Achille Falaise
achille.falaise@ens-lyon.fr

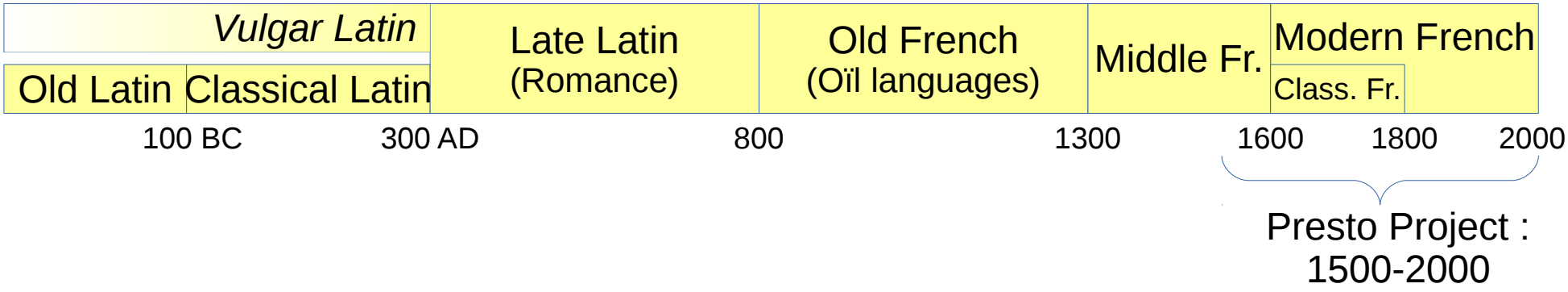
Resources Creation for an Under-Resourced Language: Classical French
(16th-18th Centuries)

http://presto.ens-lyon.fr

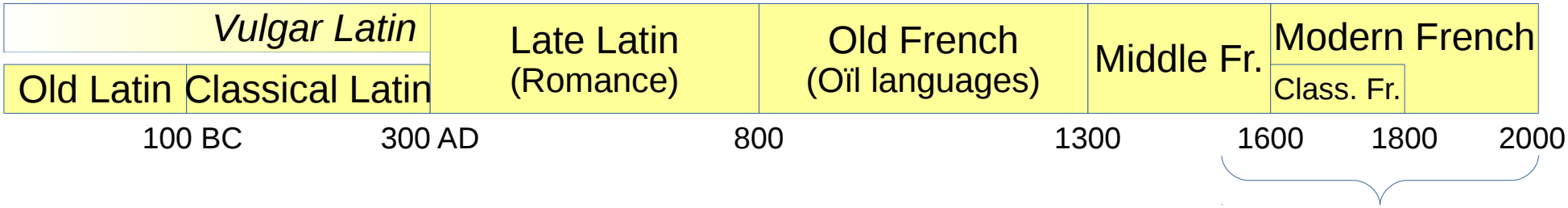
40ans2ta

icar UMR 5191 - CNRS
cnrs
ENS DE LYON
ASLAN UNIVERSITE DE LYON
Agence Nationale de la Recherche ANR
DFG Deutsche Forschungsgemeinschaft

1500-1800 : the rise of French as a national, standardised language



1500-1800 : the rise of French as a national, standardised language



- 1440 : Printing Press
- 1532 : *Pantagruel*, François Rabelais
- 1539 : Ordinance of Villers-Cotterêts − French is now France's official language
- 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 − state-sponsored standardisation of French langage
- 1751-1772 : *Encyclopédie*, Diderot & d'Alembert
- 1795 : École Normale de l'an III − standardisation of French education

XVIth Century French

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.

XVIIIth Century French

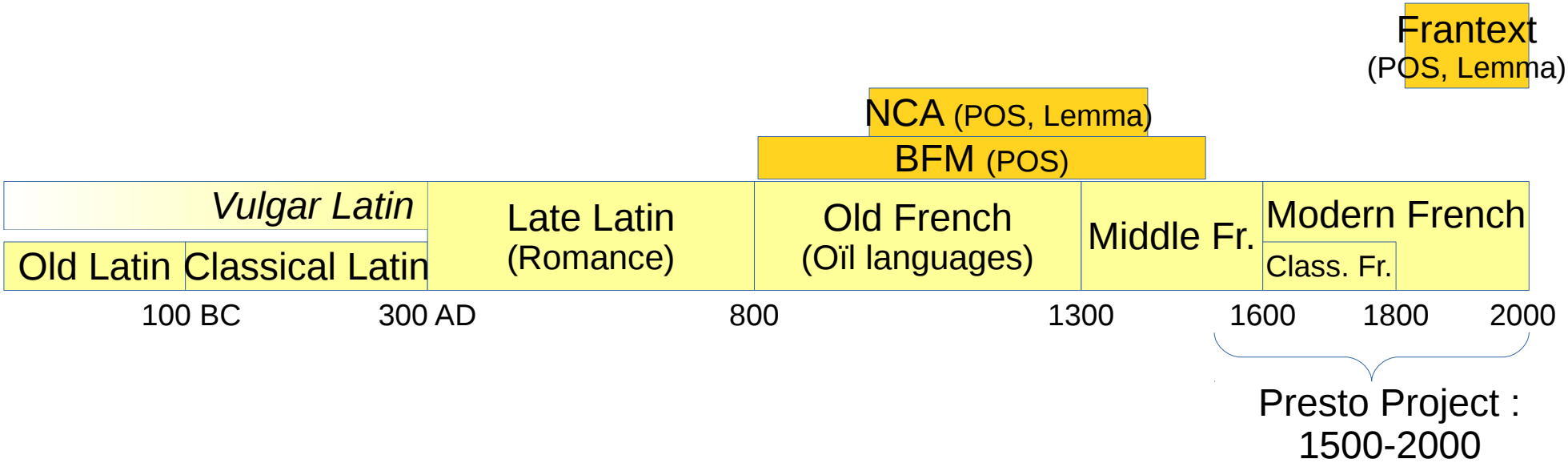
LANGAGE, s. m. (Arts. Raisonn. Philos. Metaphys.)

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.

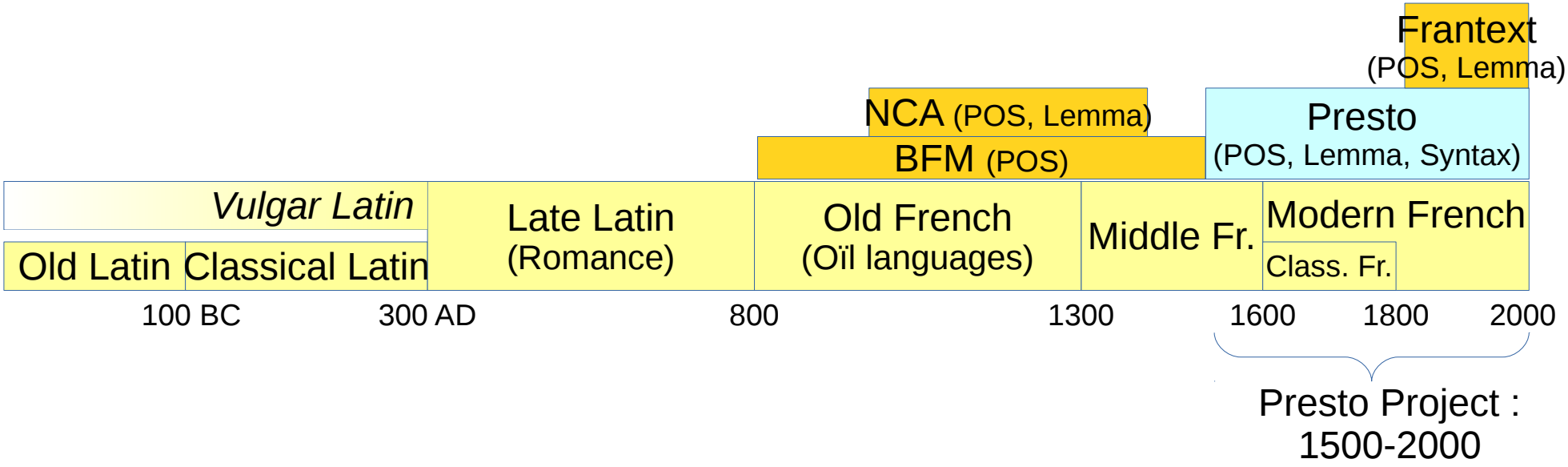
Dès que l'homme se sentit entraîné par goût, par besoin & par plaisir à l'union de ses semblables, il lui **étoit** nécessaire de développer son **ame** à un autre, & lui en communiquer les situations. Après avoir essayé plusieurs sortes d'expressions, il s'en tint à la plus naturelle, la plus utile & 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

