HAL
open science

Construction guidée de graphes de transducteurs pour l'extraction d'évènements spatio-temporellement localisés

Manal El Zant, Liliane Pellegrin, Michel Roux, Chaudet Hervé

To cite this version:

Manal El Zant, Liliane Pellegrin, Michel Roux, Chaudet Hervé. Construction guidée de graphes de transducteurs pour l'extraction d'évènements spatio-temporellement localisés. IC - 16èmes Journées francophones d'Ingénierie des Connaissances, May 2005, Nice, France. pp.109-120. hal-01023898

HAL Id: hal-01023898

https://inria.hal.science/hal-01023898v1

Submitted on 15 Jul 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Construction guidée de graphes de transducteurs pour l'extraction d'évènements spatio-temporellement localisés

Manal EL Zant[1], Liliane Pellegrin[1], Michel Roux[1], Hervé Chaudet[1,2]

[1]Laboratoire d'Informatique Fondamentale, UMR CNRS 6166
Équipe BIM, Faculté de médecine, 27 Bd Jean – Moulin,
13005 Marseille
[2]Unité de Recherche Epidémiologique, Département de Santé Publique,
Institut de Médecine Tropicale du Service de Santé des Armées,
13998 Marseille Armées
{el.zant, liliane.pellegrin, michel.roux}@medecine.univ-mrs.fr
lhcp@acm.org

Résumé : Dans le cadre du traitement des dépêches épidémiologiques pour un système d'aide à la décision, nous présentons notre approche de l'extraction de ces informations. Celle-ci repose sur des graphes de transducteurs construits par le logiciel INTEX qui sont bâtis à partir d'une théorie de représentation spatio-temporelle des évènements obtenue lors d'un travail préalable, et non pas d'une grille a priori. Ces graphes ont été appliqués sur un corpus de 100 dépêches ProMed pour extraire les évènements et leurs caractéristiques spatio-temporelles. Si ces graphes permettent effectivement d'extraire les informations spatio-temporelles, des difficultés d'application sur des macro-évènements propres au domaine épidémiologique restent à résoudre.

Mots-clés : Analyse de corpus textuel, extraction des connaissances, concept spatio-temporel, INTEX, Représentation des évènements (STEEL).

1 Introduction

Dans le cadre de la pratique de la médecine des voyages et de l'épidémiologie, de nombreuses ressources Internet mettent régulièrement des informations concernant les phénomènes épidémiologiques dans le monde à disposition des experts. Parmi celles-ci la liste de diffusion internationale ProMED-Mail (http://www.promedmail.org) transmet régulièrement des informations sous la forme de dépêches, qui sont des textes spécialisés de longueur variable décrivant l'apparition, l'évolution et les caractéristiques épidémiologiques de tels phénomènes. Les informations épidémiologiques sont utilisées par des médecins, dans le cadre d'une activité générale de veille épidémiologique, qui correspond à la collecte et l'évaluation systématiques d'informations d'intérêt médical dans l'objectif de pouvoir déterminer les risques associés à un triplet <population,lieu,date>. Cela permet de déduire les conduites préventives associées à un déplacement individuel ou de population, ou de surveiller les émergences de maladies et leurs conséquences. Ce

raisonnement de haut niveau, qui s'effectue généralement dans un contexte où les contraintes temporelles sont fortes, est orienté par les connaissances et conduit à la construction d'une représentation adéquate de la situation afin de prendre la mesure des risques et d'en déduire les décisions adéquates (Chaudet, Pellegrin et Rech, 2000). Les représentations utilisées pour décrire les phénomènes épidémiologiques, rapportés dans le cas présent par les dépêches ProMED-Mail, sont donc complexes car elles doivent prendre en compte des informations de nature diverse, aussi bien des connaissances médicales, zoologiques, que socio-économiques ou encore géographiques et temporelles.
Le besoin de systèmes de question-réponse utilisant ce type de ressources textuelles semblerait une réponse « naturelle » en proposant à ces médecins d'alléger leur tâche de recherche d'information. De tels systèmes leur permettraient d'obtenir des réponses adéquates sur les phénomènes épidémiques sans avoir à parcourir un nombre croissant de dépêches disponibles. En effet, le poids de plus en plus important de la recherche et de l'acquisition de nouvelles informations est une constante dans les différentes spécialités médicales qui justifie ainsi le développement de ce type de système (Jacquemart et Zweigenbaum, 2003 Zweigenbaum, 2003). Dans ce cadre, l'objectif du projet EpidemIA est de bâtir un système d'aide à la décision utilisant toutes les caractéristiques des épidémies décrites dans les dépêches pour assister l'utilisateur dans son activité de gestion des risques sanitaires (figure 1).

