ENS DE LYON

icar
UMR 5191 - CNRS

Annotation du corpus PRESTO: création de ressources pour l'analyse du français classique

Sascha Diwersy, Universität zu Köln
Achille Falaise, ENS de Lyon, ICAR

Journées d'étude *Presto*
Cologne, 12-13 février 2015

Plan de la présentation

1. Le corpus
2. Chaîne de traitement
   1. Stratégie de traitement
   2. Création d'un corpus d'apprentissage
   3. Création d'un modèle de langage
   4. Analyse du corpus
3. Exploitation du corpus
   1. Principes d'échantillonnage
   2. Modèles de données et méthodes
   3. Outils (démo PrimeStat)
   4. Perspectives – vers une linguistique diachronique outillée

Ressources

Le corpus Presto

- Constitution d'un corpus dans le cadre de PRESTO :
  - Prérequis pour la réalisation du projet
  - Apport majeur du projet
- **Besoins** :
  - Représentation de toutes les périodes de l'histoire du français : 9ème s. au 20ème s.
  - Présence de différents types de textes et de différents genres discursifs
  - Enrichissement linguistique : étiquetage morpho-syntaxique et lemmatisation

Le corpus Presto

- **Objectifs** :
  - Disposer d'annotations linguistiques fiables mais également de gros volumes de textes
  - Pouvoir rendre disponibles autant que possible le corpus constitué et les outils élaborés tout en respectant les contraintes juridiques
- **Contrainte** :
  - Disposer de versions numérisées de textes de bonne qualité

→Collaboration avec diverses bases textuelles existantes :

- Frantext
- BVH
- Cologne
- ARTFL
- CEPM
- …

Le corpus Presto

- **Objectifs** :
  - Disposer d'annotations linguistiques fiables mais également de gros volumes de textes
  - Pouvoir rendre disponibles autant que possible le corpus constitué et les outils élaborés tout en respectant les contraintes juridiques
- **Contrainte** :
  - Disposer de versions numérisées de textes de bonne qualité

→Collaboration avec diverses bases textuelles existantes :

- Frantext
- BVH
- Cologne
- ARTFL
- CEPM
- …

392 textes :

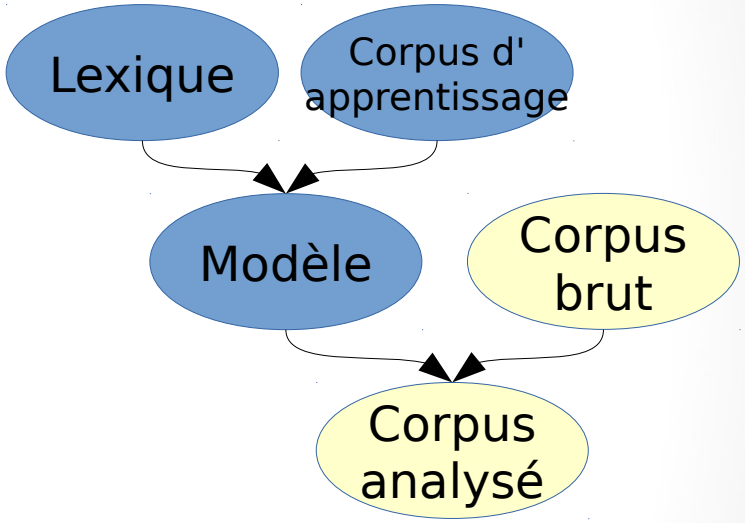
La Concorde du genre humain (1509)… La carte et le territoire (2010)

Gargantua, L'Astrée, l'Encyclopédie (tome 7)…

Chaîne 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

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

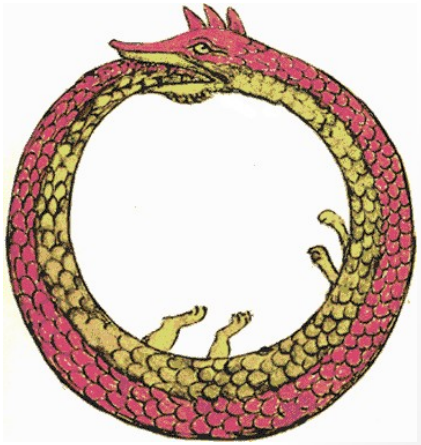
- Objectifs
  - 50k mots annotés manuellement (token, POS, lemme)
  - Par au moins 2 annotateurs
- Défis
  - C'est très chronophage !
- Méthode
  - Pré-annotation automatique du corpus

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 !