Tagged diachronic corpora for French

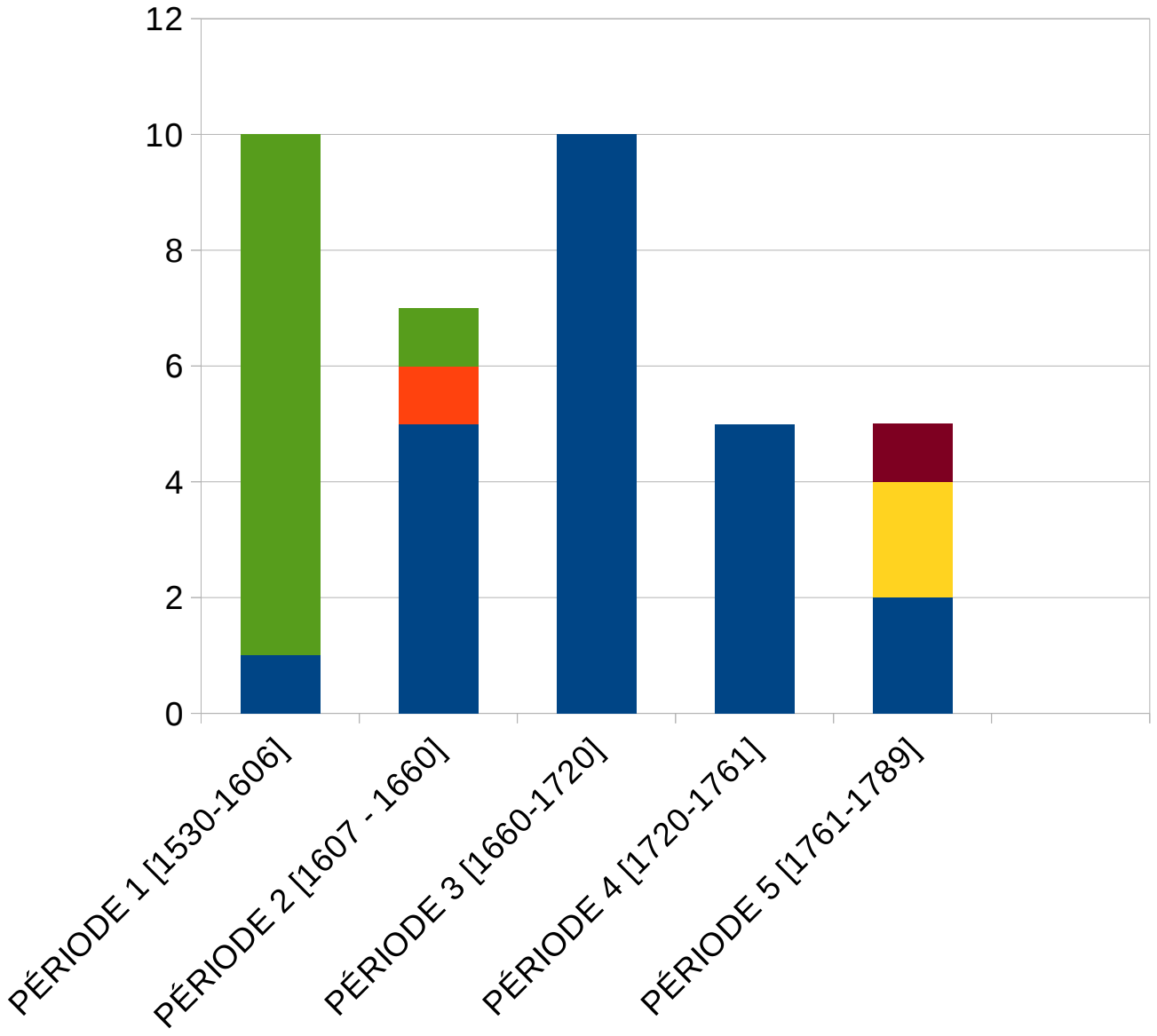


Tagged diachronic corpora for French



- Presto corpus
  - Litterature, essays, memoirs, etc.
  - 339 texts (of which 53 under Creative Commons licence)
  - 29M words

Tagged diachronic corpora for French



Frantext
S, Lemma)
Syntax)
French
0 2000
oject :
000

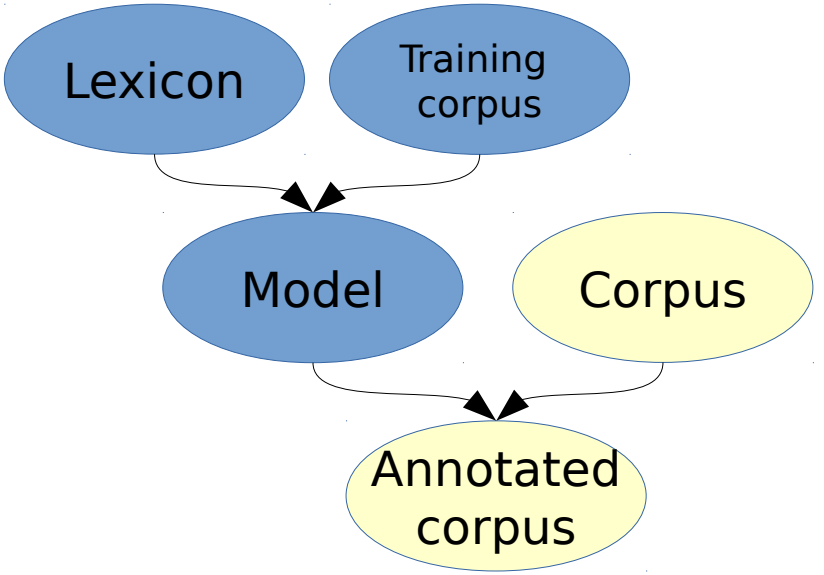
Old La

- P
  - –
  - –
  - –

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	

A classic workflow



Tokens, POS, lemmas

- Constraints :
  - *No feedback* between agents
  - *Tokens* are the minimal unit
    => tokenisation ambiguities are not allowed

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
    - Y → I
    - si (entre voyelles) alors C → SS

