

Projet EPIDEMIA

Intervention des transducteurs Nooj

M. Roux¹, M. El Zant¹, J. Royauté²

I La problématique :

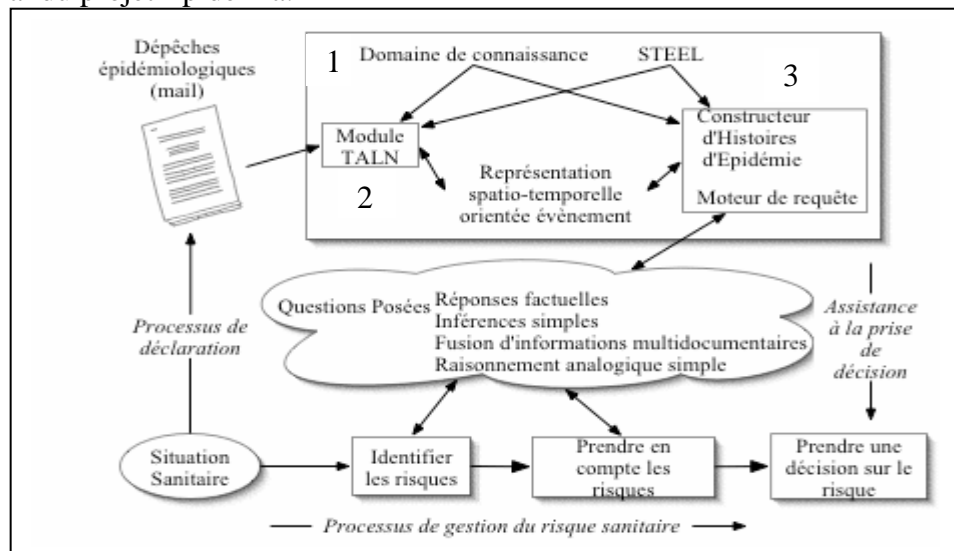
L'objectif de ce travail est la modélisation des épidémies dans le monde.

Ces épidémies sont détectées et suivies par de nombreux organismes de santé qui centralisent des messages concernant l'état des épidémies. Ces états sont constitués d'évènements épidémiologiques.

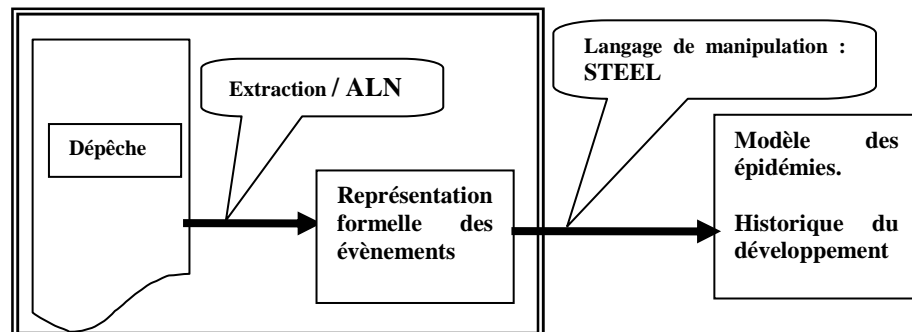
Trois tâches sont nécessaires au maintien à jour des l'état des épidémies :

1. Constitution d'une base de connaissances générales sur les épidémies (vecteurs, mode de transmission, schémas d'évolution,...) et des dictionnaires associés:
2. Dépouillement automatique des messages (dépêches épidémiologiques des organismes) permettant de maintenir à jour le modèle des épidémies en cours. Ce dépouillement relève du domaine du traitement automatique des langues.
3. Mise à la disposition des modèles ainsi constitués auprès des professionnels.

Schéma général du projet Epidemia:



Notre propos, dans ce texte, concerne la seconde tâche: le dépouillement automatique des messages.



¹ Equipe de biomathématique, informatique médicale Laboratoire d'informatique fondamentale (LIF) UMR6166.

² Equipe compréhension automatique des langues naturelles. Laboratoire d'informatique fondamentale UMR6166.

II Le contexte et enjeu médical :

Le but est de faciliter la détection et la surveillance des risques infectieux et des menaces biologiques. Notre travail a pour objectif l'analyse, la modélisation, la gestion et la restitution de connaissances et de données épidémiologiques

La seule source de dépêches PROMED [11] fournit environ 2500 dépêches par an. D'autres sources tels que l'OMS et EDISAN [1] en fournissent aussi des quantités importantes en anglais et français. Il est prévu de pouvoir traiter ces deux langues importantes de l'information médicale, en commençant par l'Anglais.

L'étape en cours de développement est l'extraction des informations à partir de ces dépêches en anglais.

III Le matériel

Nous disposons de plusieurs centaines de dépêches. Elles se répartissent en deux catégories: celles qui décrivent des évènements liés au développement des épidémies et celles qui concernent les conseils sanitaires et thérapeutiques. Nous ne nous intéressons qu'aux premières. Nous avons sélectionné 250 dépêches représentant approximativement les différents modes de rédaction. En effet l'origine et les rédacteurs de ces dépêches sont très divers. Ces dépêches ne suivent pas un plan prédéfini. Seule l'entête possède une structure stable dans la quelle on retrouve la date d'émission.

De plus, ce qui rend le problème plus ardu est le fait qu'une dépêche comporte souvent la description de plusieurs évènements épidémiologiques. Donc une dépêche peut comporter plusieurs "sous dépêches".

Enfin, ces dépêches font parfois références à des dépêches précédentes, ce qui pose le problème de la résolution de références.

IV La méthodologie :

IV-1 Les options de base.

1- L'extraction de connaissances à partir de textes peut s'envisager de plusieurs façons selon le degré de granularité souhaité pour les connaissances extraites et leur utilisation ultérieure. Nous utilisons deux approches pouvant éventuellement se compléter: une approche de traitement partiel et une approche de traitement complet. Dans l'approche de traitement partiel, il s'agit de rechercher dans les textes des motifs particuliers (maladie, virus, bactérie, etc.) au moyen d'automates et de relations entre ces motifs. Nous utilisons pour cela l'outil NOOJ [10], très bien adapté à la construction d'extracteurs permettant de construire rapidement, et de gérer des centaines de grammaires locales, dérivées des sous langages [12] mis en évidence.

2- Par ailleurs les connaissances livresques et certains travaux récents [6] suggèrent de décomposer un évènement épidémiologique en 3 champs sémantiques:

- Le type d'évènements accompagné d'éléments le précisant, éléments fonction de ce type.
- La localisation spatiale de l'évènement
- La localisation temporelle.

Ces 3 champs sont donc à isoler dans les textes.

IV-2 Les étapes de la recherche.

IV-2-1 Le corpus:

La première étape a consisté à construire le corpus comme décrit au paragraphe III.

IV-2-2 Le vocabulaire:

La seconde a consisté, à partir de ce corpus, à isoler le vocabulaire spécifique. L'outil Nooj, grâce à sa fonction de repérage des mots inconnus, ainsi que ses possibilités lexicographiques et morphologiques, nous a permis de franchir cette étape en quelques semaines. Le problème technique étant résolu par Nooj, nous nous sommes concentrés sur le "typage" des mots en grandes classes sémantiques.

