ENS DE LYON

icar
UMR 5191 - CNRS

Complexités d'un corpus en diachronie du français. Le cas de Presto.

Achille Falaise
Denis Vigier

Atelier CCC
Lyon, 4 juin 2015

Plan de la présentation

1. Le corpus
2. Chaîne de traitement
   1. Stratégie de traitement
   2. Création d'un lexique
   3. Création d'un corpus d'apprentissage
   4. Création d'un modèle de langage
   5. Analyse du corpus
3. Quelques problèmes épineux
   1. Jeu d'étiquettes : participes, gérondifs, adjectifs
   2. Les amalgames

Le corpus Presto

Le corpus Presto

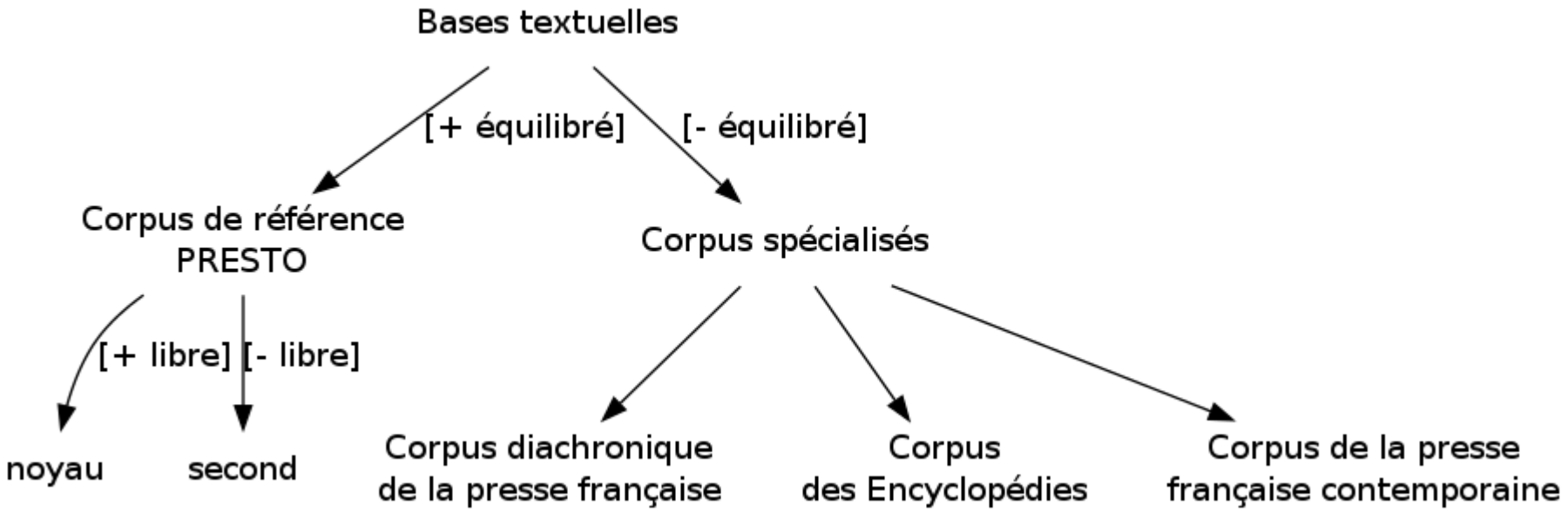
- Constitution d'un corpus dans le cadre de PRESTO :
  - Prérequis pour la réalisation du projet
  - Un des apports majeurs du projet
- Objectifs linguistiques :
  - Proposer une couverture satisfaisante de toutes les périodes de l'histoire du français : 9ème s. au 20ème s.
  - Partir des bases existantes et construire un corpus « contrôlé »
    - Critère temporel : éviter les « trous »
    - Critère générique : variété contrôlée
    - « Qualité » des éditions + orthographe non modernisée
  - => *a minima* : chaque tranche décennale : 3 txt litt (GN, Th, P) + 3 traités
  - Enrichissement linguistique : étiquetage morpho-syntaxique et lemmatisation

Le corpus Presto

- **Objectifs** en termes d' « ouverture » :
  - Rendre disponible (licences CC)
    - Une partie du corpus constitué
    - La totalité des outils élaborés

→ Collaborations avec les Bases textuelles : Frantext, BVH, ARTFL, CEPM, Cologne
  - Textes déjà sous licence CC : BVH, CEPM
  - Conventions passées
    - FRANTEXT : mise à disposition de la communauté sous licence CC d'une trentaine de textes jusqu'ici « fermés » ; mise à disposition de lexiques pour les outils PRESTO sous licence CC
    - ARTFL : mise à disposition d'un volume de l'Encyclopédie Diderot & D'Alembert

Le corpus Presto



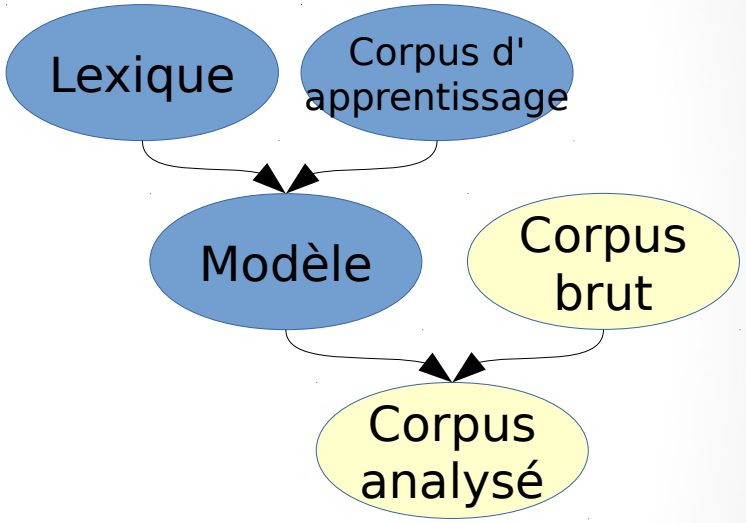
Chaîne de traitement

Chaîne de traitement

Stratégie de traitement

Une approche classique

- Approche « classique » : chaîne de traitement



- Contraintes :
  - « Massif »
  - Pas de *feedback*
  - Flux de *tokens*
    => pas d'ambiguïtés de segmentation

Des problèmes originaux

- Français classique
  - Variantes dialectales
  - Orthographe (y compris segmentation) variable, néologismes, etc.
  - Langue peu dotée
- Méthode
  - construction de ressources à partir du français moderne

