Traitement automatique du français en diachronie, retour d'expérience sur le projet Presto

Sascha Diwersy, U. Montpellier, Praxiling
Achille Falaise, ENS de Lyon, ICAR
Gilles Souvay, U. Lorraine, ATILF

CLPS 2016
9-11 mars 2016, Lyon

Le lexique PRESTO

- Lexique adapté à la langue du XVIe-XVIIe
  - orthographe non stabilisée : variation et flexion
  - segmentation
- Archaïsation du lexique moderne et projection sur un corpus textuel
  - archaïsation par règles
    - base de connaissances du lemmatiseur LGeRM
  - vérification du taux de couverture sur Frantext
    - forme, fréquence, tranche chronologique
  - expérience projet Européen IMPACT

Le lemmatiseur LGeRM

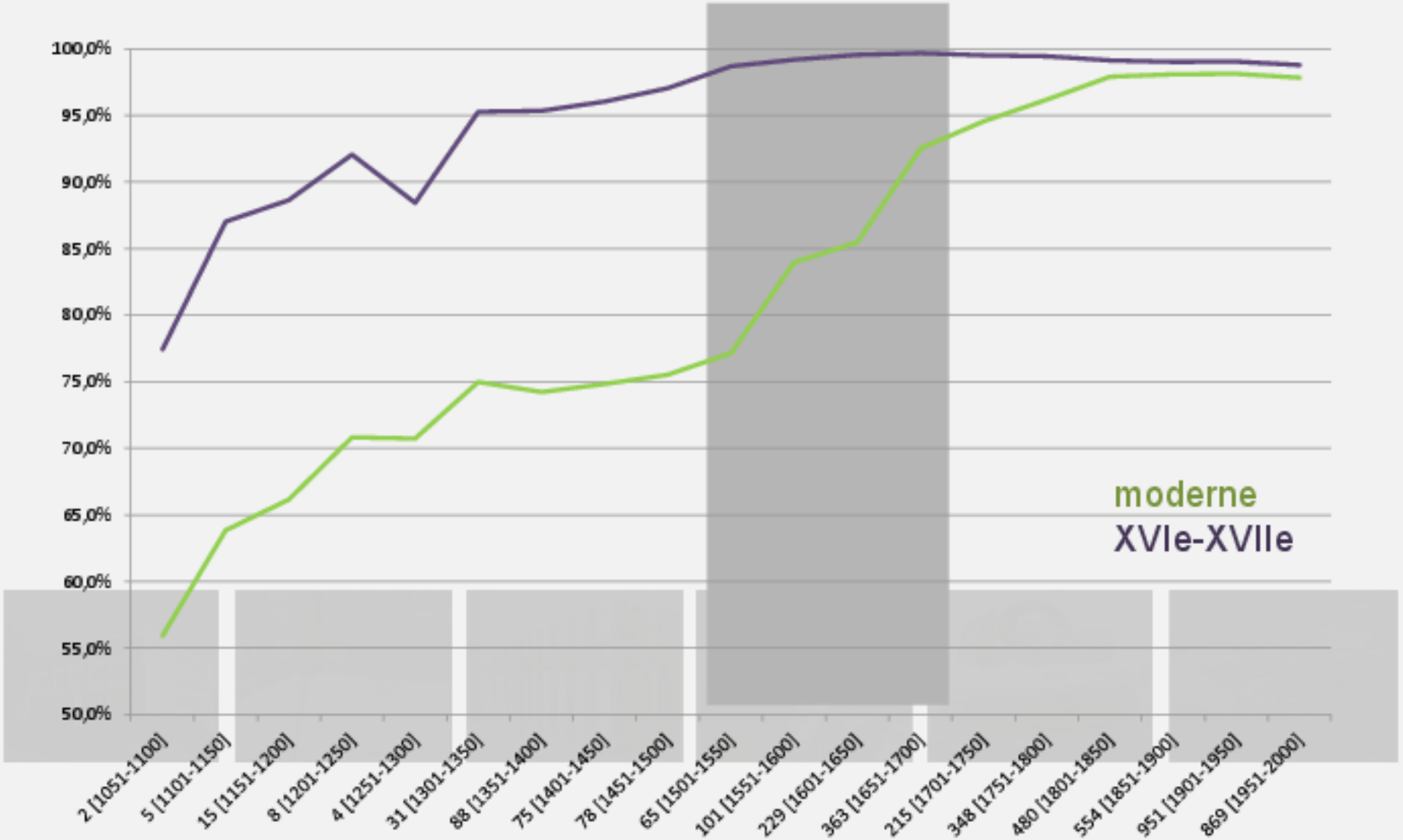
- LGeRM : **L**emmes **G**raphies **e**t **R**ègles **M**orphologiques
  - lemmatiseur et environnement de lemmatisation
  - gestion de différents états de langues
  - http://www.atilf.fr/LGeRM
- Liste de formes connues : lexique morphologique
  - gestion de « flexion et variantes » dans Frantext
- Règles d'analyse des formes inconnues
  - 6 500 règles (4/5 flexion verbale)
    - `si` (en finale) `alors` ES → EFS `finsi` — *nes* → *nefs*, NEF
    - Y → I — *fayre* → *faire*, FAIRE
    - `si` (entre voyelles) `alors` C → SS `finsi` — *mesfacent* → *mesfassent*, MÉFAIRE

Construction du lexique

- Lexique moderne LEFFF
  - étiquettes spécifiques au projet
  - lemmes complémentaires

- Construction
  - appliquer les règles d'archaïsation (3 boucles)
  - regarder les formes absentes
  - réitérer le processus
  - compléter avec les formes absentes

Taux de couverture du lexique



Perspectives

- On n'aura jamais toutes les variantes
  - intégration du lemmatiseur dans la chaîne de traitement ?
- Produire un lexique adapté à un état de langue, à un type de typographie
  - origine de la forme
  - fréquence des mots

Le corpus Presto

- Besoins
  - Représentation de toutes les périodes de l'histoire du français : 9ème s. au 21ème s.
    - Dans un premier temps : 16ème s. au 21ème 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*
  - *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)*
  - ...

Le corpus Presto

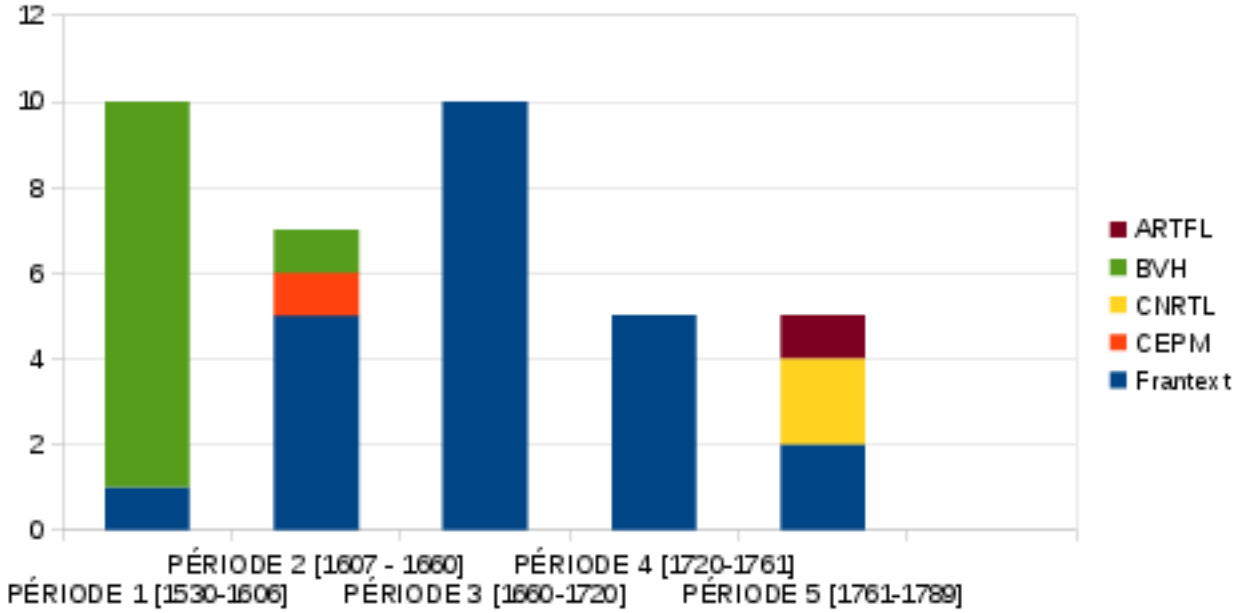
- Besoins
  - Représentation de toutes les périodes de l'histoire du français : 9ème s. au 21ème s.
    - Dans un premier temps : 16ème s. au 21ème s.
  - Présence de différents types de textes et de différents genres discursifs
  - Enrichissement linguistique : étiquetage morpho-syntaxique et lemmatisation
- Collaboration avec diverses bases textuelles existantes
  - *Frantext*
  - *BVH (Bibliothèques Virtuelles Humanistes)*
  - *Université de Cologne*
  - *ARTFL (American and French Research o*
  - *CÉPM (Corpus Électroniques de la Premi*
  - ...

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

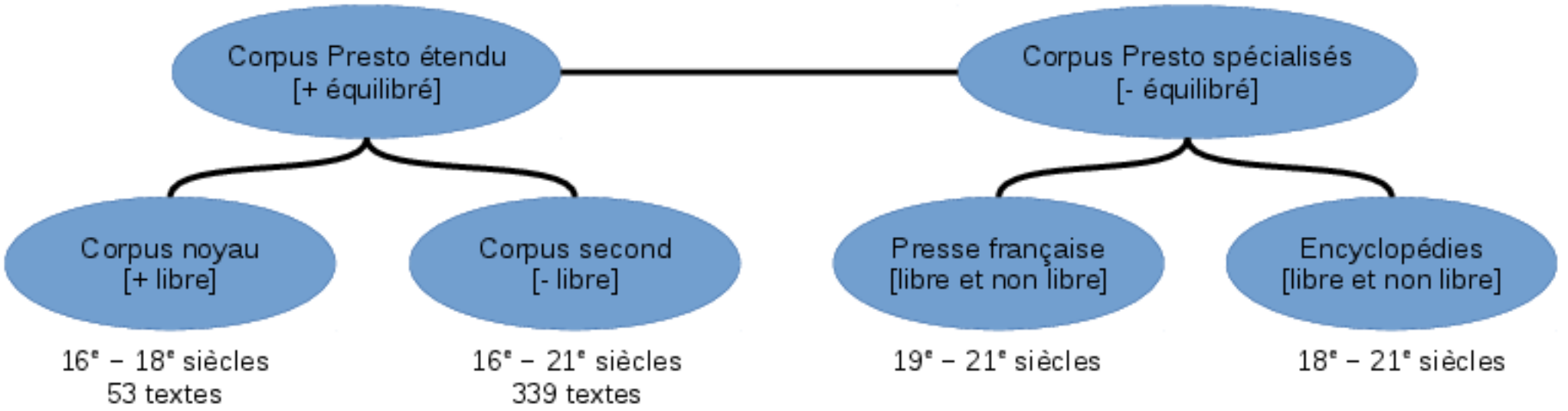
L'Astrée, Gargantua, l'Encyclopédie (vol. 7)...