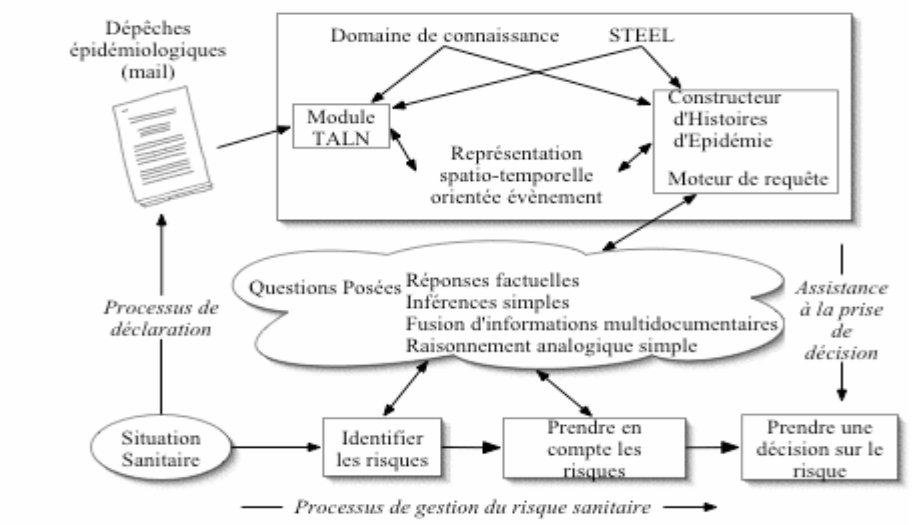

Fig.1 : Organisation générale du système EpidemIA

Ce projet comporte actuellement plusieurs volets :

- La mise au point d'un langage logique de représentation des concepts liés aux épidémies et de l'ontologie qui décrit ces concepts. Le résultat est le modèle général d'une épidémie.
- L'extraction des informations à partir des dépêches dans le module TALN

- La représentation de ces informations dans le langage logique.

Ces différents volets s'insèreront dans un système de question-réponse, qui comportera plus classiquement des modules d'analyse de questions et de génération de réponses aux utilisateurs. Dans le cas présent, il s'agira d'un système QR de domaine fermé puisqu'il est « uniquement » centré sur les phénomènes épidémiologiques (Doan-Nguyen et Kosseim, 2004).

L'hypothèse sous-jacente à ce projet est que les représentations adéquates des évènements épidémiques décrits dans ces textes ne peuvent se construire que dans le cadre d'un système fortement orienté par les connaissances du domaine. Nous avons voulu nous appuyer sur l'ingéniérie des connaissances pour construire le système en commençant par mettre au point le formalisme de représentation logique des connaissances afin de s'appuyer sur ce dernier pour bâtir la solution de TALN (traitement automatique de la langue naturelle). Un travail préalable nous a permis de développer un langage formel de représentation des connaissances (STEEL) adapté à la problématique des informations épidémiologiques, tenant compte à la fois de l'orientation évènementielle des récits, de la compositionalité des évènements et de leur localisation spatio-temporelle (Chaudet, 2004). Le formalisme de représentation logique ayant ainsi été défini préalablement, il restait à mettre au point le mécanisme de traitement des textes pour obtenir leur représentation.

L'objectif de cet article est de présenter la solution que nous avons choisie en TALN pour traiter des dépêches épidémiologiques afin de construire une représentation du contenu de ces textes dans le langage formel. Que ce soit pour le traitement des réponses, celui des documents sélectionnés, ou encore la construction d'une réponse appropriée, le choix entre divers outils d'extraction d'informations ou de TALN est un enjeu majeur auquel se trouvent confrontés les systèmes de question-réponse. Dans notre cas, nous avons choisi d'utiliser des transducteurs à état finis lors de l'analyse des dépêches, ici le système INTEX (Silberztein, 1993). La construction de ces transducteurs a été guidée par le langage de représentation STEEL pour "identifier" ou "valider" les structures sémantiques, aboutissant ainsi à une construction du module de TALN guidée par l'ingéniérie des connaissances. Dans ces structures, les informations syntaxico-sémantiques sont les éléments sémantiques nécessaires à la construction de la représentation spatio-temporelle des évènements épidémiques par STEEL.

