

# Outils et plateformes pour le TAL

Thierry Hamon

LIMSI, CNRS, Université Paris-Saclay, Orsay, France  
Université Paris 13, Sorbonne Paris Cité, Villetaneuse, France  
hamon@limsi.fr

06/10/2017

# Plan

- Introduction
- Outils pour l'analyse de données textuelles
- Plateformes de TAL
- Ressources linguistiques pour le TAL
- Bilan

# Introduction

## Outils de Traitement Automatique des Langues :

- Analyse de données textuelles
  - Détection d'éléments textuels
  - Identification de leurs propriétés
  - Mise en relation des éléments identifiés
- Données non-structurées
  - Textes libres, rédigés en langage naturel
  - Diversité des contenus, qualité rédactionnelle variable
  - Pas de structure logique explicite, au sens informatique du terme
  - Mélange d'information de différents types (dates, valeurs numériques, lieux, noms de personne, etc.)
  - Pré-traitements nécessaires

## Type de données textuelles

Textes ou collection de textes : corpus textuels

De grandes variations :

- Format électronique (texte brut, HTML, XML, PDF, Word, etc.)
- Encodage de caractères (ASCII, ISO-LATIN-1, windows-1252, UTF-8, etc.)
- Type de documents (pages web, blogs, articles scientifiques, articles de journaux, livres, tables, messages de forum, emails, SMS, etc.)
- Taille : entre quelques Kilo-octets et plusieurs Giga-octets

# Exemples de données textuelles

## Résumé d'article scientifique

Comparative Analyses of Hairpin Substrate Recognition by  
Escherichia coli and Bacillus subtilis Ribonuclease P Ribozymes.

Ando T, Tanaka T, Kikuchi Y.

Division of Bioscience and Biotechnology, Department of Ecological  
Engineering, Toyohashi University of Technology.

Previously, we reported that the substrate shape recognition of the Escherichia coli ribonuclease (RNase) P ribozyme depends on the concentration of magnesium ion in vitro. We additionally examined the Bacillus subtilis RNase P ribozyme and found that the B. subtilis enzyme also required high magnesium ion, above 10 mM, for cleavage of a hairpin substrate. The results of kinetic studies showed that the metal ion concentration affected both the catalysis and the affinity of the ribozymes toward a hairpin RNA substrate.

*Syntaxe correcte, phrases longues et complexes, abréviations, vocabulaire et sémantique spécialisés, etc.*

# Exemples de données textuelles

## Message issu de forum de discussion

Sujet: galere pour initialiser un iphone 4

Auteur: tirando

Posté le 15-09-2016 à 11:52:13

Bonjour

On ma donne un iphone 4 desimlocker et a priori depareillé du compte icloud

Je l ai reinitialisé via itunes (dabord sans carte sim et dans un 2eme essai avec une carte sim dedans)

j ai cree un compte itunes (apple)

Quand il faut me connecter a mon nouveau compte apple au demarrage de l iphone ca merde. Le serveur n a pas pu se connecte... Ca fait une semaine que je suis dessus.

*Peu de caractères accentués, apostrophes manquantes (espace, ou mots regroupés), peu de majuscules (iphone, itunes, apple), ponctuation approximative (points de fin de phrases absents), fautes d'orthographe, d'accord ou de conjugaison, franglais, sémantique approximative, niveau de langue variable, etc.*

# Exemples de données textuelles

## Transcription de l'oral – corpus ESLO

bon je crois que ça marche maintenant oui dites donc monsieur on attend un un moment si voulez-vous bien me dire comment est-ce qu' on fait une omelette chez vous est-ce que vous pouvez me décrire la façon de faire une longue lettre une omelette une omelette une omelette oui une omelette une omelette oh bien écoutez alors là vous savez moi pour faire une omelette euh il faudra plutôt vous intéresser à ma femme parce qu' à ce moment-là au point de vue cuisine vous savez ah je vous dirai tout de suite que

*Interjections, reprises, marques d'hésitations, répétitions, incisives, etc.*

## Caractéristiques des outils de TAL

- Jeu de caractères : UTF-8, parfois encore Latin1
- Peu d'homogénéité des jeux d'étiquettes utilisés
- Format d'entrée : texte brut en général, parfois du XML simple
- Format de sortie dépendant des outils
- Parfois limitation de la taille des textes à manipuler
- Très souvent : besoin de ressources (lexicales, linguistiques, terminologiques, etc.)
- Généralement, définis pour traiter des textes bien rédigés. Mais il existe aussi des outils dédiés aux textes de spécialité, aux textes particuliers (corpus oraux, SMS, etc.)



