

Le TAL en médecine

Dr Nassim ZELLAL

MCA - Département Informatique - FEI

USTHB



Le TAL

- Le Traitement Automatique des Langues (Natural Language Processing) est un carrefour entre la linguistique, l'informatique et l'intelligence artificielle.
- Modèles mathématiques, modèles probabilistes : HMM, CRF ou SVM.

Extraction d'information

- L'Extraction d'information est une tâche du TAL
- L'EI consiste à analyser des corpus (données textuelles), afin d'en extraire des informations en vue d'une application précise.
- L'EI permet de comprendre la sémantique du texte.
- L'EI permet de produire automatiquement une représentation structurée du contenu non structuré (texte brut) d'un corpus.
- L'EI est très utilisée dans le TAL.

Informations extraites par un système d'EI

- Parmi les informations extraites par un système d'Extraction d'Information :
 - Reconnaissance d'entités nommées (NER : Named Entities Recognition). Une entité nommée est un nom propre, e.g. nom de personne, nom de lieu, nom d'organisation, etc.
 - Reconnaissance de relations sémantiques entre entités nommées, e.g. relation d'acquisition, relation de contact, relation de déplacement, etc.
 - Reconnaissance d'entités médicales : posologies de traitements/médicaments, médicaments, maladies, examens, antécédents familiaux, allergies, actes médicaux, résultats d'examens, symptômes, etc.

À quoi servent ces informations ?

- Applications nombreuses :
 - ❑ indexation sémantique pour les moteurs de recherche;
 - ❑ inférence automatique (moteur sémantique);
 - ❑ analyse sémantique et exploitation d'ontologies;
 - ❑ anonymisation de documents;
 - ❑ informatique décisionnelle (business intelligence);
 - ❑ veille;
 - ❑ systèmes de questions-réponses.

Comment extraire l'information?

- Dictionnaires.
- Grammaires (graphes) d'extraction.
- Expressions régulières.

Extraction d'information - Plateformes de TAL

- UNITEX.
- NOOJ.
- HST (THALES).

Remarque : ces plateformes sont multilingues.