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

Chaîne de traitement

Création d'un lexique

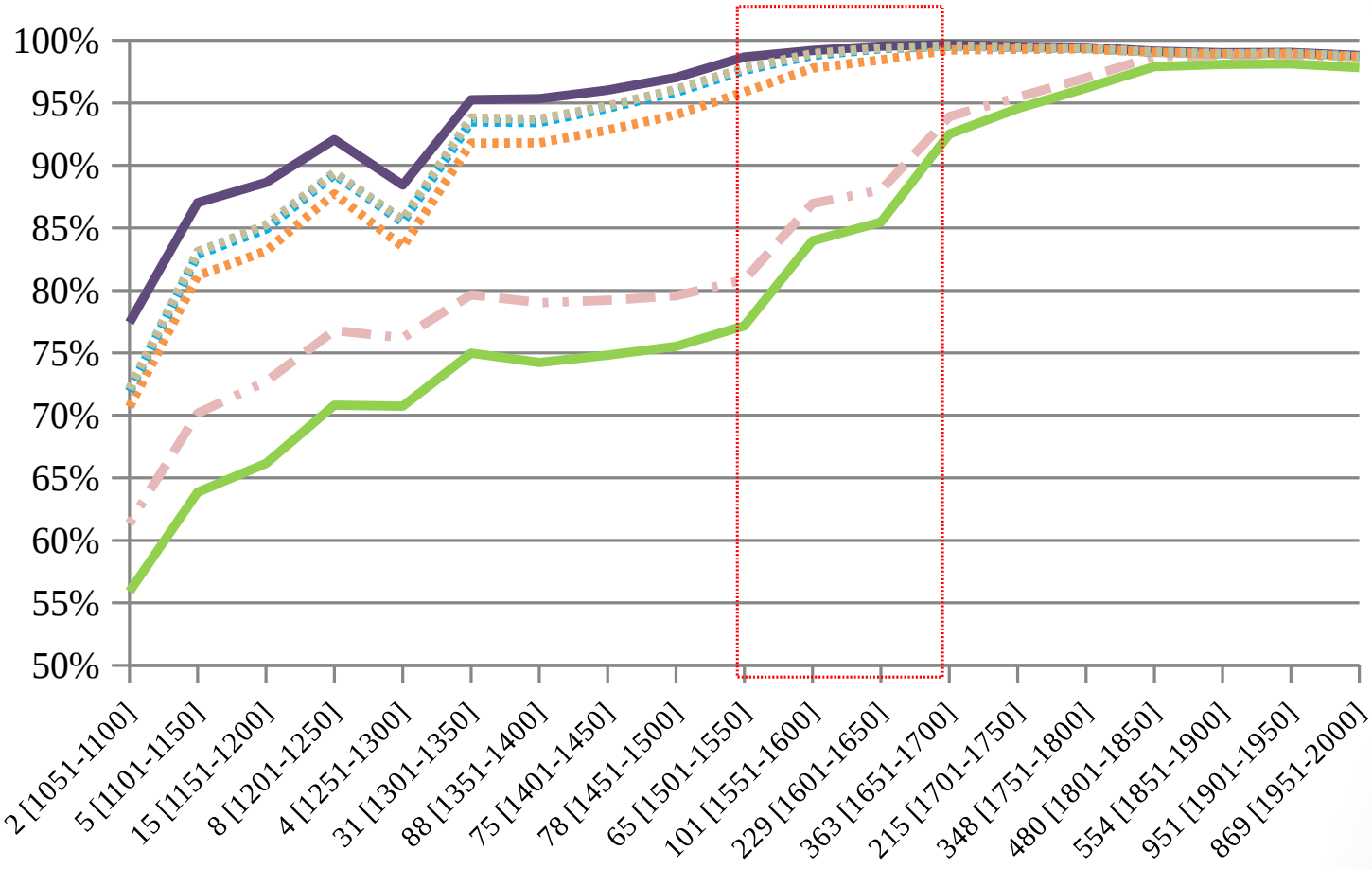
Lexique (Gilles Souvay, ATILF)

Forme ancienne => lemme moderne + POS

```
aimas;AIMER;VMIS2S0
aimasmes;AIMER;VMIS1P0
aimasse;AIMER;VMSI1S0
aimassent;AIMER;VMSI3P0
aimasses;AIMER;VMSI2S0
aimassies;AIMER;VMSI2P0
aimassiez;AIMER;VMSI2P0
aimassions;AIMER;VMSI1P0
aimassiés;AIMER;VMSI2P0
aimast;AIMER;VMSI3S0
aimastes;AIMER;VMIS2P0
```

2 735 843 entrées

Lexique (Gilles Souvay, ATILF)



Taux de couverture lexicale du lexique moderne (lexique de départ) et du lexique archaïsé (lexique final), avec étapes intermédiaires, mesurée sur le corpus Frantext (XIème - XXème siècle).

Chaîne de traitement

Création d'un corpus de référence

Choix des textes

- Sélection de 5 textes
  - 5 périodes
  - 5 genres

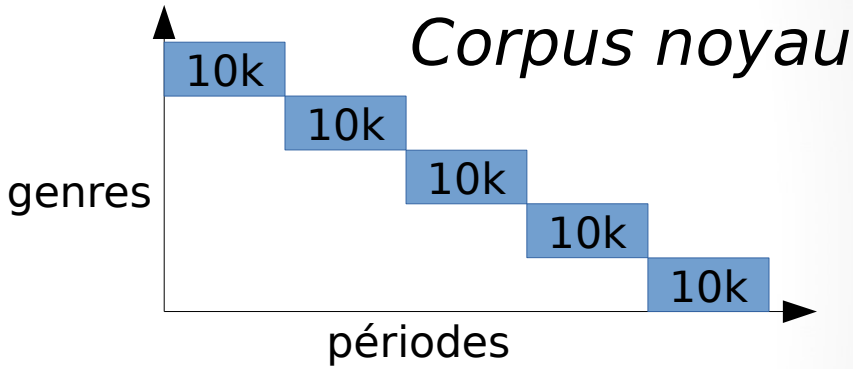
Saulsaye (1547)

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)



Total : 62k tokens

Objectifs et défis

- Objectifs
  - 50k mots annotés manuellement (token, POS, lemme)
  - Par au moins 2 annotateurs
- Défis
  - C'est très chronophage !
- Méthode
  - Pré-annotation automatique du corpus

Objectifs et défis

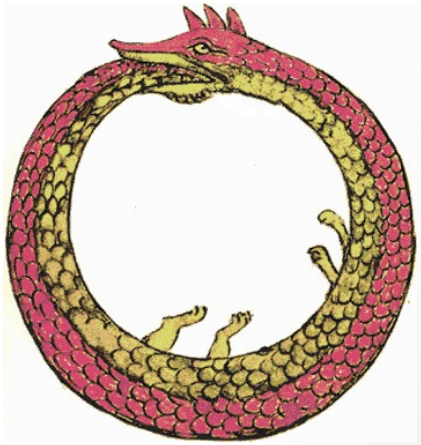
- Objectifs
  - 50k mots annotés manuellement (token, POS, lemme)
  - Par au moins 2 annotateurs
- Défis
  - C'est très chronophage !
- Méthode
  - Pré-annotation automatique du corpus

Pour créer ce corpus annoté,
il faut donc…

Objectifs et défis

- Objectifs
  - 50k mots annotés manuellement (token, POS, lemme)
  - Par au moins 2 annotateurs
- Défis
  - C'est très chronophage !
- Méthode
  - Pré-annotation automatique du corpus

Pour créer ce corpus annoté, il faut donc… un corpus annoté !