# Outils pour l'analyse de données textuelles

- Présentation d'outils disponibles librement pour les tâches suivantes :
  - Etiquetage morpho-syntaxique et lemmatisation
  - Reconnaissance d'entités nommées
  - Extraction de termes
  - Analyse syntaxique
  - Extraction de relations sémantiques

pour l'anglais et le français principalement,  
de manière simplifiée,  
sans être exhaustif

# Texte initial

Two weeks ago , the patient was started on metoprolol 12.5 mg p.o.

q.6 h. for rate control .

The patient has a history of atrial fibrillation with a slow ventricular response .

## Segmentation en mots et en phrases

Two weeks ago , the patient was started on metoprolol 12.5 mg p.o.

q.6 h. for rate control .

The patient has a history of atrial fibrillation with a slow ventricular response .

- Identification des mots et des phrases
- Tâche généralement intégrée aux outils d'étiquetage morpho-syntaxique, de lemmatisation ou de reconnaissance d'entités nommées

# Etiquetage morpho-syntaxique et lemmatisation

[https://aclweb.org/aclwiki/POS\\_Tagging\\_\(State\\_of\\_the\\_art\)](https://aclweb.org/aclwiki/POS_Tagging_(State_of_the_art))

Two week ago , the patient be start on metoprolol 12.5 mg p.o.  
*CD NNS RB DT NN VBD VBN IN FW CD NN SYM*

q.6 h. for rate control .  
*FW NP IN NN NN*

The patient have a history of atrial fibrillation with a slow ventricular response .  
*DT NN VBZ DT NN IN JJ NN IN IN DT JJ JJ NN*

- Deux tâches pouvant être couplées
- De nombreux jeux d'étiquettes morpho-syntaxiques

# Caractéristiques des étiqueteurs morpho-syntaxiques et lemmatiseurs

- Utilisation de méthodes de
  - Arbre de décision (TreeTagger)
  - Chaînes de Markov, HMM, MaxEnt, CRF (GeniaTagger, MEIt, SEM)
- Modèle dépendant de la langue mais aussi du type de texte
- Jeux d'étiquettes :
  - Universel : ADJ, NOUN, VERB, CCONJ, ...
  - Penn TreeBank POS tagset : JJ, NN, VBZ, CC, ...
  - Multext : A---s--, Ncfs--, Vmip3s--, Cc
  - ...
- Bonne précision : de l'ordre de 97% pour l'anglais et le français

# TreeTagger

<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

- Arbres de décision
- Langues : anglais, français, français parlé, allemand, italien, grec, etc.
- Jeux d'étiquettes : Penn TreeBank TagSet pour l'anglais, French TreeBank pour le français
- Apprentissage préalable sur un corpus de référence (le WSJ pour l'anglais) à partir de bigrammes  
Modèles de langues mis à jour
- Effectue également la lemmatisation
- Entrée : texte brut

## Sortie :

The	DT	the
patient	NN	patient
has	VBZ	have
a	DT	a
history	NN	history
of	IN	of
atrial	JJ	atrial
fibrillation	NN	fibrillation
with	IN	with
a	DT	a
slow	JJ	slow
ventricular	JJ	ventricular
response	NN	response
.	SENT	.

