ENS DE LYON

icar
UMR 5191 - CNRS

Construction d'un corpus annoté

Le cas du corpus Presto : 16e-21e siècles

Achille Falaise, ENS de Lyon, ICAR

École doctorale
Linguistique de corpus: regards en synchronie et en diachronie

http://presto.aiakide.net/diablerets

Pourquoi utiliser des corpus ?

- Linguistique empirique : **sciences** du langage
  - Il faut des **données** quantifiables et contrôlées pour vérifier les hypothèses

Corpus écrit

Corpus oral

Corpus multimodal

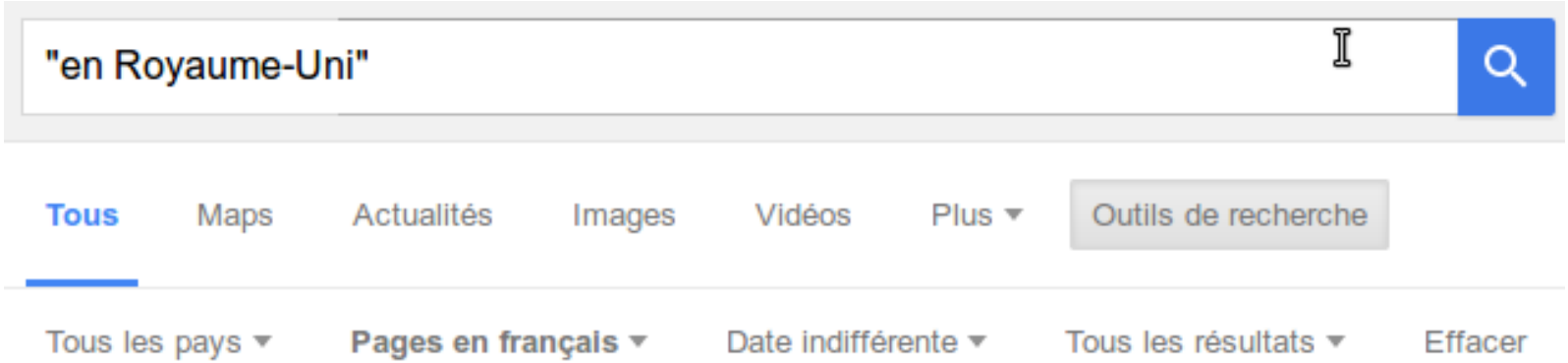
http://presto.aiakide.net/diablerets

Qu'est-ce qu'un corpus ?

- Un corpus scientifique doit être quantifiable et contrôlé
  - Taille
  - Langue
  - Date
  - Genre textuel, conditions d'énonciation
  - Auteur
- Une notion relative
- Le Web est-il un corpus ?

http://presto.aiakide.net/diablerets

Le Web est-il un corpus ?



Peut-on créer un corpus à partir du Web ?

- Sélection de sites/pages Web
  - Conversion d'un site en texte
    - ex. Bootcat : http://bootcat.sslmit.unibo.it/
- Bases libres
  - Wikipédia, Wikisource, Wikitravel, etc.
  - Conversion en texte
    - Wikipedia Extractor
      - https://github.com/bwbaugh/wikipedia-extractor
    - Wiki Extractor
      - https://github.com/attardi/wikiextractor

Choix des textes

- Définir un cadre
  - Date(s)
  - Genre(s) textuel(s)
  - D’autres variables…
  - … et essayer de bien le couvrir !
- Un corpus témoin ?
- Nombre de mots
  - Plus il y en a, mieux c’est !
    - > 1M mots
  - Équilibrer
- Licence

Le corpus Presto

- Corpus noyau (16e-20e siècles)
  - Équilibré sur le genre textuel et la date
  - Libre
  - 53 textes
- Corpus second (16e-21e siècles)
  - Équilibré sur le genre textuel et la date
  - Non-libre
  - 339 textes
- Corpus complémentaires
  - Journalistique (19e-21e)
  - Encyclopédique (18e-21e)

Le corpus Presto

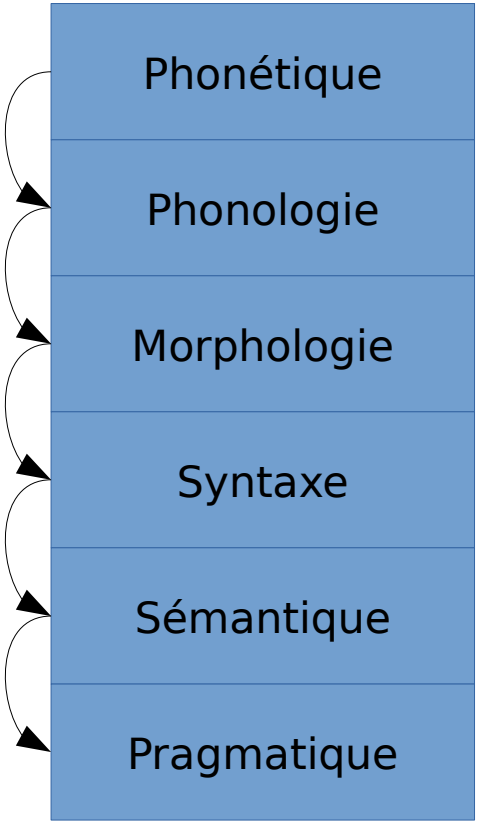
Le corpus du projet franco-allemand PRESTO a été constitué grâce aux textes issus des bases textuelles suivantes :

- FRANTEXT (http://www.frantext.fr , V. Montémont, G. Souvay)
- BVH (*Bibliothèques Virtuelles Humanistes*, http://www.bvh.univ-tours.fr - L. Bertrand, M.-L. Demonet)
- ARTFL (*American and French Research on the Treasury of the French Language*, http://artfl-project.uchicago.edu – R. Morrissey, M. Olsen)
- CEPM (*Corpus électronique de la première modernité*, http://www.cpem.paris-sorbonne.fr)

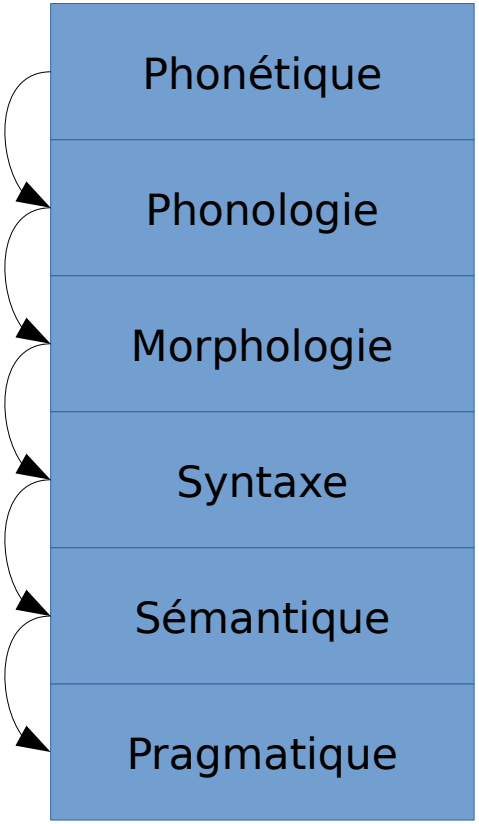
Les ressources et les outils élaborés dans PRESTO ont bénéficié des apports des logiciels LGerM (G. Souvay) et Analog (M.-H. Lay) ainsi que du lexique Morphalou.