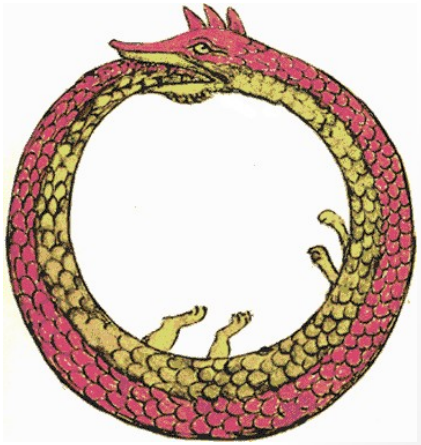


Objectifs et défis

- 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.



Choix des textes

- Sélection de 5 textes
  - 5 périodes
  - 5 genres

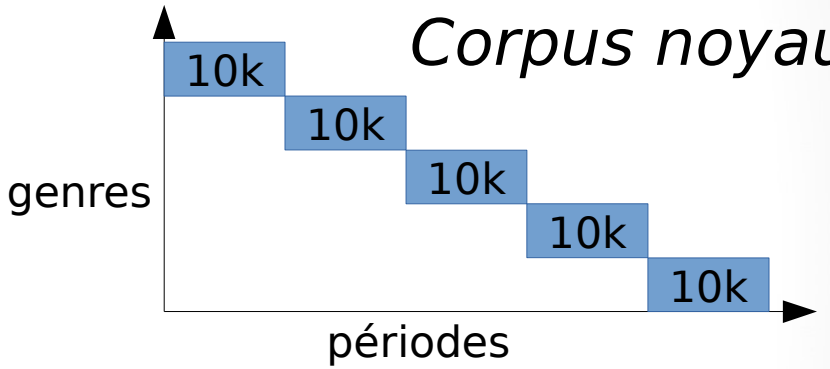
Saulsaye (1547)