Le corpus Presto

- Besoins
  - Représentation de toutes les périodes de l'histoire du français : 9ème s. au 21ème s.
    - Dans un premier temps : 16ème s. au 21ème s.
  - Présence de différents types de textes et de différents genres discursifs
  - Enrichissement linguistique : étiquetage morpho-syntaxique et lemmatisation
- Collaboration avec diverses bas
  - *Frantext*
  - *BVH (Bibliothèques Virtuelles Huma*
  - *Université de Cologne*
  - *ARTFL (American and French Rese*
  - *CÉPM (Corpus Électroniques de la*
  - ...



Le corpus Presto



Extrait

(*Sottie pour le cry de la bazoche, Anonyme, 1549*)

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 !

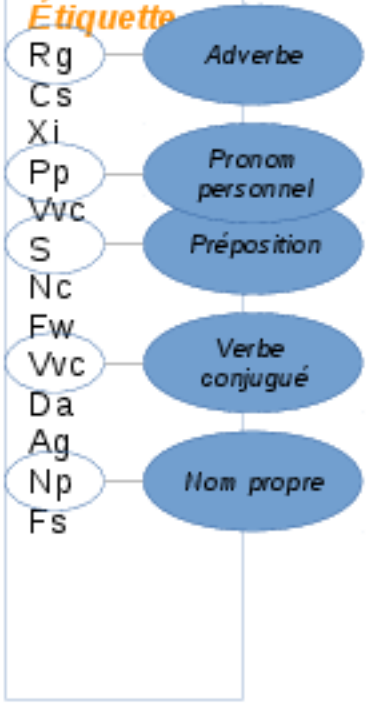
Annotation

Forme

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

Étiquette

Rg
Cs
Xi
Pp
Vvc
S
Nc
Fw
Vvc
Da
Ag
Np
Fs



Désambiguisation par TreeTagger.
Précision 94 %.

Lemme

TANT
QUE
L
ON
DIRE
PAR
HONNEUR
:
VIVRE
LE
EXCELLENT
BAZOCHE
!

Lemme moderne.
Désambiguisation et évaluation à venir.

Sème

Analyse sémantique: en cours pour les noms propres (noms de personnes vs noms de rivières vs noms de villes, etc.)

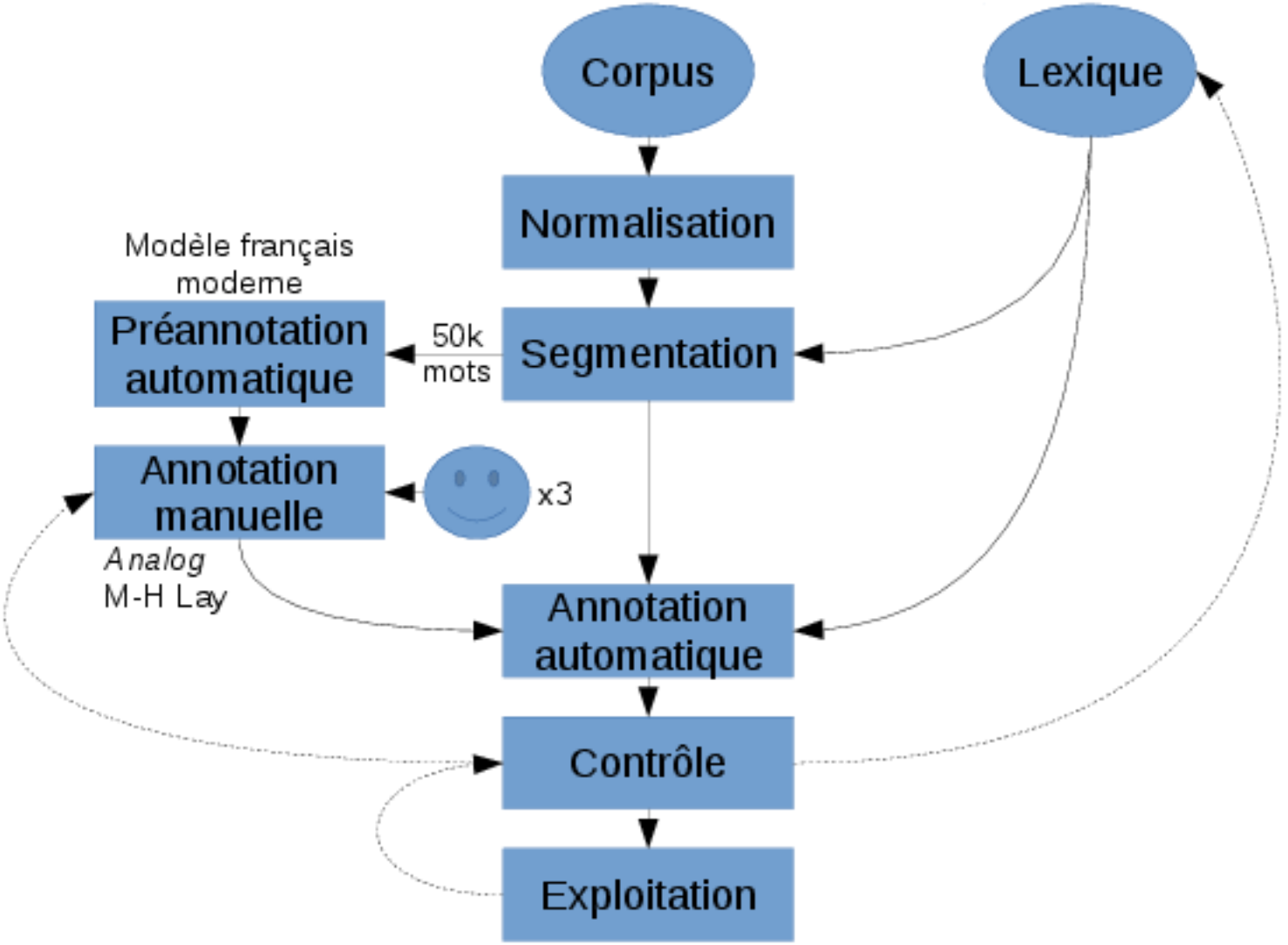
Dépendance

Analyse syntaxique de dépendances : à venir

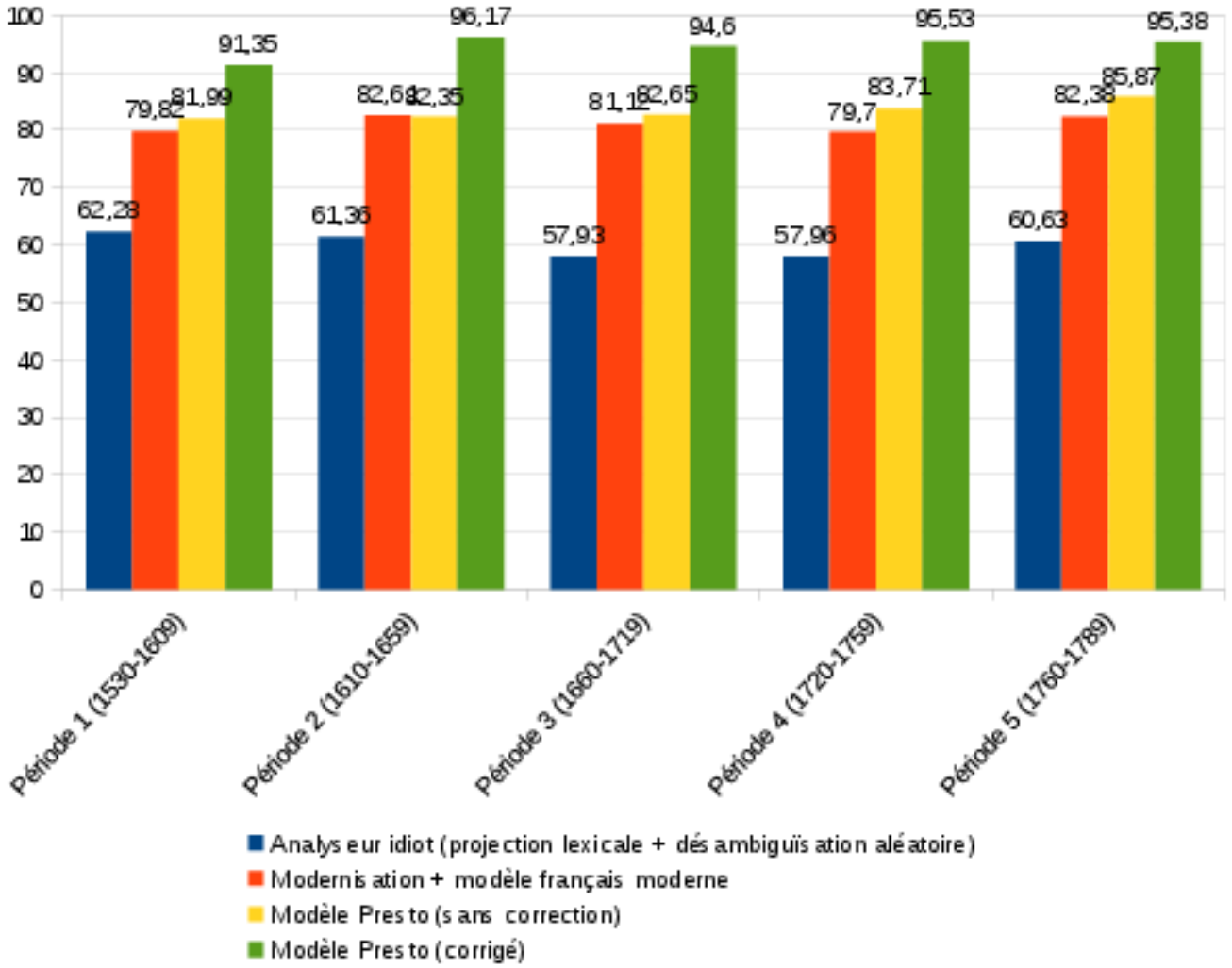
Jeu d'étiquettes

Étiquette (partie du discours 1)	Étiquette (partie du discours 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