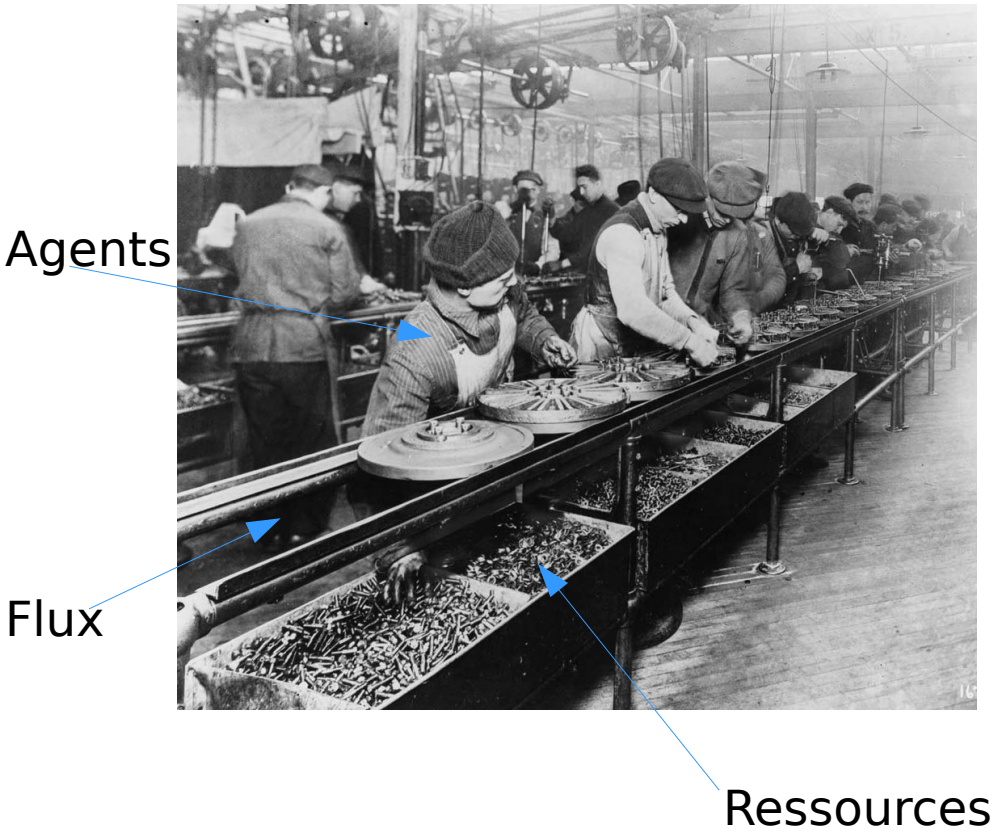
Prétraitement linguistique

Niveaux d'analyse linguistique

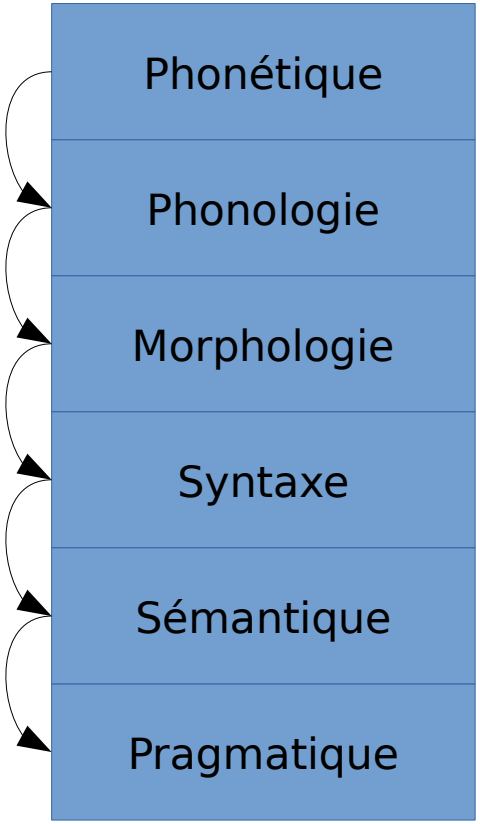


Niveaux d'analyse linguistique

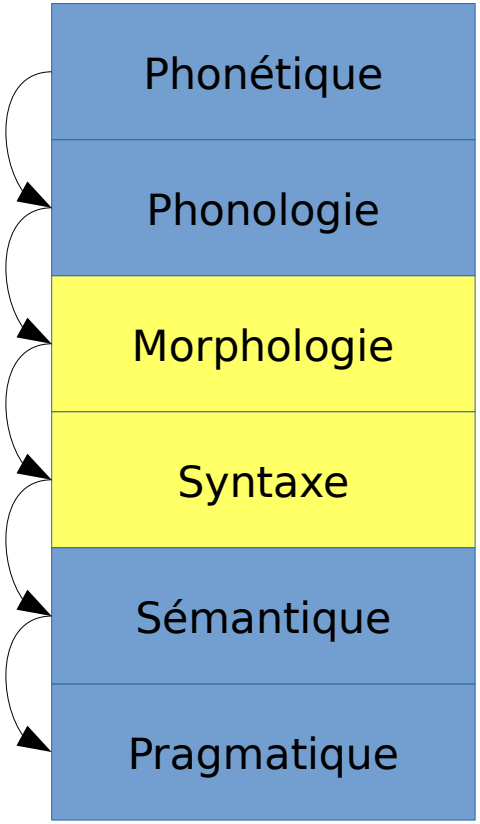




Niveaux d'analyse linguistique



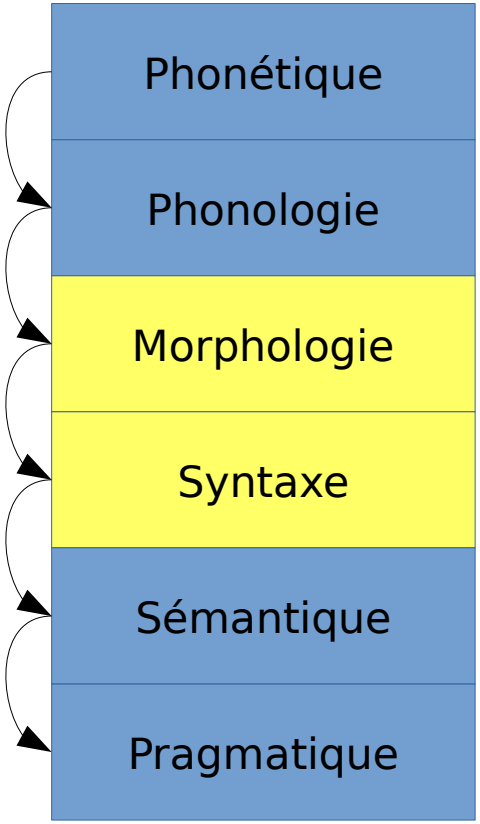
Niveaux d'analyse linguistique



Formes : *Disait* ; *lieux* ; *pour*
Partie du discours (POS) : *V* ; *N* ; *S*
Lemme : *DIRE* ; *LIEU* ; *POUR*
D'autres étiquettes...
→ *tagger*

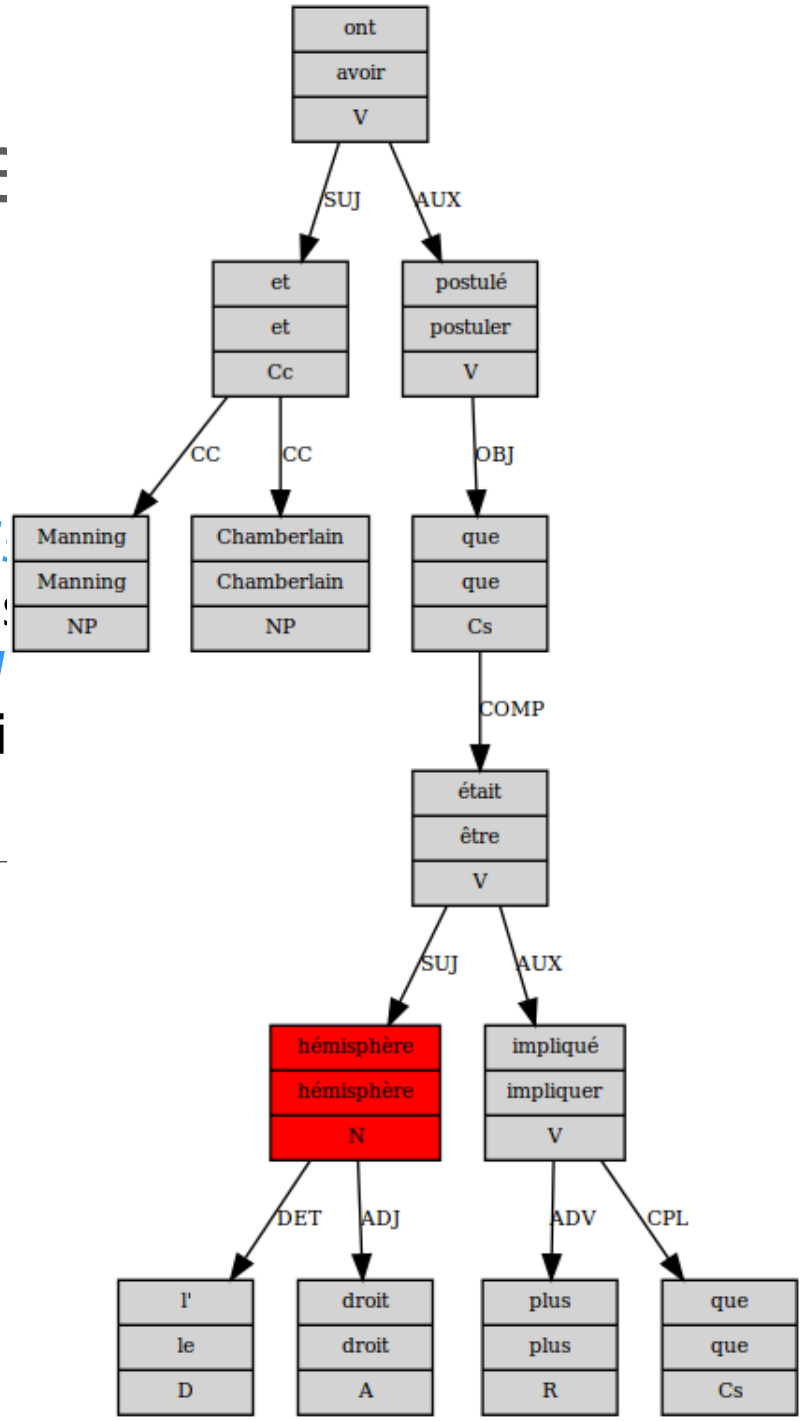
Arbre
→ *parser*

Niveaux d'analyse



Formes : *Di*
Partie du di
Lemme : *DI*
D'autres éti
→ *tagger*

Arbre
→ *parser*



Niveaux d'analyse linguistique

Ce n'est pas toujours ce découpage qui est utilisé

- Phonétique
- Phonologie
- Sémantique

- Morphologie
- Syntaxe
- Sémantique
- Pragmatique

Choix linguistiques

- Tokenisation
  - Aujourd'hui, parce que, par ce que, trèsgrand…
- Parties du discours
- Lemmes
  - le/la, il/elle, je/tu/il/nous/vous/ils, du
- Autres étiquettes
  - Traits morphologiques : mode, temps, nombre, etc.
  - Traits sémantiques : axiologiques, topologiques, etc.