Pour les champs sémantiques localisation spatiale et localisation temporelle, ce typage fût relativement aisé.

Pour le champ sémantique "type d'évènement", nous avons tenté une première classification. Pour cela nous avons isolé les verbes d'action (par opposition aux verbes d'états). Nombre d'entre eux correspondent à des évènements épidémiologiques.

IV-2-3 Les sous langages:

La troisième étape a consisté à analyser le sous langage caractérisant les concepts contenus dans ces champs sémantiques.

Les sous-langages, selon [7, 8,9] sont des discours particuliers qui portent sur des sous-domaines de la connaissance. Dans la langue générale, les contraintes d'un opérateur sur ses arguments sont relativement flous, ainsi le verbe *fuir* admet pour argument sujet un nom de type *animé*, mais aussi des noms abstraits, comme le *temps*. A l'opposé, dans les sous-langages, les opérateurs imposent des contraintes fortes sur leurs arguments. Ainsi, dans le sous-langage de l'épidémiologie, le verbe *infecter* admet des noms d'agents infectieux (virus, bactéries, vecteurs de contamination, etc.) comme argument sujet et des noms référents à des humains ou des animaux comme argument complément. [5] exploitent les propriétés d'une analyse distributionnelle couplée à un calcul de cooccurrence. Ce sont ces propriétés qui permettent de mettre en évidence de façon semi-automatique la terminologie du domaine à partir de classes de mots et des contraintes qu'exercent sur elle les différents opérateurs. Des patrons syntaxiques généraux de type Sujet-Verbe-Objet, Verbe-Sujet-Objet, Adjectif-Nom, Groupe Prépositionnel, etc. sont utilisés comme filtres distributionnels. [4] font le point sur différentes approches sur les sous-langages et détaillent deux applications reposant sur des analyseurs TAL et des grammaires (MedLEE sur des textes cliniques et GENIES sur des textes d'interaction gènes/protéines).

Cette étape est très liée à la précédente. Le typage sémantique des mots est très lié aux formes syntactico-sémantiques des membres de phrases. Les sous langages relatifs aux localisations posent peu de problèmes. Ceux liés aux différents types d'évènements en posent plus à cause de leurs diversités.

IV-2-4 Les grammaires de premier niveau:

Quatrième étape. L'analyse réalisée à l'étape précédente permet de construire les grammaires permettant de reconnaître les membres de phrases décrivant les concepts contenus dans chaque champ sémantique. Il s'agit de reconnaître les éléments "élémentaires" (atomiques) constituant les champs sémantiques et de les isoler. En effets ces éléments sont rencontrés dans le texte sans aucun ordre prédéterminé. Pour reconstituer une formulation standardisée, il est donc nécessaire d'isoler, dans un premier temps, ces éléments, avant de les regrouper et de les ordonner.

En utilisant les transducteurs de Nooj, il est possible "d'annoter" le texte d'origine à l'aide de propriétés décrivant les éléments décrits ci-dessus.

IV-2-5 Les grammaires de deuxième niveau:

Cinquième étape. Le but est de construire une formulation standardisée des éléments isolés lors de l'étape précédente. Cette étape consiste à regrouper les éléments ainsi extraits pour les mettre sous une forme standard acceptable par le langage de manipulation du modèle des épidémies (STEEL: [2; 3]).

V Les résultats.

V-1 Le corpus

Nous avons sélectionné "à la main" 250 dépêches, en incluant différents styles de rédaction, sur plusieurs milliers de textes disponibles sur le Web. Ces dépêches décrivent l'apparition de l'épidémie de SRAS de 1993 et son extension. Nous extrayons de ce corpus des sous ensembles pour tester les premiers essais de réalisations des grammaires.

V-2 Le vocabulaire et les dictionnaires.

Nous avons pris l'option de répartir notre vocabulaire en plusieurs dictionnaires, bien que Nooj puisse traiter des dictionnaires de grands volumes et de formats différents. Cette option nous facilite la mise à jour. Nous avons utilisé toutes les variantes de formats de dictionnaires offertes par Nooj (2 champs, 3 champs, mots simples, mots composés). De plus la possibilité d'associer à chaque dictionnaire un niveau de priorité, ainsi que l'option " UNAMB", possible pour chaque entrée lexicale, permet de limiter les ambiguïtés.

Exemple; Le premier mot de la ligne du dictionnaire est l'entrée lexicale:

2 champs, mots simples: (Le lemme est identique à l'entrée)

africa,N+Géo

3 champs, mots simples: (Le lemme est contenu dans le second champ)

UK,United Kingdom,N+Géo

2 champs, mots composés: (Le lemme est identique à l'entrée)

Guangdong province,N+UNAMB+Géo

3 champs, mots composés: (Le lemme est contenu dans le second champ)

hong kong special administrative region,Hong Kong SAR,N+UNAMB+Géo

Nos dictionnaires comportent un marqueur sémantique qui permet une classification en grandes catégories. Ce sont des classes d'équivalences sémantiques. Ci-dessus le marqueur sémantique est "Géo" pour "géographie". Voici la liste actuelle de ces marqueurs:

Géo + Org + Presse + Hop + Virus + Patho + Biology + Trt + Epidemie + Origin Report + Transport + Prenom + Nom + DateJour + DateMois + Spécialité + Temps;

V-3 Les sous langages.

L'analyse du sous langage associé au type d'évènement "Admission" (classe hospitalisation), que nous avons pris comme premier cas, met en évidence, outre les notions de localisations spatiale et temporelle, les notions de sujet subissant, de causes, d'objectif, d'organisme sanitaire récepteur,... En règle générale les patients sont hospitalisés dans un établissement sanitaire, pour une cause précise et pour un objectif donné.(Observation, traitement,...).

Une phrase typique:

On Saturday, two women in Leipzig, 45 years old, have been admitted to a quarantine ward, for fever, as precautionary measure.

Dans cet exemple, nous trouvons, lié au patient objet de l'admission: l'âge, le sexe, l'effectif admis.

Par ailleurs nous avons sélectionné certains verbes d'action dont l'environnement syntactico-sémantique est semblable. Ceci nous permet de créer des classes d'évènements désignés par différents verbes. Exemple : Admit et hospitalize pour l'évènement hospitalisation; travel et flight pour l'évènement déplacement.

Projet Epidemia

Equipe biomathématique, informatique médicale
Laboratoire d'informatique fondamentale.

V-4 Les différents niveaux sémantiques de traitement.

Les grammaires Nooj sont des transducteurs qui permettent "en sortie" d'ajouter des informations, produites sous forme d'annotations. Celles-ci sont liées au texte, mais sans y être incluses. Ce qui permet de "transcrire" le texte d'origine. Cette transcription peut ne concerner que des parties de phrases décrivant un élément "atomique".

L'usage des variables dans les graphes composants une grammaire, permet de formater les "sorties" selon les contraintes que doit respecter le système.