nes → nefs, NEF

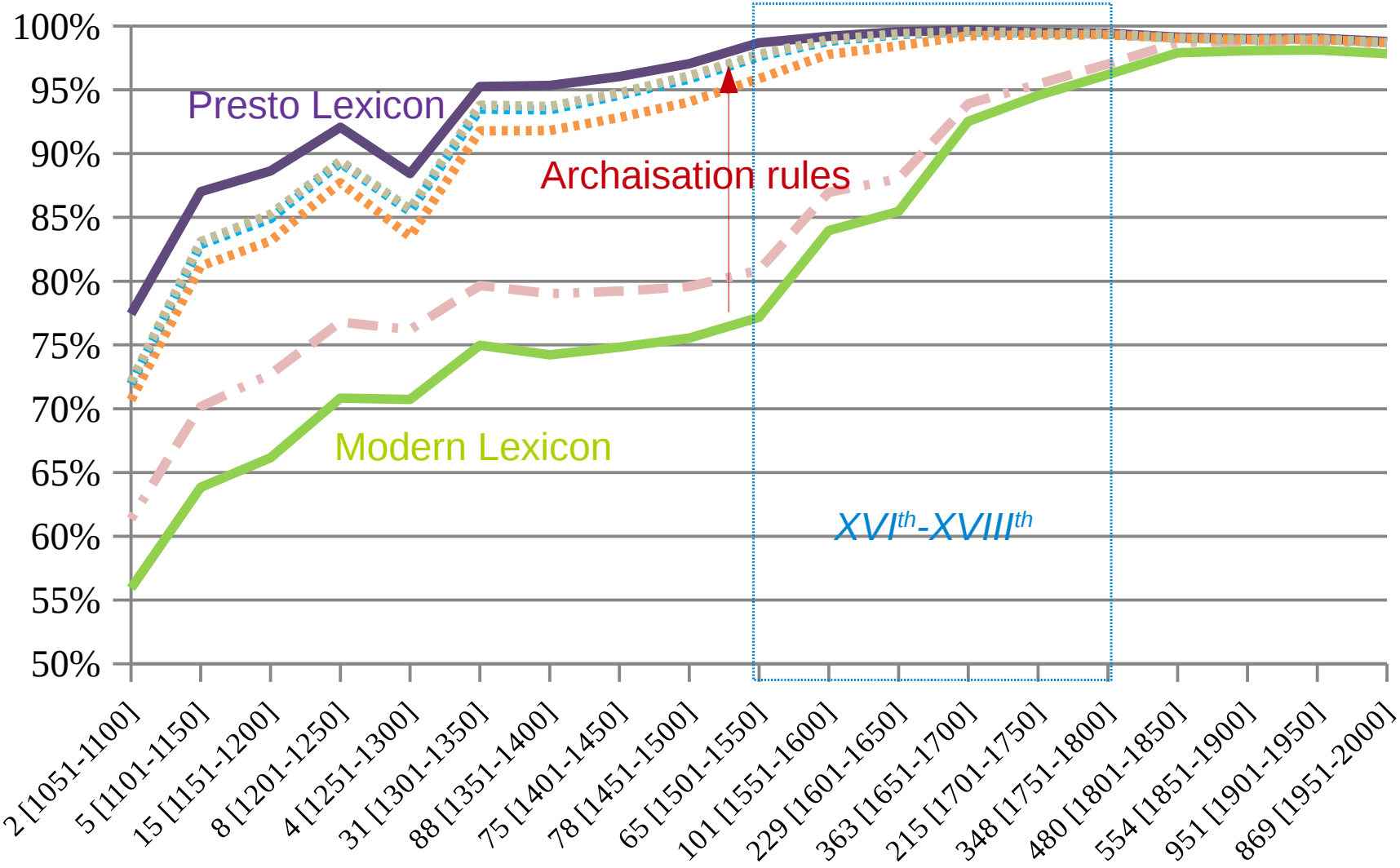
fayre → faire, FAIRE

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

Lexical coverage



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

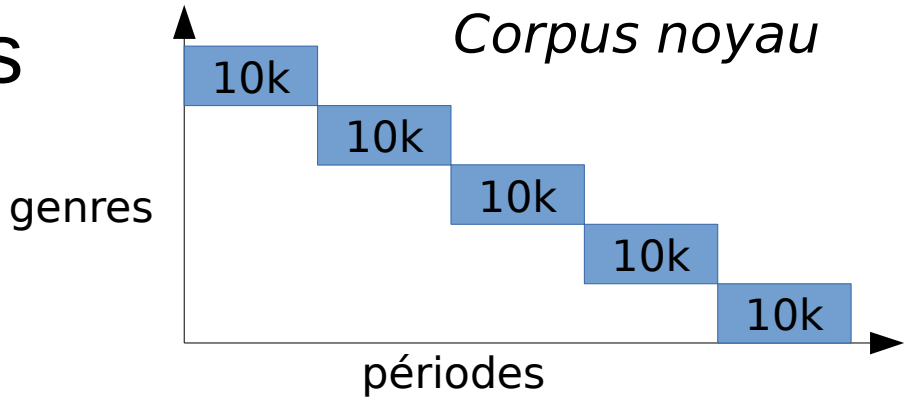
Saulsaye (1547)

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

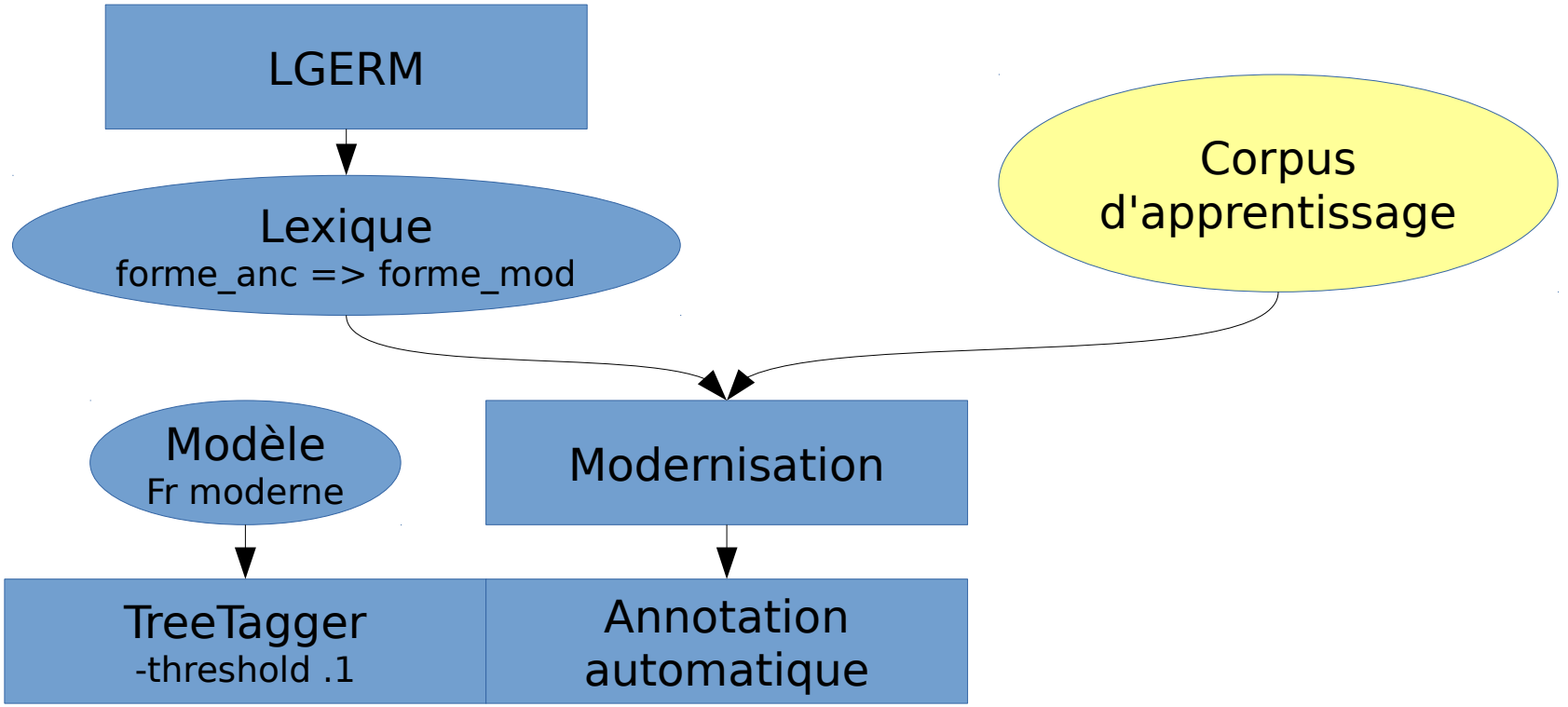
Simplification des étiquettes

```
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
```

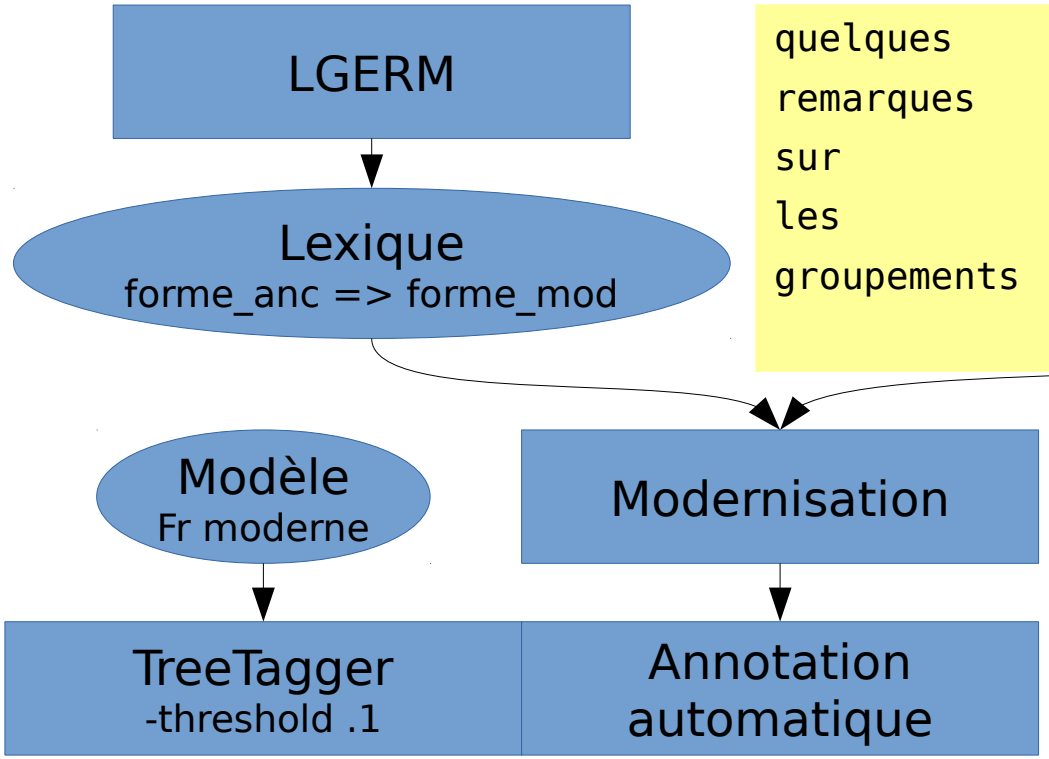
```
quelques        QUELQUE:Aq|QUELQUE:Di
remarques       REMARQUE:Nc|REMARQUER:Vvc
sur             SUR:Aq|SUR:Sp|SÛR:Aq|SÛR:R
les             LE:Da|LES:Pp|LÈS:Sp|LÉ:Nc
groupements     GROUPEMENT:Nc
```