→ projets « de référence » : MULTEXT/EAGLES, GRACE

→ manuel de segmentation, manuel d'annotation, jeu d'étiquettes

Catégorie	Définition	Sous-catégories (définitions)	
N	Nom	c (commun), p (propre)	
V	Verbe	u (être/avoir), v (autre verbe)	c (conjugué), n (infinitif)
A	Adjectif	g (général), p (possessif)	
P	Pronom	p (personnel), d (démonstratif), i (indéfini), s (possessif), t (interrogatif), r (relatif)	
D	Déterminant	a (article défini), d (démonstratif), n (article indéfini), p (article partitif), i (indéfini), r (relatif), t (interrogatif/exclamatif)	
G	Participe- Adjectif-Gérondif	a (part. présent/adjectif verbal/gérondif), e (part. passé/adjectif verbal)	
R	Adverbe	g (général), p (particule), t (interro-exclamatif)	
S	Adposition	-	
C	Conjonction	c (coordination), s (subordination)	
M	Numéral	c (cardinal), o (ordinal)	
I	Interjection		
X	Résidu	a (abréviation), e (mot étranger), s (symbole), p (préfixe), i (consonne intercalée)	
F	Ponctuation	s (forte), w (faible), o (autre)	

Choix techniques

Tokenisation (*tokeniser*)

- Méthodes heuristiques
- Méthodes lexicales
  - Chinois, vietnamien...
  - Algorithme *la plus longue chaîne d'abord*
- Méthodes stochastiques

Parties du discours + lemmes (*tagger*)

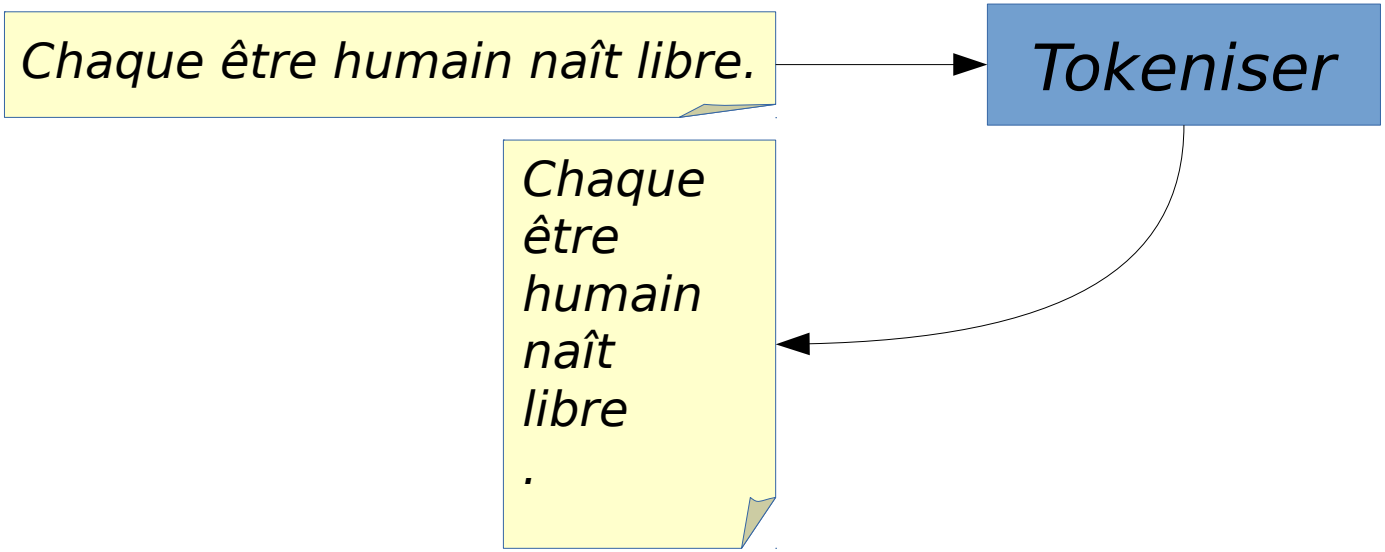
- TreeTagger
  - http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
- MElt Tagger
- ...

Dépendances (*parser*)

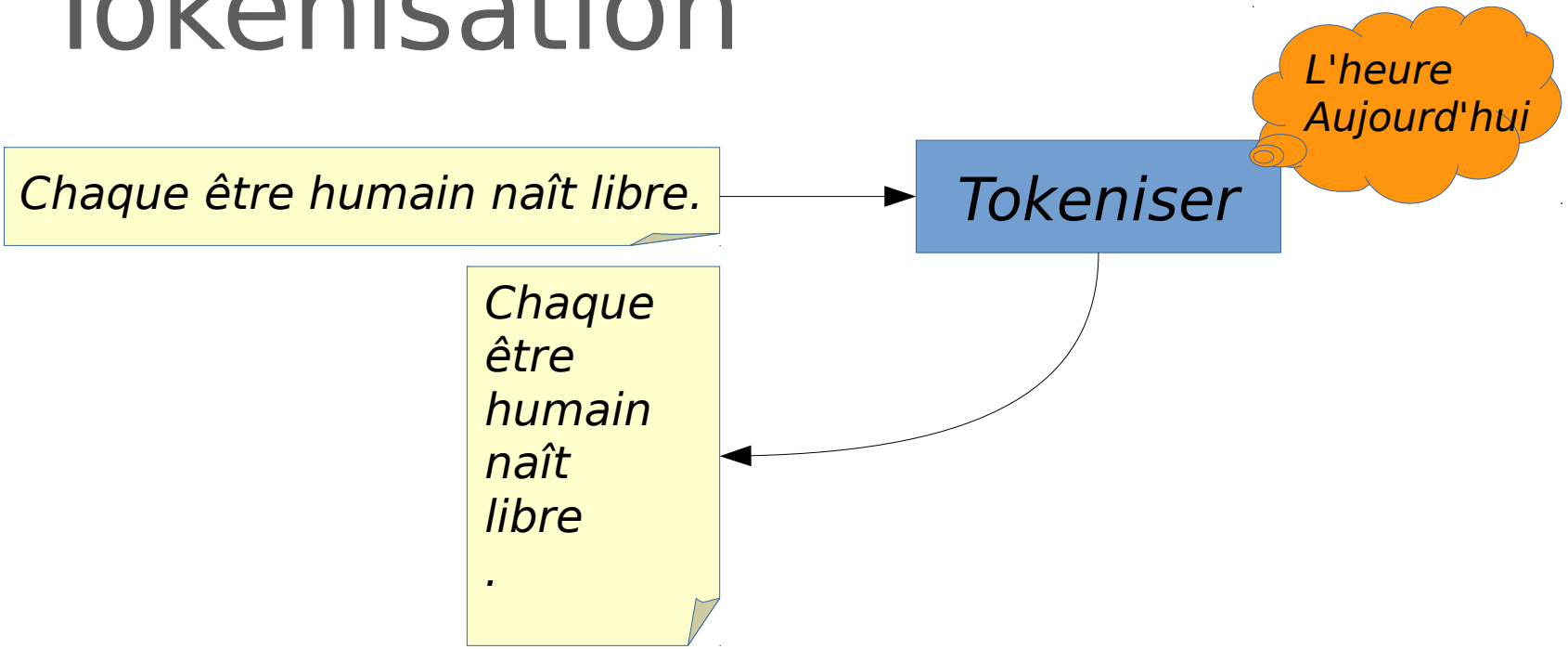
- Talismane
- Plateforme Bonsai
  - http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html
- ...