Lors	ADV	lors
de	PRP	de
l'	DET:ART	le
approche	NOM	approche
,	PUN	,
en	PRP	en
vent	NOM	vent
arrière	VER:pper	arrière
pour	PRP	pour
la	DET:ART	le
piste	NOM	piste
26L	INT	26L
,	PUN	,
à	PRO:POS	à
	DET:ART	T Hamon

# Stanford POS tagger

<https://nlp.stanford.edu/software/tagger.shtml>

- Utilisation d'un modèle log-linéaire
- Jeux d'étiquettes : Penn TreeBank TagSet pour l'anglais, French TreeBank pour le français
- Langues : anglais, français, chinois, arabe
- Pas de lemmatisation
- Entrée : texte brut ou XML
- Sortie :

```
Two_CD weeks_NNS ago_RB ,_, the_DT patient_NN was_VBD
started_VBN on_IN metoprolol_NN 12.5_CD mg_NN p.o._NN
q._NN 6_CD h._NN for_IN rate_NN control_NN ._.
```

```
The_DT patient_NN has_VBZ a_DT history_NN of_IN atrial_JJ
fibrillation_NN with_IN a_DT slow_JJ ventricular_JJ
response_NN ._.
```

## Etiqueteurs morpho-syntaxiques

- TreeTagger

<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

- GeniaTagger <http://www.nactem.ac.uk/GENIA/tagger/>

- TnT tagger <http://www.coli.uni-saarland.de/~thorsten/tnt/>

- StanfordPOSTagger

<http://nlp.stanford.edu/software/tagger.shtml>

- Flemm <https://sites.google.com/site/fiammettanamer/home/outils---ressources/flemm>

- LIA\_TAGG [http://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/chercheurs/bechet/download\\_fred.html](http://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/chercheurs/bechet/download_fred.html)

- MElt

<https://www.rocq.inria.fr/alpage-wiki/tiki-index.php?page=MElt>

- SEM <http://www.lattice.cnrs.fr/sites/itellier/SEM.html>

- ...





## Caractéristiques des outils de reconnaissance d'entités nommées

- CRF, règles+ressources
- Identification du segment et de la catégorie sémantique de l'entité nommée
- Identification de catégories
  - prédéfinies : organisation, lieu, personne (ENAMEX), dates, expressions temporelles (TIMEX), valeurs monétaires, pourcentage (NUMEX)
  - spécifiques à un domaine : espèces, protéines, gènes (biologie), médicaments, prescriptions (médecine), etc.

# Stanford NER

<https://nlp.stanford.edu/software/CRF-NER.shtml>

- Utilisation d'un classifieur CRF
- Langues : anglais, allemand, espagnol, chinois
- Catégories d'entités reconnues : *Person, Organization, Location*  
+ *Money, Percent, Date, Time*
- Entrée : texte brut
- Sortie :

```
Jacques/PERSON Dubochet/PERSON of/0 Switzerland/LOCATION ,/0  
Joachim/PERSON Frank/PERSON of/0 the/0 US/LOCATION and/0  
Richard/PERSON Henderson/PERSON of/0 the/0 UK/LOCATION were/0  
awarded/0 the/0 prize/0 for/0 cryo-electron/0 microscopy/0 ./0
```

# Heideltime

<https://dbs.ifi.uni-heidelberg.de/resources/temporal-tagging/>

- Système à base de règles et de ressources
- Langues : anglais, français, espagnol, etc.  
version 2.0 : 200 langues grâce à des ressources définies automatiquement (qualité faible)
- Normalisation des dates en fonction d'une date d'origine
- Utilisation de TreeTagger pour l'étiquetage morpho-syntaxique
- Entrée : texte brut
- Sortie : texte avec balises XML (TIMEML)

```
<TIMEX3 tid="t1" type="DATE" value="2017-W38">Two weeks ago</TIMEX3> ,  
the patient was started on metoprolol 12.5 mg p.o. q.6 h. for rate control .
```

```
Le déjeuner est offert à toutes les personnes qui se seront inscrites  
à PDIA <TIMEX3 tid="t3" type="DATE" value="2017">2017</TIMEX3> avant  
<TIMEX3 tid="t1" type="DATE" value="2017-10-03">le 03/10/17</TIMEX3> .
```

## Outils de reconnaissance d'entités nommées

- Stanford NER (pour l'anglais, l'allemand, l'espagnol et le chinois) <http://nlp.stanford.edu/ner/index.shtml>
- HeidelTime (expressions temporelles)  
<http://dbs.ifi.uni-heidelberg.de/index.php?id=129>
- GeniaTagger (entités nommées en génomique)  
<http://www.nactem.ac.uk/GENIA/tagger/>
- LIA NE (pour le français)  
<http://pageperso.lif.univ-mrs.fr/~frederic.bechet/download.html>
- ...