2 Méthodes et contexte

2.1 STEEL : le langage de représentation des connaissances

Il existe de nombreux formalismes permettant de raisonner sur le temps en fonction des concepts de base (situation, événements, actions, chroniques) et la façon de représenter le temps (instants ou intervalles, temps linéaire ou non). Si le calcul des événements (Kowalski et Sergot, 1986) est un modèle déjà bien utilisé en médecine, il ne l'a été que dans le cadre de la clinique, et jamais pour la veille épidémiologique. Par ailleurs, la localisation spatiale n'a été que peu abordée et uniquement dans le cas

de la localisation de lésions. STEEL (Spatio-Temporal Extended Event Language) (Chaudet, 2004) est un langage typé en logique du premier ordre qui est une extension du calcul des événements spécialement développée pour la modélisation qualitative d'événements épidémiologiques. Ses principales caractéristiques sont de réifier simultanément les coordonnées spatiales et temporelles des événements sous la forme de localisations spatio-temporelle basée sur des régions pouvant être référencées par des noms, et de pouvoir représenter les regroupements d'événements sous la forme d'agrégats résultant de l'application d'opérateurs de combinaisons tout en maintenant la validité spatio-temporelle de l'agrégat. À l'occasion de la création de ces agrégats, il sait créer les regroupements corrects des temps et des lieux. Au lieu de la forme traditionnelle *happens(action,time)*, STEEL utilise la représentation *happens(macroevent, <time, place>)* où *macroevent* est un regroupement d'événements élémentaires. Cette extension permet d'obtenir une représentation très proche du récit, centrée sur les événements et conservant l'organisation de ces derniers en réseaux de dépendance et d'inclusion, ce qui permet de représenter une épidémie comme un événement complexe (mécanisme d'abstraction) correspondant à l'agrégation des événements qui la compose. L'objectif alors est de pouvoir interpréter naturellement, et désigner des évènements complexes faisant intervenir le temps ainsi que le lieu, dans un langage de programmation en logique du premier ordre. Trois composantes du discours doivent donc être identifiées et représentées de façon coordonnée dans le langage de représentation : l'évènement (simple ou complexe), le temps et le lieu.

2.2 INTEX : un système de transducteurs à états finis

INTEX est un outil d'analyse, qui inclut des dictionnaires de type DELA et des grammaires, permettant l'exploration des textes à des fins d'études sur corpus, tout en utilisant une interface graphique pour construire des automates à états finis à types de transducteurs pour faire une analyse de texte. Il autorise une approche de traitement partiel qui recherche dans les textes des motifs particuliers au moyen d'automates et de relations entre ces motifs. INTEX est adapté à la construction d'extracteurs et de « shallow-parsers ». Il a la particularité de fournir aux utilisateurs un ensemble complet d'outils permettant de construire rapidement, et de gérer des centaines de grammaires locales, ainsi qu'une douzaine d'outils originaux permettant la vérification de chaque grammaire, la vérification de la cohérence entre plusieurs grammaires, leur déboguage incrémental, etc. De plus sa capacité à créer des transducteurs permet de générer des informations qui modifient le texte d'entrée (p.e. tagging).

2.3 Annotation temporelle et spatiale des événements

L'annotation précise et détaillée des expressions temporelles a commencé avec les conférences MUC 5-7 (Message Understanding Conferences) pour l'identification et la classification des entités nommées EN (Chinchor 1999). Dans la même vision Ferro et al. (2001) décrivent un ensemble de guidelines pour l'annotation des

expressions temporelles, à partir des plusieurs langues, et leur associent une représentation canonique du temps auxquels elles se réfèrent. Cependant, une autre approche d'annotation a aussi été utilisée. C'est le marquage temporel, qui vise à associer un temps du calendrier à certains ou tous les événements du texte. Filatova et Hovy (2001) décrivent une méthode pour fractionner des phrases en leurs événements constitutifs et leur assigner des marqueurs temporels. Le marquage utilise deux temps principaux : le temps de l'article et le dernier temps indiqué dans la même phrase. Dans cette approche Schilder et Habel (2001) ont développé un système d'étiquetage sémantique des expressions temporelles. Elles sont classifiées selon deux types : celles qui se rapportent à un temps du calendrier ou d'horloge et celles qui se rapportent à des événements. L'ensemble des relations temporelles proposées est équivalent aux relations d'Allen (1983). Une troisième approche (Setzer et Gaizauskas, 2000) se centre sur les relations temporelles entre les événements et le temps ou entre les événements. Cette approche prend l'identification des relations temporelles comme but, et repose sur la façon dont l'information temporelle se présente ainsi que sa relation avec le texte. Leur schéma permet de déterminer l'ordre relatif ou le temps absolu des événements. Katz et Arosio (2001) à leur tour ont proposé une annotation des informations des relations temporelles en se basant sur les relations entre évènement. Notre approche se rattache à la deuxième catégorie de travaux. L'annotation est pratiquée en tenant compte du temps de la dépêche et de celui signalé dans le récit. L'association entre l'événement et le marquage temporel se fait ultérieurement au niveau de la représentation logique.
Nous employons le terme localisation spatiale pour désigner les lieux géographiques par leurs noms de lieu (toponymes). Cette localisation spatiale peut être désignée par : des noms des villes, de villages, de provinces, des hôpitaux, des laboratoires, des aéroports, etc. Toute identification de la localisation spatiale nous intéresse puisque selon STEEL la localisation est un symbole attaché à une région de l'espace. L'annotation de la localisation spatiale, comme la temporelle relève du traitement du langage naturel MUC. Chinchor (1999) décrit une annotation, selon MUC, des différentes entités nommées où la localisation spatiale est identifiée par le type *Location*.