Fonctionnement d'un *tagger* morpho-syntaxique

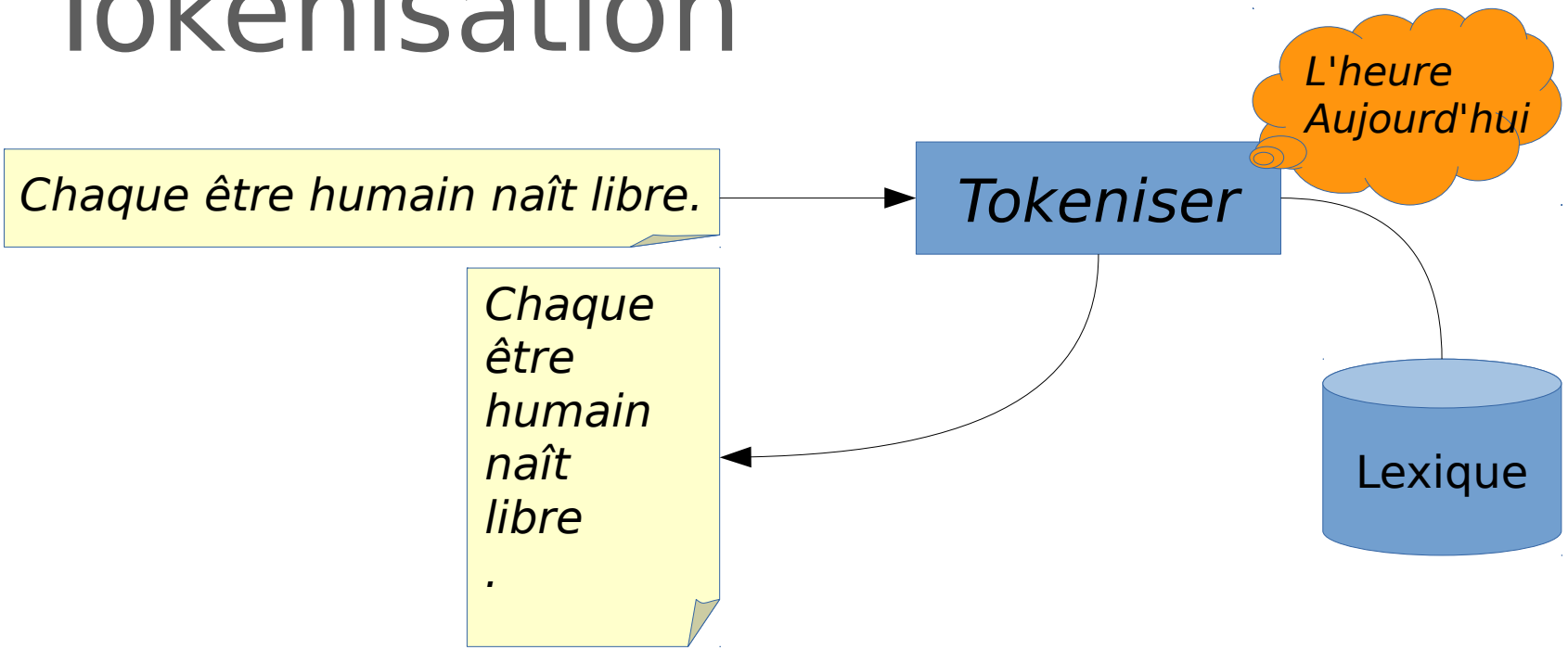
Tokenisation



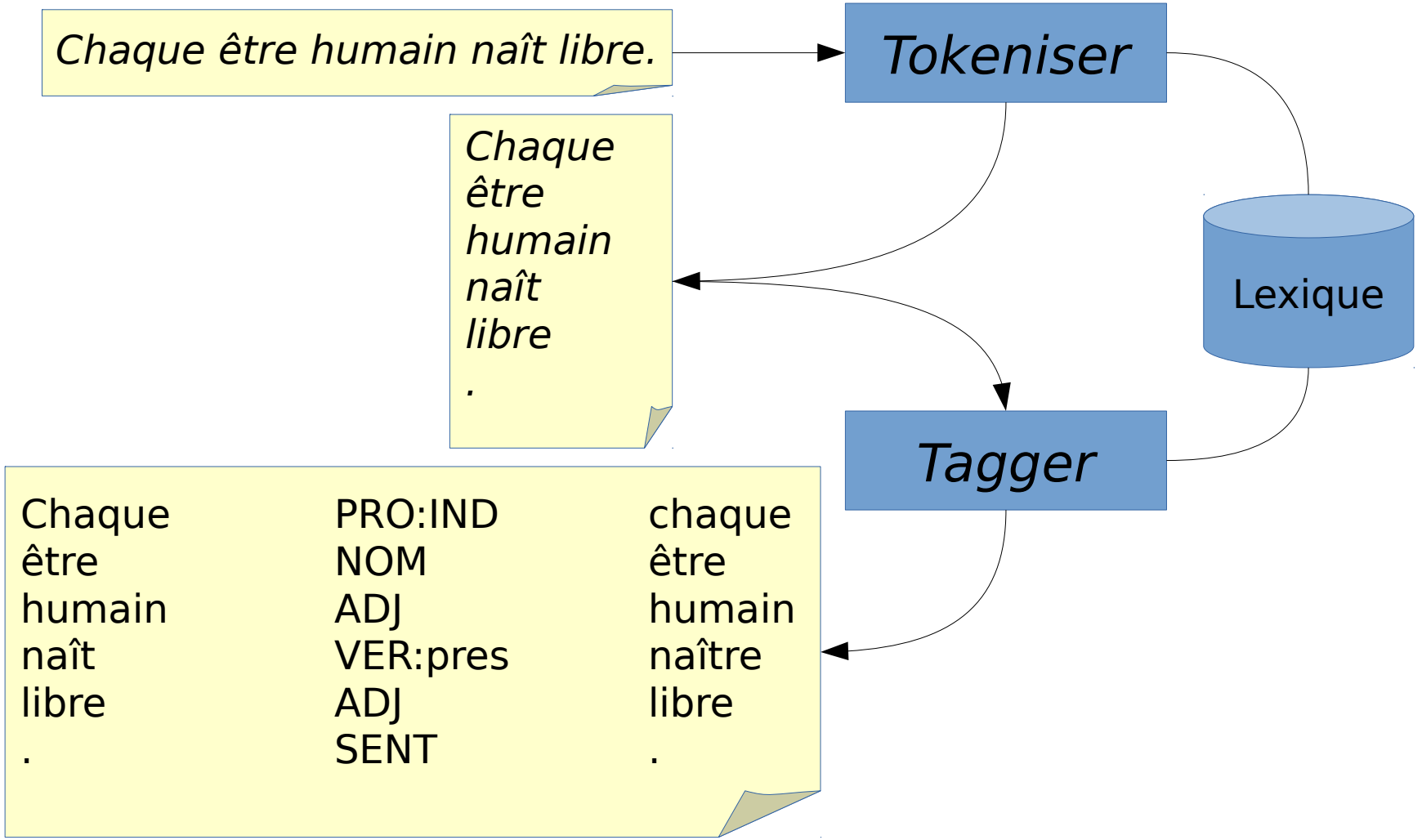
Tokenisation



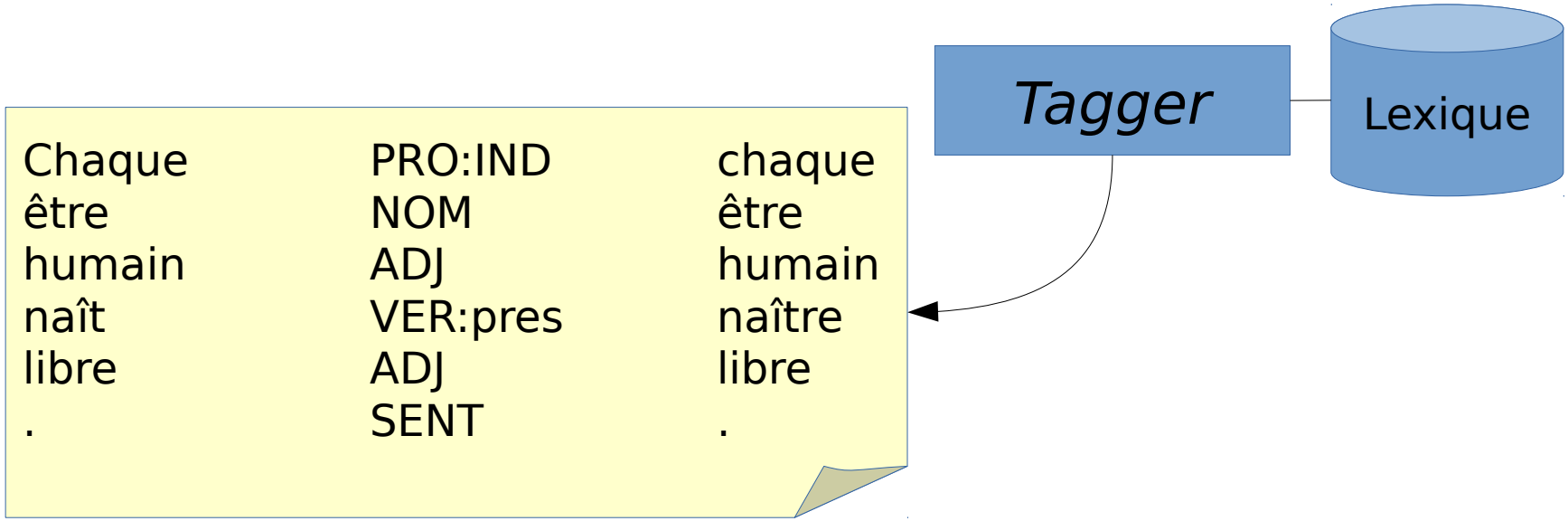
Tokenisation



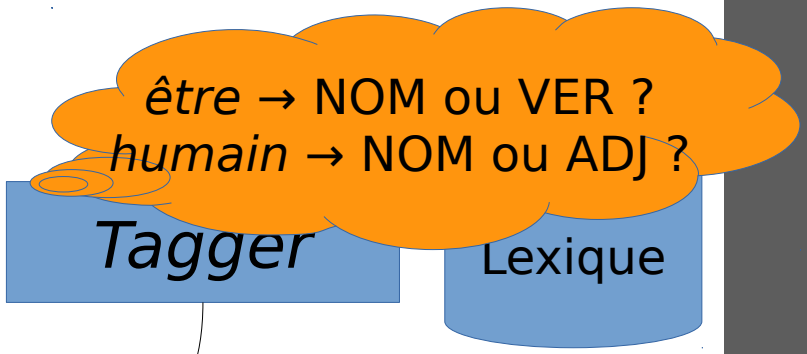
Tokenisation



Tokenisation

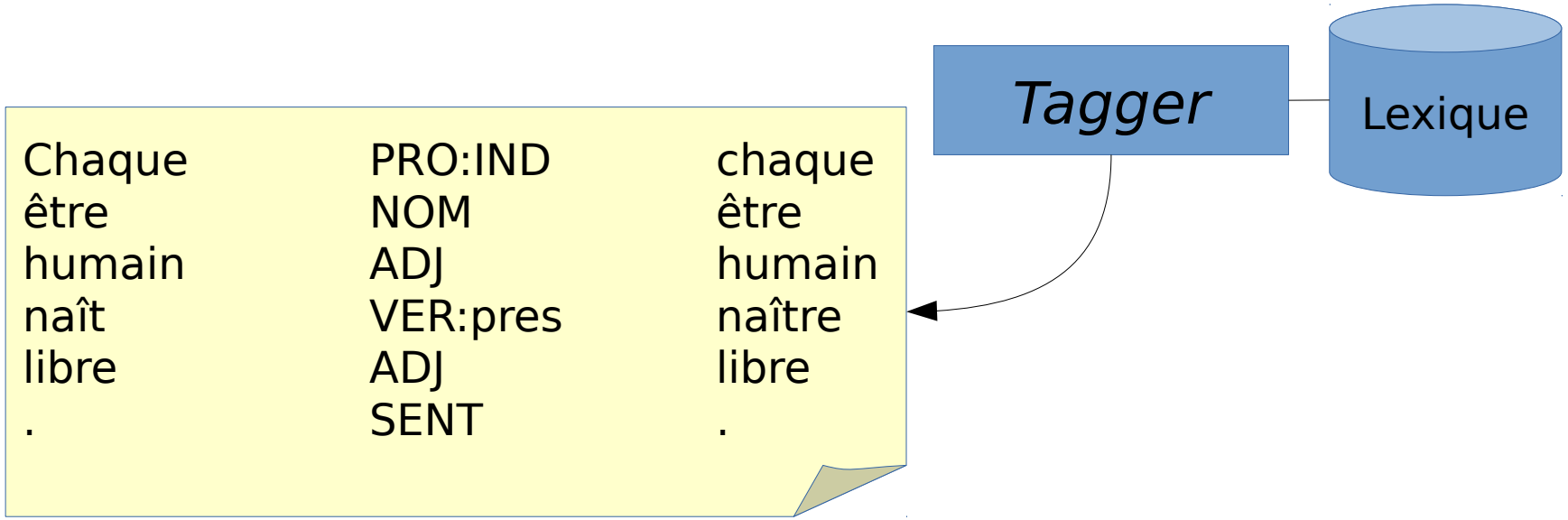


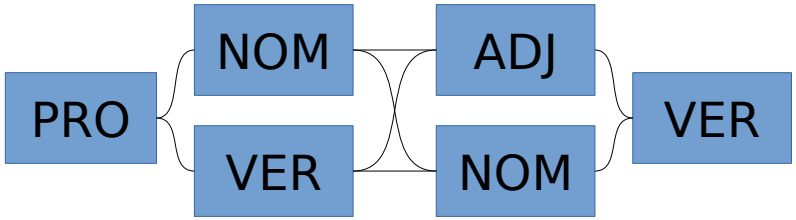
Tokenisation



Chaque	PRO:IND	chaque
être	NOM	être
humain	ADJ	humain
naît	VER:pres	naître
libre	ADJ	libre
.	SENT	.