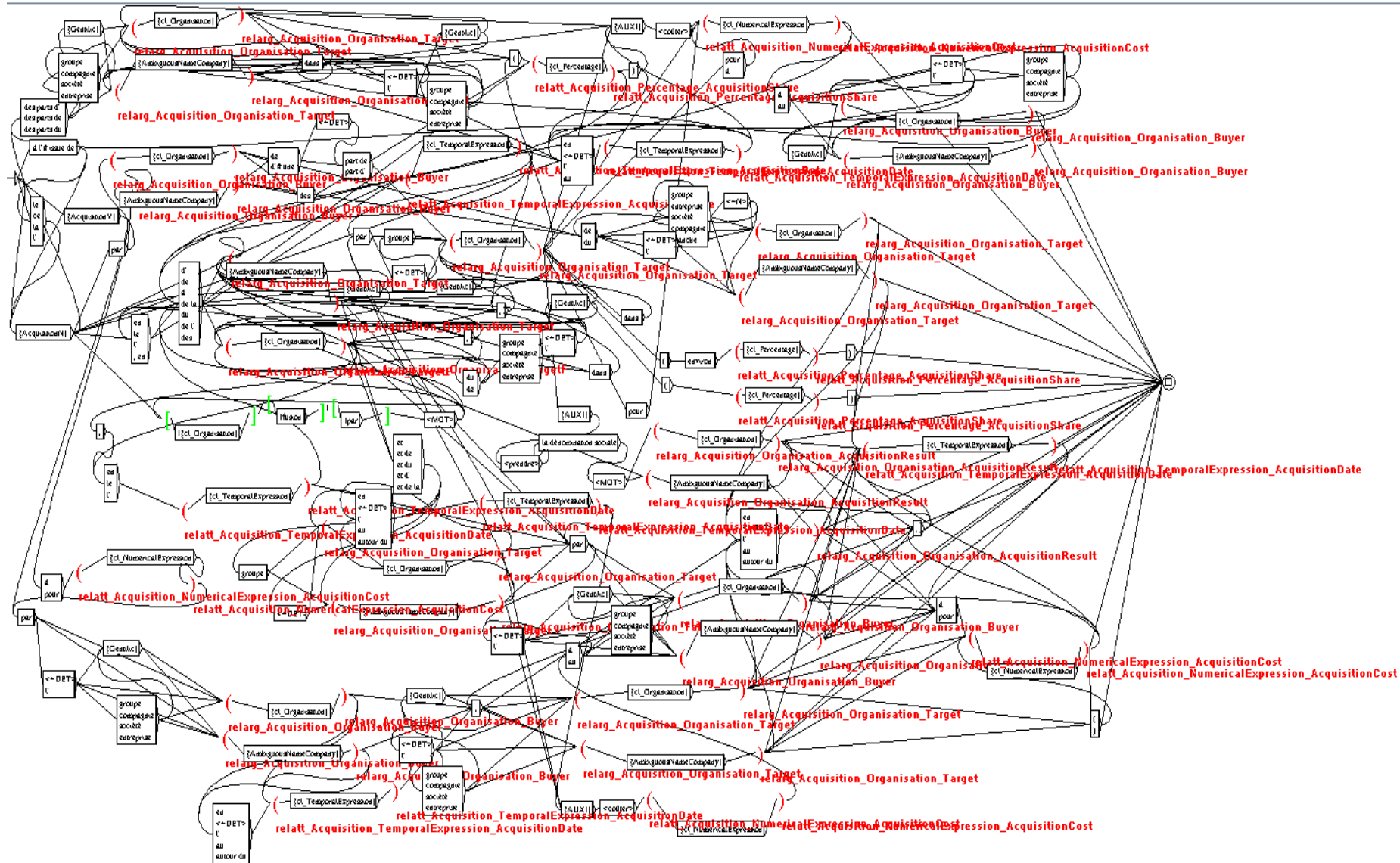
Qu'est-ce qu'UNITEX

- UNITEX est une plateforme/suite logicielle d'analyse de corpus et de TAL (traitement automatique des langues) basée sur des dictionnaires et des grammaires (graphes).
- Les grammaires et les dictionnaires sont des ressources linguistiques permettant l'extraction d'information, e.g. l'extraction d'entités nommées (noms de personnes, d'organisations, etc.), ou de relations entre entités nommées, e.g. relation d'acquisition.

Exemple d'un dictionnaire DELAF (UNITEX)

```
abacavir, .N+subst  
abatacept, .N+subst  
abciximab, .N+subst  
abiratérone, .N+subst  
acamprosate, .N+subst  
acarbose, .N+subst  
acébutolol, .N+subst  
acéclofénac, .N+subst  
acémétacine, .N+subst  
acénocoumarol, .N+subst  
acépromazine, .N+subst  
acéprométazine, .N+subst  
acétazolamide, .N+subst  
acétohexamide, .N+subst  
acétylcholine chlorure, .N+subst
```

Exemple d'un graphe d'extraction (UNITEX)



Résultat de l'extraction

<seg>de paris sur les événements organisés par ASO (Amaury Sport Organisation), à savoir le Tour de France et le Dakar ;– Winamax qui a été racheté par le chanteur Patrick Bruel et par Marc Simoncini, le fondateur et dirigeant du site de rencontres Meetic.[.] II.— UNE LIBÉRALISATION RÉGULÉE...[.] Ce projet de</seg>[.]

