XXVIII Congresso Internazionale di Linguistica e Filologia Romanza

Roma 18 - 23 luglio 2016

SEZIONE LINGUISTICA DEI CORPORA E FILOLOGIA INFORMATICA

PaLaFra:
Un projet de corpus numérique pour promouvoir la recherche sur le passage du Latin au Français

www.palafra.org

rembert.eufe@uni-tuebingen.de
sebastian.ortner@uni-tuebingen.de

La conception du projet

PaLaFra: Le **Pa**ssage du **La**tin au **Fra**nçais

objectif:

- intensifier et faciliter la recherche empirique concernant
  - le passage du latin tardif au très ancien français
  - l'écart conceptuel entre la latinité tardive et le proto-roman reconstruit à partir des langues romanes

- à l'aide d'un corpus de latin et de français à créer par
  - constitution d'un sous-corpus de latin tardif (VI^{e}-VIII^{e} siècles)
  - perfectionnement d'un sous-corpus de très ancien français existant
  - développement d'un système de requête et d'exploitation commun
  - élaboration d'un sous-corpus aligné composé de textes latins et de leur traductions en ancien français

Les partenaires de PaLaFra

- approche interdisciplinaire: collaboration de spécialistes en
  - linguistique et philologie romane et latine
  - MITIC (médias, images et technologies de l'information et de la communication, engl. *media informatics*) et textométrie

- projet binational franco-allemand
  - financement par l'*Agence nationale de la recherche française* (ANR) et la *Deutsche Forschungsgemeinschaft* (DFG)

- collaboration avec les *Monumenta Germaniae Historica* (MGH)

Les partenaires de PaLaFra

trois équipes:

- équipe de Lyon
  - Céline Guillot, Serge Heiden, Alexei Lavrentiev, Naomi Kanaoka
  - tâches principales:
    - mise à disposition et perfectionnement du corpus ancien français
    - mise à disposition et perfectionnement de la plateforme pour l'analyse textométrique TXM et formation à son utilisation

- équipe de Lille
  - Anne Carlier
  - tâches principales:
    - test du corpus
    - approfondissement des questions de recherche
    - élaboration et finalisation des résultats en vue de leur publication

Les partenaires de PaLaFra

- équipe de Ratisbonne et de Tübingen
  - Maria Selig, Christian Wolff, Lars Döhling, Rembert Eufe, Sebastian Ortner, Elisabeth Reichle
  - tâches principales:
    - constitution du corpus latin-tardif
    - collecte et enregistrement des métadonnées
    - annotation automatique et correction manuelle des résultats
    - développement d'un jeu d'étiquettes commun supplémentaire pour les deux sous-corpus latin et ancien français
    - contrôle et perfectionnement de l'utilisabilité des outils

Les sous-corpus

Le sous-corpus ancien français

47 textes de la *Base de Français Médiéval* (BFM)

- http://bfm.ens-lyon.fr/

- textes du IX^{e} au XV^{e} siècle
  - classés selon 5 domaines
    - domaine religieux, p. ex. *Séquence de sainte Eulalie*, *Vie de saint Léger*, *Vie de saint Alexis*, *Vies de sainte Geneviève*, *Passion de Clermont*
    - domaine littéraire, p. ex. *Chanson de Roland*, *Gormont et Isembart*, *Yvain*
    - domaine didactique, p. ex. *Comput* et *Bestiaire* de Philippe de Thaon, *Roman de la Rose*
    - domaine historique, p. ex. *Roman de Brut*, *Conquête de Constantinople*
    - domaine juridique, p. ex. *Serments de Strasbourg*, 3 chartes

- textes traités en partie dans le cadre du projet *Corpus représentatif des premiers textes français* (CoRPTeF, http://corptef.ens-lyon.fr/)
  - annotation suivant le système Cattex (http://bfm.ens-lyon.fr/spip.php?article176)
  - lemmatisation en cours

Le sous-corpus latin tardif

manque d'un corpus numérisé du latin tardif en Gaule adapté à la recherche en linguistique et philologie romanes