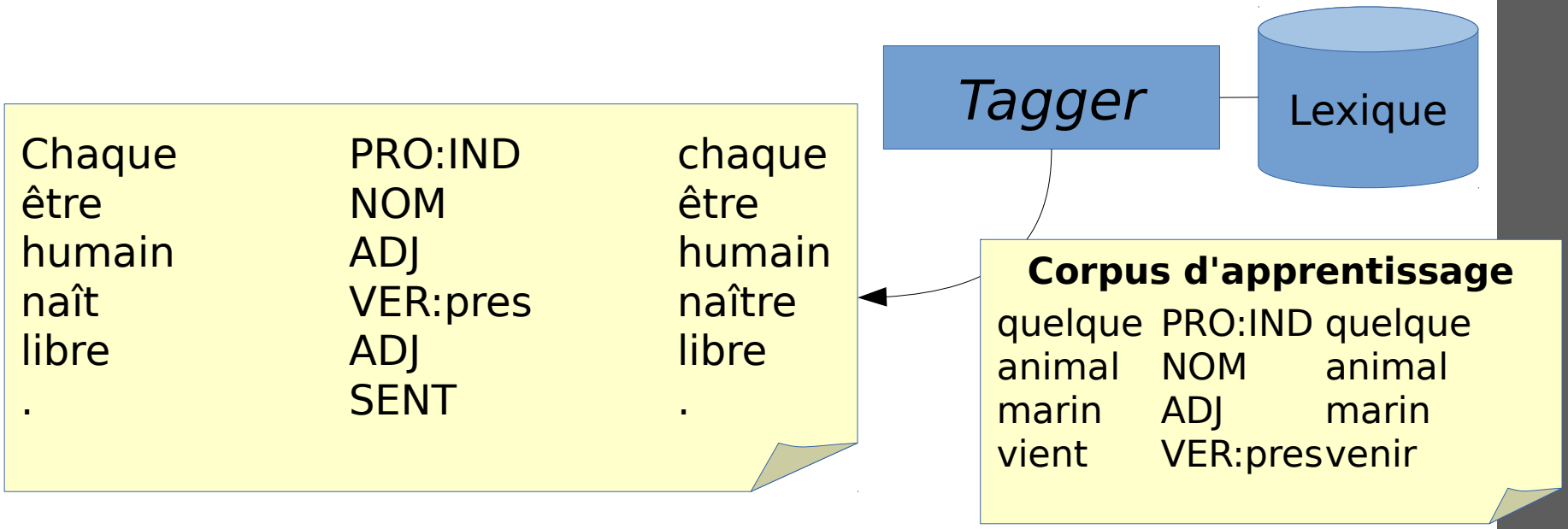
Tokenisation

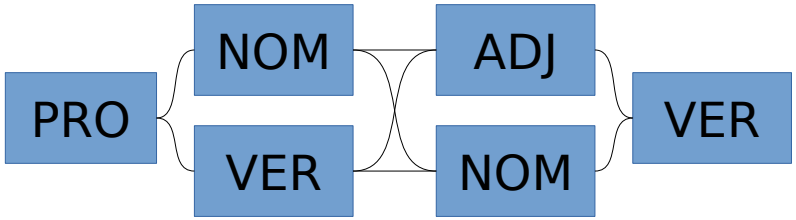




PRO	NOM	ADJ	VER
PRO	VER	ADJ	VER
PRO	NOM	NOM	VER
PRO	VER	NOM	VER

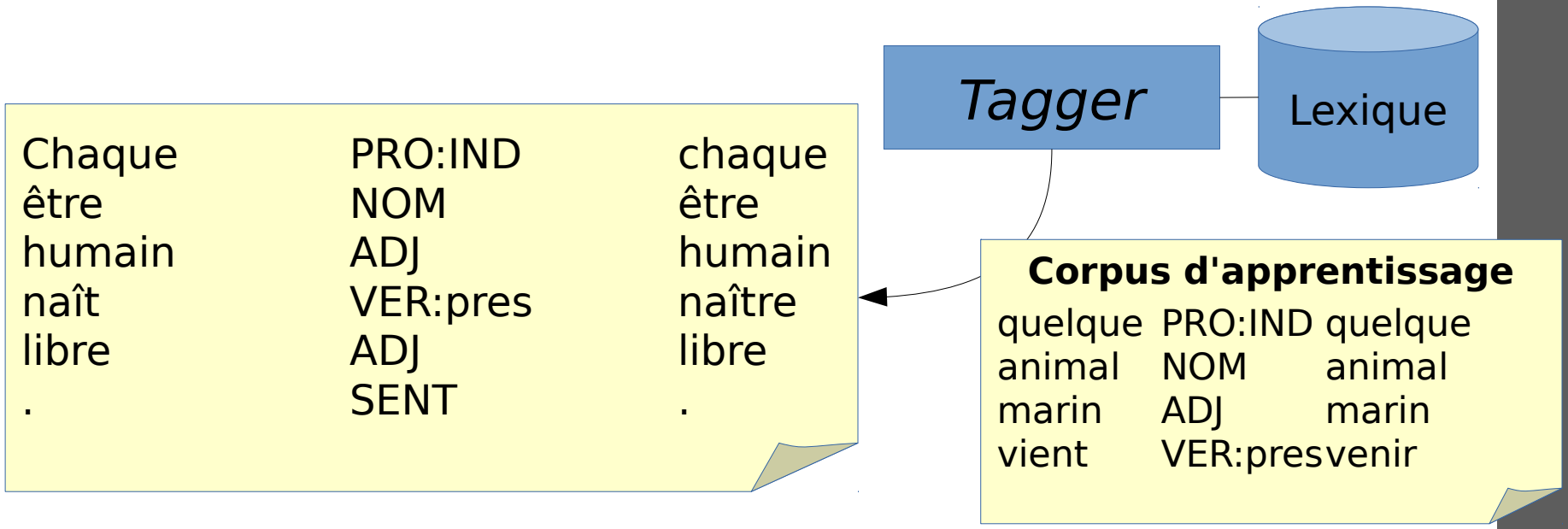
Tokenisation

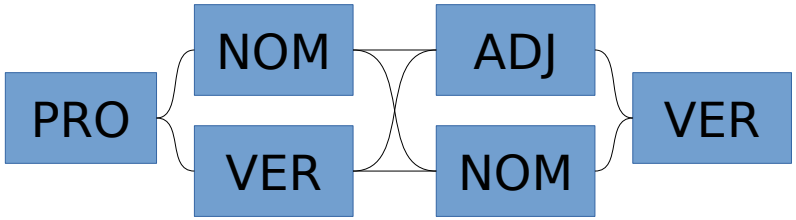




PRO	NOM	ADJ	VER
PRO	VER	ADJ	VER
PRO	NOM	NOM	VER
PRO	VER	NOM	VER

Tokenisation

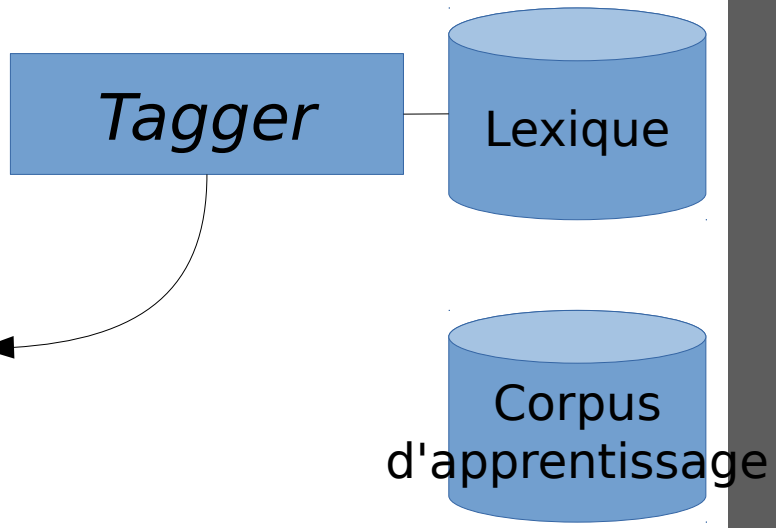


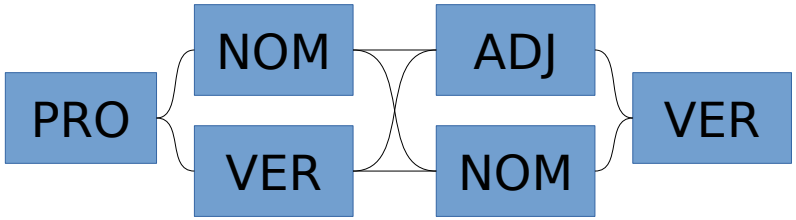


PRO	NOM	ADJ	VER	1
PRO	VER	ADJ	VER	
PRO	NOM	NOM	VER	
PRO	VER	NOM	VER	

Tokenisation

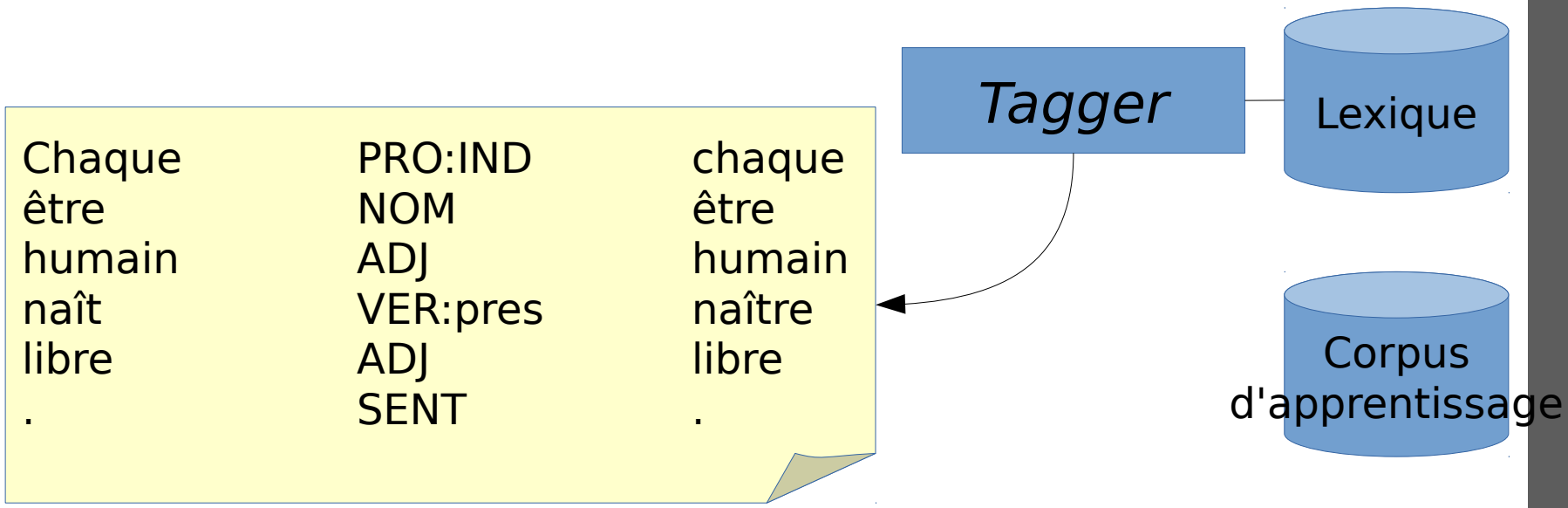
Chaque	PRO:IND	chaque
être	NOM	être
humain	ADJ	humain
naît	VER:pres	naître
libre	ADJ	libre
.	SENT	.

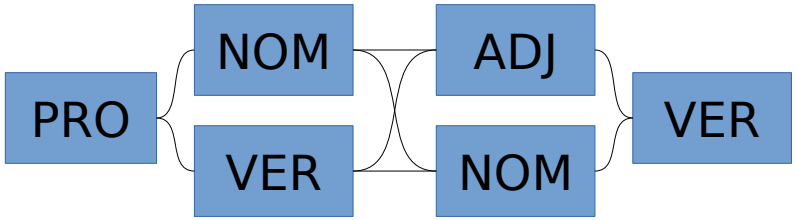


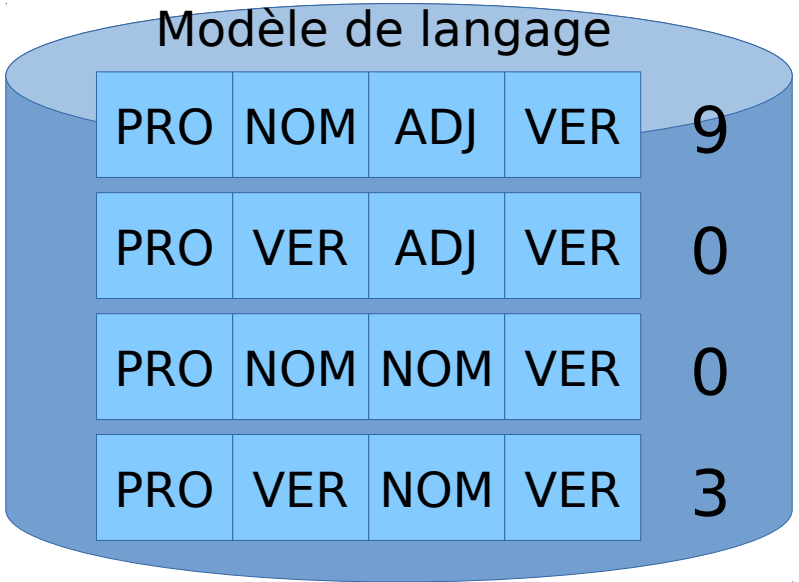