Sortie structurée XML

```
- <SemTag value="cl_Company" text="Winamax" offset="893" length="7">
  <SemTag value="cl_Organisation"/>
</SemTag>
- <SemTag value="relcl_Acquisition" text="Winamax qui a été racheté par le chanteur Patrick Bruel et par Marc Simoncini" offset="893" length="77">
  <SemTag value="relarg_Acquisition_Organisation_Target" text="Winamax" offset="893" length="7"/>
  <SemTag value="cl_Person" text="Patrick Bruel" offset="935" length="13"/>
  <SemTag value="relarg_Acquisition_Person_Buyer" text="Patrick Bruel" offset="935" length="13"/>
  <SemTag value="cl_Person" text="Marc Simoncini" offset="956" length="14"/>
  <SemTag value="relarg_Acquisition_Person_Buyer" text="Marc Simoncini" offset="956" length="14"/>
- <SemTag value="relcl_Leadership" text="Marc Simoncini, le fondateur et dirigeant du site de rencontres Meetic" offset="956" length="70">
  <SemTag value="relarg_Leadership_Person" text="Marc Simoncini" offset="956" length="14"/>
- <SemTag value="cl_Company" text="Meetic" offset="1020" length="6">
  <SemTag value="cl_Organisation"/>
</SemTag>
  <SemTag value="relarg_Leadership_Organisation" text="Meetic" offset="1020" length="6"/>
- <SemTag value="cl_Newspaper" text="La Tribune" offset="1114" length="10">
  <SemTag value="cl_MediaCompany">
    <SemTag value="cl_Company">
      <SemTag value="cl_Organisation"/>
    </SemTag>
```

Diagram illustrating the XML structure and its components, with annotations showing relationships and offsets:

- Red brackets group the XML elements into three main sections.
- Arrows indicate the mapping from text descriptions to specific XML attributes (value, text, offset, length).
- The words "Target", "Buyer", and "Buyer" are circled in red in the original image.

Extraction de posologies à partir d'un corpus médical

- Le corpus médical est composé de dossiers de patients anonymisés, issus de l'hôpital d'Amiens.
- Pour extraire les posologie, nous allons :
 - 1) construire un dictionnaire d'entités médicales par substance active à partir du site Web du « Vidal » en local ;
 - 2) exploiter un graphe d'extraction construit sous Unitex pour extraire les posologies ;
 - 3) enrichir le dictionnaire à partir du corpus médical.