- problème du nombre limité des sources de l'époque mérovingienne
- éditées en majorité dans les *Monumenta Germaniae Historica*

→ coopération avec les MGH pour le traitement de leurs textes

- avantage: gestion des droits d'auteurs
- désavantage: éditions datant de la fin du XIX^{e} et du début du XX^{e} siècle souvent héritières de la philologie du XIX^{e} s., visant à reconstruire des versions originales fictives sur la base des préconceptions des éditeurs

  - mais éditions dotées d'apparats critiques très sérieux permettant de rétablir les versions manuscrites (Van Acker 2007: 205)
    - dans beaucoup de cas, toujours les éditions de référence
    - consultables en ligne (www.dmgh.de)

Monumenta Germaniae Historica
BSB Bayerische StaatsBibliothek
Deutsche Forschungsgemeinschaft
DFG

Le sous-corpus latin tardif

choix d'environ 200 textes latins

- grand nombre de textes courts dans les collections de lettres et les formulaires
- datés du V^{e} au IX^{e} siècle
- genres et domaines comparables à ceux du sous-corpus français
  - religieux, hagiographie
  - historique, historiographie
  - historique, épistolaire
  - juridique, législatif
  - juridique, formulaire

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Textes latins

Vies de saints	religieux, hagriographie	~143 309 tokens	29 vies de saints	6ème – 8ème siècle	MGH SS rer. Merov 3-7
Grégoire de Tours, X libri, VIème livre	historique, historiographie	120 000 en tout, livre VI ca. 16 000 tokens	~50 mss., le plus ancien date du 7ème s., seulement livre VI	fin 6ème s.	MGH SS rer. Merov 1,1
Frédégaire, IVème livre	historique, historiographie	52 500 mots en tout, dont 7300 des continuationes	38 mss.; B. Krusch se base sur un manuscrit de la fin du 7ème siècle, provenant de Metz	fin 7ème s.	MGH SS rer. Merov. 2
Fréd., continuationes	historique, historiographie	ca. 7300	B. Krusch	8ème s.	MGH SS rer. Merov. 2
Liber Historia Francorum	historique, historiographie	ca. 14 000	31 mss., B. Krusch	début 8ème s.	MGH SS rer. Merov. 2
Epistolae Austrasiacae	historique, épistolaire	~ 286 193 tokens (~ 45309 types)	48 lettres, W. Gundlach	fin 6ème s.	MGH Ep. Merow. et Karol. aevi
Desiderii episcopi Cadurcensis epistolae	historique, épistolaire		36 lettres, B. Krusch	7ème s.	MGH Ep. Merow. et Karol. aevi
Epistolae aevi Merovingici collectae	historique, épistolaire		19 lettres, W. Gundlach	5ème – 7ème s.	MGH Ep. Merow. et Karol. aevi

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Textes latins

Formulae Andecavenses	juridique	formulaires	~ 294 926 tokens (~ 32624 types)	1mss., K. Zeumer	début 8ème s.	MGH Form. Merow. et Karol. aevi
Formulae Marculfi	juridique	formulaires		6 mss., K. Zeumer	fin 7ème s.	MGH Form. Merow. et Karol. aevi
Formulae Turononenses	juridique	formulaires		6 mss., K. Zeumer	8ème – 9ème s.	MGH Form. Merow. et Karol. aevi
Formulae Bituricenses	juridique	formulaires		3 mss., K. Zeumer	début 8ème s.	MGH Form. Merow. et Karol. aevi
Formulae Pithoei	juridique	formulaires		1 ms., K. Zeumer	8ème s.	MGH Form. Merow. et Karol. aevi
Formulae Salica Bignonanae	juridique	formulaires		formulaires transmis pour la plupart par un seul mss., K. Zeumer	7ème s.	MGH Form. Merow. et Karol. aevi
Formulae Salica Merkelianae	juridique	formulaires		formulaires transmis pour la plupart par un seul mss., K. Zeumer	fin 8ème s.	MGH Form. Merow. et Karol. aevi
Formulae Salica Lindenbrogianae	juridique	formulaires		formulaires transmis pour la plupart par un seul mss., K. Zeumer	fin 8ème s.	MGH Form. Merow. et Karol. aevi