PRO	NOM	ADJ	VER	9
PRO	VER	ADJ	VER	0
PRO	NOM	NOM	VER	0
PRO	VER	NOM	VER	3

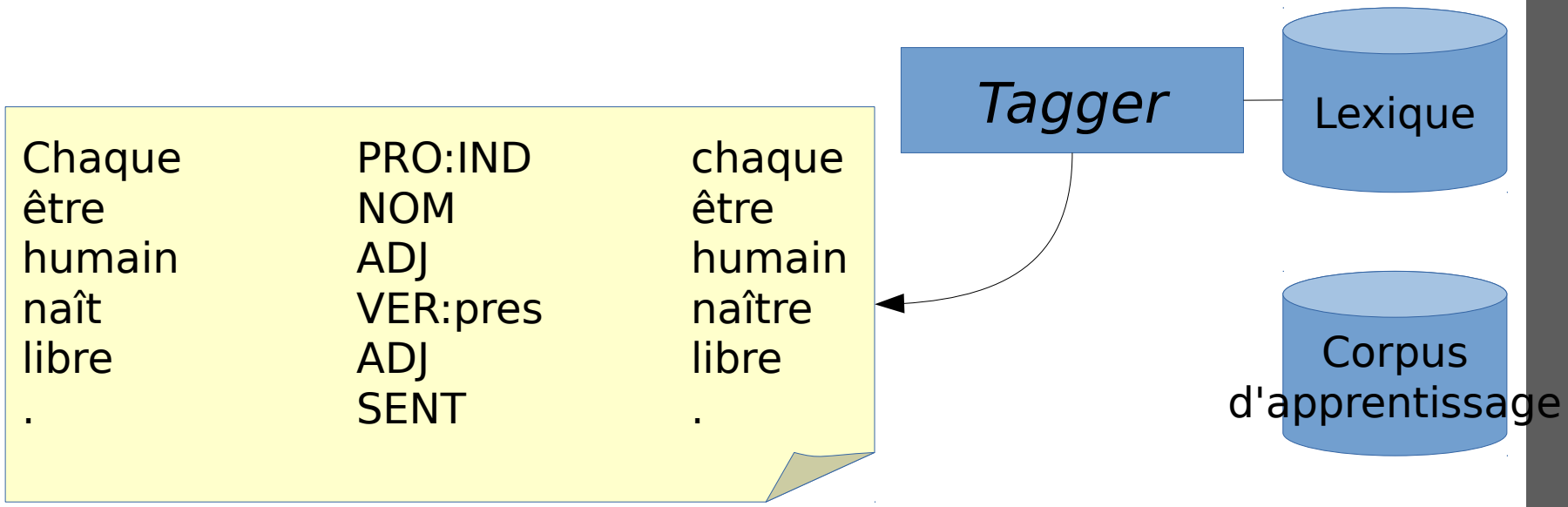
Tokenisation

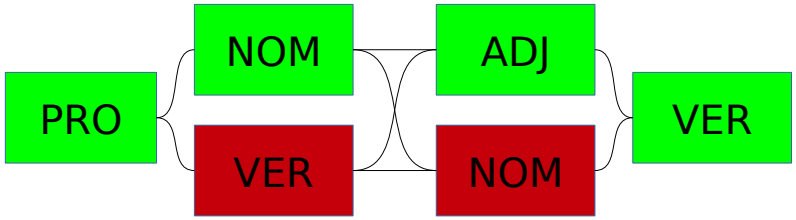


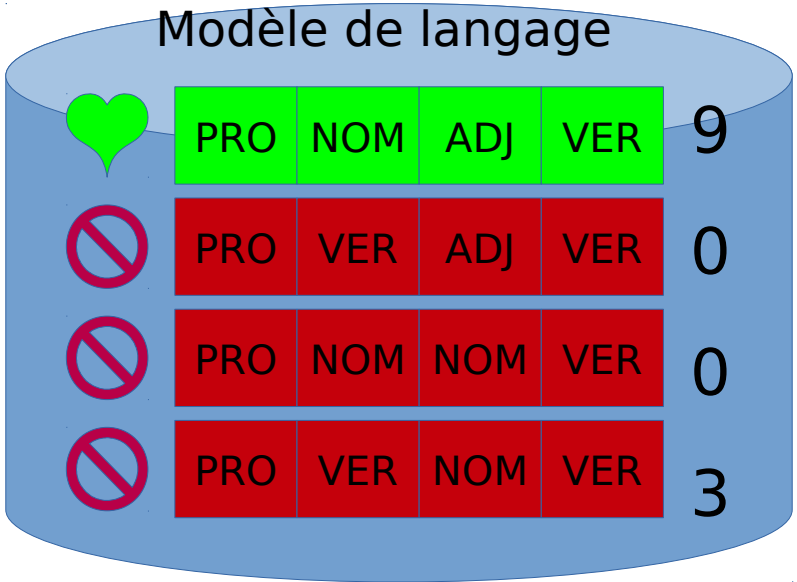




Tokenisation







Conséquences

- Le *tagger* va faire des erreurs
- Dépendant du lexique + corpus d'apprentissage
  - La tokénisation
  - Le jeu d'étiquettes
  - Les lemmes
  - Les performances du tagger
- Plusieurs modèles pour le français
  - Français standard : http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
  - Français parlé : http://cnrtl.fr/corpus/perceo/
  - Français XVIème-XVIIIème : http://presto.ens-lyon.fr/?page_id=197
  - Français IXème-XVème : http://srcmf.org/