3 Application d'INTEX sur le corpus et résultats obtenus

Notre approche permet d'analyser des membres de phrases qui peuvent comporter une séquence longue relative à un concept donné. Nous avons procédé en 6 étapes :

- Identification des mots caractérisant les trois groupes de concepts décrivant les évènements épidémiques : type, localisation géographique et localisation temporelle de l'évènement.
- Construction des dictionnaires spécifiques pour les mots caractérisant chaque groupe, en ajoutant aux informations flexionnelles et lexicales, une information sémantique. Ex Géographie, Pathologie, Biologie,…
- Pour chaque groupe de concepts, sélection par INTEX dans le corpus des membres de phrases comportant un élément du groupe.

- Analyse de la configuration syntaxique et sémantique qui entoure l'élément choisi et identification le sous langage associé à l'élément choisi. INTEX construit les graphes correspondant aux membres de phrases sélectionnés, mais donne ainsi plusieurs solutions. Les ambiguïtés sont nombreuses, dues aux dictionnaires non spécifiques du domaine. Il est donc nécessaire de construire ces graphes après une analyse humaine apportant des précisions aux lexiques.
- Construction des graphes correspondant au sous langage caractérisant chaque concept tout en testant leurs performances.

3.1 Les différents types de graphes et leurs applications sur le corpus

Le calcul de la référence spatio-temporelle joue un rôle important dans la construction de leur représentation. Notre conception du raisonnement spatio-temporel permettrait de répondre automatiquement à deux questions principales suivantes : « Quand l'événement est-il survenu? », « Où l'événement est-il survenu? ». Il s'agit d'accorder à cette orientation deux fonctions. La première consiste en l'analyse du sous-langage, vue comme l'étape initiale qui permet d'accéder aux formes spatio-temporelles de notre corpus. La seconde, qui est conditionnée à la possibilité de représenter formellement la signification des expressions, définit et formalise la sémantique d'un certain nombre d'expressions du langage des dépêches en termes spatio-temporels. Finalement, il devrait être possible de valider ces représentations en répondant à des questions portant sur le contenu de la représentation des évènements (comme les localisations spatiales et temporelles).

3.2 Le graphe de localisation temporelle

Pour créer ce graphe, nous avons étudié les différentes formes d'expressions temporelles rencontrées dans les 100 dépêches. Elles ont été regroupées en deux catégories principales que nous détaillons ici.

Premièrement, des formes langagières spécifiques aux dépêches ont été identifiées. Il s'agit de plusieurs formats non littéraires, de style rédactionnel abrégé comme les expressions suivantes : *10 Apr 2003, 10/Apr/2003, Friday [10 Apr 2003], Friday, 10/04/03, 10 Apr 2003, 2003/04/10, 10-Apr-2003 etc…*

Un graphe de transducteurs a été construit pour identifier ces cas d'expressions temporelles (Fig.2). En particulier, l'étiquette **Month($MoisNum)** entraîne l'écriture du prédicat **Month** dans le texte d'origine, avec en argument la valeur de la variable **MoisNum** qui a été retrouvée. Dans l'échantillon de 100 dépêches, 6284 séquences sont reconnues par ce graphe. Nous présentons ci dessous quelques séquences reconnues, ainsi que leurs équivalents en mode de remplacement. L'inconvénient majeur de ce graphe est que nous ne pouvons supprimer quelques cas restants d'ambiguïtés comme, par exemple, le prénom d'une personne identique au nom d'un mois (3 cas identifiés, de type *Jun Wang).*

Deuxièmement, les cas d'expressions temporelles présentant des formes plus classiquement littéraires ont été identifiées dans notre corpus comme *In mid February, after several days, during the second week of February, last Friday night.*