Le dictionnaire généré au format Unix contient 1722 substances actives

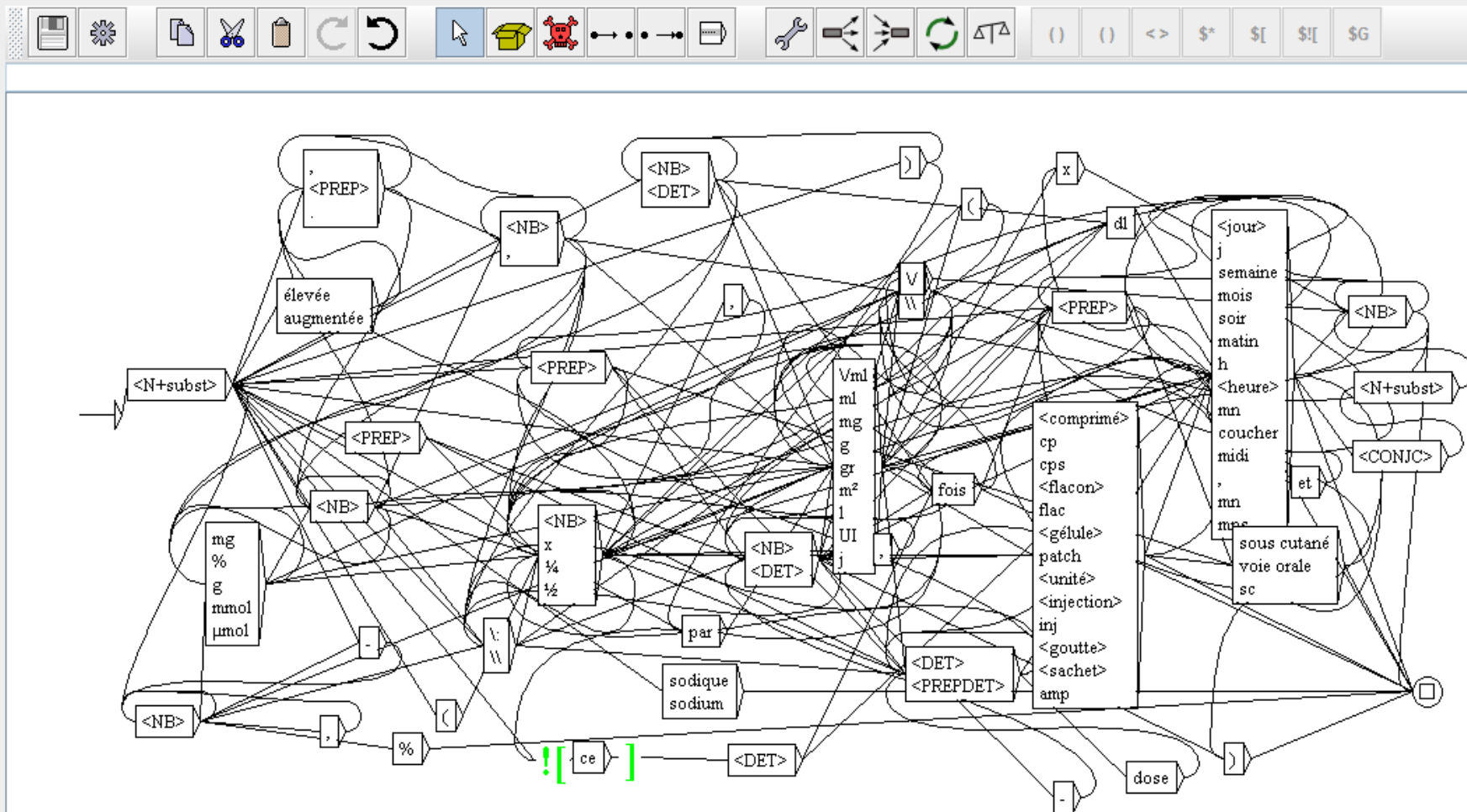
```
197 azithromycine,.N+subst
198 aztréonam,.N+subst
199 bacitracine,.N+subst
200 baclofène,.N+subst
201 bambutérol,.N+subst
202 bamifylline,.N+subst
203 baricitinib,.N+subst
204 basiliximab,.N+subst
205 baume du Pérou,.N+subst
206 bécaplermine,.N+subst
207 béclométasone,.N+subst
208 bédaquiline,.N+subst
209 bélatacept,.N+subst
210 bélimumab,.N+subst
211 bénazépril,.N+subst
212 bendamustine,.N+subst
213 bendrofluméthiazide,.N+subst
214 benfluorex,.N+subst
215 bensérazide,.N+subst
216 benzalkonium chlorure,.N+subst
217 benzathine benzylpénicilline,.N+subst
218 benzoate de benzyle,.N+subst
219 benzocaïne,.N+subst
220 benzododécinium bromure,.N+subst
221 benzydamine,.N+subst
222 benzylpénicilline,.N+subst
223 benzylthiouracile,.N+subst
224 bépridil,.N+subst
225 bêta-alanine,.N+subst
```

Graphes d'extraction de posologies sous Unitex

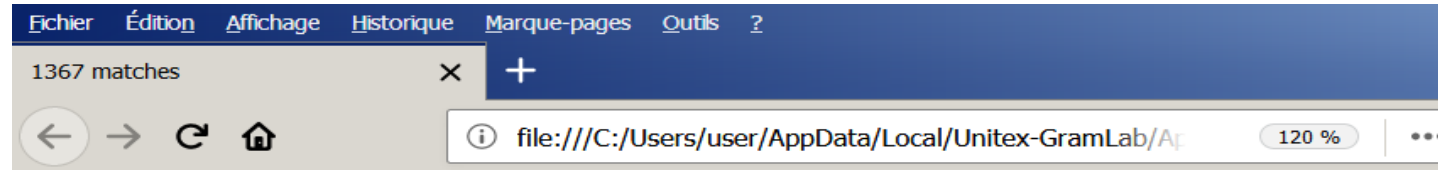
Unitex/GramLab IDE 3.1 Rev. 4314 (March 27, 2016) - current language is French