Étiquettes Presto

Désambiguïsation

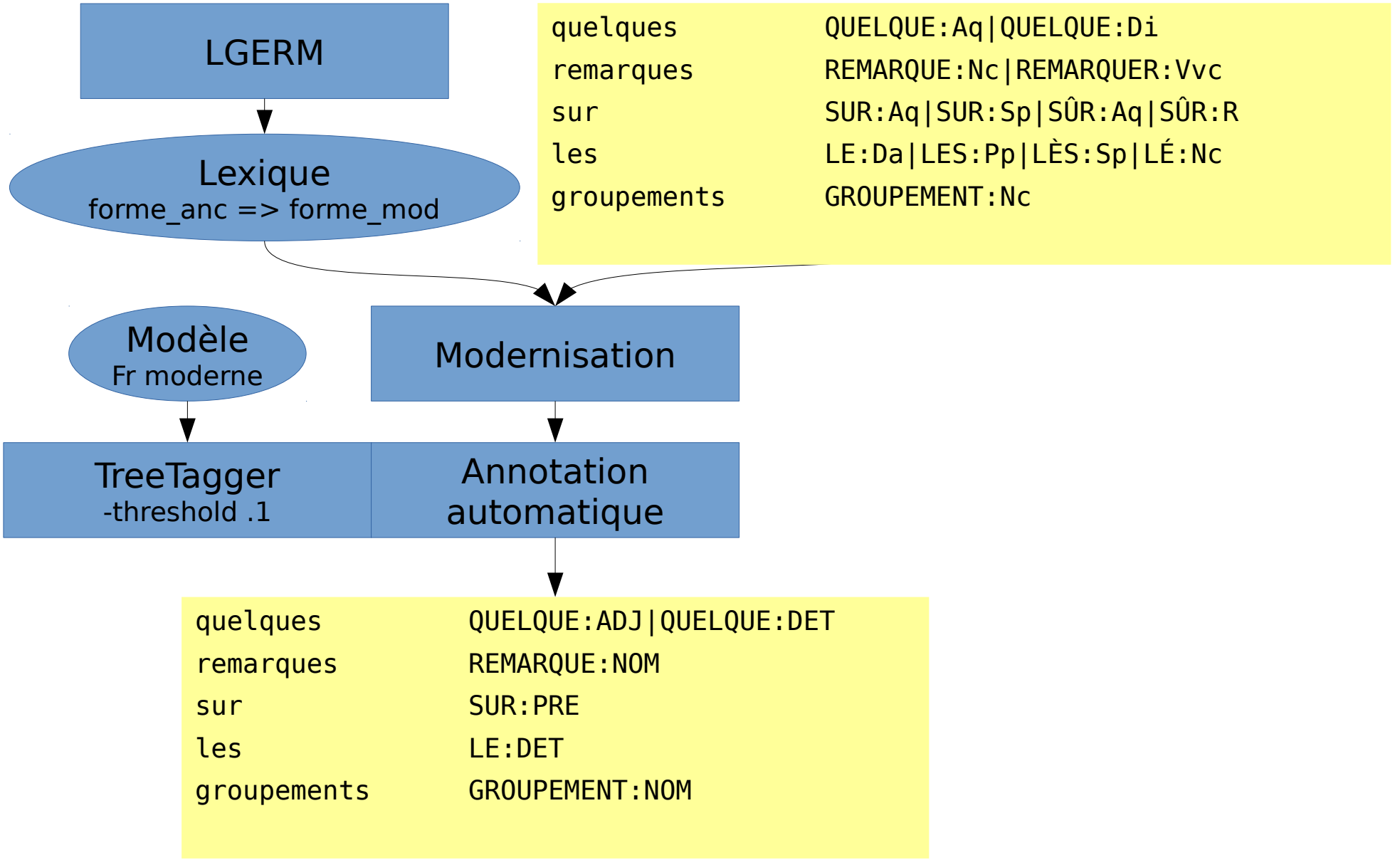


Désambiguïsation

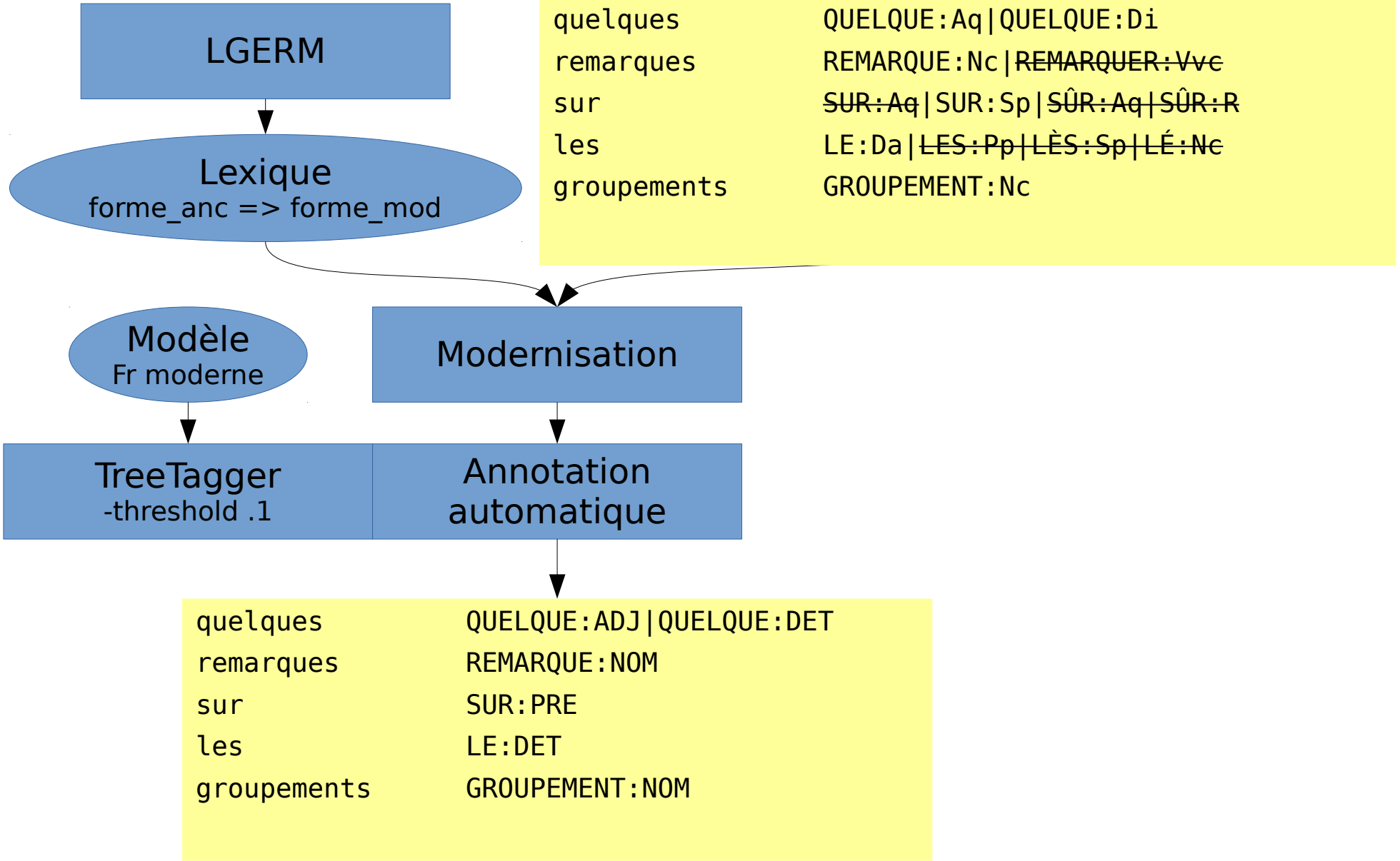


quelques	QUELQUE:Aq|QUELQUE:Di
remarques	REMARQUE:Nc|REMARQUER:Vvc
sur	SUR:Aq|SUR:Sp|SÛR:Aq|SÛR:R
les	LE:Da|LES:Pp|LÈS:Sp|LÉ:Nc
groupements	GROUPEMENT:Nc