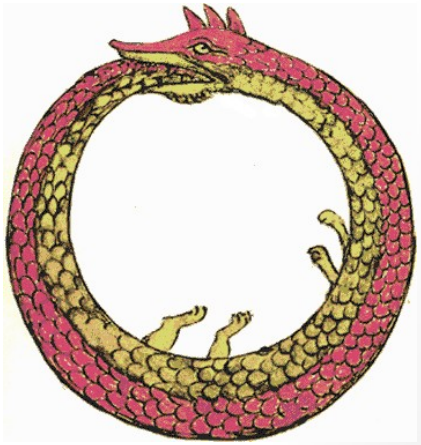


Objectifs et défis

- Objectifs
  - 50k mots annotés manuellement (token, POS, lemme)
  - Par au moins 2 annotateurs
- Défis
  - C'est très chronophage !
- Méthode
  - Pré-annotation automatique du corpus

Pour créer ce corpus annoté, il faut donc… un corpus annoté !

Mais il peut être incomplet.



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

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 !

Normalisation des textes

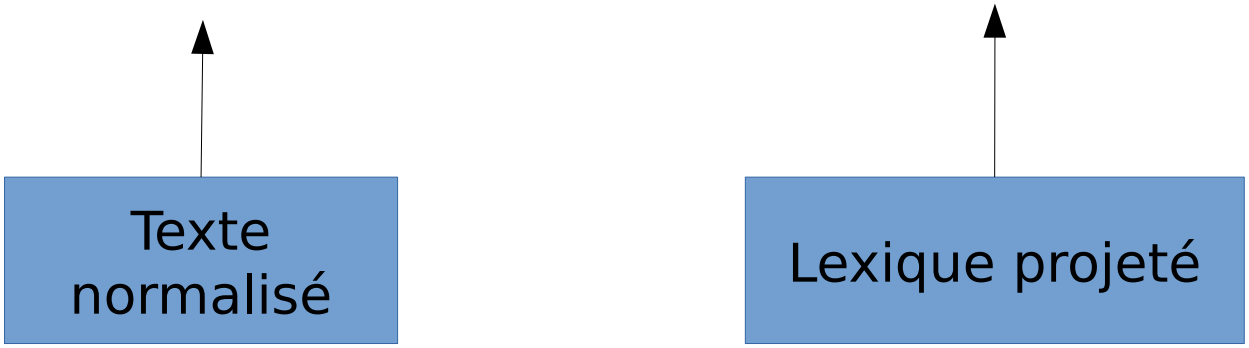
<lb/>Mais par telle legierete ne convient esti
<lb rend="hyphen"/>mer les oeuvres des humains. Car
<lb/>vous mesmes dictes, que
<choice><orig>lhabit</orig><reg>l’habit</reg></choice> ne faict
<lb/>poinct le moine: & tel est vestu
<choice><orig>dhabit</orig><reg>d’habit</reg></choice>
<lb/>monachal, qui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

Normalisation des textes

Mais par telle legierete ne convient estimer les oeuvres des humains. Car vous mesmes dictes, que l'habit ne faict poinct le moine: & tel est vestu d'habit monachal, qui au dedans n'est rien moins que moyne: & tel est vestu de cappe *(...)*

Projection lexicale

quelques	QUELQUE:AQ0CP0|QUELQUE:DI0CP0
remarques	REMARQUE:NCFP000|REMARQUER:VMIP2S0| REMARQUER:VMP00PM|REMARQUER:VMSP2S0
sur	SUR:AQ0CS0|SUR:SPS00|SÛR:AQ0MS0|SÛR:RG
les	LE:DA0CP0|LES:PP3CPA00|LÈS:SPS00|LÉ:NCMP000
groupements	GROUPEMENT:NCMP000



Projection lexicale

quelques	QUELQUE:AQ0CP0|QUELQUE:DI0CP0
remarques	REMARQUE:NCFP000|REMARQUER:VMIP2S0| REMARQUER:VMP00PM|REMARQUER:VMSP2S0
sur	SUR:AQ0CS0|SUR:SPS00|SÛR:AQ0MS0|SÛR:RG
les	LE:DA0CP0|LES:PP3CPA00|LÈS:SPS00|LÉ:NCMP000
groupements	GROUPEMENT:NCMP000

Que d'ambiguïtés !

- simplification du jeu d'étiquettes
- désambiguïsation à l'aide d'un modèle moderne

Simplification des étiquettes

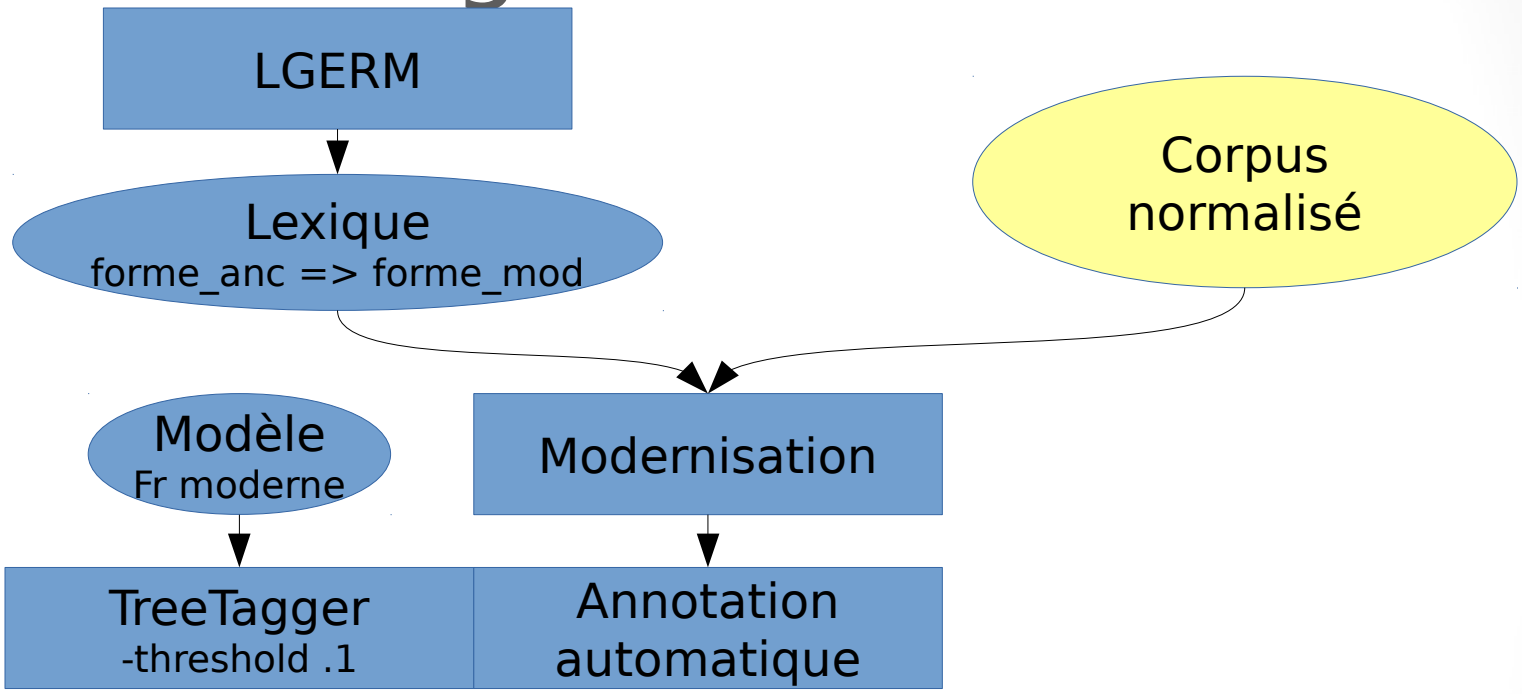
Presto_0

quelques	QUELQUE:AQ0CP0|QUELQUE:DI0CP0
remarques	REMARQUE:NCFP000|REMARQUER:VMIP2S0| REMARQUER:VMP00PM|REMARQUER:VMSP2S0
sur	SUR:AQ0CS0|SUR:SPS00|SÛR:AQ0MS0|SÛR:RG
les	LE:DA0CP0|LES:PP3CPA00|LÈS:SPS00|LÉ:NCMP000
groupements	GROUPEMENT:NCMP000