Corpus
Lexique
Normalisation
Modèle français
modeme
Préannotation
automatique
50k
mots
Segmentation
Annotation
manuelle
x3
Analog
M-H Lay
Annotation
automatique
Contrôle
Exploitation

Précision


100
90
80
70
60
50
40
30
20
10
0
62,28
79,82
81,99
91,35
61,36
82,68
82,35
96,17
57,93
81,12
82,65
94,6
57,96
79,7
83,71
95,53
60,63
82,38
85,87
95,38
Période 1 (1530-1609)
Période 2 (1610-1659)
Période 3 (1660-1719)
Période 4 (1720-1759)
Période 5 (1760-1789)
Analyseur idiot (projection lexicale + désambiguïsation aléatoire)
Modernisation + modèle français moderne
Modèle Presto (sans correction)
Modèle Presto (corrigé)

Rappel : Principes d'échantillonage

Rappel : Principes d'échantillonage

- Sélection équilibrée de textes selon différents critères (période, genre, auteur)
  - → [+ équilibré] : Corpus de référence PRESTO
  - → [- équilibre] : Corpus spécialisés
- Principes supplémentaires (Corpus de référence PRESTO):
  - statut juridique ([+ libre] / [- libre] de droits)
- Principes supplémentaires (Corpus spécialisés)

Modèles de données et méthodes

CIFRA / Romanisches Seminar
Lehrstuhl für Romanische Sprachwissenschaft (Prof. Dr. Blumenthal)
S. Diwersy

Universität zu Köln
Philosophische Fakultät

Modèles de données

Corpus de référence PRESTO

- Unités structurelles: texte ou receuil de textes, phrase
- Métadonnées:
  - texte : titre, auteur, date, genre
- Propriétés lexicales (mod. PRESTO): mot-forme, lemme, catégorie (2 degrés de granularité), position textuelle (mot → phrase [WS])

Modèles de données

Corpus diachronique de la presse française

- Echantillons de textes journalistiques parus à la fin du 19e et du 20e siècle; différenciation supplémentaire entre presse nationale et régionale
  - 19e s.: Le Figaro, le JDD (19e s.), Le Reveil de Lyon, etc.
  - 20e s.: Le Figaro, Le Monde, L'Est Républicain, Sud Ouest
- Unités structurelles: numéro de journal, phrase
- Métadonnées:
  - numéro : journal, année
- Propriétés lexicales (mod. *TreeTagger* fr.contemp.): mot-forme, lemme, catégorie (2 degrés de granularités), position textuelle (WS)

Modèles de données

Corpus de la presse française contemporaine

- **Echantillons de textes journalistiques parus à la fin du 19e et du 20e siècle; corpus de test / de démo**
  - **Le Monde (+ Le Figaro, Libération, Sud Ouest, Ouest France etc)**
- **Unités structurelles: année de journal, article, paragraphe, phrase**
- **Métadonnées:**
  - **article : titre, auteur, date, section thématique**
  - **paragraphe : position textuelle (PT)**
  - **phrase : position textuelle (ST, SP)**
- **Propriétés lexicales (mod. *Connexor*): mot-forme, lemme, catégorie (2 degrés de granularités), traits morphologiques, relation de dépendance syntaxique, fonction syntaxique, positions textuelles (WT, WP, WS)**

Modèles de données

Corpus des Encyclopédies

- Articles de l'*Encyclopédie de Diderot et d'Alembert* (18e s.) et de l'*Encyclopaedia Universalis* (20e s.)
- Unités structurelles: ouvrage, article, paragraphe, phrase
- Métadonnées:
  - article : lemme (entrée), auteur, domaine thématique
  - paragraphe : position textuelle (PT)
  - phrase : position textuelle (ST, SP)
- Propriétés lexicales (mod. PRESTO): mot-forme, lemme, catégorie (2 degrés de granularités), positions textuelles (WT, WP, WS)

Modèles de données

Bilan – Propriétés lexicales

Base textuelle	Forme	Catégorie	Lemme	Traits	Relation syntaxique	WT	WP	WS
Corpus de référence PRESTO	x	x	x	-	-	-	-	x
Corpus diachronique de la presse française	x	x	x	x	-	-	-	(x)
Presse française contemporaine	x	x	x	x	x	x	x	x
Encyclopédies	x	x	x	-	-	x	x	x
Emolex-FR	x	x	x	x	x	x	x	x
CoVaNa-FR	x	x	x	x	x	-	-	-

Visées méthodologiques

Lexical Priming (cf. Hoey 2005) – Modèle descriptif

Niveau cooccurrentiel	Niveau de schématicité distributionnelle
cooccurrences lexicales	Profil combinatoire
associations sémantiques / pragmatiques	
colligations	
collocations textuelles (chaînes lexicales)	Profil d'intégration textuelle
associations sémantiques textuelles (argumentatives)	
colligations textuelles	

Visées méthodologiques

Lexical Priming – Mise en oeuvre du modèle descriptif

Base textuelle	LCOLLOC	LSEM	LCOLLIG	TCOLLOC	TSEM	TCOLLIG
Corpus de référence PRESTO	x	(x)	(x)	-	-	(x)
Corpus diachronique de la presse française	x	(x)	(x)	-	-	(x)
Presse française contemporaine	x	x	x	-	-	x
Encyclopédies	x	(x)	(x)	-	-	x
Emolex-FR	x	x	x	-	-	x
CoVaNa-FR	x	x	x	-	-	-

Outils (*PrimeStat*.BTLC)

3.1 Boîte à outils *PrimeStat*

- Concordances KWIC
- Index (fréquences, spécificités)
- Analyse cooccurrentielle (cooccurrences lexicales, lexico-syntaxiques, colligations syntaxiques, colligations textuelles)
- Analyses multivariées (CAH, K-Means, AFC)

3.2 Plateforme *BTLC*

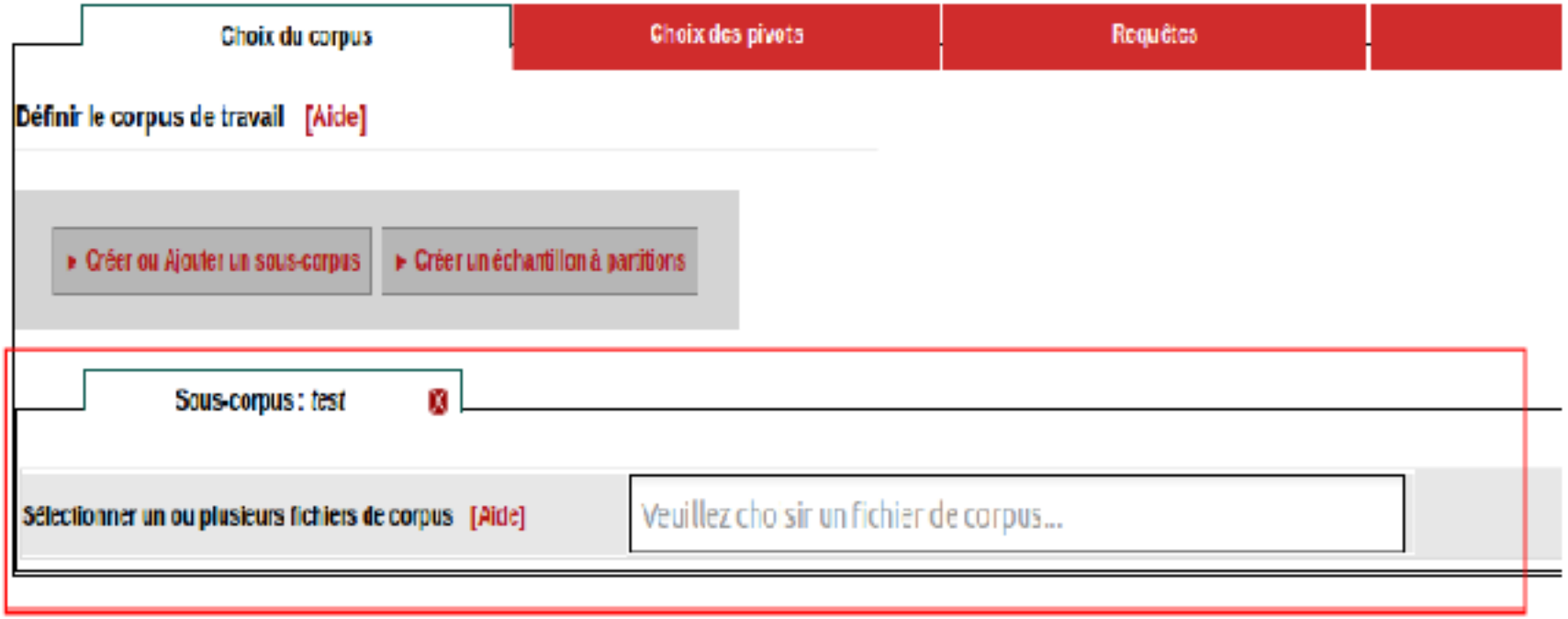
Adresse du site: http://persan.rom.uni-koeln.de/btlsc/

BTLC.*PrimeStat* – Interface graphique

- Définition d'objets de requête
  - Corpus de travail (sous-corpus / partition)
  - Expression de requête (→ pivot)
- Applications
  - Concordances KWIC
  - Calculs de fréquence (spécificités)
  - Cooccurrences lexico-syntaxiques
  - etc.