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

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>

Lexique

forme_anc => lemme + 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

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

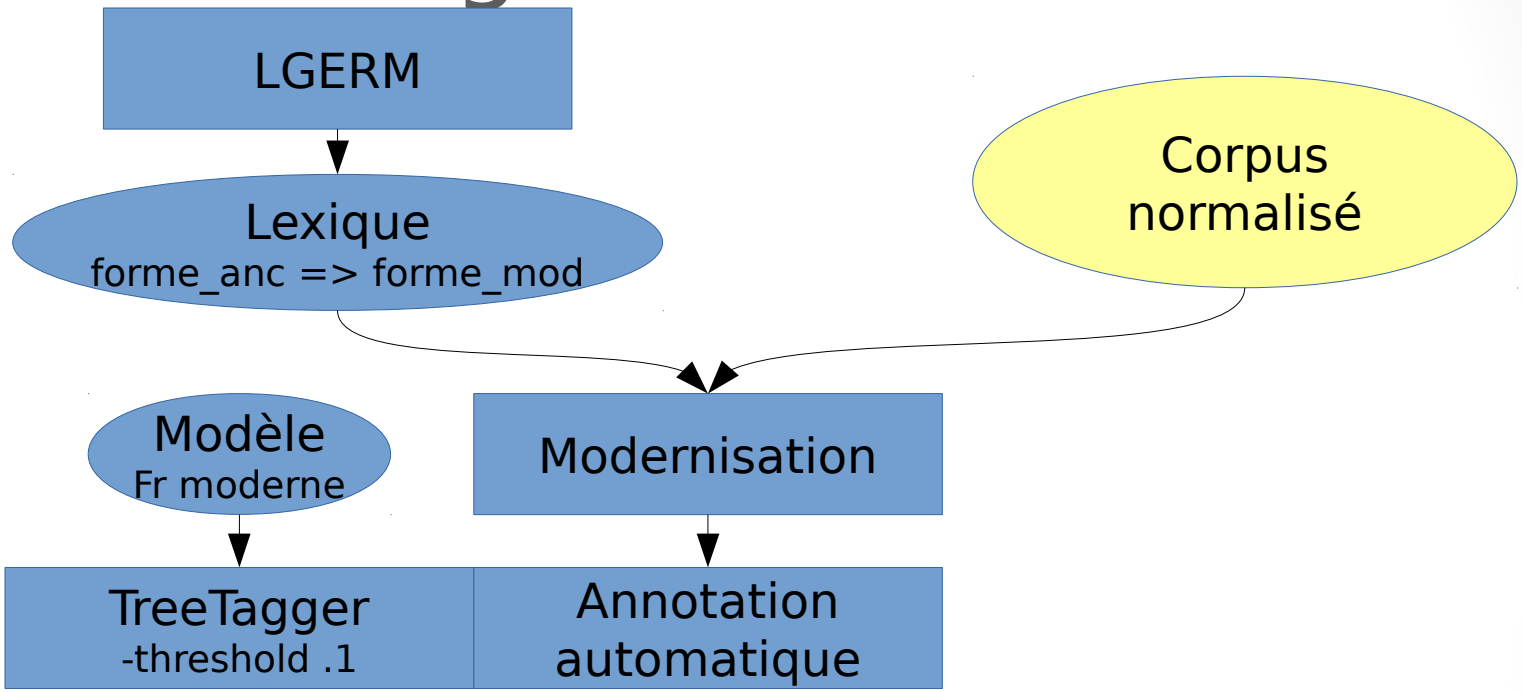
Presto_0

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

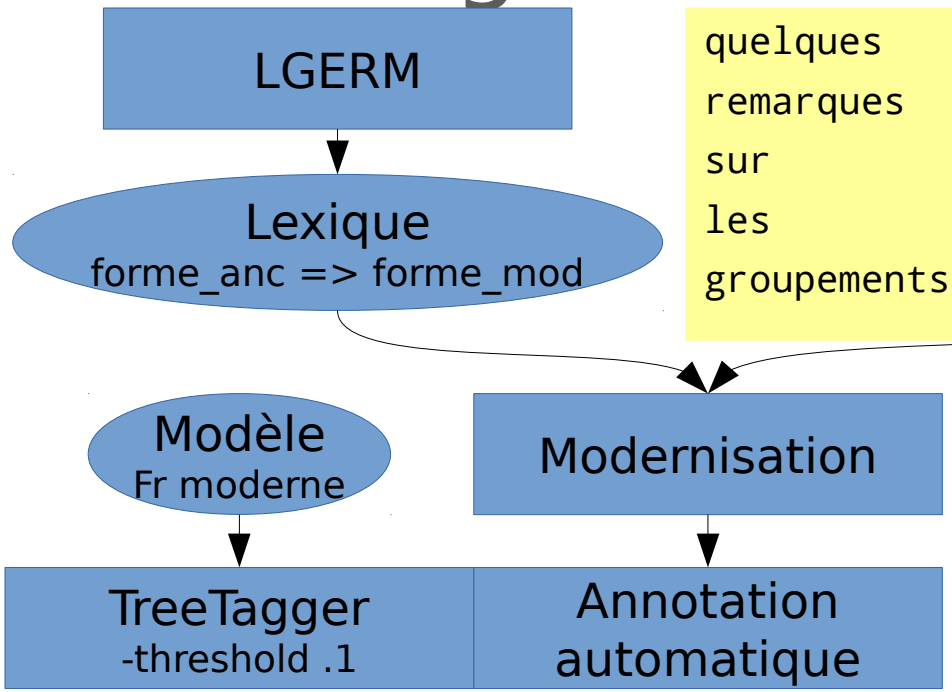
Presto_1