Textes latins

Pactus legis Salicae	juridique	législatif	~ 20 000 tokens estimés	version mérovingienne de la Lex Salica, 87 mss., K. A. Eckhardt	6ème – 7ème siècle	MGH LL nat. Germ., 4,1: Pactus legis Salicae
Lex Ribuaria	juridique	législatif	~ 15 000 tokens estimés	35 mss., K. Beyerle/ R. Buchner	623-650	MGH LL nat. Germ., 3,2: Lex Ribvaria
Chartes mérovingiennes	juridique	chartes	ca. 12 000	T. Kölzer; seulement les chartes conservées sous forme d'originaux	625-768	MGH DD Mer. 1
Chartes des Arnulfiens	juridique	chartes	< 2 000 tokens estimés	I. Heidrich; seulement les chartes conservées sous forme d'originaux (no.s 22, 23)	751	MGH DD Arnulf.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Fiches descripteurs – base de données

ID	titre	domaine_discursif	genre_discursif	aut_nom	aut_surnom	destinataire
1	Visio Baronti Monachi Longoretensis	religieux	hagiographie, mira	anonyme	NULL	NULL
2	Vita Adelphii Abbatum Habendensis	religieux	hagiographie	anonyme	NULL	NULL
3	Vita Amandi episcopi I.	religieux	hagiographie	anonyme	NULL	NULL
4	Vita Amati	religieux	hagiographie	anonyme	NULL	NULL
5	Vita Audoini episcopi rotomagensis	religieux	hagiographie	anonyme	NULL	NULL
9	Vita Desiderii Carducae urbis episcopi	religieux	hagiographie	un moine an	NULL	NULL
10	Vita Eligii episcopi Noviomagensis	religieux	hagiographie	Audoinus de	NULL	NULL
15	Vita Galli Vetustissima	religieux	hagiographie	NULL	NULL	NULL
16	Vita Galli auctore Wettino cum prologo metrico ad Gozbertum	religieux	hagiographie	Wetti	de Reichenau	NULL
17	Vita Galli Walahfrido	religieux	hagiographie	Walhafrid	Strabo	NULL

identifiant	identifiant_CT	cp_date	cp_date_debut	cp_date_fin	cp_date_formelle	cp_date_CT	cp_lieu
Baront.	selon MLW (Zitierliste)	678-79	0678-03-25	0679-12-01	0678-12-01	selon MLW (Zitierliste)	NULL
Vita Adelph.	selon Mittellateinische	VIIIe siècle	0600-01-01	0699-01-01	0650-01-01	selon MLW (op. minora)	NULL
Vita Amand.	selon MLW (op. minora	deuxième m	0750-01-01	0799-01-01	0775-01-01	selon MLW (op. minora)	NULL
Vita Amati hal	selon MLW (op. minora	VIIe/VIIIe siè	0600-01-01	0799-01-01	0700-01-01	selon MLW (op. minora)	Saint-Symph
Vita Audoini	selon MLW (op. minora	au début du	0700-01-01	0710-01-01	0705-01-01	selon MLW (op. minora)	Rouen
Vit. Desid.	selon MLW (Zitierliste)	circa 800	0790-01-01	0810-01-01	0800-01-01	cf. Krusch 1902, Heinzelr	Abbaye de S
Vita Elig.	selon MLW (Zitierliste)	circa 670/80	0670-01-01	0681-01-01	0675-01-01	(handschr. Änderung au	NULL
Vita Galli I	selon MLW (Zitierliste)	670-80	0670-01-01	0680-01-01	0675-01-01	MLW (Zitierliste *DE), cf	NULL
Vita Galli II	selon MLW (Zitierliste)	816-824	0816-01-01	0825-01-01	0820-01-01	Berschin 2012:2 (*DE Ve	Reichenau
Vita Galli III	selon MLW (Zitierliste)	833	0833-01-01	0834-01-01	0833-06-01	selon Berschin 2010:56	Reichenau

