Les corpus annotés du français, TALN 2017
lundi 26 juin 2017, Orléans

Presto, un corpus diachronique pour le français des XVI^{e}-XX^{e} siècles

http://presto.ens-lyon.fr

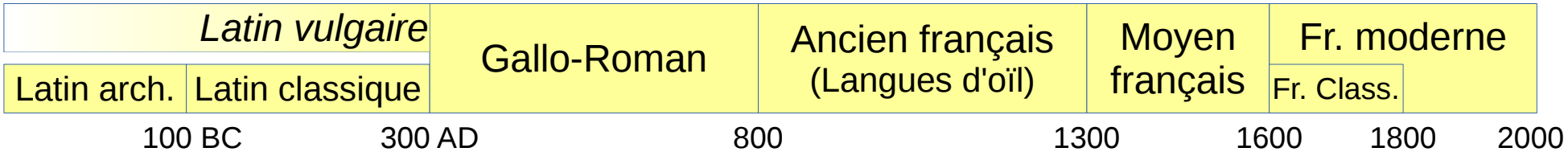
Peter Blumenthal, U. Cologne
Sascha Diwersy, U. Montpellier, Praxiling
Achille Falaise, CNRS, ICAR
Marie-Hélène Lay, FoReLL, U. Poitiers
Gilles Souvay, U. Lorraine, ATILF
Denis Vigier, U. Lyon 2, ICAR

Plan

1) Le français des XVIe-XVIIIe siècles
2) Constitution du corpus
3) Création de ressources
    1) Lexique
    2) Corpus d'apprentissage
    3) Modèle de langage
4) Annotation du corpus

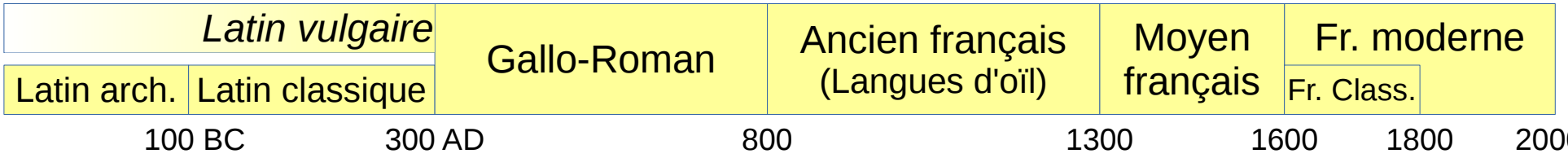
1. Le français des XVIe-XVIIIe siècles

Histoire du français



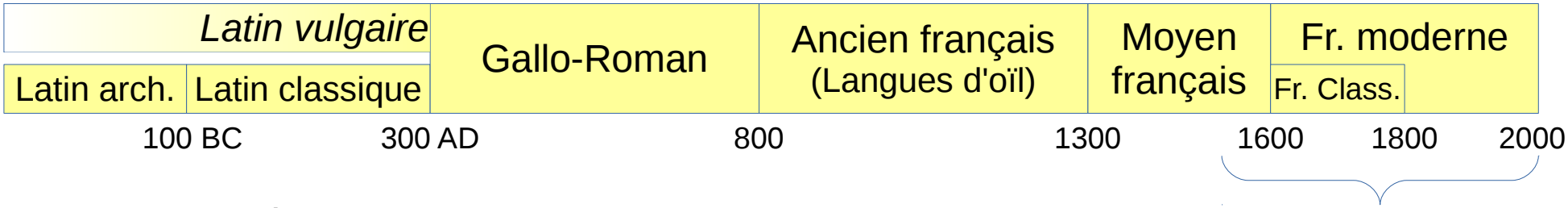
1. Le français des XVIe-XVIIIe siècles

Histoire du français



Projet Presto :
évolution du système prépositionnel du français XVIe-XXe siècles

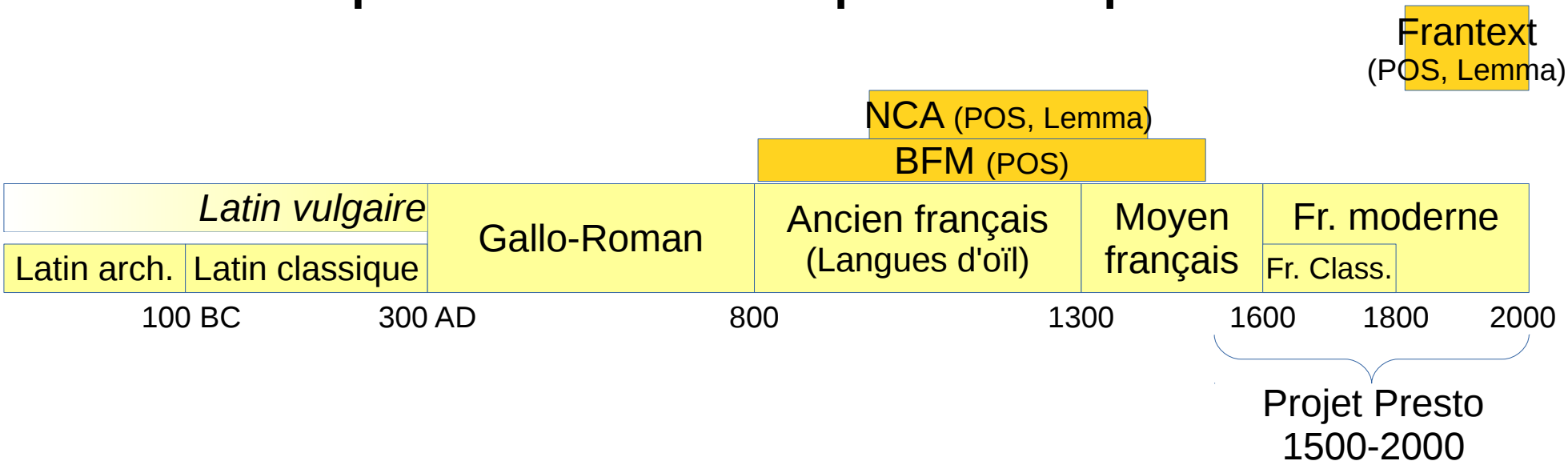
1500-1800 : l'émergence du français en temps que langue nationale standardisée



- 1440 : presse à imprimer
- 1532 : *Pantagruel*, François Rabelais
- 1539 : Ordonnace de Villiers-Cotterêts − Français langue officielle
- 1543 : *Traité des reliques*, Jean Calvin
- 1549 : *La Deffence et Illustration de la Langue Francoyse*, Joachim du Bellay
- 1550 : *Briefve collection de l'administration anatomique*, Ambroise Paré
- 1562 : *Gramere*, Pierre de la Ramée
- 1572-1592 : *Essais*, Montaigne
- 1634 : Académie française − standardisation officielle de la langue
- 1751-1772 : *Encyclopédie*, Diderot & d'Alembert
- 1795 : École Normale de l'an III − standardisation de la langue de l'enseignement