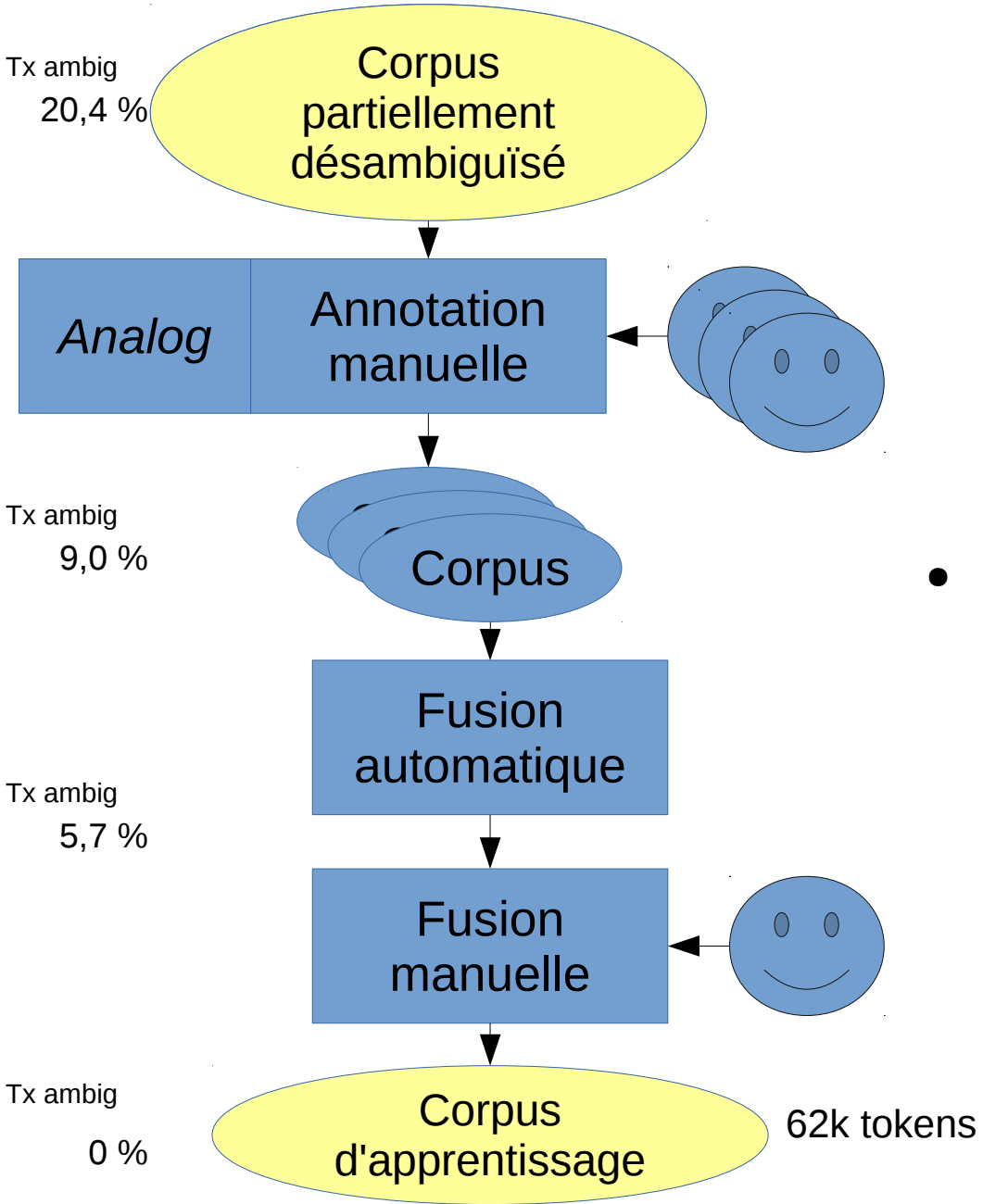
Désambiguïsation



Désambiguïsation

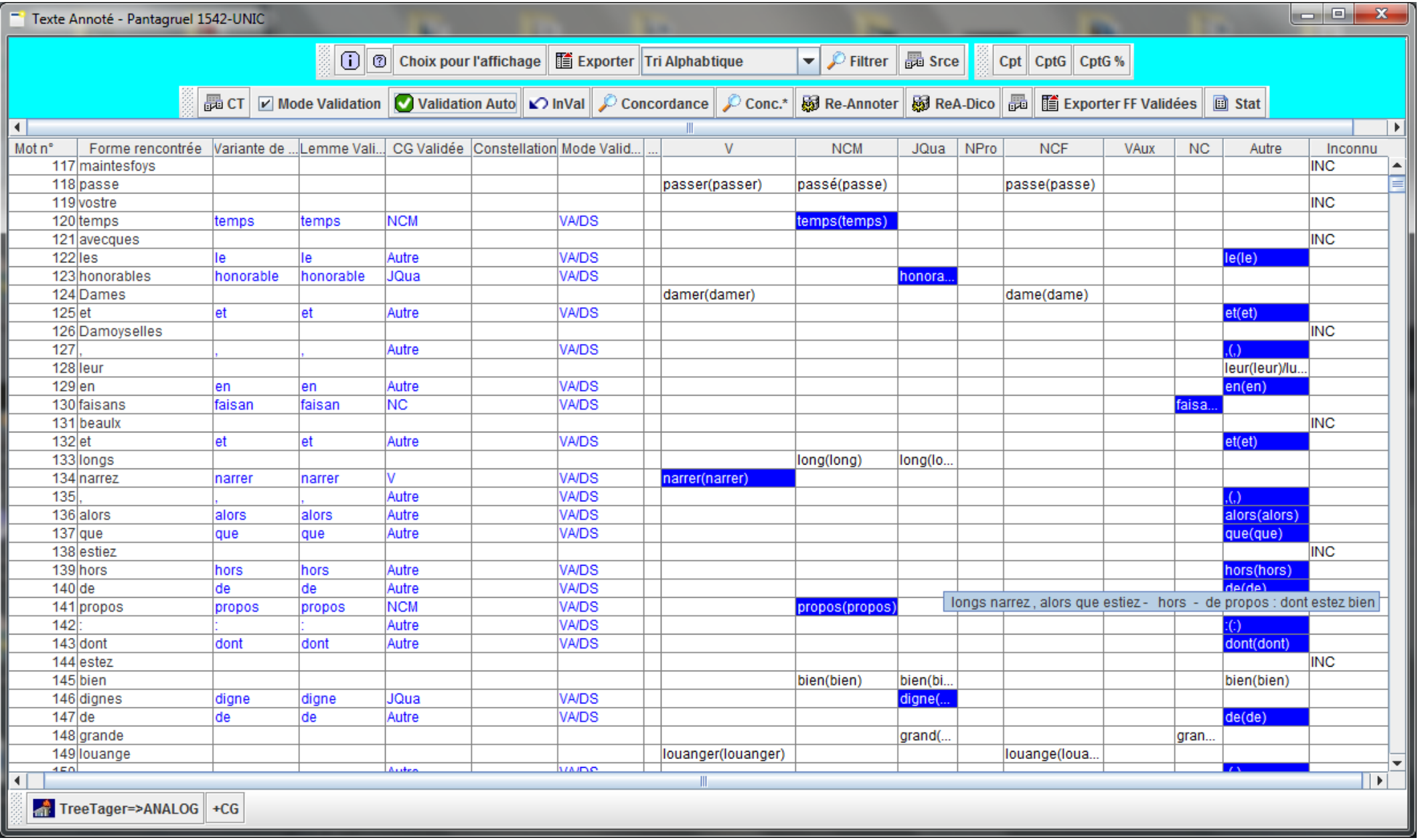


Annotation manuelle et fusion

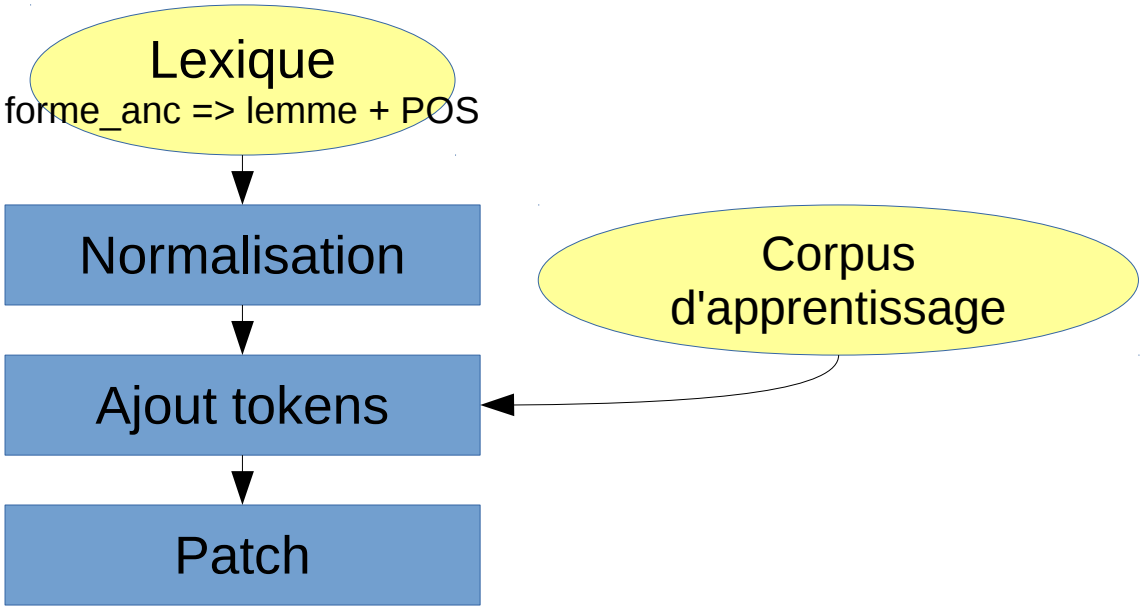


- Fusion automatique pour les cas « évidents » :
  - Au moins 2 annotateurs d'accord
  - Diacritiques