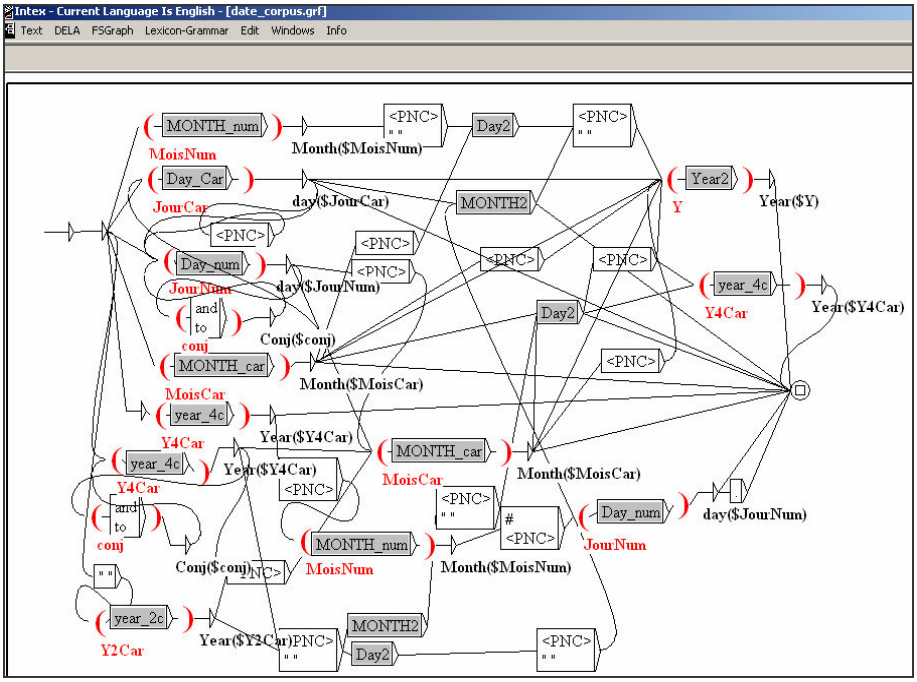

Fig. 2- Annotation des expressions temporelles

Séquence reconnues	Application du Mode de remplacement
Acute Respiratory Syndrome - Worldwide 20030315.0637 Acute respiratory …	RespiratorySyndrome – Worldwide Year(2003)Month(03)day(15) 0637Acute…
..an outbreak, as happened in February [2003]. However, the government	seed an outbreak, as happened in Month(February)Year(2003)]. However, the….
SARS patients still in the hospital on Friday [18 Apr 2003], he said, 235 of them	patients still in the hospital on day(Friday (day(18)Month(Apr)Year(2003)], he said,
…acute respiratory syndrome, and by 11-Mar-2003 similar outbreaks had been …	…acute respiratory syndrome, and by day(11)Month(Mar)Year(2003) similar…

Table 1. Exemples de séquences reconnues de résultats en mode de remplacement pour la localisation temporelle

Pour cela, nous avons bénéficié de la bibliothèque de graphes mise au point par Maurice Gross et disponible sur le site d'INTEX. Appliqués sur ces dépêches, ces graphes identifient toutes les formes littéraires de la langue anglaise présentes dans ce corpus. Cependant, il reconnaît à tort certaines expressions. Comme par exemple: « *that may have identified, from 16 countries, in a second Hong Kong hospital, in 65 cases, in terms of industries affected, on the 9th floor of the Metropole Hotel.* »

Pour réduire ces cas d'erreurs, nous avons fait des modifications dans les graphes concernés qui, ainsi modifiés, s'adaptent mieux au langage professionnel utilisé dans les dépêches.

Séquence Reconnues par INTEX	Résultat en mode de remplacement
in the evening on 10 Apr 2003	ExpTemp(in the evening)day(10)Month(Apr)) Year(2003),
, Monday evening [24 Mar 2003]	day(Monday)ExpTemp(Monday evening)day(24) Month(Mar) Year(2003),
since Monday [21 Apr 2003]	day(Monday)ExpTemp(since Monday)day(21) Month(Apr)Year(2003),
, the same day (14 Apr 2003),	ExpTemp(the same day)day(14)Month(Apr) Year(2003),

Table 2. Exemples de séquences mixtes reconnues de résultats en mode de remplacement pour la localisation temporelle

Le graphe final qui englobe l'ensemble des formes associées aux expressions temporelles des dépêches est composé des deux graphes cités ci-dessus. 7087 cas ont pu être identifiés dans notre corpus par l'application de ce graphe, qui permet de traiter les cas incluant des expressions littéraires et non littéraires.