L'annotation du sous-corpus latin

adoption du jeu d'étiquettes CompHistSem

- 15 catégories principales
  - *Adjective* (ADJ), *Adverb* (ADV), *Cardinal Number* (NUM), *Conjunction* (CON), *Distributive Number* (DIST), *Foreign Material* (FM), *Interjection* (ITJ), *Non Word* (XY), *Noun* (NN), *Ordinal Number* (ORD), *Personal Name* (NP), *Preposition* (AP), *Pronoun* (PRO), *Proper Name* (NE), *Verb* (V)

- prise en compte des particularités du latin tardif
  - regroupement de différents lemmata (p. ex. *aecclesia* et *ecclesia*, *bystia* et *bestia*) sous un superlemma (p. ex. *ecclesia*, *bestia*) pour tenir compte de la variation du latin tardif
  - cas *oblique* pour des formes accusatives sans *-m*
    - p. ex. *procedens Genovefa ad* ***cellola sua*** (Vita Genovefae)
  - marquage de formes déviantes comme *non-classical*
    - p. ex. *virginis* au lieu de *virgines*, *que* au lieu de *quae*

État de l'annotation – vies de saints latines

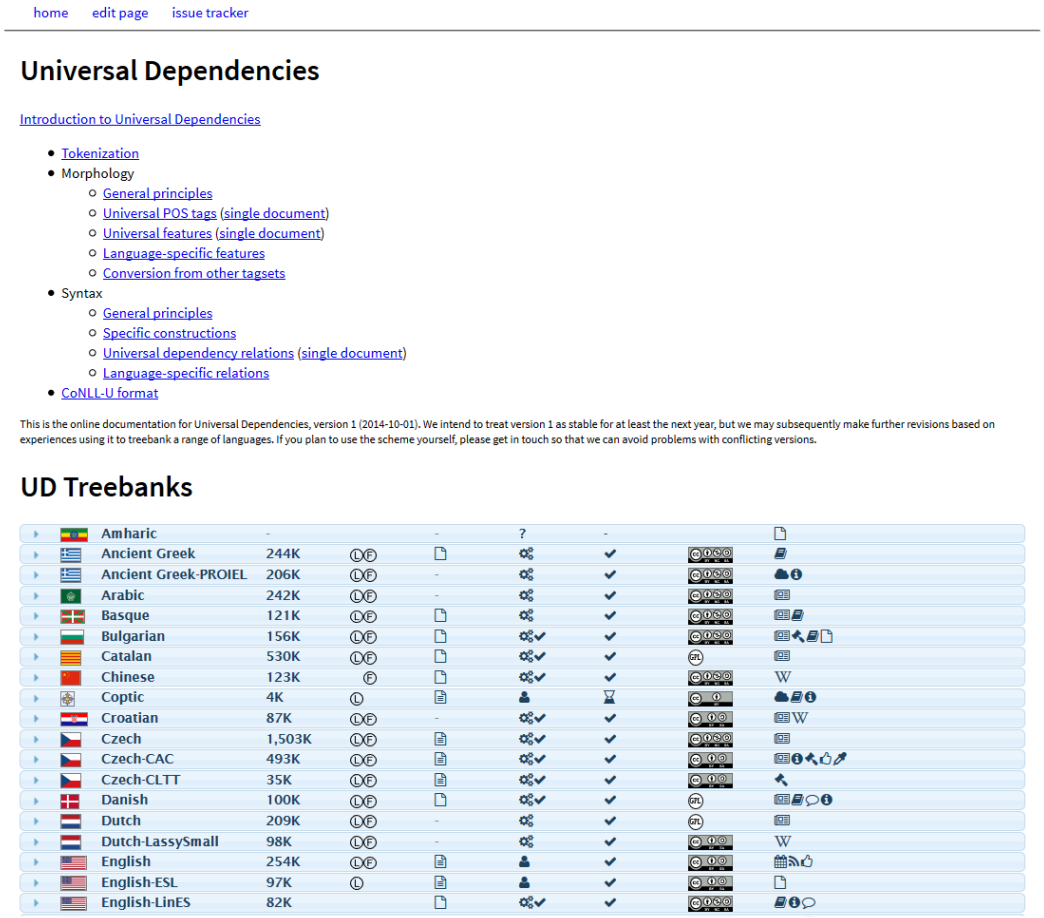
	Vita	Lematization Statistics (Stand:So, 26.06.16)
1	Vita Adelphii	100% Unique Wordform
2	Vita Bertilae	99,93%
3	Vita Trudonis	99,51%
4	Vita Galli Wettino	99,48%
5	Vita Boniti	99,31%
6	Vita Amandi	99,23%
7	Vita Galli Vetustissima	99,10%
8	Visio Baronti	98,93%
9	Vita Genovefae	98,82%
10	Vita Austrigisili Biturgi	98,71%
11	Vita Gaugerici	97,74%
12	Vita Audoini	97,71%
13	Vita Pardulfi 1	97,66%
14	Vita Filiberti B	97,55%
15	Vita Fursei	97,52%
16	Vita Filiberti A	97,31%
17	Vita Hugberti	97,23%
18	Vita Eucherii	95,94%
19	Vita Wandregiseli	95,24%