# Approches pour l'extraction de termes

- Méthodes

- Expressions régulières
- Approche contrastive
- Analyse syntaxique superficielle

puis filtrage statistique

- Production d'une liste de termes candidats
- Association de mesures statistiques aux termes candidats (fréquence, spécificité, C-Value, etc.)

# TermSuite

<http://termsuite.github.io/>

- Règles d'extraction basées sur le moteur d'expressions régulières de UIMA
- Regroupement des variantes morphologiques et syntaxiques
- Extraction de terminologies monolingues et génération de dictionnaires bilingues
- Langues : français, anglais, russe, italien, allemand, espagnol
- Utilisation de TreeTagger et de la plateforme UIMA
- Entrée : texte brut
- Sortie :

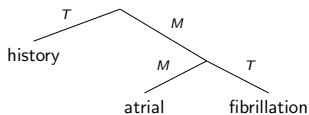
1	T	n: ribozyme	4	1	T	N N	wind turbine	5,16
2	T	nn: escherichia coli	2	1	V	N N N	horizontal-axis wind turbines	3,52
3	T	n: rnase	2	1	V	A N N N	horizontal axis wind turbine	3,50
4	T	n: catalysis	1	1	V	A N N N	vertical axis wind turbines	3,62
5	T	aan: slow ventricular response	1	1	V	A N N	smaller-scale wind turbines	2,20
6	T	npan: history of atrial fibrillation	1	1	V	A N N	on-shore wind turbines	1,90
7	T	nn: rate control	1	3	T	N N	wind energy	4,51
8	T	n: vitro	1	3	V	A N N	californian wind energy	1,90
9	T	n: metoprolol	1	3	V	A N N	offshore wind energy	3,56
10	T	npan: result of kinetic study	1	3	V	N N N	wind energy conversion	3,32



# YaTeA

<http://search.cpan.org/~thhamon/Lingua-YaTeA/>

- Analyse syntaxique superficielle (Tête / Modifieur) à l'aide de patrons minimaux d'analyse syntaxique
- Rejet des groupes nominaux non analysables
- Langues : français et anglais
- Mesures statistiques associées aux termes candidats (Fréquences, C-Value1, C-Value\*, etc.)
- Entrée : texte étiqueté par TreeTagger ou Flemma
- Sortie :



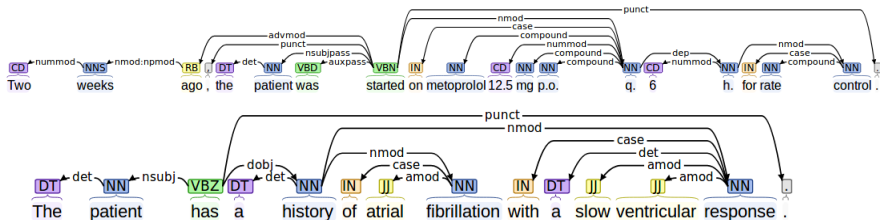
history of atrial fibrillation	1	2.32	0.2
slow ventricular response	1	2	0.25
rate control	1	1.58	0.33
patient	2	2	1
ventricular response	1	0	0.66
fibrillation	1	0	1
atrial fibrillation	1	0	0.33

## Extracteurs de termes

- TermSuite <http://termsuite.github.io/>
- YATEA <http://search.cpan.org/~thamon/Lingua-YaTeA/>
- TermoStat <http://termostat.ling.umontreal.ca/>
- TerMine <http://www.nactem.ac.uk/software/termine/>
- BioTex <http://tubo.lirmm.fr/biotex/>
- TBXTools <https://sourceforge.net/projects/tbxtools/>
- ...