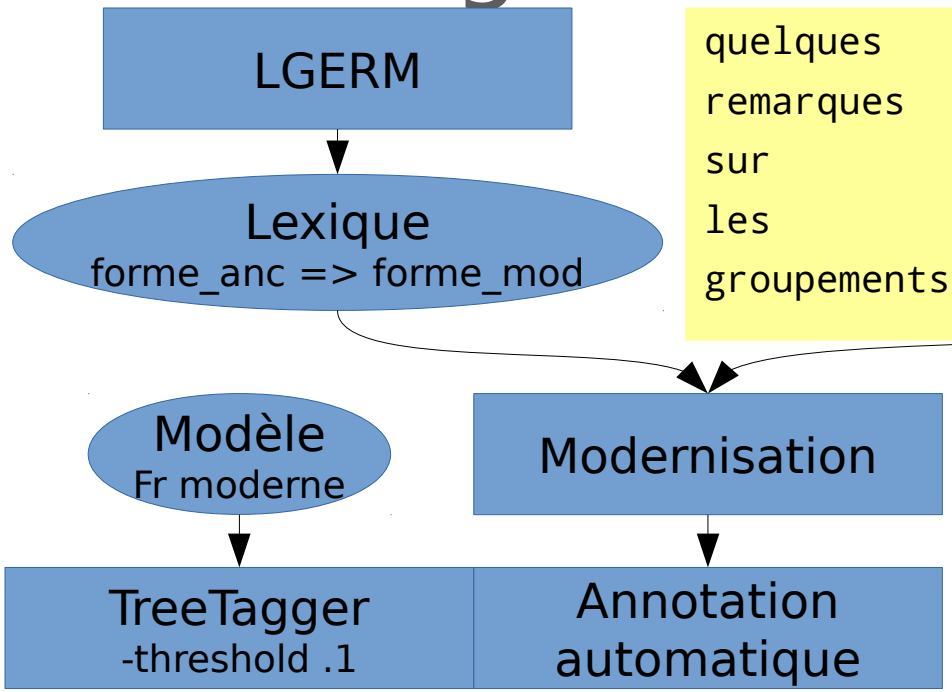
Presto_2

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

Désambiguïsation

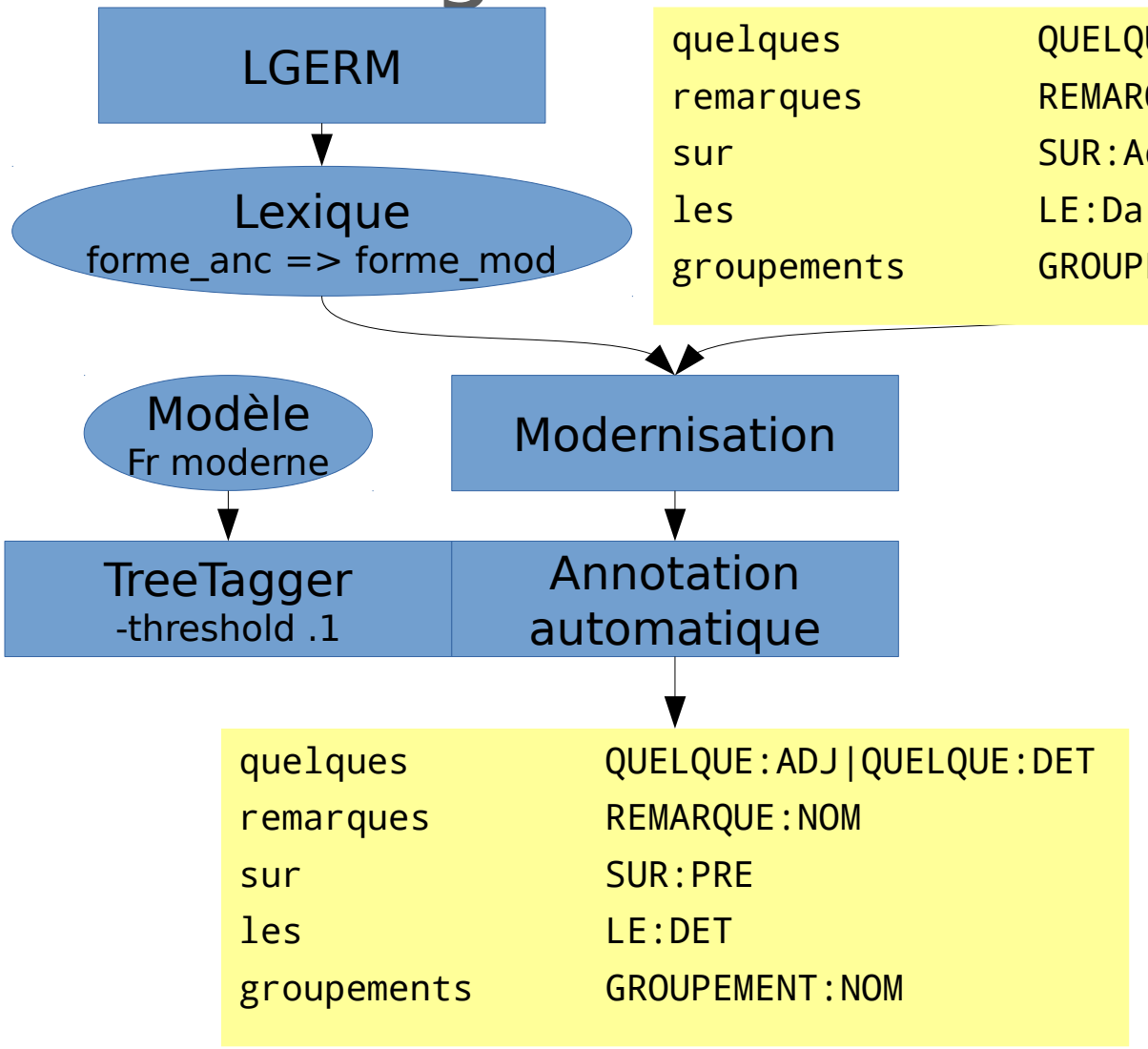


Désambiguïsation



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

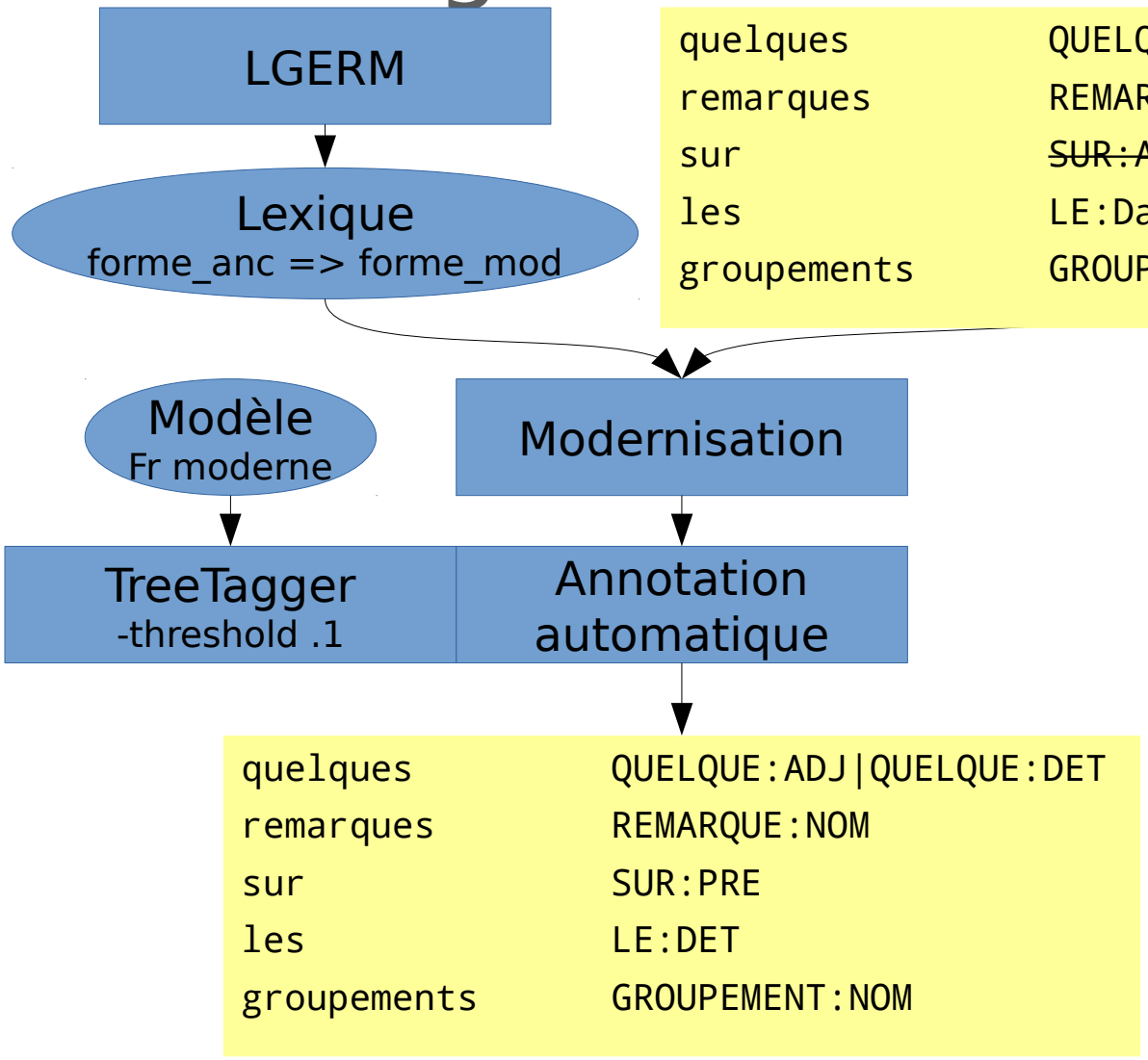
Désambiguïsation



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