Enfin la possibilité d'enchaîner l'exécution de plusieurs grammaires (dont on donne l'ordre hiérarchique de lancement), permet de traiter un texte en un seul lancement de la fonction "Linguistic analysis".

Cette fonctionnalité nous permet de réaliser plusieurs passes "sémantiques", correspondant à différents niveaux de "compréhension".

Actuellement nous avons distingués 2 niveaux:

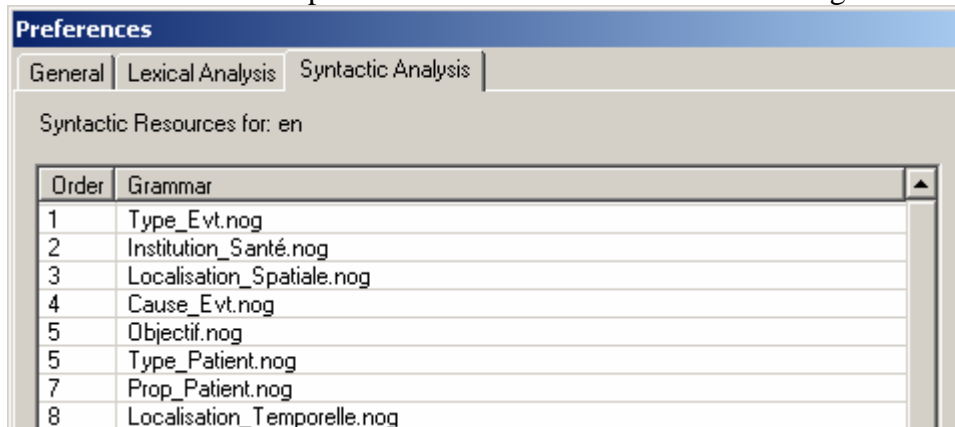
- La reconnaissance et la transcription des éléments "atomiques".
- Le regroupement structuré des éléments ainsi isolé au 1^{er} niveau, sous forme de prédicats de type logique du 1^{er} ordre.

V-4-1 . Grammaires du premier niveau.

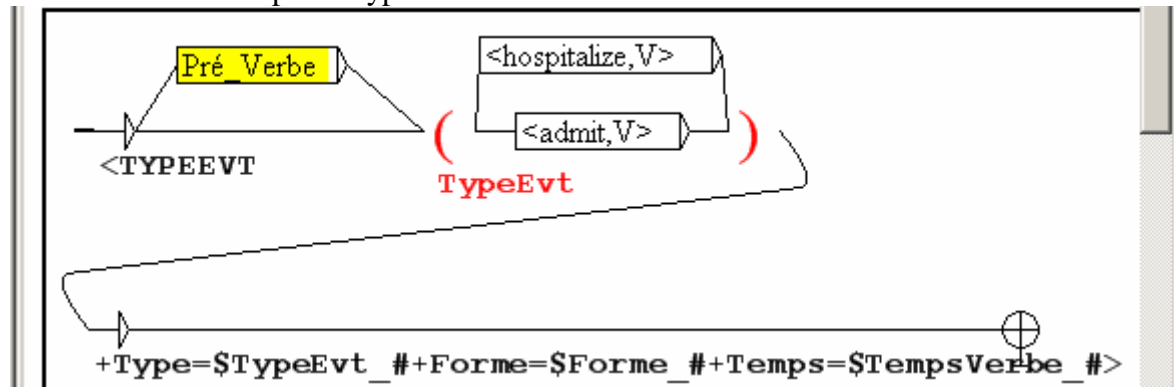
Le premier est constitué par des grammaires qui reconnaissent les segments de phrases correspondants aux concepts génériques élémentaires ("atomiques"): Localisation spatiale, temporelle, type d'évènement, concepts liés au type.

Pour le premier niveau nous utilisons 8 grammaires qui sont lancées selon l'ordre affiché.

La fenêtre "Preferences" permet de fixer l'ordre de lancement des grammaires.



Grammaire 1: Exemple de type d'évènement.

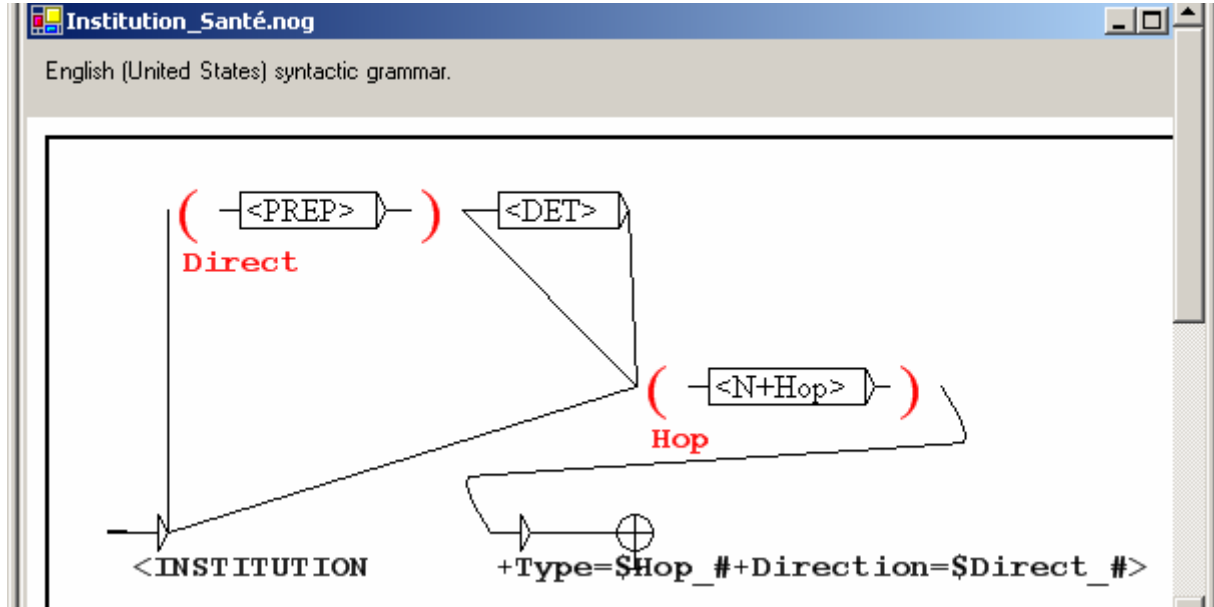


Remarquons ici la structure de la "sortie" du transducteur: TYPEEVT sera traité dans les annotations comme une catégorie syntaxique (Ex. Comme N ou V,...). Les paramètres