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

Désambiguïsation

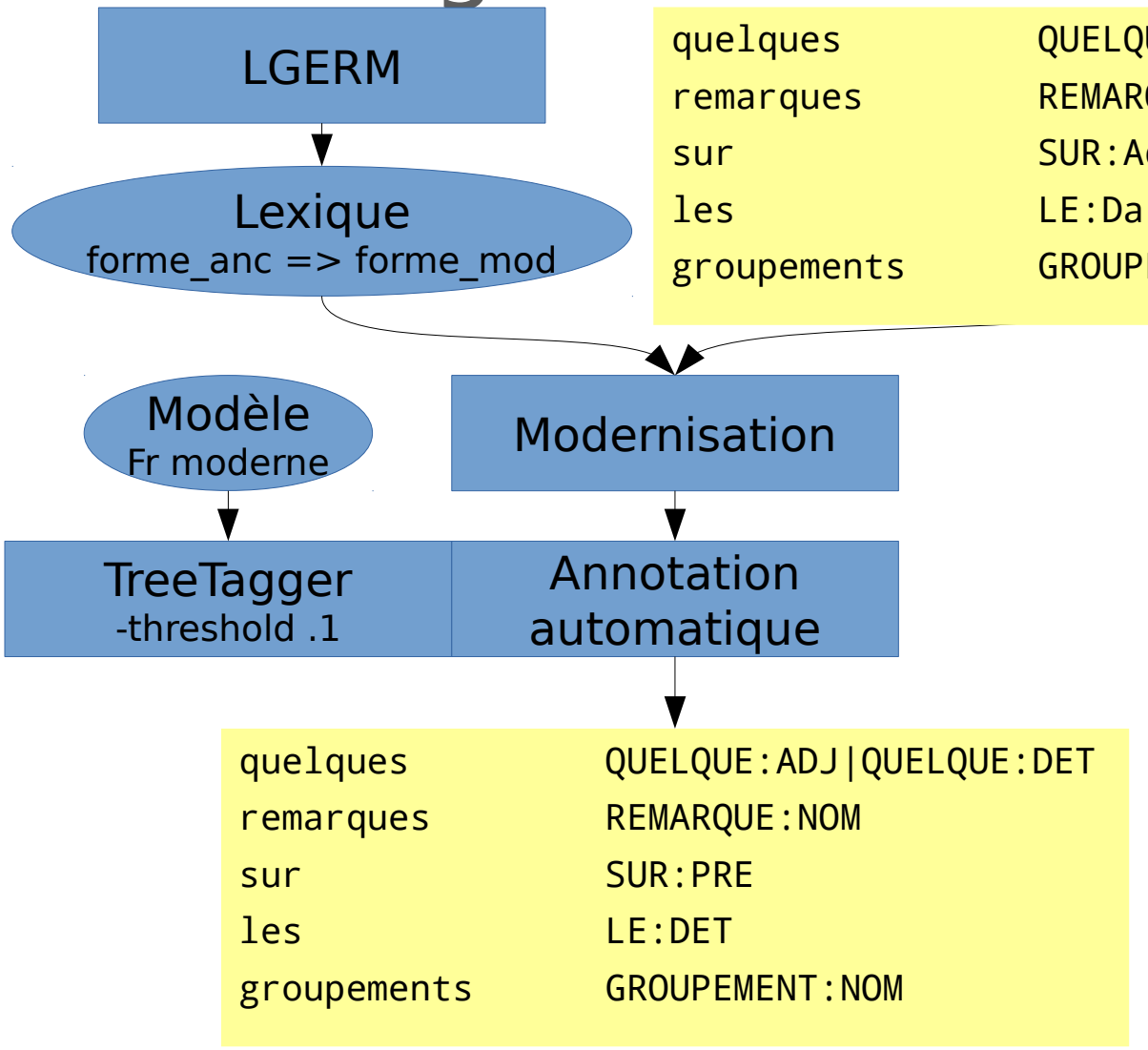


Désambiguïsation



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

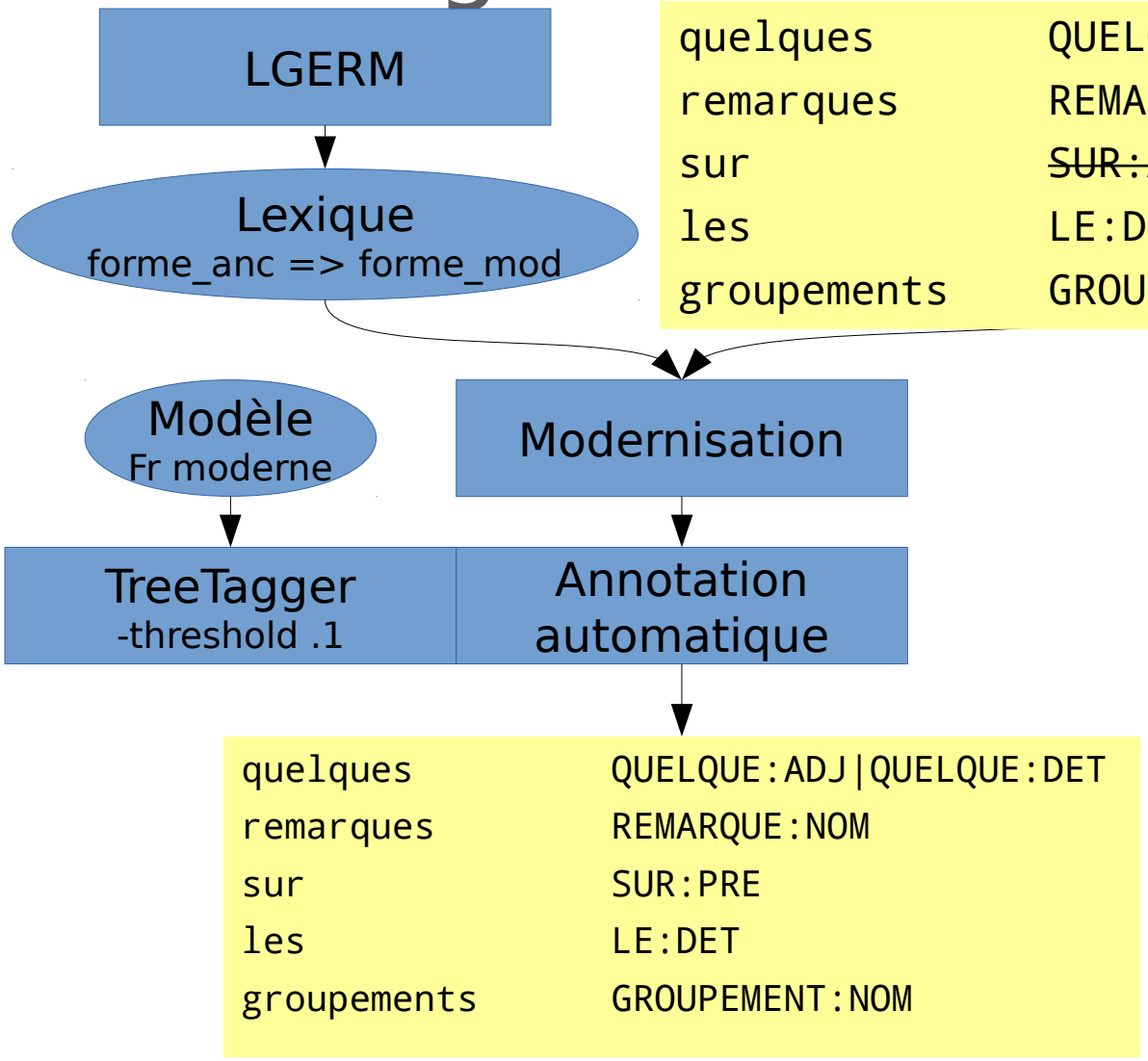
Désambiguïsation



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

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

Désambiguïsation