Analog


Texte Annoté - Pantagruel 1542-UNIC
Choix pour l'affichage
Exporter
Tri Alphabtique
Filtrer
Srce
Cpt
CptG
CptG %
CT
Mode Validation
Validation Auto
InVal
Concordance
Conc.*
Re-Annoter
ReA-Dico
Exporter FF Validées
Stat
Mot n°
Forme rencontrée
Variante de ...
Lemme Vali...
CG Validée
Constellation
Mode Valid...
...
V
NCM
JQua
NPro
NCF
VAux
NC
Autre
Inconnu
117 maintesfoys INC
118 passe passer(passer) passé(passe) passe(passe)
119 vostre INC
120 temps temps temps NCM VA/DS temps(temps)
121 avecques INC
122 les le le Autre VA/DS le(le)
123 honorables honorable honorable JQua VA/DS honora...
124 Dames damer(damer) dame(dame)
125 et et et Autre VA/DS et(et)
126 Damoyselles INC
127 , , , Autre VA/DS ,(,)
128 leur leur(leur)/lu...
129 en en en Autre VA/DS en(en)
130 faisans faisan faisan NC VA/DS faisa...
131 beaulx INC
132 et et et Autre VA/DS et(et)
133 longs long(long) long(lo...
134 narrez narrer narrer V VA/DS narrer(narrer)
135 , , , Autre VA/DS ,(,)
136 alors alors alors Autre VA/DS alors(alors)
137 que que que Autre VA/DS que(que)
138 estiez INC
139 hors hors hors Autre VA/DS hors(hors)
140 de de de Autre VA/DS de(de)
141 propos propos propos NCM VA/DS propos(propos)
longs narrez , alors que estiez - hors - de propos : dont estez bien
142 : : : Autre VA/DS :(:)
143 dont dont dont Autre VA/DS dont(dont)
144 estez INC
145 bien bien(bien) bien(bi... bien(bien)
146 dignes digne digne JQua VA/DS digne(...
147 de de de Autre VA/DS de(de)
148 grande grand(... gran...
149 louange louanger(louanger) louange(loua...
TreeTager=>ANALOG
+CG