Chaîne de traitement : le cas de Presto

Difficultés

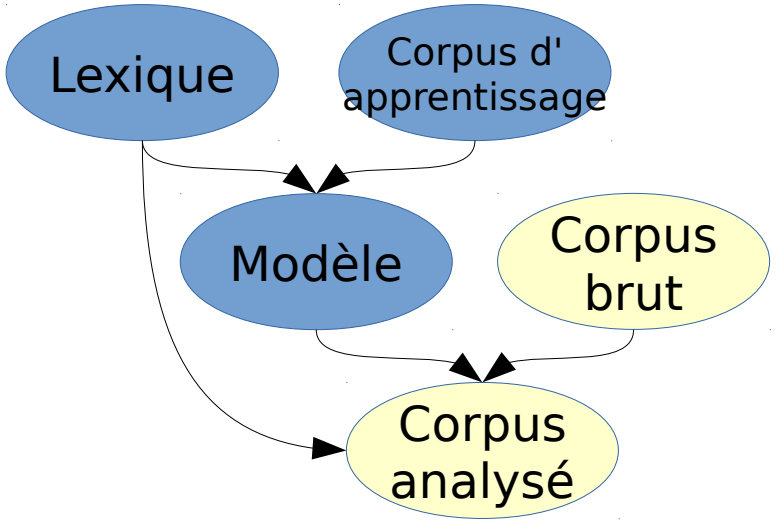
- Français 16e-18e
  - Variantes dialectales
  - Orthographe (y compris segmentation) variable, néologismes, etc.
  - Langue peu dotée

C'est assez dict pour ceste foys.
Quand sçavoir en vous s'assocye,
Monsieur Rien, l'on vous remercye
Du bien qu'avons aprins de vous.
Bazochiens, entendez tous :
Je veulx en triumphant arroy
Eslire et faire ung nouveau roy,
Comme il est coustume de faire ;
Pourtant chacun pense a l'affaire,
Autant les grandz que les petitz,
Et faire les preparatifz ;
Car, ainsi comme liberalle,
Je tendz a monstre generalle
Qui, l'esté qui vient, sera faicte.
En honneur du triumphe et feste,
Ne faillez monstrer vos bons cueurs
Qui font de la vertu approche,
Tant que l'on dye par honneurs :
Vive l'excellente Bazoche !

Sottie pour le cry de la bazoche, 1549

Une approche classique

- Chaîne de traitement



Ressources à créer

Lexique

- Avec l'ATILF (G. Souvay)

Corpus d'apprentissage

- 50k mots annotés manuellement (token, POS, lemme)
- Par au moins 2 annotateurs

Lexique Presto

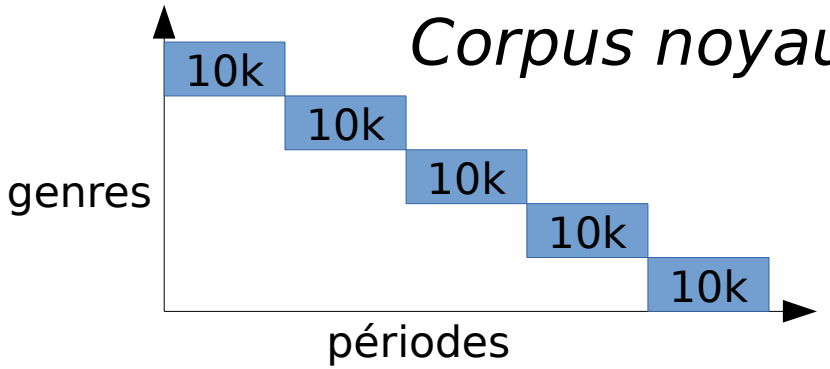
abasourdi;ABASOURDIR;Ge;LEFFF:VMP00SM;0;36,2103;
abasourdie;ABASOURDIR;Ge;LEFFF:VMP00SF;0;36,2104;
abasourdies;ABASOURDIR;Ge;LEFFF:VMP00PF;0;36,2105;
abasourdiez;ABASOURDIR;Ge;LEFFF:VMP00PF@abasourdies;0;36,2106;
abasourdimes;ABASOURDIR;Vvc;LEFFF:VMIS1P0@abasourdîmes;0;36,2107;
abasourdimez;ABASOURDIR;Vvc;LEFFF:VMIS1P0@abasourdîmes@abasourdimes;0;36,2108;
abasourdimés;ABASOURDIR;Vvc;LEFFF:VMIS1P0@abasourdîmes@abasourdimes@abasourdimez;0;36,2109;
abasourdir;ABASOURDIR;Vvn;LEFFF:VMN0000;0;36,2110;
abasourdira;ABASOURDIR;Vvc;LEFFF:VMIF3S0;0;36,2111;
abasourdirai;ABASOURDIR;Vvc;LEFFF:VMIF1S0;0;36,2112;
abasourdiraient;ABASOURDIR;Vvc;LEFFF:VMIC3P0;0;36,2113;
abasourdirais;ABASOURDIR;Vvc;LEFFF:VMIC2S0;0;36,2114;
abasourdirait;ABASOURDIR;Vvc;LEFFF:VMIC3S0;0;36,2115;
abasourdiraiz;ABASOURDIR;Vvc;LEFFF:VMIC2S0@abasourdirais;0;36,2116;
abasourdiras;ABASOURDIR;Vvc;LEFFF:VMIF2S0;0;36,2117;
abasourdiray;ABASOURDIR;Vvc;LEFFF:VMIF1S0@abasourdirai;0;36,2118;
abasourdirays;ABASOURDIR;Vvc;LEFFF:VMIC2S0@abasourdirais;0;36,2119;
abasourdirayz;ABASOURDIR;Vvc;LEFFF:VMIC2S0@abasourdirais@abasourdirays;0;36,2120;

Corpus d'apprentissage Presto

2;AVANT-PROPOS;AVANT-PROPOS;Nc
3;qui;QUI;Pr
4;contient;CONTENIR;Vvc
5;le;LE;Da
6;plan;PLAN;Nc
7;de;DE;S
8;cet;CE;Dd
9;ouvrage;OUVRAGE;Nc

Corpus d'apprentissage

- Sélection de 5 textes
  - 5 périodes
  - 5 genres



Total : 62k tokens

Saulsaye (1547)

Lisandre et Caliste (1631)

Les Lettres de messire Roger de Rabutin, comte de Bussy (1681)

Essay sur l'histoire generale et sur les moeurs et sur l'esprit des nations (1756)

Le Paysan perverti ou les Dangers de la ville (1776)

Normalisation des textes

C'est assez dict pour ceste foys.<lb/>
Quand sçavoir en vous s'assocye,<lb/>
Monsieur *Rien, l'on vous remercye<lb/>
Du bien qu'avons aprins de vous.<lb/>
Bazochiens, entendez tous :<lb/>
Je veulx en triumphant arroy<lb/>
Eslire et faire ung nouveau roy,<lb/>
Comme il est coustume de faire ;<lb/>
Pourtant chacun pense a l'affaire,<pb n="267"/>
Autant les grandz que les petitz,<lb/>
Et faire les preparatifz ;<lb/>
Car, ainsi comme liberalle,<lb/>
Je tendz a monstre generalle<lb/>
Qui, l'esté qui vient, sera faicte.<lb/>
En honneur du triumphe et feste,<lb/>
Ne faillez monstrer vos bons cueurs<lb/>
Qui font de la vertu approche,<lb/>
Tant que l'on dye par honneurs :<lb/>
Vive l'excellente *Bazoche !</p>

Normalisation des textes

<lb/>Mais par telle legierete ne convient esti
<lb rend="hyphen"/>mer les oeuvres des humains. Car
<lb/>vous mesmes dictes, que
<choice><orig>lhabit</orig><reg>l’habit</reg></choice> ne faict
<lb/>poinct le moine: & tel est vestu
<choice><orig>dhabit</orig><reg>d’habit</reg></choice>
<lb/>monachal, qui au dedans
<choice><orig>nest</orig><reg>n’est</reg></choice> rien
moins
<lb/>que moyne: & tel est vestu de cappe hes-
<fw place="bot-center" type="sig">A iij</fw>

Projection lexicale

quelques
remarques
sur
les
groupements

Projection lexicale

quelques	QUELQUE:Ag|QUELQUE:Di
remarques	REMARQUE:Nc|REMARQUER:Vvc
sur	SUR:Ag|SUR:Sp|SÛR:Ag|SÛR:R
les	LE:Da|LES:Pp|LÈS:Sp|LÉ:Nc
groupements	GROUPEMENT:Nc