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

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

Simplification des étiquettes

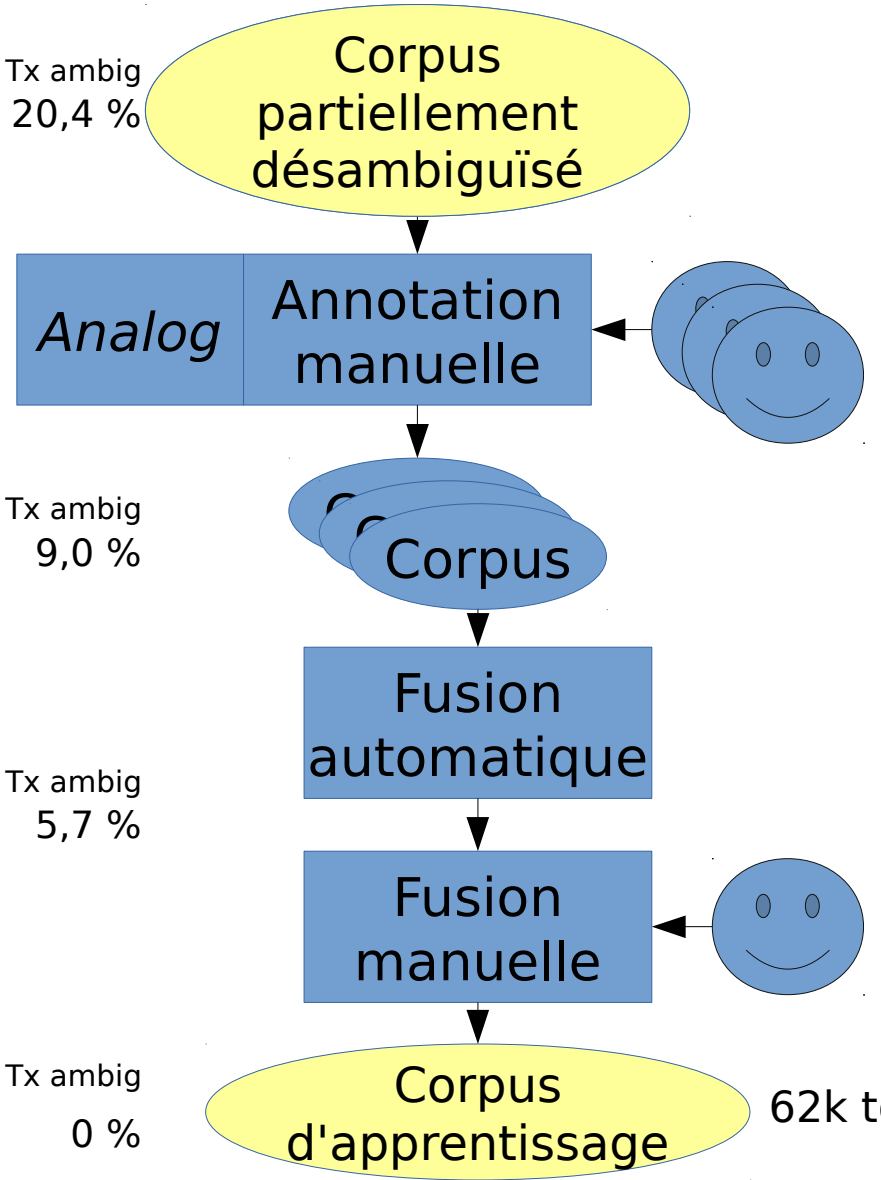
quelques	QUELQUE:AQ0CP0|QUELQUE:DI0CP0
remarques	REMARQUE:NCFP000|REMARQUER:VMIP2S0| REMARQUER:VMP00PM|REMARQUER:VMSP2S0
sur	SUR:AQ0CS0|SUR:SPS00|SÛR:AQ0MS0|SÛR:RG
les	LE:DA0CP0|LES:PP3CPA00|LÈS:SPS00|LÉ:NCMP000
groupements	GROUPEMENT:NCMP000