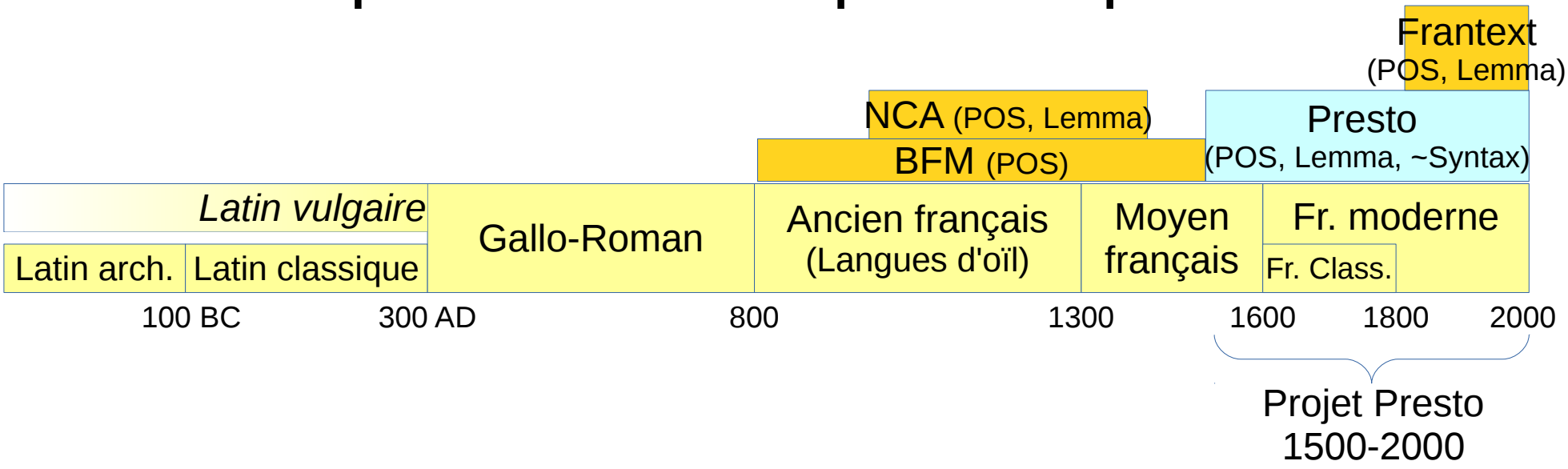
1. Le français des XVIe-XVIIIe siècles

Corpus diachroniques étiquetés



1. Le français des XVIe-XVIIIe siècles

Corpus diachroniques étiquetés



Le français du XVIe siècle

SI LA NATURE (dont quelque Person- de grand'renommée non sans rayson a douté, si on la devoit appeller Mere, ou Maratre) eust donné aux Hommes un commun vouloir, & consentement, outre les innumerables commoditez, qui en feussent procedées, l'Inconstance humaine, n'eust eu besoing de se forger tant de manieres de parler. Laquéle diversité, & confusion, se peut à bon droict appeller la Tour de Babel.

Début de *La deffence, et illustration de la langue francoyse*, Joachim du Bellay, 1549.

Le français du XVIIIe siècle

LANGAGE, s. m. (Arts. Raisonn. Philos. Metaphys.)

modus & usus loquendi, **maniere** dont les hommes se communiquent leurs pensées, par une suite de paroles, de gestes & d'expressions adaptées à leur génie, leurs **moeurs** & leurs climats.

Dès que l'homme se sentit entraîné par goût, par besoin & par plaisir à l'union de ses semblables, il lui **étoit** nécessaire de développer son **ame** à un autre, & lui en communiquer les situations. Après avoir essayé plusieurs sortes d'expressions, il s'en tint à la plus naturelle, la plus utile & la plus étendue, celle de l'organe de la voix. Il **étoit aise** d'en faire usage en toute occasion, à chaque instant, & sans autre peine que celle de se donner des **mouvemens** de respiration, si doux à l'existence.

Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, Diderot & d'Alembert (dir.), 1751-1765

2. Constitution du corpus

Le corpus Presto

- Besoins
  - Représentation du français des XVI^{e} s. au XX^{e} s.
  - Présence de différents types de textes et de différents genres discursifs : narratif, poésie, théâtre, traité
  - Enrichissement linguistique : étiquetage morpho-syntaxique et lemmatisation
- Collaboration avec diverses bases textuelles existantes
  - *Frantext, CNRTL*
  - *BVH (Bibliothèques Virtuelles Humanistes)*
  - *Université de Cologne*
  - *ARTFL (American and French Research on the Treasury of the French Language)*
  - *CÉPM (Corpus Électroniques de la Première Modernité)*
  - …

2. Constitution du corpus

Le corpus Presto

- Besoins
  - Représentation du français des XVIe s. au XXe s.
  - Présence de différents types de textes et de différents genres discursifs : narratif, poésie, théâtre, traité
  - Enrichissement linguistique : étiquetage morpho-syntaxique et lemmatisation
- Collaboration avec diverses bases textuelles existantes
  - *Frantext, CNRTL*
  - *BVH (Bibliothèques Virtuelles Humanistes)*
  - *Université de Cologne*
  - *ARTFL (American and French Resea*
  - *CÉPM (Corpus Électroniques de la P*
  - ...

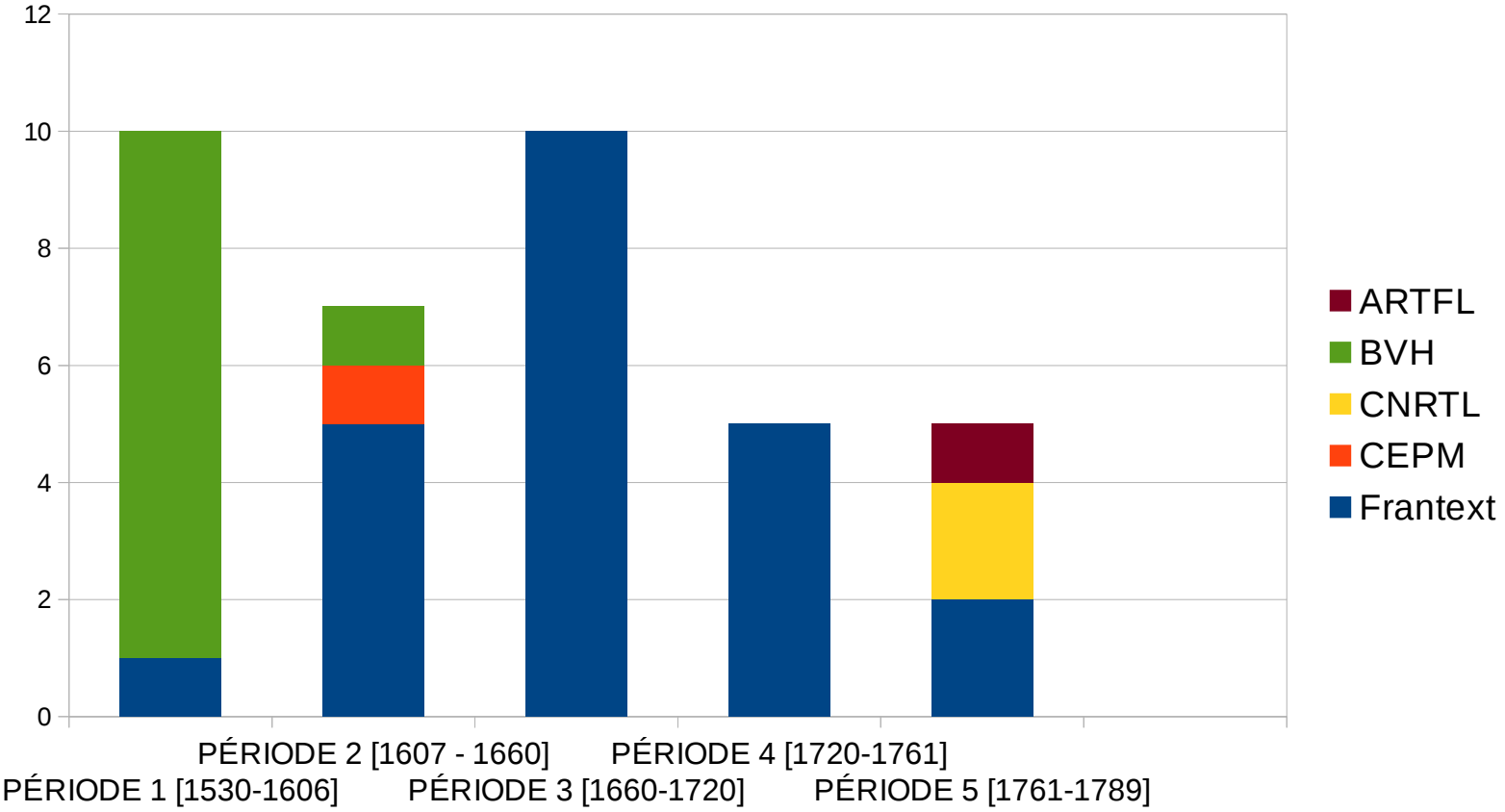
340 textes (1509-2010), dont 53 libres :

L'Astrée, Gargantua, l'Encyclopédie (vol. 7)...