Le jeu d'étiquettes commun

Universal Dependencies

- catalogue d'étiquettes commun servant de base pour l'annotation syntax. de différentes langues
  - un projet pour le français moderne
  - 3 projets pour le latin (Perseus, IT, Proiel)

- catalogue fixe de 17 étiquettes POS
- *features* extensibles



home edit page issue tracker

Universal Dependencies

Introduction to Universal Dependencies

- Tokenization
- Morphology
  - General principles
  - Universal POS tags (single document)
  - Universal features (single document)
  - Language-specific features
  - Conversion from other tagsets
- Syntax
  - General principles
  - Specific constructions
  - Universal dependency relations (single document)
  - Language-specific relations
- CoNLL-U format

This is the online documentation for Universal Dependencies, version 1 (2014-10-01). We intend to treat version 1 as stable for at least the next year, but we may subsequently make further revisions based on experiences using it to treebank a range of languages. If you plan to use the scheme yourself, please get in touch so that we can avoid problems with conflicting versions.

UD Treebanks

Amharic	-		-	?	-	
Ancient Greek	244K	LF				
Ancient Greek-PROIEL	206K	LF	-			
Arabic	242K	LF	-			
Basque	121K	LF				
Bulgarian	156K	LF				
Catalan	530K	LF				
Chinese	123K	F				
Coptic	4K	L				
Croatian	87K	LF	-			
Czech	1,503K	LF				
Czech-CAC	493K	LF				
Czech-CLTT	35K	LF				
Danish	100K	LF				
Dutch	209K	LF	-			
Dutch-LassySmall	98K	LF	-			
English	254K	LF				
English-ESL	97K	L				
English-LinES	82K					