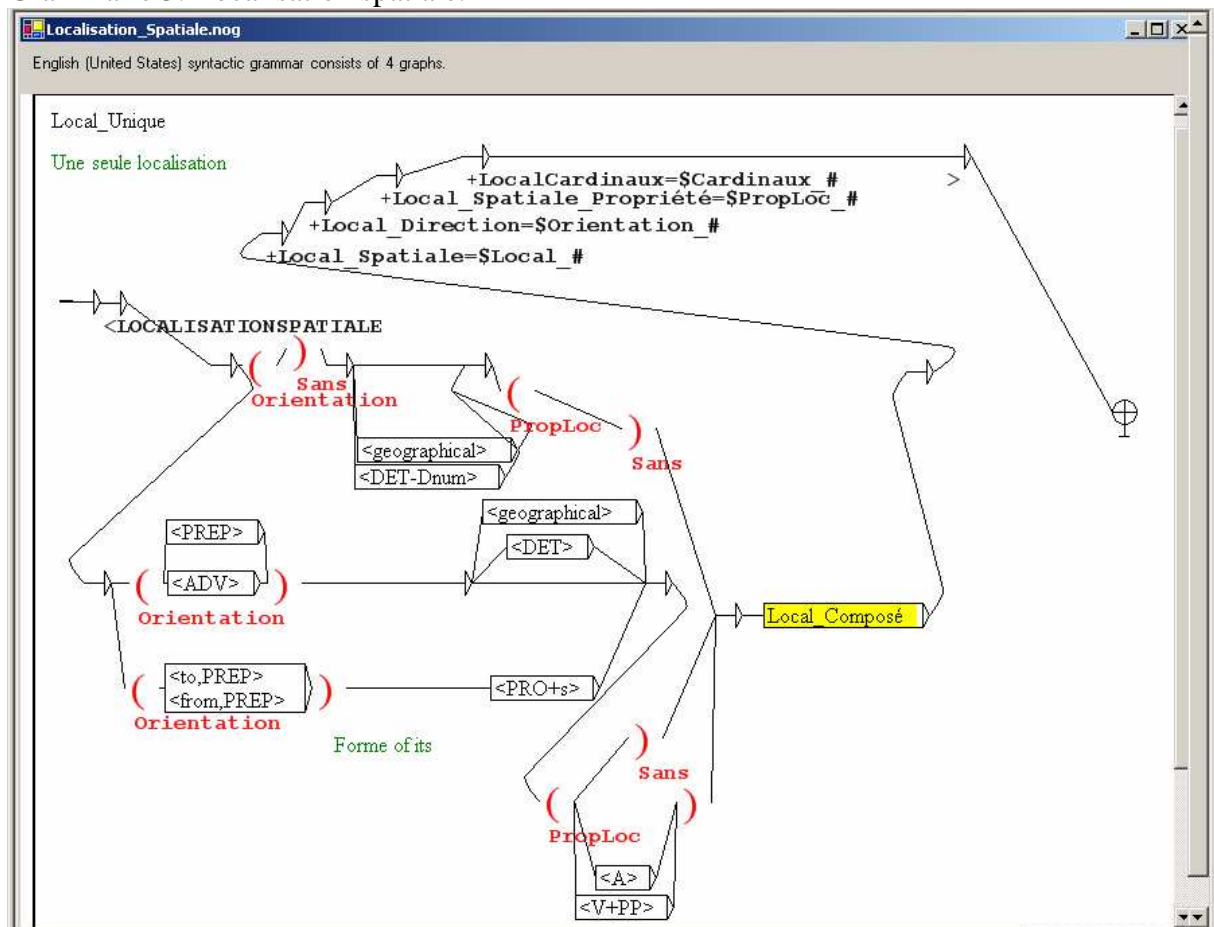
précédés d'un + sont considérés comme des propriétés liées à la catégorie TYPEEVT, et seront utilisables par les grammaires suivantes.

On remarque, de plus, que l'on affecte à ces propriétés des valeurs provenant du texte par l'intermédiaire des variables Nooj.

Grammaire 2: institutions de santé.

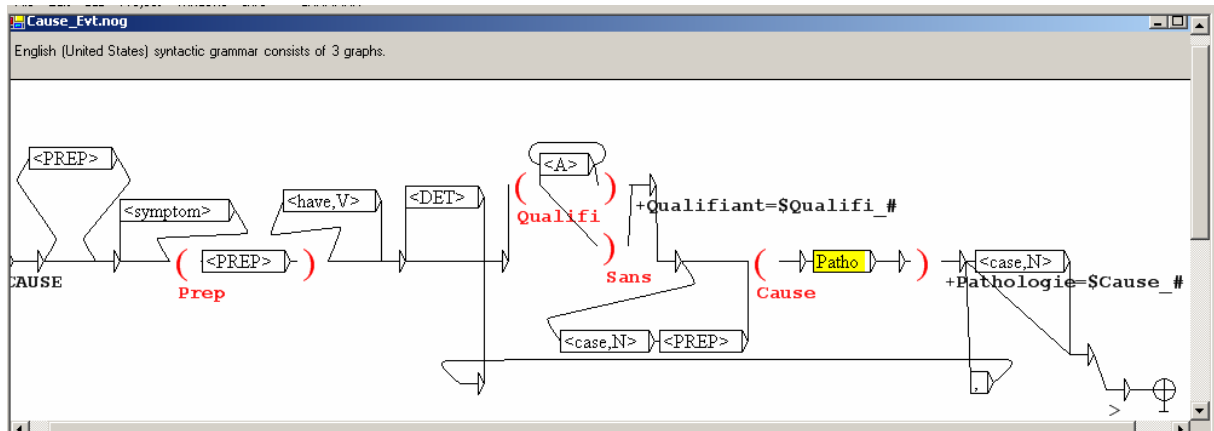


Grammaire 3: Localisation spatiale.



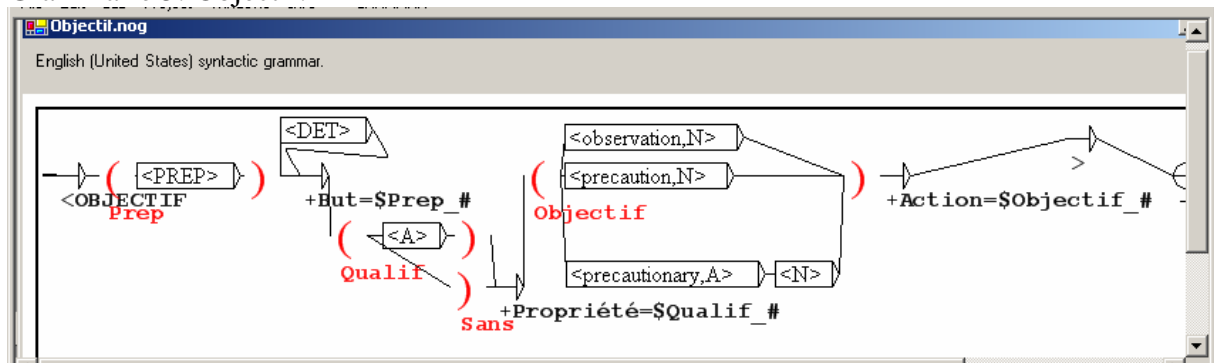
A cette occasion, remarquons la possibilité de donner des valeurs par défaut aux variables syntaxiques de Nooj.

Grammaire 4: les causes.



