figure 4.2 (page 76) offre un premier exemple de variété de vocabulaire dans la presse. Il est tiré d'un article du *Figaro*[43]. Il y est question d'un nouveau virus, qui possède des symptômes proches du SRAS et appartient à la famille des coronavirus. Trois termes différents font office de déclencheurs : SRAS, nouveau virus et coronavirus. Nous pouvons y ajouter l'anaphore grammaticale « cette maladie » (paragraphe 7).

Pour autant, il n'est pas impératif d'identifier ces quatre formes pour identifier l'évènement décrit dans l'article. La variété du vocabulaire permet au lecteur de déterminer qu'il est question de maladie, sans connaître l'intégralité du vocabulaire, ni même décoder l'anaphore nominale. Nous avons donc ici une isotopie au sens de Greimas ([Greimas-1970]) ; c'est à dire des redondances qui assurent la cohérence et facilitent la démarche d'interpré-

43. `http://sante.lefigaro.fr/actualite/2012/09/24/19135-decouverte-dun-nouveau-virus-proche-sras` consultée le 12 octobre 2013

tation du texte. Au-delà de l'isotopie purement sémantique, telle que la concevait Greimas à l'origine, nous concevons les isotopies en tant que traces du chemin à suivre pour interpréter un texte. Certaines isotopies sont conçues par l'auteur de façon intentionnelle, d'autres peuvent résulter d'interprétations qui n'étaient pas envisagées initialement.

Nous proposons dans la figure 4.3 (page 77) la même expérience que dans la figure 4.2 mais sur un texte en grec. Remarquons que cette fois la forme initiale (en rouge) a une distribution plus faible. Ceci n'est pas étonnant dans la mesure où le grec est une langue à déclinaisons. Le terme « $\gamma\rho\iota\pi\eta$ » (grippe) connaît ainsi deux formes différentes dans ce texte, avec ou sans ζ en finale. La variété en vocabulaire est par contre comparable à notre exemple en français.

Segment	exemple en français		exemple en grec	
	Début	Fin	Début	Fin
Début de texte	**N**	**S**		**N**
Tête		**N S S**		**N S**
Corps	**S S S S**	**N S N**	**S S S S**	**F F**
Pied	**S N**	**S S S S**	**F**	**S**

Tableau 4.5 – Représentation des occurrences de noms de maladies dans deux exemples en français et en grec. En **rouge** la maladie principalement décrite dans l'article, en **bleu** les termes qui la rappellent (S pour les synonymes et F pour les variations de forme du nom initial). Les autres noms de maladies apparaissent en **vert**

4.2.3 L'articulation lexique-texte

Nous proposons dans le tableau 4.5 une représentation structurée des occurrences de nom de maladie dans nos deux exemples. Pour calculer la démarcation débuts–fins utilisée dans ce tableau nous cherchons le milieu du segment concerné pour avoir deux parties de taille égale à un grain donné. Nous utilisons successivement les grains suivants pour identifier une dichotomie valable : paragraphes puis phrases puis caractères. Le pied et le chapeau sont donc chacun découpés en deux paragraphe distincts. Le titre est découpé en deux chaînes de caractères. Si le corps comporte un nombre impair de paragraphes, le paragraphe central est alors découpé en phrases ou à défaut en caractères.

Nous pouvons observer certaines constantes dans la position des répétitions de noms de maladie. Le premier terme déclencheur utilisé dans le texte permet de dessiner des relations entre le titre, le chapeau et le pied. Les synonymes utilisés pour relayer ce terme initial sont présents dans le chapeau et le corps. La variété du vocabulaire est donc plus grande que la variété des positions. Les relations découvertes permettent de discriminer le sujet réel de l'article de ce qui est accessoire. Dans notre cas, il s'agit des autres noms de maladie qui sont recensées mais ne définissent pas le thème, le sujet central de l'article. Nous pouvons donc énoncer le principe suivant : de même qu'il n'est pas nécessaire de

TITRE :Découverte d'un **nouveau virus** proche du **Sras**

Paragraphe 1/7 Un homme est hospitalisé et deux autres sont décédés après un séjour en Arabie saoudite. Les autorités sanitaires se veulent toutefois rassurantes car aucune contagion d'homme à homme n'a encore été détectée.
Paragraphe 2/7 Un **nouveau virus** appartenant à la famille du **Sras** (**syndrome respiratoire aigu sévère**), responsable de la mort de 800 personnes en 2002, a été identifié sur un Qatarien hospitalisé à Londres dans un état grave après avoir séjourné récemment en Arabie saoudite. C'est le troisième malade qui contracte ce **coronavirus** lors d'un séjour au Moyen-Orient, mais les autorités sanitaires se veulent rassurantes car aucune contamination d'homme à homme n'a pour l'instant été identifiée.
Paragraphe 3/7 C'est l'Organisation mondiale de la santé (OMS) qui a annoncé lundi la nouvelle via son système «d'alerte et de réponse globale». «Le patient est toujours en vie, mais d'après ce que nous savons, dans un état grave», a déclaré Gregory Hartl, porte-parole de l'OMS. L'homme de 49 ans souffre d'insuffisance respiratoire et rénale.
Paragraphe 4/7 Il existe un grand nombre de **coronavirus**. En général, ils provoquent des **rhumes** chez les humains. Mais une forme particulière de **coronavirus** à l'origine du **Sras** en 2002 avait tué 774 personnes dans le monde, dont 349 en Chine. Plus de 8000 personnes avaient été infectées.

INTERTITRE Vigilance à l'approche du pèlerinage de La Mecque

Paragraphe 5/7 À Londres, l'Agence de protection de la santé (Health Protection Agency, HPA) a indiqué que ce **nouveau virus** était «différent de ceux qui avaient jusqu'à présent été identifiés chez l'être humain». Elle a noté toutefois que les premières investigations n'avaient révélé aucune contamination des personnes ayant été en contact avec le malade, y compris le personnel de santé. Le ministère saoudien de la Santé a confirmé que deux autres patients ont été diagnostiqués dans le pays. Tous deux sont décédés.
Paragraphe 6/7 Pour Peter Openshaw, directeur du centre des **infections respiratoires** à l'Imperial College de Londres, le **nouveau virus** ne semble pas, à ce stade, devoir être un sujet de préoccupation publique. «Pour le moment, la vigilance s'impose mais pas l'inquiétude», a-t-il dit.
Paragraphe 7/7 Même si elle n'a donné aucune consigne de restriction sur les déplacements, l'OMS surveille avec attention l'émergence potentielle de **cette maladie** alors que de nombreux voyageurs commencent à affluer en Arabie saoudite pour le pèlerinage (Hadj) de La Mecque. Ce pèlerinage rassemble environ 2,5 millions de fidèles chaque année et a donné lieu à des épidémies par le passé, notamment de **grippe**, de **méningite** ou de **poliomyélite**.

FIGURE 4.2 – Richesse du vocabulaire en français journalistique. En **rouge** la maladie principalement décrite dans l'article, en **bleu** les termes qui la rappellent. Les autres noms de maladies apparaissent en **vert**

Titre:Συναγερμός στο Χονγκ Κονγκ για τη **γρίπη** των πτηνών
Paragraphe 1/9 Εμπάργκο στις εισαγωγές πουλερικών.
Paragraphe 2/9 Οι αρχές του Χονγκ Κονγκ αναβάθμισαν τον συναγερμό για την **γρίπη** των πτηνών και επέβαλαν προσωρινό εμπάργκο στις εισαγωγές ζωντανών πουλερικών μετά το θάνατο τριών πτηνών στα οποία ανιχνεύτηκε ο ιός **H5N1**.
Paragraphe 3/9 Το επίπεδο συναγερμού αναβαθμίστηκε σε «σοβαρό», στην τρίτη βαθμίδα της κλίμακας με πέντε βαθμίδες.
Paragraphe 4/9 Ο επικεφαλής των Υπηρεσιών Υγείας του Χονγκ Κονγκ Γιορκ Τσόου ανακοίνωσε την απαγόρευση, για προληπτικούς λόγους, των εισαγωγών ζωντανών πουλερικών και τη θανάτωση 17.000, μετά τον εντοπισμό σε αγορά του Χονγκ Κονγκ κοτόπουλου που είχε προσβληθεί από τη **νόσο**.
Paragraphe 5/9 Το πτηνό βρέθηκε θετικό στον ιό **H5N1** μετά τον έλεγχο ρουτίνας που πραγματοποιήθηκε σε αγορά, τόνισε ο ίδιος και πρόσθεσε πως θετικά στον ιό **βρέθηκαν** και άλλα δύο πτηνά-ένα σπουργίτι και ένας **γλάρος**.
Paragraphe 6/9 Ο άντρας που βρήκε τον γλάρο στην αυλή ενός σχολείου και ο γιος του εμφάνισαν συμπτώματα της **γρίπης** και νοσηλεύτηκαν για μικρό χρονικό διάστημα.
Paragraphe 7/9 Σύμφωνα με την κρατική τηλεόραση PTHK, περίπου είκοσι μαθήτριες ενός σχολείου θηλέων ηλικίας έξι και επτά χρόνων παρουσιάζουν τα συμπτώματα της **γρίπης**, αλλά δεν νοσηλεύονται.
Paragraphe 8/9 Το Χονγκ Κονγκ ήταν η πρώτη χώρα που ανακοίνωσε το 1997 επιδημία της **γρίπης** των πτηνών, με έξι νεκρούς, η οποία προκλήθηκε από την μετάλλαξη του μέχρι τότε άγνωστου ιού. Εκατομμύρια πουλερικά είχαν θανατωθεί τότε.
Paragraphe 9/9 Η μετάδοση του**ιού H5** γίνεται από τα ζώα στον άνθρωπο, όμως οι επιστήμονες φοβούνται πως μια μετάλλαξη θα επιτρέψει στον ιό να μεταδίδεται από άνθρωπο σε άνθρωπο, προκαλώντας μια θανατηφόρα πανδημία.

FIGURE 4.3 – Richesse du vocabulaire sur un article en grec. En **rouge** la maladie principalement décrite dans l'article, en **bleu** les termes qui la rappellent. Les autres noms potentiellement déclencheurs apparaissent en **vert**

connaître tous les mots d'une phrase pour la comprendre, il n'est pas obligatoire d'obtenir une représentation sémantique de toutes les phrases pour analyser un texte.

Le fait d'utiliser des indices de plus haut niveau permet de restreindre l'utilisation des ressources et donc de garantir, dans une certaine mesure, un statut égal à toutes les langues qui doivent être traitées pour une tâche donnée. Au contraire avec des approches coûteuses, les langues de diffusion plus faible ou représentant une moindre force de frappe économique sont ainsi de fait laissées de côté. Il faudrait alors séparer les corpus multilingues en diverses sous–catégories de langues plus ou moins facilement traitables. L'autre solution consiste à se contenter de ressources de taille raisonnable, ressources que l'on espère obtenir à moindre coût et pour un maximum de langues. Les résultats sur une seule langue peuvent être moins probants mais, les possibilités de couverture accrue ou de complémentarité deviennent très intéressantes (voir par exemple [Ferret-2006] sur la segmentation thématique).

Plusieurs travaux portant sur les questions multilingues ont fait le constat que la dépendance aux ressources externes était problématique. Une tradition caennaise s'est ainsi construite ces dernières années autour d'une analyse guidée par un modèle de document et basée sur un traitement de chaînes de caractères. Deux thèses ([Brixtel-2011, Lecluze-2011]) portant sur l'alignement de documents multilingues (ou multi-documents) ont ainsi utilisé pour fondement ce couplage modèle de document–analyse au grain caractère.

Ce type d'approche constitue en quelque sorte une synthèse de deux voies envisagées pour élaborer des traitements multilingues. Chez Kando ([Kando-1999]) c'est la structure du document qui est mise en avant pour faciliter l'accès à l'information. Mc Namee ([McNamee-2004]) utilise des n-grammes de caractères pour des tâches de recherche d'information. Peu de travaux ont opéré cette synthèse, et aucun à notre connaissance ne l'a appliquée à la veille.

Nous décrivons dans la section suivante comment l'analyse au grain caractère permet de respecter les principes de factorisation et de parcimonie.

4.3 Le grain caractère comme unité d'analyse

Le mot reste l'unité d'analyse la plus couramment utilisée en traitement de la langue, que ce soit dans l'approche fondée sur les données (dite aussi approche statistique) ou dans l'approche fondée sur des règles (ou approche symbolique). La segmentation en mots est alors une étape indispensable à tout traitement automatique [44]. Pourtant, le mot n'existe pas dans toutes les langues. Mounin ([Mounin-1974]) écrit ainsi que **le mot n'est pas une réalité de linguistique générale**. En effet, la notion de mot est liée au système d'écriture. Elle n'est pas liée à la faculté de langage à proprement parler.

La segmentation en mots est principalement justifiée par des aspects lexicographiques : c'est l'unité d'entrée dans les dictionnaires. Il est alors tentant d'utiliser le critère sémantique pour motiver la segmentation en mots. Ce serait une unité facile à détecter, possédant par ailleurs une grande valeur pour la « construction du sens ». Mais la segmentation elle-même pose problème (Section 4.3.1). Dans une visée multilingue, il semble donc plus pertinent de s'intéresser aux chaînes de caractères (Section 4.3.2). Les propriétés des chaînes utilisées sont décrites dans la Section 4.3.3 et nous montrons quelques applications de l'analyse au grain caractère dans la Section 4.3.4.

4.3.1 Les difficultés posées par l'extraction des mots graphiques

L'extraction des mots graphiques suppose de disposer d'un *segmenteur*. Usuellement, un segmenteur repose sur un certain nombre de règles. L'approche la plus simple d'un point de vue multilingue est de recenser dles séparateurs adaptés à un grand nombre de langues :

44. On en viendrait même à ce que ce découpage soit tellement implicite qu'il ne soit plus utile de le mentionner. Dire que l'on découpe en mots serait aussi banal que de dire la première étape du traitement est d'ouvrir des fichiers.

blanc typographique, virgule, point et autres signes de ponctuation. Certains cas limites peuvent se présenter, par exemple des noms propres (G.W.Bush) ou des informations chiffrées (100.000). Ces cas peuvent toutefois être considérés comme suffisamment rares pour ne pas remettre en cause le modèle.

Un premier problème sera constitué par la recomposition des mots graphiques en unités de sens. Par exemple, reconstruire « wake up » à partir de « wake » et « up » ou « pomme de terre » à partir de « pomme » « de » et « terre », n'est pas trivial. La quantité importante de travaux qui s'intéressent aux unités multi-mots (*multiword units*), y compris sur le plan monolingue, en est une preuve.

Le problème des langues flexionnelles, et parmi elles des langues agglutinantes, est plus notable encore. Dans les langues agglutinantes, les mots graphiques comportent un nombre important de morphèmes. De ce fait, un mot graphique dans une langue agglutinante peut correspondre à plusieurs mots dans une autre langue. Il est difficile dès lors de conserver le critère « d'unité sémantique » pour justifier la segmentation en mots. De plus, la phase de découpage en mots n'exclut pas des recompositions ultérieures comme la lemmatisation. L'automatisation de ce procédé est encore mal maîtrisé dans la majorité des langues. L'argument de la fiabilité opératoire du découpage en mots est donc peu pertinent.

Enfin, la segmentation en mots graphiques pose problème dans des langues qui ne connaissent pas le mot, comme le chinois. Il faut alors trouver des règles permettant de segmenter ces langues en l'absence du séparateur le plus commun : l'espace typographique. Cette solution est la plus couramment utilisée : le mot étant le grain d'analyse « incontournable », il faudra en obtenir des équivalents en chinois ou en thaï. Une solution plus souple pourrait être l'utilisation du mot typographique quand il existe ou d'une autre unité dans le cas contraire. Le caractère est légitime pour prendre cette place. Nous défendons ici une solution plus radicale : utiliser le caractère quelle que soit la langue. En effet, cette unité est directement accessible dans un texte sous forme électronique. Ceci est vrai quelle que soit la langue traitée, le grain caractère est compatible avec notre objectif de factorisation maximale.

4.3.2 L'apport du grain caractère

Le grain qui nous semble le plus compatible avec une vision multilingue est le grain caractère, plus précisément les chaînes de caractères. L'hypothèse est que ce sont les suites de caractères, **mots ou non–mots**, qui vont constituer la meilleure unité d'analyse d'un point de vue multilingue. Isoler les caractères est techniquement plus aisé qu'isoler les mots puisqu'il est possible d'utiliser le format *Unicode*. La segmentation est donc directement gérée au niveau du système d'encodage des caractères. Cette technologie est éprouvée et offre une grande couverture : plus de 650 langues peuvent être écrites avec *Unicode*. Ceci permet de régler la question technique.

La question suivante est de savoir comment il est possible d'analyser au grain caractère. Nous nous intéressons aux chaînes de caractères qui sont répétées. Nous faisons

l'hypothèse que la répétition est un facteur important pour déterminer la pertinence d'une chaîne de caractères pour extraire de l'information. Il ne s'agit pas de chercher l'effectif (ou *frequency*) le plus grand. Un élément (un mot par exemple) ayant un effectif important dans un document est probablement peu pertinent pour décrire ce document. Cette propriété se vérifie pour les chaînes de caractères ([Lecluze-2013]).

Par contre, le fait qu'un élément soit répété est un indice important. C'est le passage de la situation d'*hapax* (effectif de 1) à la situation de répétition (effectif de 2) qui est importante. Church ([Church-2000]) a décrit cet effet : la probabilité qu'un terme soit répété n'est pas toujours le carré de la probabilité qu'il soit présent une fois. Autrement dit, la probabilité de répétition est conditionnelle. Ceci est d'autant plus vrai si c'est un mot plein. C'est ce que Church nomme « la mesure d'adaptation ». Les mots pleins, pertinents pour des tâches de fouille de textes, ont une meilleure mesure d'adaptation que les mots vides. La probabilité de la seconde occurrence du mot plein est nettement supérieure à la probabilité de la première occurrence. Au contraire, les mots vides ont une probabilité d'apparition globalement constante. La répétition est donc un critère important pour juger de la pertinence d'un mot pour décrire le document. Nous faisons la même hypothèse concernant les chaînes de caractères. De façon à réduire le nombre de chaînes à analyser, il est utile de se restreindre aux chaînes répétées qui sont maximales.

4.3.3 Propriétés des chaînes de caractères répétées maximales

Les chaînes de caractères que nous utilisons sont des motifs sans trous tels que décrits par ([Ukkonen-2009]). Nous appelons « $rstr_{max}$ » ces chaînes de caractères répétées maximales.

Les deux caractéristiques constitutives des $rstr_{max}$ sont les suivantes :

la répétition : la chaîne a un effectif de 2 au minimum ;

la maximalité : la chaîne n'est pas strictement incluse dans une chaîne plus longue et d'effectif égal.

Pour expliciter ces caractéristiques, prenons comme exemple la chaîne « MISSISSIPPI ». Le tableau 4.6 ci-après présente les sous–chaînes répétées dans cette chaîne de caractères, leurs positions (offsets dans la chaîne) et leurs effectifs respectifs.

À partir de ces données, il est possible de vérifier la condition de maximalité. Cette opération est effectuée par ordre croissant de tailles et d'offsets. Les trois répétitions de longueur 1 (I, S et P) sont maximales, car non–incluses dans une chaîne répétée plus longue (c'est le cas de P) de même effectif (c'est le cas de I, S et P). Les deux répétitions de longueur 2 IS et SI ne sont pas maximales, car elles sont strictement incluses dans d'autres chaînes répétées de même effectif : respectivement ISS et SSI. ISS et SSI ne sont pas non plus maximales, car elles sont strictement incluses dans ISSI qui a le même effectif. ISSI est maximale car elle n'est pas incluse dans une chaîne répétée de taille supérieure. Les $rstr_{max}$ de MISSISSIPPI sont donc I, S, P et ISSI. Les $rstr_{max}$ possèdent

Sous-chaîne	Longueur	Offsets dans la chaîne	Effectif
I	1	1,4,7,10	4
IS	2	1-2,4-5	2
ISS	3	1-3,4-6	2
ISSI	4	1-4,4-7	2
S	1	2,3,5,6	4
SI	2	2-3,5-6	2
SSI	3	2-4,5-7	2
P	1	8,9	2

Tableau 4.6 – Sous-chaînes répétées de « Mississippi », offsets et effectifs

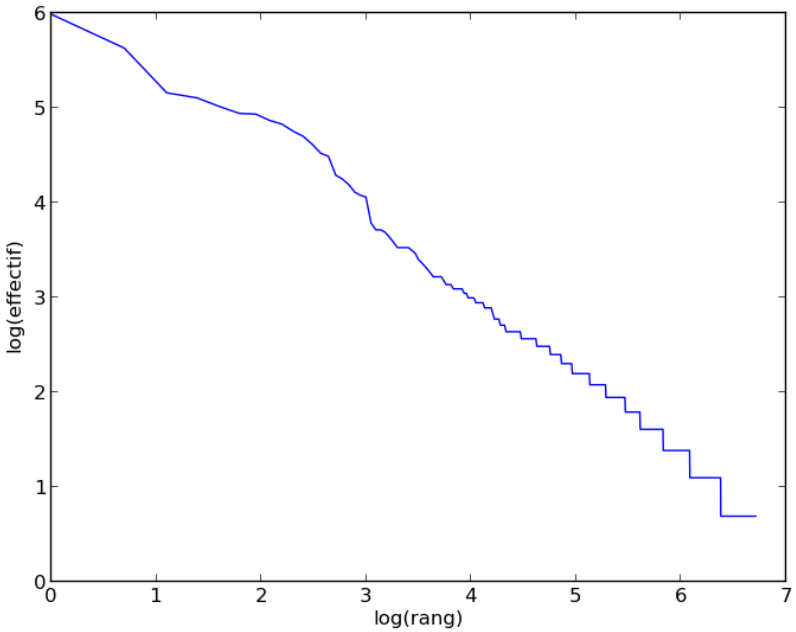


FIGURE 4.4 – Effectifs des $rstr_{max}$ en fonction de leur rang, analyse effectuée sur l'exemple présenté dans la figure 4.2

des caractéristiques intéressantes du point de vue de l'analyse. Ainsi, la fréquence des $rstr_{max}$ d'un texte est comparable à une loi de Zipf sur les mots, comme nous pouvons le voir sur la figure 4.4. La relation entre le caractère informatif d'un élément et sa position sur la courbe est supposée proche de celle qui existe pour les mots.

Au sein d'un énoncé, l'existence de ces répétitions n'est pas anodine. Elle l'est d'autant moins que ces répétitions sont repérables à des positions clés. La section suivante présente quelques applications de cette analyse en chaînes de caractères.

Longueur	Chaîne	Positions
35	es_autorités_sanitaires_se_veulent_	[1, 2]
29	_rassurantes_car_aucune_conta	[1, 2]
24	ion_d'homme_à_homme_n'a_	[1, 2]
23	_aucune_contamination_d	[2, 6]
21	_en_Arabie_saoudite._	[1, 2]
19	_dans_un_état_grave	[2, 3]
16	nt_été_identifié	[2, 6]
16	n_nouveau_virus_	[0, 2]
15	e_respiratoire_	[2, 3]
15	**_nouveau_virus_**	**[0, 2, 6]**
14	e_coronavirus_	[2, 4]
14	_été_identifié	[2, 6]
14	_de_la_santé_(	[3, 6]
13	e_coronavirus	[2, 4]
13	00_personnes_	[2, 4]

Tableau 4.7 – Longueurs et positions (en paragraphes) des 15 plus longues $rstr_{max}$ du texte présenté dans la figure 4.2

4.3.4 Applications de l'analyse au grain caractère

Nous montrons ici comment l'utilisation de la répétition de caractères permet de sélectionner des segments de textes utiles pour la veille épidémiologique. L'analyse au grain caractère est couplée à l'analyse positionnelle présentée dans la section 4.1.1. Un premier exemple d'analyse est proposé à l'aide du texte de la figure 4.2 (page 76). Dans le tableau 4.7 figurent les quinze plus longues $rstr_{max}$ de ce texte. Les positions présentées sont des n-uplets [45]. Les chiffres entre crochets représentent le numéro d'ordre du paragraphe, « 0 » représentant le titre. Par exemple, la chaîne « _nouveau_virus_ » longue de 15 caractères est présente dans le titre, le deuxième et le sixième paragraphe.

Le tableau 4.8 (page 83) représente la même mesure en ne gardant pour chaque position que la chaîne la plus longue. Nous constatons que certaines positions sont beaucoup plus productives : elles comptent des $rstr_{max}$ de plus grande taille. Le nombre de $rstr_{max}$ de taille supérieure à 10 caractères chute très vite lorsque l'on limite la sélection à une seule $rstr_{max}$ par position. Les $rstr_{max}$ ainsi sélectionnées fournissent des informations utiles sur l'évènement décrit dans le texte.

45. Ici il y a seulement des couples et des triplets. Certaines $rstr_{max}$ sont présentes à plus de positions mais aucune ne figure dans les 15 plus longues $rstr_{max}$.

Longueur	Chaîne	Positions
35	es_autorités_sanitaires_se_veulent_	[1, 2]
23	_aucune_contamination_d	[2, 6]
19	_dans_un_état_grave	[2, 3]
16	n_nouveau_virus_	[0, 2]
15	_nouveau_virus_	[0, 2,], [0, 6], [2, 6]
14	e_coronavirus_	[2, 4]
14	_de_la_santé_(	[3, 6]
13	_sont_décédés	[1, 6]
13	_aucune_conta	[1, 2], [1, 6], [2, 6]
12	_personnes_a	[4, 6]
11	_personnes_	[2, 4], [2, 6], [4, 6]
10	proche_du_	[0, 5]
9	e_du_Sras	[0, 2], [0, 4], [2, 4]
7	e.</p>\n	[1, 2], [1, 3], [2, 3]
7	_de_la_	[2, 3], [2, 6], [3, 6]

Tableau 4.8 – $rstr_{max}$ la plus longue pour chaque jeu de positions

4.3.5 Conclusion sur l'utilisation du grain caractère

Nous avons pu remarquer que l'utilisation du caractère était parfois considéré comme artificielle ou peu élégante en regard d'une segmentation en mots. Pourtant, la segmentation en mots n'est pas plus légitime qu'une autre segmentation. Sauf peut-être à considérer que les unités manipulées par la machine doivent nécessairement être interprétables par l'humain.

D'autres éléments de réflexion sur les problèmes posés par la segmentation en mots figurent dans deux thèses récentes soutenues à Grenoble ([Denoual-2006] et [Cromieres-2009]). Les auteurs montrent, respectivement en traduction automatique et en alignement, l'intérêt des méthodes en caractères et la plus-value offerte vis-à-vis des méthodes se basant sur les mots. Des travaux plus récents, menés au laboratoire GREYC de l'Université de Caen, ont exploré de façon approfondie l'intérêt du traitement au grain caractère pour l'alignement multilingue ([Brixtel-2011, Lecluze-2011, Lecluze-2013]). Ces recherches portaient sur des corpus multilingues de documents alignés, ou « multi-documents », diffusés par les institutions européennes. Là aussi, l'analyse au grain caractère s'est avérée tout à fait exploitable pour le traitement de la langue et spécialement pertinente pour la perspective multilingue.

Nous avons défendu l'idée que la segmentation en mots n'est pas une étape obligatoire pour le traitement des langues. Bien au contraire, elle est même inadaptée pour certaines langues. Une manière de surmonter cet obstacle, sans chercher des mots graphiques dans n'importe quelle langue, est de traiter des chaînes de caractères. Toute langue écrite est en effet analysable par ce biais. De plus, il n'est pas nécessaire de manipuler des

unités directement interprétables par l'humain ou correspondant à une certaine vision des langues pour obtenir un résultat de qualité. Il est alors possible d'utiliser des techniques d'**algorithmique du texte** au profit de l'extraction de connaissances à partir des textes en langue naturelle.

Synthèse

Dans notre approche, l'analyse de la place des mots dans la phrase est remplacée par l'analyse de la **position de chaînes de caractères** dans les textes. Les propriétés communicationnelles des textes doivent permettre une grande parcimonie dans la description linguistique en exploitant les stratégies économes utilisées par les destinataires des textes. En effet, dans les textes, **le global détermine le local**. Les propriétés textuelles permettent de guider efficacement un processus parcimonieux d'analyse. L'utilisation du niveau textuel comme niveau de décision permet de limiter la description des phénomènes locaux propres à chaque langue. L'utilisation du grain caractère comme grain d'analyse minimal permet de s'affranchir de l'encombrante notion de mot, gênante voire inutilisable dans de nombreuses langues.

Ces principes permettent de jeter les bases d'une approche réellement multilingue pour l'analyse de la presse. Cette approche est fondée sur un modèle de document dépendant des propriétés du genre, académique ou journalistique, plutôt que sur des propriétés locales des langues impliquées. Nous mettons ainsi en avant la dimension textuelle, globale par rapport à la dimension locale afin de favoriser des méthodes endogènes donc peu coûteuses. Ceci est favorisé par une approche différentielle, non descriptive, plus à même de permettre de manipuler des langues ayant des caractéristiques « locales » bien différentes.

Conclusion de la deuxième partie

Dans cette partie, nous avons présenté les caractéristiques de notre approche destinée à lever les verrous inhérents à l'analyse multilingue de la presse. Nous avons relié ces caractéristiques à l'importance de la factorisation pour des tâches requérant une grande généricité. Nous avons montré en quoi l'analyse séquentielle en vigueur dans de nombreux domaines du TAL trouve ses limites dès lors que pour la grande majorité des langues les modules et ressources nécessaires sont peu fiables voire inexistants. Le coût prohibitif de leur constitution et de leur mise à jour nous a mené à la recherche d'invariants. Nous avons identifié ces invariants au niveau textuel, en se fondant sur des propriétés communicationnelles et leurs implications positionnelles. Nous avons ainsi proposé de tenir compte spécifiquement des invariants spécifiques au genre journalistique. Nous avons mis en lumière l'importance de la répétition et de la position dans la transmission de l'information entre le journaliste et son lecteur. Ces principes nous ont permis de définir une méthode nécessitant suffisamment peu de description linguistique, pour permettre une couverture multilingue à coût raisonnable.

Troisième partie

Veille épidémiologique multilingue : évaluation et implantation

Introduction

Dans cette troisième et dernière partie, nous présentons en détail le corpus d'étude que nous avons constitué, les modalités d'évaluation que nous avons déterminé, ainsi que le système de veille que nous avons implanté. Puis nous montrons comment le traitement de nouveaux corpus est dépendant du genre textuel à traiter plutôt que des langues dans lesquelles sont rédigés les documents impliqués.

Dans le chapitre 5, nous présentons DAnIEL (Data Analysis for Information Extraction in any Language) : le système de veille que nous avons développé. Ce système se base sur les principes à vocation multilingue présentés dans les chapitres 3 et 4. De façon à pallier l'absence de corpus d'étude disponible dans le domaine, nous avons collecté un grand nombre de documents sur différents fils de presse du domaine de la "santé" et tout particulièrement sur ceux de *Google News*. Nous avons ainsi constitué un corpus comprenant des langues appartenant à différentes familles. Ce corpus a été annoté par des locuteurs de chaque langue afin de pouvoir comparer, dans une fenêtre de temps donné, les performances de notre système et les résultats de la veille humaine. L'analyse des résultats de la veille humaine nous a invité à réfléchir à des modalités d'évaluation à même de refléter les besoins de l'utilisateur final de notre veille. Les résultats de DAnIEL sur ce corpus sont comparés aux attendus en terme d'extraction d'évènements. Nous examinons dans quelle mesure DAnIEL est robuste à la variation en langue.

Dans le chapitre 6, nous présentons différentes expérimentations menées sur le genre « articles scientifiques ». Nous montrons que notre approche parcimonieuse fondée sur le genre est à même de traiter d'autres types de textes que les articles de journaux. Ceci renforce notre hypothèse : la variation en genre est plus importante que la variation en langue. Notre modèle est robuste à l'extension multilingue, dès lors que nous restons au sein d'un même genre textuel. Au contraire, il appelle des aménagements lorsque l'on passe à un autre genre textuel : le modèle de document doit être adapté.

Dans le chapitre 7, nous examinons comment notre approche fondée sur le modèle peut être sensible à la qualité des documents analysés. Nous nous interrogeons donc sur la problématique du nettoyage des pages Web ou détourage. Pour ce faire, nous proposons une évaluation de différents détoureurs disponibles en ligne. Cette évaluation est faite dans deux directions différentes : (I) une évaluation classique fondée sur la comparaison avec un jeu de données de référence, (II) une évaluation par la tâche en examinant l'influence du détourage sur un système qui se situe en aval de la chaîne de traitement.

Chapitre 5

DAnIEL : notre système de veille multilingue

Sommaire

Ce chapitre est consacré au système que nous avons bâti à partir des principes exposés dans les chapitre 3 et 4 : un système pensé pour une utilisation multilingue, qui se base sur les propriétés du genre de textes étudié pour assurer une factorisation maximale. Cette factorisation est possible par le recours au critère de genre. Dans notre approche la variable est le genre et non la langue.

Ce système est baptisé DAnIEL pour « *Data Analysis for Information Extraction in any Language* ». Il est articulée autour d'un noyau central qui recherche des chaînes de caractères répétées maximales à certaines positions, ce noyau étant commun à toutes les langues traitées. DAnIEL s'appuie sur la présence de chaînes de caractères à certaines positions dans le texte pour analyser les articles de presse. Par rapport aux approches classiques en TAL, DAnIEL utilise un grain d'analyse plus petit (les caractères et non les mots). A l'opposé, le grain de décision utilisé par DAnIEL est plus gros (le texte et non la somme des phrases qui le composent). L'utilisation de propriétés textuelles et l'absence de description grammaticale des langues permettent à DAnIEL d'être fortement **indépendant des langues traitées**.

Sa très large couverture (53 langues traitées à ce jour [46] dont 17 pour lesquelles il a été formellement évalué) lui permet de détecter la plupart des évènements dès le premier article publié à son sujet. DAnIEL est très peu influencé par la langue dans laquelle cet article est rédigé. Il est, par contre, spécialisé sur le genre « articles de presse » ; pour analyser d'autres genres textuels des adaptations du modèle de document utilisé sont nécessaires.

Son architecture complète est détaillée dans la Section 5.1. De façon à limiter la dépendance aux ressources externes, DAnIEL a été conçu pour fonctionner avec une base de noms du domaine et de termes géographiques de taille limitée, collectée automatiquement et aisée à mettre à jour. Ceci permet à DAnIEL de traiter un grand nombre de langues avec un coût minimal.

La première évaluation proposée est une comparaison des résultats extraits de DAnIEL avec les rapports émis par le système manuel de veille épidémiologique ProMED dans la section 5.2. Puis, DAnIEL est évalué sur un jeu de données de référence que nous avons constitué (Section 5.3).

5.1 Description de l'architecture générale de DAnIEL

Nous développons ici le fonctionnement de notre système de veille épidémiologique à visée multilingue, DAnIEL. Ce système est conçu pour être aussi indépendant de la langue que possible. Deux principes ont été appliqués dans la conception de DAnIEL pour remplir cet objectif : *I* la **parcimonie** dans l'utilisation des ressources externes et *II* la **factorisation** maximale des étapes de traitement.

46. Ce chiffre correspond à l'ensemble des langues pour lesquelles des documents sont diffusés sur *EMM*.

DAnIEL utilise une base de connaissances minimale (Section 5.1.1) ; sa chaîne de traitement centrale comprend quatre phases :

1. Segmentation de l'article (Section 5.1.2) ;
2. Extraction de motifs (Section 5.1.3) ;
3. Filtrage de ces motifs (Section 5.1.4) ;
4. Détection des paires maladie–lieu (Section 5.1.5).