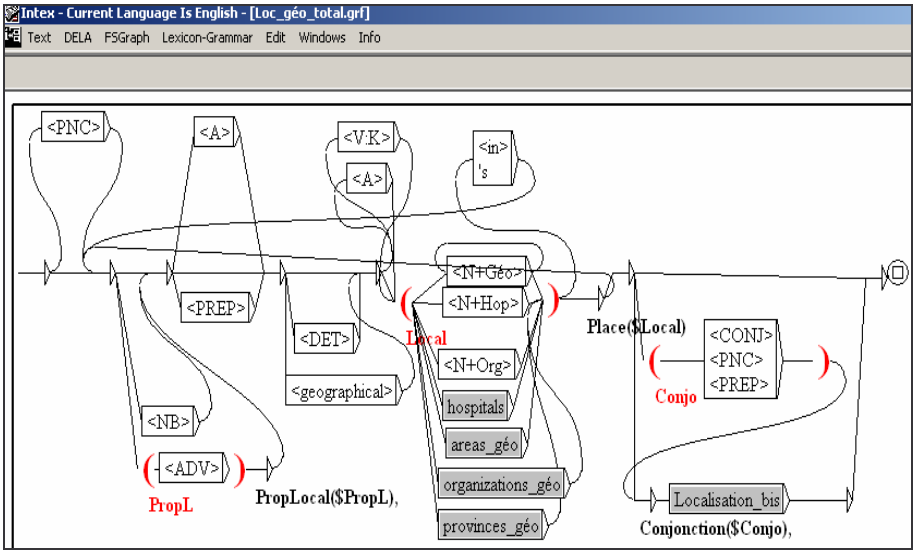

Fig.3 – Extraction des expressions spatiales.

3.3 Le graphe de localisation spatiale

Pour la création de ce graphe nous avons étudié les différentes indications géographiques existantes dans les dépêches. L'analyse nous a révélé que trois éléments sont regroupés :

1. le site géographique au sens classique comme le Nom de ville, de région, d'un pays, etc. *(Hong Kong, China, Hanoi, Canada, Toronto...);*
2. le nom d'un hôpital, d'une clinique, d'une laboratoire, etc. (*Kwong Wah Hospital, Alice Ho Miu Ling Nethersole Hospital (AHNH)*;
3. le nom d'un organisme *(Department of Health(DH), Ministry of Health (MOH), Centers for Disease Control and Prevention (CDC)).*

Nous avons typé les mots correspondant avec un attribut sémantique dans les dictionnaires. L'analyse du sous langage correspondant (grammaire locale) met en évidence la position des déterminants, des adverbes, des adjectifs ainsi que la possible association de plusieurs localisations spatiales, reliées ou non par des conjonctions ou des ponctuations (Fig.4). Par exemple : « *7 patients died among Alice Ho Miu Ling Nethersole Hospital (AHNH), Kwong Wah Hospital, Princess Margaret Hospital (PMH), Queen Elizabeth Hospital (QEH), Tseung Kwan O Hospital (TKO)and Tuen Mun Hospital (TMH) ».*

L'étiquette **Place ($Local)** provoque l'écriture du prédicat **Place** avec un argument, la valeur de la variable **Local**, valeur trouvée dans le texte. Ce graphe est récursif pour permettre la reconnaissance des plusieurs lieux cités successivement. Dans l'échantillon de 100 dépêches, 14656 séquences sont reconnues par ce graphe.

Séquence Reconnues par INTEX	Résultat en mode de remplacement
hospital in Hohhot, China ;	Place(hospital in Place(Hohhot)) Conjonction($Conjo),Localisation2(China);
Centers for Disease Control and Prevention in Atlanta;	Place(the Centers for Disease Control and Prevention in Localisation2(Atlanta));
BEIJING - The World Health Organization;	Place(BEIJING) Conjonction($Conjo),Org(The World Health Organization),;
from Guangdong province to Hong Kong.	Place(Guangdong province) Conjonction($Conjo),Localisation2(Hong Kong),;

Table 3. Exemples de séquences reconnues de résultats en mode de remplacement pour la localisation spatiale.

3.4 Les graphes d'évènements épidémiques

Le problème est nettement plus complexe. De nombreux évènements attendus par STEEL, sont évoqués par des verbes d'action comme « admettre », « hospitaliser » *(Admit, hospitalize).* D'autres peuvent être identifiés par des noms communs issus du langage professionnel comme « épidémie » ou d'autres plus généraux mais appliqués dans un contexte linguistique spécifique comme le terme « lien » dans l'exemple suivant : *So far, no **link** has been found between these cases and the outbreak in Hanoi.* C'est bien « *link* » qui caractérise le fait de « la non-association entre les cas observés et l'épidémie en cours à Hanoi ». Nous avons abordé en priorité les verbes d'action à partir de la liste de tous les verbes trouvés dans le corpus (2854 formes verbales, flexions comprises). Ainsi, autour de ces verbes d'action, on trouve, outre

des notions de temps et de lieu qui sont traités à part, des descriptions de pathologies, de patients, de nombre de cas. Actuellement nous avons traité seulement 94 formes verbales différentes en 40 graphes.

A titre d'exemple, le graphe de la figure 4 représente le verbe d'action « *admit* » et décrit les différentes associations d'informations possibles autour du verbe. Ce graphe peut identifier des séquences comme : *admit new patients, admitted as suspect SARS cases, admitted for observation etc.* Il sélectionne 332 séquences dans notre échantillon de 100 dépêches.