3.2 *BTLC.PrimeStat* – Interface graphique

Définition du corpus de travail

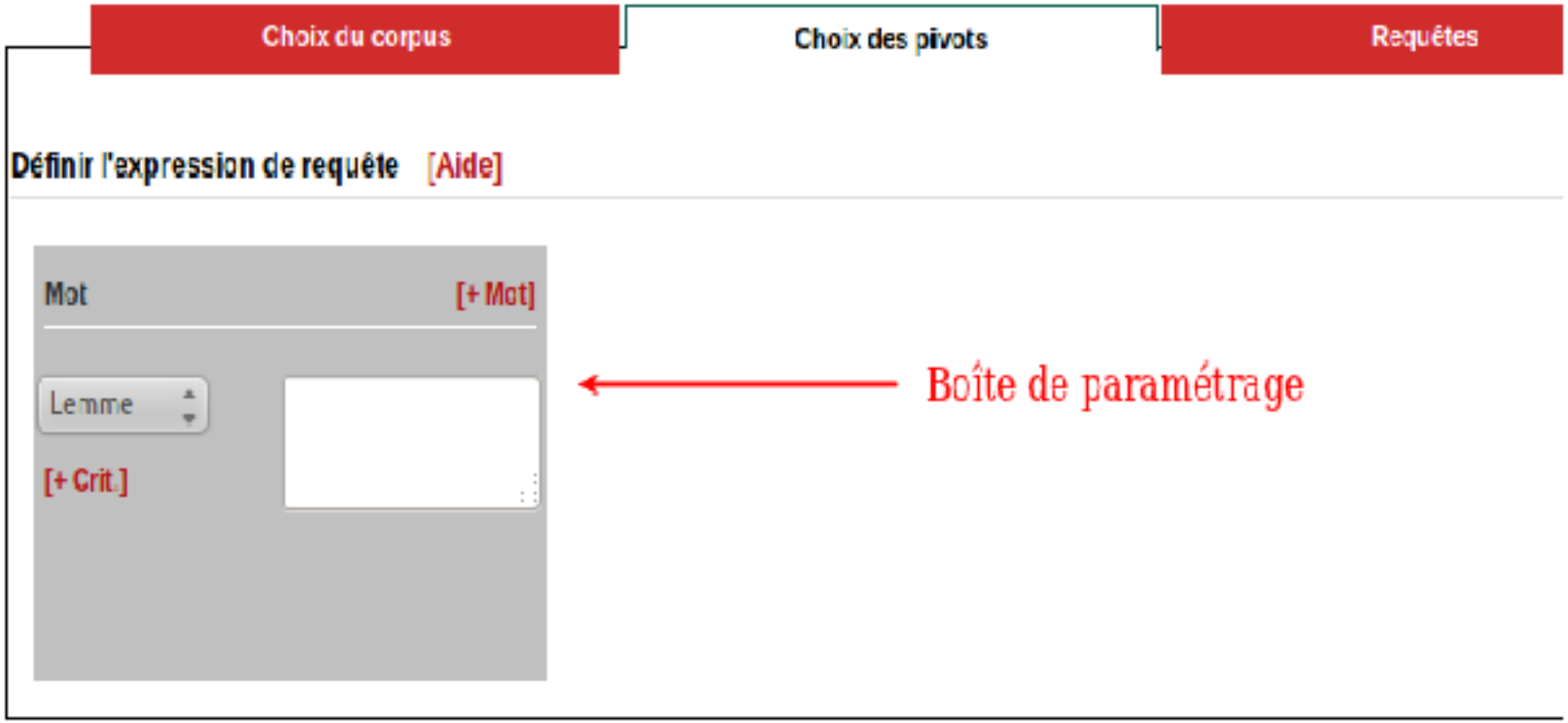


3.2 *BTLC.PrimeStat* – Interface graphique

Définition du corpus de travail

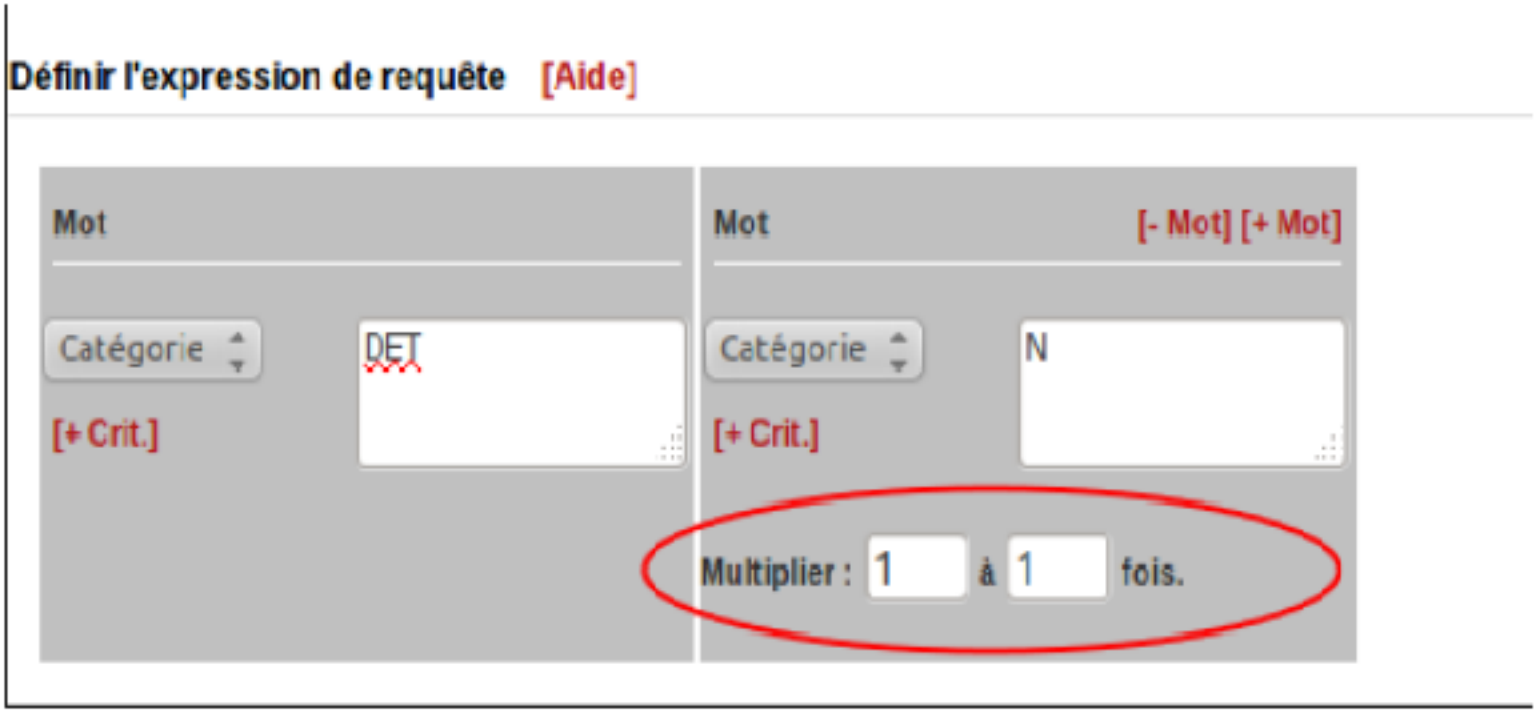
3.2 *BTLC.PrimeStat* – Interface graphique

Définition de l'expression de requête



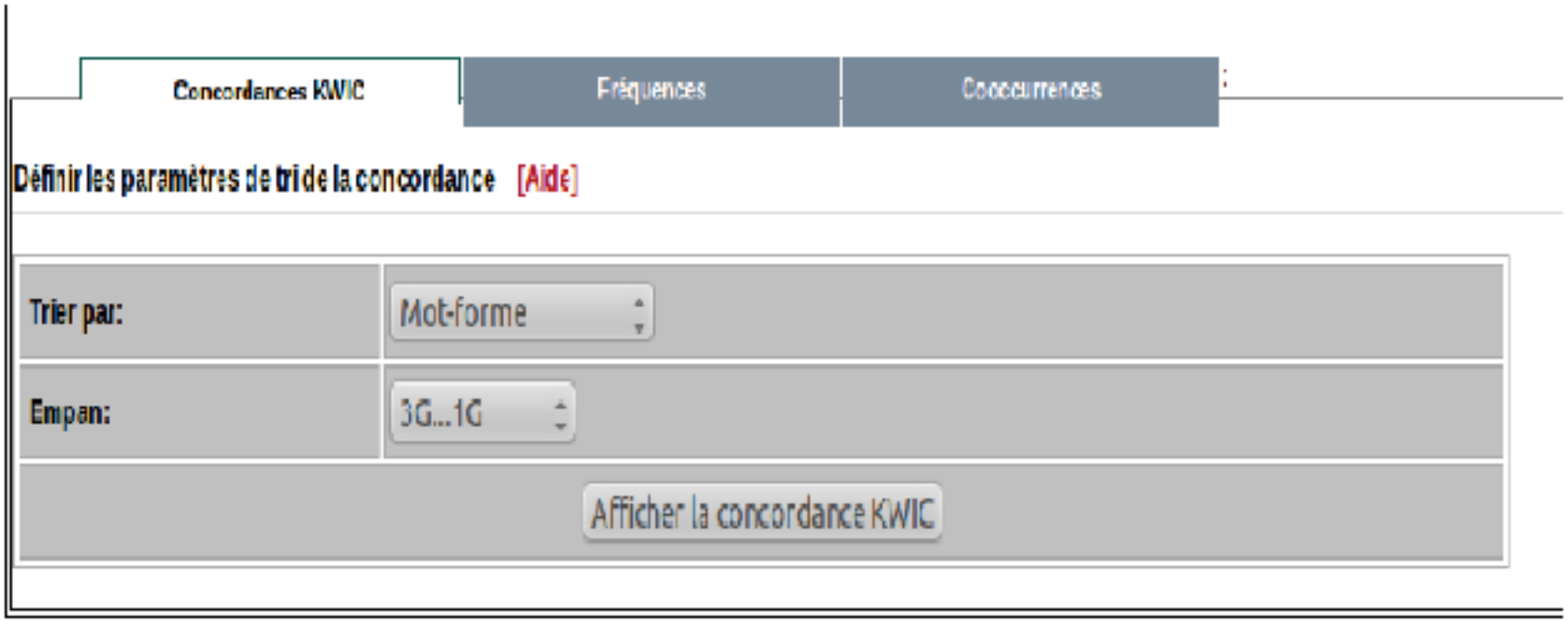
3.2 *BTLC.PrimeStat* – Interface graphique