L'exploitation du corpus

TXM

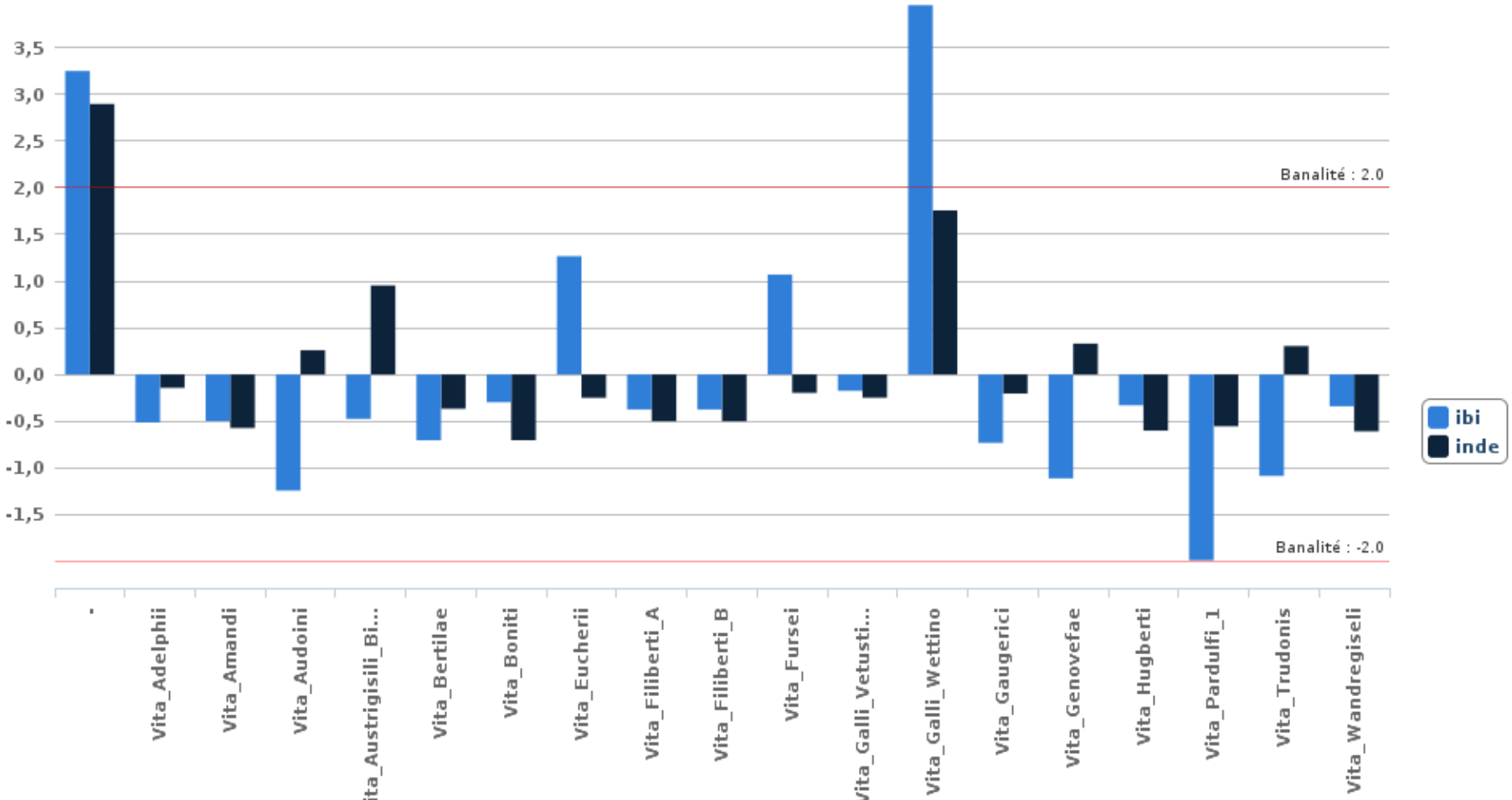
plateforme open-source pour les chercheurs en Sciences humaines

- linguistes, géographes, historiens, etc.

- possibilité de produire
  - des concordances kwic à partir de recherches de motifs lexicaux complexes
  - calculs
    - du vocabulaire d'ensemble d'un corpus
    - d'index à partir de recherches de motifs lexicaux complexes
    - de cooccurrents statistiques
  - calculs de comparaison
    - à partir de différents tableaux de contingence
    - des mots ou de leurs propriétés particulièrement présents dans une partie du corpus (spécificités statistiques)
    - par analyse factorielle des correspondances
    - par classification hiérarchique ascendante