2. Constitution du corpus

Le corpus Presto

- Besoins
  - Représentation du français des XVIe s. au XXe s.
  - Présence de différents types de textes et de différents genre
  - Enrich
    synta
- Collaboratio
  - *Frantext,*
  - *BVH (Bib*
  - *Universit*
  - *ARTFL (*
  - *CÉPM (C*
  - …



2. Constitution du corpus

Le corpus Presto

Corpus	Période	Genres	Taille	Licence
Noyau	XVIe-XXe s.	narratif, poésie, théâtre, et traité	53 textes, 6,8 M mots	CC 3.0 BY-SA-NC
Équilibré			162 textes, 5,4 M mots	Sous droits
Étendu	XVIe-XXIe s.		340 textes, 35,5 M mots	

19e − 21e siècles

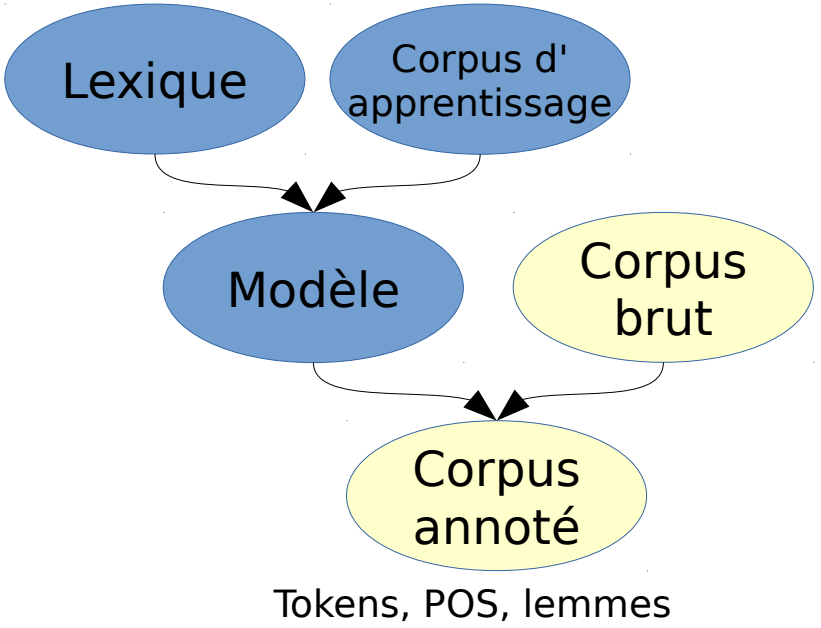
18e − 21e siècles

Jeu d'étiquettes (basé sur Multext/Eagles + Grace)

Étiquette niv 1	**Étiquette niv 2**	**Flexion**
Nom (N)	commun, propre	
Verbe (V)	être/avoir, autre	conjugué, infinitif
Adjectif (A)	général, possessif	
Pronom (P)	personnel, démonstratif, indéfini, possessif, interrogatif, relatif	
Déterminant (D)	article défini, démonstratif, article indéfini, article partitif, indéfini, relatif, interrogatif/exclamatif	
Participe-Adjectif-Gérondif (G)	part_présent/adjectif_verbal/gérondif, part_passé/adjectif_verbal	
Adverbe (R)	général, particule, interro-excalamatif	
Adposition (S)		
Conjonction (C)	coordination, subordination	
Numéral (M)	cardinal, ordinal	
Interjection (I)		
Résidu (X)	abréviation, mot étranger, symbole, préfixe, consonne intercalée	
Ponctuation (F)	forte, faible, autre	

Traitement : une approche classique

- Approche « classique » : chaîne de traitement



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

3a. Ressources – Lexique

Le lemmatiseur LGeRM

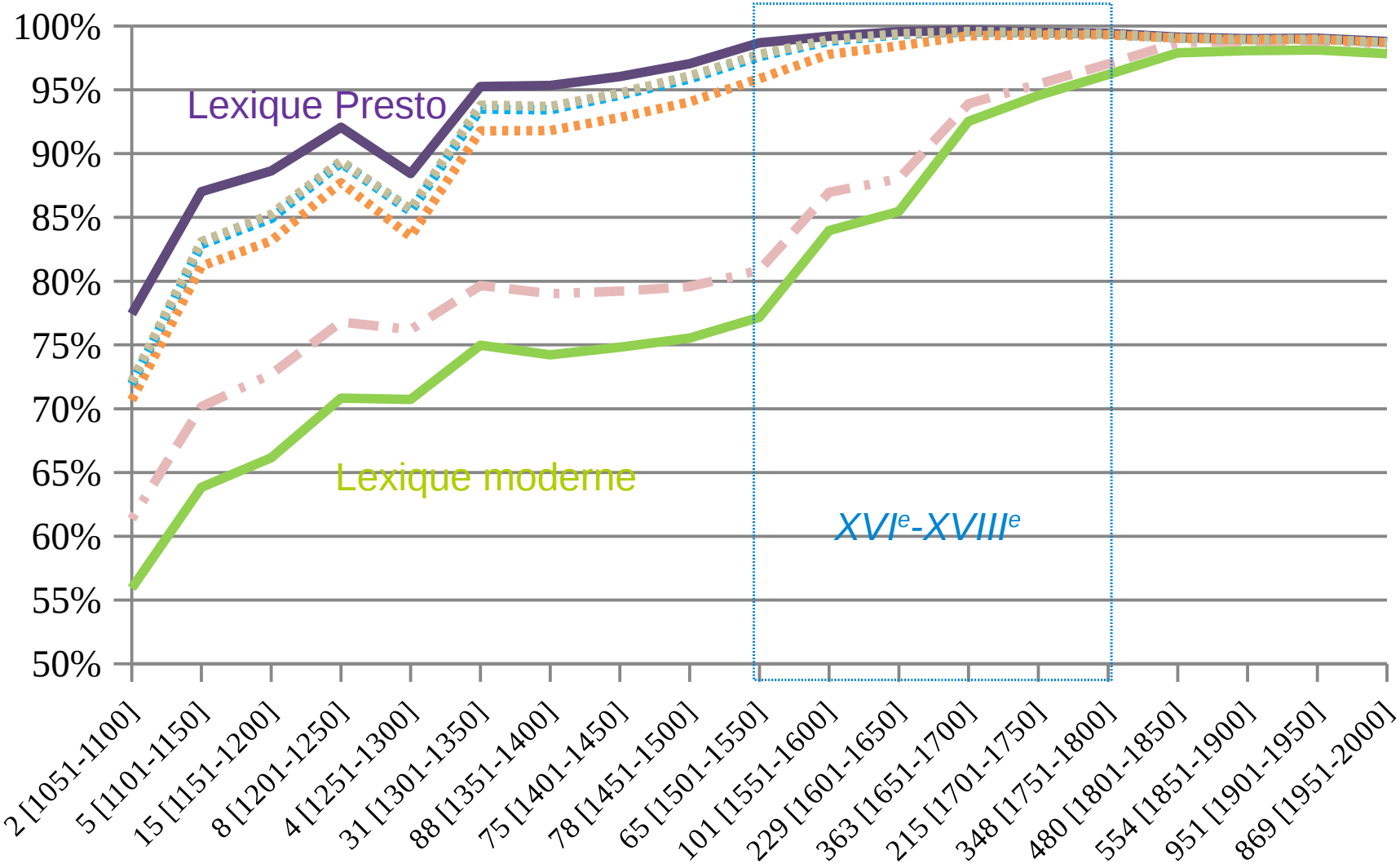
- LGeRM : Lemmes Graphies et Règles Morphologiques
  - lemmatiseur et environnement de lemmatisation
  - gestion de différents états de langues
  - http://www.atilf.fr/LGeRM
- Liste de formes connues : lexique morphologique
  - gestion de « flexion et variantes » dans Frantext
- Règles d'analyse des formes inconnues
  - 6 500 règles (4/5 flexion verbale)
    - si (en finale) alors ES → EFS — nes → nefs, NEF
    - Y → I — fayre → faire, FAIRE
    - si (entre voyelles) alors C → SS — mesfacent → mesfassent, MÉFAIRE

Construction du lexique

- Lexique moderne LEFFF
  - étiquettes spécifiques au projet
  - lemmes complémentaires issus du TLF/Morphalou et du DMF
- Construction
  - appliquer les règles d’archaïsation (3 boucles)
  - regarder les formes absentes
  - réitérer le processus
  - compléter avec les formes absentes
- Forme ancienne → forme moderne → POS + lemme

3a. Ressources – Lexique

Couverture lexicale



Couverture lexicale, mesurée sur le corpus *Frantext*, pour le lexique moderne (vert), les 3 itérations d'archaïsation (pointillés), et le lexique Presto final (violet).