Presto_0

Presto_1

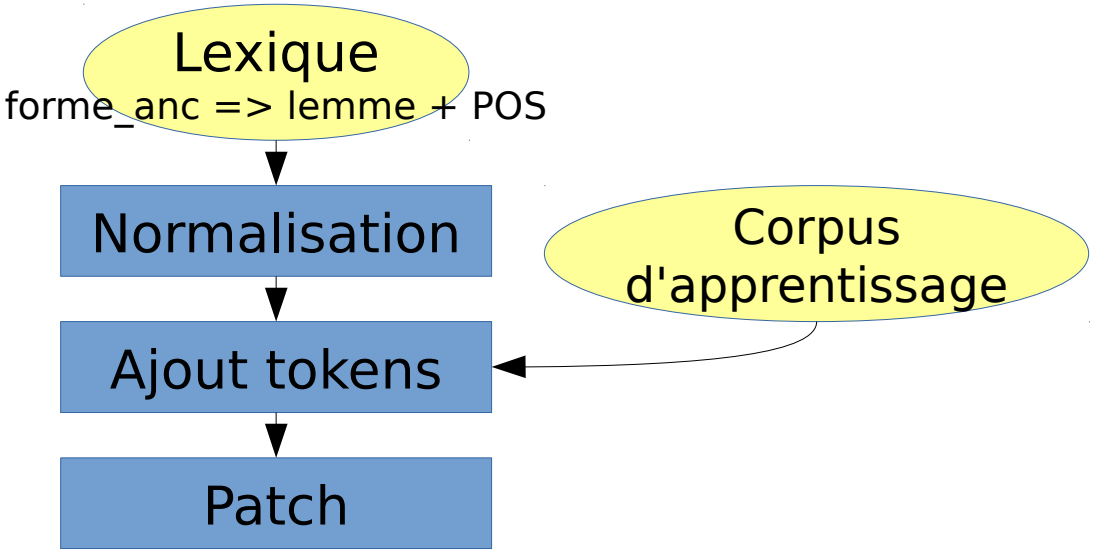
quelques	QUELQUE:Aq|QUELQUE:Di
remarques	REMARQUE:Nc|REMARQUER:Vvc
sur	SUR:Aq|SUR:Sp|SÛR:Aq|SÛR:R
les	LE:Da|LES:Pp|LÈS:Sp|LÉ:Nc
groupements	GROUPEMENT:Nc