Préparation du lexique

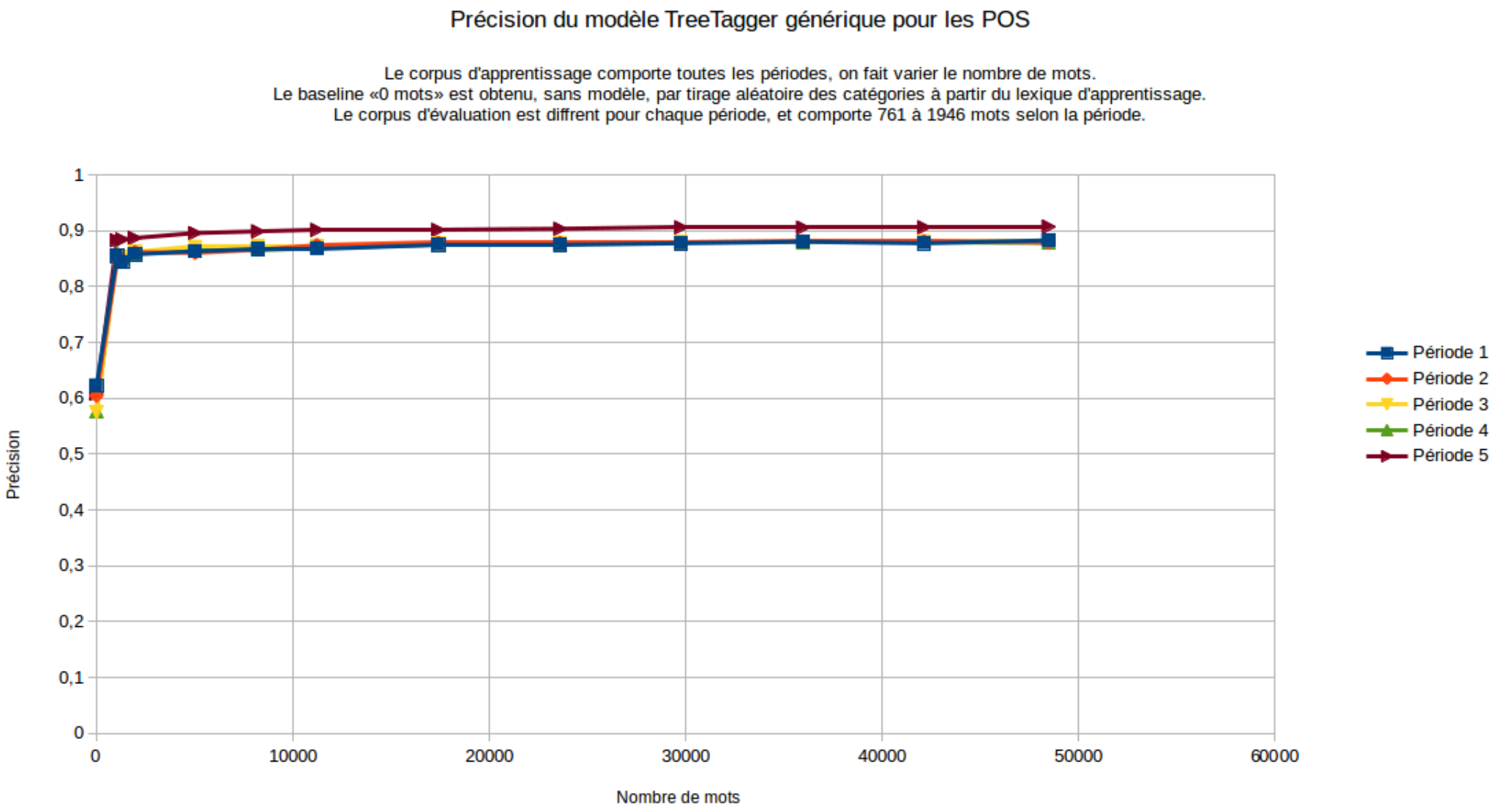


- Patch :
  - Listes de tokens
    - À ajouter
    - À enlever
  - Règles ad hoc

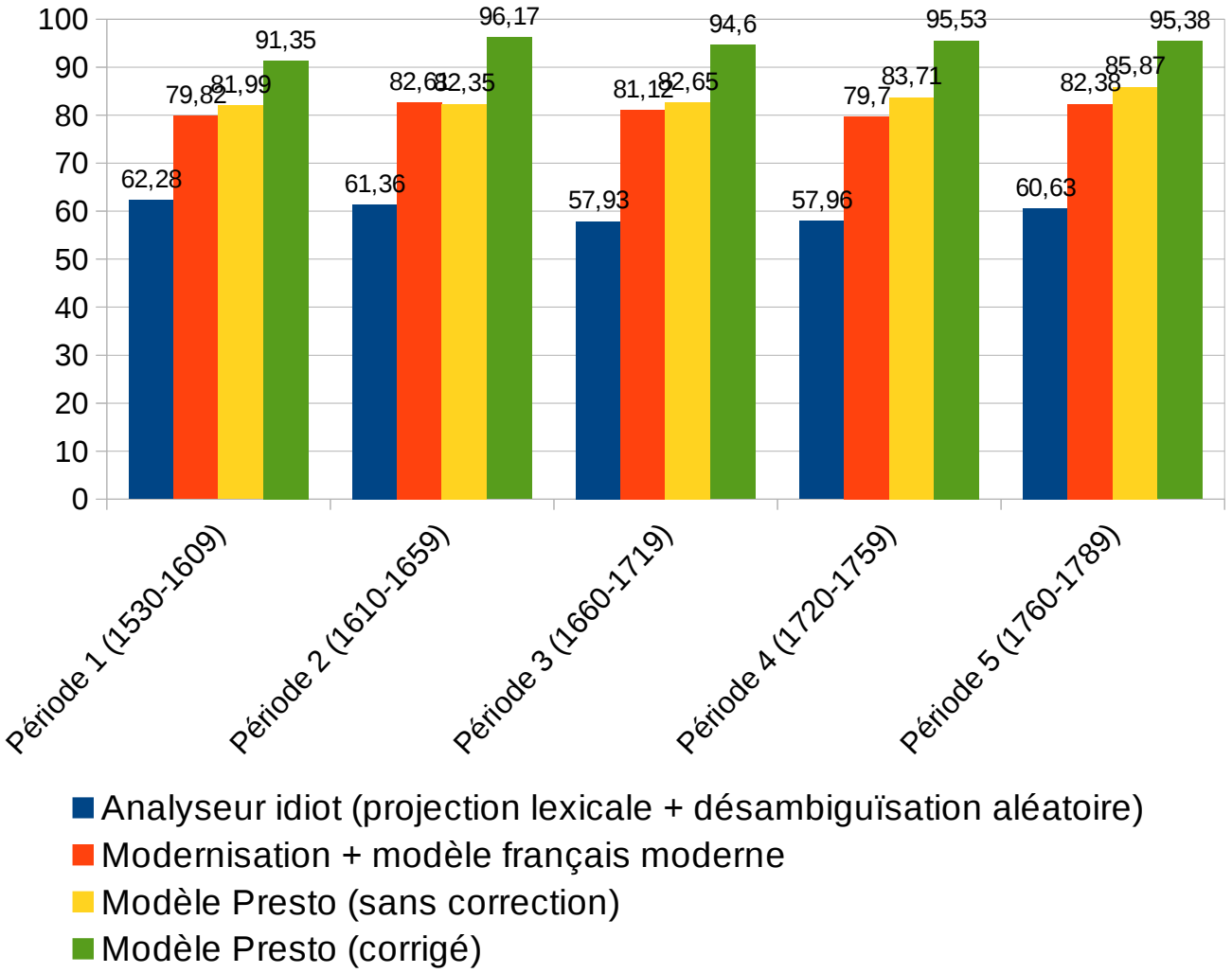
Model creation

- Training corpus divided in 3 parts
  - Train (80% − 49 630 tokens)
  - Dev (10% − 6 164 tokens)
  - Eval (10% − 6 110 tokens)
- *Autotuning* to find the best parameters for TreeTagger
  - cl 2 ; dtg 0,5 ; sw 1 ; ecw 0,06 ; atg 1,15
  - Precision gain : +0,05 %
- Evaluation
  - Train : 95,77 %
  - Dev : 94,28 %
  - **Eval : 94,46 %**

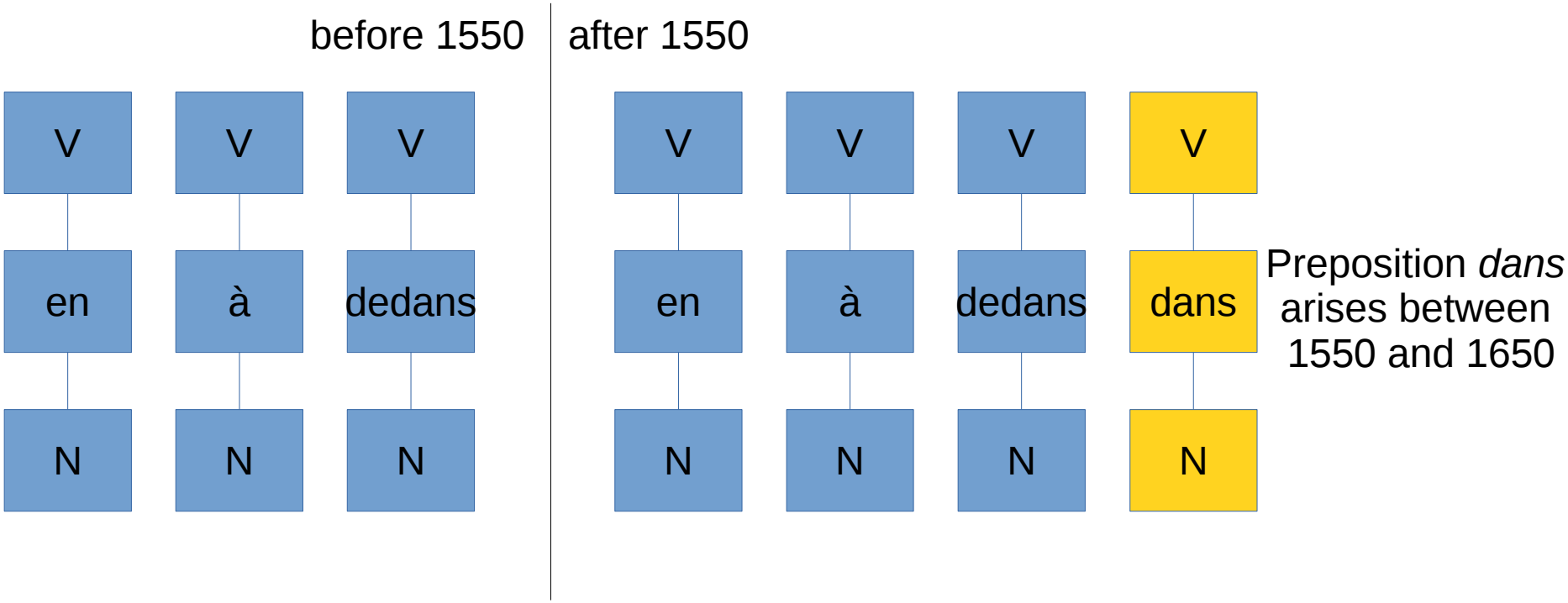
Évaluation du modèle



Évaluation du modèle



Prepositions dynamics



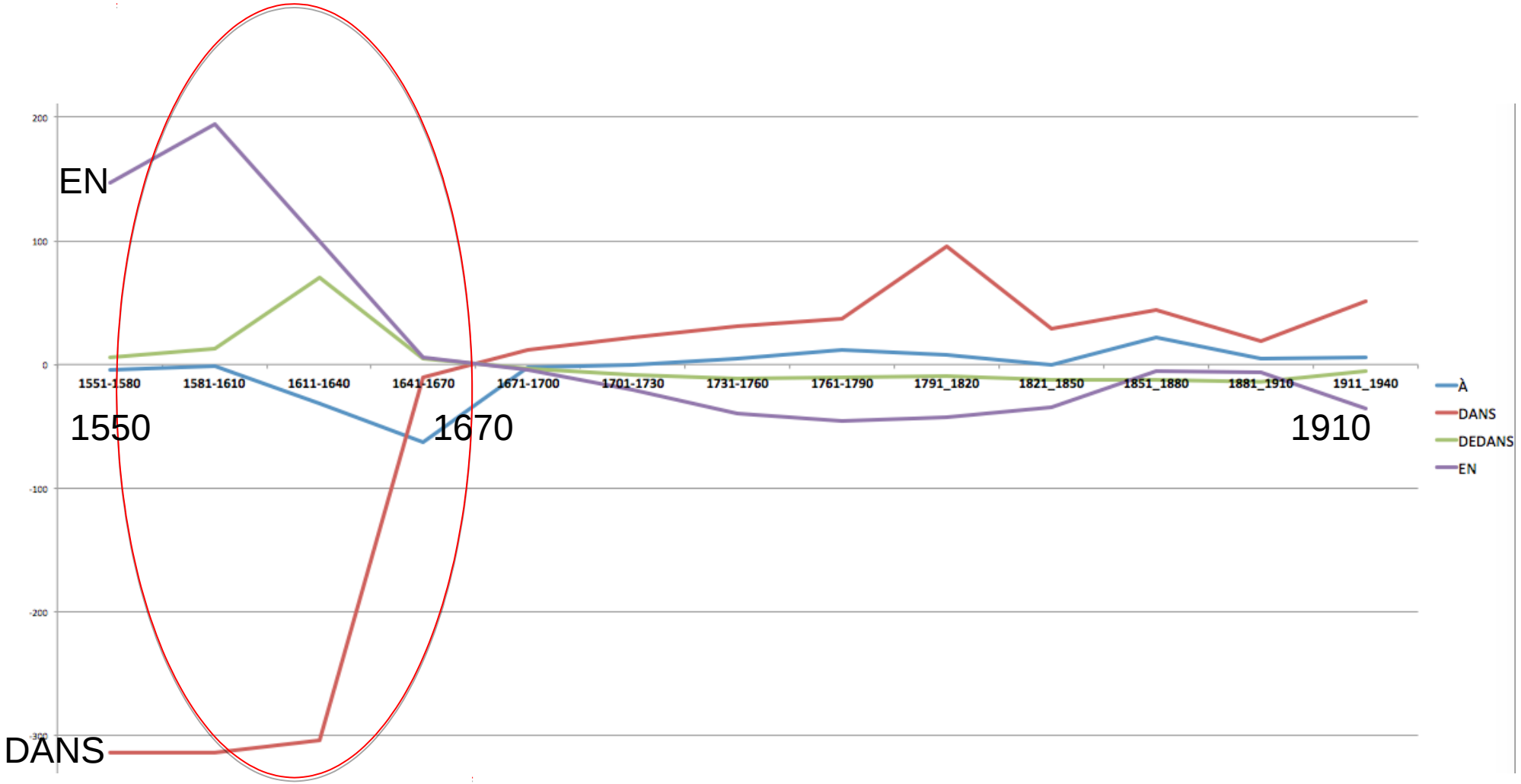


Diagramme 1. Évolution des scores de spécificité de Laffon affectés aux prépositions *en, dans*, *dedans* entre 1551 et 1900. Corpus Presto, partitionné en 13 tranches de 30 ans.