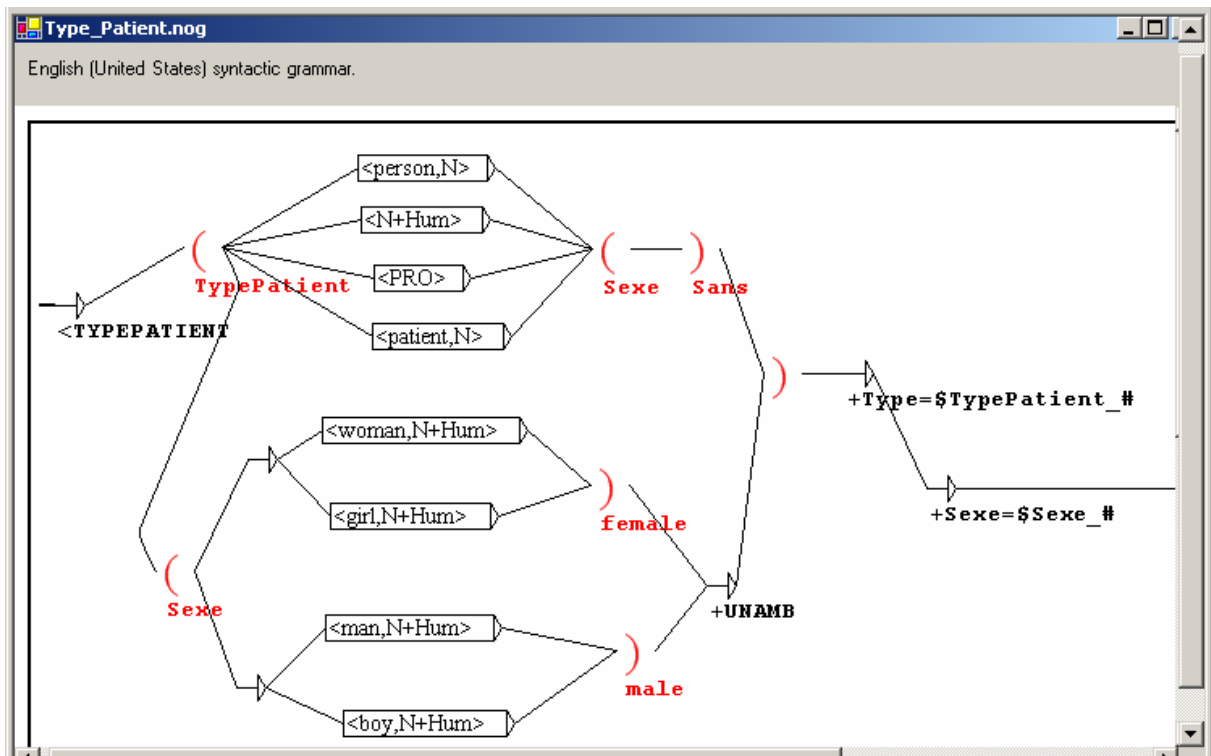
A noter: Appel au "sous graphe" Patho et la possibilité de plusieurs causes (Boucle). Attention aux valeurs de variables modifiées à chaque passage. Il faut les "mémoriser" en les écrivant.

Grammaire 5: Objectif.



Dans ces deux grammaires (ici un graphe), on remarque l'usage des valeurs par défaut qui permettent lors de la transformation finale d'avoir toujours le même nombre d'arguments pour les prédicats, même si une information est absente.

Grammaire 6: Type de patient: On détermine le type du patient et son sexe (implicite).



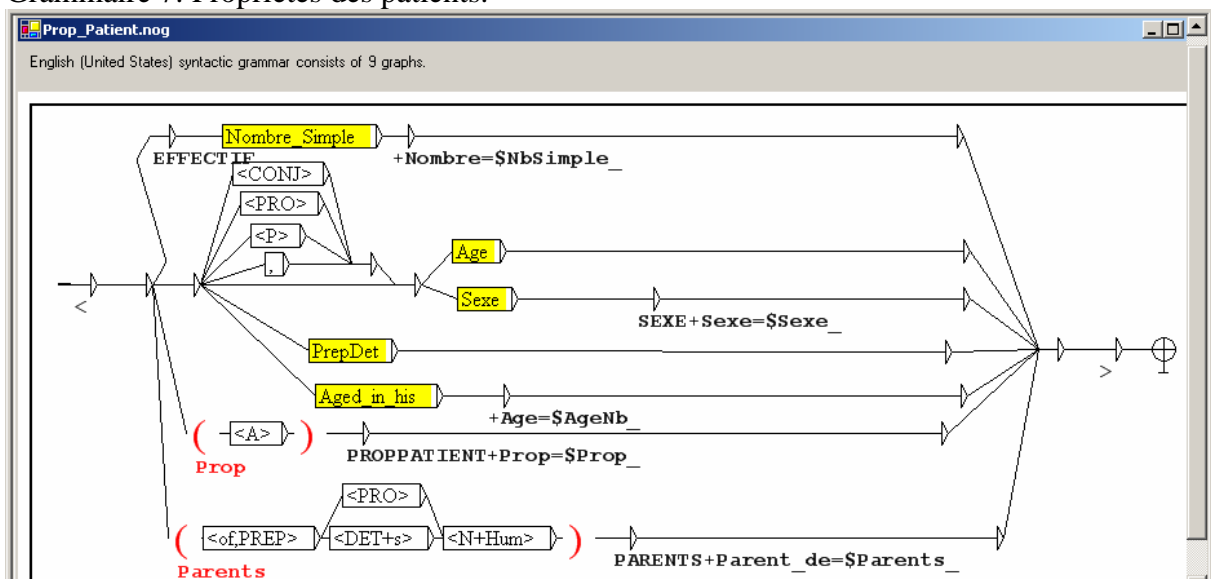
On notera ici l'écriture de la propriété +UNAMB qui a la même fonction qu'au niveau des dictionnaires. Outre l'écriture de +UNAMB dans les annotations, le passage dans cette branche du graphe exclue la deuxième solution (Branche où figure <N+Hum> qui reconnaît aussi "woman").

En d'autres termes:

Lorsque parmi les solutions d'analyse, une au moins contient la marque +UNAMB, alors :