Annotation manuelle et fusion



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

Préparation du lexique



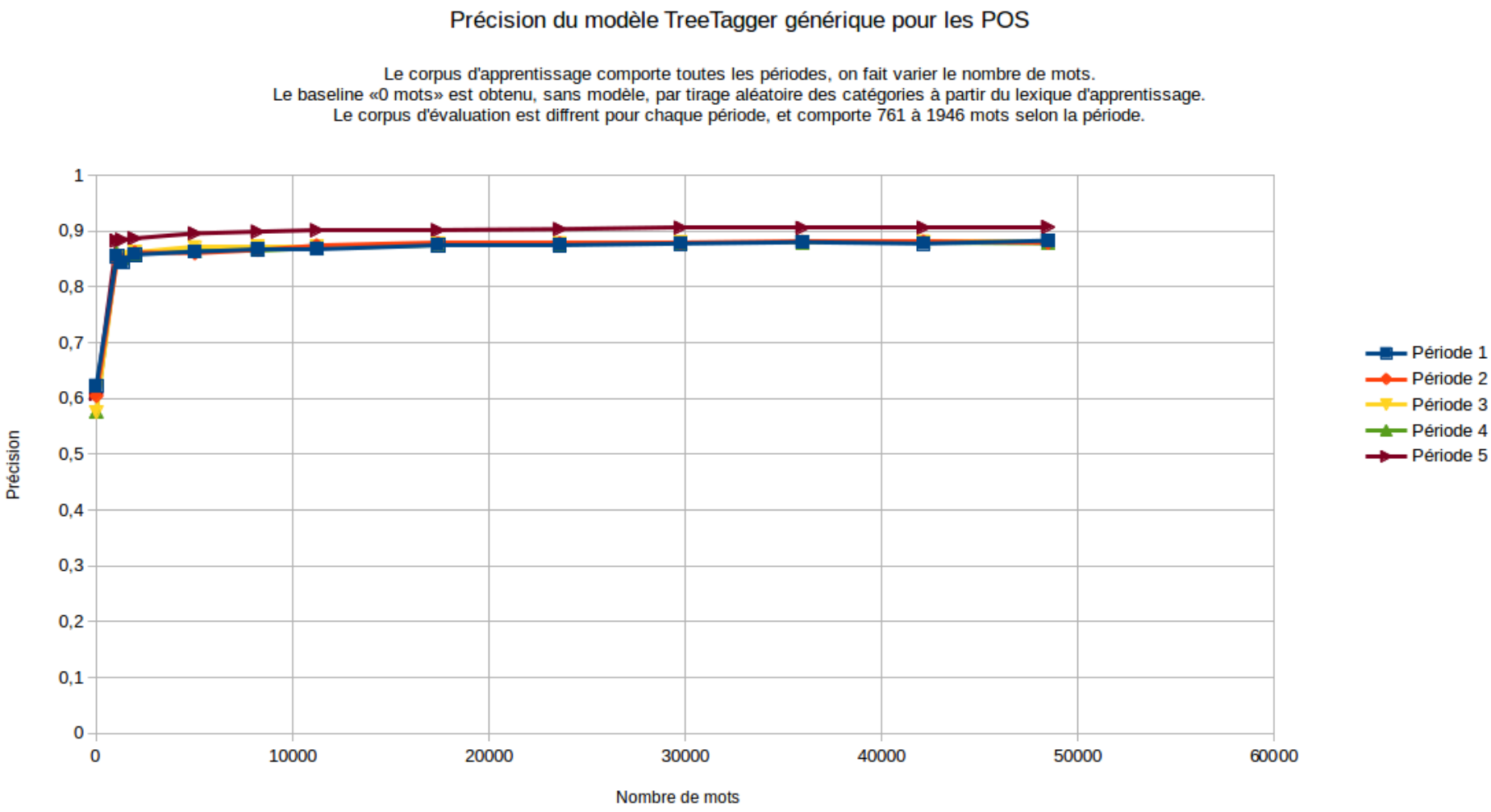
- Patch :
  - Listes de tokens
    - À ajouter
    - À enlever
  - Règles ad hoc

Création du modèle

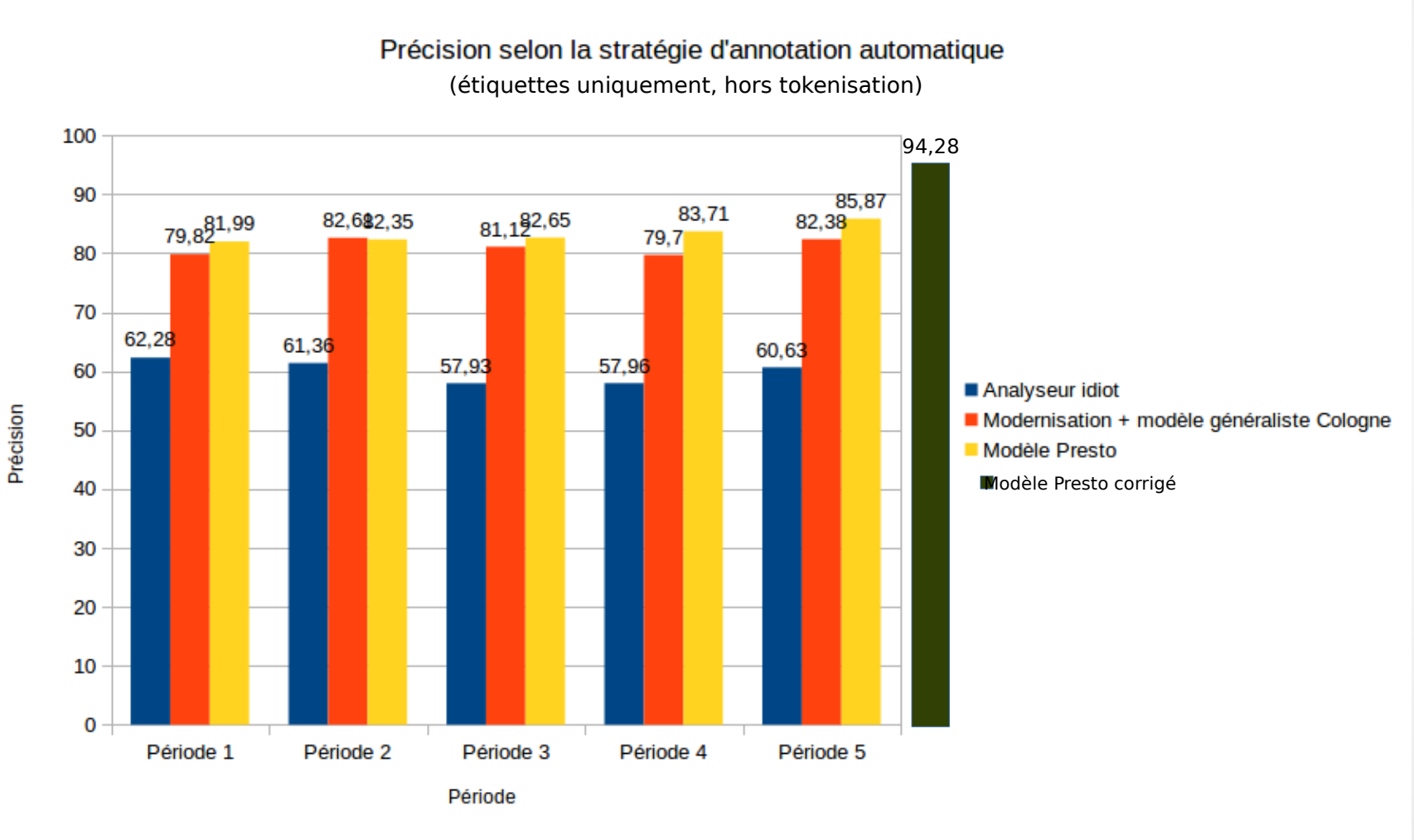
- Division du corpus d'apprentissage en trois
  - Corpus d'entraînement (80% − 49 630 tokens)
  - Corpus de développement (10% − 6 164 tokens)
  - Corpus de référence (10% − 6 110 tokens)
- *Autotuning* pour trouver les meilleurs paramètres pour TreeTagger
  - cl 2 ; dtg 0,5 ; sw 1 ; ecw 0,06 ; atg 1,15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