Définition de l'expression de requête



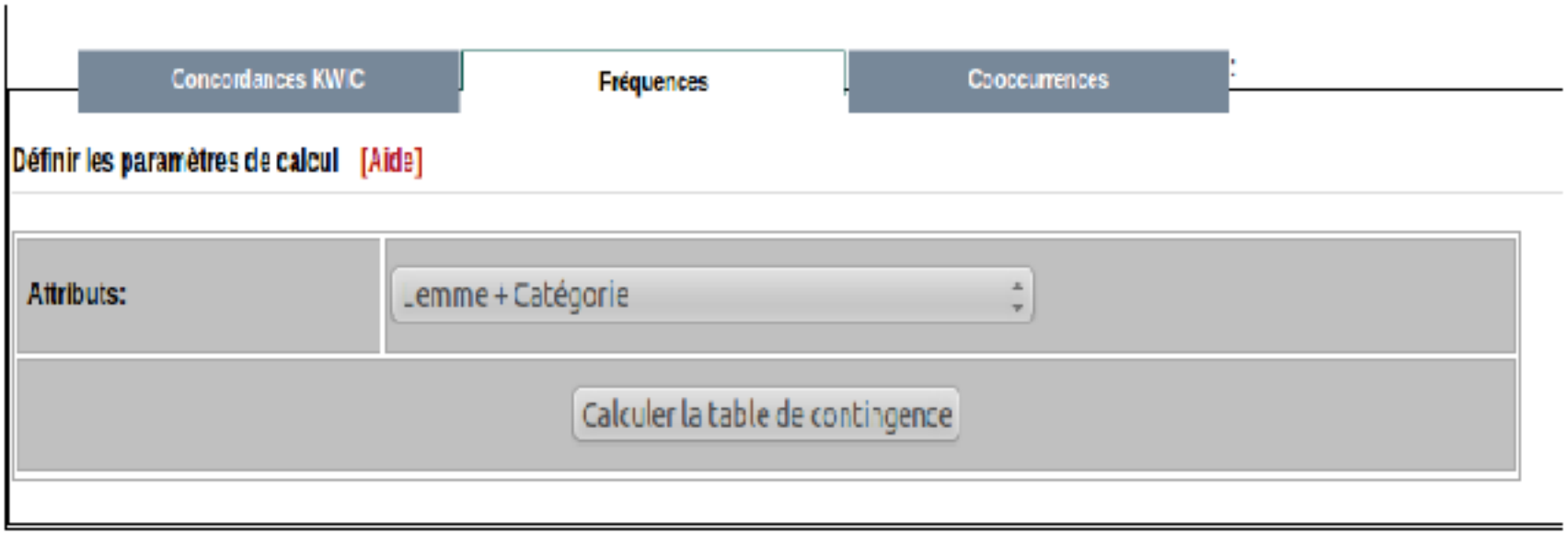
3.2 *BTLC.PrimeStat* – Interface graphique

Applications: Concordances KWIC



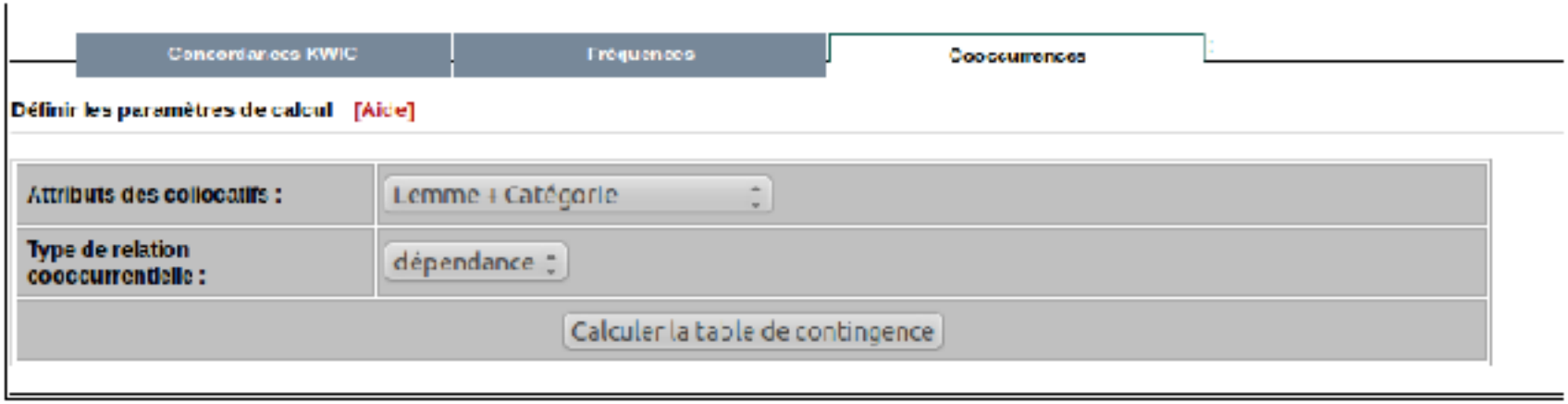
3.2 *BTLC.PrimeStat* – Interface graphique

Applications: Calculs de fréquence



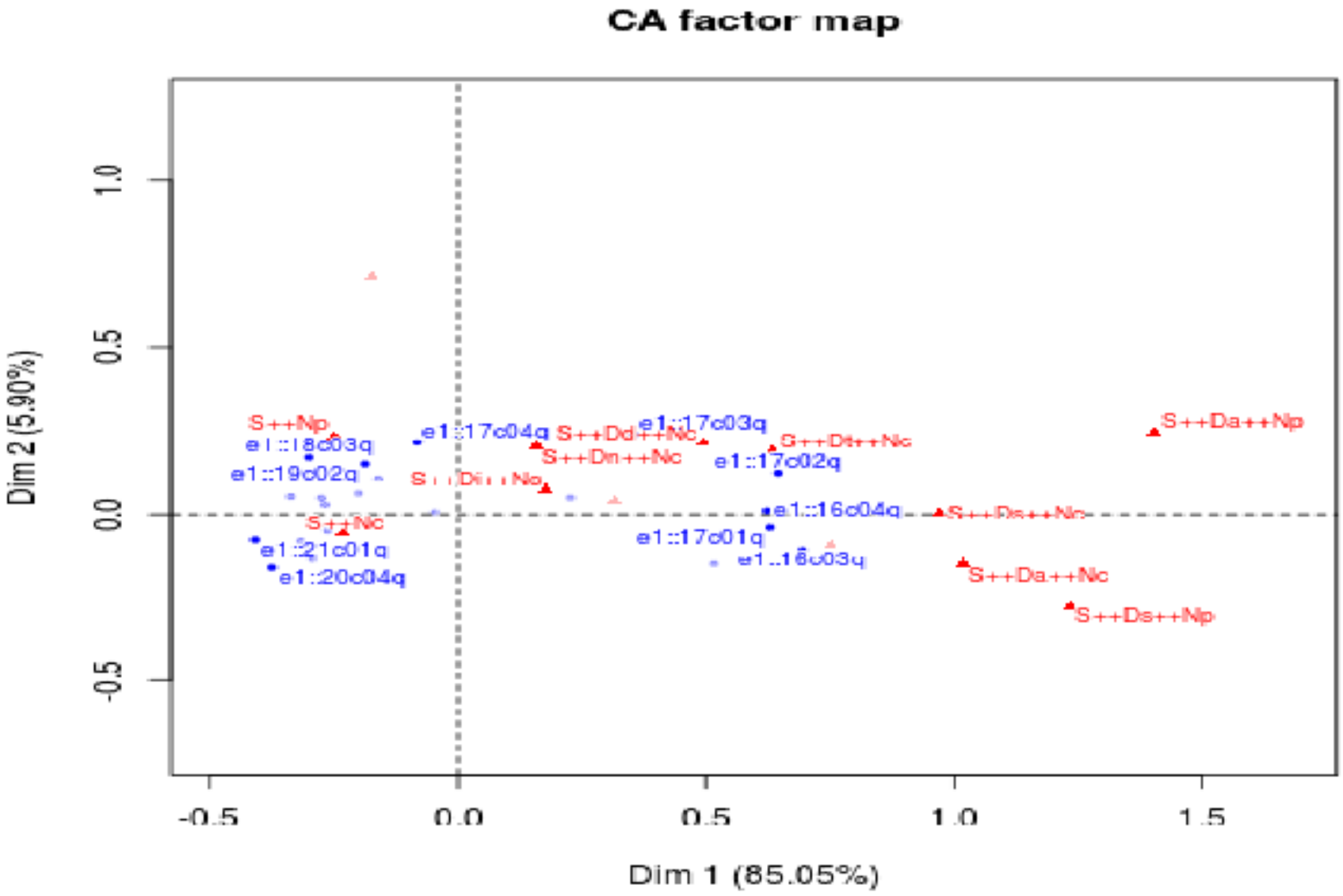
3.2 *BTLC.PrimeStat* – Interface graphique