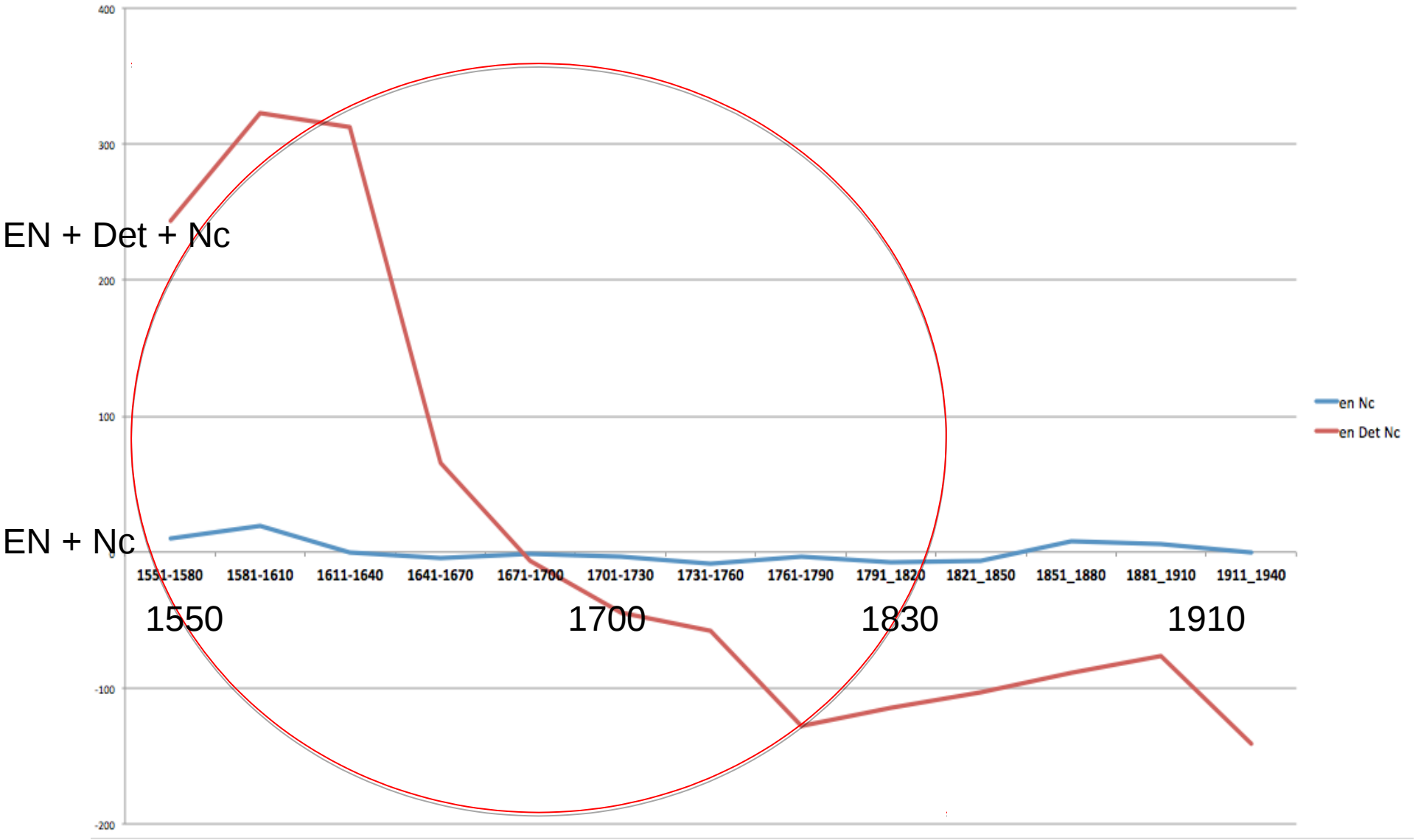
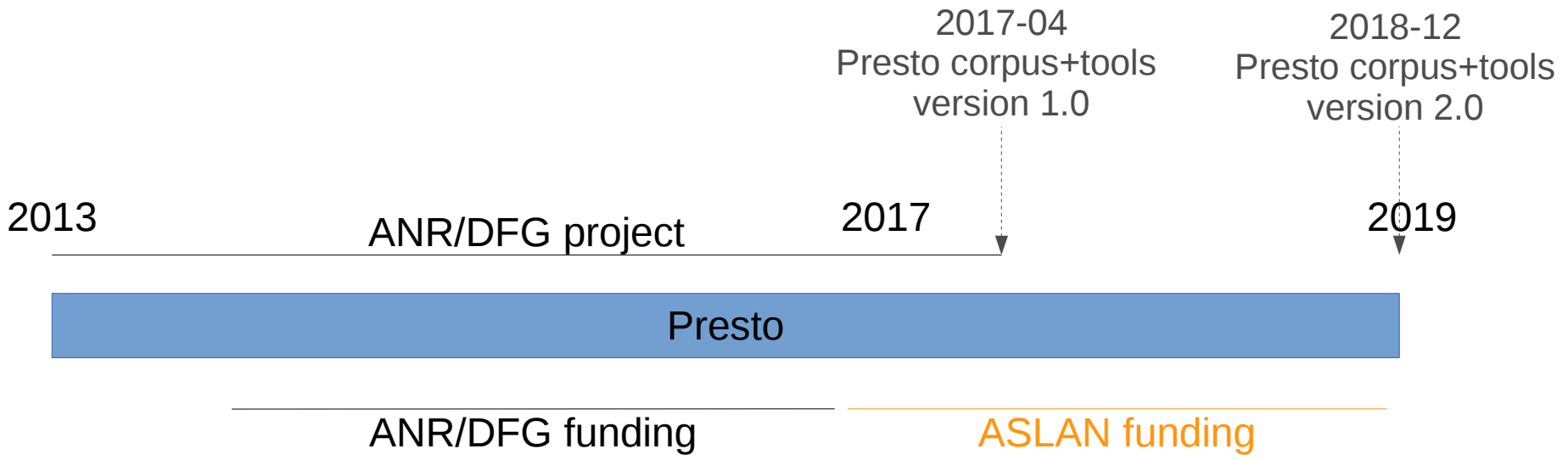


Diagramme 1. Évolution des scores de spécificité de Laffon affectés aux prépositions *en, dans*, *dedans* entre 1551 et 1900. Corpus Presto, partitionné en 13 tranches de 30 ans.

Presto project



- 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)

Most specific nouns by century

16th century	17th century	18th century	19th century	20th century (<1940)
dit	lettre	marquis	don	bacille
seigneur	amour	nation	médecine	infection
église	dessein	peuple	lady	tuberculose
jeunesse	madame	animal	phénomène	lésion
écriture	vertu	génie	habitude	môme
suppôt	contentement	ouvrage	avoué	ganglion
dame	esprit	nature	scène	assassin
sentence	gloire	objet	idée	politique
cerveau	fortune	plaisir	lord	humanisme
propos	sujet	idée	société	civilisation
hôte	proposition	goût	moment	cobaye
damoiseau/selle	mal	talent	physiologie	porte
chose	discours	baron	individu	temporel

Most specific verbs by century

16th century	17th century	18th century	19th century	20th century (<1940)
advenir	faire	former	exister	aller
requérir	savoir	instruire	comprendre	rester
bailler	aimer	élever	rester	apparaître
ouïr	obliger	jouir	modifier	comprendre
vouloir	témoigner	convenir	rappeler	regarder
confesser	mander	naître	admettre	poser
endurer	voir	apercevoir	rentrer	réaliser
engendrer	estimer	greffer	indiquer	bouger
ensuivre	rendre	rassembler	agir	constater
pendre	croire	prouver	entourer	lever
apparaître	assurer	raisonner	écrier	surveiller
faillir	servir	flatter	retrouver	remonter
départir	trouver	rendre	aller	partir