5.1.1 Utilisation parcimonieuse des ressources en mémoire

DAnIEL utilise comme ressource une simple liste des principaux noms de maladies et de lieux, de l'ordre de 500 items au maximum pour chaque langue. Il y a donc une différence d'ordre de grandeur vis-à-vis des systèmes classiques qui se basent sur des milliers d'items ([Collier-2006, Steinberger-2008a]). Les noms de maladies sont les termes spécifiques du domaine qui permettent de trier les documents selon leur pertinence. DAnIEL offre un **filtrage des documents** en deux classes : pertinents ou non-pertinents pour la veille épidémiologique. Au sein de la classe des documents pertinents, une classe pour chaque maladie détectée est crée.

Pour les documents jugés pertinents, DAnIEL va localiser les faits décrits. La base de noms de lieux permet de situer ces faits. Ainsi, pour les documents pertinents DAnIEL aboutit à une description sous la forme d'une Paire Maladie-Lieu (PML).

Langues	DAnIEL	PULS	BioCaster
#Noms de maladies et termes déclencheurs	109	2400	1498
#Noms de lieux	386	1500	4015

Tableau 5.1 – Nombre moyen de termes utilisés par langue par DAnIEL et deux systèmes de l'état de l'art

Les connaissances utilisées par DAnIEL sont extraites automatiquement sur *Wikipedia* avec une brève vérification manuelle. Ceci favorise l'extension rapide du système vers de nouvelles langues (Tableau 5.1). Nous estimons que traiter une nouvelle langue avec DAnIEL occasionne un coût inférieur à une demi–journée–homme[47]. La structure des deux bases de connaissance est extrêmement simple : il s'agit de listes de termes avec, si possible, le terme équivalent en anglais. Le terme anglais est utilisé pour favoriser le regroupement des alertes par PML. Cela permet aussi la comparaison des évènements extraits par DAnIEL avec ceux extraits par d'autres systèmes.

L'économie de moyens est à la base de la conception de DAnIEL, elle est permise par l'aspect textuel de l'analyse. En effet, les noms « scientifiques » des maladies ne sont pas les plus fréquents à l'échelle du texte. Les journalistes utilisent dans leurs articles des

47. La majorité de ce temps est consacrée au recensement des sources à traiter.

termes du langage commun de manière à s'assurer que l'information est bien décodée par le lecteur. Au contraire, au grain phrase il serait nécessaire de connaître un grand nombre de termes pour pouvoir décoder l'information.

La localisation des faits décrits est réalisée à l'échelle du pays. Une localisation plus précise serait très coûteuse à obtenir, si l'on veut la proposer dans toutes les langues. Cela impliquerait la collecte et le stockage d'un très grand nombre de noms de lieux pour chaque langue à traiter. Disposer par exemple d'une base aussi complète que la base *GeoNames*[48] pour 42 langues ne semble pas réaliste à ce jour. Les moyens nécessaires pour parvenir à un tel résultat seraient en effet colossaux.

Nous remarquons d'ailleurs les difficultés rencontrées dans cette tâche précise par des systèmes monolingues ([Keller-2009]). Notre choix est également justifié par le fait que le degré de précision recherché par les principaux systèmes de veille épidémiologique reste le pays. Localiser les épidémies à ce grain facilite la comparaison.

Ces deux listes, noms de maladies et termes géographiques, sont les seuls éléments dépendants de la langue dont a besoin DAnIEL pour fonctionner. DAnIEL nécessite donc une faible description lexicale. C'est cet aspect parcimonieux qui le rend compatible avec une couverture massivement multilingue[49].

5.1.2 Segmentation des articles

L'analyse effectuée par DAnIEL se base sur un modèle de document fondé sur le style collectif des journalistes. Ce modèle a pour but de déterminer quelles **positions** sont pertinentes dans le texte.

Nous avions dans un premier temps utilisé un seul modèle de document exploitantune dichotomie entre la tête de l'article et le reste du document ([Lejeune-2010a, Lejeune-2010b]). La tête du document est composée idéalement du titre et du chapeau. Ce modèle s'est avéré peu souple. En effet, ignorer la structure particulière des petits articles impactait négativement le rappel obtenu avec ce modèle.

Dans les petits articles, le chapeau a une moins grande importance. A l'opposé, dans des articles très longs c'était la précision qui souffrait du manque de souplesse de notre modèle initial. Parfois, des articles relatant des faits secondaires pouvaient être considérés comme pertinents ; le corps de texte étant disproportionné par rapport à la tête. Les positions pertinentes doivent être déterminées selon la taille des documents impliqués.

Au sein même du genre « articles de presse », il existe différentes catégories selon le degré de description des faits relatés. L'indice que nous utilisons pour différencier ces catégories est la longueur des textes en paragraphes. Les mécanismes de répétition et de position sont légèrement différents selon la longueur des articles concernés, la longueur des

48. Disponible sur `http://www.geonames.org/`, la base comporte 10 millions d'entrées mais seulement en anglais

49. Une couverture massivement multilingue signifie pour nous pouvoir couvrir toutes les langues pour lesquelles des articles de presse en ligne sont disponibles.

articles est reliée à une différence de « fonction » dans la transmission des informations.

Nos observations nous amènent à considérer trois modèles d'articles :

petite taille : dépêches (exposés factuels)

taille moyenne : articles de facture classique (évolution d'un évènement)

grande taille : articles d'analyse ou de synthèse (évènements bénéficiant d'un certain recul)

Dans les petits articles se trouvent une plus grande proportion de faits nouveaux. Les dépêches contiennent souvent des informations nouvellement disponibles. Celles-ci seront plus tard développées dans les articles de facture classique (taille moyenne). Ces formats d'articles permettent d'assurer un suivi des évènements,et de développer des informations secondaires. Les articles de synthèse se placent à un stade de description encore différent. Ils sont généralement conçus pour mettre en perspective différentes composantes d'un évènement pour lequel on dispose déjà d'un certain recul.

Cette tripartition du modèle de document permet d'améliorer les résultats de manière sensible. En effet le thème et le rhème sont proportionnels ([Lucas-2005]). La taille dévolue au rhème est dépendante de celle dévolue au thème. Localiser le thème implique donc une segmentation adaptée. Comparer simplement la tête (titre + chapeau) et le reste du texte (tronc et pied) ne convient pas pour tous les articles. Pour les articles de petite taille, cette tripartition améliore le rappel (on donne la même importance à tous les paragraphes). Le bruit dû aux articles de grande taille est limitée, ne sont comparées que la tête et le pied (Tableau 5.2).

Type d'article	# paragraphes	Segments comparés
Court	3 et moins	Tous les paragraphes
Moyen	4 à 10	Tête VS Tronc+pied
Long	11 et plus	Tête VS Pied

Tableau 5.2 – Segmentation des articles en fonction de leur taille

Les positions remarquables sont déterminées selon la taille des documents à traiter. À partir de ces positions, DAnIEL va extraire des éléments sous la forme de chaînes de caractères mots ou non-mots.

5.1.3 Extraction des motifs

Les motifs que nous utilisons sont des $rstr_{max}$ tels que définies dans la Section 4.3.3. Nous rappelons simplement ici les deux caractéristiques de ces motifs :

répétition : les motifs ont deux occurrences au moins ;

maximalité : les motifs ne peuvent être étendus sans perdre une occurrence.

Le nombre de motifs figurant dans un texte est inférieur à la longueur en caractères de ce texte. Ces motifs sont détectés en temps linéaire, $O(n)$ avec n la taille du texte en caractères, en utilisant les tableaux de suffixes augmentés décrits par Kärkkäinen ([Karkkainen-2006]). Une description plus complète de ces motifs figure dans la section 4.3.2 (page 79). Pour le calcul des motifs, nous utilisons une implantation en *Python* développée principalement par Romain Brixtel. Cette implantation est librement téléchargeable [50].

La recherche de motifs est effectuée de manière strictement identique dans toutes les langues, elle constitue le cœur de l'algorithme *relevant_ content* utilisé par DAnIEL (Algorithme 1).

D'un point de vue strictement opératoire, le filtrage par la segmentation et le filtrage par la connaissance s'effectuent en même temps. Toutefois, par souci de clarté et afin de mieux visualiser l'impact de chacun, nous décrivons séparément ces deux sous-tâches.

Data: Texte à analyser et base de noms de maladies
Result: Diagnostic du document : Pertinent/Non-pertinent
begin
  $Positions\_importantes \leftarrow SegmenterTexte$;
  `/* Calcul des positions remarquables dans le texte */`
  $Contenu\_pertinent \leftarrow RstrPositions\_remarquables\&Basedemaladies$;
  `/* Recherche des motifs (`m`) du texte communs à une position remarquable et à une entrée (`e`) de la base de noms de maladies */`
  for *sous-chaîne* m *d'une maladie* e **do**
    if $\frac{len(m)}{len(e)} > \theta$ **then**
      $Diagnostic \leftarrow pertinent$;
    else
      $Diagnostic \leftarrow non - pertinent$;
    end
  end
end

Algorithme 1: L'algorithme *relevant_ content* : méthode pour déterminer la pertinence d'un article pour la veille épidémiologique

Exemples d'application de *relevant_ content*

À titre d'exemple, nous présentons tout d'abord deux documents en anglais. Le premier a pour titre « CDC Urges Travelers to Israel to Protect Themselves from Measles [51] » et le second « Democracy at a discount [52] ». Une simple lecture en diagonale pourrait suffire à un locuteur pour juger de la pertinence de ces deux documents pour une activité de veille.

50. `http://code.google.com/p/py-rstr-max`
51. `http://www.cdc.gov/media/pressrel/2008/r080414a.htm`, consulté le 12 octobre 2013
52. `http://euobserver.com/7/114308`, consulté le 12 octobre 2013

Le premier (Document 1) contient bien un évènement épidémiologique : Rougeole–Israël (*measles–Israel*). Le second (Document 2) ne mentionne un nom de maladie qu'à titre métaphorique (choisir entre la peste et le choléra).

L'examen des $rstr_{max}$ présentes à des positions remarquables des textes permet de déterminer la pertinence. Le Tableau 5.3 contient les 10 plus longs motifs répétés à des positions clés dans chacun de ces deux articles. Ces motifs sont triés par longueur en caractères décroissante. Les chaînes de caractères détectées nous semblent pertinentes pour caractériser rapidement ces documents. Nous observons que les noms de maladies utilisés comme métaphore dans le document 2 ne sont pas répétés. La simple répétition offre ici un diagnostic efficace pour déterminer la classe des deux documents. Nous pourrions multiplier les exemples ; notons simplement que la répétition permet d'éliminer un grand nombre de faux positifs potentiels.

Document 1 (pertinent)	Document 2 (non-pertinent)
_to_Israel_for_Passover_	the_ eurozone_crisis_is_
_Americans_travel_	al_institutions_
ses_of_measles_	_the_European
_to_Israel_for_	_the_eurozone
_s_of_measles	c_institutions_
_travelers	the_european_
_to_Israel_	_the grecs
_Travelers	_political_
Passover,	_democracy
Jerusalem	commission

Tableau 5.3 – Les dix plus longs motifs d'un document pertinent (Document 1) et d'un non-pertinent (Document 2). "_" représente un espace typographique.

Le tableau 5.4 (page 98) présente cette fois l'application de *relevant_ content* sur un document pertinent en polonais et montre l'influence du filtrage par zones. Ce document fait partie de ceux qui nous ont permis d'éprouver le système. Selon nos annotateurs polonais ce document est pertinent pour la veille épidémiologique. La PML concernée est $denga - Tajlandia$ (dengue–Thaïlande). Des sous-chaînes du nom de maladie concerné (« denga ») sont répétées à des positions clés. Le filtrage positionnel couplé au filtrage lexical permet de détecter la sous-chaîne « deng ». C'est le segment commun aux différentes formes de « denga » rencontrées dans le texte.

Le tableau 5.5 (page 98) présente les différentes formes que peut prendre le terme en polonais. Nous remarquons que la chaîne extraite par DAnIEL correspond à la racine la plus courante du terme [53].

53. Dans ce cas spécifique la racine **den**-dz- ne poserait pas de problèmes : le vocatif est inusité en polonais tandis que le datif définit le complément d'objet indirect (probablement peu pertinent dans le cas de la veille épidémiologique).

Sans filtrage	Filtrage positionnel	Filtrage positionnel et lexical
1 _czarnych_legginsów,_	1 _czarnych_legginsów,_	1 deng
2 _wiceminister	2 _w_Tajlandii_	2 iła
3 _w_Tajlandii_	3 _tym_roku_	3 ła
4 _komarów.	4 _43_osoby	4 ró
5 _tym_roku_	5 _przenosz	5 ra
...	...	...
39 _deng	31 _deng	N/A

Tableau 5.4 – Motifs les plus longs détectés dans un document pertinent en polonais selon le type de filtrage appliqué

Cas	Nominatif	Génitif	Datif	Accusatif	Instrumental	Locatif	Vocatif
Forme	**deng**-a	**deng**-i	**den**-dz-e	**deng**-ę	**deng**-ą	**den**-dz-e	**deng**-o

Tableau 5.5 – La déclinaison de « denga » dans les différents cas du polonais

Ceci montre comment l'analyse au grain caractère permet de se passer de description morphologique, ce qui est particulièrement intéressant pour les langues à morphologie riche. Bien entendu, cela ne signifie pas l'absence d'erreur dans la reconnaissance des termes par DAnIEL. Certains phénomènes morphologiques tels que la métathèse[54] pourraient amener des erreurs de détection. Ces erreurs sont toutefois extrêmement rares. Développer une analyse morphologique plus poussée est trop coûteux par rapport au bénéfice escompté.

La combinaison de critères positionnels et d'analyse au grain caractère nous permet donc d'éviter une coûteuse description : I des règles morphologiques de la langue et II de la structure syntaxique des phrases de la langue.

5.1.4 Filtrage des motifs

Il est important de pouvoir filtrer le grand nombre de motifs extraits. En effet de nombreux motifs ne sont pas pertinents pour établir le diagnostic du document. La position du motif dans le document étudié permet de limiter l'extraction de faux positifs. C'est ici qu'intervient la segmentation décrite plus haut, le nombre de motifs est réduit avec un silence minimal. Des motifs obtenus, DAnIEL ne conserve que ceux qui sont une sous-chaîne d'un terme de sa base de connaissance.

54. La métathèse est la permutation des sons dans la chaînee parlée. Un exemple prototypique est le latin « formaticum » qui a donné « fromage » en français. Une trace savoureuse de cette origine est encore présente dans la langue sous la forme de la Fourme d'Ambert. La métathèse se manifeste donc à l'écrit par des permutations de lettres qui gêne l'extraction de racines uniques.

$\frac{len(m)}{len(e)}$	sous-chaîne	terme polonais	traduction	position
0.8	deng	denga	dengue	[1, 4]
0.75	iła	kiła	syphilis	[1, 3]
0.5	ła	kiła	syphilis	[1, 2, 3, 5]
0.5	ró	róża	érysipèle	[2, 5]
0.5	ra	odra	rougeole	[2, 3, 5]
0.5	os	ospa	variole	[1, 2, 4]
0.5	od	odra	rougeole	[1, 2, 4, 5]
0.5	ister	listerioza	listeria	[2, 3]
0.5	dr	odra	rougeole	[0, 2]
0.44	rowi	norowirus	norovirus	[0, 2]
0.4	że	tężec	tetanos	[2, 5]
0.4	ry	grypa	grippe	[1, 5]
0.4	rowi	astrowirus	astrovirus	[0, 2]
0.4	ol	ebola	ebola	[1, 2]
0.4	ng	denga	dengue	[1, 4]

Tableau 5.6 – Motifs extraits après filtrage positionnel et lexical classés par taille décroissante

Filtrage positionnel

Les motifs n'apparaissant pas dans deux segments différents, selon la segmentation définie dans le Tableau 5.2 (page 95), sont écartés. Ce filtrage permet d'écarter 85% des motifs présents, une vérification manuelle sur quelques documents a montré que 99% des motifs écartés n'étaient pas pertinents pour notre tâche.

Filtrage lexical

Les motifs restants sont comparés à la liste de maladies pour déterminer si le texte contient un évènement. Nous avons déterminé empiriquement qu'à partir de $\frac{4}{5}$ de caractères consécutifs en commun l'extraction était fiable. De façon plus formelle l'algorithme est énoncé comme suit : pour un motif m et une entrée e dans la liste de maladies $\frac{len(m)}{len(e)} > \frac{4}{5}$ avec *len* la longueur en caractères de m et e.

Nous nommons ce seuil θ, son influence précise sur l'extraction est examinée dans la Section 5.3.4.

5.1.5 Localisation de l'évènement

La localisation de l'évènement détecté est elle aussi basée sur les principes du genre journalistique. La première hypothèse est que le lieu de l'évènement est répété et figure à des positions remarquables : il y a **localisation explicite**. L'algorithme *relevant_ content*

est à nouveau exploité, la liste de noms de maladies est simplement remplacée par la liste de noms de lieux (voir algorithme 1 page 96).

Il existe de nombreux cas où aucun lieu n'est détecté avec la localisation explicite. Nous formulons donc une seconde hypothèse, la **localisation implicite** : si le journaliste ne mentionne pas explicitement un lieu c'est que le lieu concerné correspond à la source de l'article. Si aucun nom de pays n'est répété dans l'article alors la localisation de la source et celle de l'évènement sont jugées équivalentes. Par exemple, nous considérons que par défaut un évènement relaté dans *le Figaro* a lieu en France. Dans le cas contraire, la localisation serait explicitement mentionnée et répétée. L'efficacité de cette règle est évaluée dans la Section 5.3.7.

5.1.6 Exemples de sortie du système

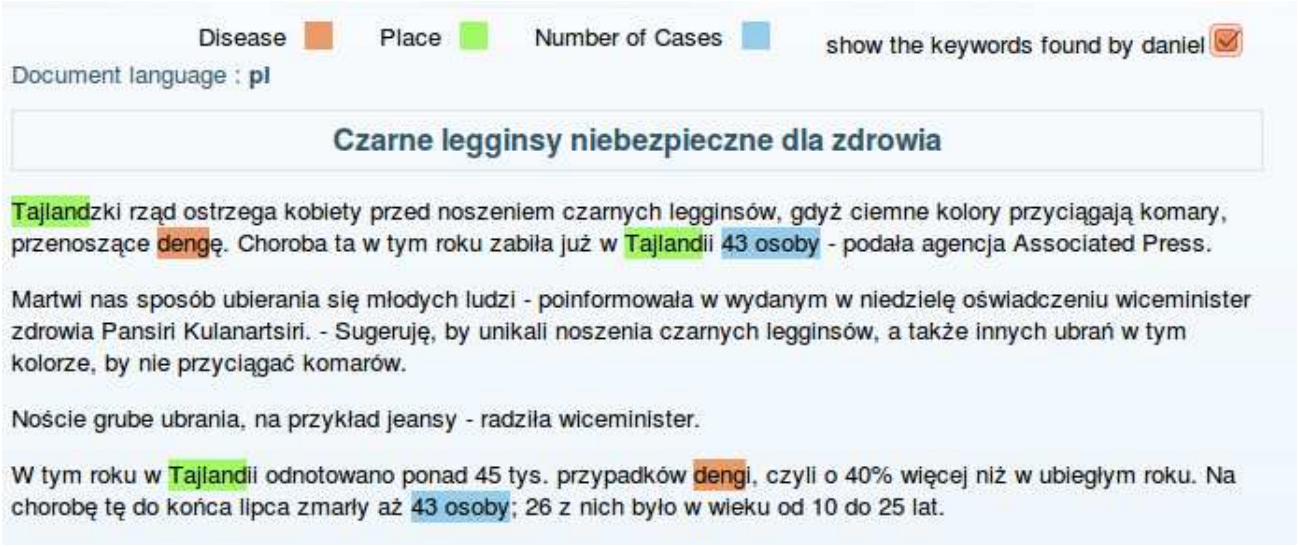
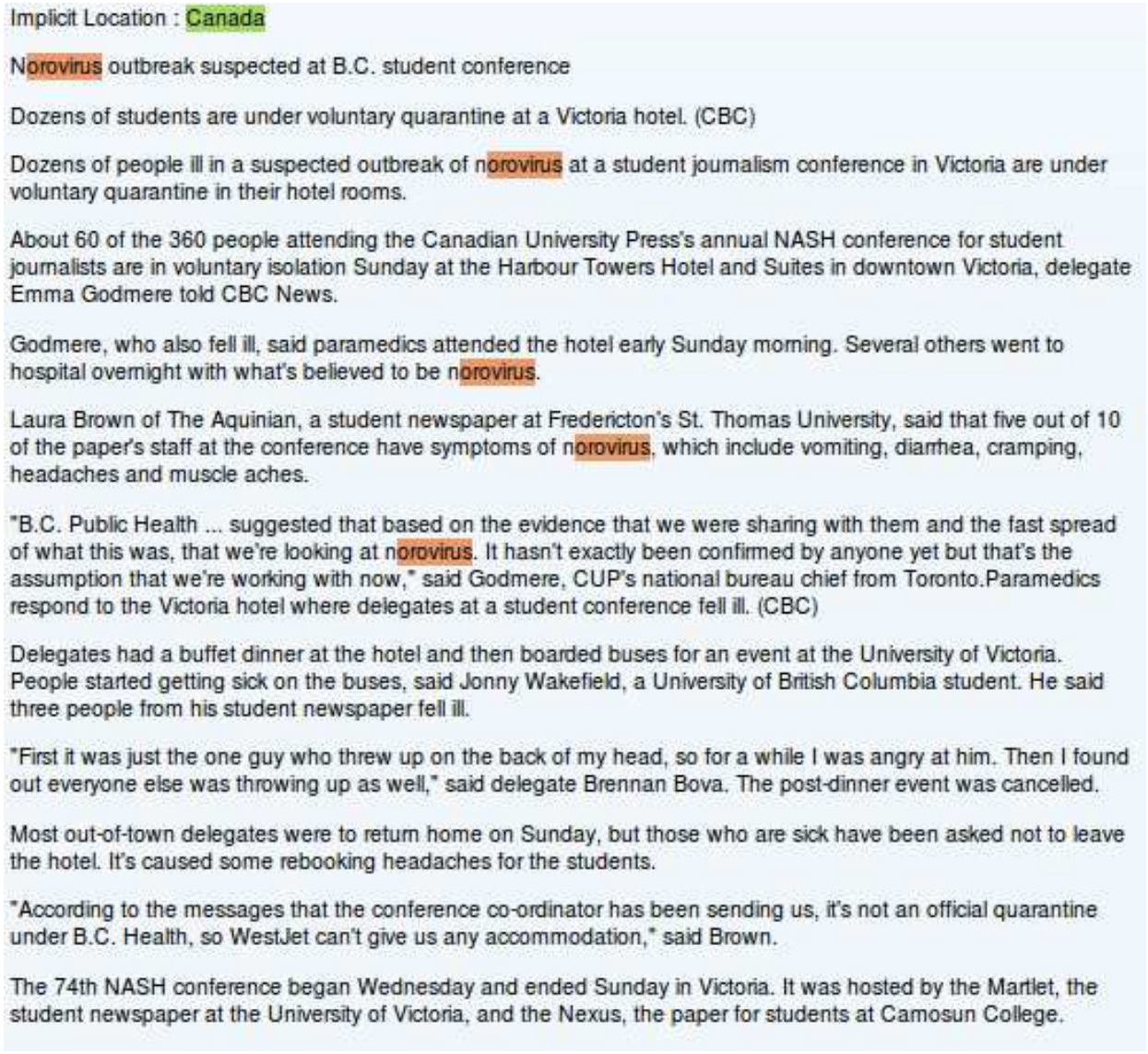


FIGURE 5.1 – Exemple d'extraction automatique de PML par DAnIEL sur un article en polonais.

Le phénomène de répétition utilisé par les journalistes pour transmettre l'essence du message est bien visible dans la Figure 5.1. La PML est présente dans le premier et le troisième paragraphe. La racine « deng∼ » est répétée avec deux flexions différentes et parfois rappelée par l'hyperonyme « chorob∼ » (maladie). La chaîne « Tajland∼ » correspondant au nom de pays est elle aussi repérée dans le chapeau et dans le corps.

Ici l'évènement est énoncé (quelle maladie, dans quel pays) puis développé. C'est une manifestation très claire de la structuration thème/rhème des articles d'actualité.

La figure 5.2 (ci-contre) présente un document en anglais dans lequel DAnIEL a détecté la paire norovirus-Canada. La chaîne « *orovirus* » a été détectée à des positions remarquables du texte : titre, chapeau et corps.


Implicit Location : Canada

Norovirus outbreak suspected at B.C. student conference

Dozens of students are under voluntary quarantine at a Victoria hotel. (CBC)

Dozens of people ill in a suspected outbreak of norovirus at a student journalism conference in Victoria are under voluntary quarantine in their hotel rooms.

About 60 of the 360 people attending the Canadian University Press's annual NASH conference for student journalists are in voluntary isolation Sunday at the Harbour Towers Hotel and Suites in downtown Victoria, delegate Emma Godmere told CBC News.

Godmere, who also fell ill, said paramedics attended the hotel early Sunday morning. Several others went to hospital overnight with what's believed to be norovirus.

Laura Brown of The Aquinian, a student newspaper at Fredericton's St. Thomas University, said that five out of 10 of the paper's staff at the conference have symptoms of norovirus, which include vomiting, diarrhea, cramping, headaches and muscle aches.

"B.C. Public Health ... suggested that based on the evidence that we were sharing with them and the fast spread of what this was, that we're looking at norovirus. It hasn't exactly been confirmed by anyone yet but that's the assumption that we're working with now," said Godmere, CUP's national bureau chief from Toronto.Paramedics respond to the Victoria hotel where delegates at a student conference fell ill. (CBC)

Delegates had a buffet dinner at the hotel and then boarded buses for an event at the University of Victoria. People started getting sick on the buses, said Jonny Wakefield, a University of British Columbia student. He said three people from his student newspaper fell ill.

"First it was just the one guy who threw up on the back of my head, so for a while I was angry at him. Then I found out everyone else was throwing up as well," said delegate Brennan Bova. The post-dinner event was cancelled.

Most out-of-town delegates were to return home on Sunday, but those who are sick have been asked not to leave the hotel. It's caused some rebooking headaches for the students.

"According to the messages that the conference co-ordinator has been sending us, it's not an official quarantine under B.C. Health, so WestJet can't give us any accommodation," said Brown.

The 74th NASH conference began Wednesday and ended Sunday in Victoria. It was hosted by the Martlet, the student newspaper at the University of Victoria, and the Nexus, the paper for students at Camosun College.

FIGURE 5.2 – Exemple d'extraction automatique de PML par DAnIEL sur un article en anglais avec application de la règle de « localisation implicite ».


Implicit Location : 中国

流感又增4死 皆未接種流感疫苗 春節注意防疫

春節在即，大家都忙著採買年貨、送禮，沒法顧及「不要到人多的公共場所」的流行病防護建言，眼見衛生署預估的流感最高峰即將來到，衛署疾病管制局今（17）日再度發布新增4例流感死亡病例，全都是B型病毒，而且4人都未接種流感疫苗。

疾管局流感疫情監測資料顯示，上周新增145例流感併發重症，較上上周的147例略少兩例。本流感季自100年7月1日起，迄今累計共587例流感併發症確定病例，和上個流感季同期的851例相較，併發症病例少3成1。本流行季共有28例死亡，上個流感季同期是85例，流感併發症發生率為每百萬人口25，死亡率為1.2，預估於農曆年前後達到最高峰。

疾管局副局長周志浩表示，新增的4名流感死亡病例，分別是2歲女童、81歲女性及64歲男性，以及台北市48歲男性，其中3例成人都有高血壓慢性病史，另64歲男性有慢性阻塞性肺病，而女童則有潛在之腦部疾病。4例均未接種流感疫苗，皆感B型病毒，發病後分別因心肺衰竭、敗血性休克及細菌感染肺炎等因素死亡。

農曆春節期間返鄉及旅遊人潮眾多，為避免流感病毒傳播，周志浩提醒民眾要注意「咳嗽禮節」，如有呼吸道症務必戴上口罩。為降低感染流感後併發嚴重併發症機率，家中若有發生流感併發症的高危險族群，如嬰幼兒、年長者及具慢性疾病患者，於感染流感後應提高警覺。

必須特別注意是否出現如呼吸短促、呼吸困難、發紺、血痰或濃痰、胸痛、意識改變、低血壓或高燒持續72小時等危險症狀，應即刻前往大醫院治療，掌握黃金治療時間。春節期間流感防疫也不打烊，疾管局已請各縣市協調約210家急診服務合約醫院，於春節期間開診持續提供公費藥劑予民眾使用，民眾如不知可到何處求治，可洽所在地衛生局，或撥打24小時免費防疫諮詢專線「1922」詢問。

圖說:為避免流感病毒傳播，疾管局提醒民眾要注意「咳嗽禮節」，如有呼吸道症務必戴上口罩。（記者盧宥君攝）

FIGURE 5.3 – Exemple d'extraction automatique de PML par DAnIEL sur un article en chinois avec application de la règle de localisation implicite.

Nous pouvons tout d'abord remarquer que ce n'est pas la chaîne *norovirus* qui a été détectée. Ceci est dû au fait que le texte est analysé tel quel, sans changement de la casse. Or, l'occurrence figurant dans le titre (position remarquable) est capitalisée. DAnIEL a ensuite utilisé la règle de « localisation implicite ». En effet, les données géographiques répétées dans le texte (*Victoria et Columbia*) sont inconnues de DAnIEL. Le système a donc situé l'évènement décrit au Canada du fait de la source : `www.cbc.ca`[55].

La figure 5.3 montre l'exploitation de la même règle en chinois. Cet exemple illustre toutefois un manque dans la granularité que nous utilisons pour la localisation, pour des pays très grands (Russie, Canada, États-Unis...). Un grain plus fin serait sans doute plus parlant. Ce grain pourrait être une subdivision administrative telle que l'« état » ou la « province ».

5.1.7 Synthèse sur le fonctionnement du système

DAnIEL permet donc une analyse multilingue parcimonieuse des articles de presse grâce à son utilisation du grain caractère comme grain d'analyse et du grain texte comme grain de décision. Le système est rapide ([Lejeune-2013b]) puisqu'il permet de traiter 2000

55. D'autres exemples figurent en ligne sur le site du projet : `https://daniel.greyc.fr`

documents en moins de 15 secondes [56]. À puissance de calcul égal, c'est le temps utilisé dans un système classique pour la simple phase d'étiquetage morpho-syntaxique (voir section 2.2.2).

La vitesse de DAnIEL est donc compatible avec la surveillance en temps réel de même qu'avec l'exploration d'archives de presse. DAnIEL propose un filtrage des documents en deux classes : pertinents ou non–pertinents pour la veille épidémiologique. Pour les documents classés dans la première catégorie, DAnIEL propose un étiquetage plus fin sous la forme de paires maladie-lieu. Extraire une paire X–Y signifie pour le système l'existence d'une épidémie de la maladie X dans le pays Y.

À partir de ces extractions, nous souhaitons maintenant confronter le système à des jeux de données de référence. Tout d'abord, nous comparons la réactivité de DAnIEL dans la détection d'évènements épidémiologiques avec celle du système manuel de référence ProMED (Section 5.2).

5.2 Détection accélérée de faits épidémiologiques grâce à DAnIEL

Dans cette section nous présentons les résultats de DAnIEL sur un corpus multilingue collecté sur une fenêtre de temps assez grande. L'objectif est de voir dans quelle mesure DAnIEL apporte une plus-value à la détection manuelle ; comment la couverture multilingue apporte des bénéfices en terme de vitesse de détection. Cette comparaison ne permet pas d'obtenir des résultats très fins en terme de qualité. Il est en effet difficile de savoir précisément dans quelle mesure nos hypothèses sur la structuration des documents sont valides. Au vu de la taille du corpus, il serait difficile (ou au moins très coûteux) de savoir ce que DAnIEL a pu « manquer ». *A priori*, ProMED devrait recenser, même si c'est avec retard, l'intégralité des épidémies survenues.

Les documents qui constituent notre corpus ont été collectés entre le premier octobre 2011 et le 31 janvier 2012. Les rapports émis par ProMED dans la période constituent la référence, ce jeu de données est décrit dans la Section 5.2.1. L'analyse de DAnIEL est évaluée sur un corpus de référence que nous avons constitué et que nous présentons dans la Section 5.2.2. Les résultats des deux systèmes sont confrontés dans la Section 5.2.3.

5.2.1 Jeu de données issu des rapports ProMED

Nous avons collecté les rapports diffusés sur ProMED-mail 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, russe). Quelques comptes–rendus en thaï et en vietnamien ont également été produits.

56. Sur un ordinateur portable avec un processeur 2 cœurs cadencé à 2,4 Ghz et 2 Go de ménoire vive.

	anglais	français	portugais	russe	espagnol	thaï	vietnamien
# rapports	819	148	129	127	220	25	78
# Novembre 2011	285	3	26	49	68	25	78
# Décembre 2011	291	33	15	28	78	0	0
# Janvier 2012	193	62	48	37	37	0	0
# Février 2012	54	50	40	33	38	0	0

Tableau 5.7 – Répartition des rapports ProMED pour chaque langue et chaque mois de la période d'étude

Chacun des rapports collecté expose des faits relatifs à un ou plusieurs évènements épidémiologiques, sous la forme d'un triplet maladie–lieu–date. Pour cette expérience nous nous intéressons spécifiquement au **premier signalement**. Nous disposons par ailleurs d'informations sur la source qui a permis l'émission du rapport. Ceci nous permet d'établir la langue de la source qui a permis la détection de l'évènement.

Le corpus ainsi créé est présenté dans le tableau 5.7 (page 104). Nous pouvons remarquer que 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 monolingue. 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 5.9 (page 105). Notons toutefois que 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 un plus grande variété de maladies et de lieux (tableau 5.8). Toutefois 40% des évènements décrits se concentrent simplement sur trois pays : États-Unis, Australie et Royaume-Uni).

	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
# paire maladie-lieu	366	63	40	55	46	12	26

Tableau 5.8 – Détails sur les rapports ProMED : répartition par maladie, lieux et 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 une répartition par langue similaire à celle présentée dans le tableau 5.8. Seuls deux de ces rapports provenaient d'une source rédigée dans une langue autre que celles

traitées par ProMED. Pour faciliter la comparaison, nous avons donc considéré que la langue du rapport ProMED correspondait à la langue de la source qui a permis l'émission du rapport.

5.2.2 Corpus d'articles de presse utilisé par DAnIEL

Nous avons principalement constitué ce corpus à partir de la catégorie santé de *Google News*. Le choix de se limiter à cette catégorie a été fait pour des raisons d'annotation développées dans la section 5.3. Des documents ont été collectés pour toutes les langues qui disposaient de cette catégorie : arabe (ar), tchèque (cs), anglais (en), français (fr), allemand (de), italien (it), norvégien (no), portugais (pt), russe (ru), espagnol (es), suédois (sv), turc (tr), et chinois (zh) Nous avons également collecté des documents à partir de fils RSS relatifs à la santé et des catégories santé de grands journaux nationaux pour le finnois (fi), le grec (el) et le polonais (pl).

Ce 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 5.9. Remarquons que 40% de ces documents sont rédigés dans des langues non couvertes par ProMED.