Applications: Cooccurrences lexico-syntaxiques



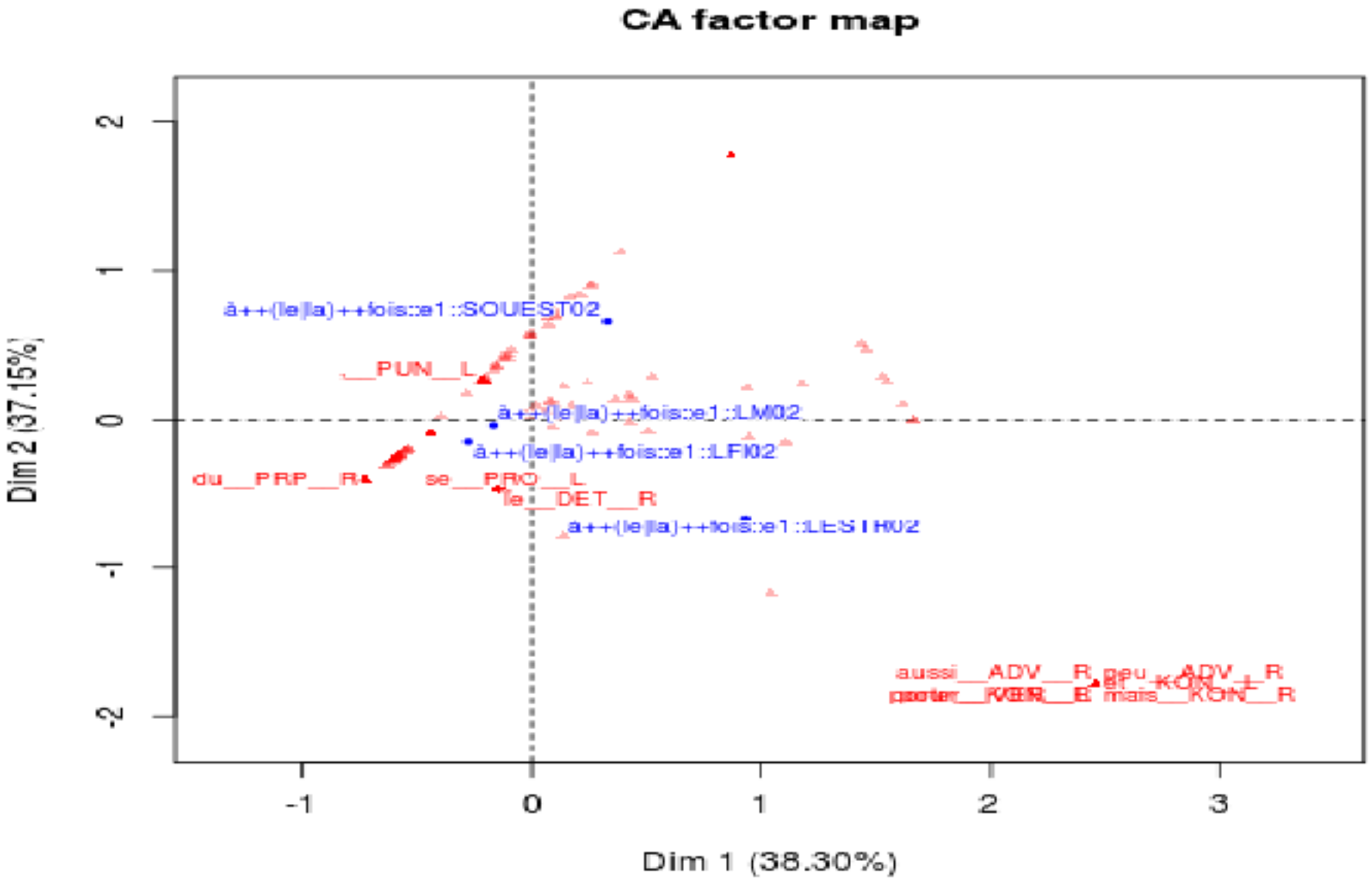
3.2 *BTLC.PrimeStat* – Applications

AFC – ***en*** (+ D) + N



3.2 *BTLC.PrimeStat* – Applications

AFC – *à la fois* (Presse 20e s.)



Perspectives
-
Vers une linguistique diachronique outillée

CIFRA / Romanisches Seminar
Lehrstuhl für Romanische Sprachwissenschaft (Prof. Dr. Blumenthal)
S. Diwersy

Universität zu Köln
Philosophische Fakultät

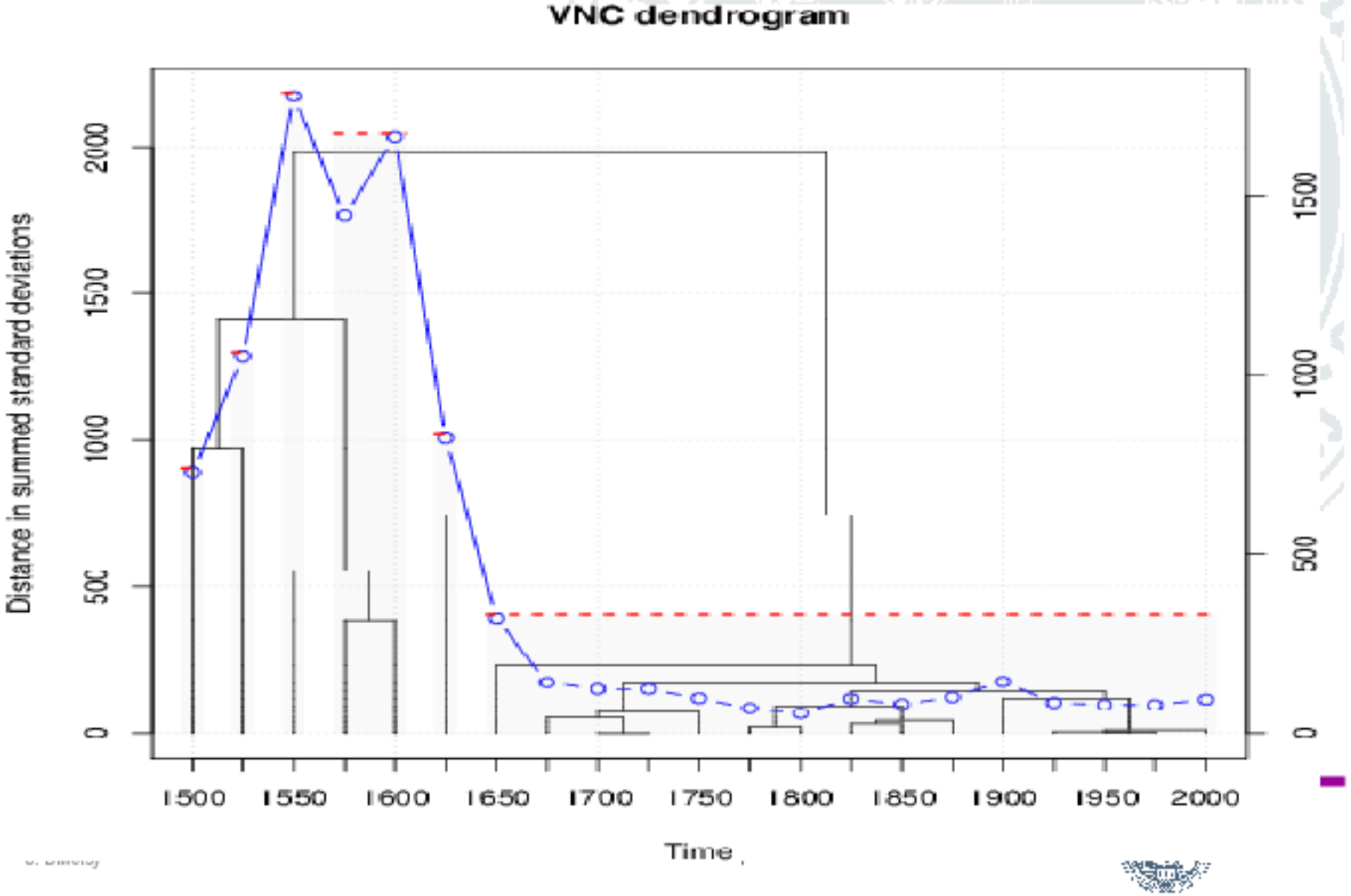
Vers une linguistique diachronique outillée

Méthodes d'exploitation de corpus dans une perspective diachronique

- **VNC (*Variability-based Neighbour Clustering*: Gries & Hilpert 2008, Hilpert & Gries 2009, Hilpert 2013)**
- **Calculs de productivité (Hilpert 2012)**
- **HCFA (*Hierarchical Configural Frequency Analysis*: Gries 2004, Hilpert 2013)**

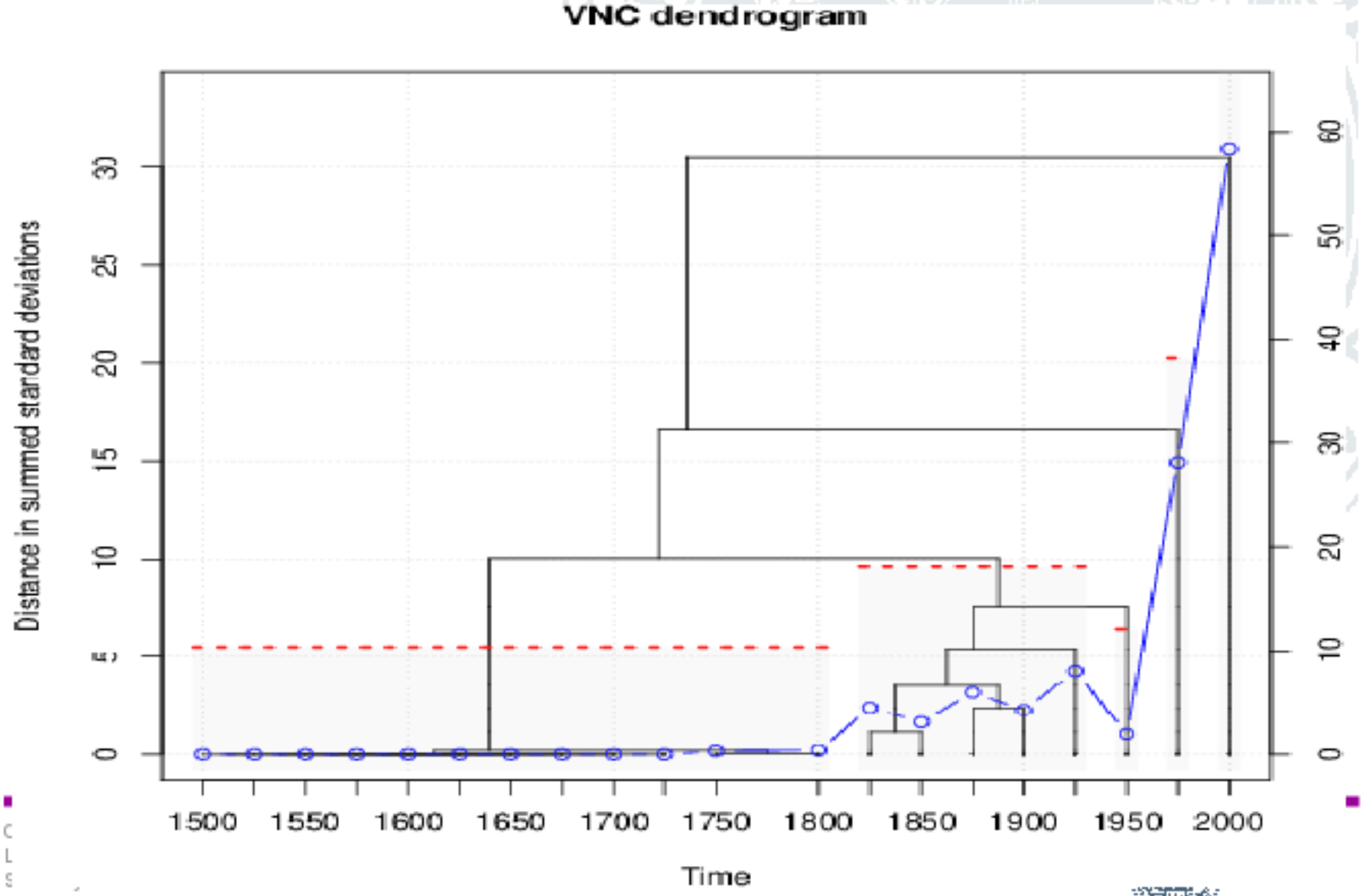
Vers une linguistique diachronique outillée