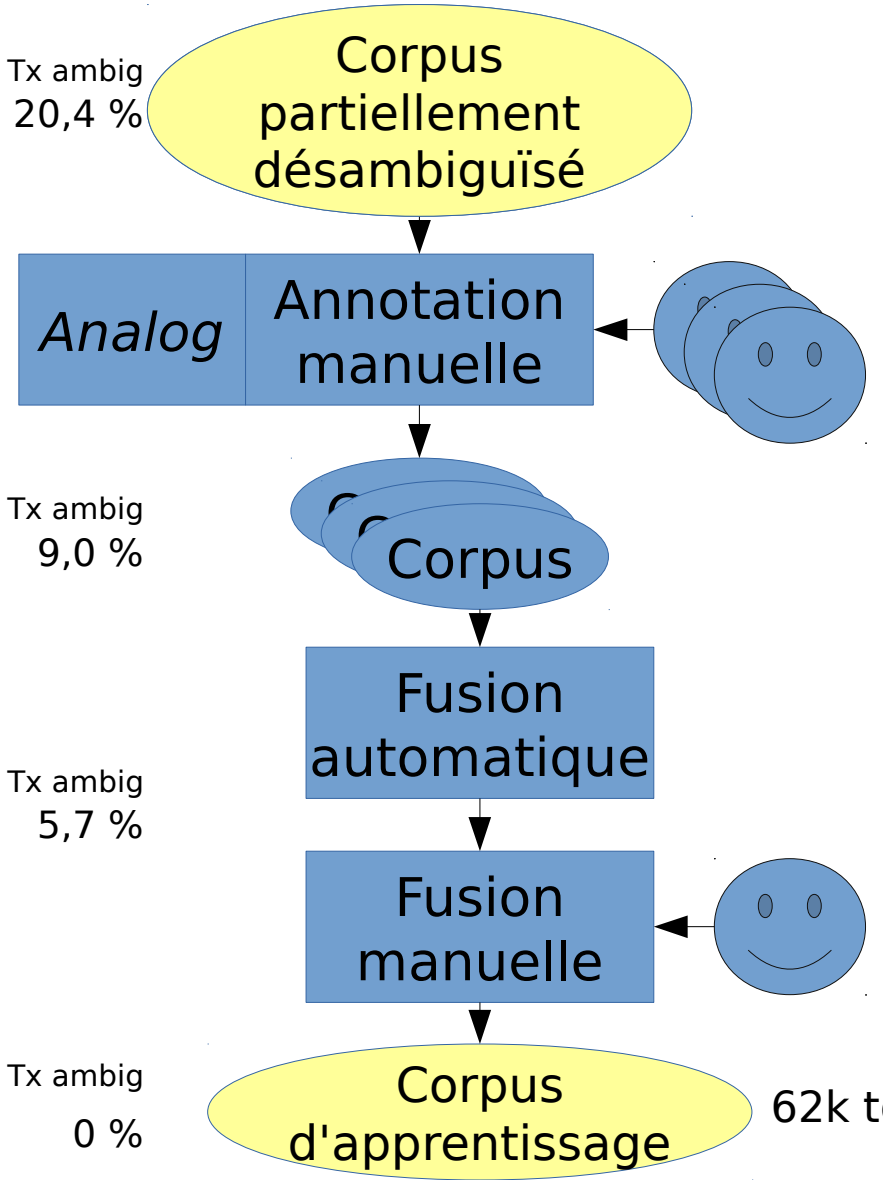
Projection lexicale

quelques	QUELQUE:Ag|QUELQUE:Di
remarques	REMARQUE:Nc|REMARQUER:Vvc
sur	SUR:Ag|SUR:Sp|SÛR:Ag|SÛR:R
les	LE:Da|LES:Pp|LÈS:Sp|LÉ:Nc
groupements	GROUPEMENT:Nc

Utilisation d'un modèle de langage moderne

quelques	QUELQUE:Aq|QUELQUE:Di
remarques	REMARQUE:Nc|~~REMARQUER:Vvc~~
sur	~~SUR:Aq~~|SUR:Sp|~~SÛR:Aq|SÛR:R~~
les	LE:Da|~~LES:Pp|LÈS:Sp|LÉ:Nc~~
groupements	GROUPEMENT:Nc

Annotation manuelle et fusion



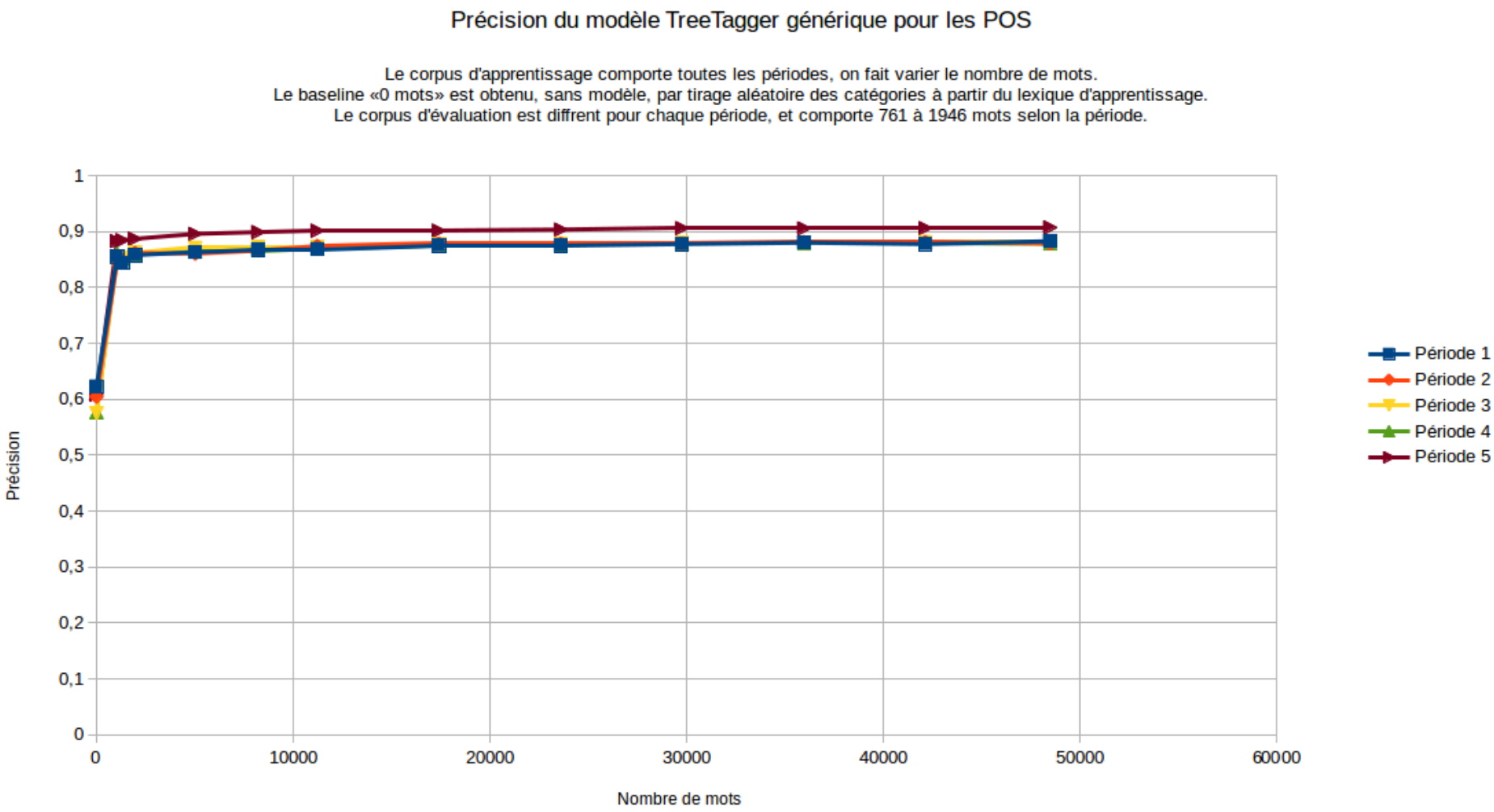
- Fusion automatique pour les cas « évidents » :
  - Au moins 2 annotateurs d'accord
  - Diacritiques

Création du modèle

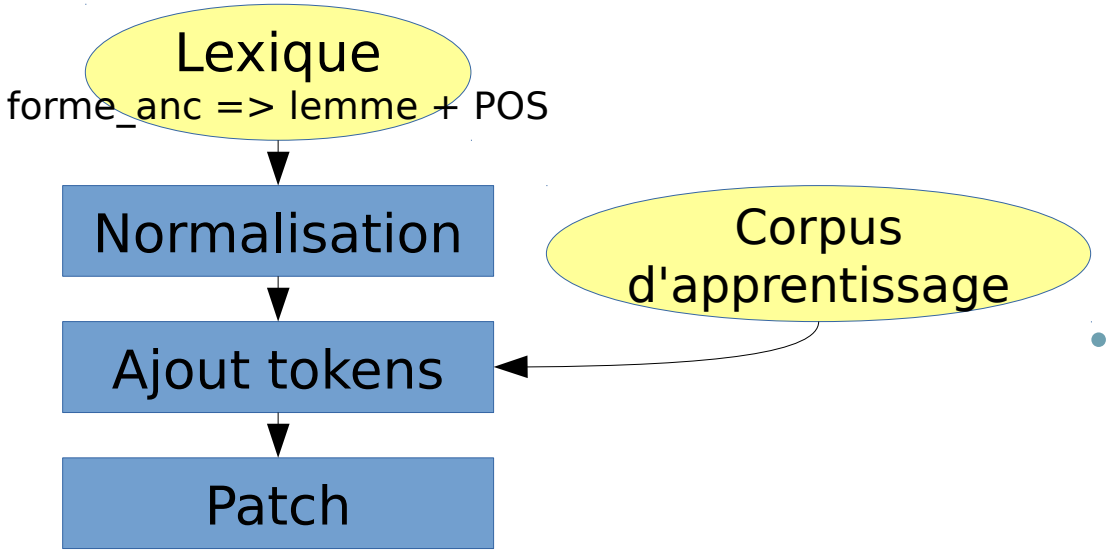
- Division du corpus d'apprentissage en trois
  - Corpus d'entraînement (80% − 49 630 tokens)
  - Corpus de développement (10% − 6 164 tokens)
  - Corpus de référence (10% − 6 110 tokens)

3. Création d'un modèle de langue pour le français classique

Évaluation du modèle



Préparation du lexique



- Patch :
  - Listes de tokens
    - À ajouter
    - À enlever
  - Règles ad hoc

Évaluation du modèle