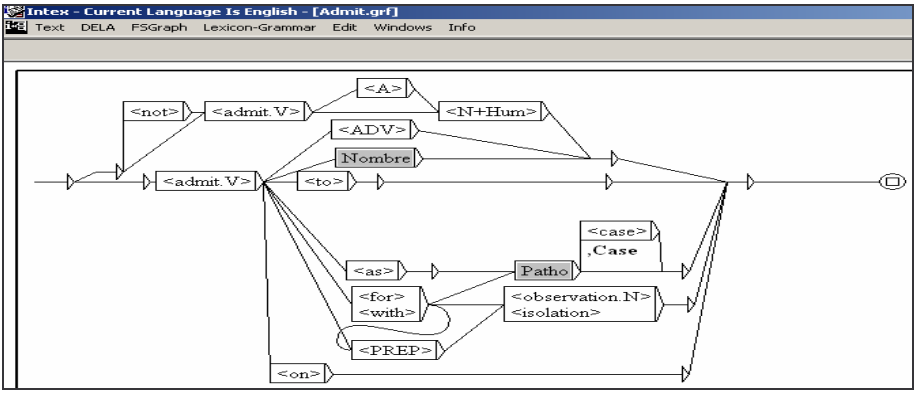

Fig.4 – Extraction des évènements engendrés par l'événement « Admission » associé au verbe *Admit*.

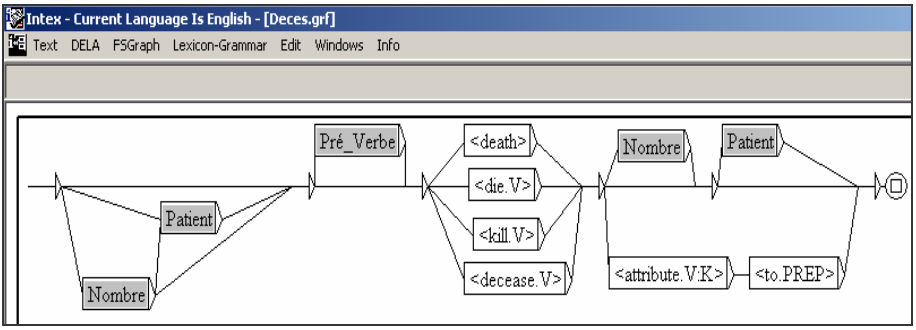

Fig.5 - Extraction des évènements engendrés par l'évènement « décès » associé aux expressions *death/kill/die…*

Un autre événement identifié dans les textes est le fait « décès » qui prend plusieurs formes linguistiques possibles (*death, kill, decease..*), elles-mêmes associées à d'autres informations comme « patients », « nombre d'individus » le type de personnes affectées, etc.. (*patient*, *34 poeple*, *46-years-old female).* Le graphe représentant les différentes formes de « Décès » a sélectionné dans le corpus 894 séquences. Voici quelques séquences sélectionnées : *a 46-year-old female died, an American businessman died, deaths Health care workers, more than 50 deaths, have been deaths in individuals, has killed at least 34 people, at least 9 deaths.*

4 Conclusion et Discussion

Les précédents travaux similaires portant sur le traitement des dépêches épidémiologiques (Damiano et col., 2002, Grishman et col., 2002) reposaient sur des systèmes d'extraction d'information demandant la définition à priori de grilles d'information utilisées par des automates d'identification de formes régulières. Cette approche limite à ces seules grilles l'information qui est extraite, ce qui restreint les services qui peuvent être rendus. Nous avons préféré introduire un certain relâchement dans les contraintes d'extraction d'information en faisant reposer la représentation non plus sur une grille mais sur une théorie, et nous avons répercuté les contraintes de cette théorie sur le dispositif d'analyse. Nous avons pu extraire un ensemble d'expressions temporelles et spatiales présentes dans le corpus, qui sont la base des descriptions de macro-évènements du langage STEEL. L'avantage non négligeable d'INTEX est qu'il se fonde sur l'unité lexicale et applique ses transducteurs phrase à phrase. Cependant, de nombreux problèmes restent à résoudre. Nos premiers résultats montrent aussi la difficulté de composer des évènements complexes, spécifiques au domaine d'application concerné, à partir d'éléments syntaxiques comme les verbes ou les noms communs même si ceux-ci se rapportent à de concepts majeurs dans l'ontologie du domaine. En effet, non seulement le système doit être capable d'identifier des évènements, mais il doit aussi organiser les liens entre ces évènements et la composition entre certains d'entre eux, dans un sens correct vis à vis du domaine. D'autant que ces évènements ne sont pas systématiquement présents dans une seule phrase, ni même un seul paragraphe ou encore un seul texte mais peuvent se retrouver dans plusieurs textes. En effet, la description de l'évolution d'une épidémie et de ses caractéristiques cliniques, de diffusion dans une population, ne se fait pas en une seule dépêche mais au fil de l'ensemble des dépêches la concernant. Nous devons donc dans l'avenir élaborer une procédure spécifique qui transformera les résultats fournis par les transducteurs d'INTEX en langage STEEL. Ce dernier sera à même de synthétiser les informations issues de plusieurs dépêches.