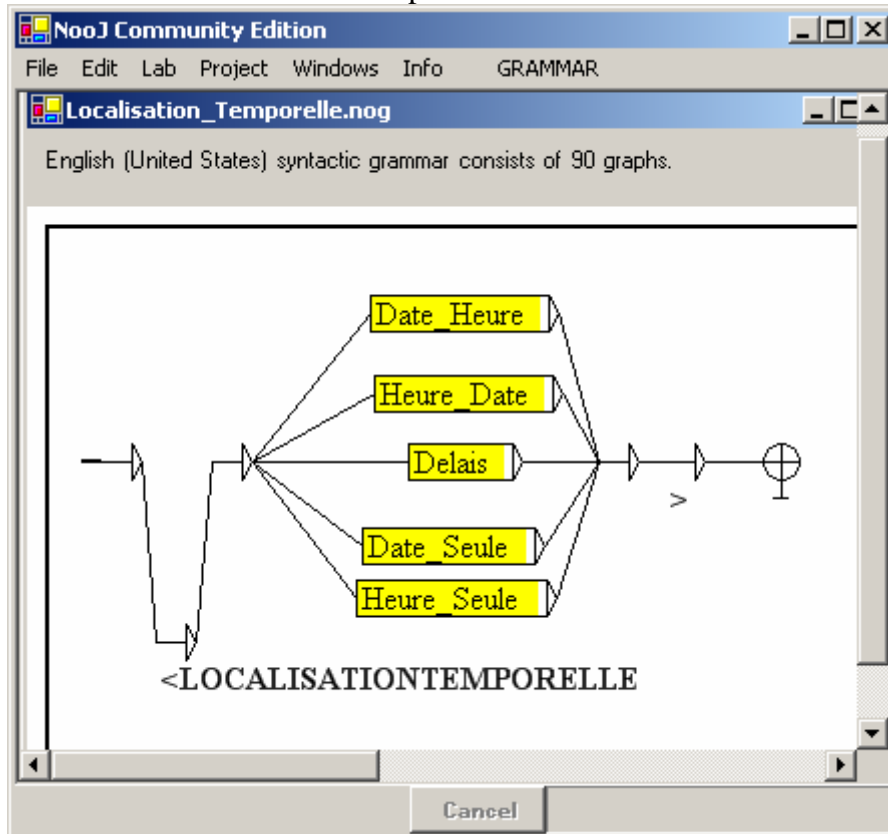
- toutes les solutions qui ne contiennent pas de +UNAMB sont effacées
- parmi toutes les solutions marquées +UNAMB, seules les plus longues sont gardées.

Grammaire 7: Propriétés des patients.

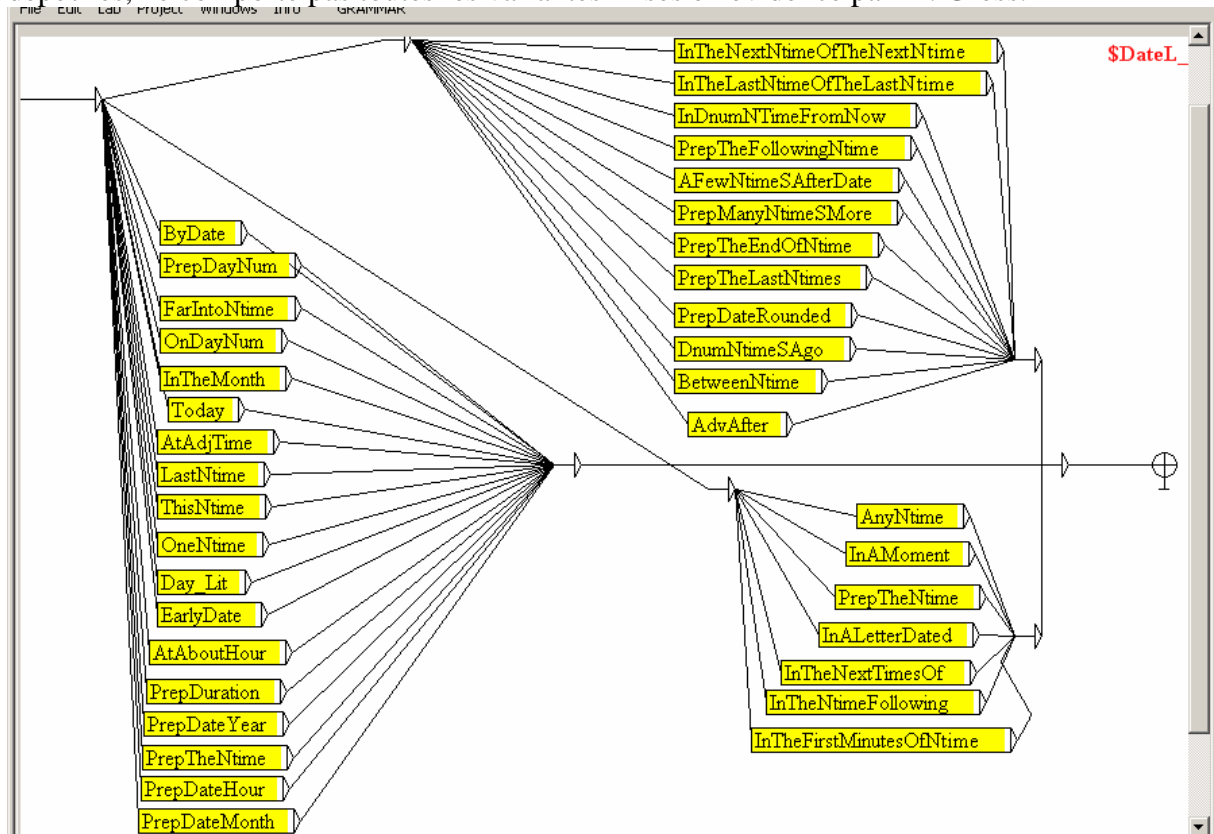


Le graphe "Sexe" est utilisé quand le sexe du patient est exprimé explicitement.

Grammaire 8: Localisation temporelle



Une grande partie de cette grammaire est héritée des travaux de Maurice GROSS. Mais il nous faut la reprendre pour la simplifier. En effet le sous-langage temporel utilisé dans nos dépêches, ne comporte pas toutes les variantes mises en évidence par M. Gross.



Voici les annotations résultant de l'application de ces 8 grammaires sur la phrase donnée.

The screenshot displays the NooJ Community Edition interface with the following annotations for the sentence: "On saturday, two women in Leipzig, 45 years old, have been admitted to a quarantine ward, for fever, as precautionary measure."

Annotations 1-17:

- 13: two, DET+Dnum+p
- 17: woman, N+Hum+p
- LOCALISATIONTEMPORELLE+Jour=saturday
- two, N+s
- EFFECTIF+Nombre=two
- TYPEPATIENT+UNAMB+Type=woman+Sexe=female

Annotations 23-38:

- 23: LOCALISATIONSPATIALE+Local_Spatiale=Leipzig+Local_Direction=in+Local_Spatiale_Propriété=Sans+LocalCardinaux=Sans
- 33-38: AGE+Age=45+Unité=year
- in, PREP
- Leipzig, N+Géo
- year, N+Ntime+p
- old, N+s
- old, A+N
- have, V+AUX+PR+1+2+s
- have, V+AUX+PR+1+2+3+p
- have, V+AUX+INF
- be, V+AUX+PP
- admit, V+PT+1+2+3+s+p
- admit, V+PP
- admitted, A
- to, PR

Annotations 44-68:

- 44: old, N+s
- 49: TYPEEVT+Type=admit+Forme=Affirmation+Temps=Passé
- 54: be, V+AUX+PP
- 59: admit, V+PT+1+2+3+s+p
- 68: INST

Annotations 68-104:

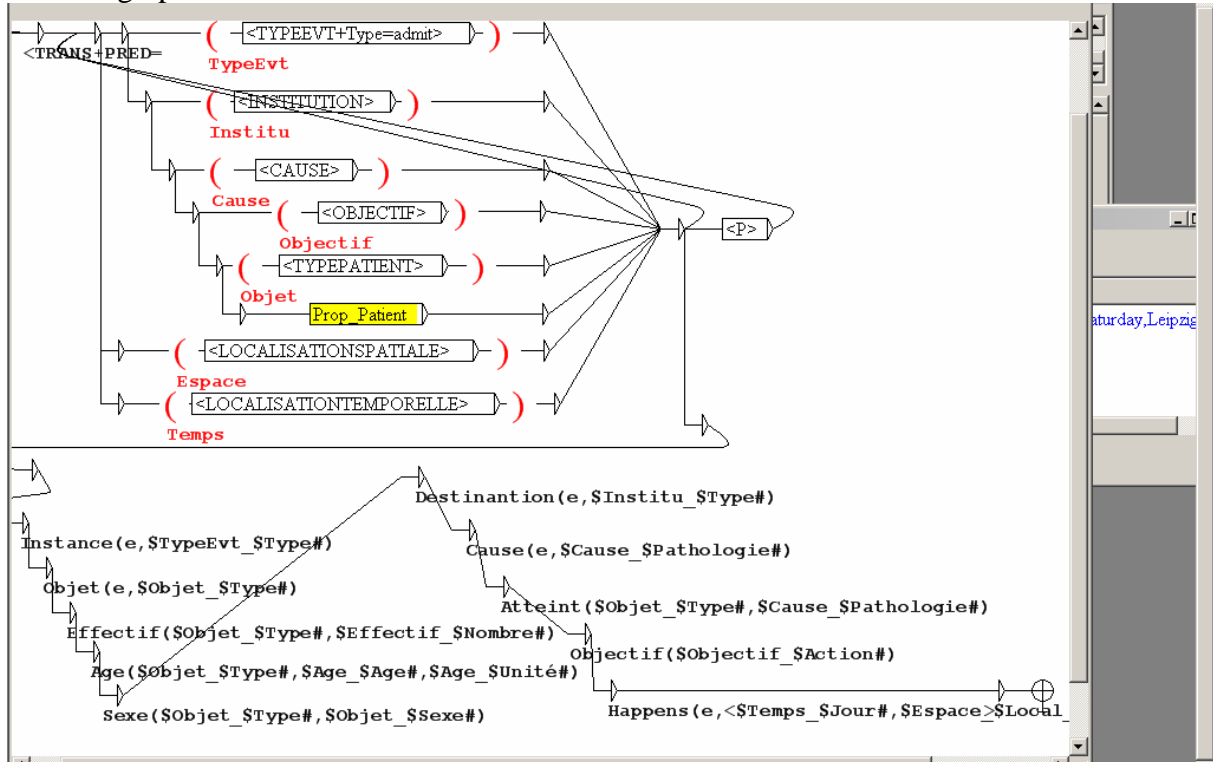
- 68: INSTITUTION+Type=quarantine_ward+Direction=to
- 71: a, DET+s
- 73: quarantine_ward, N+Hop
- 90: CAUSE+Qualifiant=Sans+Pathologie=fever
- 94: fever, N+Patho+s
- 101: OBJECTIF+But=as+Propriété
- 104: PROPPATIENT+Prop=precautionary
- to, PREP
- for, PREP
- for, CONJ
- as, PREP
- as, ADV+A
- as, CONJ
- precautionary, A
- precautionary, A
- precautionary, A

Annotations 101-118:

- 101: OBJECTIF+But=as+Propriété=Sans+Action=precautionary_measure
- 118: measure, N+s
- measure, V+INF
- measure, V+PR+1+2+s
- fever, N+Patho+s
- as, PREP
- as, ADV+A
- as, CONJ
- precautionary, A
- precautionary, A
- precautionary, A

V-4-2 Les grammaires du second niveau.

Le second niveau permet le regroupement des informations prétraitées, selon un format pré-établi. Cette étape est en cours. La solution actuelle n'est pas définitive. Un troisième niveau sera peut-être nécessaire dans le cas où le second ne pourrait pas résoudre toutes les références sémantiques. (Cas des propriétés liées à un patient, mais dispersées dans la phrase). Voici le graphe actuel:



On remarquera:

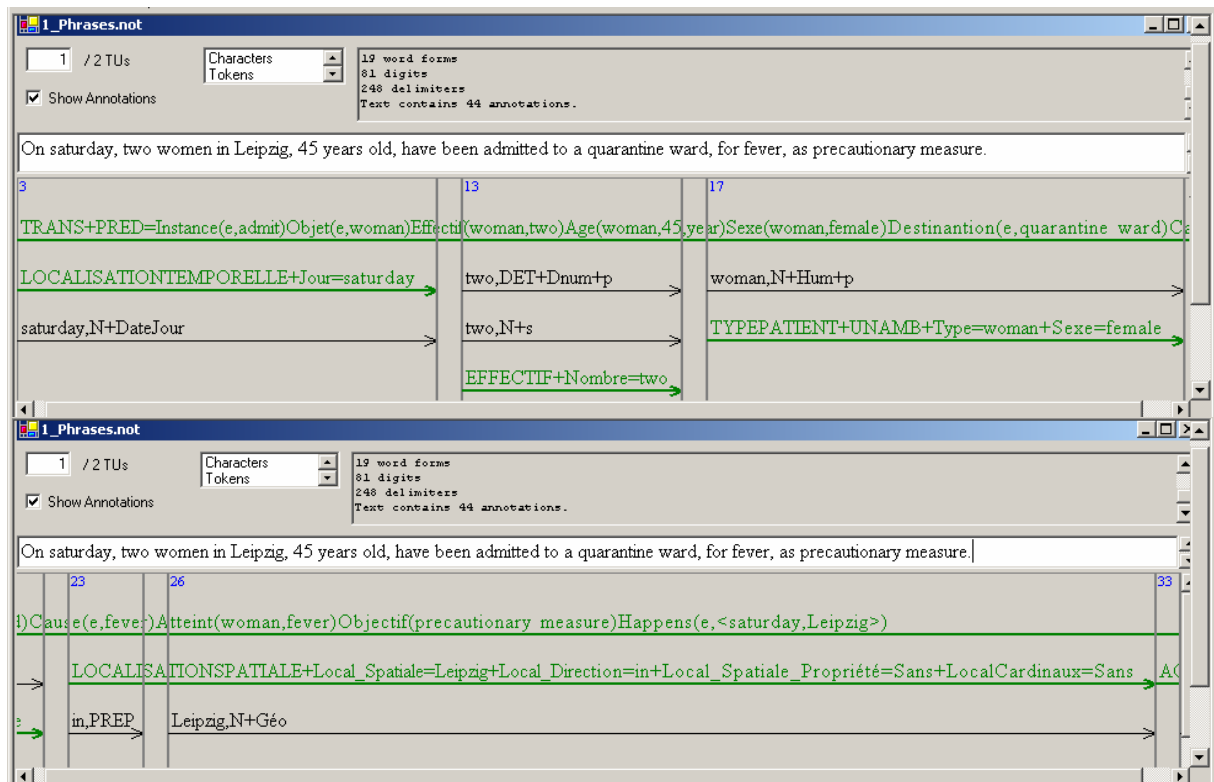
- Les catégories de la forme <CAUSE> type une séquence d'annotations du texte (Voir les fenêtres présentant les résultats du 1er niveau).
- Les catégories comportent des propriétés désignées dans les annotations par le caractère +.
- Nooj qui repère les catégories, permet de récupérer la valeur de la propriété, valeur qui a été mémorisée sous forme d'annotation à l'étape précédente, ce qui est très puissant
- Donc d'une grammaire à l'autre il est possible de passer des "arguments".
- Ces arguments sont typés par la catégorie et par la propriété associée à cette catégorie.
- Ex. ci-dessus: Récupération de "fever" dans la variable \$Cause_ \$Pathologie.