quelques	QUELQUE:ADJ|QUELQUE:DET
remarques	REMARQUE:NOM
sur	SUR:PRE
les	LE:DET
groupements	GROUPEMENT:NOM

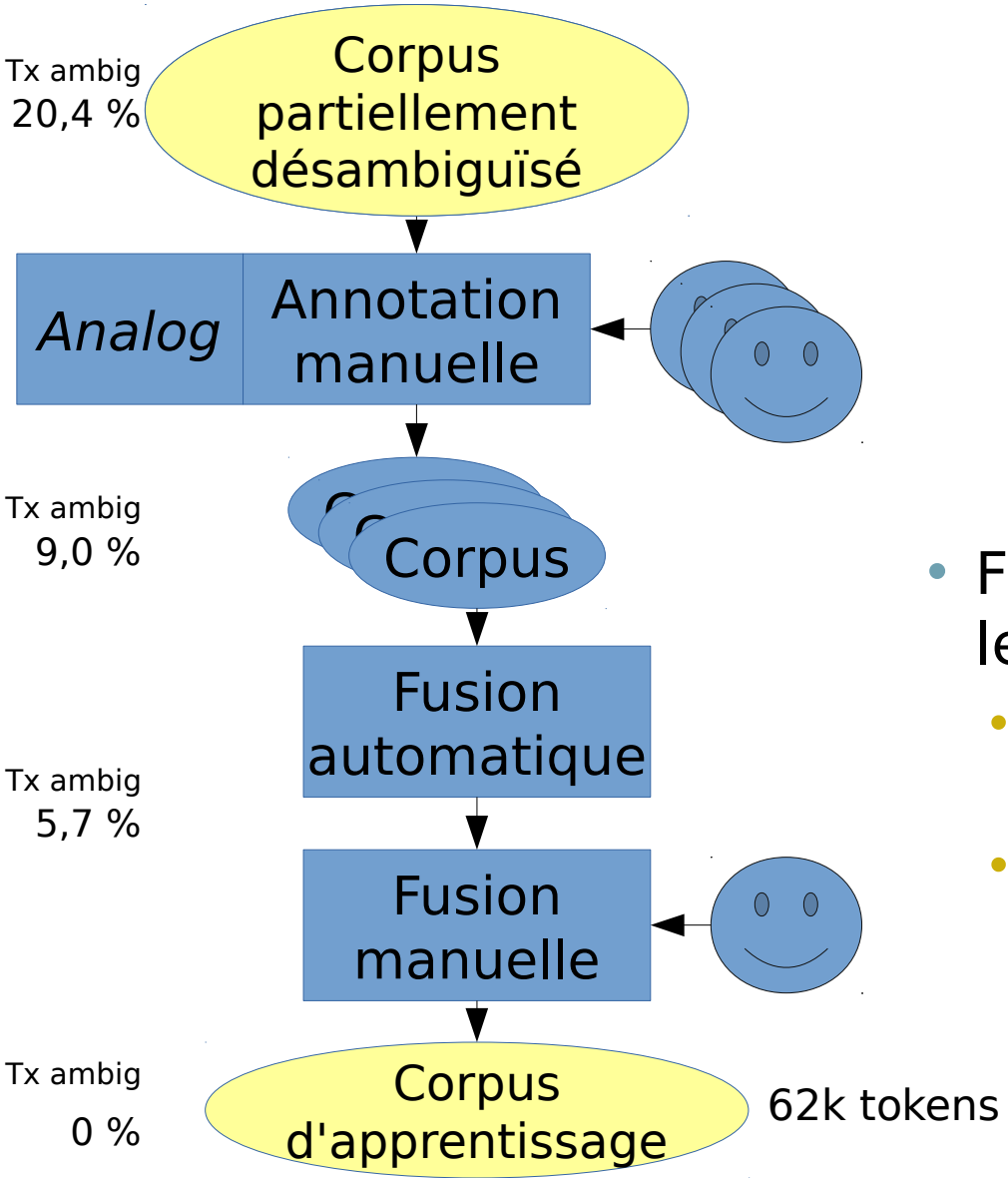
Désambiguïsation



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

quelques	QUELQUE:ADJ|QUELQUE:DET
remarques	REMARQUE:NOM
sur	SUR:PRE
les	LE:DET
groupements	GROUPEMENT:NOM