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

Tableau 5.9 – Nombre d'articles par langue et par mois

Les documents composant ce corpus sont en html *brut*. Nous avons donc utilisé un outil interne de nettoyage de contenu afin d'extraire le contenu textuel. À partir de ce corpus, DAnIEL a extrait 1571 signalements relatifs à la veille épidémiologique. Ces signalements

Langues	# A	# S	$100 * S/A$	Oct. 2011	Nov. 2011	Déc. 2011	Jan. 2012
ar	3093	30	0,97%	3	5	12	10
cs	208	15	7,21%	2	7	3	3
de	2509	63	2,51%	7	13	24	19
el	1380	83	6,01%	17	25	18	23
en	4742	285	6,01%	63	75	67	80
es	4389	230	5,24%	42	62	71	55
fi	132	7	5,3%	2	0	3	2
fr	2132	142	6,66%	17	50	48	27
it	703	54	7,68%	12	19	15	8
nl	876	24	2,74%	2	4	12	6
no	311	11	3,53%	0	4	4	3
pl	801	140	17,47%	15	37	36	52
pt	1362	92	6,75%	30	22	25	15
ru	1896	296	15,61%	49	84	54	109
sv	196	26	13,26%	2	10	9	5
tr	239	0	0	0	0	0	0
zh	1122	73	6,51%	12	25	14	22

Tableau 5.10 – 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.

sont décrits dans les tableaux 5.10 et 5.11 (page 107). Parmi les 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 grand 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. Nous pouvons compléter en rappelant la variabilité en contenu des fils santé selon les langues (section 2.2.2 page 39). 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. Nous en déduisons que le choix de collecter des documents sur la catégorie santé de*Google News* n'était pas pertinent pour le turc : le filtre est trop strict.

Dans le tableau 5.11 (page 107), nous présentons le nombre de maladies et de lieux différents concernés par les signalements émis par DAnIEL. Nous remarquons la grande variété en nombre de lieux offerte par les langues de grande diffusion, ceci était un résultat attendu : les langues de grande diffusion reportent 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

Langues	#Rapports	#Maladies	#Lieux	#Paires maladie–lieu
ar	30	7	3	12
cs	15	6	2	9
de	63	12	19	32
el	83	13	7	25
en	285	33	55	161
es	230	29	35	115
fi	7	6	2	4
fr	142	32	39	85
it	54	22	9	28
nl	24	9	7	11
no	11	6	1	6
pl	140	19	45	83
pt	92	23	14	50
ru	296	21	70	141
sv	26	7	2	10
tr	0	0	0	0
zh	73	16	6	23

Tableau 5.11 – Nombre de maladies, de lieux et de paires maladie–lieu impliqués dans les signalements produits par DAnIEL

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. C'est l'objet de la Section 5.2.3 dans laquelle nous proposons une évaluation de la plus–value offerte par DAnIEL.

5.2.3 Évaluation de la plus-value offerte par DAnIEL

Nous étudions dans cette section le bénéfice que peut offrir un système de veille épidémiologique automatique tel que DAnIEL. Nous évaluons dans quelle mesure la couverture accrue en nombre de langues offre réellement des bénéfices dans l'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 [57].

Parmi les PML extraites, 167 ont été extraites par les deux systèmes. Afin de mesurer les différences en terme de délai de détection nous avons pour chaque PML comparé la

57. 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 DAnIEL.

Paire		ProMED		DAnIEL		Décalage
Maladie	Lieu	Lg.	Date	Lg.	Date	(jours)
Choléra	Zimbabwe	**en**	2011-12-18	**en**	2012-01-30	43
Pneumonie	États-Unis	**en**	2011-12-11	**en**	2012-01-17	37
Rougeole	Europe	**es**	2011-12-01	*el*	2012-01-06	36
Grippe	Hong Kong	ru	2011-12-24	fr	2012-01-23	30
Grippe	Canada	**en**	2011-11-04	**en**	2011-12-01	27
Pneumonie atypique	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
Grippe	Népal	en	2011-11-29	en	2011-12-06	6
Syphilis	Espagne	**es**	2011-11-29	**es**	2011-12-01	2

Tableau 5.12 – Exemples de PML pour lesquelles le premier signalement a été effectué par ProMED. Pour chaque paire nous indiquons la langue et la date de détection par chacun des systèmes ainsi que le décalage (en jours) de DAnIEL. En gras, les langues de détection qui sont des langues officielles du pays, en italique les langues non-couvertes par ProMED.

date du plus ancien rapport issu de chaque système durant la période.

Le tableau 5.12 présente des exemples de PML signalées en premier lieu par ProMED, le tableau 5.13 (page 109) 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).

Parmi les PML extraites par les deux systèmes, dans 37% des cas DAnIEL a fourni le tout premier signalement (Tableau 5.14 page 109). 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.

La **plus-value** offerte par DAnIEL est relativement faible en Amérique du Nord pour les mêmes raisons. Ce phénomène est plus contrasté en Afrique où DAnIEL tire avantage du traitement de l'arabe dans certaines régions. Globalement, il 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.

Paire		ProMED		DAnIEL		Plus-value
Maladie	Lieu	Lg.	Date	Lg.	Date	(jours)
Grippe	Kazakhstan	pt	2012-02-28	*pl*	2012-01-06	53
Norovirus	Russie	**ru**	2011-12-27	**ru**	2011-11-28	29
Hépatite	Pays-Bas	en	2012-02-12	*nl*	2012-01-17	26
Encéphalite japonaise	Inde	**en**	2011-11-02	**en**	2011-10-11	22
Méningites	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	Royaume-Uni	**en**	2012-02-02	**ru**	2012-01-23	10
Salmonelle	Russie	**ru**	2012-01-14	**ru**	2012-01-06	8
Grippe	Espagne	**es**	2011-12-14	**es**	2011-12-09	5
Leptospirose	Philippines	ru	2012-01-08	*el*	2012-01-03	5
Dengue	Brésil	**pt**	2011-12-03	**pt**	2011-12-01	2

Tableau 5.13 – Exemples de PML pour lesquelles le premier signalement vient de DAnIEL. Pour chaque paire nous indiquons la langue et la date de détection par chacun des systèmes ainsi que la plus-value (en jours) par rapport à ProMED. En gras, les langues de détection qui sont des langues officielles du pays, en italique les langues non-couvertes par ProMED.

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

Tableau 5.14 – Localisation des Premiers Signalements (PS) de chacun des systèmes

Nous présentons dans le tableau 5.15 (page 110) 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 portuguais.

	ar	cs	de	el	en	es	fi	fr	it	nl	no	pl	pt	ru	sv	tr	zh
ProMED	-	-	-	-	54	27	-	5	-	-	-	-	15	6	-	-	-
DAnIEL	1	2	3	4	8	8	2	8	3	3	0	2	4	9	1	0	3

Tableau 5.15 – Repartition par langue des premiers signalements de ProMED et DAnIEL. "-" signale une langue non couverte

5.2.4 Conclusions sur la comparaison ProMED–DAnIEL

La comparaison des sorties des deux systèmes a permis de mettre en lumière les bénéfices que l'on pouvait attendre d'un traitement multilingue. 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 explication est que la qualité du nettoyage du code source était très variable selon les sources. Il est dès lors difficile d'évaluer si tous les évènements signalés par ProMED auraient pu être « identifiés » dans le corpus par DAnIEL. Enfin, nous n'avons pas sur cette expérience d'évaluation de l'influence de chaque phase du traitement opéré par DAnIEL.

Les expériences visant à corriger ces biais sont exposées dans la Section 5.3. Le corpus de référence multilingue que nous avons constitué est présenté dans la Section 5.3, et son annotation par des locuteurs de chaque langue dans la Section 5.3.2.

5.3 Résultats de DAnIEL sur un corpus de référence

Dans cette section figure une évaluation des différents aspects de notre approche. Tout d'abord l'efficacité du principe de répétition est étudiée. Puis, les résultats de DAnIEL et ceux de l'annotation manuelle de référence sont comparés. Ensuite, nous examinons quel est le regard des annotateurs sur les sorties de DAnIEL et nous proposons une évaluation fondée sur le nombre de PML extraites plutôt que sur le nombre de documents concernés. Enfin, une étude complète de l'impact des différents paramètres pouvant affecter DAnIEL est réalisée.

5.3.1 Construction du corpus

L'absence de données annotées de référence librement disponibles nous a conduit a construire notre propre référence. Pour les expériences qui suivent, une partie du corpus décrit dans la Section 5.2.2 a été utilisée. Le filtrage par la catégorie santé qui a été utilisé permet d'obtenir un corpus de taille significative à moindre coût. En effet, sur une source filtrée les documents jugés pertinents par les annotateurs représentaient de 6 à 10% du total de documents examinés. Sur des sources non-filtrées, cette proportion était inférieure à 0,5%. De ce fait, obtenir un nombre significatif de documents pertinents était très coûteux. Faute d'une quantité significatives de données pertinentes, l'impact des

expériences aurait été grandement affecté. La campagne d'annotation s'est concentrée sur l'anglais, chinois, le grec, le polonais et le russe de manière à offrir une variété importante en familles de langues.

5.3.2 Instructions d'annotation

Afin de mesurer le rappel et la précision sur le choix des documents pertinents et la caractérisation des PML, des locuteurs de chaque langue ont été chargé d'annoter des jeux de documents. Pour chaque langue a été constitué un corpus d'au moins 500 documents. Parmi eux, 100 étaient utilisés pour vérifier si des ajustements de l'algorithme étaient nécessaires, le reste étant conservé pour l'évaluation elle même.

Les annotateurs [58] devaient juger si le document était pertinent pour la surveillance des maladies infectieuses. Pour les documents jugés pertinents, ils devaient préciser de quelle maladie dans quel pays il était question. Les annotations utilisées figurent en ligne sur `https://daniel.greyc.fr` de même que les résultats obtenus par DAnIEL sur ce corpus [59]. Les caractéristiques du corpus figurent dans le tableau 5.16. La répartition des documents par langue est relativement homogène. Le nombre de documents en polonais est toutefois plus faible que celui des autres langues du fait de la présence d'un certain nombre de doublons dans le corpus annoté.

	anglais	chinois	grec	polonais	russe	ensemble
#documents	475	446	390	352	426	2089
dont pertinents	31 (6,5%)	16 (3,6%)	26 (6,7%)	30 (8,5%)	41 (9,6%)	144 (6,9%)
#paragraphes	6791	4428	3543	3512	2891	21165
moy. ±écart-type	14.29±7.23	9.9±10.5	9.08±7.78	9.97±6.95	6.78±6.11	10.13±8.3
#caractères (10^6)	1.35	1.14	2.05	1.04	1.56	7.17
moy. ± écart-type	2568±2796	2858±1611	5264±5489	2971±2188	3680±5895	3432±4085

Tableau 5.16 – Caractéristiques du corpus annoté : nombre de documents et leur taille en paragraphes et en caractères

La proportion de documents pertinents est de 6,9% au total mais il existe une certaine variété entre les langues. En chinois, les annotateurs ont sélectionné 3,6% de documents pertinents alors qu'en russe ce taux avoisine les 10%. Cette variabilité est due à deux effets principaux. Tout d'abord, de façon évidente, le nombre d'évènements reportés dans deux langues n'est pas nécessairement le même dans une période de temps donné. Ensuite, la qualité du filtrage qui sélectionne des documents pour le fil santé prête à discussion. Selon nos annotateurs, un nombre important de documents figurant sur le fil santé en chinois

58. Nos quinze annotateurs étaient pour la plupart locuteurs natifs des langues qu'ils avaient à traiter. Nous précisons qu'ils n'étaient pas impliqués dans le développement de DAnIEL.

59. Les documents sont disponibles sur la plateforme et classés par catégories. Les annotations au format *JSON* (*JavaScript Object Notation*) sont également disponibles.

avaient peu à voir avec la santé. Enfin, la notion de pertinence semble être variable selon les annotateurs. Ce dernier point sera discuté dans la Section 5.3.9.

5.3.3 Filtrage des documents pertinents

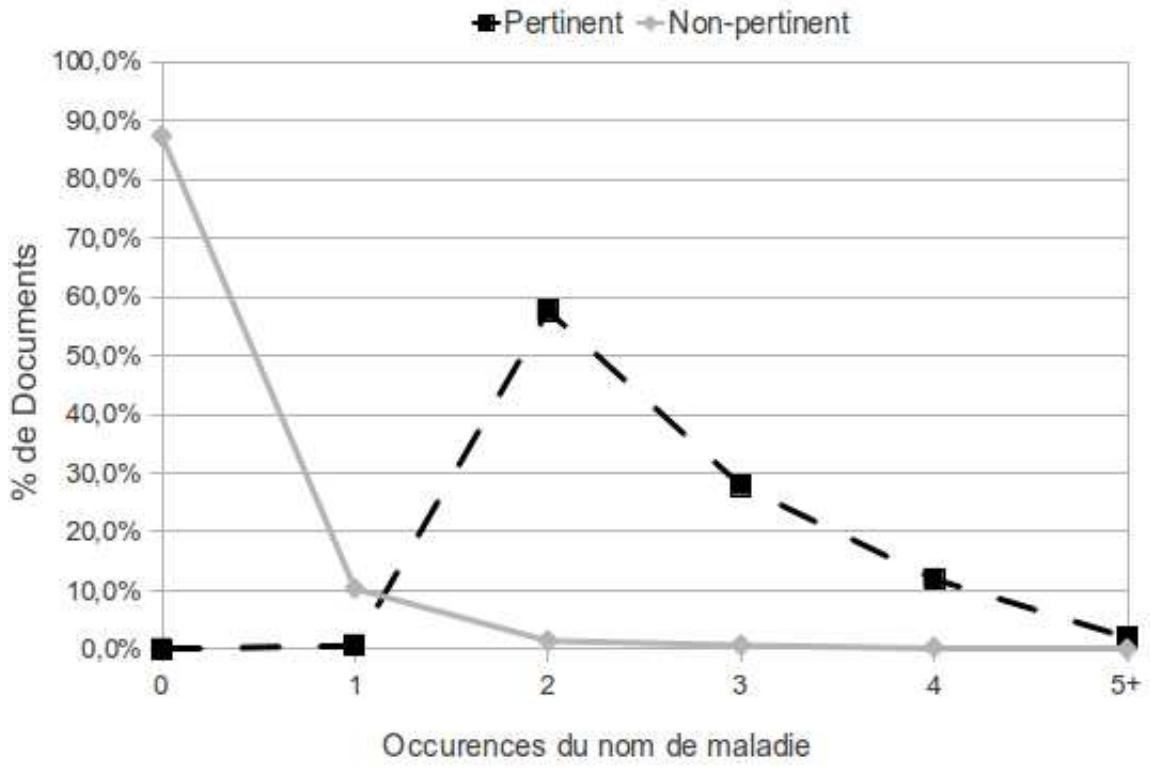


FIGURE 5.4 – Occurrences d'un même nom de maladie dans des articles pertinents et non-pertinents

La Figure 5.4 montre le nombre d'occurrences d'un même nom de maladie dans les articles pertinents et non-pertinents de notre corpus annoté. Ce qui discrimine réellement ces deux classes de documents n'est ni la présence d'un nom de maladie, ni sa grande fréquence, mais simplement la différence entre une fréquence de 1 (terme *hapax*) et une fréquence de 2 ou plus (terme répété).

La répétition du nom de maladie est très discriminante ; elle permet de filtrer 97% des documents non pertinents. Un seul des documents jugés pertinents par nos annotateurs ne contenait pas de répétition du nom de maladie (soit 0,7%).

Le tableau 5.17 (page 113) montre qu'un nombre importants de motifs sont filtrés grâce à l'utilisation des critères de position. Il est important de mesurer à quel point ce filtrage est justifié et peut avoir un impact sur le rappel et la précision.

Langues	anglais	chinois	grec	polonais	russe
#Documents	396	415	159	192	90
#Motifs sans segmentation (moy.)	1101,45	271,72	1242,81	1128,12	1311,07
#Motifs avec segmentation (moy.)	114,67	120,70	148,33	129.05	159.72
Taux de filtrage	9.60	2,62	8.67	8.74	8.20

Tableau 5.17 – Impact du filtrage par position sur le nombre de motifs pour les articles de type moyen et long

Le Tableau 5.18 montre de manière détaillée l'impact des critères de répétition et de position utilisés par DAnIEL. Notons que pour ces *baselines*, le vocabulaire était constitué de l'intégralité des noms de maladies existants dans la base de connaissance de DAnIEL à laquelle nous avons ajouté les noms de maladies présents dans les annotations. Il s'agit donc du vocabulaire nécessaire et suffisant à un rappel parfait. L'analyse est effectuée au grain caractère mais avec $\theta = 1$. Au grain mot, les résultats de la *baseline* en terme de rappel sont inférieurs pour le grec comme nous l'avons montré dans [Lejeune-2012b].

		anglais	chinois	grec	polonais	russe	ensemble
Baseline 1 (B1) *présence*	Précision	0,30	0,47	0,41	0,39	0,59	0,41
	Rappel	1,0	1,0	0,96	0,90	0,88	0,94
	F_1-mesure	0,47	0,64	0,57	0,54	0,71	0,57
	F_2-mesure	0,69	0,82	0,76	0,71	0,80	0,74
Baseline 2 (B2) *répétition*	Précision	0,44	0,76	0,57	0,50	0,76	0,57
	Rappel	0,91	1,0	0,92	0,60	0,78	0,83
	F_1-mesure	0,59	0,86	0,69	0,55	0,77	0,68
	F_2-mesure	0,75	0,94	0,8	0,58	0,78	0,76
Baseline 3 (B3) *répétition+position*	Précision	0,63	0,80	0,74	0,63	0,76	0,71
	Rappel	0,71	1,0	0,93	0,33	0,76	0,72
	F_1-mesure	0,67	0,89	0,82	0,43	0,76	0,71
	F_2-mesure	0,69	0,95	0,88	0,37	0,76	0,72

Tableau 5.18 – Évaluation de trois *baseline* : précision, rappel, F_1-mesure et F_2-mesure

Grâce à ce vocabulaire, trois *baselines* ont été construites :

B1 utilise simplement la **présence** d'un nom de maladie dans le document ;

B2 sélectionne un document en cas de **répétition** d'un nom de maladie ;

B3 combine la **répétition** et la **position**.

B1 est une *baseline* très naïve qui illustre la difficulté de traiter des langues à morphologie riche. Le rappel ne compense pas la très mauvaise précision. En anglais et en chinois, le rappel est de 1. C'est un résultat attendu puisque se passer de lemmatisation pour ces deux langues est plus facile, il est très probable de trouver la forme attendue dans le

texte. Le grec fait exception parmi les langues à morphologie riche, le terme canonique est fréquemment présent dans les documents.

B2 montre le gain en précision offert par le simple critère de répétition (indépendamment de la position). Globalement cela permet de gagner 16 points de précision et 11 points de F_1-mesure. Le rappel en grec reste proche de la B1, la forme canonique est souvent répétée.

Avec la *baseline* B3, la précision augmente de 14 points mais la F_1-mesure n'augmente que de quatre points. Le gain en F-mesure (F_1 ou F_2) offert par l'utilisation des critères de répétition et de position est important. Toutefois, la perte en rappel est assez grande pour le polonais et pour le russe du fait de l'absence de prise en compte de la morphologie.

5.3.4 Évaluation du seuil θ

Nous proposons ici une évaluation du seuil de comparaison θ utilisé pour décider de la pertinence du document (cf. Section 5.1.3). Le seuil θ détermine à partir de quelle valeur du ratio $len(motif)/len(terme)$ DAnIEL considère qu'un document est pertinent pour la veille épidémiologique.

Nous observons que les courbes de performance en fonction de ce seuil ont globalement la même allure pour chaque langue. Augmenter la valeur de θ permet d'améliorer la précision au prix d'une perte en rappel par paliers. Avec $\theta < 0.5$, la précision est très faible. En effet, de trop nombreux documents sont sélectionnés à tort du fait que la contrainte de comparaison des chaînes est trop relaxée. Les sous-chaînes trouvées à des positions pertinentes sont trop petites en proportion de la taille des termes recherchés. Des exemples de chaînes extraites pour des valeurs faibles de θ figurent dans le tableau 5.6 (page 99).

Lorsque la valeur du seuil θ augmente, la précision augmente. Les chaînes qui permettent la sélection des documents sont beaucoup plus proche de la forme canonique du nom de maladie. Aux alentours de $\theta = 0,8$, le rappel se met à chuter : la contrainte devient trop forte et certaines variations morphologiques ne sont plus « acceptées » par le système. La chute est d'autant plus grande que la langue est riche morphologiquement.

Pour le chinois, le rappel reste constant à $1,0$. Le rappel pour le grec est peu affecté par l'augmentation de θ, de même, mais dans une moindre mesure, que le rappel pour l'anglais. À l'opposé pour le russe et surtout pour le polonais, le rappel chute à mesure que la contrainte du seuil devient plus forte, se rapprochant ainsi des résultats obtenus par la *baseline* B3 (Tableau 5.18). Pour ces deux langues, aucun gain en précision n'est mesuré lorsque l'on fait varier θ au-delà de la valeur par défaut du système $(0,8)$.

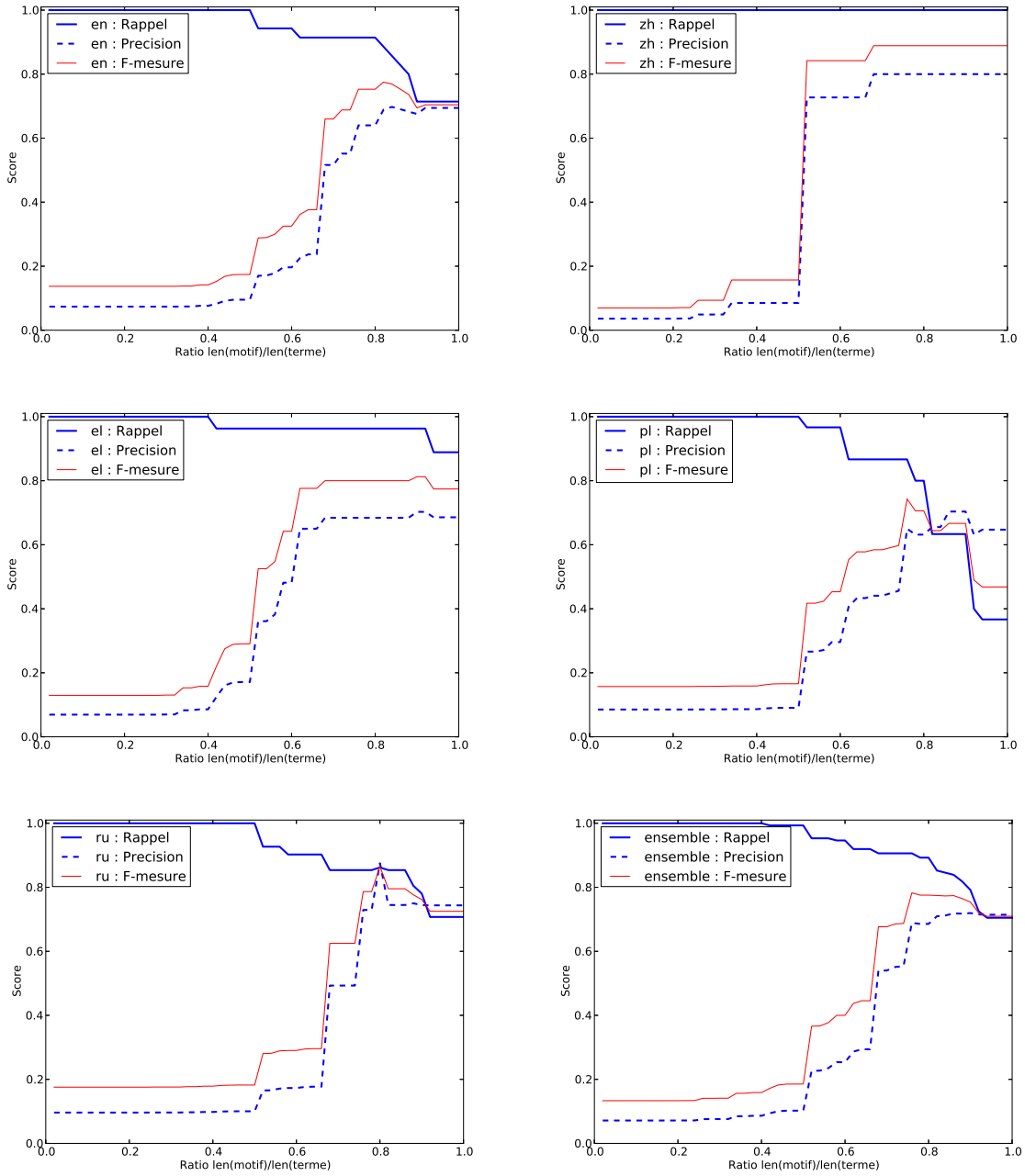


FIGURE 5.5 – Rappel, précision et F_1-mesure en fonction du seuil θ (par langue (anglais, chinois, grec, polonais et russe) et pour le corpus cumulé

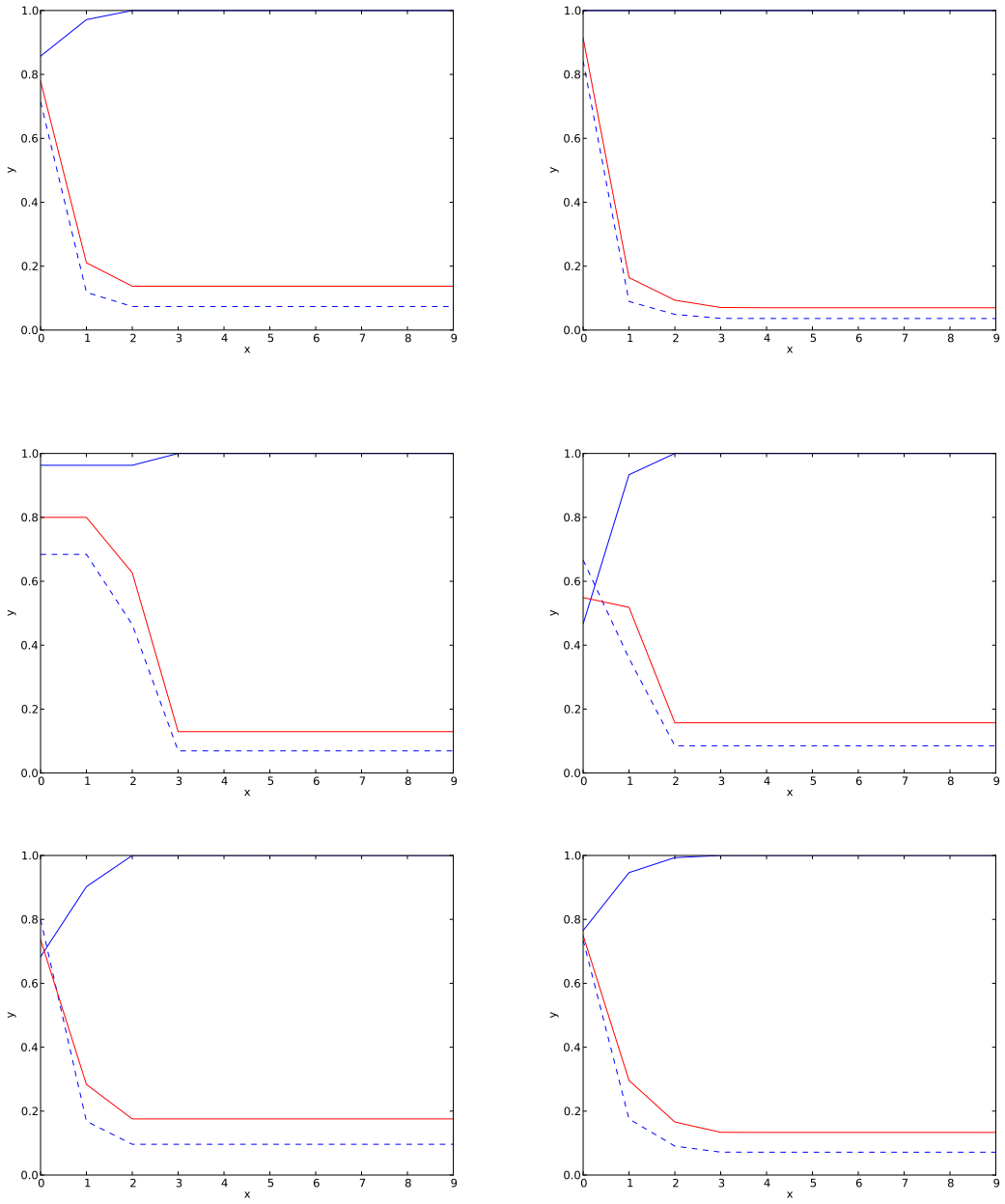


FIGURE 5.6 – Rappel, précision et F_1-mesure avec l'application d'une seuil absolu (par langue (anglais, chinois, grec, polonais et russe) et pour le corpus cumulé. La valeur en abscisse représente la différence admise (en nombre de caractères) entre le nom de maladie et les sous-chaînes identifiées dans le texte.

La figure 5.6 montre que le choix d'un seuil absolu serait moins efficace que celui d'un ratio.. La perte en précision serait trop grande dès que l'on s'éloigne de la forme lexicalisée. De ce fait, la possibilité de paramétrer finement le système pour favoriser spécifiquement le rappel ou la précision serait plus difficile.

5.3.5 Évaluation du rappel et de la précision

	anglais	chinois	grec	polonais	russe	Ensemble	
θ	0.8	0.75	0.75	0.76	0.85	0.8	θ optimal par langue
Précision	0.70	0.84	0.68	0.65	0.88	0.69	0.75
Rappel	0.89	1.0	0.96	0.87	0.86	0.89	0.91
F_1-mesure	0.78	**0.91**	**0.80**	0.77	0.87	0.78	0.82
F_2-mesure	0.84	**0.96**	**0.89**	0.86	0.87	0.84	0.88

Tableau 5.19 – Filtrage des documents : précision, rappel et F_2-mesure pour le meilleur θ (valeur minimale) individuel, pour la valeur par défaut et pour la combinaison des meilleures valeurs

Pour un système basé sur des ressources lexicales minimales, la question du rappel est *a priori* la plus problématique. Les expérimentations montrent au contraire que le rappel est globalement élevé et surtout supérieur à la précision (Tableau 5.19). L'optimisation de la valeur θ a principalement pour effet d'améliorer la précision.

Le compromis rappel-précision est souvent évalué à l'aide de la F-mesure. Une autre manière de l'évaluer est d'utiliser une courbe *ROC* (*Receiver Operating Characteristics*). Celle ci permet de représenter les performances d'un classifieur binaire pour différents jeux de paramètres. En ordonnée figure le pourcentage de vrais positifs (le rappel ou sensibilité). Il est également tenu compte de la spécificité (ou précision) : en abscisse figure le pourcentage de faux positifs (égal à 1 moins spécificité). Plus un point est proche de 1 en ordonnée, plus la sensibilité plus grande. Plus l'abscisse d'un point est proche de 1, plus le bruit est élevé (la spécificité est faible).

Un excellent classifieur atteint une sensibilité très élevée sans que sa spécificité faiblisse trop vite. La courbe ROC « idéale » a donc une pente très forte dès les faibles valeurs de bruit (gauche de la courbe). La diagonale qui part du point $(0, 0)$ et qui va vers le point $(1, 1)$ représente la *baseline*. Cette *baseline* est un classifieur binaire naïf qui possède deux configurations. Soit il considère tous les documents comme non-pertinents : sensibilité = 0 et bruit =0. Soit il considère tous les documents comme pertinents : sensibilité = 1 et bruit proche de 1.

Plus la courbe d'un système se situe au-dessus de cette *baseline*, meilleure est sa performance. L'aire sous la courbe (*Area Under the Curve*) en constitue l'évaluation chiffrée. Pour un système parfait elle est de 1 : le système atteint une sensibilité de 1 en conservant un bruit (1-spécificité) égal à 0. Pour la *baseline* l'aire sous la courbe est de $0,5$. L'aire représente la probabilité d'identifier correctement des documents pertinents lorsque le système est également confronté à des documents non-pertinents. Pour la *baseline*, cette probabilité est donc égale à $0,5$

Il est possible calculer l'aire sous la courbe en utilisant une intégrale ou en l'approximant avec la « méthode des trapèzes ». Dans la mesure où il y a peu de points, la fonction

est affine par morceaux, on obtient avec cette méthode une valeur exacte avec un faible coût calculatoire. La méthode consiste simplement à prendre la liste des coordonnées (x, y) de la courbe triée par ordre croissant. Pour chaque élément i de la liste, le couple de coordonnées (x_i, y_i) et le couple de coordonnées (x_{i+1}, y_{i+1}) servent à former un trapèze. Ses quatre coins se situent aux coordonnées $(x_i, 0)$,(x_i, y_i),(x_{i+1}, y_{i+1}) et $(x_{i+1}, 0)$. L' aire sous la courbe est la somme des aires de tous les trapèzes formés.

La figure 5.7 présente, sous forme de courbe ROC, les performances de DAnIEL sur notre jeu de données de référence en cinq langues (anglais, chinois, grec, polonais et russe).

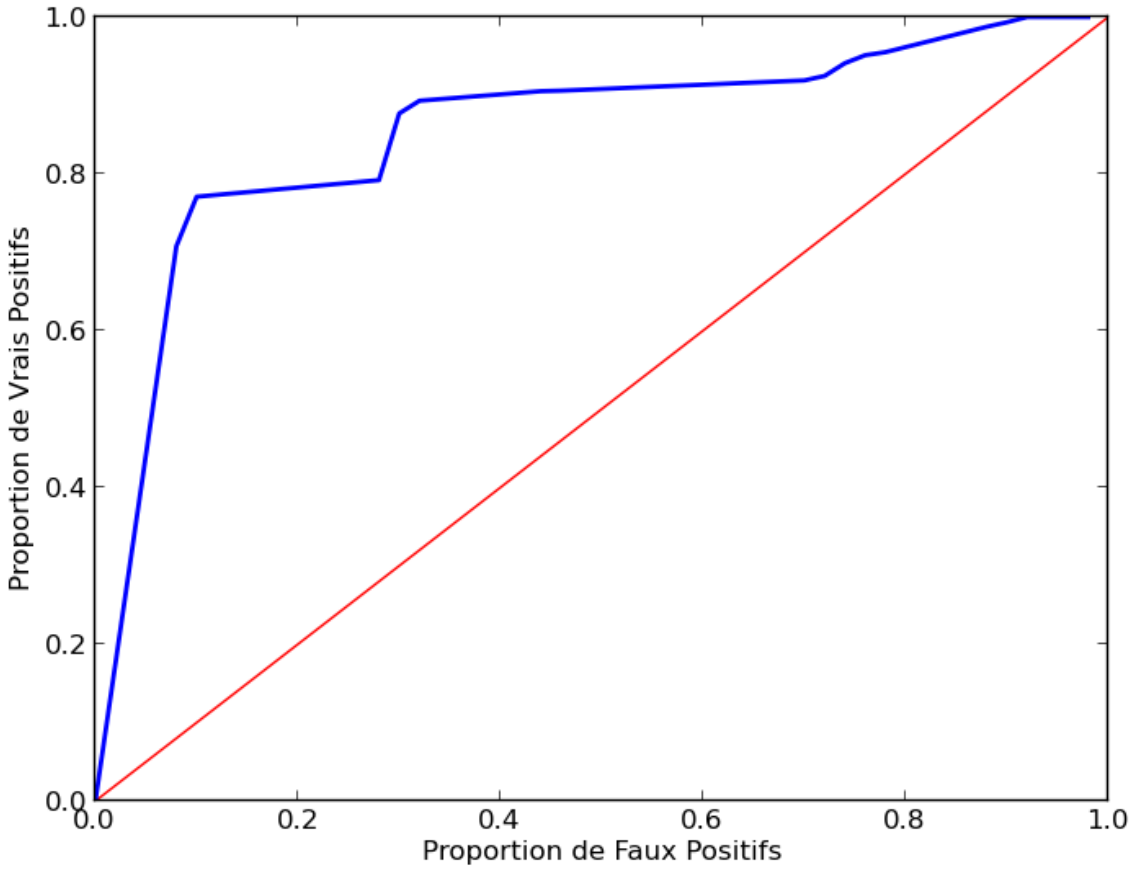


FIGURE 5.7 – Courbe ROC du système DAnIEL (bleu) sur le jeu de données de référence. La *baseline* apparaît en rouge. L'aire sous la courbe est de 0,86.