Une solution serait, afin d'améliorer l'identification de ces différentes expressions, de passer à une forme générale, qui est obtenue grâce à la fonction de génération des transducteurs. On substituerait à toutes séquences d'origine, des séquences générées à partir de marqueurs définis (figés) dans les graphes et de mots ou marqueurs sémantiques récupérés par INTEX dans des variables, à partir du texte d'origine. Cette forme serait alors, dans une prochaine, l'étape traduite dans le langage STEEL.

Une autre solution serait aussi de revenir vers une optique plus classique en TALN qui utiliserait des grammaires de type LinkParser permettant de faire plus facilement le lien entre les différentes structures syntaxiques support aux informations recherchées.

Références

ALLEN, J.F. (1983). Maintaining Knowledge about Temporal Intervals. *Communications of the ACM, 26*, 11, p. 832-843.

CHAUDET H. (2004). Une extension du Calcul des Evènements pour la representation de récits épidémiologiques. *15 ème journées francophones d'Ingénierie des Connaissances IC'2004,* p.285-296.

CHAUDET H., PELLEGRIN L., RECH M. (2000). Polymorphisme des besoins d'informations dans le cadre d'une consultation assistée par hypertexte, in M. FIESCHI, O. BOUHADDOU, R. BEUSCART, R. BAUD, Eds;, *L'informatique au service du patient,* Comptes rendus des Huitièmes Journées Francophones d'Informatique Médicale Marseille, 30-31 Mai 2000, Springer Verlag Editions. p.157-166. France.

CHINCHOR N., BROWN, E., FERRO L., ROBINSON P. *(*1999). Named Entity Recognition Task Definition, *MITRE, 1999.*

DAMIANOS L., DAY D. et col. (2002). Real users, real data, real problems : the MiTAP system for monitoring bio events. *Proceedings of BTR2002.* The University of Mexico, March 2002.

DOAN-NGUYEN H. & KOSSEIM L. (2004). Amélioration de la précision dans un système de question-réponse de domaine fermé. *Actes des 7èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT-2004)*. p. 325-333. Louvain-la-Neuve.

FERRO, L., MANI, I., SUNDHEIM, B., WILSON, G. *(2001).* TIDES Temporal Annotation Guidelines Draft - Version 1.02. *MITRE Technical Report MTR MTR 01W000004. McLean, Virginia: The MITRE Corporation.*

FILATOVA E. & HOVY E.H. (2001). Assigning Time-Stamps to Event-Clauses. *In Proceedings of the Workshop on Temporal and Spatial Reasoning at the Conference of the ACL. Toulouse, France.*

GRISHMAN R., HUTTUNEN S., YANGARBER R. (2002), Information extraction for enhanced access to disease outbreak reports *Journal of Biomedical Infoirmatics*. 35. p.236-46.

JACQUEMART P. & ZWEIGENBAUM P. (2003). Towards a medical question-answering system: a feasibility study. In R. BAUD, M. FIESCHI, P. LE BEUX, P. RUCH, Eds. Actes du colloque *The New Navigators: from Professionals to Patient Medical Informatics Europe*, Studies in Health Technology and Informatics, IOS Press, vol.95, p. 463-468, Amsterdam.

KATZ G. & AROZIO, F. (2001). The Annotation of Temporal Information in Natural Language Sentences. *Workshop on Temporal and Spatial Information Processing at the 39th Annual Meeting of the Association for Computational Linguistics (ACL'2001),* Toulouse, p. 104--111. France.

KOWALSKI R. & SERGOT M. (1986). A Logic-based Calculus of Event, *New Generation Computing,* 4, p.67-95

LAVENUS K. & LAPALME G. (2002) Evaluation des systèmes de question réponse. Aspects méthodologiques. *Traitement automatique des langues*, 43(3), p. 181-208.

SCHILDER F. & HABEL C. (2001). From Temporal Expressions To Temporal Information: Semantic Tagging Of News Messages. *In Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing, ACL-2001. Toulose, France, 6-11 July. pp. 65-72.*

SETZER A. & GAIZAUSKAS G. (2000). Annotating Events and Temporal Information in Newswire Texts. *In Proceedings of the Second International Conference on Language Resources and Evaluation,* p. 1287-1294.

SILBERZTEIN M. (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*. Masson: Paris.

ZWEIGENBAUM, P. (2003) Question answering in biomedicine. In M. DE RIJKE & B. WEBBER, Eds. *Actes du Workshop Natural Language Processing for Question Answering,* EACL- 2003, pages 1-4, Budapest.