Text DELA FSGraph Lexicon-Grammar XAlign File Edition Windows Help Info

posologie.grf (C:\Users\user\Documents\Unitex-GramLab\Unitex\French\Graphs)



Résultat de l'extraction sur un navigateur Web : 1367 extractions



ce Maker : Non Traitement hospitalier : [simvastatine 20mg: 1-0-0](#), Inexium 40mg: 0-0-1 innohep 0
hospitalier : simvastatine 20mg: 1-0-0, [Inexium 40mg: 0-0-1 innohep 0,5mL: 1 inj SC à 16h](#), pipé
tatine 20mg: 1-0-0, Inexium 40mg: 0-0-1 [innohep 0,5mL: 1 inj SC à 16h](#), pipéracilline-tazobactam
ep 0,5mL: 1 inj SC à 16h, pipéracilline-[tazobactam 4g/500mg: 1flac/8h](#), Particularités cliniques
ription(s) médicale(s) de médicaments : [INEXIUM 40 mg](#), cpr gastro-résistant Statut : EC A compt
par Voie ORALE Posologie : 1 cpr à 19 h [SIMVASTATINE 20 MG](#) ARROW, CPR Statut : EC A compter du
porteur de : Perfusion Type de Stomie: [simvastatine 20mg: 1-0-0](#), Inexium 40mg: 0-0-1 innohep 0
pe de Stomie: simvastatine 20mg: 1-0-0, [Inexium 40mg: 0-0-1 innohep 0,5mL: 1 inj SC à 16h](#), pipé
tatine 20mg: 1-0-0, Inexium 40mg: 0-0-1 [innohep 0,5mL: 1 inj SC à 16h](#), pipéracilline-tazobactam
ep 0,5mL: 1 inj SC à 16h, pipéracilline-[tazobactam 4g/500mg: 1flac/8h](#), Surveillance particulièr
ce Maker : Non Traitement hospitalier : [simvastatine 20mg: 1-0-0](#), Inexium 40mg: 0-0-1 innohep 0
hospitalier : simvastatine 20mg: 1-0-0, [Inexium 40mg: 0-0-1 innohep 0,5mL: 1 inj SC à 16h](#), pipé
tatine 20mg: 1-0-0, Inexium 40mg: 0-0-1 [innohep 0,5mL: 1 inj SC à 16h](#), pipéracilline-tazobactam
ep 0,5mL: 1 inj SC à 16h, pipéracilline-[tazobactam 4g/500mg: 1flac/8h](#), Particularités cliniques
ription(s) médicale(s) de médicaments : [INEXIUM 40 mg](#), cpr gastro-résistant Statut : EC A compt
par Voie ORALE Posologie : 1 cpr à 19 h [SIMVASTATINE 20 MG](#) ARROW, CPR Statut : EC A compter du
gramme sanguin : natrémie à 133 mmol/l, [potassium à 4,4 mmol/l](#), protides à 74 g/l, calcium à 2,
assium à 4,4 mmol/l, protides à 74 g/l, [calcium à 2,01 mmol/l](#), phosphore à 1,02 mmol/l, créatin
otides à 74 g/l, calcium à 2,01 mmol/l, [phosphore à 1,02 mmol/l](#), créatininémie à 118 µmol/l, ur
,02 mmol/l, créatininémie à 118 µmol/l, [urée à 8,9 mmol/l](#), BNP à 359 ng/l. Bilan hépatique norm
/l. NFS : globules blancs à 13 500/mm3, [hémoglobine à 11,4 g/dl](#), VGM à 84,1 µ3, plaquettes à 32
EP 10 000 unités une seringue par jour, [INEXIUM 40 1 cp par jour le soir](#), PRIMPERAN 10 mg 2 ml
nausées et vomissements, PIPERACILLINE, [TAZOBACTAM 4 g 500 mg 3 flacons par jour](#) IV, LACTULOSE