VNC appliqué sur un motif : *en* + D + N



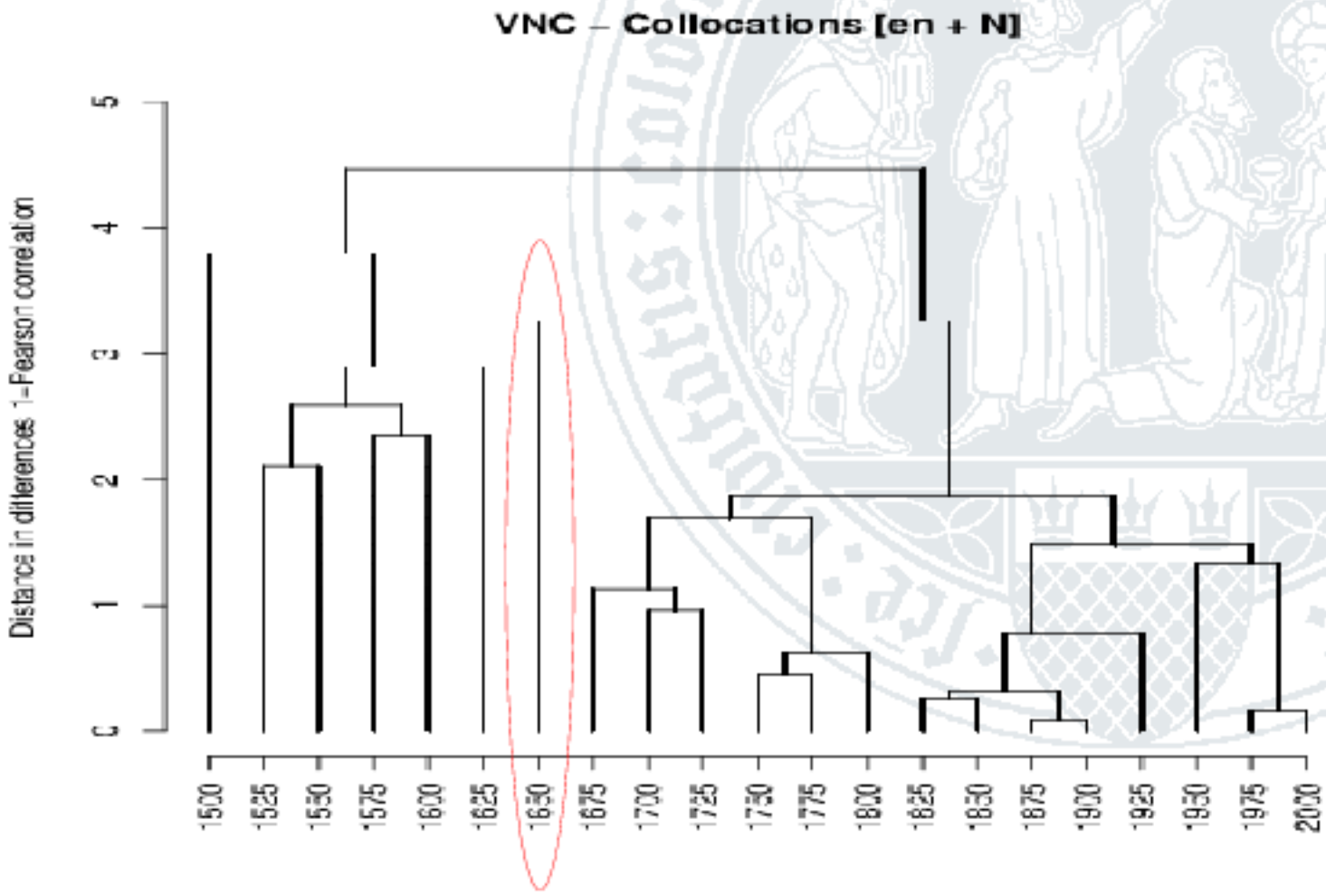
Vers une linguistique diachronique outillée

VNC appliqué sur une locution – *dans le cadre de*



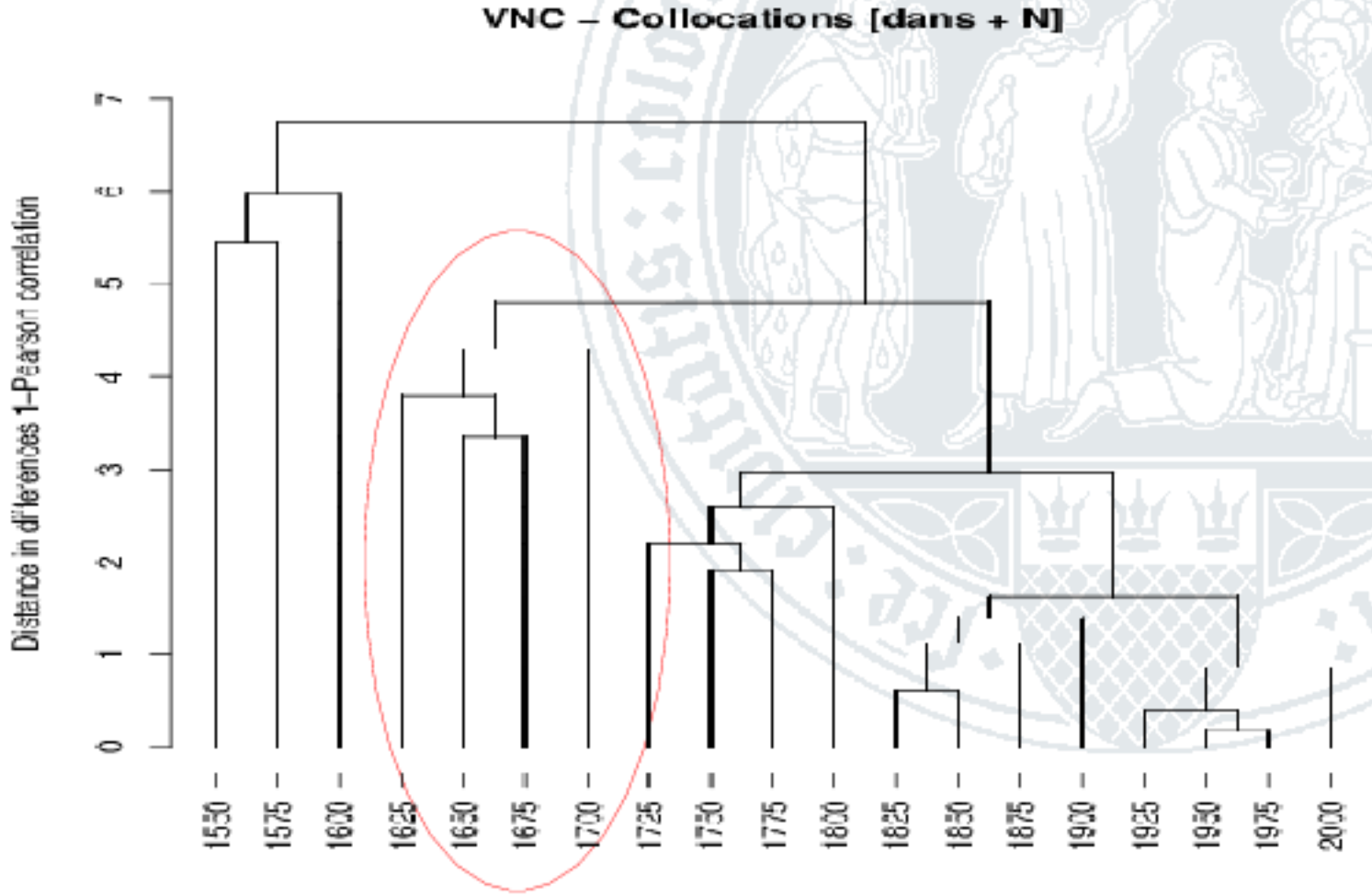
Vers une linguistique diachronique outillée