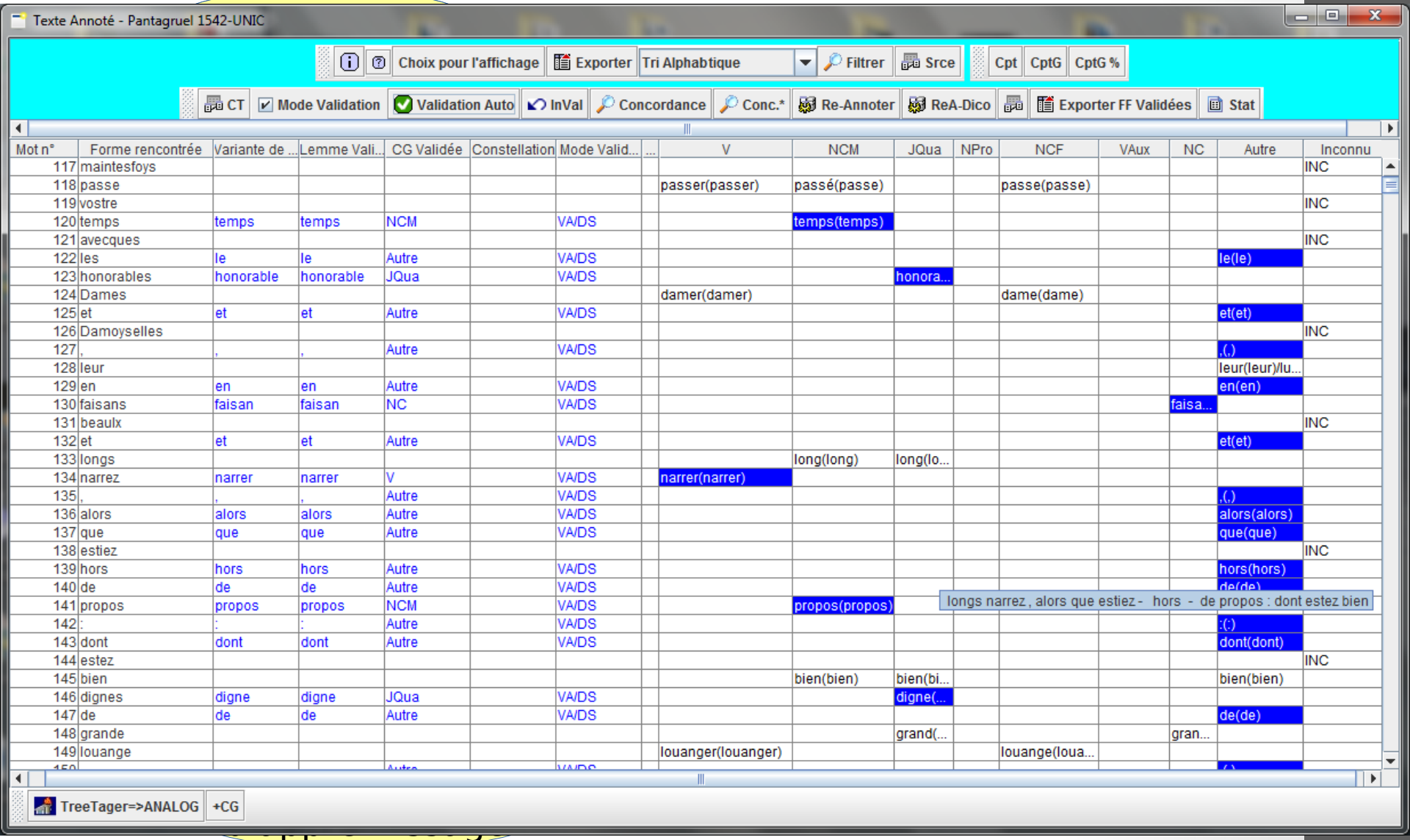
Annotation manuelle et fusion



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

2. Création d'un corpus de référence

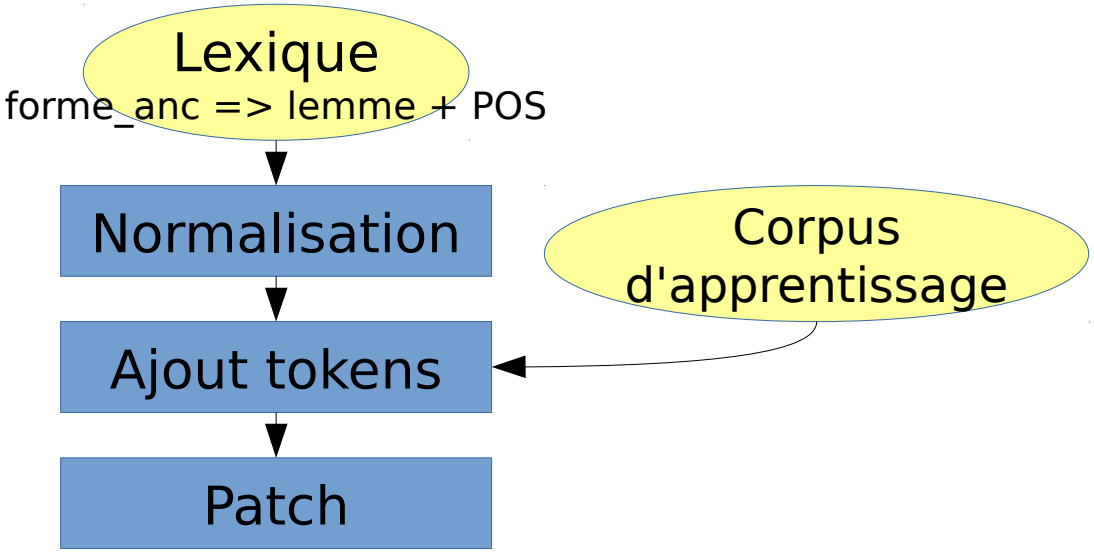
Analog (MH Lay, Poitiers)