La courbe de DAnIEL 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. C'est le point où DAnIEL obtient la meilleure $f_{0.5}$-mesure avec $0,88$.

Le second palier se situe aux coordonnées $(0.31, 0.91)$. À partir de ce point, la pente de la courbe devient très faible. Autrement dit, tout gain en sensibilité devient prohi-

bitif au regard du bruit engendré. Ainsi, un gain de $0,04$ en sensibilité occasionne une augmentation du bruit de $0,41$.

5.3.6 Typage des erreurs impactant le rappel

Le Tableau 5.20 propose une classification des faux négatifs selon le type de phénomène qui occasionne l'erreur de DAnIEL. Il s'agit principalement d'évaluer l'impact de la parcimonie des ressources. Par exemple, l'absence dans le lexique du terme « maladie respiratoire » en russe a occasionné deux faux négatifs. Le Tableau 5.19 montre que le rappel est sensiblement supérieur à la précision. Notre choix de se baser sur un vocabulaire minimal a très peu impacté cette mesure. De fait, l'utilisation systématique des termes spécialisés dans les articles de presse grand public est très rare. Quand le terme spécialisé est employé, son équivalent « générique » est généralement utilisé en complément, afin de s'assurer de la compréhension par tous les lecteurs. Il peut arriver que le terme spécialisé devienne un terme courant, c'est le cas de « H5N1 » et « H1N1 ».

Ce ne sont pas les erreurs dues à la petitesse du lexique qui ont été les plus fréquentes mais celles dues à la comparaison entre la longueur du motif trouvé et la longueur de l'entrée dans la base de connaissances.

	anglais	chinois	grec	polonais	russe	ensemble
Manque de lexique	1	0	0	1	3	5
Absence de répétition	0	0	0	1	0	1
Répétition non-détectée	1	0	0	0	1	2
Erreur de comparaison	1	0	1	2	2	6
Silence	3	0	1	4	6	14
#documents pertinents	31	16	26	30	41	144

Tableau 5.20 – Erreurs affectant le rappel lors du filtrage des documents pour la valeur de θ qui optimise la F_1-mesure

Ces erreurs sont principalement provoquées par des noms de maladies « courts », pour lesquels le ratio autorisé par défaut ($\theta = 0.8$) est trop faible. Toutefois la relaxation de cette contrainte affecte grandement la précision. La valeur par défaut de θ est un point d'équilibre fiable du système.

La règle de répétition n'affecte que très peu le rappel puisqu'un seul document pertinent (en polonais) ne comportait pas de répétition du nom de maladie. À deux reprises, l'erreur est venue d'une répétition non détectée du fait du modèle de document. Dans les deux documents en question (un en russe et un en anglais), les paragraphes étaient nombreux mais ne comportaient chacun qu'une phrase. Le modèle de document au grain paragraphe s'est avéré trop rigide dans ces deux cas, les documents ont été considérés à tort comme des articles d'analyse.

5.3.7 Localisation de l'évènement

Nous avons sélectionné des documents où la localisation n'était pas mentionnée explicitement. L'objectif est de mesurer si relier les maladies extraites à la localisation de la source était pertinent (Tableau 5.21), si la règle de localisation implicite était efficace[60].

	anglais	grec	polonais	russe
# documents sélectionnés	93	188	213	230
Dont documents sans localisation explicite	46	33	35	51
Localisation identique à la source	78.3%	81.8%	82.9%	78.4%
Localisation différente de la source	21.7%	18.2%	17.1%	21.6%
Taux global d'erreur	12.2%	3%	2.8%	4.8%

Tableau 5.21 – Résultats de la règle de localisation implicite

Dans notre corpus, 70% des épidémies sont localisés explicitement, la localisation est présente dans le texte. L'application de la règle de localisation implicite sur les évènements restants est efficace, particulièrement si l'on tient compte de son coût très faible en calcul comme en ressources ([Lejeune-2012b]). En effet, parmi les 22% de documents pertinents en russe sans localisation explicite, 78% sont correctement localisés. Ce qui laisse au total seulement 4,8% d'évènements mal localisés au total. La majorité des erreurs vient d'une mauvaise localisation de la source (un journal ukrainien considéré comme russe par DAnIEL par exemple) ou encore de faits dépassant le cadre du seul pays (Grippe dans le monde par exemple).

5.3.8 Évaluation de la localisation explicite

Bien que la règle de localisation implicite soit efficace, elle ne suffit pas à localiser de façon sûre un nombre significatif de faits épidémiologiques. La règle de localisation explicite est appliquée avec la même méthode que la détection du nom de maladie. Ce sont les sous-chaînes répétées de noms de lieux apparaissant à des positions clés qui servent d'indices. La règle de localisation explicite est influencée par le seuil θ choisi pour la comparaison entre les chaînes de caractères trouvées dans le texte et le terme canonique. L'évaluation de ce seuil est difficilement détachable de celle du θ utilisé pour la sélection des documents. En effet, il n'est pas possible d'évaluer la localisation des évènements épidémiologiques indépendamment des documents pertinents effectivement retrouvés par DAnIEL[61].

Le seuil utilisé pour la sélection du nom de maladie est nommé θ_1, celui utilisé pour le nom de lieu est θ_2. La figure 5.8 montre la carte de chaleur (*heatmap*) du rappel pour les PML extraites selon l'évolution de θ_1 et θ_2. Dans cette figure, le bleu foncé représente les

60. Cette expérience n'a malheureusement pas pu être menée sur le chinois du fait du coût humain.
61. Il serait intéressant d'évaluer cette règle sur d'autres domaines que l'épidémiologie.

résultats les plus mauvais, le rouge foncé représente les meilleurs résultats. Plus la couleur « tire » vers le bleu, plus les résultats sont dégradés, inversement pour le rouge.

Le résultat correspond à la proportion de PML correctement repérées sur le total de PML à repérer. Seules les PML où les deux attributs (maladie et lieu) sont identiques, dans la référence humaine et dans la sortie de DAnIEL, sont comptabilisées comme Vrais Positifs.

La zone rouge foncée correspond à la combinaison de paramètres suivante : $0.55 <= \theta_1 <= 0.8$ et $0.8 <= \theta_2 <= 0.85$. La détermination de θ_2 influe moins sur les résultats du fait de la règle de localisation implicite. Quand $\theta_2 >= 0.5$, peu de lieux « aberrants » sont extraits et peu de PML sont affectées par ce paramètre.

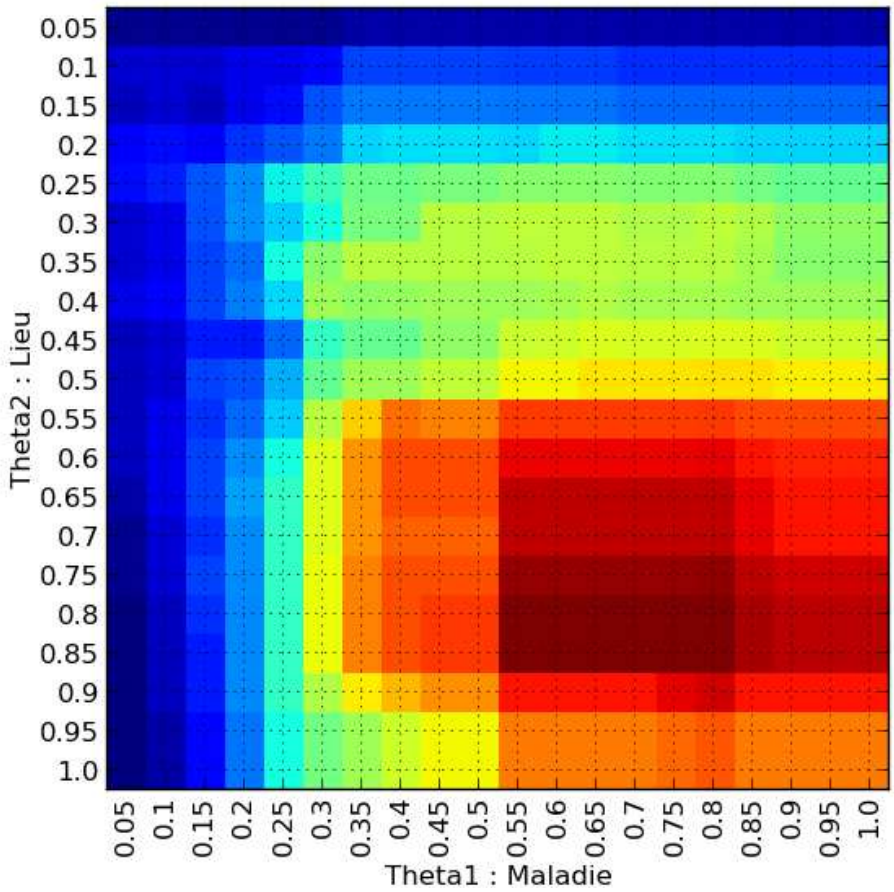


FIGURE 5.8 – Évaluation par PML, rappel en fonction de θ_1 (maladie) et θ_2 (lieu) pour les langues suivantes : anglais, chinois, grec, polonais et russe.

La figure 5.9 propose une évaluation de la précision sur les PML. Il s'agit donc de la proportion de PML correctes (sur les deux attributs) vis à vis du total de PML retournées

par le système. La zone où la précision est la meilleure représente une surface proche de la zone optimale pour le rappel (figure 5.8). Toutefois, les combinaisons optimales de θ_1 et θ_2 sont moins nombreuses. L'optimisation de la précision de la méthode est donc plus difficile que l'optimisation du rappel, le paramétrage doit être plus fin. Ce résultat n'est pas étonnant, dans la mesure où il fait écho aux résultats sur le filtrage présentés dans la figure 5.19. À nouveau, la parcimonie de la méthode n'a pas d'impact négatif sur le rappel : avoir beaucoup moins de termes stockés en mémoire n'entraîne pas significativement plus de silence. Sur cette évaluation des PML, c'est à nouveau la précision qui est plus problématique. Une évaluation quantitative est menée dans la section 5.3.9 pour mieux appréhender cet aspect.

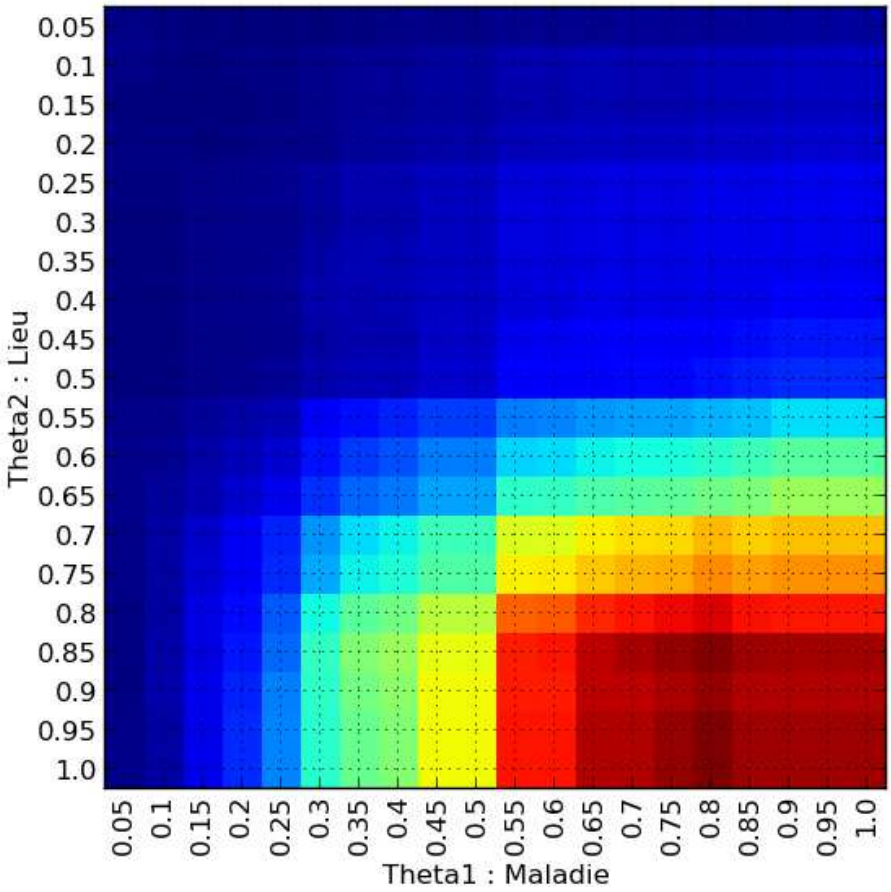


FIGURE 5.9 – Évaluation par PML, précision en fonction de θ_1 (maladie) et θ_2 (lieu) pour les langues suivantes : anglais, chinois, grec, polonais et russe.

Enfin, la figure 5.10 propose une carte de chaleur sur la F_1-mesure. Les valeurs numériques utilisées sont présentées dans le tableau 5.22 (page 124). Cette représentation est en quelque sorte une synthèse des deux figures précédentes ; il n'est donc pas surprenant d'y trouver des points communs sur les valeurs optimales des seuils utilisés pour la détection des maladies (θ_1) et des lieux (θ_2). Le premier point commun est que lorsque $\theta_1 > 0.55$ et $\theta_2 > 0.6$, les résultats sont plus fiables : la F_1-mesure est strictement supérieure à 0,4. Le second est que pour $\theta_1 > 0.65$ et $\theta_2 > 0.6$, on a $F_1 > 0.6$. Les meilleurs combinaisons sont observées pour la combinaison ($0.55 < \theta_1 <= 1, \theta_2 = 0.85$), avec comme résultat : $0.74 <= F_1 <= 0.8$.

Le jeu de paramètres peut donc être ajusté selon les besoins de l'utilisateur final bien que DAnIEL fonctionne efficacement avec les paramètres par défaut $\theta_1 = \theta_2 = 0.8$.

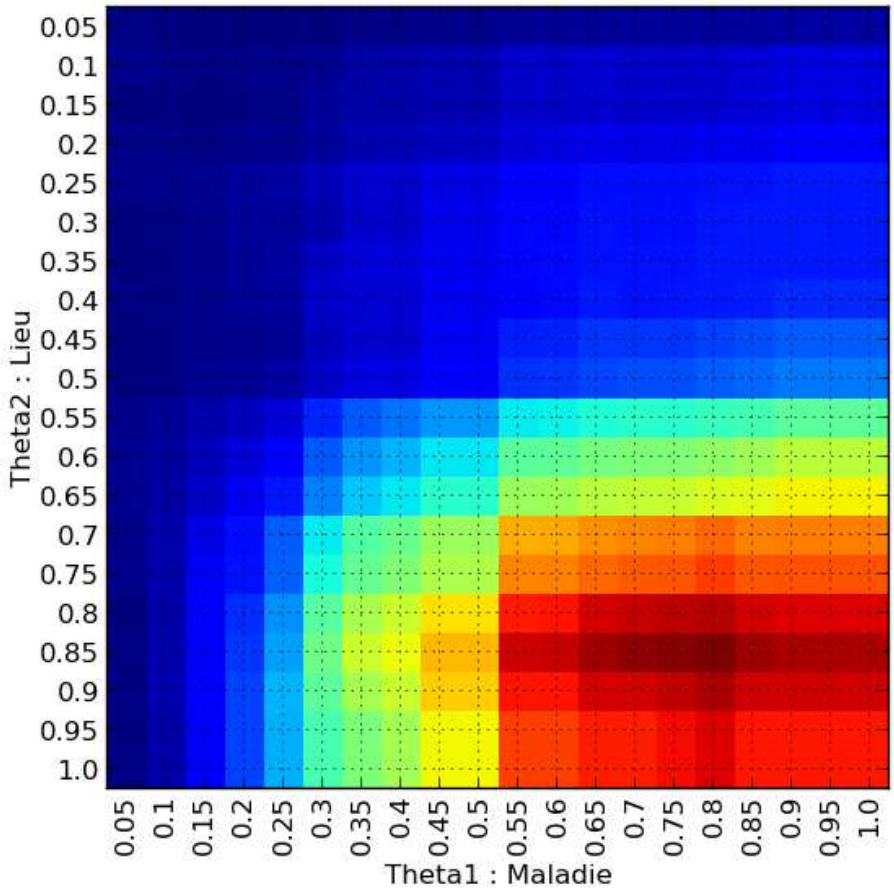


FIGURE 5.10 – Évaluation par PML, F_1-mesure en fonction de θ_1 (maladie) et θ_2 (lieu) pour les langues suivantes : anglais, chinois, grec, polonais et russe.

θ_1 \ θ_2	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	**80**	85	90	95	100
05	1	1	0	0	0	0	1	1	1	1	2	2	2	2	2	2	3	3	3	3
10	1	1	1	1	1	2	3	3	4	4	5	5	6	6	6	6	6	7	7	7
15	1	1	0	1	1	2	3	3	4	4	5	5	6	6	6	6	7	7	7	7
20	1	1	1	1	1	2	4	4	5	5	7	7	8	8	8	8	8	9	9	9
25	1	1	2	2	3	4	6	6	8	8	10	10	11	11	11	12	12	12	12	12
30	1	1	1	2	2	4	6	6	8	8	11	11	12	11	11	12	12	12	12	12
35	1	1	1	2	3	5	7	7	8	8	11	11	12	12	12	12	12	12	12	12
40	1	1	1	2	3	6	6	6	8	8	11	11	12	12	12	12	13	14	14	14
45	1	1	1	1	2	5	6	6	9	9	12	13	14	14	14	15	16	17	17	17
50	1	1	2	2	3	6	7	7	10	10	14	14	16	16	16	17	19	20	20	20
55	1	2	3	5	7	13	17	2	22	22	29	29	31	32	32	34	35	37	37	37
60	2	3	5	7	9	17	22	24	29	29	37	37	4	40	41	42	44	46	46	46
65	2	3	6	9	12	20	26	28	32	32	43	44	46	47	47	49	50	52	52	52
70	1	4	7	11	18	29	36	38	43	43	58	58	60	61	62	64	62	62	62	62
75	1	4	8	12	18	31	38	40	45	45	61	62	64	65	65	67	65	66	66	66
80	1	3	9	14	22	37	45	48	54	54	70	71	74	75	75	76	74	73	73	73
85	1	4	9	15	23	39	48	51	57	57	74	75	77	78	79	**80**	78	77	77	77
90	1	4	9	15	24	37	44	47	55	55	71	71	74	74	75	77	74	74	74	74
95	1	3	9	16	24	35	40	44	52	52	67	67	70	70	72	73	70	70	70	70
100	1	3	9	16	24	35	40	44	52	52	67	67	70	70	72	73	70	70	70	70

Tableau 5.22 – Évaluation par PML, F_1-mesure en fonction de θ_1(maladie) et θ_2(lieu) pour les langues suivantes : anglais, chinois, grec, polonais, russe (toutes les valeurs sont indiquées en pourcentage).

5.3.9 Évaluation qualitative des PML extraites

La grande majorité du bruit dans les résultats de DAnIEL provient de faits secondaires : par exemple la campagne de vaccination, corollaire d'une épidémie précédente. L'extraction de véritables faux négatifs (évènements anciens, historiques ou erronés) est beaucoup plus rare. Pour mesurer précisément cette répartition, les sorties de DAnIEL ont été examinées par des annotateurs. L'objectif est de mesurer plus finement la précision. Une classification proche de celle proposée par Yangarber ([Yangarber-2011a]) a été utilisée. Il s'agissait pour les annotateurs de juger si l'évènement décrit par DAnIEL était :

non pertinent : évènement erroné ou évènement ancien ;

assez pertinent : article de synthèse (évènement en cours), réactions à une épidémie ;

très pertinent : nouvelles informations, mise à jour d'un évènement.

Chaque évènement extrait a été jugé par 3 annotateurs et le jugement majoritaire a été conservé [62]. Le Tableau 5.23 (page 125) synthétise ces résultats. Nous observons qu'environ 6% des documents sélectionnés par DANIEL ont été étiquetés « non-pertinents » par les annotateurs. Plus de 83% ont été étiquetés « très pertinents ». La majorité du bruit, les deux-tiers dans cette expérience, généré par DAnIEL est de moyenne importance : il s'agit

62. Les annotateurs affichaient régulièrement des désaccords mais, de façon intéressante, aucun document n'amenait d'indécision totale (1 vote dans chaque catégorie).

	grec	polonais	russe
# Évènements	168	185	243
Très pertinent	82.7%	85.4%	82.3%
Assez pertinent	11.9%	8.1%	11.5%
Non pertinent	5.4%	6.5%	6.2%

Tableau 5.23 – Jugement des annotateurs sur les résultats de DAnIEL

de documents assez pertinents. Ce sont très majoritairement des documents relatant des informations « secondaires » pour la veille épidémiologique :

- campagnes de vaccination ;
- découvertes scientifiques ;
- bilans d'épidémies passées ou déjà sous contrôle.

Au grain document, il semble légitime de dire que ce genre de document n'est pas pertinent. En effet, ils n'apportent que peu d'informations nouvelles au veilleur. Pour autant, il ne s'agit pas de documents complètement erronés. Ces documents sont souvent reliés à des PML déjà détectées par ailleurs dans la même période. Au grain PML, ces documents ne créent donc pas un bruit véritablement problématique : le document assez pertinent est rattaché à une PML validée humainement. La classification binaire document pertinent ou non-pertinent n'est sans doute pas la plus adéquate pour juger la performance pour la tâche de veille. Il y a en effet différents niveaux de bruit.

Ceci amène à réexaminer l'annotation humaine avec cette question : peut-on améliorer la fiabilité de l'annotation humaine ? Nous avons donc évalué si un découpage en 3 catégories (très pertinent, assez pertinent ou non pertinent) était plus juste. De fait, l'accord inter–annotateurs (Fleiss Kappa) était sensiblement meilleur avec 3 catégories (0.81) qu'avec 2 (0.76). La classification binaire des documents semble donc difficile, y compris pour un humain. Cela tend à prouver l'existence d'une troisième classe, intermédiaire en pertinence : les documents « moyennement pertinents ».

Il serait intéressant de voir dans quelle mesure le jugement des annotateurs est influencé par ce qui a été annoté précédemment. Il est possible que lorsque la classe d'un document est douteuse, le fait de disposer d'autres documents sur le même évènement « rassurait » l'annotateur : s'il a déjà sélectionné des documents pertinents avec cette PML, il a pu sembler moins grave de considérer non-pertinent d'autres documents de facture proche traitant de la même PML. Dans notre corpus de référence, de nombreux documents, surtout en anglais, traitaient de la disparition de la poliomyélite en Inde. L'accord entre DAnIEL et les annotateurs sur ces documents était plus faible que sur le reste du corpus. Les annotateurs ont globalement jugé que c'était un évènement sans pour autant être tout à fait d'accord sur la pertinence des différents documents qui le traitent. Il est possible que les annotateurs ait eu une lecture des *guidelines* orientée « corpus » plutôt que document. Une fois qu'une PML a été détecté dans de nombreux documents,

les annotateurs deviennent sans doute plus strict avant d'étiqueter de nouveaux documents avec cette PML. Nous ne disposons pas de données pour évaluer plus finement ce problème d'annotation. Le choix de simplifier la tâche d'annotation a permis d'obtenir une large couverture. La charge d'annotation s'est retrouvée allégée au prix d'une plus grande latitude laissée aux annotateurs. Il aurait été intéressant de mesurer l'accord intra-annotateurs pour estimer le degré de récurrence des choix des annotateurs sur les mêmes documents. Un processus d'annotation « agile » ([Fort-2013]) aurait permis d'affiner les différences de jugement entre annotateurs sans contraindre outre mesure la tâche d'annotation. La durée de la campagne d'annotation (plusieurs mois) explique sans doute aussi ces variations : les annotateurs se sont approprié les règles d'annotation et ont exprimé un jugement sur la tâche elle-même.

Dans le cadre de notre tâche, il est apparu qu'évaluer l'efficacité d'un système par le nombre de documents n'était pas toujours juste. En effet, analyser correctement 99 documents parlant plus ou moins du même évènement (Grippe en Espagne par exemple) tout en étant silencieux sur un évènement qui ne serait décrit que dans un document (Ebola au Congo) ne peut pas pour une autorité sanitaire mériter un score de 99%. DAnIEL a été évalué sur cet aspect. À cet effet, un recensement des PML uniques figurant dans notre corpus annoté a été effectué Ce comptage ne tient pas compte du nombre de documents dans lesquels ces paires sont trouvées. Il s'agit donc une évaluation par classes de PML, elle est destinée à mesurer quelles classes auraient dû être peuplées mais ne l'ont pas été.

Dans le Tableau 5.24 figurent les résultats de cette évaluation. De façon globale seuls 3 PML (sur 57) n'ont pas été détectées. Les résultats sont ici meilleurs que dans l'évaluation par document. DAnIEL tire parti de la dimension « corpus ». Le fait de traiter plusieurs langues est là aussi un avantage, puisque des PML non détectées dans une langue peuvent l'avoir été dans une autre ([Brixtel-2013]). Notons que le nombre total de PML uniques n'est pas la somme des PML existant dans chaque langue puisqu'une même PML peut avoir été détectée dans différentes langues.

	Total PML uniques	PML Extraites	PML Non-extraites
anglais	15	14	1 (6,6%)
chinois	5	5	0 (0%)
grec	17	17	0 (0%)
polonais	28	26	2 (7,1%)
russe	23	21	2 (8,6%)
Total	62	59	3 (5,2%)

Tableau 5.24 – Évaluation par PML

Synthèse sur DAnIEL

Nous avons présenté dans ce chapitre l'architecture générale de DAnIEL, notre système multilingue de veille épidémiologique. DAnIEL répond aux impératifs de factorisation maximale fixés, son noyau central d'analyse est indépendant de la langue traitée. La seule contrainte est la constitution d'une base de noms de maladies usuels et d'une base de noms de lieux. Pour les langues de notre étude ces listes ont été constituées automatiquement mais la faible taille requise (200 termes suffisent pour avoir de très bons résultats) rend une constitution manuelle tout à fait réaliste.

Les résultats bruts en terme de filtrage et de typage des documents montrent que DAnIEL offre un bon compromis coût–efficacité. En effet, s'il semble moins efficace que des approches plus classiques lorsque l'on traite l'anglais, il présente une grande fiabilité sur d'autres langues peu ou pas traitées par ces systèmes. Par ailleurs, l'évaluation par Paire Maladie-Lieu (PML) a montré l'importante plus-value offerte par la couverture multilingue de DAnIEL. De nombreuses PML ne sont détectables que sur peu de documents et dans des langues rarement traitées. Une plus grande couverture en terme de langues s'est donc traduite par une meilleure couverture en terme de faits épidémiologiques détectés.

Enfin, nous avons évalué le jugement des annotateurs sur des évènements extraits par DAnIEL. Nous avons constaté que DAnIEL détecte très peu d'évènements erronés. Ses erreurs proviennent le plus souvent d'évènements secondaires. Le bruit généré par DAnIEL est donc assez limité. De plus, le système offre un excellent rappel, ce qui est capital en veille épidémiologique. Ce rappel implique un faible un coût au niveau de la précision : la mission de filtrage de DAnIEL est donc remplie. Nos expériences sur les PML ont par ailleurs montré que les performances du système dans la tâche de catégorisation étaient elles aussi très bonnes.

Chapitre 6

Variations sur le genre

Sommaire

Afin de valider notre approche fondée sur le genre, il nous a paru nécessaire de nous pencher sur d'autres genres textuels. En effet, nous avons montré dans le chapitre précédent qu'un modèle de document adapté au genre permettait de diminuer considérablement le coût marginal de traitement de nouvelles langues. Une question naturelle est alors de vérifier dans quelle mesure ceci est vrai pour de nouveaux genres. Nous montrons dans ce chapitre qu'avec notre méthode, la variété en genre a plus d'impact que la variété en langue. C'est en effet l'adaptation du modèle de document à un nouveau genre qui est la plus coûteuse : il faut disposer d'un modèle de document adapté au genre étudié.

Nous présentons deux applications de notre méthode dans le cadre de deux campagnes du DÉfi Fouille de Textes (DEFT) : Appariement de résumés et d'articles scientifiques (Section 6.1) et Indexation libre et contrôlée d'articles scientifiques (Section 6.2). Les travaux menés dans le cadre du défi sont le fruit de collaborations actives avec des chercheurs

de l'équipe *HULTECH* (Romain Brixtel, Emmanuel Giguet, Nadine Lucas et Gaël Dias) ainsi que des étudiants du master Langue Image et Documents (Gaëlle Doualan et Mathieu Boucher) de l'Université de Caen. Le « nous » utilisé dans ce chapitre possède donc une valeur collective.

6.1 Appariement de résumés et d'articles scientifiques

Une des tâches du « DÉfi Fouille de Textes 2011 » ([Grouin-2011]) consistait à apparier des résumés avec l'article scientifique pour lequel ils ont été rédigés [63]. Ces articles provenaient de revues de sciences humaines publiées sur la plateforme « Érudit » [64]. Deux pistes étaient proposées. Dans la première, les articles étaient complets, dans la seconde ils étaient tronqués ; l'introduction et la conclusion avaient été retirées. Un premier jeu de données, ou corpus d'apprentissage, était fourni aux participants dès l'inscription. Le second jeu de données, ou corpus de test, a été rendu disponible plusieurs semaines plus tard, chaque équipe disposant de trois jours pour communiquer ses résultats. Nous avons conçu notre système sur des principes les plus simples possibles. L'idée était de mesurer l'adaptabilité de notre méthode à une nouvelle tâche. Nous avions également comme objectif de pouvoir nous adapter au traitement d'une nouvelle langue, sans intervention extérieure ([Lejeune-2011]).

6.1.1 Principes mis en œuvre

Nous avons choisi un appariement séquentiel, sans remise en cause des appariements précédemment effectués. L'hypothèse sous-jacente est qu'un résumé et un article sont en quelque sorte indissociables. Il n'est pas nécessaire d'envisager une quelconque ambiguïté d'appariement entre un article et plusieurs résumés ou entre plusieurs articles et un résumé.

Dans cette même recherche d'une solution simple, d'un point de vue calculatoire, nous avons considéré qu'il était plus efficace de rechercher pour chaque document son résumé, plutôt que de rechercher pour chaque résumé son document. L'espace de recherche de la collection de résumés est en effet plus petit que l'espace de recherche de la collection d'articles. Par ailleurs, nous supposons qu'un article contient toutes les informations importantes disponibles dans le résumé, alors que l'inverse n'est pas vrai. Chercher à quel article correspond tel résumé serait alors potentiellement générateur d'« ambiguïtés », artificiellement engendrées par la démarche.

Notre approche prend le contrepied des applications de recherche d'information classique, où le résumé serait envisagé comme la requête posée à un moteur travaillant sur la collection d'articles indexés. Cette approche aurait été pertinente en terme de réutilisation

63. L'autre tâche du défi proposait une étude diachronique d'un corpus journalistique : dater des articles à partir d'extraits de 300 ou 500 mots.

64. `www.erudit.org/`

de technologies disponibles mais, peut être plus discutable en terme d'adéquation au problème. Ce choix d'utilisation de techniques proches de l'état de l'art a été fait par d'autres équipes qui ont proposé des approches basées sur l'utilisation de techniques classiques de recherche d'information ([Bestgen-2011]), ou d'apprentissage ([Raymond-2011]).

Nous n'avons pas non plus cherché a pondérer les fréquences des éléments recherchés en fonction de la collection, par exemple par une métrique de type *TF-IDF*. L'approche la moins coûteuse en calcul a été utilisée. Elle consiste à considérer que la simple cooccurrence de séquences communes au résumé et à l'article constitue un indice de corrélation suffisamment fiable pour un appariement de qualité. Nous montrons que cet indice est d'autant plus fiable qu'il est cohérent avec les positions définies dans le modèle d'article attendu.

Du point de vue linguistique, nous avons adopté une tripartition des articles : introduction, développement et conclusion. La mise en œuvre informatique consiste à calculer cette tripartition. La segmentation est déduite de la structure physique dont la trace se manifeste par la présence d'éléments *XML* « titre », qu'il s'agisse de titre ou de sous-titre. La segmentation ne repose donc pas sur la recherche de mots-clés comme « introduction », « conclusion » qui induirait une dépendance à la langue [65] ou aux variations de libellé, le titre « discussion » pouvant annoncer la conclusion.

Le premier segment, du début du texte à la première balise titre est associé à l'introduction ; le dernier segment, de la dernière balise titre à la fin du texte, est associé à la conclusion. Le reste est associé au développement, le corps du texte. Cette mise en œuvre simple part de l'hypothèse que l'introduction et la conclusion sont ordinairement non découpées en sous-sections, contrairement au développement de l'article.

Nous supposons que le résumé contient des reprises à l'identique de l'introduction, et que le contexte, la thématique et les perspectives sont des points communs que partage le résumé avec le couple introduction–conclusion. De fait, l'implantation réalisée traduit ces hypothèses. Elle tient compte de deux critères :

1. la recherche de la plus longue séquence de caractères répétée dans l'article et partagée avec un seul des résumés ;

2. la plus forte corrélation en terme de séquences de caractères partagées entre l'article et un seul résumé.

La plus longue chaîne (critère 1) est attendue dans l'introduction et *hapax* dans la collection de résumés. Un *hapax* est un terme (ici, une chaîne de caractères) que l'on ne rencontre qu'une seule fois dans une collection donnée (ici la collection de résumés). Les séquences répétées uniques (critère 2) sont attendues principalement dans l'introduction et la conclusion.

65. Le corpus est monolingue mais nous avons souhaité que le système soit d'emblée conçu pour une utilisation multilingue.

6.1.2 Description de l'approche et terminologie

Nous définissons la finalité de l'appariement comme la formation de couples entre un résumé et l'article qu'il résume. Le fait que deux documents puissent constituer un couple résumé-article provient de certaines connections que l'on peut trouver entre eux. Nous avons nommé ces connections, ces points communs, des **affinités**. Nous considérons comme affinités les chaînes de caractères, mots ou non-mots, communes aux deux documents. Dans la terminologie utilisée par la suite, **chaque article est un célibataire qui possède un certain nombre de prétendants : les résumés**. Pour former des couples nous formulons une hypothèse contrastive : parmi une collection de résumés nous recherchons celui qui possède les meilleures affinités avec un article. La proximité entre un article et un résumé ne se juge donc pas localement, mais par rapport à la collection.

La tâche est redéfinie de la façon suivante : **rechercher, à partir d'un corpus de célibataires d'une part et d'un corpus de prétendants de l'autre, le plus de couples corrects résumé-article**. Le bon prétendant pour un célibataire donné sera le résumé qui partagera le plus grand nombre d'affinités avec cet article.

Nous aurions pu utiliser simplement des mots mais dans la lignée des principes décrits plus haut nous avons souhaité :

- ne pas nous baser sur des pré-traitements (lemmatisation par exemple) pour pouvoir effectuer certaines opérations (associer « traduction » et « traductions » par exemple) [66] ;
- favoriser, bien que le corpus soit finalement monolingue, une méthode qui soit facilement réutilisable pour des corpus multilingues.

Plus généralement, nous n'avons stocké aucune information à l'issue de la « phase d'apprentissage » [67], ni utilisé aucune ressource externe. La phase initiale nous a servi simplement à éprouver le système. Dans la même idée de généricité, pour faciliter le passage à l'échelle, nous n'avons pas souhaité utiliser les informations concernant la revue. Le système que nous présentons va donc chercher le résumé correspondant à un article donné parmi toute la collection de prétendants sans pré-filtrage. Toutefois, et nous le verrons plus loin, une revue a posé plus de problèmes que les autres.