Corpus d'apprentissage

- Sélection de 5 textes
  - 5 périodes
  - 5 genres

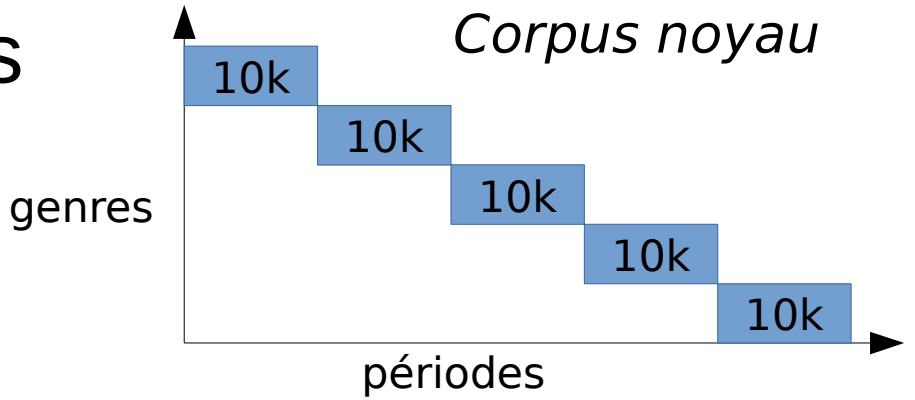
Saulsaye (1547)

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

Préannotation : 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

Préannotation : 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

3b. Ressources – Corpus d'apprentissage

Simplification des étiquettes

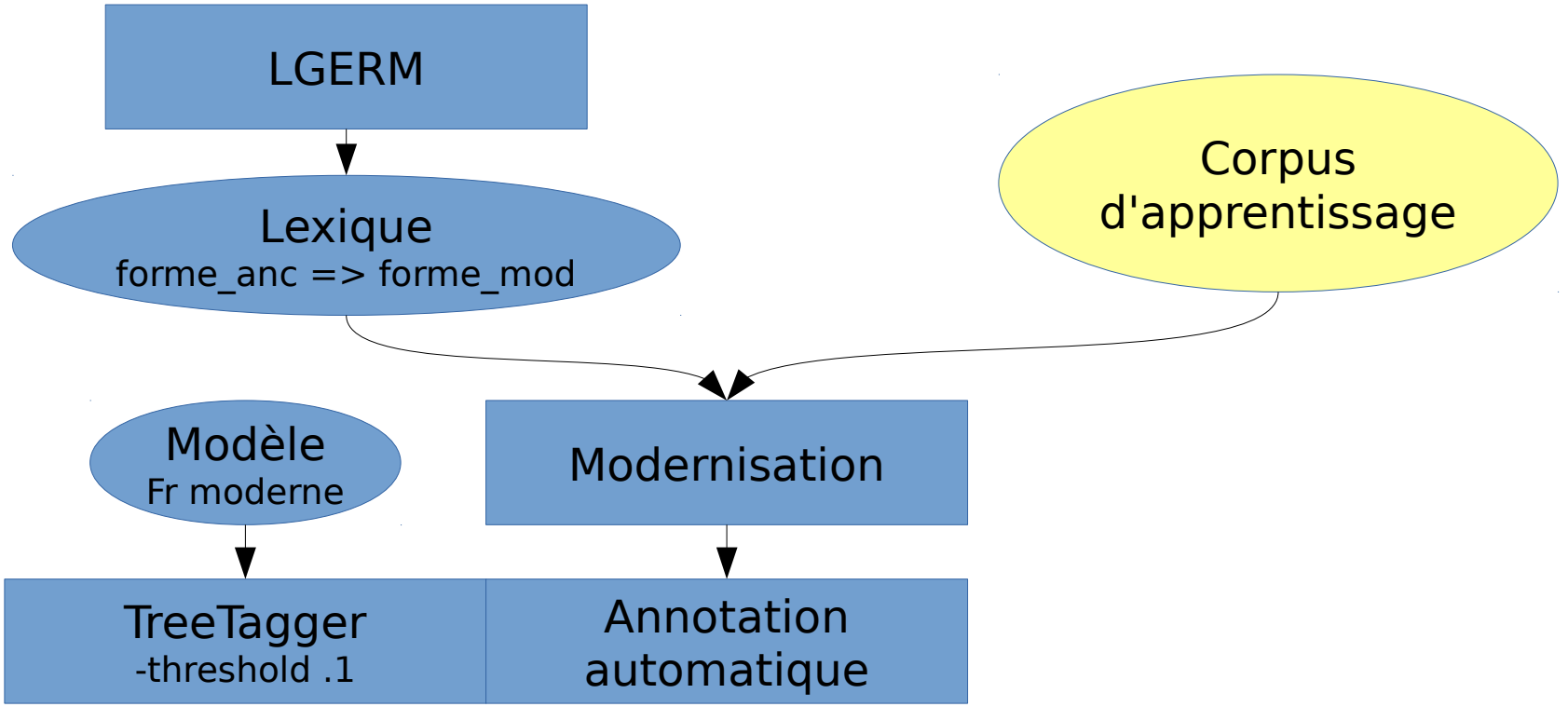
quelques	QUELQUE:AQ0CP0|QUELQUE:DI0CP0
remarques	REMARQUE:NCFP000|REMARQUER:VMIP2S0| REMARQUER:VMP00PM|REMARQUER:VMSP2S0
sur	SUR:AQ0CS0|SUR:SPS00|SÛR:AQ0MS0|SÛR:RG
les	LE:DA0CP0|LES:PP3CPA00|LÈS:SPS00|LÉ:NCMP000
groupements	GROUPEMENT:NCMP000