VNC appliqué sur un inventaire cooccurrentiel – *en* + N



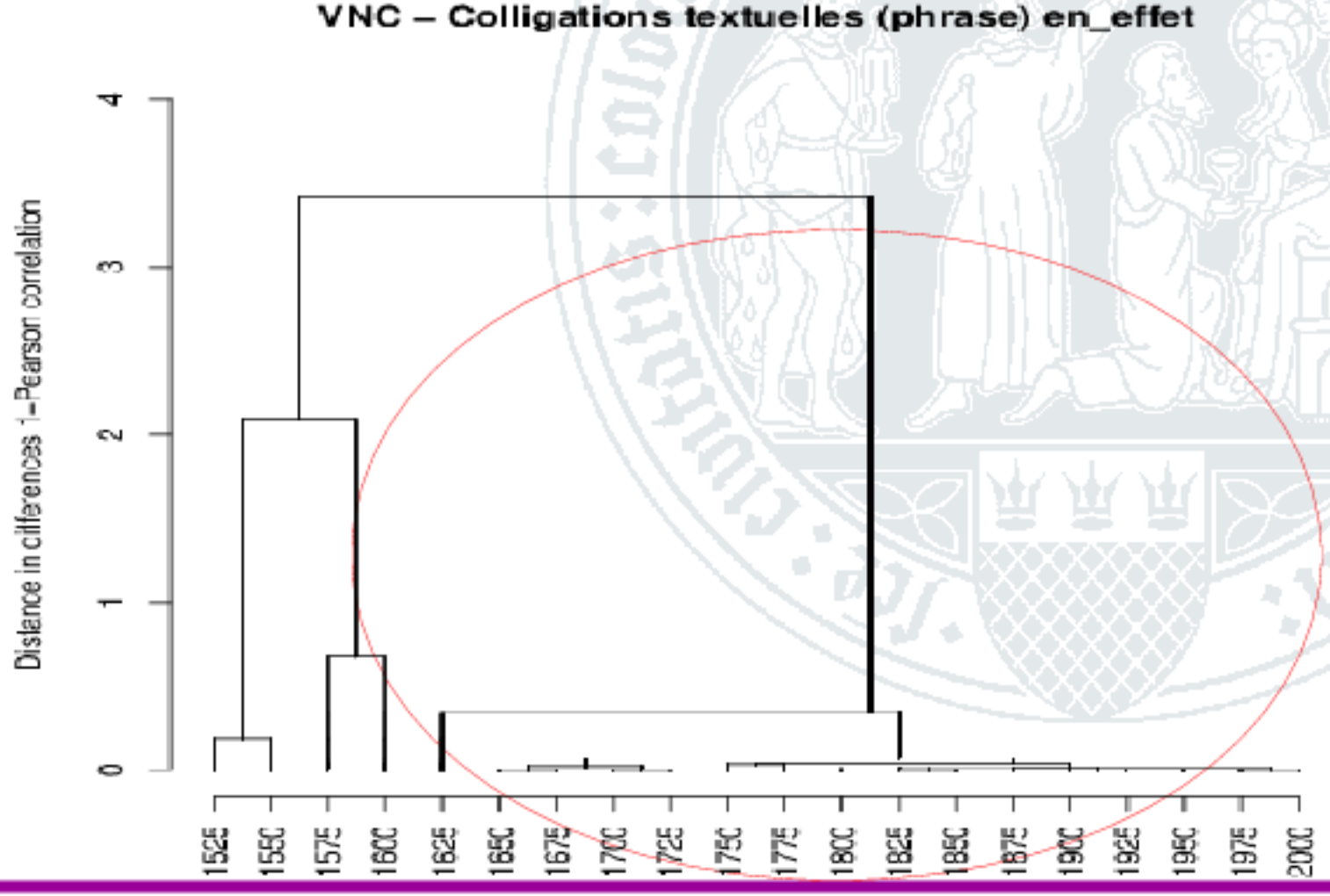
Vers une linguistique diachronique outillée

VNC appliqué sur un inventaire cooccurrentiel – *dans* + N



Vers une linguistique diachronique outillée

VNC appliqué sur l'association préférentielle à des positions texuelles – *en effet* (en début, milieu, fin de phrase)

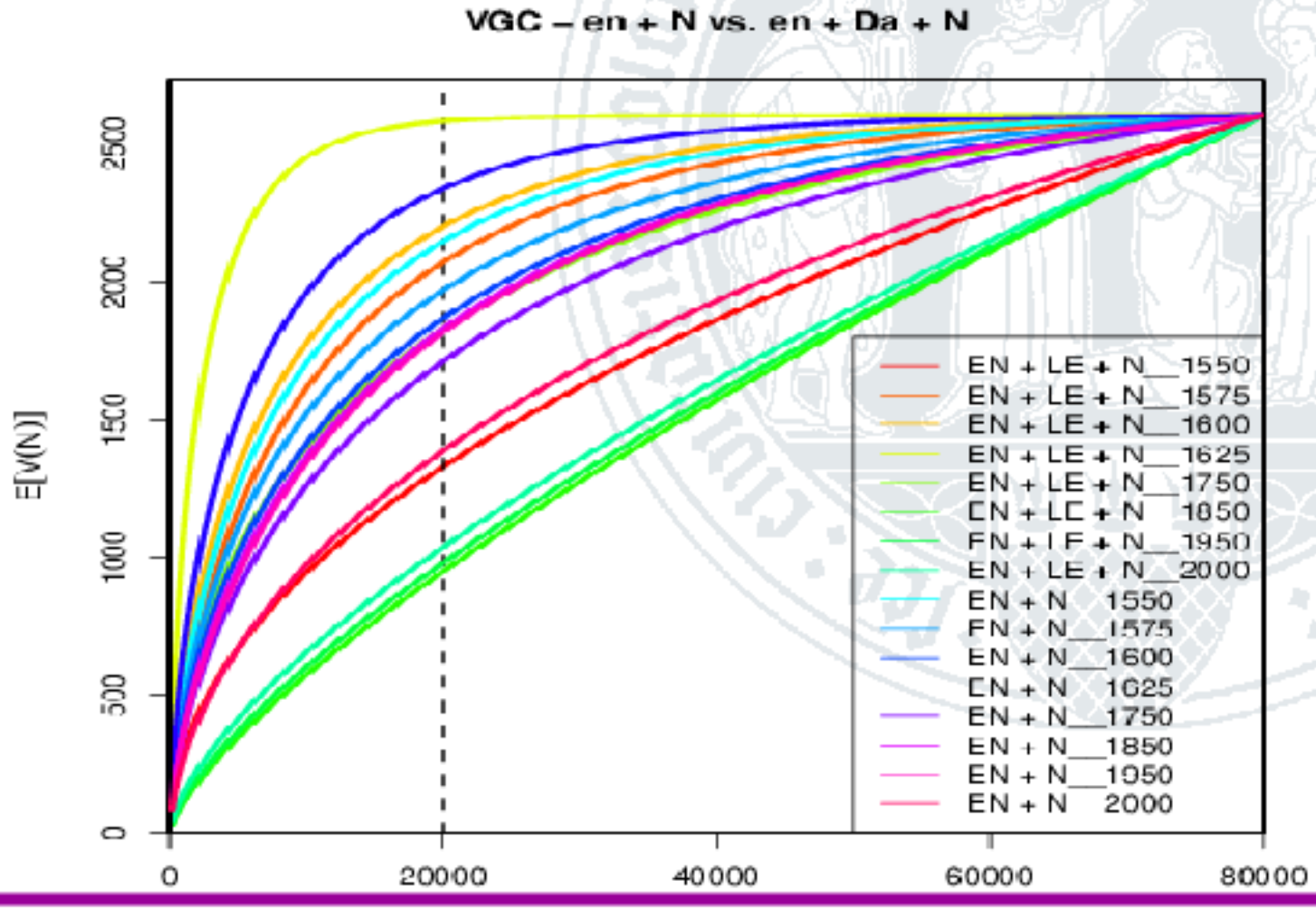


Vers une linguistique diachronique outillée

- Etablir, de façon inductive, la saturation d'une position paradigmatique au sein d'un motif
- Analyses quantitatives de la productivité
  - morphologique (Baayen & Lieber 1991, Baayen 1992, Evert & Lüdeling 2005)
  - (morpho-)syntaxique (Zeldes 2012, Desaguiler à paraître)
- Méthodes de calcul et outils
  - modélisation LNRE (Baayen 2001, Evert 2004)
  - zipfR (Evert & Baroni 2006, 2007)

Vers une linguistique diachronique outillée

Courbes d'accroissement lexical (Vocabilary Growth Curves) – *en* + N vs. *en* + Da + N



Vers une linguistique diachronique outillée

HCFA – colligations textuelles de *en effet* (selon période, genre)

l1	l2	time	genre	Freq	Exp	Cont.chisq	Obs-exp	Dec
EN EFFET	WS_B	2000	encyclopédie	247	41,5909	1014,4743	>	***
EN EFFET	WS_B	1875	traité	518	326,7645	111,9186	>	***
EN EFFET	WS_B	1750	encyclopédie	195	34,2458	754,6009	>	***
EN EFFET	WS_B	1800	traité	377	239,009	79,6686	>	***
EN EFFET	WS_B	2000	presse	178	52,3541	301,5407	>	***
EN EFFET	WS_B	1675	traité	264	141,265	106,6356	>	***
EN EFFET	WS_B	1975	presse	157	46,4102	263,5219	>	***
EN EFFET	WS_B	1900	traité	264	158,7448	69,7891	>	***
EN EFFET	WS_B	1825	traité	330	245,7868	28,8537	>	***
EN EFFET	WS_B	1850	traité	292	204,0495	37,9089	>	***

Vers une linguistique diachronique outillée

HCFA – colligations textuelles de *en effet* (selon période, genre)

l1	l2	time	genre	Freq	Exp	Cont.chisq	Obs-exp	Dec
EN EFFET	WS_B	1975	traité	195	275,7521	23,6477	<	***
EN EFFET	WS_B	1750	narratif	10	78,5027	50,7765	<	***
EN EFFET	WS_B	1875	narratif	57	100,151	18,392	<	**
EN EFFET	WS_B	1950	traité	77	125,5689	18,786	<	**
EN EFFET	WS_M	1975	traité	30	79,2768	30,6294	<	***
EN EFFET	WS_B	1675	narratif	3	43,2967	37,5046	<	***
EN EFFET	WS_B	1825	narratif	34	75,3319	22,6773	<	***
EN EFFET	WS_B	1825	presse	5	41,3069	31,9712	<	***
EN EFFET	WS_B	1975	traité	9	50,4256	34,0319	<	***
EN EFFET	WS_B	1900	narratif	14	48,6542	24,6826	<	***
EN EFFET	WS_M	1750	narratif	1	22,569	20,6133	<	***

1	Je	Pp	JE
2	vous	Pp	VOUS
3	remercie	Vvn	REMERCIER
4	de	S	DE
5	votre	Ds	VOTRE
6	attention	Nc	ATTENTION
7	!	Fs	!