# Analyse syntaxique

[https://aclweb.org/aclwiki/Parsing\\_\(State\\_of\\_the\\_art\)](https://aclweb.org/aclwiki/Parsing_(State_of_the_art))

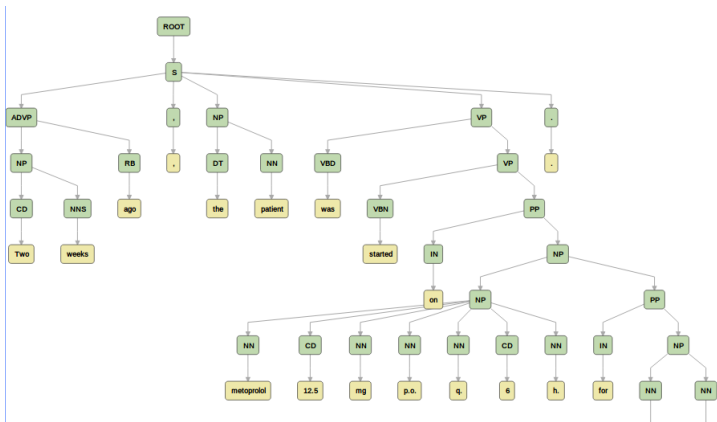


- Construction de l'arbre syntaxique des phrases
- Deux types d'analyse : en dépendances ou en constituants

# Analyse syntaxique

[https://aclweb.org/aclwiki/Parsing\\_\(State\\_of\\_the\\_art\)](https://aclweb.org/aclwiki/Parsing_(State_of_the_art))

Exemple d'analyse en constituants :



# Caractéristiques des analyseurs syntaxiques

[https://aclweb.org/aclwiki/Parsing\\_\(State\\_of\\_the\\_art\)](https://aclweb.org/aclwiki/Parsing_(State_of_the_art))

- Méthodes :
  - Règles d'analyse syntaxique définies manuellement
  - Apprentissage statistique
- Modèles syntaxiques généralement issus des grammaires génératives
- Temps d'analyse variable suivant le type de méthode et/ou la complexité des phrases
- Assez dépendant du type de texte

# Link Grammar Parser

<https://www.abisource.com/projects/link-grammar/>

- Système à base de règles
- Production de toutes les analyses possibles
- Langues : anglais, russe, arabe, persan, vietnamien, etc.
- Entrée : texte brut
- Sortie : arbres syntaxiques (+constituants)



# Talismane

<http://redac.univ-tlse2.fr/applications/talismane.html>

- Analyse syntaxique statistique  
Classification statistique et intégration de connaissances linguistiques à travers la sélection de traits et des règles spécifiques à chaque langue
- Langues : français, anglais, occitan
- Entrée : texte brut
- Sortie :

1	Le	le	DET	DET	n=s g=m	2	det	2	det
2	déjeuner	déjeuner	NC	NC	n=s g=m	4	sub	4	sub
3	est	être	V	V	n=s t=P p=3	4	aux_pass	4	aux_pass
4	offert	offrir	VPP	VPP	n=s g=m t=K	0	root	0	root
5	à	à	P	P		4	mod	4	mod
6	toutes	tout	ADJ	ADJ	n=p g=f	8	mod	8	mod
7	les	les	DET	DET	n=p	8	det	8	det
8	personnes	personne	NC	NC	n=p g=f	5	prep	5	prep
9	qui	qui	PROREL	PROREL	n=s	12	sub	12	sub
10	se	se	CLR	CLR	n=p,s p=3	12	aff	12	aff
11	seront	être	V	V	n=p t=F p=3	12	aux_pass	12	aux_pass
12	inscrites	inscrire	VPP	VPP	n=p g=f t=K	8	mod_rel	8	mod_rel
13	à	à	P	P		12	mod	12	mod
14	PDIA	-	NPP	NPP		13	prep	13	prep
15	2017	2017	NC	NC		14	mod	14	mod
16	avant	avant	P	P		12	mod	12	mod
17	le	le	DET	DET	n=s g=m	18	det	18	det
18	03	03	NC	NC		16	prep	16	prep
19	/	/	PONCT	PONCT		18	nonct	18	nonct