Chaîne de traitement

Création d'un modèle de langue pour le français classique

Préparation du lexique



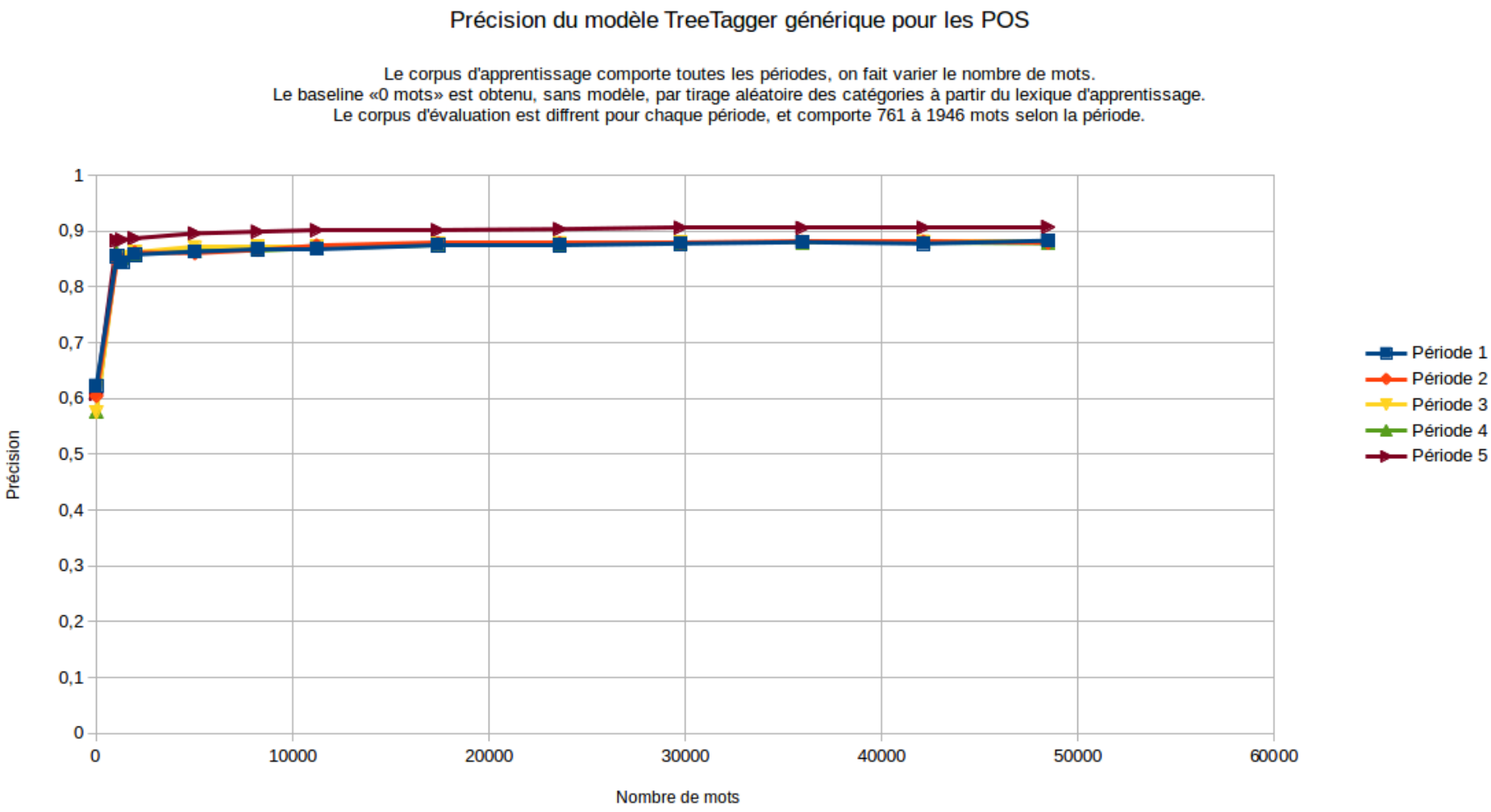
- Patch :
  - Listes de tokens
    - À ajouter
    - À enlever
  - Règles ad hoc

Création du modèle

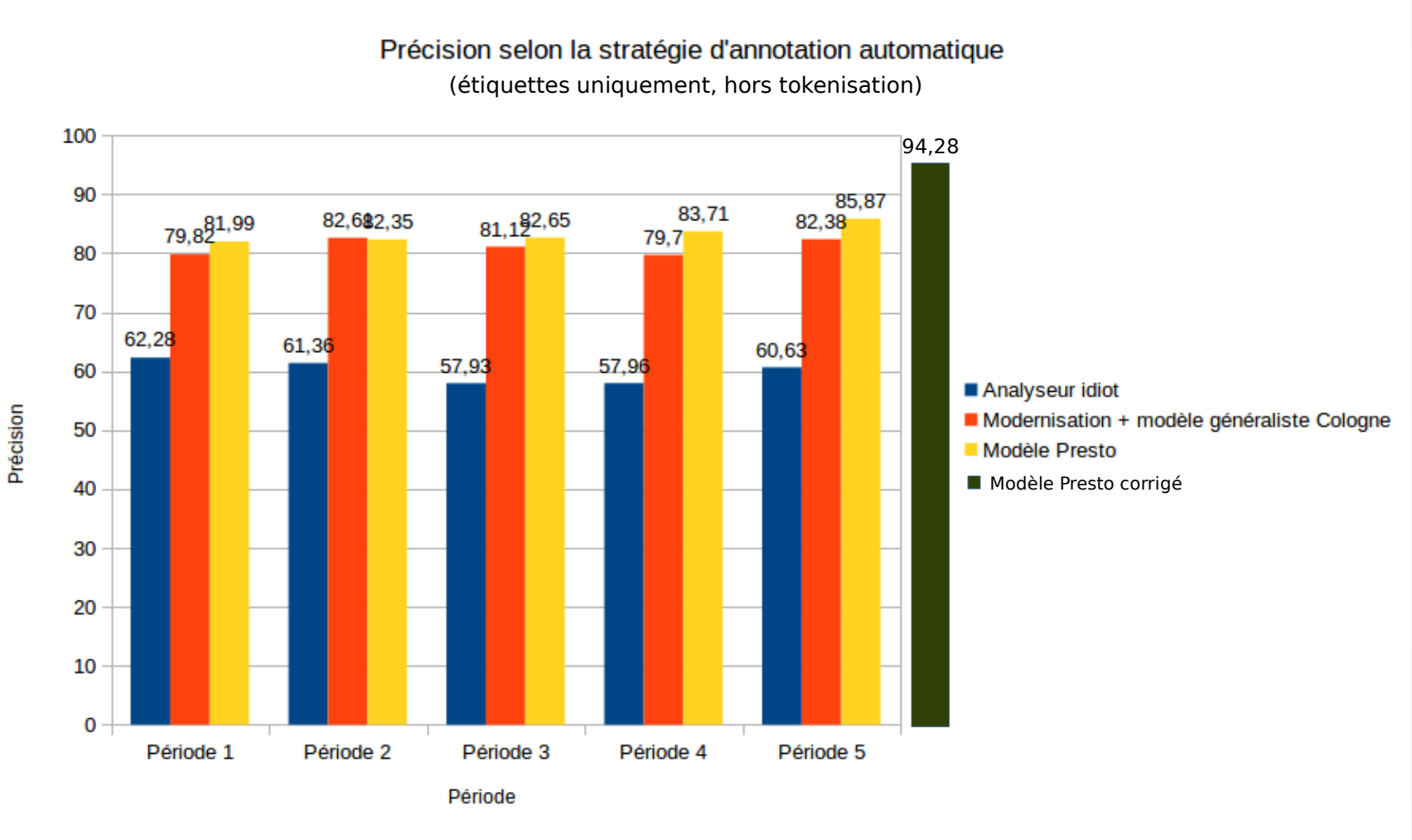
- Division du corpus d'apprentissage en trois
  - Corpus d'entraînement (80% − 49 630 tokens)
  - Corpus de développement (10% − 6 164 tokens)
  - Corpus de référence (10% − 6 110 tokens)
- *Autotuning* pour trouver les meilleurs paramètres pour TreeTagger
  - cl 2 ; dtg 0,5 ; sw 1 ; ecw 0,06 ; atg 1,15
  - Précision +0,05 %
- Précision
  - Corpus d'entraînement : 95,77 %
  - Corpus de développement : 94,28 %
  - Corpus de référence : 94,46 %