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

Étiquettes Presto

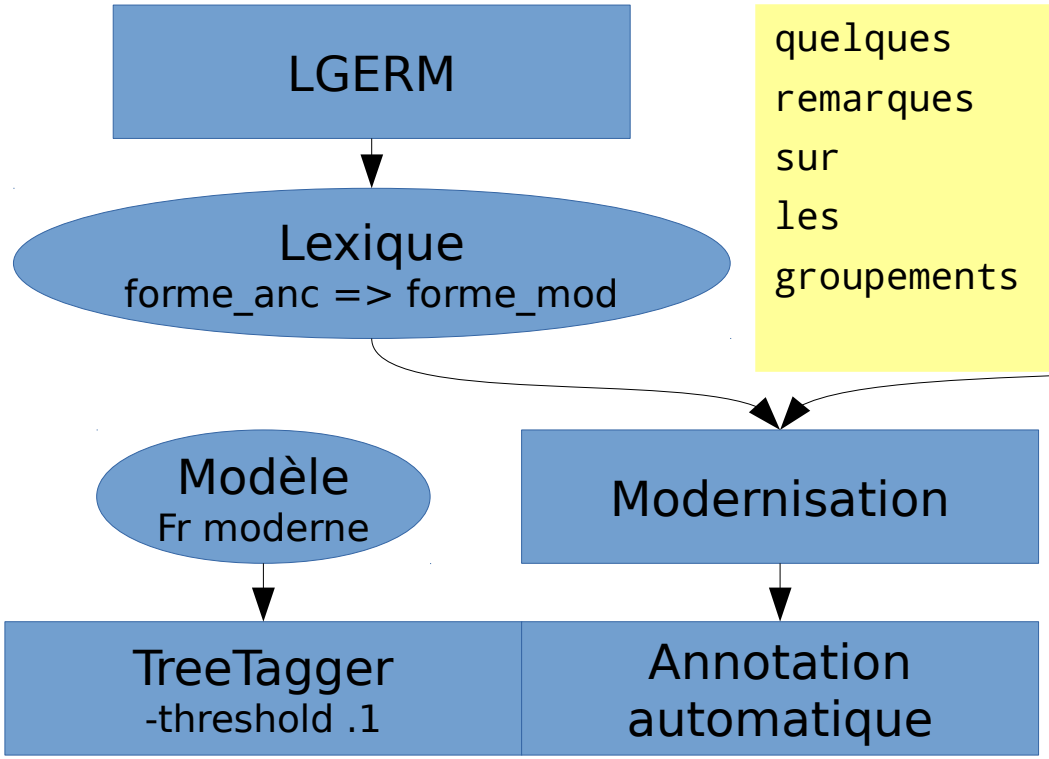
3b. Ressources – Corpus d'apprentissage

Désambiguïsation



3b. Ressources – Corpus d'apprentissage

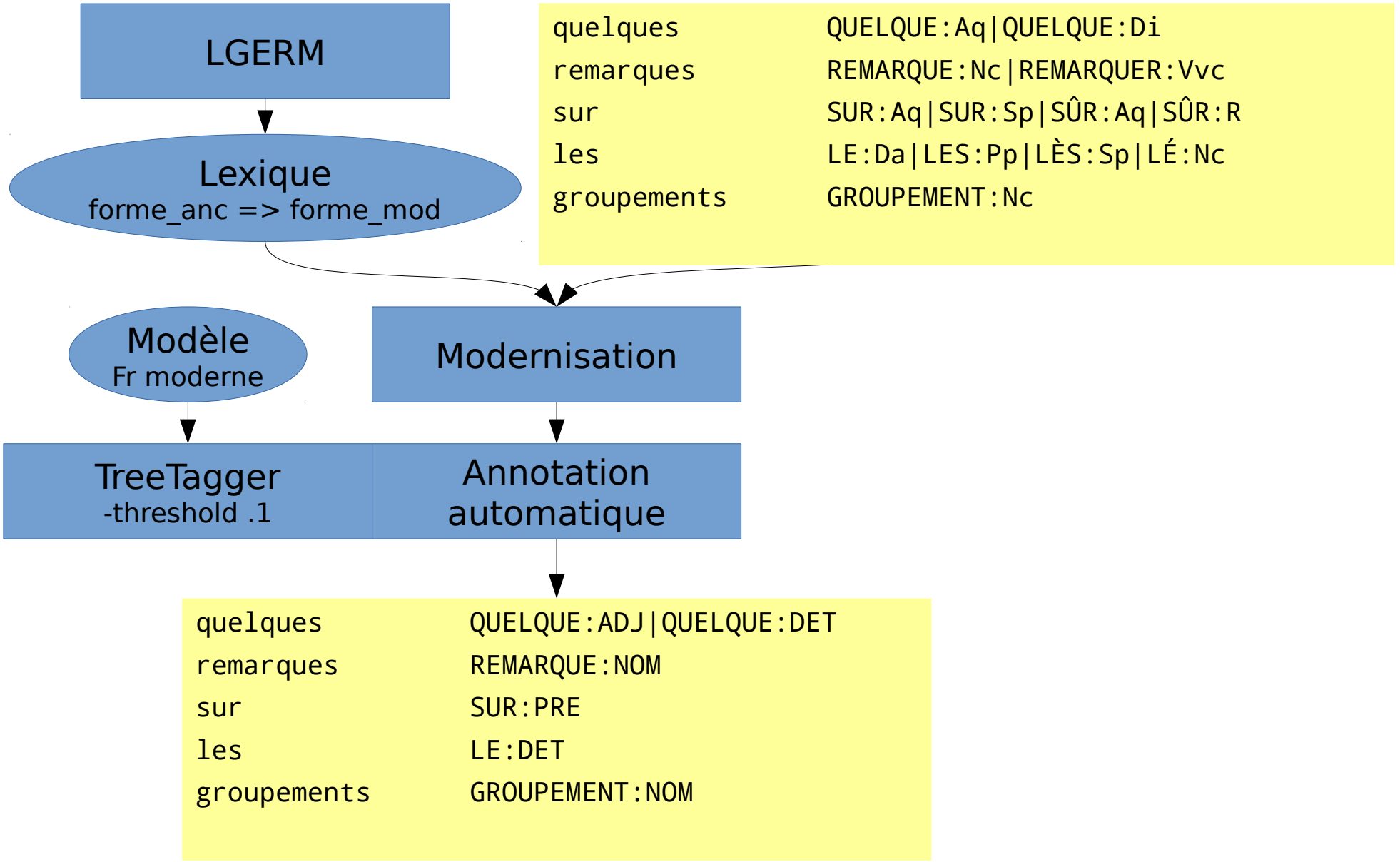
Désambiguïsation




```
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
```