Nous cherchons ici, à nouveau, à valoriser les principes de généricité et de parcimonie. Si le but du concours était bien entendu de faire le meilleur score, nous nous sommes attachés dans notre démarche à ne pas créer un système trop complexe ou trop paramétrable. Au contraire, c'est le même système que nous avons fait fonctionner pour le *run* de référence de chaque piste. Enfin, notons qu'aucun nettoyage du code source *XML* n'a été pratiqué. Nous ne souhaitons pas effectuer des traitements, même les plus « naturels »,

66. La recherche de sous-chaînes peut être utilisée pour lemmatiser mais ici nous ne recherchons pas des lemmes. Toute chaîne de caractères présentant les caractéristiques recherchées est utilisée y compris si elle chevauche plusieurs mots, de la mise en forme ou de la ponctuation.

67. Nous mettons ici des guillemets car pour notre approche cette phase est uniquement une phase de calibrage, de développement.

s'ils ne sont pas justifiés. L'idée est là encore de s'appuyer sur l'implémentation la plus simple possible et de profiter des indices présents dans les documents. Les pré-traitements ont, de notre point de vue, une tendance à écraser les observables. Or, par exemple, nous n'avions aucune raison objective de considérer que les balises risquaient de détériorer les résultats. Aucun débalisage n'est donc effectué dans le processus.

6.1.3 Définition des affinités

Des segments présents dans l'article sont repris par l'auteur dans l'écriture du résumé. Selon les stratégies mises en place par l'auteur pour construire son résumé, la recopie pourra être plus ou moins prononcée. Cette recopie pourra être un mot, un groupe de mots voire une phrase ou une proposition. L'utilisation des chaînes de caractères répétées maximales ($rstr_{max}$) permet de repérer des unités qui se rapprochent dans une certaine mesure des unités multi-mots. L'intérêt de ces unités dans le cadre des tâches de recherche d'information est reconnu ([Doucet-2006b]).

Pour rechercher le résumé correspondant à un article nous recherchons des points d'ancrage. Ces points d'ancrage sont des $rstr_{max}$ entre un résumé R et un article A. Nous faisons l'hypothèse qu'il faut apparier R et A s'ils ont des segments en commun longs et nombreux : les **affinités**. Plus un couple (R, A) possède d'affinités, si possible de grande taille, plus il y a de chances qu'il constitue un appariement correct.

Les chaînes de caractères reflétant les affinités entre résumés et articles sont repérées à l'aide de l'implantation *Python* de Romain Brixtel déjà évoquée [68].

Grâce à cette implantation, en comparant un célibataire à tous ses prétendants nous obtenons une structure de données donnant pour chaque $rstr_{max}$, les documents du corpus dans lesquels elle apparaît. La fréquence de ces segments aux sein des documents ne nous intéresse pas ici ; c'est leur position qui est primordiale. D'autre part nous ne conservons que les affinités entre le célibataire et ses prétendants, les affinités entre prétendants ne sont pas prises en compte.

6.1.4 Filtrage des affinités

Deux mesures sont utilisées pour effectuer les appariements. Pour chaque article à apparier, on calcule pour tous les résumés disponibles :

1. la taille en caractères de la plus grande affinité (**affinité-max**)
2. le nombre total d'affinités hapax (**card-affinités**).

Pris isolément, le critère affinité-max n'est pas suffisamment fiable, mais il est complémentaire avec card-affinités. Le nombre total d'affinités peut quant à lui souffrir de la sur-représentation d'affinités peu significatives. En effet le nombre de $rstr_{max}$ pour un

68. `http://code.google.com/p/py-rstr-max/`

« resse » « ymbol » « ssib » « est_p » « ns_et » « la_mise_en » « s_donné » « ntifi »
« qu'elle » d'une_co » « e_ap » « les,_ » « s_qua » « ur_l'a » « lum » « ns_f »

FIGURE 6.1 – Exemples d'affinités banales, « _ » représente une espace typographique

document donné est potentiellement très élevé. Sur des documents en langue naturelle, ce nombre est quadratique en la taille des documents.

Nous ne conservons que les affinités *hapax* dans la collection de prétendants. Nous supposons que ce qui est rare peut avoir une grande valeur. En l'occurrence, n'est tenu compte d'une affinité entre un célibataire et un de ses prétendants que si cette affinité n'est pas partagée par d'autres prétendants. Ce critère d'exclusivité évite par ailleurs la sur-génération de motifs qui pourrait intervenir avec des $rstr_{max}$.

Si un article a une affinité commune avec plusieurs résumés, cette affinité n'est pas considérée comme significative. De cette façon, nous cherchons à ne pas tenir compte de celles pouvant être trop peu discriminantes. Notre hypothèse est qu'un couple réussi doit partager des affinités « originales ». En pratique, non seulement nous écartons des termes trop génériques ou trop largement distribués, mais nous filtrons aussi de nombreuses affinités qui tiennent plus de la langue concernée que du thème de l'article [69]. Nous en donnons quelques exemples dans la figure 6.1.

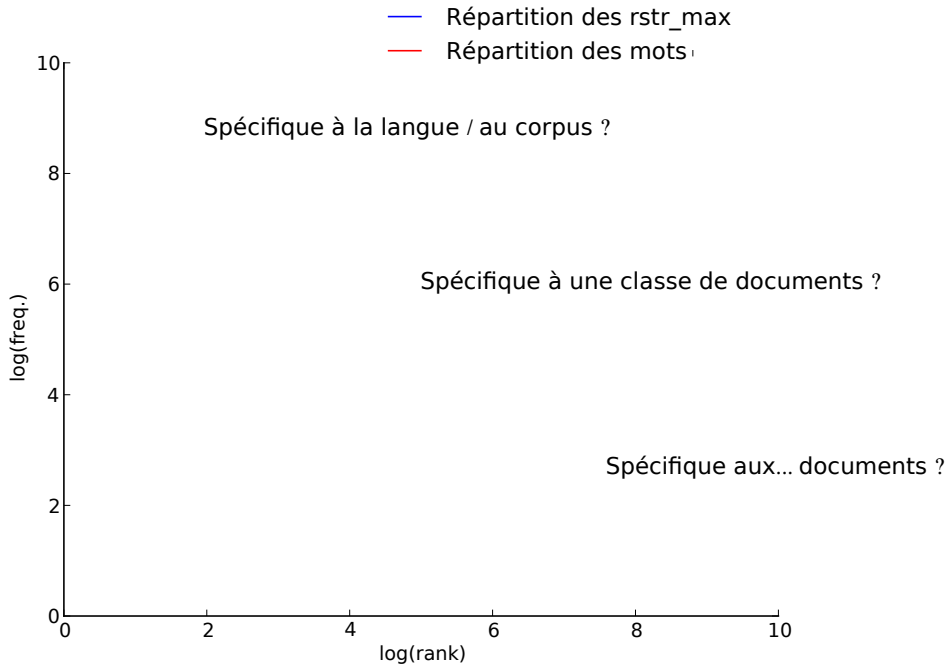


FIGURE 6.2 – Loi de Zipf sur des $rstr_{max}$ et sur des mots

69. Sans doute pourrait-on appeller ces affinités des « affinités vides ».

« a_philosophie_politique_d » « s_les_organisations » « r_la_reconnaissance_des » « des_organisations_internationales » « s_les_années_1970 » « établissements »

FIGURE 6.3 – Exemples d'affinités hapax, « _ » représente une espace typographique

Nous avons déjà remarqué que la loi de Zipf s'applique très bien aux $rstr_{max}$ (section 4.4). Nous faisons l'hypothèse qu'il est possible de caractériser des chaînes de caractères « vides ». Nous utilisons ici le terme « vide » dans la même acception que celle qu'il possède dans l'opposition mot-plein/mot-vide ou terme-plein/terme-vide. En haut à gauche de la courbe de la figure 6.2 (page 134) nous trouvons des $rstr_{max}$ très courtes et très fréquentes : des affixes, des suites de caractères typiques de la langue et des mots courts.

Au contraire, les affinités rares, particulièrement les *hapax*, sont des chaînes plus facilement reliées à un sens. Le but n'est pas de manipuler des unités interprétables, toutefois cette propriété c'est avérée utile pour analyser des erreurs du système. Cette observation nous a semblé être un pas intéressant vers la validation de notre hypothèse : les couples obtenus grâce aux affinités hapax semblent formés pour de « bonnes » raisons (Figure 6.3).

La figure 6.4 illustre l'importance de l'utilisation du critère de fréquence. La fréquence 2 en abscisse indique la présence de l'affinité dans l'article et dans un seul résumé, on a donc une affinité qui est hapax dans la collection de résumés. Nous observons sur cette courbe que lorsque l'on relâche la contrainte d'effectif, les résultats chutent. Par exemple, tenir compte des affinités présentes dans 2 résumés (fréquence 3) fait passer les résultats sur le corpus d'entraînement de 0.97 à 0.78.

Les chaînes de caractères utilisées comme affinités ne signifient rien en elles-mêmes pour notre système. C'est au niveau de l'improbabilité relative de la présence d'une chaîne répétée que se situe le critère de décision. La figure 6.5 montre l'existence d'un réglage optimal de la taille minimale des affinités prises en compte. Fixer cette taille entre 7 et 12 caractères permet d'optimiser les résultats. En dessous de cet intervalle le score reste constamment supérieur à 0.9. Le filtrage par les hapax offre donc de bons résultats sans paramétrage ; l'ajustement du seuil de longueur permet simplement une optimisation pour s'approcher d'un score de 100%. Par contre au delà de l'intervalle $[7; 12]$, et spécialement à partir d'un seuil de 20 caractères, les résultats sont en chute libre. Le nombre d'affinités à observer est trop faible pour former un nombre suffisant de couples.

6.1.5 Fonctionnement et résultats

Prenons en entrée la liste des articles et des résumés à apparier. Chaque célibataire (article à apparier) est comparé à tous ses prétendants (résumés à apparier). L'implémentation py-rstr-max calcule les chaînes de caractères répétées maximales ($rstr_{max}$). Seules les affinités *hapax* dans le corpus de prétendants et de taille supérieure à 8 sont conservées. Ce seuil a été fixé pour le français de manière empirique. Il a été validé par le calcul mais

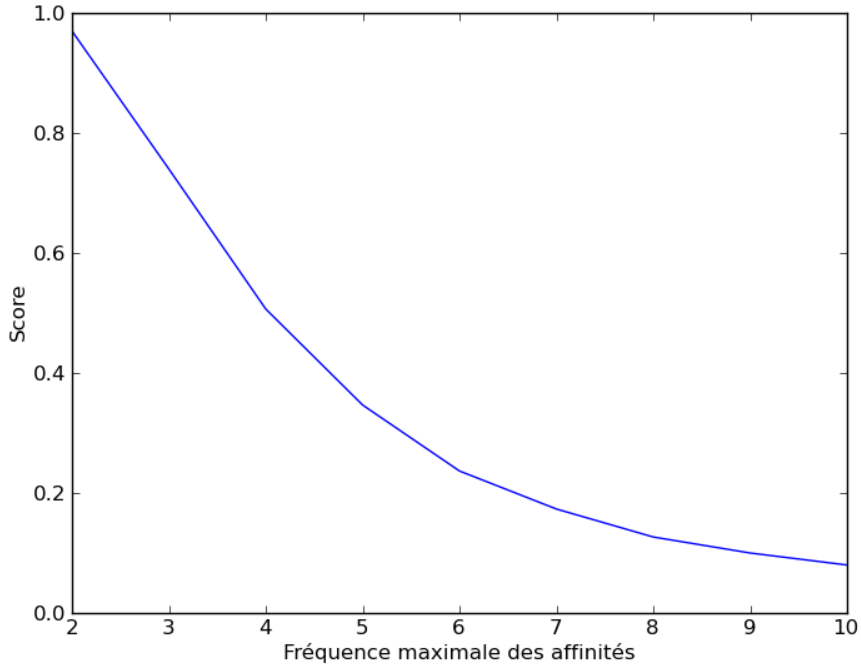


FIGURE 6.4 – Évolution de la proportion de bons appariements selon la fréquence maximale des affinités prises en compte

pourrait être recalculé pour chaque collection quelle que soit la langue. L'absence de seuil donnerait toutefois des résultats tout à fait corrects, comme le montre la Figure 6.5.

Description locale

Chaque article est comparé à la collection de résumés, un couple est formé chaque fois qu'il est significativement relié par des affinités. Pour ce faire il faut qu'ils partagent l'**affinité-max** trouvée dans la collection de prétendants et un nombre significatif d'affinités hapax (**card-affinités**). Les prétendants sont classés par déciles de nombre d'affinités avec le célibataire.

Si un prétendant est unique dans un décile, si ce décile est isolé des autres (Figure 6.6), nous considérons que le prétendant se détache : un couple est formé. Le célibataire et le prétendant concernés ne sont alors plus confrontés aux autres.

Il convient de se demander dans quelle mesure le nombre d'affinités d'un prétendant et d'un célibataire est significatif par rapport au nombre que partageraient un autre couple potentiel. Nous avons remarqué en comparant l'ensemble des affinités d'un couple correct (A_1, R_1) avec celles de tous les autres couples possibles A_i, R_i (figure 6.6) que les affinités hapax étaient réparties en trois tiers globalement équivalents en terme d'effectifs :

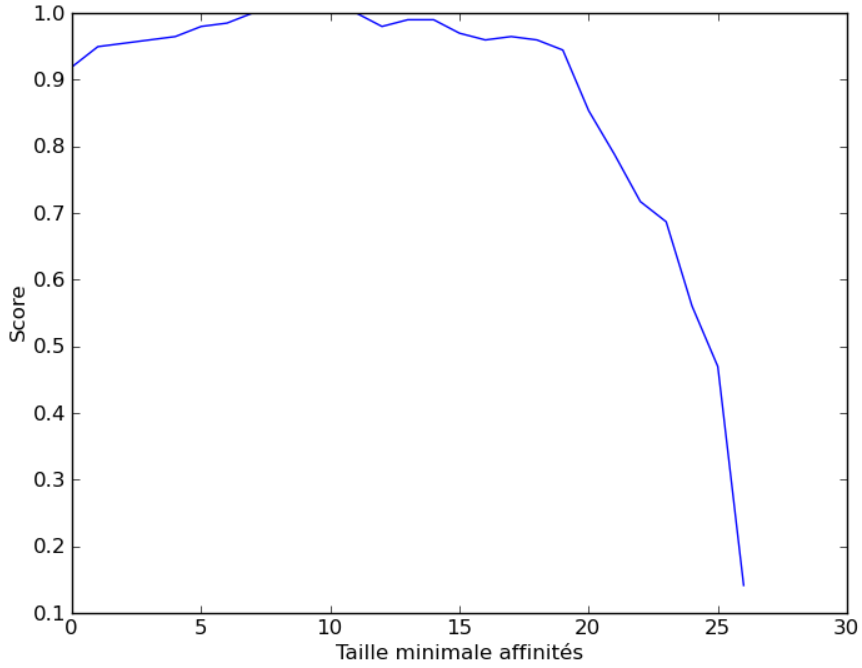


FIGURE 6.5 – Évolution de la proportion de bons appariements selon la taille minimale des affinités prises en compte

- des "affinités vides" non filtrées par le critère d'exclusivité ;
- des affinités peu significatives et non discriminantes, très proches d'un document à l'autre ;
- des affinités pouvant décrire les centres d'intérêt du célibataire, les thématiques spécifiques de l'article.

Nous avons observé que le nombre d'affinités existant entre un article et son résumé était 1,5 fois supérieur à celui des autres prétendants. Le prétendant est donc unique dans son décile et son décile est détaché des autres (Figure 6.6). Quand ce critère n'est pas respecté nous considérons qu'il y a jalousie potentielle : aucun prétendant ne se détache, il existe donc un risque d'erreur dans la constitution du couple. Le célibataire sera alors laissé de côté et attendra une phase ultérieure pour être apparié.

L'ordre dans lequel nous traitons les célibataires pourrait avoir une influence sur le résultat. Le tableau 6.1 montre différents résultats selon l'ordre de tirage des célibataires. Varier les ordres de tirage des célibataires sur le corpus d'entraînement amène des différences peu significatives. L'ordre d'apparition n'amène des changements que dans les dernières phases d'appariement, lorsqu'il ne reste que peu de documents à coupler.

Nous considérons que lorsqu'un couple est formé, le prétendant ne doit plus être présenté aux autres célibataires. De cette façon, pour les célibataires ayant du mal à trouver

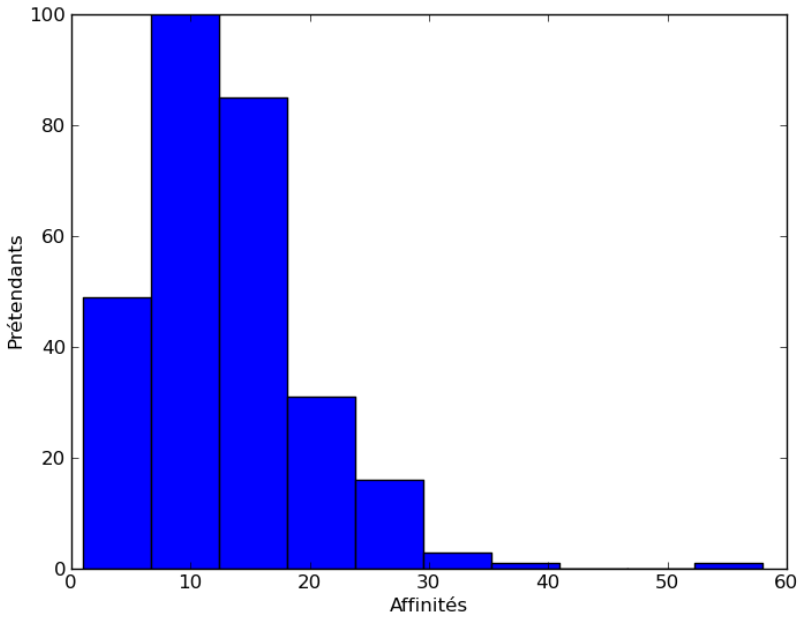


FIGURE 6.6 – Classement des prétendants par déciles de nombre d'affinités. Un prétendant se détache

Run	1	2	3	4	5	6	7	8	9	10
Score	0.97	0.967	0.97	0.97	0.973	0.97	0.967	0.967	0.97	0.963

Tableau 6.1 – Corpus d'entrainement, tirage aléatoire de l'ordre d'apparition des célibataires dans la boucle

leur prétendant, la tâche est facilitée : le nombre d'affinités hapax est plus grand, le critère devient plus discriminant tout en restant très efficace. Tant qu'il reste des célibataires, ils sont comparés aux prétendants disponibles. La constitution de tous les couples nécessite 5 à 6 phases. Les erreurs d'appariement ont deux origines. La première est constitué par des appariements erronés dans la phase précédente (cas rare). La seconde est la faiblesse du nombre d'affinités disponibles, qui rend les derniers appariements moins convaincants (cas le plus fréquent).

Nos différents tests ont montré que quels que soient les jeux de données, la première phase permet d'apparier 80% des célibataires avec une précision supérieure à 99%. Si deux phases consécutives n'amènent pas d'appariements, le système s'arrête de manière à éviter des appariements « par défaut ».

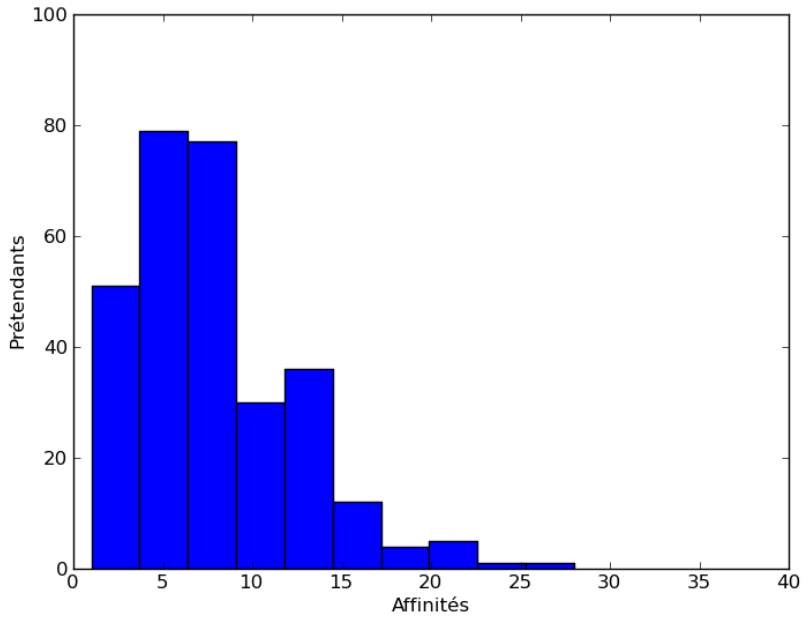


FIGURE 6.7 – Classement des prétendants par déciles de nombre d'affinités. Aucun prétendant ne se détache

6.1.6 Résultats

Nous montrons ici les résultats obtenus sur le corpus d'apprentissage fourni pour le DEFT, sur le corpus de test qui a servi au classement des systèmes en compétition et sur la concaténation des deux corpus. Le système est rigoureusement le même pour chacune des pistes et pour chacun des corpus. Nous remarquons que les résultats obtenus sur les articles tronqués sont moins bons. C'était attendu puisque le phénomène de recopie que nous cherchons est moins visible dans le développement que dans l'introduction et la conclusion. Nous n'avons pas souhaité concevoir un système différent selon les jeux de données, de ce fait le modèle de document attendu (avec introduction et conclusion) détériore quelque peu les résultats dans la piste 2.

Nous avons défini une *baseline* qui consiste à former des couples uniquement selon le critère affinité-max. Ses résultats sont faibles. Nous constatons avec intérêt la complémentarité des critères affinité-max et card-affinités. Le critère affinité-max permet d'éviter quelques mauvaises décisions, basées uniquement sur card-affinités (Tableau 6.2).

Le tableau 6.3 montre les résultats par corpus. Il est intéressant de noter que nos résultats sur le corpus de test sont supérieurs à ce que nous escomptions sur la piste 1 à l'issue de la phase d'apprentissage/calibrage. Nous avons donc fait un test en combinant les deux corpus. Remarquons que le résultat de ce test (corpus combinés) n'est pas la simple moyenne des résultats des corpus pris isolément.

Critère	Affinité-max	Card-affinités	Combinaison
Piste 1 : Articles complets	0.626	0.975	1
Piste 2 : Articles tronqués	0.48	0.934	0.959

Tableau 6.2 – Résultats obtenus sur le corpus de test, score selon les critères utilisés

	Corpus d'apprentissage	Corpus de test	Corpus concaténés
Piste 1 : Article-résumé	0.97	1	0.978
Piste 2 : Texte-résumé	0.96	0.959	0.96

Tableau 6.3 – Résultats selon les corpus

Le système n'utilise pas l'information sur la revue, mais les résultats n'en souffrent pas. Un article de la revue X est toujours apparié à un résumé de la même revue. Le système est donc indépendant de la revue. Il est toutefois intéressant de noter qu'une des revues, *Meta*, a concentré la très grande majorité des erreurs d'appariements, cela quels que soient les jeux de données. Le faible nombre d'affinités hapax rencontrés dans les articles de cette revue a été un facteur déterminant. La distribution des séquences utilisées a semblé plus homogène dans les différents articles issus de cette revue et a fortement nui aux appariements. La piètre qualité de la transformation de documents au format *PDF* en documents au format *XML* en constitue l'explication. Des sauts de ligne apparaissaient au milieu de phrases, gênant ainsi la détection de longues séquences, capitale dans notre méthode.

Les erreurs les plus fréquentes sur la seconde tâche provenaient là aussi du plus faible nombre d'affinités détectées par le système : sans l'introduction et la conclusion, le nombre d'affinités *hapax* diminue (Tableau 6.4). La différence entre un bon couple et un couple erroné tend à s'estomper et la qualité des résultats s'en ressent. Comme nous l'avons évoqué, quelques paramètres bien choisis auraient sans doute permis d'atteindre un meilleur score dans cette piste mais nous avons voulu garder la simplicité et la reproductibilité comme objectifs primordiaux. Par ailleurs cette chute des résultats corrobore l'hypothèse des linguistes : les débuts et fins de segments sont en soi intéressants à exploiter, à différents niveaux de granularité ([Lucas-2009b]).

ID Résumés corpus de test	013.res	066.res	073.res	154.res	155.res
Card-affinités du bon couple (article complet)	58	54	76	71	49
Card-affinités du bon couple (article tronqué)	42	42	49	33	37
Évolution du nombre d'affinités	-28%	-22%	-36%	-54%	-24%

Tableau 6.4 – Nombre d'affinités du bon couple résumé-célibataire selon que l'article est complet ou tronqué

6.1.7 Robustesse au changement de langue

Nous supposons qu'une approche fondée sur le modèle et utilisant des chaînes de caractères permet de traiter de nouvelles langues efficacement et à faible coût. Pour valider cette hypothèse, nous avons constitué un corpus d'articles scientifiques en polonais. Il nous a semblé en effet pertinent de traiter une langue à morphologie plus riche, supposée plus difficile à traiter au grain mot du fait de la moindre disponibilité d'analyseurs et autres ressources externes. Ce corpus est constitué de documents publiés par la revue de sciences humaines *Kultura i Historia*[70]. Nous avons conservé les documents qui comportait un résumé identifiable (en polonais : *Abstrakt*), nous avons ainsi obtenu 106 documents. Les résumés ont été séparés des articles. Nous avons utilisé le même système pour effectuer les appariements.

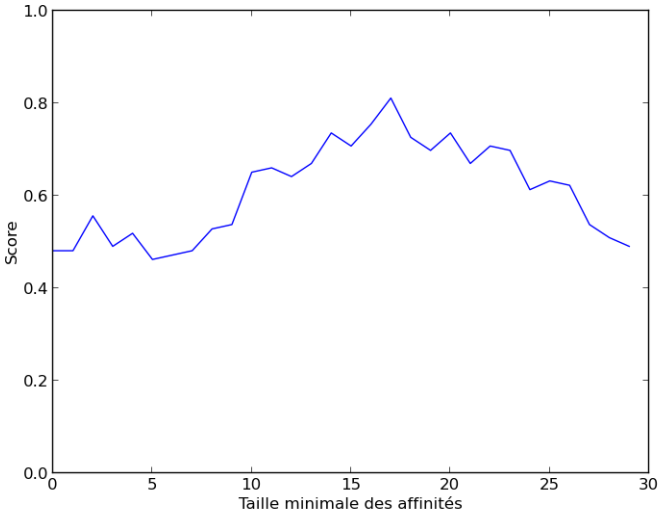


FIGURE 6.8 – Évolution de la proportion de bons appariements sur le corpus polonais selon la taille minimale des affinités prises en compte, utilisation du seul critère affinité-max.

La figure 6.8 présente les résultats obtenus, en utilisant seulement le critère affinité-max. Nous observons que les résultats sont faibles, mais comparables avec ceux obtenus avec le même critère sur le corpus en français proposé dans le DEFT 2011. Sur la figure 6.9 (page 142) nous pouvons voir les résultats obtenus cette fois avec le critère card-affinités seul. Pour un seuil de taille minimale des affinités fixé entre 11 et 13, les résultats sont supérieurs à 90%. Par contre, les résultats sont très faibles lorsque ce seuil est inférieur à 5. Enfin, dans la figure 6.10 (page 142) sont présentés les résultats obtenus en combinant les deux critères card-affinités et affinité-max. La mesure affinité-max est plus efficace en polonais qu'en français. Pour certaines tailles d'affinités elle donne de meilleurs résultats

70. `http://www.kulturaihistoria.umcs.lublin.pl/`

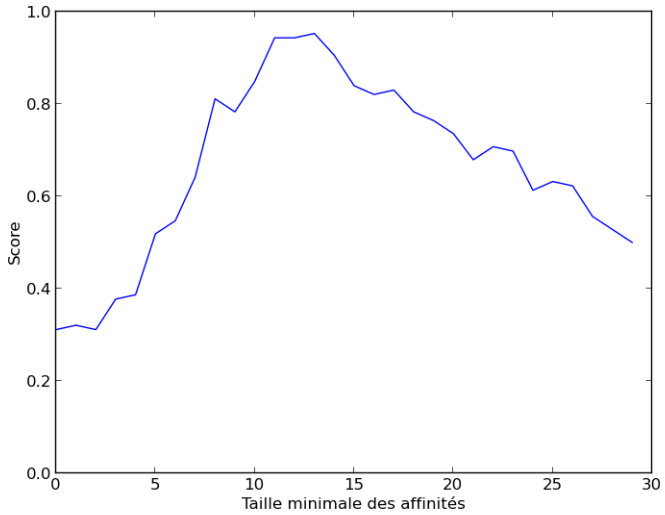


FIGURE 6.9 – Évolution de la proportion de bons appariements sur le corpus polonais selon la taille minimale des affinités prises en compte, utilisation du seul critère card-affinités.

que card-affinités. Comme dans le cas du corpus en français (figure 6.5) les deux mesures sont complémentaires. La méthode a permis d'obtenir jusqu'à 96% d'appariements corrects sans qu'aucune phase d'apprentissage ni aucun paramêtrage ne soient nécessaires.

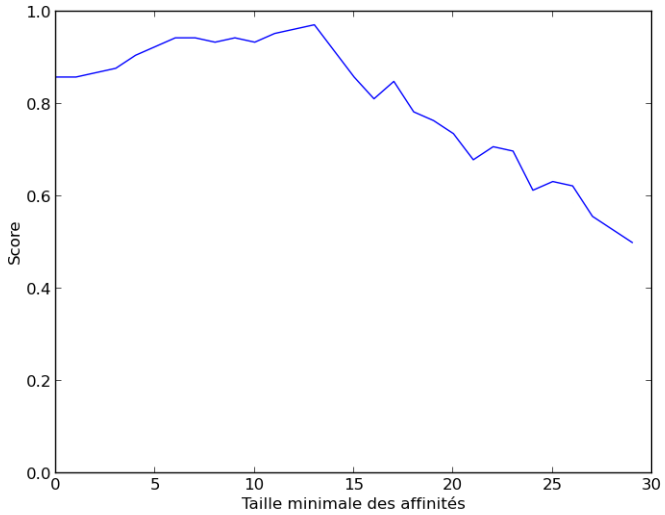


FIGURE 6.10 – Évolution de la proportion de bons appariements sur le corpus polonais selon la taille minimale des affinités prises en compte sur le corpus polonais, utilisation combinée des critères card-affinités et affinité-max.

6.1.8 Synthèse sur l'appariement résumé-article

Nous avons développé pour cette expérience une méthode d'appariements d'articles scientifiques et de leurs résumés basée sur des distributions de chaînes de caractères. Cette méthode a eu de très bons résultats sur la piste qui concernait les articles complets. Le phénomène de recopie que nous cherchions à utiliser était par contre moins prégnant sur les documents tronqués de la seconde piste, ce qui corrobore notre hypothèse fondée sur le distributionnalisme linguistique. Il nous semble que ces résultats apportent une pierre à l'édification de modèles alternatifs au « tout interprétable ».

L'utilisation des chaînes de caractères mots ou non-mots revient à considérer l'espace typographique comme un caractère comme les autres et non comme la frontière délimitant une unité d'analyse immuable.

Notre participation a porté sur la tâche 2, consistant à rapprocher un article de son résumé, et non sur la tâche 1, consistant à dater un extrait d'article. Nous disposions d'un modèle de structuration des articles académiques suite aux travaux de Lucas ([Lucas-2004]), et d'un modèle de structuration des articles de presse ([Giguet-2004]) dans la lignée des travaux de Van Dijk ([VanDijk-1988]). Mais aucun de ces modèles n'étaient adapté aux documents fournis pour la tâche 1 du fait du caractère tronqué (morceaux de 300 ou 500 mots) de ces documents.

De notre point de vue, les bons résultats obtenus dans cette campagne constituent une validation de notre approche. Nous avons combiné un modèle de document adapté au genre et une analyse au grain caractère. Cette méthode s'est révélée aussi efficace que des méthodes basées sur des techniques de l'état de l'art. La plus-value apportée se décline en deux points.

Le premier est la parcimonie dans l'utilisation des ressources externes. Hormis le modèle de document utilisé, aucune connaissance spécifique n'est utilisée. Le corpus d'apprentissage a été exploité pour éprouver le système, nous n'en avons pas tiré de connaissances particulières stockées en mémoire.

Le second point, corollaire du premier, est que la méthode est nativement adaptée au traitement multilingue. Si le modèle de document utilisé est dépendant du genre, le grain d'analyse est indépendant de la langue. Nous avons ainsi montré que la méthode utilisée s'était également montré très efficace sur des documents en polonais, sans qu'aucune modification n'ait été nécessaire.

Nous avons montré que la variété en langue était moins importante que la variété en genre : un même modèle est utilisé pour plusieurs langues sans modification mais il faut l'adapter lorsque l'on change de genre en passant d'articles de journaux à des articles académiques.

Nous présentons dans la section suivante d'autres expériences menées avec la même approche mais pour une tâche différente : l'extraction de mots-clés. Le genre textuel étant toujours l'article académique, nous montrons que pour notre méthode la variation de la tâche est secondaire.

6.2 Extraction de mots-clés

La tâche proposée dans le cadre du Défi Fouille de Textes 2012 consiste à retrouver dans des articles de sciences humaines les mots-clés proposés par les auteurs. Le corpus de travail est scindé en deux pistes, la première comportant 140 articles et la seconde 141. Une terminologie qui regroupe tous les mots-clés des articles est proposée avec la première piste. Nous proposons ici une méthode qui utilise des informations de la terminologie et une autre qui ne les utilise pas. Nous nous appuyons dans cette édition du DEFT sur l'algorithme de recherche de chaînes répétées maximales $rstr_{max}$.

Dans la section 6.2.1, nous procédons à une analyse du corpus qui nous permet d'appréhender le matériau sur lequel nous travaillons. Dans la section 6.2.2, nous détaillons la méthode utilisée pour l'approche au grain caractère puis dans la section 6.2.3 celle utilisée pour l'approche au grain mot. Ensuite, nous présentons les résultats et des éléments de discussion dans la section 6.2.4.

6.2.1 Description du corpus

Le corpus utilisé comporte des articles de sciences humaines provenant de quatre revues diffusées sur le site Érudit [71]. Nous présentons ici plus précisément les articles à traiter et les mots-clés qui leur sont associés.

Le corpus DEFT 2012 est constitué de 300 articles répartis sur 4 revues de sciences humaines :

- *Anthropologie et Société* (AS) ;
- *Revue des Sciences de l'Education* (RSE) ;
- *Traduction, Terminologie et Rédaction* (TTR) ;
- *Méta*, journal des traducteurs (META).

Configuration des articles

Les articles sont au format *XML*. Ils sont constitués d'un identifiant, de la liste des mots-clés fournis par l'auteur, d'un résumé et du corps de l'article lui-même. Le nom de la revue n'apparaît pas dans le contenu *XML*, mais dans le nom du fichier. De même, le nom de l'auteur et le titre de l'article ne figurent pas dans le fichier *XML*. Ceci a rendu plus complexe la recherche des mots-clés du fait notamment que le nom de l'auteur figurait systématiquement parmi les mots-clés des articles de la revue *Anthropologie et Société*.

Nous présentons dans la figure 6.11 un exemple d'article du corpus, afin de montrer la configuration et la structure des documents proposés. Notons que les titres et sous-titres des articles n'étaient pas disponibles. Nous pouvons formuler pour ce corpus les mêmes préoccupations sur la structure des documents « nettoyés » que celles exprimées dans la section 6.1.8 pour ce qui était du corpus du défi 2011. La composition du corpus est décrite

71. `http://www.erudit.org`

<?XML version="1.0" encoding="UTF-8" ?>
-<doc id="0001">
-<motscles>
<nombre>4</nombre>
<mots>Labrecque ;économie politique ;féminisme ;ethnographie</mots>
</motscles>
-<article>
-<resume>
Tout en poursuivant l'objectif de la présentation du numéro,
...
la consolidation de la théorie.
</resume>
-<corps>
Qui sape l'ethnographie ébranle la théorie
...
d'une anthropologie engagée, d'autre part.
</corps>
</article>
</doc>

FIGURE 6.11 – Un exemple d'article du jeu d'entraînement

dans le tableau 6.5. Le nombre moyen de paragraphes ne varie pas particulièrement en fonction de la revue, à l'exception de certains articles de *META*, pours lequels le découpage en paragraphes était de mauvaise qualité.