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

Évaluation du modèle



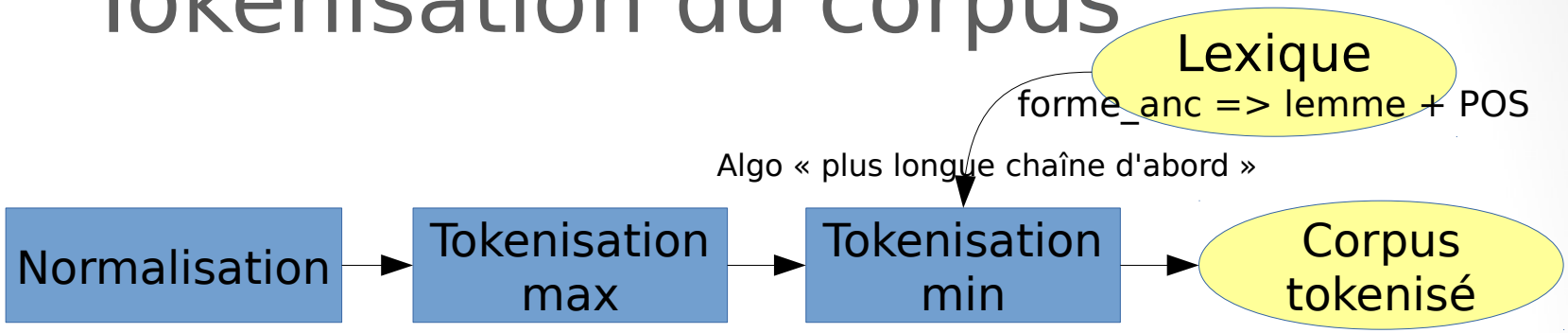
Évaluation du modèle



Chaîne de traitement

Analyse du corpus

Tokenisation du corpus



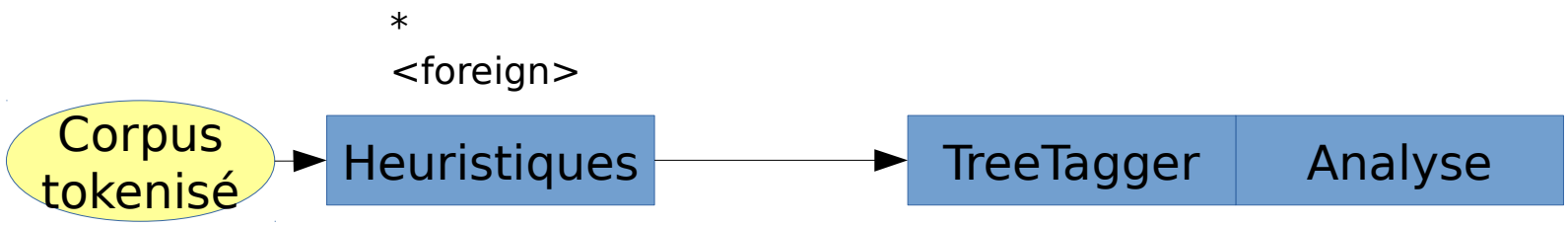
Tant que l'on dye par honneurs: Vive L'excellente *Bazoche!

Tant
que
l
'
on
dye
par
honneurs
:
Vive
L
'
excellente
*Bazoche
!

Tant que l'
Tant que l
Tant que
Tant
...
L'on dye
L'
on
...
L'excellente
L'
...

Tant
que
l'
on
dye
par
honneurs
:
Vive
L'
excellente
*Bazoche
!

Traitement du corpus



Tant	Tant	Tant	Rg	TANT
que	que	que	Cs	QUE
l'	l'	l'	Xi	L
on	on	on	Pp	ON
dye	dye	dye	Vvc	<unknown>
par	par	par	S	PAR
honneurs	honneurs	honneurs	Nc	HONNEUR
:	:	:	Fw	:
Vive	Vive	Vive	Vvc	VIVRE
l'	l'	l'	Da	LE
excellente	excellente	excellente	Ag	EXCELLENT
*Bazoche	Bazoche Np	Bazoche	Np	<unknown>
!	!	!	Fs	!

Implémentation

- Programmation
  - Format de données tabulaire
  - Commandes Unix (« *Unix for Poets* »), Bash, Perl
  - Chaîne de montage : *Pipes* Unix
    - ~multitâche
    - Souple, facile à déboguer
- Vitesse (création du modèle, màj lexique, tokenisation, analyse)
  - Pour un lexique de 2,7 M tokens
  - Pour un corpus de 28,3 M tokens
  - Processeur Xeon 4*3,2 Ghz
  - => 10 minutes

Perspectives

- Pêche aux erreurs
  - Vérification des ressources

Quelques problèmes épineux

Jeu d'étiquettes : La catégorie G (participe, adjectifs, gérondifs)

Ga : Participes présents, adjectifs verbaux, gérondifs

Ge : participes passés, adjectifs

- **En synchronie** : cas pb -> tests (Riegel & al. 2009 : 737-738) -> chances de divergences entre annotateurs. En outre :
  - *Selon les cas (le type de verbe, le contexte), ils [les participes] sont sentis comme plus ou moins «verbaux» ou «adjectivaux» (avec une marge appréciable de liberté d'interprétation)* (P. le Goffic 1993, § 134: 201)
- **Dimension diachronique** :
  - *En français classique, la tripartition des formes en –ant ne va pas de soi (...) le gérondif, invariable, se distingue mal du participe (au masculin singulier) du fait qu'il n'est pas régulièrement précédé de en; le participe qui peut être variable en genre et en nombre, se distingue mal de l'adjectif verbal.* (N. Fournier, 2002, § 421: 291-292)

Les amalgames

- **en+le > enl, el, on, ou, au, hou, hu**
  - speciallement que pour toy estre liee **ou** lien de mariage (PRESTO_1535, La déplourable fin de Flamete, J. de Flores)
  - ainsi que escript cesar en ses commentaires, & jean de gravot **on** mythologies gallicques. (PRESTO_1542_Pantagruel_Rabelais)
- ***en+les > ens, ans, eins, ons, es, ès, eus, aus, aux***
  - mais par ce moyen de propagation seminale demoure **es** enfans ce que estoit de perdu **es** parens (PRESTO_1542_Pantagruel_Rabelais)
  - la quele il faudroit chercher en ces vieux grecz, & latins, non point **és** aucteurs francoys. (PRESTO_1549_Deffence_J. du Bellay)

  Importance pour les calculs de fréquence, de cooccurrence lexicale ...