\$Cause_ \$Pathologie : Composée

- de la catégorie "Cause" et
- de la propriété "Pathologie".

On Saturday, two women in Leipzig, 45 years old, have been admitted to a quarantine ward, for fever, as precautionary measure.									
68	71	73	90	94	101	104			
INSTITUTION+Type=quarantine ward+Direction=to			CAUSE+Qualifiant=Sans+Pathologie=fever		OBJECTIF+But=as+Propriété				
		quarantine ward,N+Hop	for,PREP	fever,N+Patho+s	as,PREP	precautionary,.			
to,PREP	a,DET+g		for,CONJ		as,ADV+A	PROPPATIENT			
					as,CONJ				

En immatriculant cette grammaire en 9^{ème} position dans le fichier "Preferences", on obtient:



Cette grammaire réordonne en sortie les éléments reconnus dans le texte (ou les lemmes), sous une forme choisie par l'utilisateur. Pour nous, c'est la forme de prédicats de logique du 1er ordre.

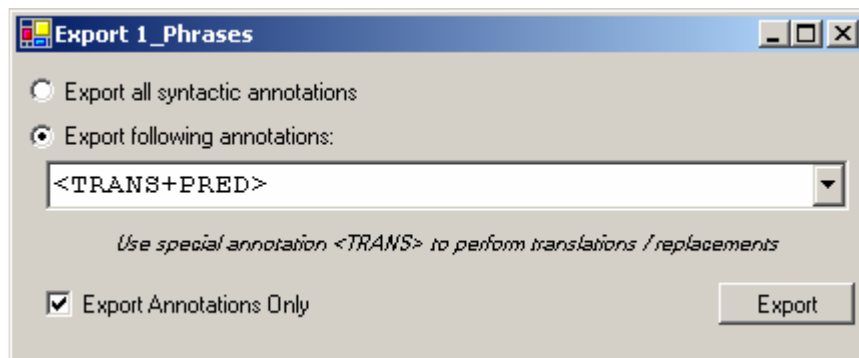
En faits, les annotations figurants dans cette fenêtre, sont issues de toutes les grammaires lancées, que ces grammaires traitent le texte d'origine ou les annotations générées par les grammaires lancées avant elle (ou les deux).

On notera la présence d'annotations "superposées" correspondant au deux niveaux de traitement.

V-4-3 Extraction des informations:

C'est le résultat de la grammaire ci-dessus: exportation des annotations (sans le texte d'origine).

L'utilisation de la catégorie TRANS qui autorise, à l'exportation des résultats du traitement du texte (donc des annotations), de ne retenir que les séquences générées, sans le texte d'origine.



Voici le résultat après exportations des annotations par les 9 grammaires:

```
<PRED>Instance(e,admit)Objet(e,woman)Effectif(woman,two)Age(woman,45,year)
Sexe(woman,female) Destinantion(e,quarantine ward)Cause(e,fever)Atteint(woman,fever)
Objectif(precautionary measure) Happens(e,<saturday,Leipzig></PRED>
```

Ce qui est le résultat du traitement de la phrase:

On saturday, two women in Leipzig, 45 years old, have been admitted to a quarantine ward, for fever, as precautionary measure.

VI Conclusion.

VI-1 Sur Epidémia:

Nous sommes à l'étape intermédiaire.

Les graphes de premier niveau qui reconnaissent les différents éléments d'un évènement doivent être affinés et doivent être modifiés pour prendre en compte tous les éléments constitutifs des différents types d'évènements.

Les graphes de deuxième niveau, qui doivent reconstituer les évènements dans leurs globalités, en prenant en compte tous les éléments reconnus au premier niveau, doivent être mis au point.

VI-2 Sur Nooj:

Notre projet n'aurait pas avancé sans certaines des fonctionnalités de Nooj:

- Pour les variables:
 - Valeurs par défaut.
 - Récupération de valeurs du texte d'origine variables standard: \$Var.
 - Récupération de valeurs des propriétés des annotations: \$Catégorie_\$Propriété.
- Le débogage des grammaires.
- Le lancement automatique des grammaires dans un ordre prédéfini.
- L'exportation des résultats avec l'option <TRANS> qui permet une substitution des éléments générés à la place du texte d'origine.

VII Bibliographie succincte.

- [1]. Bourgeade A., Chaudet H. (2005) Station EDISAN®. Système de documentation sanitaire pour centres de conseils aux voyageurs. Version 3. CD-ROM, Paris: CD Conseil, 1989-2005
- [2]. Chaudet H. (2004). STEEL: A Spatio-Temporal Extended Event Language for Tracking Epidemic Spread from Outbreak Reports. In: U Hahn, Eds, Proceedings of the First International Workshop on Formal Biomedical Knowledge Representation – KR-MED 2004, 1 June 2004, Whistler, Canada, BC. CEUR Workshop Proceedings, Vol 102.
- [3]. Chaudet H. (2005) Extending the Event Calculus for tracking epidemic spread. Artificial Intelligence in Medicine 2005 Jul 30; [Epub ahead of print].
- [4]. C. Friedman, P. Kra, A. Rzhetsky (2002) *Two biomedical sublanguages: a description based on the theories of Zellig Harris*. J Biomed Inform. 2002 Aug;35(4):222-35.
- [5]. R. Grishman, L. Hirschman, N. T. Nhan (1986) *Discovery Procedures For Sublanguage Selectional Patterns: Initial Experiments*, Computational Linguistics, Volume 12, Number 3, July-September 1986.
- [6]. Grishman R, Huttunen S, Yangarber R. (2002). Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, 35, 236–246.
- [7]. Z. Harris, (1968), *Mathematical Structures of Language*, New-York, John Wiley & Sons.

Projet Epidémia

**Equipe biomathématique, informatique médicale
Laboratoire d'informatique fondamentale.**

- [8]. Z. Harris, M. Gottfried, T. Ryckman, P. Mattick, A. Daladier, T. N. Harris & S. Harris, (1989) *The Form of Information in Science: Analysis of an immunology*.
- [9]. Z. Harris, (1991), *A Theory of Language and Information: A mathematical approach*. Oxford & New York: Clarendon Press, xii, 428 pp.
- [10]. [Nooj] <http://perso.wanadoo.fr/rosavram/pages/noojpag.html>
- [11]. [PROMED] <http://www.promedmail.org>
- [12]. [Schokkenbroek C] Schokkenbroek C. (1999). News Stories - Structure, time and evaluation. *Time and Society*, 8(1), 59-98. Woodall J. Official versus unofficial outbreak reporting through the Internet. *International Journal of Medical Informatics* 1997;47:31-4.