	Nombre de documents	Taille moyenne en paragraphes	Taille moyenne en caractères
Piste 1	94	67,8	41235
Piste 2	93	80,2	39153

Tableau 6.5 – Statistiques sur les documents du corpus d'évaluation

Les mots-clés

Les articles qui composent le corpus ne comportent pas le même nombre de mots-clés : en moyenne 5,4 sur la piste 2 et 5,7 sur la piste 1. Il y a de grandes disparités d'un texte à l'autre. Les article comportaient de 1 à 10 mots-clés. Nous avons noté que le premier mot-clé est systématiquement le nom de l'auteur de l'article pour la revue *Anthropologie et Société*.

Nature des mots-clés

Nous proposons une classification des mots-clés proposés par les auteurs :

- Noms propres : nom de l'auteur (ex : Labrecque), auteur faisant l'objet de l'article (ex : Jack Kerouac), lieu géographique (ex : Japon) ;
- Noms communs : des noms communs seuls ou parfois accompagnés d'adjectifs, mais jamais de verbes ni d'adverbes (ex : féminisme, économie politique) ;
- Noms et compléments du noms : avec des motifs tels que celui-ci : N de art N (ex : traitement de l'information sociale) ;
- Noms coordonnés : par exemple traduction scientifique et technique.

Globalement, plus les mots-clés sont longs, moins ils sont présents tels quels dans le texte. Quand ils y figurent, ils sont néanmoins peu fréquents. Globalement 79% des mots-clés sont présents tels quels dans le corps du texte, 44,5% dans le résumé et 42% à la fois dans le corps et dans le résumé.

6.2.2 Une approche au grain caractère

Nous reprenons ici les principes de la méthode utilisée pour le DEFT 2011 ([Lejeune-2011]). Nous supposons que les segments communs entre le résumé et le reste du texte constituent des mots-clés pertinents. Pour sélectionner les mots-clés pertinents, nous nous fondons sur leur proximité avec des éléments de la terminologie. Chaque article est donc découpé en deux parties : résumé et corps de l'article.

Nous cherchons cette fois les affinités entre les positions importantes de l'article et les éléments de la terminologie. Nous effectuons l'opération inverse de celle utilisée dans le DEFT 2011. Dans le cadre du DEFT 2011 les affinités servaient à caractériser un lien entre le résumé et les parties importantes du document. Ici, c'est le lien existant entre le résumé et les parties importantes qui permet de considérer les affinités comme de bons mots-clés pour le document. Nous comparons les deux segments textuels (résumé et corps) et l'ensemble de la terminologie en une seule opération. Nous conservons les $rstr_{max}$ apparaissant dans ces deux segments et dans un élément de la terminologie. Seuls les motifs respectant un critère de longueur donné sont considérés comme pertinents. Pour tenir compte des variations morphologiques du français, nous avons fixé la proximité minimale entre un motif trouvé et un élément de la terminologie à 0.9. Autrement dit, un élément t de la terminologie est considéré comme mot-clé du texte s'il existe une chaîne c telle que :

- c est présente dans le résumé et dans le corps de l'article ;
- c est une sous chaîne de t ;
- $\frac{len(c)}{len(t)} \geq \frac{9}{10}$ avec len le nombre de caractères dans c et t.

Nous n'avons pas appliqué notre méthode à la seconde piste car la sélection de chaînes de caractères adaptées à l'évaluation était malaisée. Une fonction de lemmatisation était utilisée pour l'évaluation mais les participants n'y avaient pas accès. Nous avons préféré modifier le système aussi peu que possible. En effet, le seul pré-traitement effectué est le découpage en deux segments textuels (résumé et corps). Aucun outillage linguistique (lemmatisation, étiquetage...) n'est nécessaire. Par ailleurs, aucun post-traitement n'est effectué.

Nous comparons ces résultats à une *baseline* utilisant uniquement la mesure tf-idf et à une méthode plus complexe développée par un trinôme d'étudiants (section 6.2.3) dans le cadre de Travaux Pratiques de *Master* informatique que nous avons encadrés.

6.2.3 Approche au grain mot

Dans cette approche, un découpage classique en mots est utilisé. La détection des $rstr_{max}$ est utilisée en l'appliquant cette fois sur des mots, plutôt que sur des caractères. La méthode est conçue pour fonctionner en l'absence de terminologie de référence.

L'algorithme $rstr_{max}$ est appliqué à l'intégralité de l'article. Une liste de « chaînes de mots » répétées et maximales est obtenue. Un grand nombre de motifs sont détectés, dont certains sont partiellement redondants. Par exemple, à partir des motif $ABCD$ et $BCDF$, on ne souhaiterait garder que la partie commune BCD. L'idée est d'optimiser le compromis rappel-précision, de diminuer le nombre de motifs tout en gardant les plus pertinents. Une application de $rstr_{max}$ sur les chaînes déjà extraites permet de d'écarter certaines de ces redondances. Pour chacune des chaînes ainsi extraites on calcule l'IDF (*Inverse Document Frequency*) :

$$TF \times IDF = \frac{freq(C,D)}{t(D)} \times -\log_2 \frac{nd(C)}{N}$$

Avec :

- freq(C,D) le nombre d'occurrences de la chaîne C dans le document D ;
- t(D) le nombre de mots du document D ;
- nd(C) le nombre de documents contenant C dans le corpus ;
- N la taille du corpus en documents.

L'importance des chaînes est ensuite pondérée selon les critères suivants :

- effectif de la chaîne dans l'article ;
- effectif de la chaîne dans le résumé ;
- longueur de la chaîne ;
- présence de la chaîne dans le premier paragraphe (*a priori* : introduction) ;
- présence de la chaîne dans la dernier paragraphe (*a priori* : conclusion).

À chacune de ces mesures est attribué un coefficient qui pondère leur importance. Nous avons effectué des statistiques sur le corpus afin d'anticiper les places occupées par les mots-clés dans les articles. Une chaîne qui est fréquente dans le résumé a davantage de chance d'être un mot-clé qu'une autre chaîne. Différents poids ont été expérimentés pour ces mesures en fonction de leur capacité à traduire le comportement des mots-clés. Notons que l'absence des titres dans les documents analysés rend difficile la détection des segments introductifs et conclusifs.

Les chaînes sont triées par ordre décroissant de poids, les sept premières chaînes sont sélectionnés comme mots-clés. Ce seuil a été fixé à partir des meilleurs résultats obtenus sur le corpus d'entraînement.

6.2.4 Résultats

	Résultat piste 1	Résultat piste 2
Approche 1 : $rstr_{max}$ au grain caractère	**0.44**, 3e/10	⊘
Approche 2 : $rstr_{max}$ au grain mot	0,12	**0,13**, 7e/9
Baseline : tf-idf simple	0,08	0,07

Tableau 6.6 – Résultats et rangs pour les deux approches et baseline

La première approche donne de bons résultats en raison de l'appui de la terminologie. Ces résultats sont logiquement meilleurs que ceux de l'approche par poids (Tableau 6.6). Sans doute ces résultats auraient pu être encore améliorés avec quelques heuristiques, par exemple : chercher à affecter chaque mot-clé de la terminologie à au moins un document. Nous avons considéré qu'utiliser ce genre de règles exploitait les biais de l'évaluation. Par ailleurs, nous n'avons pas souhaité complexifier la procédure utilisée. Nous constatons l'importance de la plus-value de l'approche par poids par rapport à la *baseline*. Il y a

aussi une réelle différence avec la seconde méthode que nous avons décrite. De notre point de vue cela justifie le choix de la simplicité dans les méthodes utilisées. Notre méthode basée sur la présence de $rstr_{max}$ à certaines positions a obtenu le troisième rang lors de cette campagne. Les deux méthodes qui ont obtenu les meilleurs résultats étaient fondées sur de l'apprentissage ([ElGhali-2012, Claveau-2012]). Au contraire, notre méthode 1 n'exploite pas de données autres que la terminologie. Ce résultat montre donc l'intérêt d'une démarche parcimonieuse.

Par ailleurs, nous avons cherché à utiliser une chaîne de traitement très proche de celle que nous avions conçu pour le DEFT 2011. Cette approche de réutilisation ne garantissait pas d'emblée de bons résultats comme l'ont montré Ahat *et al.*([Ahat-2012]). Nous considérons que cela traduit la robustesse de notre méthode : une tâche supplémentaire implique un coût marginal minimal dans la mesure où l'on travaille sur le même genre de texte.

Synthèse

Nous avons présenté deux évaluations de notre méthode fondée sur le genre textuel. Ces évaluations ont été effectuées à travers les compétitions de fouille de texte DEFT 2011 et 2012. C'est par la mise en œuvre la plus simple et la plus immédiate des deux caractéristiques de notre méthode, à savoir traitement au grain caractère et approche guidée par un modèle du genre textuel, que nous avons choisi d'aborder ces deux campagnes successives. Nous avons montré que notre modèle de structuration des articles scientifiques fournit de bons résultats d'une part et qu'il est robuste à la variation en tâche d'autre part. Les corpus du DEFT étaient composés uniquement de documents en français. Nous avons donc constitué un corpus de référence en polonais pour reproduire l'expérience du DEFT 2011 sur une autre langue. Ces expériences illustrent la pertinence de la méthode pour différentes tâches.

Chapitre 7

Variations sur le document

Sommaire

L'utilisation d'un modèle de document pour l'analyse automatique requiert une certaine qualité dans la structure des documents fournis en entrée du système. L'analyse des positions demande en effet autre chose qu'un texte réduit à une simple suite de phrases, dépossédé de certaines caractéristiques de mise en page essentielles à sa compréhension (présentes dans le code source sous la forme de balises notamment). La transformation du document brut en document interprétable par la machine est donc capitale. C'est une problématique fréquemment évoquée pour les documents au format *PDF*, pour pouvoir

utiliser des fonctions de recherche de motifs ([Benel-2012]) ou pouvoir accéder à la structure du document ([Doucet-2011]). Le format *HTML* pose lui aussi un certain nombre de problèmes. Le code source est généralement bruité par des éléments non-informatifs qu'il convient de détecter pour faciliter les analyses ultérieures. Cette opération porte parfois le nom de nettoyage de page Web (*Web Page Cleaning*) ou de détection de modèle de page (*Web Page Template Detection*). Le terme de nettoyage nous semble impropre et quelque peu réducteur vis-à-vis de l'importance de cette étape vis-à-vis des traitements ultérieurs. Nous utilisons pour notre part le terme de **détourage**, terme issu de la photographie. Ce terme désigne le fait de ne conserver qu'une partie d'une illustration pour une tâche bien précise. Le détourage de pages Web consiste à extraire le texte recherché à partir des données brutes, tout en conservant certaines données de structure (titraison, paragraphes). Cette opération est effectué par un **détoureur**.

Extraire le texte du code source (*HTML* ou autre) n'est pas une tâche triviale. En effet, obtenir automatiquement des corpus où l'on a à sa disposition le texte, tout le texte et rien que le texte impose une grande robustesse. Être en mesure de traiter un grand nombre de sources impose de tenir compte de la variabilité des feuilles de style. Cette variabilité rend les approches par règles inopérantes.

Nous proposons ici deux types d'évaluation pour le détourage. Tout d'abord, nous utilisons une **évaluation fondée sur le contenu** du texte détouré. Nous exploitons à cette fin les métriques de la compétition *Cleaneval*, en ajoutant une évaluation au grain caractère. D'autre part, nous proposons une **évaluation par la tâche** dont la problématique est la suivante : comment la qualité d'un détoureur peut-elle se mesurer en fonction des résultats obtenus par un système qui utilise les documents nettoyés ? Différents détoureurs sont utilisés pour l'évaluation, nous proposons par ailleurs plusieurs possibilités de combinaison entre ces détoureurs.

Dans les expériences précédentes, sur DAnIEL et sur le DEFT, nous étions en quelque sorte en **conditions de laboratoire**. Les textes comportaient les indications de mise en forme nécessaires à l'analyse et avaient une structure assez fiable. Il était dès lors possible de fonder l'analyse sur un certain nombre d'attendus, la qualité des textes était assez peu variable. Cette qualité avait été garantie par un post-traitement manuel des textes comme dans le corpus utilisé pour tester DAnIEL. Nous avons ainsi pu nous assurer que les documents détourés possédaient la même structure visuelle que les originaux et que l'intégralité des segments de texte était bien conservée. Pour les campagnes DEFT, les documents étaient issus d'une seule source et un détoureur spécialisé sur cette source avait été utilisé. Le rendu n'était pas parfait puisque des tableaux et des titres avaient disparus par exemple, mais la qualité était connue d'avance : tous les documents détourés avaient les mêmes « défauts ».

Dans la section 7.1 nous exposons la problématique du détourage. Puis nous détaillons les caractéristiques des différents détoureurs que nous comparons dans la section 7.2. Nous présentons ensuite dans la section 7.2.5 le corpus utilisé pour l'évaluation par la tâche. La

section 7.3 est consacrée aux modalités d'évaluation utilisées pour la campagne *Cleaneval*. L'évaluation des détoureurs est présentée dans la section 7.4.

7.1 La problématique du détourage des pages Web

Pour l'humain, destinataire supposé des documents publiés sur le Web, la détection du contenu purement textuel ne semble pas poser de difficulté. L'ergonomie visuelle des sites Web est de qualité variable, néanmoins nous pouvons remarquer que sur les sites de presse elle est suffisante pour assurer un accès rapide à l'information. Nous ne disons pas que l'ergonomie soit systématiquement optimale mais il apparaît qu'il y a finalement assez peu de variation sur la présentation des articles de presse eux-mêmes. Le contenu apparaît la plupart du temps au centre et le titre permet de fixer rapidement un point de départ pour la lecture. Au contraire, ce sont plutôt les pages de rubriques et la « Une » où les variations sont les plus fréquentes et les moins justifiées. Ce point de vue est également exprimé dans l'étude critique du nouveau modèle du *New York Times* menée par le *web-designer* Andy Ruledge [72].

Automatiser ce processus reste un défi à l'heure actuelle. Il est revenu au premier plan du fait de l'émergence de nouveaux supports ([Baluja-2006]) tels que les *smartphones* et autres tablettes. En effet, la diminution de la taille de l'écran (par rapport à un ordinateur de bureau « classique ») peut rendre la page difficilement lisible pour l'humain : le format de la page Web n'est plus adapté [73]. Deux approches peuvent être envisagées pour favoriser la lisibilité des sites Web :

- Une approche « côté serveur » qui consiste à prévoir une version spéciale du site Web adaptée au terminal avec lequel l'utilisateur se connecte ;
- Une approche « côté client » où c'est un programme dédié qui adapte automatiquement le format de la page pour le terminal concerné.

Les applications en tous genres disponibles pour les *smartphones* tiennent à la fois des deux approches : l'application est dédiée à la fois à un site et à un terminal. C'est le cas des applications de navigation sur *Facebook* et autres *Twitter* qui sont spécifiques à chaque smartphone [74]. Ce problème se pose pour la plupart des sites Web, l'approche dédiée n'est pas réaliste.

Nous nous concentrons spécifiquement sur le cas des sites où sont publiés des articles de presse. Nous évaluons les méthodes de détourage uniquement sur ce genre de documents. Il nous semble en effet que la spécialisation sur un genre textuel n'est pas gênante dans

72. `http://andyrutledge.com/news-redux.php` consulté le 12 octobre 2013

73. Une autre contrainte existe : avec l'essor du haut-débit, les pages Web sont plus gourmandes en bande passante, ce qui se révèle mal adapté aux réseaux mobiles. Nous ne traitons pas ce point ici bien que nous y voyions, là encore, un manque de parcimonie dans l'utilisation des ressources.

74. La petite taille de l'écran et la faible qualité de la bande passante rendant souvent impossible la navigation sur ces sites avec le navigateur Web du *smartphone*

la mesure où le nombre de genres textuels que l'on est amené à traiter est limité, à défaut d'être à proprement parler fini. L'indépendance vis-à-vis du genre textuel est toutefois un objectif pour certains chercheurs ([Kohlschutter-2010]).

Nous employons ici la terminologie la plus usitée dans le domaine pour désigner ce qui est constitutif du texte proprement dit de ce qui lui est extérieur. Dans cette terminologie, le **contenu informatif** du document (le texte lui-même) est opposé au **contenu non-informatif** (publicités, menus, liens externes...).

7.1.1 Différenciation du contenu informatif et du contenu non-informatif

La figure 7.1 (page 156) présente une page Web [75] avec une proposition de sélection d'éléments informatifs :

- en bleu, les segments faisant partie du contenu informatif : titre, chapeau, sous-titre et paragraphes ;
- en orange, les segments potentiellement intéressants : informations sur l'auteur, date de l'article et légende de la photographie.

Les autres éléments font partie du contenu non-informatif :

- menus de navigation (en haut) ;
- articles connexes (en bas) ;
- image (à droite) ;
- publicité (en bas à droite).

Nous avons donc une tripartition, avec des segments appartenant de façon certaine à l'une ou l'autre des classes et d'autres dont le classement est plus difficile à justifier. Les informations sur la date ou l'auteur peuvent être importantes pour l'interprétation du document. Toutefois, ces éléments sont souvent considérés comme non-informatifs. Par ailleurs, considérer que l'image ne fait pas partie du contenu informatique est lié à une approche purement littérale.

Une vision non-binaire a également été utilisée par les concepteurs de *Boilerpipe* pour la constitution de leur corpus de référence ([Kohlschutter-2010]). L'annotation comprenait les segments suivants par ordre décroissant d'informativité ; entre parenthèses figure la proportion de blocs concernés dans le corpus de référence constitué :

1. titre, sous-titre(s), chapeau et corps de texte (13%) ;
2. autres segments de l'article, par exemple les légendes (3%) ;
3. commentaires des lecteurs (1%) ;
4. contenu connexe, par exemple liens vers d'autres articles (4%) ;

75. `http://sante.lefigaro.fr/actualite/2013/04/25/20419-fish-pedicure-nest-pas-sans-risque` consulté le 12 octobre 2013

Tag	Contenu
<h>	La « fish pedicure » n'est pas sans risque
<auteur>	Par Dephine Chayet
<date>	25/04/2013
<legende>	La « fish pedicure » est apparue en France en 2010.
<p>	L'Agence nationale de sécurité sanitaire demande un encadrement [. . .]
<p>	Se laisser grignoter les peaux mortes des pieds par des petits poissons [. . .]
<p>	Apparue en France en 2010, la « fish pedicure » n'est aujourd'hui [. . .]
<h>	Poissons d'élevage
<p>	Même si aucun cas documenté n'a pour l'instant été rapporté,[. . .]
<p>	Dans une eau qui ne peut par définition être désinfectée,[. . .]
<p>	Elle recommande aussi une information « objective » du public [. . .]

Tableau 7.1 – Attendu de détourage sur l'article du *Figaro* intitulé : La « fish pedicure » n'est pas sans risque

Les éléments non classés dans l'une de ces catégories sont partie intégrante de la catégorie des segments « non-informatifs ». Dans l'évaluation menée par les auteurs, ils représentent 79% des blocs présents dans les documents du corpus. Dans notre étude nous nous concentrons sur le premier élément de la liste : ce qui concerne l'article lui même, puisque cela correspond à ce que nous recherchons dans notre évaluation par la tâche.

La fonction de détourage peut être résumée à deux sous-tâches :

- Le nettoyage :
  - Du code (*javascript*, feuille de style) ;
  - Du squelette de page (menus, liens, entêtes et pieds de page).
- L'annotation maîtrisée de la structure :
  - Titres (<h>) ;
  - Paragraphes (<p>) ;
  - Listes (<ol>, <ul>).

Nous proposons dans le tableau 7.1 une référence de détourage pour l'exemple présenté dans la figure 7.1 (page 156). En l'absence de consensus sur le statut à donner aux éléments secondaires comme l'auteur, la date et les légendes, nous les avons conservé, mais avec une balise typante (<auteur>, <date> et <legende>). Dans la section 7.1.2 nous examinons les caractéristiques communément utilisées pour obtenir ce genre de résultat.

7.1.2 Les caractéristiques utilisées pour le détourage

L'approche la plus naturelle pour détourer les pages Web est l'exploitation du *Document Object Model* (ou DOM). Il existe une familiarité évidente entre le détourage et la

FIGURE 7.1 – Exemple de page Web du site du *Figaro*, les éléments textuels importants sont entourés en bleu. En orange figurent les éléments potentiellement intéressants

segmentation automatique de pages Web à partir du DOM comme chez Chakrabarti *et al.* ([Chakrabarti-2008]).

Au grain DOM, exploiter les similarités existant entre différentes pages issues du même site Web ([Vieira-2006]) est une approche séduisante. Ce qui est commun à plusieurs pages est supposé correspondre à la structure et aux publicités (le contenu non-informatif). Ce qui est différent constituant le contenu informatif. Une approche connexe est de représenter la spécificité des pages Web d'un même site sous la forme d'un arbre. La position relative des nœuds dans l'arbre permet alors de classer les différents segments dans la catégorie « informatif » ou « non-informatif » ([Shine-2012]).

La densité en balises *HTML* est une autre façon d'exploiter le code source de la page pour classer les segments comme chez Ferraresi ([Ferraresi-2008]). D'autres approches plus « statistiques » proposent l'exploitation la fréquence des n-grammes de caractères comme dans *NCleaner* ([Evert-2008]) ou *Victor* ([Spousta-2008]). La combinaison de ces deux principes a été proposée par Pasternack *et al.* ([Pasternack-2009]).

7.2 Caractéristiques des détoureurs utilisés

Dans cette section nous détaillons la méthode utilisée par chacun des détoureurs utilisé pour nos expériences. Le code *HTML* est utilisé dans la quasi-totalité des squelettes rencontrés sur les sites de presse. Nous utilisons par souci de lisibilité le terme *HTML* afin de désigner le code source qu'il soit *HTML* ou XHTML.

Globalement, nous pouvons considérer que les détoureurs peuvent faire appel à des caractéristiques de différents ordres. Nous proposons une classification en fonction de la granularité. Dans cette approche les critères peuvent être rattachés à quatre niveaux d'analyse différents. Nous présentons ces niveaux dans la Figure 7.2.

1. le **site Web** où l'on observe les caractéristiques inhérentes à différentes pages d'un même site ;
2. l'**image de page** où l'on observe le rendu de la page tel qu'il est donné par un navigateur ;
3. le **code *HTML*** lui même avec notamment les informations de hiérarchie entre les blocs ;
4. le **contenu des blocs** *HTML* : phrases, mots et caractères.

FIGURE 7.2 – Catégorisation des indices exploitables pour le détourage des pages Web

7.2.1 Notre *baseline* : *Html2Text*

Html2Text est un utilitaire disponible en ligne de commande sur les systèmes Unix [76]. Cet outil constitue une *baseline* intéressante puisqu'il se base uniquement sur le rendu du code *HTML* [77], ce qui correspond au second point de la classification de la Figure 7.2. *Html2Text* va donc restituer sous forme textuelle l'intégralité de ce qu'un utilisateur humain pourra lire grâce à un navigateur. Le résultat qu'il offre est attendu comme référence en terme de rappel : l'intégralité des éléments textuels devant être présents. *A contrario*, la précision devrait être faible : tous les éléments visuels du squelette de page ainsi que les commentaires sont conservés.

7.2.2 *Boilerpipe*

Boilerpipe ([Kohlschutter-2010]) est un outil gratuit disponible sur le Web [78]. Il se base sur une combinaison de différents critères concernant le contenu supposé des blocs *HTML*. Les concepteurs ont délibérément choisi de ne pas tenir compte des niveaux « image de page » et « site ». Pour le premier niveau, les auteurs estiment que le calcul du rendu de l'image de page est trop coûteux d'un point de vue calculatoire, spécifiquement pour traiter une grande quantité de fichiers. Pour le second niveau, ils invoquent une volonté d'indépendance vis-à-vis du domaine ou du type de squelette utilisé. Par ailleurs, ils considèrent que cela implique un traitement trop différencié entre des sites contenant un nombre de pages potentiellement très différent. Un autre cas difficile est celui où l'on a une seule page disponible pour un site Web donné. Toutefois, une dimension « corpus » est conservée dans la mesure où des statistiques sur différents éléments du contenu textuel sont calculées au grain corpus.

La structure dégagée par le code *HTML* est très peu utilisée dans *Boilerpipe*. À nouveau, les auteurs invoquent le coût calculatoire du calcul du rendu de la page pour écarter l'idée d'utiliser la structure du DOM. Seules les balises considérées comme les plus communes des zones textuelles sont utilisées, à savoir les balises titres (<h1> à <h6>), paragraphes (<p>) et le conteneur . La balise de liens (<a>) est utilisée de façon assez classique pour identifier les zones qui ne sont probablement pas des zones de texte.

La caractéristique principale dont tient compte *Boilerpipe* est la longueur moyenne des mots graphiques (suite de caractères sans espace ou signe de ponctuation). Cette caractéristique est complétée par des indices locaux et des indices contextuels.

Les indices locaux sont la proportion de mots capitalisés, de liens hypertextes, de points (en tant que marqueurs supposés de fin de phrase) ainsi que celle de « | » (tube ou en anglais *pipe*). Les indices contextuels se fondent sur une hypothèse de position relative des blocs de texte et des blocs de squelette. Les auteurs postulent que les blocs de textes sont

76. Une documentation figure en ligne à l'adresse `http://www.mbayer.de/html2text/`
77. Ce qui revient manuellement à faire un copier-coller du contenu complet de la fenêtre du navigateur.
78. `http://code.google.com/p/boilerpipe/`

consécutifs et qu'il en est de même pour les blocs faisant partie du squelette. L'étiquette « informatif » ou « non-informatif » d'un bloc est donc fortement dépendante de l'étiquette du segment précédent. Une mesure de densité permet de juger si l'étiquette doit changer. La densité de chaque bloc est mesurée en divisant le nombre de *tokens* contenus dans un bloc par le nombre de lignes de ce bloc. Le *token* utilisé est le mot graphique. Le nombre de lignes est obtenu par simulation du rendu visuel : la largeur maximale d'une colonne considérée être de 80 caractères. C'est donc le nombre de colonnes de 80 caractères pouvant être remplies avec les mots du bloc. Une colonne incomplète n'est considérée que si elle est la seule colonne du bloc.

7.2.3 *NCleaner*

Le détoureur *NCleaner*[79] ([Evert-2008]) est basé sur des modèles de langues en n-grammes de caractères. La méthode se base sur un entraînement à partir des données de référence. Elle exploite un modèle qui mesure la probabilité qu'un caractère donné appartienne à la langue sachant les caractères qui le précèdent. C'est donc une probabilité conditionnelle de la forme $Pr(c_i|c_1...c_{i+1})$, où c_i représente le caractère à l'offset i. *NCleaner* cherche à identifier les n-grammes (avec n indéfini) qui maximisent la probabilité d'appartenance d'un bloc au contenu informatif. Ces n-grammes sont identifiés par apprentissage pour chaque langue opéré sur les données de référence. L'utilisation du grain caractère est censée assurer une certaine indépendance vis-à-vis de la langue. En contrepartie, l'existence d'un corpus d'apprentissage est contraignante. *NCleaner* n'utilise pas d'heuristiques à vocation généraliste comme le font d'autres détoureurs. Dès lors, ses performances par défaut (i.e. sans données d'entrainement pour une langue spécifique) peuvent être faibles.

7.2.4 *Readability*

Readability est disponible sous la forme d'une interface de programmation (en anglais *API*). Son fonctionnement est partiellement décrit sur le site Web du projet (`http://lab.arc90.com/`). En l'absence de publication scientifique dédiée à cette interface, nous avons collecté des informations sur des sites Web dédiés à la programmation[80] ou dans le code source de différentes adaptations de *Readability* (en *Python* notamment). L'approche est fondée sur la notion de « candidats ». Un certain nombres d'indices permettent au système de mesurer la probabilité qu'un bloc de texte fasse partie ou non du contenu informatif. Des indices positifs sont utilisés : la présence dans les identifiants de blocs de termes comme : *article, body, main, content...* . D'un autre côté des indice négatifs sont également exploités : par exemple la présence d'identifiants de blocs contenant *comment*,

79. `http://webascorpus.sourceforge.net/PHITE.php?sitesig=FILES&page=FILES_10_Software`
80. Voir par exemple l'intervention du concepteur de *Boilerpipe* sur `http://stackoverflow.com/questions/3652657/what-algorithm-does-readability-use-for-extracting-text-from-urls`

foot ou *disqus*[81]. Ce faisceau d'indices est complété par une série d'heuristiques classiques telles que la densité en liens ou en images.

7.2.5 Corpus de référence

Nous utilisons comme corpus de référence les documents qui ont servi dans le chapitre 4 à évaluer DAnIEL. Ce corpus permettra une évaluation classique par le contenu ainsi qu'une évaluation par la tâche. Il s'agira de voir dans quelle mesure la qualité du détourage a une influence sur la qualité des extractions de DAnIEL. Nous pourrons également mesurer si le meilleur détoureur sur des mesures classiques sur le contenu est aussi le meilleur détoureur par rapport à une tâche donnée.

7.3 La campagne *Cleaneval* : motivations, description et évaluation

Nous décrivons dans cette section les modalités d'évaluation du détourage proposées dans le cadre de la campagne *Cleaneval* ([Baroni-2008]). Les motivations de cette campagne sont évidemment très proches de celles de ce chapitre mais plus encore, c'est spécifiquement le constat initial effectué par les promoteurs de cette campagne qui justifie l'intérêt accordé à cette campagne d'évaluation. Les organisateurs font en effet le constat que le détourage est une tâche capitale pour toute analyse linguistique ultérieure mais qu'elle reste « peu glamour » (pour reprendre les termes mêmes des auteurs[82]).

Nous exposons dans un premier temps le format du *Gold Standard* utilisé pour la compétition (Section 7.3.1). Ce format a des conséquences directes sur les choix en termes d'évaluation (Section 7.3.2). Enfin, nous proposons une discussion sur ces modalités et proposons des améliorations (Section 7.3.3).

7.3.1 Le format de texte utilisé pour *Cleaneval*

Le corpus de référence a été obtenu par le travail d'annotateurs humains munis d'instructions d'annotation précises[83]. Les annotateurs, étudiants de l'université de Leeds, étaient invités à nettoyer les textes de manière à faciliter leur traitement automatique ultérieur. Ceci impliquait en premier lieu d'enlever les traces du squelette de page (*boilerplate*) ainsi que le code *HTML* ou *Javascript*. En second lieu, il s'agissait de conserver une structure de texte simplifiée utilisant trois balises :

81. Service Web de commentaires.

82. *Cleaning webpages is a low-level, unglamorous task and yet it is increasingly crucial : the better is done, the better the outcomes.*

83. Bien que l'adresse d'origine soit désormais inaccessible, il est encore possible d'accéder aux instructions d'annotation complètes par le biais de *Web Archive* : `http://web.archive.org`

– <**h**> pour les titres et sous-titres
– <**p**> pour les paragraphes
– <**l**> pour les éléments de listes

Par ailleurs, aucun encapsulage n'est considéré : chaque balise ouvrante ferme la précédente balise ouverte. Le corpus comprenait 741 documents en anglais et 713 documents en chinois. Les organisateurs ont donc souhaité un format de texte simple, format qui nous semble adapté à une tâche d'annotation d'envergure importante. Un format plus fouillé aurait, selon nous, eu deux inconvénients majeurs.

Le premier serait d'augmenter le risque de désaccord entre annotateurs. Selon notre expérience de l'annotation dans le domaine épidémiologique, faire annoter un trop grand nombre de traits implique des différences très grandes entre les annotateurs.

Le second est le corolaire du premier : plus l'annotation est contraignante, plus il est difficile d'obtenir une grande quantité de données fiables. Les annotations doivent être contrôlées pour garantir une certaine constance dans le processus. Et le coût en temps de l'annotation augmente.

Le choix opéré dans *Cleaneval* a néanmoins été critiqué, car le lien entre le balisage original et le balisage de référence pouvait disparaître. Par exemple, un paragraphe dont le contenu serait *tototiti* marqué par l'usage de la balise de retour à la ligne *tototiti* <*br*/> dans le document source se trouvait « traduit » dans la référence par <*p*> *tototiti*.

7.3.2 Modalités d'évaluation

Le format de texte choisi amène deux aspects à évaluer :

– Le nettoyage de ce qui n'est pas du contenu ;
– La conservation de la structure du texte d'origine.

Le script d'évaluation, en *Python*, fourni pour la campagne *Cleaneval* considère donc deux grains :

– Les éléments du contenu textuel, les mots ;
– Les éléments de structure, les balises.

Pour la structure, deux modalités d'évaluation sont proposées : *labelled* qui tient compte du type de balise, et *unlabelled* qui n'en tient pas compte. En mode *unlabelled*, la série « <p> <p> <l> » est strictement équivalente à la série « <p> <p> <p> ». Pour chaque fichier, la version détourée automatiquement est comparée avec la version figurant dans le corpus de référence. Chaque version est normalisée de la façon suivante :

– remplacement des caractères de contrôle (sauts de lignes, tabulations...) par des espaces ;
– normalisation des espaces [84].

84. Chaque série d'espaces consécutifs est remplacé par un seul espace.

Chaque version est ensuite découpée en une série unique de tokens obtenue en découpant à chaque signe de ponctuation ou espace. Afin de comparer les deux séries de tokens, la méthode *SequenceMatcher* de la librairie *difflib* est utilisée. L'algorithme utilisé dans *SequenceMatcher* est inspiré de l'agorithme de Ratcliff ([Ratcliff-1988]). Cet algorithme est de complexité cubique dans le pire cas. Il est conçu pour calculer les séquences communes maximales entre deux séries. *SequenceMatcher* permet d'obtenir ces séquences communes ainsi que la série d'opérations permettant de passer d'une série à l'autre : effacements, remplacements et insertions. L'algorithme maximise la longueur des sous-séquences communes, la similarité entre deux séquences est le double de la longueur cumulée des sous-séquences communes (sans chevauchement) divisé par la longueur totale des deux chaînes. Soient deux chaînes $s_1 =$ "*totititoti*" et $s_2 =$ "*tototiti*". L'algorithme de Ratcliff permet d'obtenir la liste d'opérations suivantes pour transformer s_1 en s_2 (l'opération de « conservation » est donnée à titre indicatif.

Opération	Offset dans la chaîne	Séquence concernée
Insertion	0	"*to*"
Conservation	2	"*totiti*"
Suppression	6	"*toti*"

Tableau 7.2 – Résultats de l'algorithme de Ratcliff pour transformer la séquence de caractères $s_1 =$ "*totititoti*" en $s_2 =$ "*tototiti*".

Ici la similarité entre les deux chaînes est donc $(2 * 6)/(10 + 8) = \frac{2}{3}$.

L'évaluation *Cleaneval* n'utilise pas ce calcul de similarité mais les opérations identifiées afin de calculer rappel et précision. Ainsi, en comparant la séquence issue résultant d'un détoureur et celle issue de la référence on effectue les opérations suivantes :

- La conservation d'une séquence de longueur n revient à augmenter le nombre de **vrais positifs** (VP) de n ;
- L'insertion de n éléments revient à augmenter de n les **faux négatifs** (FN) ;
- La suppression de n éléments, augmente le nombre de **faux positifs** (FP) de n.

Dans l'exemple précédent, nous avons donc $VP = 6$, $FN = 2$ et $FP = 4$ soit un rappel de 0.75 et une précision de 0.6.

Ces mesures sont calculées séparément pour les balises ; par contre la mesure sur les mots ne différencie pas mots et balises.