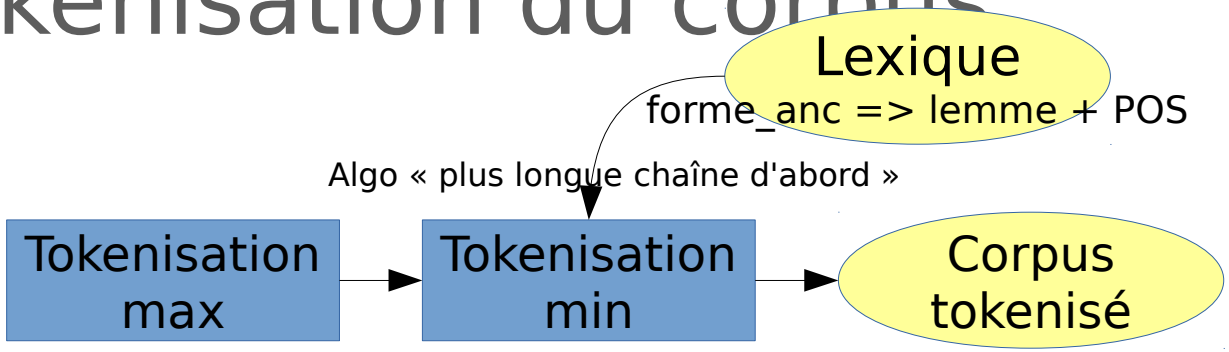
Évaluation du modèle



Évaluation du modèle



Tokenisation du corpus



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

*
<foreign>

Tant	Tant		Tant	Rg	TANT
que	que		que	Cs	QUE
l'on	l'on		l'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
  - Chaîne de montage : *Pipes* Unix
    - ~multitâche
    - Souple, facile à déboguer
  - Shell, Perl
- 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

Perspectives

- Pêche aux erreurs
  - Vérification des ressources
- Annotation manuelle ciblée
- Règles de segmentation
  - *Trèsgrand, trèsbel*
  - *Par ce que*
- Détection de la langue
- Désambiguïsation sémantique pour les lemmes
  - 1,8 % des lemmes du corpus sont ambigus
  - *fil*|*fils*
  - *congre*|*congrès*
- Analyse syntaxique

ENS DE LYON

icar
UMR 5191 - CNRS

Annotation du corpus PRESTO: création de ressources pour l'analyse du français classique

Sascha Diwersy, Universität zu Köln
Achille Falaise, ENS de Lyon, ICAR

Journées d'étude *Presto*
Cologne, 12-13 février 2015