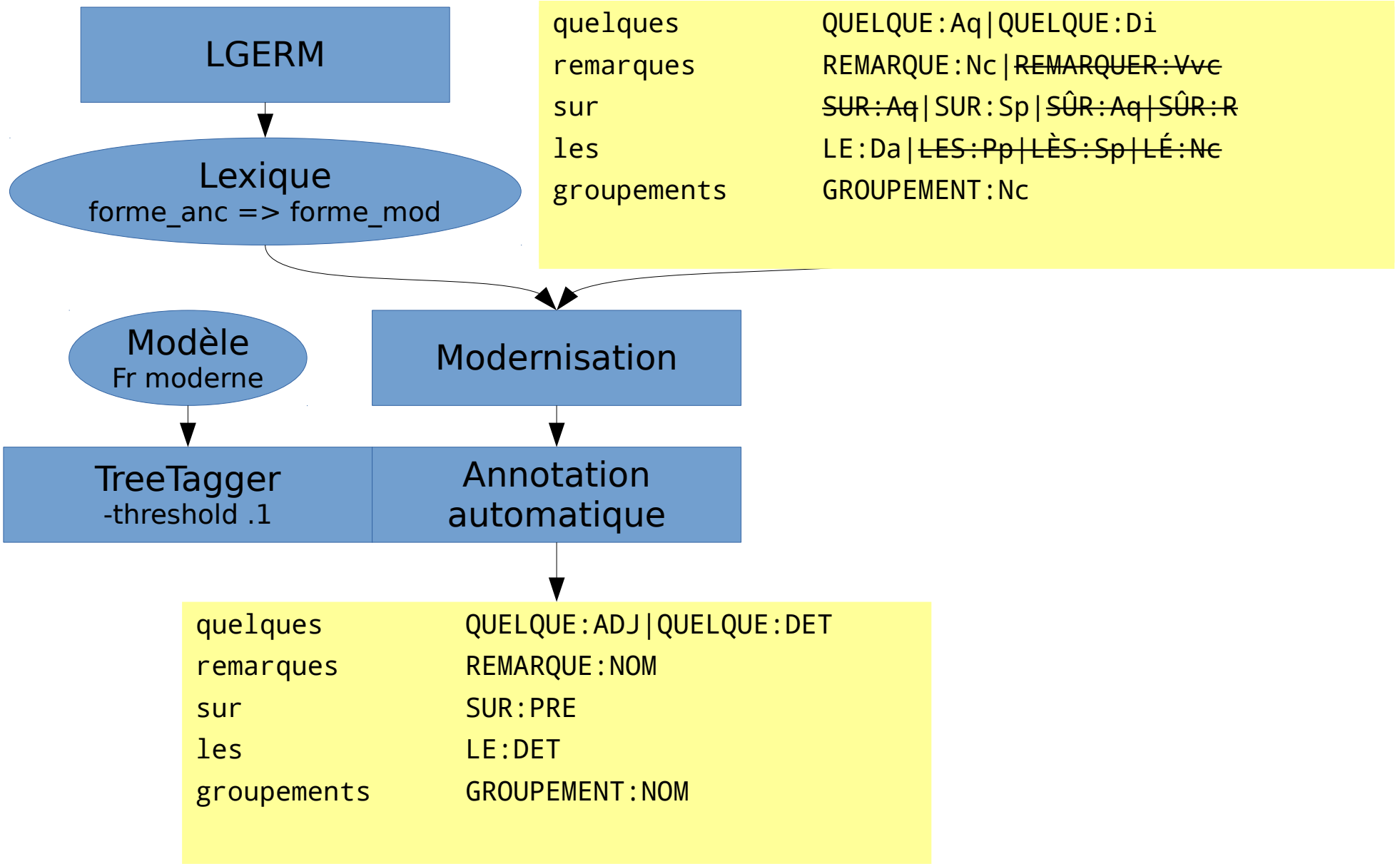
3b. Ressources – Corpus d'apprentissage

Désambiguïsation



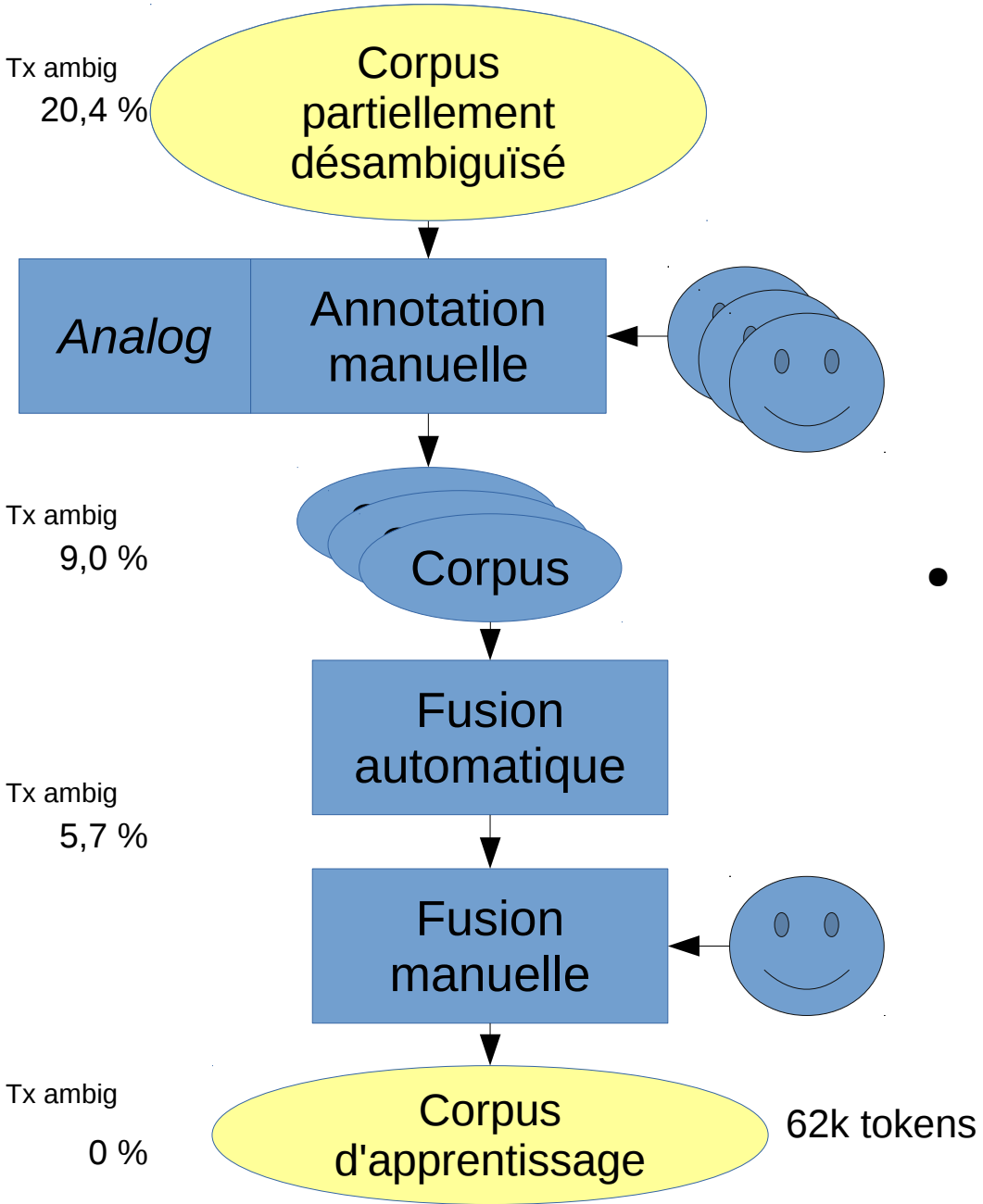
3b. Ressources – Corpus d'apprentissage

Désambiguïsation



3b. Ressources – Corpus d'apprentissage

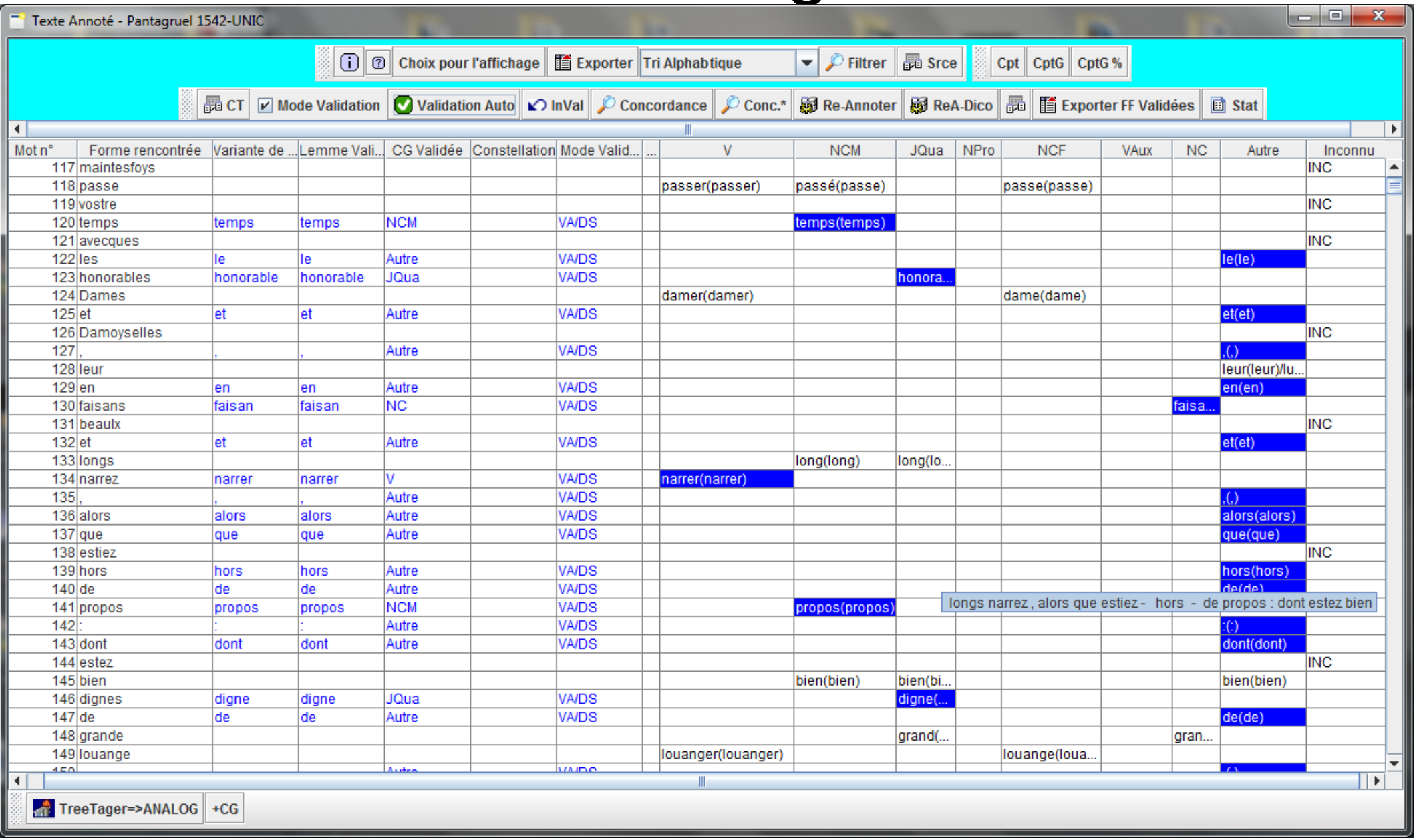
Annotation manuelle et fusion



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

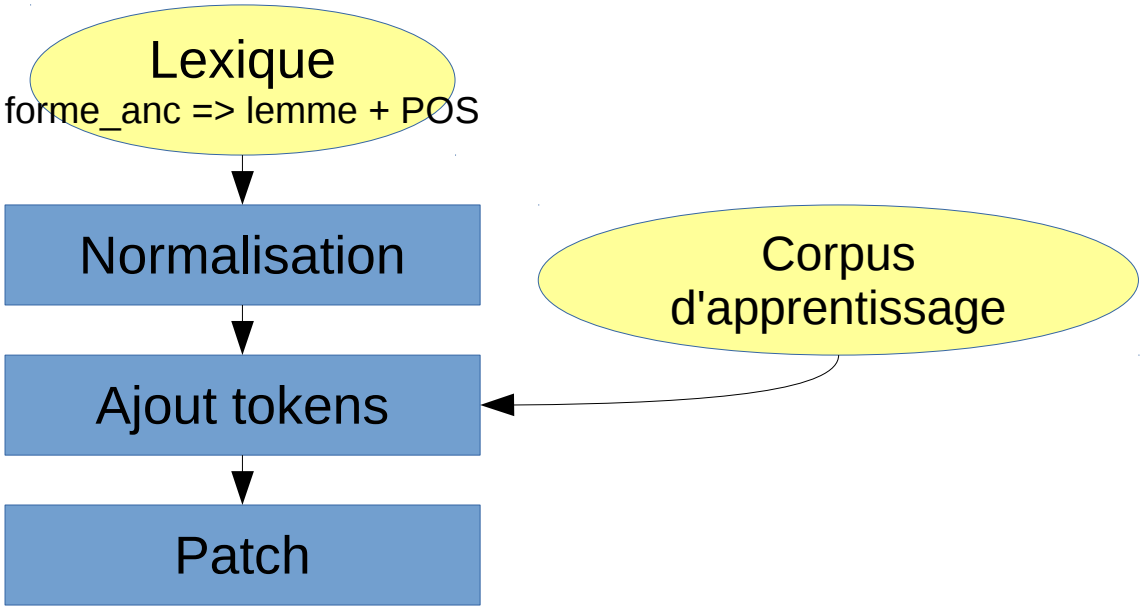
3b. Ressources – Corpus d'apprentissage

Analog



3c. Ressources – Modèle de langage

Préparation du lexique