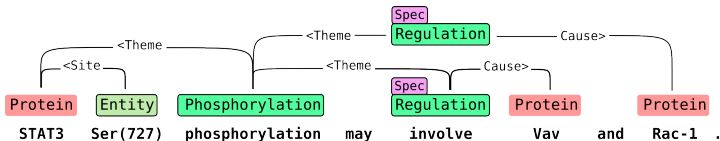


# Syntactic parsing

- **LinkParser (Carnegie Mellon University)**  
`https://www.abisource.com/projects/link-grammar/`
- **Memory-based Shallow parser (ILK)**  
`ilk.uvt.nl/cgi-bin/tstchunk/demo.pl`
- **Stanford parser (Stanford University)**  
`https://nlp.stanford.edu/software/lex-parser.shtml`
- **LoPar** `http://www.ims.uni-stuttgart.de/tcl/SOFTWARE/LoPar.html`
- **Charniak&Johnson parser** `https://github.com/BLLIP/bllip-parser`  
`https://nlp.stanford.edu/~mcclosky/selftraining.html`
- **Talismane** `https://github.com/joliciel-informatique/talismane`
- **Berkeley Parser** `https://github.com/slavpetrov/berkeleyparser`
- **MST Parser** `https://sourceforge.net/projects/mstparser/`  
`http://alpage.inria.fr/statgram/frdep/fr\_stat\_dep\_mst.html`
- **Malt parser** `http://www.maltparser.org/`  
`http://alpage.inria.fr/statgram/frdep/fr\_stat\_dep\_malt.html`
- **Bonsai** `http://ressources.labex-efl.org/bonsai`  
`http://alpage.inria.fr/statgram/frdep/fr\_stat\_dep\_bky.html`
- **frmg** `http://alpage.inria.fr/frmgwiki/book/export/html/4307`

## Extraction de relations sémantiques

- Assez peu d'outils disponibles librement
- Outils dédiés généralement à un domaine particulier
- Exemple : Turku Event Extraction System (TEES)  
<http://jbjorne.github.io/TEES/>
  - Extraction d'événement à partir de textes biomédicaux
  - Basé sur une analyse syntaxique, un classifieur SVM et un post-traitement sémantique à base de règles
- Exemple :



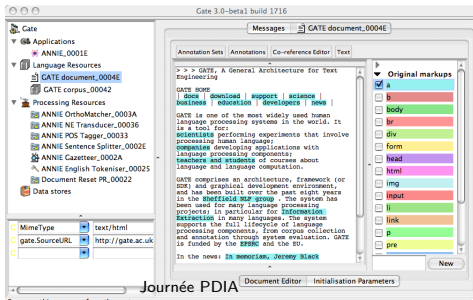
# Plateformes de TAL

- La plupart des outils intègrent plusieurs traitements, mais
  - ils sont dédiés à une tâche en particulier
  - ils n'offrent pas vraiment la possibilité d'agencer des outils de TAL ou d'intégrer des ressources suivant les besoins applicatifs
- Nombreuses plateformes disponibles
  - Interopérabilité entre des outils de TAL existants ou réimplémentation de méthodes
  - Réflexion sur les formats de stockage des annotations (souvent déportées)
  - Proposition des API

# GATE

<http://gate.ac.uk/>

- Architecture orientée objet (Java) avec un environnement de développement graphique
- Utilisateurs : principalement ingénieurs linguistes, etc.
- Intégration d'outils existants (Protégé, Jena, Yale, ...)
- Nombreux composants : raciniseurs, étiqueteurs morpho-syntaxiques, analyseurs syntaxiques, outils d'apprentissage, résolution de la coréférence nominale + outils de visualisation des annotations, des ontologies, etc.