7.3.3 Discussion

Le fait que les métriques sur les mots et sur les balises ne soient pas totalement « étanches » gêne la lisibilité des résultats. Deux détoureurs très proches dans leurs résultats au grain « mot » peuvent avoir des résultats au grain « balise » qui sont complètement

différents. Or, de bons résultats sur le balisage permettent d'améliorer le résultat sur les mots.

Si nous prenons un exemple absurde, un détoureur qui n'extrairait que la structure et aucun des mots du texte aurait un score non-nul au grain mot. Nous proposons donc d'évaluer véritablement séparément ces deux aspects et de conserver une métrique unifiée qui soit pleinement assumée.

De plus, l'utilisation du grain mot n'est pas sans poser problème pour des langues telles que le chinois. Par ailleurs, les pré-traitements opérés sur les textes (détouré automatiquement et référence) étaient insuffisants pour évaluer correctement la conformité des balises. Il reste difficile enfin d'estimer ce que peut être « un bon résultat » avec les métriques *Cleaneval*. En effet, si l'on compare simplement les versions brutes *HTML* et le corpus de référence, le rappel n'est pas de 100%. Ceci est compréhensible pour les balises, puisqu'il y a une sorte de convention dans l'utilisation des balises (remplacer les </br> par des <p> par exemple). Nous pourrions attendre toutefois qu'au grain mot, tous les *token* recherchés soient bel et bien présents dans le fichier source.

Nous proposons donc une version modifiée du script d'évaluation destinée à combler ces manques. Nous introduisons notamment une évaluation par quadrigrammes de caractères, destinée à mieux évaluer les performances des détoureurs, sur le chinois. Les séquences de quadrigrammes (référence et version détourée) sont obtenus à l'aide d'une fenêtre glissante de quatre caractères appliquée sur l'intégralité des documents. Dans le cadre du stage de Pierre-Philippe Berenguer [85], étudiant de licence 3, une version plus raffinée du script d'évaluation a été développée. Dans cette version, l'étape de nettoyage préalable à l'évaluation a été améliorée, de façon à corriger certains biais de l'évaluation. Par exemple, lorsque un mot était « collé » à une balise de bloc dans un fichier à évaluer (i.e. sans espace typographique entre eux), le script ne faisait pas la coupure. Ceci amenait à donner des scores relativement décevants à des fichiers détourés automatiquement qui étaient pourtant très proches de la référence. Nous avons ainsi des scores plus proches de 100% au grain mot lorsque le source *HTML* est comparée à la référence manuelle.

7.4 Comparaison des différents détoureurs

7.4.1 Évaluation globale

Le tableau 7.3 présente les résultats de différents détoureurs et de leurs combinaisons en utilisant les métriques originales de *Cleaneval* auxquelles nous avons ajouté une évaluation par quadrigrammes de caractères. Pour les tableaux suivants nous utilisons les abréviations *tag* pour balises et *quadri* pour quadrigrammes. Les mesures sont le rappel (R), la précision (P) et la F_1-mesure (F). Elles ont appliquées sur les mots, les balises et les quadrigrammes de caractères. Pour ce tableau et les suivants, les abréviations suivantes

85. Encadré par Emmanuel Giguet de la société *Semiotime*.

	Mots			Balises			Quadrigrammes		
	F.	P.	R.	F.	P.	R.	F.	P.	R.
H2TXT	36,9	26,6	66,1	0,1	18,2	0,03	37,48	24,8	80,7
BR2	46,14	30,4	**95,71**	19,17	10,64	**97,01**	46,47	30,72	**95,34**
NC	13,1	11,52	15,19	10,3	8,79	12,43	9,35	8,46	10,44
BP	84,39	81,18	87,88	72,89	63,92	84,79	85,97	83,7	88,36
BP-BR2	**85,92**	82,88	89,19	**76,17**	68,79	85,33	**86,82**	84,50	89,27
BP-NC	51,85	**89,99**	36,42	42,94	**88,02**	28,4	35,54	**90,82**	22,09
BR2-BP	62,66	50,73	81,91	45,44	31,27	83,09	61,59	50,45	79,04
BR2-NC	44,16	50,12	39,47	41,32	57,54	32,23	32,5	50,06	24,06

Tableau 7.3 – Comparaison des détoureurs et de plusieurs combinaisons sur les métriques de la campagne *Cleaneval* ainsi que sur une évaluation par quadrigrammes de caractères calculée avec les mêmes formules

sont utilisées : H2TXT pour la *baseline Html2Text*, BR2 pour notre détoureur local (issu du travail de master 1 de Benoit Romito [86]), NC pour *NCleaner* et BP pour *BoilerPipe*. Les combinaisons de détoureurs consistent à opérer un premier détourage avec un détoureur et d'exploiter ce résultat avec un second détoureur. Il s'agit de voir dans quelle mesure les détoureurs peuvent être complémentaires.

Nous pouvons voir que les performances de *BoilerPipe* sont les meilleures parmi les détoureurs évalués. Notre détoureur local BR2 étant celui qui offre le meilleur rappel dans les trois grains d'évaluation (mots, balises et quadrigrammes), nous avons souhaité évaluer s'il pouvait apporter une plus-value aux résultats de *BoilerPipe*.

Il apparaissait fondé de faire fonctionner en premier lieu BR2 puis de voir comment BP pouvait affiner les résultats en améliorant les résultats. Pourtant les résultats du chaînage BR2-BP sont décevants. BP utilise les balises et les indices locaux pour typer les segments, or nombre de ces observables sont écrasés par le traitement opéré par BR2. Le chaînage inverse, BP-BR2, offre par contre des résultats significativement meilleurs que BP. Cela permet notamment d'améliorer la précision au grain balises de 4,8 points de pourcentage.

Le tableau 7.4 présente cette fois les résultats de l'évaluation par la tâche. Nous indiquons dans la colonne « REF DAnIEL » le résultat attendu sur un détourage idéal. Nous utilisons les modèles d'articles de presse présentés dans le chapitre 5 (tableau 5.2 page 95) pour établir trois grilles d'analyse des textes.

Avec *dan1*, chaque paire de segments de texte est considérée comme une position remarquable, indépendamment du nombre de segments que compte le texte.

Avec *dan2*, pour les documents comprenant quatre segments ou plus, les positions remarquables sont celles qui impliquent l'un des deux premiers segments, dans les autres cas *dan1* est appliqué.

Dan3 est le modèle utilisé par DAnIEL dans le chapitre 5. Les documents comportant

86. Encadré par Nadine Lucas à l'Université de Caen.

	Dan3			Dan2			Dan1		
	F.	**P.**	**R.**	**F.**	**P.**	**R.**	**F.**	**P.**	**R.**
H2TXT	1,6	33,3	8,0	2,3	33,3	1,2	26,6	15,4	93,3
BR2	7,7	100	4,0	14,5	28,9	18,2	30,4	18,3	98,4
NC	21,7	37,7	15,1	25,6	40,3	18,7	27,7	27,7	27,7
BP	61,6	**69,0**	55,6	70,2	63,2	79,1	63,7	**48,1**	94,3
BP-BR2	**64,9**	67,2	**62,8**	**70,8**	**63,4**	**80,1**	**63,9**	**48,1**	**95,0**
BP-NC	46,4	46,4	46,4	50,0	45,5	55,3	40,7	31,6	57,1
BR2-BP	41,3	66,6	30,0	56,6	60,9	50,3	50,5	35,8	86,0
BR2-NC	41,9	43,5	34,5	42,1	44,3	40,2	31,4	24,5	43,7
REF DAnIEL	**76,9**	**66,1**	**91,9**	**75,4**	**62,6**	**94,6**	**66,3**	**50,3**	**97,3**

Tableau 7.4 – Comparaison des différents détoureurs, évaluation par DAnIEL (Dan.1. Dan.2. Dan.3) avec 1, 2 et 3 jeux de positions attendues.

plus de 11 segments n'ont comme positions remarquables que celles qui impliquent deux segments parmi les trois premiers et les deux derniers. Dans les autres cas, *dan2* est appliqué.

Nous nous attendons à ce que le relâchement de contraintes de *dan1* amène un plus grand rappel alors que le resserrement impliqué par *dan3* doit améliorer la précision.

Nous observons que BP est le détoureur le plus efficace et que le chaînage BP-BR2 reste le plus prometteur. La méthode *dan1* amène de très bons résultats en terme de rappel, au prix d'une précision légèrement inférieure au score « attendu ». C'est lorsque les contraintes de position sont les plus faibles que les détoureurs automatiques permettent d'obtenir les résultats les plus proches de ce que permet le détourage manuel. Nous pouvons voir qu'il est difficile d'améliorer la précision sans que le rappel souffre trop. Ainsi, pour BP-BR2, le meilleur détoureur dans cette évaluation également, l'application de *dan2* fait gagner en précision l'équivalent de ce qui est perdu en rappel (de l'ordre de 15 points). Ceci ne semble pas adapté à notre tâche où le rappel est capital.

Il y a deux façons d'interpréter ce résultat. La première est de considérer que le modèle enrichi utilisé par DAnIEL est inadapté à la réalité du détourage automatique. La seconde est de considérer que ce sont les détoureurs qui ont des résultats finalement moins fiables que l'on pourrait le croire au premier abord. Des résultats de très bonne qualité sur des critères de contenu textuel peuvent masquer la disparition d'un certain nombre de traits, notamment des traits structurels inhérents au genre. Par ailleurs, nous pouvons nous interroger sur les performances pour chaque langue : dans quelle mesure les détoureurs présentés sont-ils robustes à la variation en langue ?

	Multi	el	en	pl	ru	zh
F.mot	36,9	50,28	49,66	46,97	34,45	3,57
P.mot	16,6	42,02	34,54	32,54	22,09	2,00
R.mot	66,1	**62,56**	**88,30**	84,40	78,20	17,01
F.tag	0,1	0,00	0,27	0,00	0,00	0,00
P.tag	18,2	0,00	90,91	0,00	0,00	0,00
R.tag	0,03	0,00	0,14	0,00	0,00	0,00
F.quadri	37,48	39,88	39,07	38,67	27,86	41,92
P.quadri	24,8	28,99	25,11	25,05	16,67	28,16
R.quadri	80,7	**63,88**	**88,04**	84,77	84,70	82,02

Tableau 7.5 – Performances par langues de Html2Text, Multi étant le corpus cumulé des cinq langues

7.4.2 Évaluation par langue

Le tableau 7.5 représente les performances par langues de notre *baseline*. Les performances sur les balises sont anecdotiques puisque Html2Text ne cherche nullement à segmenter correctement. Les performances par langues éclairent d'autres aspects. Bien que disposant d'une option « Utf-8 », *Html2Text* éprouve des difficultés à analyser certains documents qui ne sont pas en anglais. Notamment, certains fichiers « détourés » en grec sont en fait vides, ce qui impacte fortement les résultats. En examinant les tableaux 7.6 et 7.8 nous constatons que les textes en chinois et en russe sont moins bien détourés que ceux des autres langues sur les métriques initiales de *Cleaneval*. Toutefois, l'introduction de l'évaluation par quadrigrammes de caractères amène le détourage du chinois à des chiffres plus conformes à ceux des autres langues. Par contre, les résultats sur les textes en russe restent inférieurs quelle que soit la mesure utilisée. Ceci semble lié à des caractéristiques du source *HTML* lui même : ce que nous pouvons voir sur l'évaluation par balises.

Finalement, seul le détoureur BR2 parvient à des résultats en russe proche de ceux des autres langues. Cette performance est principalement intéressante en termes de rappel car dans le même temps la précision est très faible.

Le tableau 7.9 représente les résultats par langues du chaînage BP-BR2. Les résultats sont en amélioration par rapport à BP, nous pouvons noter qu'il n'y a pas à proprement parler de « rattrapage » sur le russe. Les améliorations sont au mieux équivalentes à celles qui sont obtenues sur le corpus complet, que ce soit en valeur absolue (points de pourcentage) ou relative (ratios de points de pourcentage). Le russe est, sur notre corpus de référence, une langue qui pose des problèmes de détourage particuliers.

Les performances variables des détoureurs selon les langues invitent à se demander si la solution est d'avoir des détoureurs « spécialisés » sur certaines langues. Cette solution n'est évidemment pas compatible avec un objectif de parcimonie et de multilinguisme. Nous

	Multi	el	en	pl	ru	zh
F.mot	46,14	53,35	49,67	48,33	33,02	26,21
P.mot	30,40	**36,75**	33,15	32,94	**20,17**	15,25
R.mot	95,71	97,29	99,04	90,70	91,11	93,33
F.tag	19,17	15,70	24,96	17,57	9,89	24,59
P.tag	10,64	8,53	14,28	9,71	5,22	14,06
R.tag	97,01	97,75	99,21	92,74	92,97	97,80
F.quadri	46,47	54,08	47,46	47,40	33,39	51,62
P.quadri	30,72	37,26	31,20	31,96	20,29	36,16
R.quadri	90,19	98,59	99,14	91,72	94,23	95,34

Tableau 7.6 – Performances par langues de BR2, Multi étant le corpus cumulé des cinq langues

	Multi	el	en	pl	ru	zh
F.mot	13,10	0,2	31,51	4,33	1,54	1,44
P.mot	11,52	**0,13**	29,3	9,03	1,98	0,89
R.mot	15,19	0,36	**34,08**	2,85	1,27	3,76
F.tag	10,30	0,25	24,24	2,16	4,38	3,00
P.tag	8,79	0,16	20,48	3,24	4,20	4,26
R.tag	12,43	0,63	29,7	1,62	4,57	2,31
F.quadri	9,35	0,62	24,3	4,95	2,25	4,46
P.quadri	8,46	0,49	18,79	7,15	3,08	4,82
R.quadri	10,44	0,87	34,39	3,79	1,77	4,15

Tableau 7.7 – Performances par langues du détoureur *NCleaner* (NC), Multi étant le corpus cumulé des cinq langues

	Multi	el	en	pl	ru	zh
F.mot	84,39	93,16	88,7	83,09	66,05	55,5
P.mot	81,18	91,21	85,81	82,38	57,76	58,93
R.mot	87,88	**95,19**	91,79	83,81	77,11	**52,45**
F.tag	72,89	77,06	79,63	72,34	49,55	74,98
P.tag	63,92	66,58	69,29	62,91	35,69	84
R.tag	84,79	91,45	93,59	85,08	81,01	67,71
F.quadri	85,97	95,42	89,83	85,4	70,12	84,36
P.quadri	83,7	93,4	86,82	84,68	60,81	93,09
R.quadri	88,36	**97,54**	93,06	86,13	82,8	**77,12**

Tableau 7.8 – Performances par langues du détoureur BoilerPipe (BP), Multi étant le corpus cumulé des cinq langues

BP-BR2	Multi	el	en	pl	ru	zh
F.mot	85,92 (+1,5)	94,08 (+0,9)	89,69 (+0,9)	85,27 (+2,1)	67,5 (+1,4)	63,73 (+8,2)
P.mot	82,88 (+1,7)	92,39 (+1,1)	86,78 (+0,9)	84,49 (+2,1)	59,65 (+1,8)	71,41 (+12,4)
R.mot	89,19 (+1,3)	95,83 (+0,6)	92,8 (+1,0)	86,05 (+2,2)	77,73 (+0,6)	57,54 (+5,0)
F.tag	76,17 (+3,2)	82,2 (+5,1)	82,89 (+3,2)	75,76 (+3,4)	51,98 (+2,4)	77,65 (+2,6)
P.tag	68,79 (+4,8)	74,62 (+8,0)	74,24 (+4,9)	67,63 (+4,7)	38,25 (+2,5)	91,72 (+7,7)
R.tag	85,33 (+0,5)	91,49 (+0,0)	93,83 (+0,2)	86,12 (+1,0)	81,09 (+0,0)	67,33 (-0,3)
F.quadri	86,82 (+0,8)	95,8 (+0,3)	90,39 (+0,5)	87,05 (+1,6)	70,97 (+0,8)	85,66 (+1,3)
P.quadri	84,5 (+0,8)	93,75 (+0,3)	87,34 (+0,5)	86,14 (+1,4)	62,12 (+1,3)	94,13 (+1,0)
R.quadri	89,27 (+0,9)	97,93 (+0,3)	93,68 (+0,6)	87,97 (+1,8)	82,77 (-0,0)	78,58 (+1,4)

Tableau 7.9 – Performances par langues du chaînage BP-BR2, Multi étant le corpus cumulé des cinq langues. Entre parenthèse la plus-value par rapport à BP seul.

BR2-NC	Multi	el	en	pl	ru	zh
F.mot	44,16	0,48	64,47	12,78	1,32	11,80
P.mot	50,12	35,82	50,23	52,6	25,22	39,17
R.mot	39,47	**0,24**	**89,97**	7,27	0,68	6,94
F.tag	41,32	0,07	68,21	11,15	1,18	10,86
P.tag	57,54	2,44	59,56	41,12	7,60	55,67
R.tag	32,23	0,04	79,78	6,45	0,64	6,02
F.quadri	32,5	9,64	64,04	12,40	1,08	9,52
P.quadri	50,06	0,03	50,4	54,04	27,73	44,55
R.quadri	24,06	396	87,8	7	0,55	5,33

Tableau 7.10 – Performances par langues du chaînage BR2-NC, Multi étant le corpus cumulé des cinq langues

BP-NC	Multi	el	en	pl	ru	zh
F.mot	51,85	0,48	86,81	11,16	1,23	6,88
P.mot	89,99	75,73	**90,29**	87	75,84	59,36
R.mot	36,42	**0,24**	**83,58**	5,96	0,62	3,65
F.tag	42,94	0,07	80,71	7,54	0,68	2,43
P.tag	88,02	50	89,41	80,26	29,17	43,65
R.tag	28,4	0,04	73,56	3,96	0,34	1,25
F.quadri	35,54	1,99	86,25	10,55	0,39	7,87
P.quadri	90,82	0,01	90,99	89,57	53,16	91,52
R.quadri	22,09	25,48	81,99	5,6	0,2	4,11

Tableau 7.11 – Performances par langues du chaînage BP-NC, Multi étant le corpus cumulé des cinq langues

BR2-BP	Multi	el	en	pl	ru	zh
F.mot	62,66	63,12	73,93	61,88	34,01	42,82
P ,mot	50,73	50,77	62,49	51,09	24,19	35,24
R.mot	81,91	83,41	**90,51**	78,44	57,24	**54,57**
F.tag	45,44	36,7	55,44	37,86	19,03	71,27
P.tag	31,27	23,46	39,57	24,82	11,02	68,82
R.tag	83,09	84,24	92,54	79,77	69,48	73,89
F.quadri	61,59	65,74	72,95	61,93	36,74	65,59
P.quadri	50,45	53,13	60,56	50,51	26,05	64,71
R.quadri	79,04	86,19	91,71	80,03	62,3	66,5

Tableau 7.12 – Performances par langues du chaînage BR2-BP. Multi étant le corpus cumulé des cinq langues

	F	P	R	F.tag	P.tag	R.tag	F.quadri	P.quadri	R.quadri
BR2	53,35	36,75	**97,29**	15,70	8,53	**97,75**	54,08	37,26	**98,59**
NC	0,20	0,13	0,36	0,25	0,16	0,63	0,62	0,49	0,87
BP	93,16	91,21	95,19	77,06	66,58	91,45	95,42	93,40	97,54
BP-BR2	**94,08**	**92,39**	95,83	**82,20**	**74,62**	91,49	**95,80**	93,75	97,93
BP-NC	0,48	75,73	0,24	0,07	50,00	0,04	25,48	0,01	199
BR2-NC	0,48	35,82	0,24	0,07	2,44	0,04	9,64	0,03	396
BR2-BP	63,12	50,77	83,41	36,70	23,46	84,24	65,74	53,13	86,19

Tableau 7.13 – Performances des détoureurs sur le grec

en déduisons que la volonté d'indépendance du domaine et du genre de texte invoquée par les concepteurs de la plupart des détoureurs cache en fait une forte dépendance à la langue (Tableaux 7.13 à 7.17).

7.4.3 Discussion

L'évaluation du détourage automatique est une tâche difficile. Dans le cadre de la campagne *Cleaneval*, deux modalités d'évaluation ont été proposées : le grain mot et le grain balise. Il y a donc une volonté de ne pas se concentrer sur un texte brut *stricto*

	F	P	R	F.tag	P.tag	R.tag	F.quadri	P.quadri	R.quadri
BR2	49,67	33,15	**99,04**	24,96	14,28	**99,21**	47,46	31,20	**99,14**
NC	31,51	29,30	34,08	24,24	20,48	29,70	24,30	18,79	34,39
BP	88,70	85,81	91,79	79,63	69,29	93,59	89,83	86,82	93,06
BP-BR2	**89,69**	86,78	92,80	**82,89**	74,24	93,83	**90,39**	87,34	93,68
BP-NC	86,81	**90,29**	83,58	80,71	**89,41**	73,56	86,25	**90,99**	81,99
BR2-NC	64,47	50,23	89,97	68,21	59,56	79,78	64,04	50,40	87,80
BR2-BP	73,93	62,49	90,51	55,44	39,57	92,54	72,95	60,56	91,71

Tableau 7.14 – Performances des détoureurs sur l'anglais

	F	P	R	F.tag	P.tag	R.tag	F.quadri	P.quadri	R.quadri
BR2	48,33	32,94	**90,70**	17,57	9,71	**92,74**	47,40	31,96	**91,72**
NC	4,33	9,03	2,85	2,16	3,24	1,62	4,95	7,15	3,79
BP	83,09	82,38	83,81	72,34	62,91	85,08	85,40	84,68	86,13
BP-BR2	**85,27**	84,49	86,05	**75,76**	67,63	86,12	**87,05**	86,14	87,97
BP-NC	11,16	**87,00**	5,96	7,54	**80,26**	3,96	10,55	**89,57**	5,60
BR2-NC	12,78	52,60	7,27	11,15	41,12	6,45	12,40	54,04	7,00
BR2-BP	61,88	51,09	78,44	37,86	24,82	79,77	61,93	50,51	80,03

Tableau 7.15 – Performances des détoureurs sur le polonais

	F	P	R	F.tag	P.tag	R.tag	F.quadri	P.quadri	R.quadri
BR2	33,02	20,17	91,11	9,89	5,22	92,97	33,39	20,29	94,23
NC	1,54	1,98	1,27	4,38	4,20	4,57	2,25	3,08	1,77
BP	66,05	57,76	77,11	49,55	35,69	81,01	70,12	60,81	**82,80**
BP-BR2	**67,50**	59,65	**77,73**	**51,98**	**38,25**	**81,09**	**70,97**	**62,12**	82,77
BP-NC	1,23	**75,84**	0,62	0,68	29,17	0,34	0,39	53,16	0,20
BR2-NC	1,32	25,22	0,68	1,18	7,60	0,64	1,08	27,73	0,55
BR2-BP	34,01	24,19	57,24	19,03	11,02	69,48	36,74	26,05	62,30

Tableau 7.16 – Performances des détoureurs sur le russe

	F	P	R	F.tag	P.tag	R.tag	F.quadri	P.quadri	R.quadri
BR2	26,21	15,25	**93,33**	24,59	14,06	**97,80**	51,62	36,16	**90,19**
NC	1,44	0,89	3,76	3,00	4,26	2,31	4,46	4,82	4,15
BP	55,50	58,93	52,45	74,98	84,00	67,71	84,36	93,09	77,12
BP-BR2	**63,73**	**71,41**	57,54	**77,65**	**91,72**	67,33	**85,66**	**94,13**	78,58
BP-NC	6,88	59,36	3,65	2,43	43,65	1,25	7,87	91,52	4,11
BR2-NC	11,80	39,17	6,94	10,86	55,67	6,02	9,52	44,55	5,33
BR2-BP	42,82	35,24	54,57	71,27	68,82	73,89	65,59	64,71	66,50

Tableau 7.17 – Performances des détoureurs sur le chinois

sensu. Toutefois, la mesure en terme de rappel et précision proposée n'est pas pleinement satisfaisante. Elle ne répond pas véritablement à la question « qu'est-ce qui constitue un bon ou un mauvais résultat ? ». L'évaluation permet, à défaut, d'opérer un classement des systèmes, on obtient donc une mesure relative de l'efficacité de chaque système. Le problème de l'interprétation des résultats n'est pas spécifique à cette évaluation, mais il nous semble ici qu'il est particulièrement important. En effet, dans le cadre d'une chaîne de traitement de la langue, la fonction de détourage est moins que tout autre « détachable » des traitements qui sont effectués en aval. La dénaturation du document traité, qu'elle concerne sa structure ou son contenu textuel, a des conséquences opératoires.

Notre contribution a consisté ici à proposer une nouvelle évaluation, par la tâche, de cette fonction de détourage. La veille épidémiologique multilingue, tâche centrale de cette thèse, nous a semblé être une tâche particulièrement adaptée à cette évaluation. En effet, DAnIEL utilise à la fois des propriétés du contenu textuel et de la structure. Il est donc influencé par les deux grains mesurés dans la campagne *Cleaneval* : les balises et les mots. Notre questionnement a été le suivant : dans quelle mesure le détourage a une influence sur les résultats de DAnIEL. Nous avons montré que les meilleurs détoureurs automatiques amenaient une perte significative dans les résultats de DAnIEL. Cette perte s'est avérée moins grande lorsque nous avons appauvri le modèle de document utilisé par DAnIEL. Nous avons montré que ce résultat pouvait s'interpréter de deux façons : (I) un manque de robustesse du modèle de document utilisé par DAnIEL face à la *réalité du détourage* ou (II) une incapacité des détoureurs de l'état de l'art à conserver dans le document détouré un aspect fidèle au document original.

Nous pourrions dans ce dernier cas conclure la chose suivante : à partir d'un excellent détoureur, DAnIEL fonctionnera très bien. Nous avions abordé dans le chapitre 3 la problématique des erreurs en cascade inhérente aux chaînes de traitement de TAL. Il nous semble que nous sommes dans ce cas. Nous avons proposé de raccourcir la chaîne de manière à rendre l'ensemble du processus plus robuste. Ici, nous pouvons considérer que nous arrivons à une limite de notre approche : nous sommes arrivé à un nombre minimal de composants et il semble difficile d'avoir un système plus épuré.

Pour une analyse complémentaire de ces résultats, nous pouvons reprendre les termes utilisés par les organisateurs de *Cleaneval* pour décrire le détourage. Le détourage est une tâche peu gratifiante et il est sans doute aussi peu gratifiant d'évaluer un système de traitement automatique conjointement avec le détoureur dont il dépend. Il nous semble au contraire tout à fait justifié d'évaluer l'intégralité du processus de traitement. En effet, l'utilisateur final ne souhaite sans doute pas que la qualité des résultats soit influencée par une contrainte qui paraît relever du domaine de l'ingénierie. Surtout que, de son point vue, le navigateur Web qu'il utilise n'a aucune difficulté à donner un rendu correct à ces documents, comment expliquer que l'on ne puisse le faire à des fins de traitement ?

Évaluer comment les *conditions de laboratoire* influent sur les résultats devrait alors être un souci plus constant des travaux de TAL. S'il n'existe pas de méthode, d'algorithme

permettant de détourer efficacement les pages Web, y compris de façon conditionnelle[87], alors il convient d'admettre que le détourage n'appartient pas encore au domaine de l'ingénierie, mais reste du domaine de la recherche.

87. On pourrait avoir une méthode de détourage parfaite sur l'anglais ou sur le domaine de la presse par exemple, sans que cette méthode prétende traiter plus.

Conclusion de la troisiéme partie

Dans cette dernière partie, nous avons présenté nos principales contributions sous la forme d'un corpus de référence et d'un système d'analyse automatique nommé DAnIEL. A l'heure où nous écrivons, le corpus annoté que nous avons constitué est le seul à être mis à disposition de la communauté des chercheurs en veille épidémiologique. Avec plus de 2000 documents annotés en cinq langues nous pouvons obtenir une photographie assez fiable des évènements épidémiologiques survenus sur une grande partie de la planète dans notre période d'étude (Novembre 2011-Janvier 2012). Cette photographie nous autorise à évaluer la correspondance entre les attentes des autorités sanitaires et ce que peut proposer le traitement manuel. Nous avons ainsi pu proposer différentes modalités d'évaluation, afin notamment de dépasser les biais engendrés par l'évaluation au grain « document ».

Nous avons ensuite présenté l'architecture générale de DAnIEL avec sa structure massivement factorisée adaptée au traitement multilingue. Avec DAnIEL, le coût marginal de traitement d'une nouvelle langue est en effet très limité. De cette manière, l'intervention humaine nécessaire est minimale. Ceci permet de laisser les spécialistes du domaines ou les locuteurs de la langue dans le rôle de juge de la qualité du système. Cela est plus valorisant que le rôle, ingrat, d'informateur ou d'expert chargé de remplir des bases de données. Nous pensons donc que DAnIEL offre un compromis coût-efficacité tout à fait intéressant et qu'il apporte une réelle plus-value pour la veille épidémiologique. Les principes qui ont guidé son fonctionnement sont adaptés à la dimension multilingue requise pour notre problème. Notre modèle de document, conçu pour le genre journalistique, montre une forte robustesse au traitement de nouvelles langues. Il permet de diminuer considérablement, et à performances égales, le **coût marginal** de traitement de nouvelles langues.

Nous avons montré ensuite que lors du traitement de nouveaux genres textuels, des articles scientifiques en l'occurrence, le modèle de document devait être ajusté. C'est la preuve que pour notre méthode, le genre textuel est une variable plus importante que la langue. Ces expérimentations ont montré à nouveau qu'un modèle générique, adapté au genre traité, était une solution efficace pour obtenir une couverture réellement multilingue.

Enfin, nous avons montré en quoi le détourage, ou nettoyage, de pages Web était une contrainte réelle pour les analyses effectuées en aval dans une chaîne de traitement. Nous avons montré que les meilleurs détoureurs sur des métriques classiques n'étaient pas nécessairement les plus adaptés pour une tâche donnée. Cette évaluation par la tâche nous permet de disposer de l'ensemble des informations nécessaires pour exploiter DAnIEL de façon efficace.

Conclusions et perspectives

Notre travail a été guidé par un objectif de traitement multilingue par le biais d'universaux et d'invariants du langage. Cet objectif était nourri par la volonté d'explorer de nouvelles facettes du traitement automatique des langues tout en proposant un système adapté à un thème particulier : la veille épidémiologique multilingue. L'exigence de couverture et donc de multilinguisme de ce domaine particulier nous a conduit à rechercher des solutions empruntant des voies peu ou pas empruntées par les approches de l'état de l'art. Nous avons montré qu'une approche multilingue raisonnée, c'est à dire compatible avec la réalité des ressources disponibles, devait se baser sur des universaux, des invariants de la langue. Cette approche est fondée sur deux principes : la factorisation et la parcimonie.

Nous avons fait le constat que les approches classiques se traduisent par deux caractéristiques très handicapantes dans une perspective multilingue. La première est la trop grande contrainte que constitue une analyse locale poussée, dépendante de modules très spécifiques aux langues. De ce fait, le coût marginal de traitement d'une nouvelle langue est important. La seconde est la prise en compte très parcellaire des aspects pragmatiques et communicationnels des énoncés en langue naturelle. Nous avons montré que la prise en compte du genre textuel analysé permet justement d'avoir une analyse indépendante de la langue traitée.

C'est en nous tournant vers des travaux de linguistique textuelle, de rhétorique et de pragmatique que nous avons trouvé les outils permettant de factoriser. La généricité recherchée a été obtenue par l'utilisation de propriétés des articles de presse dont nous avons montré la stabilité à travers les langues. Ces propriétés définissent un modèle du genre, modèle sur lequel DAnIEL se fonde pour extraire de l'information à partir des textes. La manifestation principale de la généricité de DAnIEL est le faible coût de traitement d'une nouvelle langue. Les modules classiques d'analyse locale tels que les lemmatiseurs, étiqueteurs, analyseurs syntaxiques ont été mis de côté au profit d'un noyau d'analyse indépendant des langues. Ce noyau utilise comme unité de base le caractère en lieu et place du mot ou du lemme afin de faciliter le traitement de langues avec des caractéristiques morphologiques variées.

Nous avons ainsi mené des expériences concluantes sur des langues à morphologie pauvre (chinois, anglais) comme sur des langues à morphologie riche (grec, polonais et russe) en passant par des langues à richesse morphologique plus modérée (allemand, fran-

çais et espagnol).

Le noyau central de notre système permet d'extraire le « contenu pertinent » de l'article en se fiant à des critères de position et de répétition. Il a montré sa robustesse puisque nous présentons une efficacité proche de l'état de l'art au prix d'un coût nul en modules d'analyse locale et d'un coût minimal en ressources lexicales.

Au-delà de l'aspect d'économie sur le traitement de langues données, notre méthode s'est révélée particulièrement adaptée à des langues pour lesquelles les modules d'analyse locale pouvaient être inexistants ou peu performants (en grec par exemple).

Ce noyau d'analyse central, épaulé par ses lexiques dédiés de petite taille constitue l'essence du système DAnIEL (pour *Data Analysis for Information Extraction in Any Language*). DAnIEL est aujourd'hui en mesure de traiter de nouvelles langues sans apprentissage, au sens classique du terme, ni paramétrage. DAnIEL se fonde seulement sur des propriétés du genre textuel. Dans le cadre de la veille épidémiologique, il nécessite simplement l'intégration de ressources lexicales aisées à collecter automatiquement et de taille compatible avec une collecte manuelle. Enfin, la simplicité et la généricité de DAnIEL nous permettent d'envisager le traitement d'autres domaines de la veille avec un minimum d'intervention humaine.

Par ailleurs, nous avons montré que pour une approche fondée sur le genre, le changement de langue importait peu. Nous avons pris pour exemple deux tâches de fouille de textes scientifiques proposées dans le cadre des campagnes d'évaluation DEFT. En changeant simplement le modèle de document utilisé par DAnIEL, nous avons pu obtenir de très bons résultats. Nous avons reproduit ces expériences sur le même type de textes mais une autre langue (le polonais) en conservant le même modèle. Les résultats obtenus étaient très proches des résultats obtenus sur le DEFT en français ce qui laisse à penser que l'approche fondée sur le genre est pertinente également dans le cas des articles scientifiques. Les invariants de genre semblent donc constituer d'excellents candidats pour un traitement des corpus multilingues efficace et à coût raisonnable.

Enfin, nous avons exploré la question du nettoyage des documents ou « détourage » et de leur influence sur les traitements ultérieurs. Pour une approche fondée sur le genre, la structure des documents analysés est en effet capitale. Nous avons proposé d'évaluer des outils de détourage de l'état de l'art non pas seulement sur le contenu des documents retournés, mais aussi en fonction de la tâche recherchée. Nous avons ainsi pu mesurer l'influence de ce pré-traitement sur les résultats de DAnIEL.

Le modèle de document utilisé par DAnIEL pourrait être développé et affiné de manière à offrir un cadre d'analyse plus souple, permettant de nouvelles utilisations. Nous avons choisi de restreindre la variété en genre textuel en en tâche afin de traiter au mieux la dimension multilingue. Dès lors, lever ces restrictions offrirait de nombreuses pistes à explorer. Dans le cadre de la veille épidémiologique il serait profitable de s'intéresser aux réseaux sociaux et notamment à *Twitter* qui est très prisé des épidémiologistes. La détection de nouveauté dans les documents traités offrirait notamment une plus-value sup-

plémentaire pour les autorités sanitaires. Toutefois, plus que l'étude de nouveaux genres ou de nouvelles tâches, c'est l'application de nos méthodes à d'autres domaines de la veille qui constitue pour nous la piste de recherche la plus intéressante.

Bibliographie

[Ahat-2012] Murat Ahat, Coralie Petermann, Yann Vigile Hoareau, Soufian Ben Amor, and Marc Bui. Algorithme automatique non supervisé pour le DEFT 2012. In *DEFT 2012*, pages 73–80, 2012. (Cité à la page 149.)