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

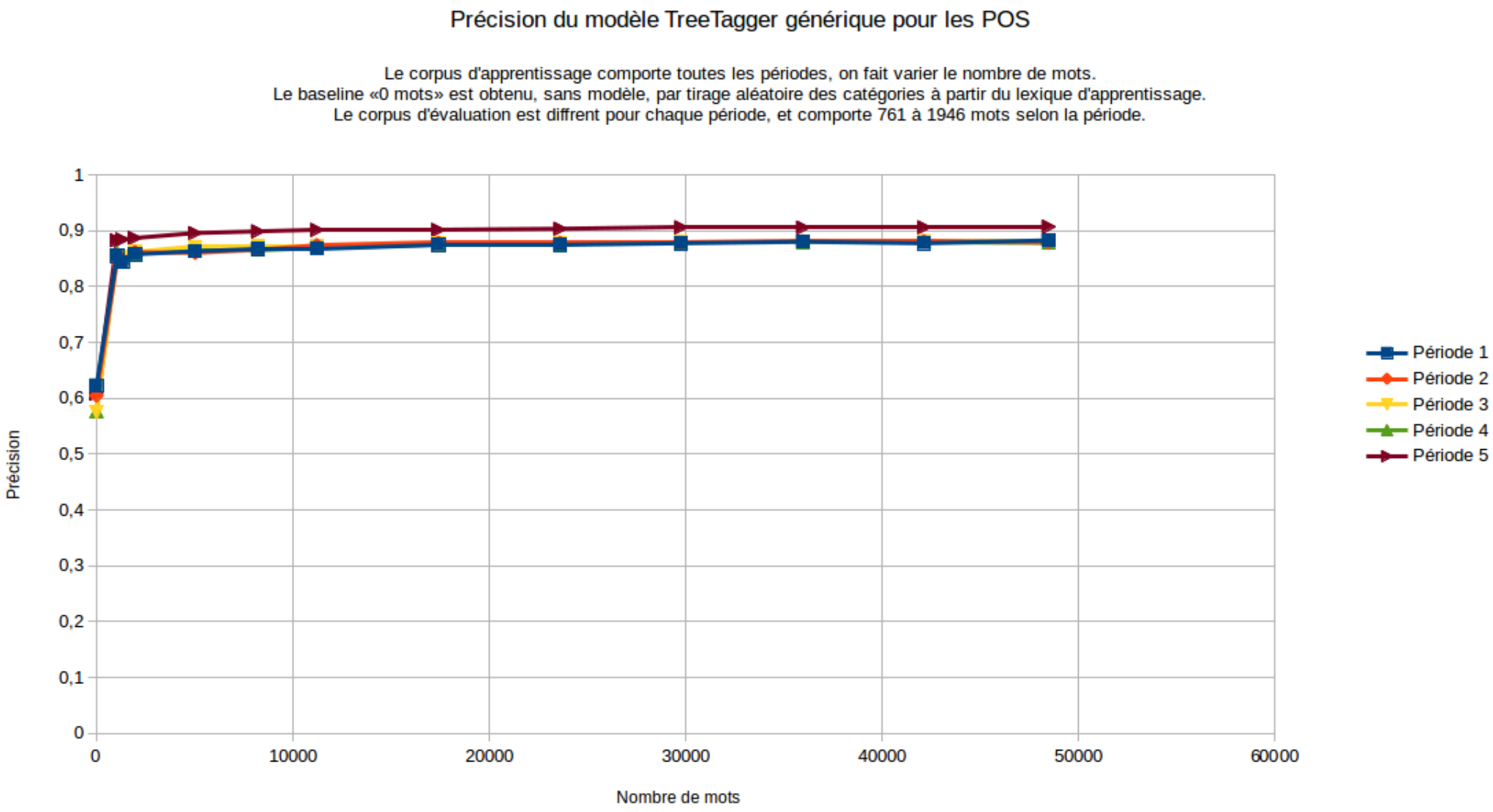
3c. Ressources – Modèle de langage

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 %

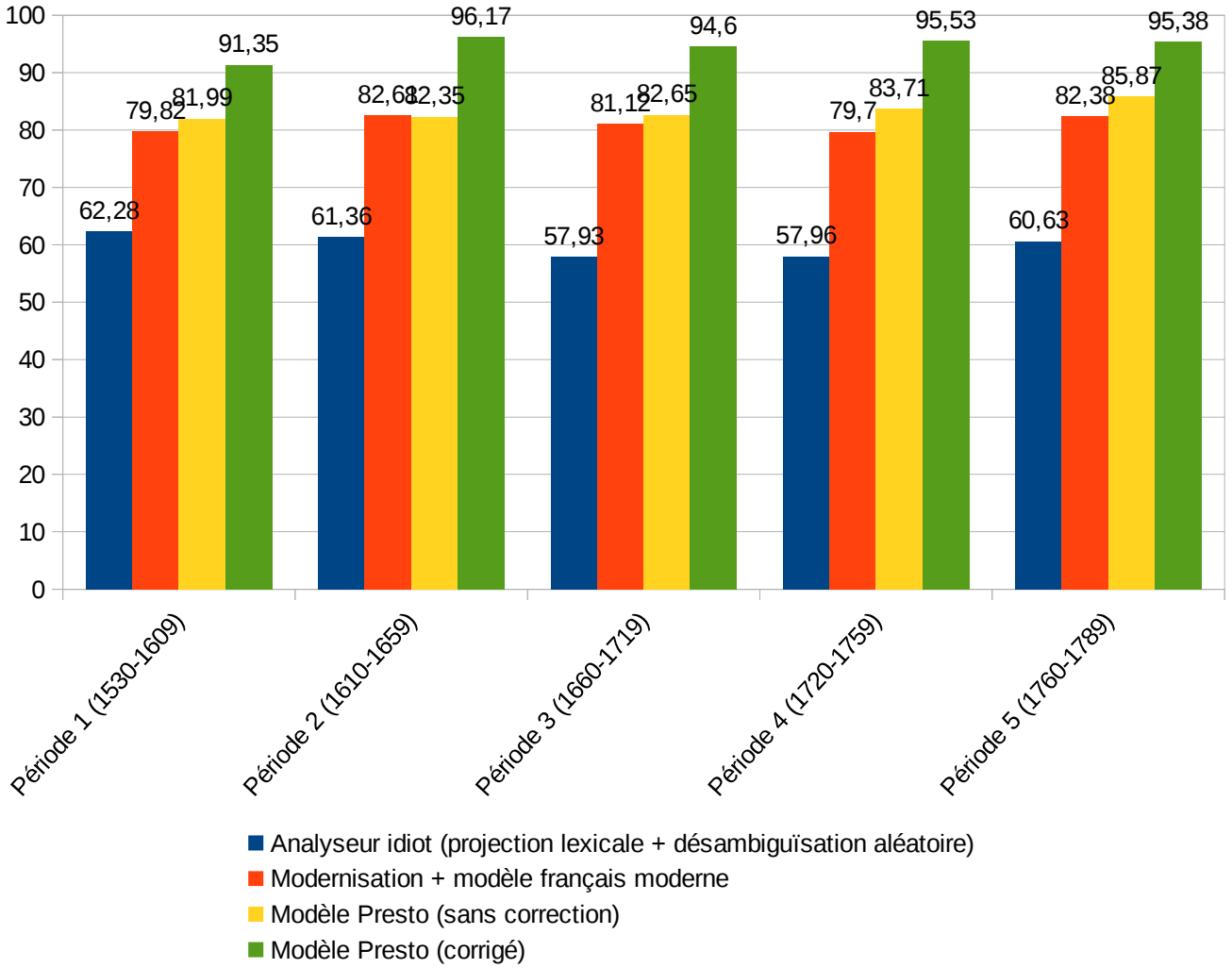
3c. Ressources – Modèle de langage

Évaluation du modèle



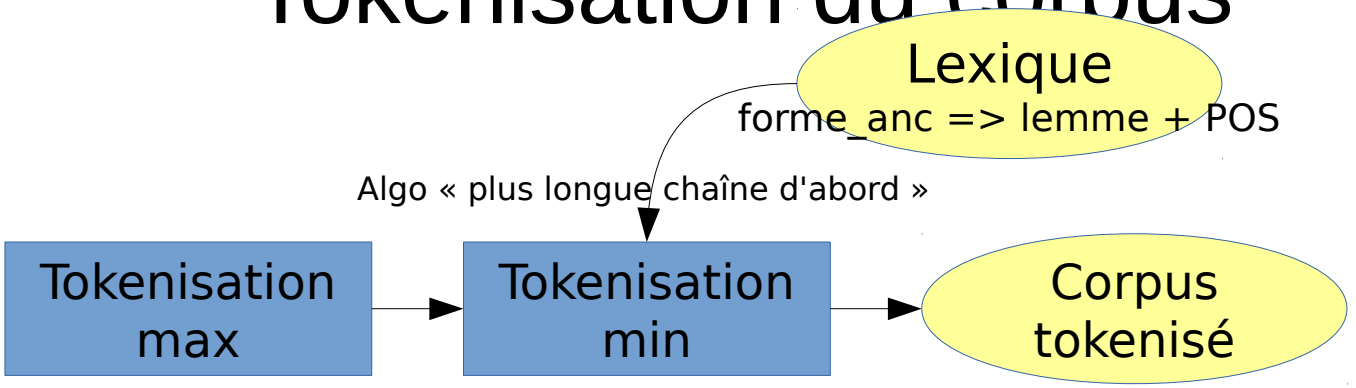
3c. Ressources – Modèle de langage

Évaluation du modèle



4. Traitement du corpus

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

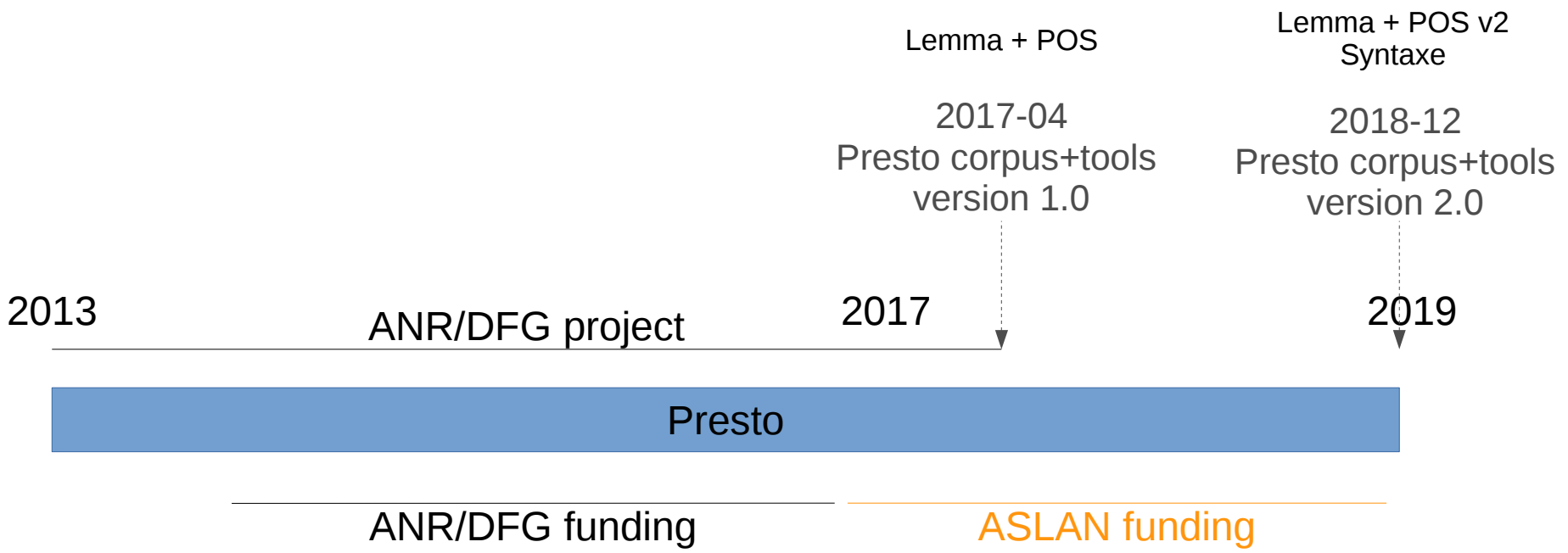
*
<foreign>

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 traitement : *Pipes* Unix
    - ~multitâche
    - Souple, facile à déboguer
  - Shell, logithèque Unix, 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

État du projet et perspectives



- Direction : Peter Blumenthal (Köln), Denis Vigier (Lyon)
- Digitalised XML texts offered by Frantext (Nancy), Bibliothèques Virtuelles Humanistes (Tours), ARTFL (Chicago), Corpus Électronique de la Première Modernité (Paris)
- Corpus processing in collaboration with Sascha Diwersy (Köln / Montpellier), Marie-Hélène Lay (Poitiers), Gilles Souvay (Nancy)