# UIMA

<http://uima.apache.org/>

- Plateforme initialement développée par IBM, désormais projet Apache
- Points communs et interopérabilité avec GATE mais le nombre de composants est plus réduit
- Plutôt à destination des ingénieurs en informatique et développeurs API Java, C++, accessible dans Eclipse
- Architecture permettant un déploiement en parallèle
- Composants : raciniseur (Snowball), segmenteur en mot, projection de dictionnaires, étiqueteur morpho-syntaxique (HMM), etc. OpenNLP (segmenteur en mots et en phrases, étiqueteur morpho-syntaxique, REN, analyseur syntaxique) cTAKES (traitement de données cliniques)
- Plutôt difficile d'utilisation

# NLTK

<http://www.nltk.org/>

- Plutôt pour la recherche et l'éducation  
Succès grandissant
- Plutôt une bibliothèque de composants  
Pas d'interface graphique ni d'outils en ligne de commande  
API Python (paquetage python)
- Composants : calcul de collocations, étiqueteur morpho-syntaxique, analyseur syntaxique, etc.  
+modules permettant l'interopérabilité avec les outils de Stanford

# Plateformes de TAL

- GATE <http://gate.ac.uk/>
- UIMA <http://uima.apache.org/>
- NLTK <http://www.nltk.org/>
- Stanford coreNLP <https://stanfordnlp.github.io/CoreNLP/>
- LingPipe <http://alias-i.com/lingpipe/>
- OpenNLP <http://incubator.apache.org/opennlp/>
- Treex [ufal.mff.cuni.cz/treex](http://ufal.mff.cuni.cz/treex)
- Lima <https://github.com/aymara/lima/wiki>
- Ogmios <http://search.cpan.org/~thhamon/Lingua-Ogmios/>
- Plateforme Alvis (INRA)  
<http://bibliome.jouy.inra.fr/demo/ontobio/alvisir2/webapi/search>
- FreeLing <http://nlp.lsi.upc.edu/freeling/>

# Ressources

- Fournissent les informations linguistiques nécessaires aux outils de TAL
  - Liste de mots « vides »
  - Ressources lexicales
  - Ressources terminologiques
  - Réseaux sémantiques
  - Lexiques divers (émotions, etc.)

Mais aussi des corpus annotés



# WordNet

<https://wordnet.princeton.edu/>

- Réseau lexical de l'anglais (env. 150 000 mots – noms, verbes, adjectifs, adverbes)  
Basé initialement sur une approche psycholinguistique
- Lexique et sens organisés autour de synsets (ensemble de synonymes)
- Organisation hiérarchique  
Relation sémantique principale : hyperonymie
- Relations sémantiques entre synsets
- Ressources similaires dans plusieurs langues
- Exemple :

```
Sense 1
patient
=> case
    => person, individual, someone, somebody, mortal, soul
    ...
=> sick person, diseased person, sufferer
    => unfortunate, unfortunate person
        => person, individual, someone, somebody, mortal, soul
```

```
Sense 1
patient
=> case
=> sick person, diseased person, sufferer
```


# BabelNet

<http://babelnet.org/>

- Réseau sémantique multilingue
- Fusion des Wik\*, plusieurs WordNet, etc.
- 271 langues, 6 millions de concepts, 7,7 millions d'entités nommées
- Expression d'un même concept dans plusieurs langues
- Traduction automatique des synsets
- Mise en correspondance des sens de WordNet avec les articles Wikipedia
- Désambiguïsation à l'aide des informations issues de WordNet et de Wikipedia
- API pour l'interrogation de la ressource
- Utilisé dans le système de désambiguïsation Babelfy (<http://babelfy.org/>)