[Arnold-2013] Carrie Arnold. 10 years on, the world still learns from sars. *The Lancet Infectious Diseases*, 13(5) :394–395, 2013. (Cité à la page 23.)

[Atkinson-2013] Martin Atkinson, Mian Du, Jakub Piskorski, Hristo Tanev, Roman Yangarber, and Vanni Zavarella. Techniques for multilingual security-related event extraction from online news. In *Computational Linguistics*, pages 163–186. Springer, 2013. (Cité à la page 26.)

[Baker-2007] Michael G. Baker and Andrew M Forsyth. The new International Health Regulations : a revolutionary change in global health security. *Journal of the New Zealand Medical Association*, 120(1267) :U2872, 2007. (Cité à la page 23.)

[Baluja-2006] Shumeet Baluja. Browsing on small screens : recasting web-page segmentation into an efficient machine learning framework. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 33–42, New York, NY, USA, 2006. ACM. (Cité à la page 153.)

[Baroni-2008] Marco Baroni, Francis Chantree, Adam Kilgarriff, and Serge Sharoff. Cleaneval : a competition for cleaning web pages. In *Actes du 4ème Workshop Web as Corpus, LREC 2008*. European Language Resources Association, 2008. (Cité à la page 160.)

[Baylon-2005] Christian Baylon. *Initiation à la linguistique*. Armand Colin, 2005. (Cité à la page 71.)

[Benel-2012] Aurélien Bénel, Sylvie Calabretto, Véronique Eglin, Jérôme Gensel, Elisabeth Murisasco, Jean-Marc Ogier, Thierry Paquet, Jean-Yves Ramel, Florence Sèdes, and Nicole Vincent. *Information Interaction Intelligence le point sur le i3* , chapter Vers un « CTRL+F amélioré » pour tout type de document numérique ? Techniques et enjeux de la recherche de motifs. Cépaduès, Toulouse, 2012. (Cité à la page 152.)

[Bestgen-2011] Yves Bestgen. LSVMA : au plus deux composants pour apparier des

résumés à des articles. In *DEFT 2011*, pages 105–114, 2011. (Cité à la page 131.)

[Brants-1998] Thorsten Brants. TnT – Statistical Part-of-Speech Tagging. `http://www.coli.uni-saarland.de/~thorsten/tnt/`, 1998. (Cité à la page 41.)

[Breton-2010] Didier Breton, Mathieu Roche, Pascal Poncelet, and François Marques. Analyse de dépêches pour l'épidémiologie. In *21èmes Journées Francophones d'Ingénierie des Connaissance, Démonstrations*, pages 1–3, 2010. (Cité à la page 36.)

[Breton-2012] Didier Breton, Sandra Bringay, François Marques, Pascal Poncelet, and Mathieu Roche. Epimining : Using Web News for Influenza Surveillance. In *Workshop on Data Mining for Healthcare Management*, 2012. (Cité à la page 37.)

[Brixtel-2010] R. Brixtel, M. Fontaine, B. Lesner, C. Bazin, and R. Robbes. Language-independent clone detection applied to plagiarism detection. In *Source Code Analysis and Manipulation (SCAM), 2010 10th IEEE Working Conference on*, pages 77–86, 2010. (Cité à la page 57.)

[Brixtel-2011] Romain Brixtel. *Alignement endogène de documents, une approche multilingue et multi-échelle*. PhD thesis, Université de Caen, 2011. (Cité aux pages 78 et 83.)

[Brixtel-2013] Romain Brixtel, Gaël Lejeune, Antoine Doucet, and Nadine Lucas. Any Language Early Detection of Epidemic Diseases from Web News Streams. In *International Conference on Healthcare Informatics (ICHI)*, 2013. (Cité à la page 126.)

[CDC-2004] James W. Buehler, Richard S. Hopkins, Jean-Marc Overhage, Daniel M. Sosin, and Van Tong. Framework for Evaluating Public Health Surveillance Systems for early detection of Outbreaks. Technical report, Center for Disease Control, 2004. (Cité aux pages 33 et 46.)

[Cataldi-2010] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on Twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 1–10, New York, NY, USA, 2010. ACM. (Cité à la page 25.)

[Chakrabarti-2008] Deepayan Chakrabarti, Ravi Kumar, and Kunal Punera. A graph-theoretic approach to webpage segmentation. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 377–386, New York, NY, USA, 2008. ACM. (Cité à la page 157.)

[Chan-2010] Emily H. Chan, Timothy F.Brewer, Lawrence C. Madoff, Marjorie P. Pollack, Any L.Sonricker, Mikaela Keller, Clark C. Freifeld, Michael Blench,

Abla Mawudeku, and John S. Brownstein. Global Capacity for Emerging Infectious Disease Detection. *Proceedings of the National Academy of Sciences*, 107(50) :21701–21706, 2010. (Cité aux pages 33 et 36.)

[Chanlekha-2010] Hutchatai Chanlekha, Ai Kawazoe, and Nigel Collier. A Framework for Enhancing Spatial and Temporal Granularity in Report-based Health Surveillance Systems. *BMC Medical Informatics& ca cDecision Making*, 10(1) :1+, 2010. (Cité à la page 37.)

[Charnois-2009] Thierry Charnois, Marc Plantevit, Christophe Rigotti, and Bruno Crémilleux. Fouille de données séquentielles pour l'extraction d'information dans les textes. *Revue TAL*, pages 59–87, 2009. (Cité à la page 60.)

[Chen-2010] Hsinchun Chen, Daniel Zeng, and Ping Yan. Argus. *Infectious Disease Informatics*, 21 :177–181, 2010. (Cité à la page 40.)

[Church-2000] Kennet Church. Empirical Estimates of Adaptation : The chance of Two Noriega's is closer to p/2 than p 2. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 173–179, 2000. (Cité à la page 80.)

[Claveau-2012] Vincent Claveau and Christian Raymond. Participation de l'IRISA à DEFT 2012 : recherche d'information et apprentissage pour la génération de mots-clés. In *DEFT 2012*, pages 53–64, 2012. (Cité à la page 149.)

[Collier-2006] Nigel Collier, Kawazoe Ai, Lihua Jin, et al. A multilingual ontology for infectious disease surveillance : rationale, design and challenges. *Journal of Language Resources and Evaluation*, pages 405–413, 2006. (Cité aux pages 36, 60, et 93.)

[Collier-2010] Nigel Collier, Ai Kawazoe, Lihua Jin, M. Shigematsu, D. Dien, R. Barrero, K. Takeuchi, and A. Kawtrakul. An ontology-driven system for detecting global health events. In *Proc. 23rd International Conference on Computational Linguistics (COLING)*, pages 215–222, 2010. (Cité à la page 60.)

[Collier-2011] Nigel Collier. What's unusual in online disease outbreak news ? *Journal of Biomedical Semantics*, 1(2), 2011. (Cité à la page 43.)

[Coursil-2000] Jacques Coursil. *La fonction muette du langage*. Ibis Rouge, 2000. (Cité aux pages 10 et 72.)

[Cowie-1996] James R. Cowie and Wendy G. Lehnert. Information Extraction. *Commun. ACM*, 39(1) :80–91, 1996. (Cité à la page 35.)

[Cromieres-2009] Fabien Cromières. *vers un plus grand lien entre alignement, segmentation et structure des phrases*. PhD thesis, Université de Grenoble, 2009. (Cité à la page 83.)

[DeBusser-2006a] Rik de Busser and Marie-Francine Moens. *Information extraction and information technology*, pages 1–22. Springer, Berlin, Heidelberg, 2006. (Cité à la page 35.)

[DeBusser-2006b] Rik de Busser and Marie-Francine Moens. *Information Extraction from an historical perspective*, pages 23–46. Springer, Berlin, Heidelberg, 2006. (Cité à la page 35.)

[Debili-2006] F. Debili, Z.B. Tahar, and E. Souissi. Analyse automatique vs analyse interactive : un cercle vertueux pour la voyellation, l'étiquetage et la lemmatisation de l'arabe. In *Traitement Automatique des Langues Naturelles (TALN) 2006*, pages 347–356, 2006. (Cité à la page 59.)

[Denecke-2012] Kerstin Denecke. Surveillance Methods. In *Event-Driven Surveillance*, Springer Briefs in Computer Science, pages 25–53. Springer Berlin Heidelberg, 2012. (Cité à la page 31.)

[Denoual-2006] Etienne Denoual. *Méthodes en caractères pour le traitement automatique des langues*. PhD thesis, Université de Grenoble, 2006. (Cité à la page 83.)

[Doan-2008] Doan Son, Hung-Ngo Quoc, Kawazoe Ai, and Nigel Collier. Global Health Monitor - a Web-based system for detecting and mapping infectious diseases. *Proc. International Joint Conference on Natural Language Processing (IJCNLP)*, pages 951–956, 2008. (Cité à la page 43.)

[Doucet-2006a] Antoine Doucet and Miro Lehtonen. Unsupervised classification of text-centric xml document collections. In *Comparative Evaluation of XML Information Retrieval Systems, Fifth International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006*, volume 4518 of *Lecture Notes in Computer Science*, pages 497–509. Springer, 2007. (Cité à la page 67.)

[Doucet-2006b] Antoine Doucet. Advanced document description, a sequential approach. *ACM SIGIR Forum*, 40(1) :71–72, 2006. (Cité à la page 133.)

[Doucet-2010] Antoine Doucet and Helena Ahonen-Myka. An efficient any language approach for the integration of phrases in document retrieval. *International Journal of Language Resources and Evaluation*, 44(1-2) :159–180, 2010. (Cité à la page 57.)

[Doucet-2011] Antoine Doucet, Gabriella Kazai, and Jean-Luc Meunier. ICDAR 2011 Book Structure Extraction Competition. In *Proceedings of the Eleventh International Conference on Document Analysis and Recognition (ICDAR'2011)*, pages 1501–1505, Beijing, China, September 2011. (Cité à la page 152.)

[Du-2011] Mian Du, Peter Von Etter, Mikhail Kopotev, Mikhail Novikov, Natalia Tarbeeva, and Roman Yangarber. Building support tools for Russian-language information extraction. In *Proceedings of the 14th international conference on Text, speech and dialogue*, TSD'11, pages 380–387, Berlin, Heidelberg, 2011. Springer. (Cité aux pages 43 et 47.)

[ECDC-2006] ECDC. Framework for a Strategy for Infectious Disease Surveillance in Europe (2006-2008). Technical report, European Center for Disease prevention and Control, 2006. (Cité aux pages 33 et 46.)

[Eco-1985] Umberto Eco. *Lector in fabula ou La Coopération interprétative dans les textes narratifs*. Grasset, Paris, 1985. (Cité à la page 71.)

[Efimenko-2004] Irina Efimenko, Vladimir Khoroshevsky, and Victor Klintsov. Ontosminer family : Multilingual IE systems. In *SPECOM 2004 : 9th Conference Speech and Computer*, 2004. (Cité à la page 43.)

[ElGhali-2012] Adil El Ghali, Daniel hromada, and Kaoutar El Ghali. Enrichir et raisonner sur des espaces sémantiques pour l'attribution de mots-clés. In *DEFT 2012*, pages 81–93, 2012. (Cité à la page 149.)

[Elhadj-2012] Ali Ait Elhadj, Mohand Boughanem, Mohamed Mezghiche, and Fatiha Souam. Using structural similarity for clustering XML documents. *Knowledge and Information Systems*, 32(1) :109–139, juillet 2012. (Cité à la page 67.)

[Etzioni-2011] Oren Etzioni, Anthony Fader, Janara Christensen, and Stephen Soderland. Open Information Extraction : The Second Generation. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 3–10, 2011. (Cité à la page 43.)

[Evert-2008] Stefan Evert. A lightweight and efficient tool for cleaning web pages. In *Actes du 4ème Workshop Web as Corpus, LREC 2008*, 2008. (Cité aux pages 157 et 159.)

[Ferraresi-2008] Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Actes du 4ème Workshop Web as Corpus, LREC 2008*, 2008. (Cité à la page 157.)

[Ferret-2006] Olivier Ferret. Approches endogène et exogène pour améliorer la segmentation thématique de documents. *TAL*, 47, 2006. (Cité à la page 77.)

[Fort-2013] Bruno Guillaume and Karën Fort. Expériences de formalisation d'un guide d'annotation : vers l'annotation agile assistée. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 628–635, 2013. (Cité à la page 126.)

[Foudon-2008] Nadège Foudon. *L'acquisition du langage chez les enfants autistes : Étude longitudinale*. PhD thesis, Université Lyon-2, 2008. (Cité à la page 71.)

[Freifeld-2008] Clark C. Freifeld, Kenneth D. Mandl, Ben Y. Reis, and John S. Brownstein. HealthMap : Global infectious disease monitoring through automated classification and visualization of Internet media reports. *Journal of American Medical Informatics Association*, December 2007. (Cité à la page 41.)

[Gaussier-2004] E. Gaussier, J. m. Renders, I. Matveeva, C. Goutte, and H. Déjean. A geometric view on bilingual lexicon extraction from comparable corpora. In *In Proceedings of ACL-04*, pages 527–534, 2004. (Cité à la page 61.)

[Gey-2009] Fredric Gey, Jussi Karlgren, and Noriko Kando. Information access in a multilingual world : transitioning from research to real-world applications. *SIGIR Forum*, 43(2) :24–28, 2009. (Cité à la page 48.)

[Giguet-2004] Emmanuel Giguet and Nadine Lucas. *La détection automatique des citations et des locuteurs dans les textes informatifs*, pages 410–418. J. M. López-Muñoz, S. Marnette, L. Rosier, 2004. (Cité à la page 143.)

[Greimas-1970] Algirdas Julien Greimas. *Du sens, essais sémiotiques*. Éditions du Seuil, 1970. (Cité à la page 74.)

[Grice-1975] Paul Grice. Logic and Conversation. In *Syntax and semantics, vol 3.*, Syntax and semantics, vol 3. Cole, P. and Morgan, J. (eds.), 1975. (Cité à la page 71.)

[Grishman-2002] Ralph Grishman, Silja Huttunen, and Roman Yangarber. Information extraction for enhanced access to disease outbreak reports. *J. of Biomedical Informatics*, 35(4) :236–246, August 2002. (Cité à la page 43.)

[Grog-2012] Le Réseau des Groupes Régionaux d'Observation de la Grippe (Réseau des GROG). Surveillance de la Grippe en France, Avril 2012. (Cité à la page 26.)

[Grouin-2011] Cyril Grouin, Dominique Forest, Patrick Paroubek, and Pierre Zweigenbaum. Présentation et résultats du défi fouille de textes DEFT2011. In *DEFT 2011 (TALN 2011)*, pages 3–14, 2011. (Cité à la page 130.)

[Hazem-2013] Amir Hazem and Emmanuel Morin. Extraction de lexiques bilingues à partir de corpus comparables par combinaison de représentations contextuelles. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 243–256, 2013. (Cité à la page 60.)

[Heymann-2004] DAvid L.Heymann and Guenael Rodier. SARS : A Global Response to an International Threat. *Brown Journal of World Affairs*, X(2), 2004. (Cité à la page 23.)

[Hobbs-1993] Jerry R. Hobbs. The generic information extraction system. In *Proceedings of the 5th conference on Message understanding*, MUC5 '93, pages 87–91, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics. (Cité aux pages 35 et 38.)

[Hobbs-2010] Jerry R. Hobbs and Ellen Riloff. Information Extraction. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010. ISBN 978-1420085921. (Cité à la page 36.)

[Itule-2006] Bruce Itule and Douglas Anderson. *News Writing and Reporting for Today's Media*. McGraw-Hill Humanities, 2006. (Cité à la page 72.)

[Kabadjov-2013] Mijail Kabadjov, Josef Steinberger, and Ralf Steinberger. Multilingual Statistical News Summarization. In Thierry Poibeau, Horacio Saggion, Jakub Piskorski, and Roman Yangarber, editors, *Multi-source, Multilingual Information Extraction and Summarization*, Theory and Applications of Natural Language Processing, pages 229–252. Springer Berlin Heidelberg, 2013. (Cité à la page 47.)

[Kanayama-2004] Kanayama Hiroshi, Nasukawa Tetsuya, and Watanabe Hideo. Deeper sentiment analysis using machine translation technology. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Strouds-burg, PA, USA, 2004. Association for Computational Linguistics. (Cité à la page 47.)

[Kando-1999] Noriko Kando. Text Structure Analysis as a Tool to Make Retrieved Documents Usable. In *In Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages*, pages 126–135, 1999. (Cité à la page 78.)

[Karkkainen-2006] Juha Kärkkäinen, Peter Sanders, and Stefan Burkhardt. Linear work suffix array construction. *Journal of the ACM*, 53(6) :918–936, 2006. (Cité à la page 96.)

[Katsiavriades-2007] Kryss Katsiavriades and Talaat Qureshi. The 30 Most Spoken Languages of the World, 2007. (Cité à la page 45.)

[Keller-2009] Mikaela Keller, Clark Freifeld, and John Brownstein. Automated vocabulary discovery for geo-parsing online epidemic intelligence. *BMC Bioinformatics*, 10(1) :385, 2009. (Cité à la page 94.)

[Kohlschutter-2010] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 441–450, New York, NY, USA, 2010. ACM. (Cité aux pages 154 et 158.)

[Lardilleux-2010] Adrien Lardilleux. *Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle.* PhD thesis, Université de Caen, 2010. (Cité à la page 56.)

[Lecluze-2011] Charlotte Lecluze. *Alignement de documents multilingues sans présupposé de parallélisme.* PhD thesis, Université de Caen, 2011. (Cité aux pages 78 et 83.)

[Lecluze-2013] Charlotte Lecluze, Romain Brixtel, Loïs Rigouste, Emmanuel Giguet, Régis Clouard, Gaël Lejeune, and Patrick Constant. Détection de zones parallèles à l'intérieur de bi-documents pour l'alignement multilingue. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 381–394, 2013. (Cité aux pages 80 et 83.)

[Lehtonen-2007] Miro Lehtonen and Antoine Doucet. Phrase detection in the wikipedia. In *Focused access to XML documents, Sixth International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007*, volume 4862 of *Lecture Notes in Computer Science*, pages 115–121. Springer, 2008. (Cité à la page 67.)

[Lejeune-2009a] Gaël Lejeune. Ce que le texte peut dire au TAL. In *Ce que le texte fait à la phrase*, Caen, 2009. Crisco. (Cité à la page 43.)

[Lejeune-2009b] Gaël Lejeune. Structure patterns in Information Extraction : a multilingual solution ? *Advances in methods of Information and Communication Technology, AMICT09*, 11 :105–111, May 2009. (Cité à la page 43.)

[Lejeune-2010a] Gaël Lejeune, Antoine Doucet, and Nadine Lucas. Tentative d'approche multilingue en Extraction d'Information. In *JADT 2010*, pages 1259–1268. JADT, 2010. (Cité à la page 94.)

[Lejeune-2010b] Gaël Lejeune, Antoine Doucet, Roman Yangarber, and Nadine Lucas. Filtering news for epidemic surveillance : towards processing more languages with fewer resources. In *4th Workshop on Cross Lingual Information Access*, pages 3–10, 2010. (Cité à la page 94.)

[Lejeune-2011] Gaël Lejeune, Romain Brixtel, Emmanuel Giguet, and Nadine Lucas. Deft2011 : appariement de résumés et d'articles scientifiques fondé sur les chaînes de caractères. In *Défi Fouille de Textes/TALN 2011*, pages 53–64, 2011. (Cité aux pages 130 et 146.)

[Lejeune-2012a] Gaël Lejeune and Christine Durieux. Pour une approche cibliste en TAL : le cas de l'analyse automatique de la presse. In *Rhétorique et Traduction*, 2012. à paraître. (Cité à la page 71.)

[Lejeune-2012b] Gaël Lejeune, Romain Brixtel, Antoine Doucet, and Nadine Lucas. DAnIEL : Language Independent Character-Based News Surveillance. In *JapTAL*, pages 64–75, 2012. (Cité aux pages 113 et 120.)

[Lejeune-2013a] Gaël Lejeune, Romain Brixtel, Charlotte Lecluze, Antoine Doucet, and Nadine Lucas. Added-value of automatic multilingual text analysis for epidemic surveillance. In *Artificial Intelligence in Medicine (AIME)*, 2013. (Cité à la page 44.)

[Lejeune-2013b] Gaël Lejeune, Romain Brixtel, Charlotte Lecluze, Antoine Doucet, and Nadine Lucas. DAnIEL : Veille épidémiologique multilingue parcimonieuse (démonstration). In *TALN 2013*, pages 787–788, 2013. (Cité à la page 102.)

[Linge-2009] Jens Linge, Ralf Steinberger, Thomas Weber, Roman Yangarber, Erik van der Goot, Delilah Al Khudhairy, and Nikolaos Stilianakis. Internet surveillance systems for early alerting of threats. *Eurosurveillance*, 14(13), 2009. (Cité aux pages 20, 36, 41, et 47.)

[Lucas-2000] Nadine Lucas. Le rôle de la citation dans la structuration des articles de presse. In *Actes du premier colloque d'études japonaises de l'Université Marc Bloch*, pages 215–244, 2000. (Cité à la page 10.)

[Lucas-2004] Nadine Lucas. The enunciative structure of news dispatches, a contrastive rhetorical approach. *Language, culture, rhetoric*, pages 154–164, 2004. (Cité aux pages 66 et 143.)

[Lucas-2005] Nadine Lucas and Emmanuel Giguet. UniTHEM, un exemple de traitement linguistique à couverture multilingue. In *In Conférence Internationale sur le Document Electronique (CIDE 8)*, pages 115–132, 2005. (Cité à la page 95.)

[Lucas-2009a] Nadine Lucas. *Modélisation différentielle du texte, de la linguistique aux algorithmes*. PhD thesis, Habilitation à Diriger les recherches, Université de Caen, 2009. (Cité à la page 66.)

[Lucas-2009b] Nadine Lucas. Discourse Processing for Text Mining. In *Information Retrieval and Biomedicine : Natural Language Processing for Knowledge Integration*, pages 229–262. V.Prince and M.Roche, 2009. (Cité à la page 140.)

[Lucas-2012] Nadine Lucas. *Stylistic devices in the news, as related to topic recognition*, volume 26 of *Łódź, Studies in language*, pages 301–316. Peter Lang, Frankfurt am Main, 2012. (Cité à la page 10.)

[Lyon-2011] A. Lyon, M. Nunn, G. Grossel, and M. Burgman. Comparison of Web-Based Biosecurity Intelligence Systems : BioCaster, EpiSPIDER and HealthMap. *Transboundary and Emerging Diseases*, 2011. (Cité à la page 44.)

[MUC-1991] MUC. *Proceedings of the 3rd Conference on Message Understanding, MUC 1991, San Diego, California, USA, May 21-23, 1991*. ACL, 1991. (Cité à la page 35.)

[MUC-1992] MUC. *Proceedings of the 4th Conference on Message Understanding, MUC 1992, McLean, Virginia, USA, June 16-18, 1992*, 1992. (Cité à la page 35.)

[MUC-1993] MUC. *Proceedings of the 5th Conference on Message Understanding, MUC 1993, Baltimore, Maryland, USA, August 25-27, 1993*, 1993. (Cité à la page 35.)

[Maingueneau-2005] Dominique Maingueneau. *Analyser les txetes de communication*. Dunod, 2005. (Cité à la page 71.)

[McCallum-2006] Andrew McCallum. Information extraction, data mining and joint inference. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 835–845, New York, NY, USA, 2006. ACM. (Cité à la page 59.)

[McNamee-2004] Paul McNamee and James Mayfield. Character N-Gram Tokenization for European Langauge Text Retrieval. *Information Retrieval*, 2004. (Cité à la page 78.)

[Mondor-2012] Luke Mondor, John S. Brownstein, Emily H. Chan, Lawrence C. Madoff, Marjorie P. Pollack, David L.Buckeridge, and Timothy F.Brewer. Timeliness of Nongovernemental versus Governemental Global Outbreak Communications. *Emerging Infectious Diseases*, 2012. (Cité à la page 25.)

[Mougin-2013] Fleur Mougin and Natalia Grabar. Using a cross-language approach to improve the mapping between biomedical terminologies. . In *Artificial Intelligence in Medicine (AIME)*, 2013. (Cité à la page 60.)

[Mounin-1974] Georges Mounin. *Dictionnaire de la linguistique*. Presses universitaires de France, 1974. (Cité à la page 78.)

[OIF-2007] Organisation internationale de la Francophonie. *La Francophonie dans le monde 2006-2007*. Nathan, 2007. (Cité à la page 45.)

[OMS-2005] Organisation Mondiale de la Santé. *International Health Regulations (2005)*, 2005. (Cité à la page 21.)

[Pasternack-2009] Jeff Pasternack and Dan Roth. Extracting article text from the web with maximum subsequence segmentation. In *WWW*, pages 971–980, 2009. (Cité à la page 157.)

[Perkins-2010] Edward Perkins. Part of Speech tagging with NLTK Part 4 - Brill tagger VS Classifier Taggers. `http://streamhacker.com/2010/04/12/pos-tag-nltk-brill-classifier/`, 2010. (Cité à la page 41.)

[Piskorski-2011] Jakub Piskorski, Jenya Belyaeva, and Martin Atkinson. On Refining Real-Time Multilingual News Event Extraction through Deployment of Cross-Lingual Information Fusion Techniques. In *Proceedings of European Intelligence and Security Informatics Conference (EISIC), 2011, Athens, Greece.*, pages 38–45, 2011. (Cité à la page 47.)

[Pottker-2003] Horst Pöttker. News and its communicative quality : the inverted pyramid—when and why did it appear ? . *Journalism Studies*, 4(4), 2003. (Cité à la page 72.)

[Poudat-2004] Céline Poudat. Recension et présentation comparative d'étiqueteurs pour le français et l'anglais. *Texto !*, IX(4), 2004. (Cité à la page 41.)

[Poulard-2011] Fabien Poulard, Erwan Moreau, and Laurent Audibert. Vers des outils robustes et interopérables pour le TAL : la piste UIMA. In *Traitement Automatique des Langues Naturelles (TALN) 2011*, page ?, 2011. (Cité à la page 59.)

[Rastier-2002] François Rastier. Enjeux épistémologiques de la linguistique de corpus. In *2ème journées de la linguistique de corpus*, 2002. (Cité à la page 65.)

[Rastier-2008] François Rastier. Que cachent les données textuelles ? In *Actes des 9es Journées internationales d'Analyse statistique des Données Textuelles*

(JADT 2008), Lyon, 12-14 mars 2008 9es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2008). Presses Universitaires de Lyon, 2008. (Cité à la page 10.)

[Ratcliff-1988] John W. Ratcliff and David E. Metzener. Pattern matching : The gestalt approach. *Dr. Dobbs Journal*, 13(7) :46, 47, 59–51, 68–72, July 1988. (Cité à la page 162.)

[Raymond-2011] Christian Raymond and Vincent Claveau. Participation de l'IRISA à DEFT 2011 : expériences avec des approches d'apprentissage supervisé et non supervisé. In *DEFT 2011*, pages 19–27, 2011. (Cité à la page 131.)

[Reilly-2008] Aimee R. Reilly, Emily A. Iarocci, Carrianne M. Jung, David M. Hartley, and Noele P. Nelson. Indications and warning of pandemic influenza compared to seasonal influenza. *Advances in disease surveillance*, 5 :190, 2008. (Cité à la page 37.)

[Riloff-1999] Ellen Riloff and Jeffrey Lorenzen. Extraction-based text categorization : Generating domain-specific role relationships automatically. *Natural language Information retrieval*, pages 167–196, 1999. (Cité à la page 36.)

[Rouach-2010] Daniel Rouach. *La veille technologique et l'intelligence économique*. Presses Universitaires de France, 2010. (Cité à la page 7.)

[Roy-2007] Thibault Roy. *Visualisations interactives pour l'aide personnalisée à l'interprétation d'ensembles documentaires*. PhD thesis, Université de Caen, 2007. (Cité aux pages 56 et 59.)

[Shine-2012] Shine N. Das, Pramod K. Vijayaraghavan, and Midhun Mathew. Article : Eliminating noisy information in web pages using featured dom tree. *International Journal of Applied Information Systems*, 2(2) :27–34, May 2012. Published by Foundation of Computer Science, New York, USA. (Cité à la page 157.)

[Smolinski-2003] Kelly J. Henning. *Microbial Threats to Health : Emergence, Detection, and Response*. Mark S. Smolinski, Margaret A. Hamburg, and Joshua Lederberg, Washington DC : National Academy Press, 2003. (Cité à la page 21.)

[Sperber-1998] Dan Sperber and Deirdre Wilson. *Relevance : Communication and cognition*. Blackwell press, Oxford U.K, 1998. (Cité à la page 71.)

[Spousta-2008] Miroslav Spousta, Michal Marek, and Pavel Pecina. Victor : the Web-Page Cleaning Tool. In *Actes du 4ème Workshop Web as Corpus, LREC 2008*, 2008. (Cité à la page 157.)

[Steinberger-2008a] Ralf Steinberger, Flavio Fuart, Erik van der Goot, Clive Best, Peter von Etter, and Roman Yangarber. Text Mining from the web for medical intelligence. In *Mining massive data sets for security*, pages 295–310. OIS Press, 2008. (Cité aux pages 41, 42, 59, et 93.)

[Steinberger-2008b] Ralf Steinberger, Bruno Pouliquen, and Camelia Ignat. Using language-independent rules to achieve high multilinguality in Text Mining. In *Mining massive data sets for security*, pages 217–240. OIS Press, 2008. (Cité aux pages 47 et 48.)

[Steinberger-2011] Ralf Steinberger. A survey of methods to ease the development of highly multilingual text mining applications. *Language Resources and Evaluation*, pages 1–22, 2011. (Cité à la page 9.)

[Tolentino-2007] Herman Tolentino, Raoul Kamadjeu, Paul Fontelo, Fang Liu, Michael Matters, Marjorie P. Pollack, and Larry Madoff. Scanning the Emerging Infectious Diseases Horizon - Visualizing ProMED Emails Using EpiSPIDER. *Advances in disease surveillance*, 2 :169, 2007. (Cité à la page 42.)

[Tulechki-2013] Nikola Tulechki and Ludovic Tanguy. Similarité de second ordre pour l'exploration de bases textuelles multilingues. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 651–658, Les Sables d'Olonne, France, 2013. (Cité à la page 26.)

[Ukkonen-2009] Esko Ukkonen. Maximal and minimal representations of gapped and non-gapped motifs of a string. *Theorie in Computer Science*, 410(43) :4341–4349, 2009. (Cité à la page 80.)

[VanDijk-1988] T.A Van Dijk. *News as discourse*. Lawrence Erlbaum Associates, Hillsdale N.J, 1988. (Cité à la page 143.)

[Vergne-2004a] Jacques Vergne. Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource. In *Journées d'Analyse des Données Textuelles (JADT)*, pages 1157–1163, 2004. (Cité à la page 56.)

[Vergne-2004b] Jacques Vergne. Un exemple de traitement "alingue" endogène : extraction de candidats termes dans des corpus bruts de langues non identifiées par étiquetage mot vide - mot plein, 2004. (Cité à la page 56.)

[Vieira-2006] Karane Vieira, Altigran S. da Silva, Nick Pinto, Edleno S. de Moura, João M. B. Cavalcanti, and Juliana Freire. A fast and robust method for web page template detection and removal. In *ACM international conference on Information and knowledge management*, CIKM '06, pages 258–267, New York, NY, USA, 2006. ACM. (Cité à la page 157.)

[Wendt-1979] Lloyd Wendt. *Chicago Tribune : The Rise of a Great American Newspaper*. Rand McNally (Chicago), 1979. (Cité à la page 72.)

[Wilkens-2008] Matt Wilkens. Evaluating POS Taggers : Speed. `http://mattwilkens.com/2008/11/08/evaluating-pos-taggers-speed/`, 2008. (Cité à la page 41.)

[Wilson-2007] James M. Wilson. Argus : A Global Detection and Tracking System for Biological Events. *Advances in disease surveillance*, 4 :1953, 2007. (Cité à la page 40.)

[Wilson-2008] James M. Wilson. Golbal Argus - Indications and Warnings to Detect and Track Biological Events. Oral Presentation, 2008. (Cité à la page 40.)

[Wilson-2009] Kumanan Wilson, Christopher McDougall, and Alan Forster. The responsibility of healthcare institutions to protect global health security. *Health Quarterly*, 12(1) :56–60, 2009. (Cité aux pages 21 et 23.)

[Yangarber-2008] Roman Yangarber, Peter von Etter, and Ralf Steinberger. Content Collection and Analysis in the Domain of Epidemiology. *Proceedings of DrMED-2008 : International Workshop on Describing Medical Web Resources*, 2008. (Cité aux pages 8, 34, et 36.)

[Yangarber-2011a] Silja Huttunen, Vihavainen Arto, Peter von Etter, and Roman Yangarber. Relevance Prediction in Information Extraction using Discourse and Lexical Features. In *Nordic Conference on Computational Linguistics, Nodalida 2011*, pages 114–121, 2011. (Cité aux pages 65 et 124.)

[Yangarber-2011b] Roman Yangarber. Discovering Complex Networks of Events and Relations in News Surveillance. In *EISIC'2011*, pages 7–7, 2011. (Cité à la page 35.)

Table des figures

Liste des tableaux

Veille épidémiologique multilingue : une approche parcimonieuse au grain caractère fondée sur le genre textuel

Cette thèse explore la problématique du multilinguisme en recherche d'information. Nous présentons une méthode de veille sur la presse adaptée au traitement du plus grand nombre de langues possible. Le domaine spécifique d'étude est la veille épidémiologique, domaine pour lequel une couverture la plus large possible est nécessaire. La méthode employée est différentielle, non-compositionnelle et endogène. Notre but est de maximiser la factorisation pour traiter de nouvelles langues avec un coût marginal minimal. Les propriétés du genre journalistique sont exploitées, en particulier la répétition d'éléments à des positions clés du texte. L'analyse au grain caractère permet d'être indépendant des contraintes posées par le mot graphique dans de nombreuses langues. Nous aboutissons à l'implantation du système DAnIEL (*Data Analysis for Information Extraction in any Language*). DAnIEL analyse les documents pour déterminer s'ils décrivent des faits épidémiologiques et les regrouper par paires maladie-lieu. DAnIEL est rapide et efficace en comparaison des systèmes existants et nécessite des ressources très légères. Nous montrons d'autres applications de DAnIEL pour des tâches de classification et d'extraction de mots-clés dans des articles scientifiques. Enfin, nous exploitons les résultats de DAnIEL pour évaluer des systèmes de nettoyage de page web.

Mots-clés : Langage naturel, traitement du (informatique) ; Multilinguisme ; Recherche d'information ; Recherche de l'information ; Extraction d'Information.

Multilingual epidemic surveillance : a parsimonious caracter-based approach

In this dissertation we tackle the problem of multilingual epidemic surveillance. The approach advocated here which is differential, endogenous and non-compositionnal. We maximise the factorization by using genre properties and communication principles. Our local analysis does not rely on classical linguistic analyzers for morphology, syntax or semantics. The distribution of character strings at key positions is exploited, thus avoiding the problem of the definition of a "word". We implemented DAnIEL (Data Analysis for Information Extraction in any Language), a system using this approach. DanIEL analyzes press articles in order to detect epidemic events. DAnIEL is fast in comparison to state-of-the-art systems. It needs very few additional knowledge for processing new languages. DAnIEL is also evaluated on the analysis of scientific articles for classification and keyword extraction. Finally, we propose to use DAnIEL outputs to perform a task-based evaluation of boilerplate removal systems.

Keywords : Natural Language Processing ; Information Extraction ; Multilingualism ; Information Retrieval.

Discipline : Informatique et applications

Laboratoire GREYC Campus Côte de Nacre — BP 5186 — 14032 Caen Cedex