# BabelNet

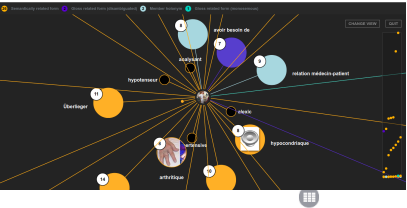
## Exemple



**BabelNet**

- Dictionnaire
- Images
- Traductions
- Sources
- Catégories
- Mots composés
- Autres forms
- Liens externes

bn:00061017n-



**Traductions**

- en** patient, patient en consultation externe, Patients client, milieu hospitalier, patients hospitalisés, procédure ambulatoire
- ar** مريض, المرضي, المرض, المرضى, عيادة خارجية, مرضى, مريض, مريض, Patient, أجزاء العيادات الخارجية, أي والمريض, المرضى الخارجيين, مريضاً
- zh** 患者, 病患, 病人, 疾病患者, 病者, 门诊患者, 门诊病人, 住院, 患病, 门诊, 病人, 患者, 门诊, 门诊手术
- fr** patient, outpatient, Outpatient Department, Day patient, Discharge planning, Health consumer, In-patient, Inpatient admissions, Inpatient procedure, Inpatients, Out-patient, Outpatient procedure, Outpatients, Patients
- de** patient, ambulanter patient, ambulante, patienten, stationär, stationäre, verfahren
- el** ασθενής, εξωτερικός ασθενής, περιπτωση, ασθενής, διαδικασία εξωτερικών ασθενών, σε ασθενή
- hi** मरीज, बीमार, मरीज, inpatients, अतिरिक्त विभाग, बाहर मरीज, में अति, अतिरिक्त
- it** paziente, ammalato, assistito, cliente, degente, malato, paziente ambulatoriale, Pazienti, ambulatoriali, pazienti ricoverati, procedure ambulatoriale
- ja** 患者, クランケ, 外来, 病人, 外来処置, 外来患者
- ru** пациент, амбулаторный пациент, больной, больной, пациент, пациента, амбулаторная процедура, пациентом
- es** paciente, Pacientes, paciente ambulatorio, pacientes hospitalizados, procedimiento ambulatorio
- pt** paciente, paciente, ambulatória, ambulatória, paciente

# Ressources pour le TAL

- Lexiques :

- DELA <http://infolingu.univ-mlv.fr/DonneesLinguistiques/Dictionnaires/telechargement.html>
- Morphalou <http://www.cnrtl.fr/lexiques/morphalou/>
- Lefff <http://alpage.inria.fr/~sagot/lefff.html>
- GLAFF <http://redac.univ-tlse2.fr/lexiques/glaff.html>
- GLAWI <http://redac.univ-tlse2.fr/lexiques/glawi.html>

- Autres ressources linguistiques :

- FrameNet <https://framenet.icsi.berkeley.edu/fndrupal/>
- WoNef <https://wonef.fr/>
- UMLS <https://uts.nlm.nih.gov/home.html>

- Corpus :

- Penn Treebank <http://www.cis.upenn.edu/~treebank/>
- French TreeBank <http://ftb.linguist.univ-paris-diderot.fr/>

# Bilan

- Nombreux outils pour les différentes tâches du TAL
  - Assez peu interoperables directement (formats et jeux d'étiquettes différents)
  - Mais très souvent : intégration de plusieurs tâches de TAL en amont de la finalité de l'outil
- Plusieurs plateformes de TAL pour des utilisateurs et des applications variées
  - Accès aux tâches d'analyse de données textuelles dans des applications à « moindre coût »

## Qu'est-ce qui n'a pas été couvert ?

- Résolution de coréférence
- Systèmes de dialogue
- Recherche de réponses à des questions
- Génération de textes
- Résumé automatique
- Extraction de mots-clés
- Traduction automatique

Outils commerciaux : Synapse, Synomia, Proxem, Syllabs, etc.

# Sites annuaires

- [https://aclweb.org/aclwiki/Main\\_Page](https://aclweb.org/aclwiki/Main_Page)
- <http://www.atala.org/-Outils-pour-le-TAL->
- <http://www.elra.info/en/catalogues/lre-map/>
- <https://nlp.stanford.edu/links/statnlp.html>
- [tangra.cs.yale.edu/newaan/index.php](http://tangra.cs.yale.edu/newaan/index.php)
- <http://ortolangx.hypotheses.org/>
- <http://registry.dfki.de/>
- <http://multital.inalco.fr/>