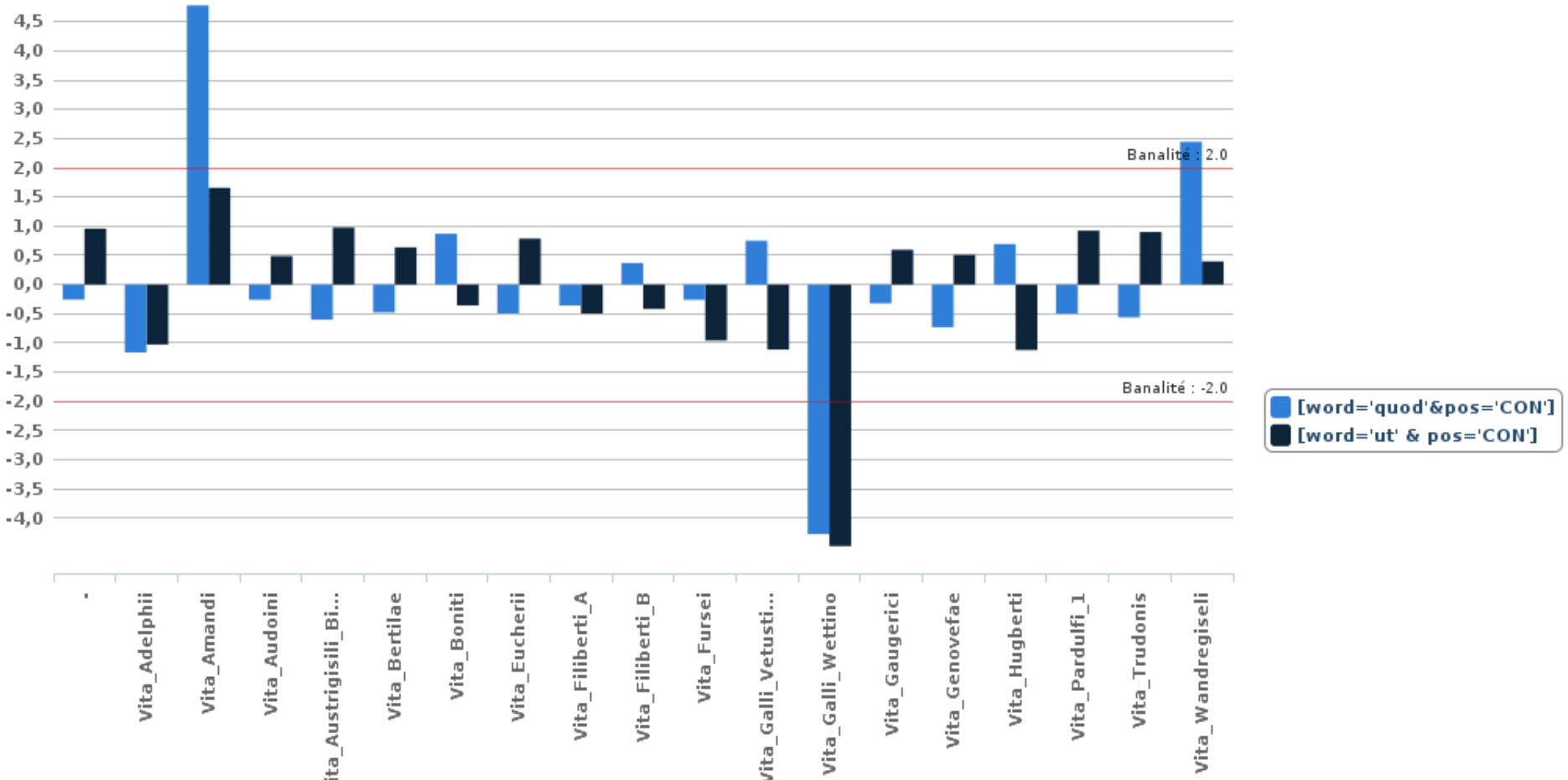
(http://textometrie.ens-lyon.fr/?lang=fr)

IBI et INDE dans 19 vies de saints



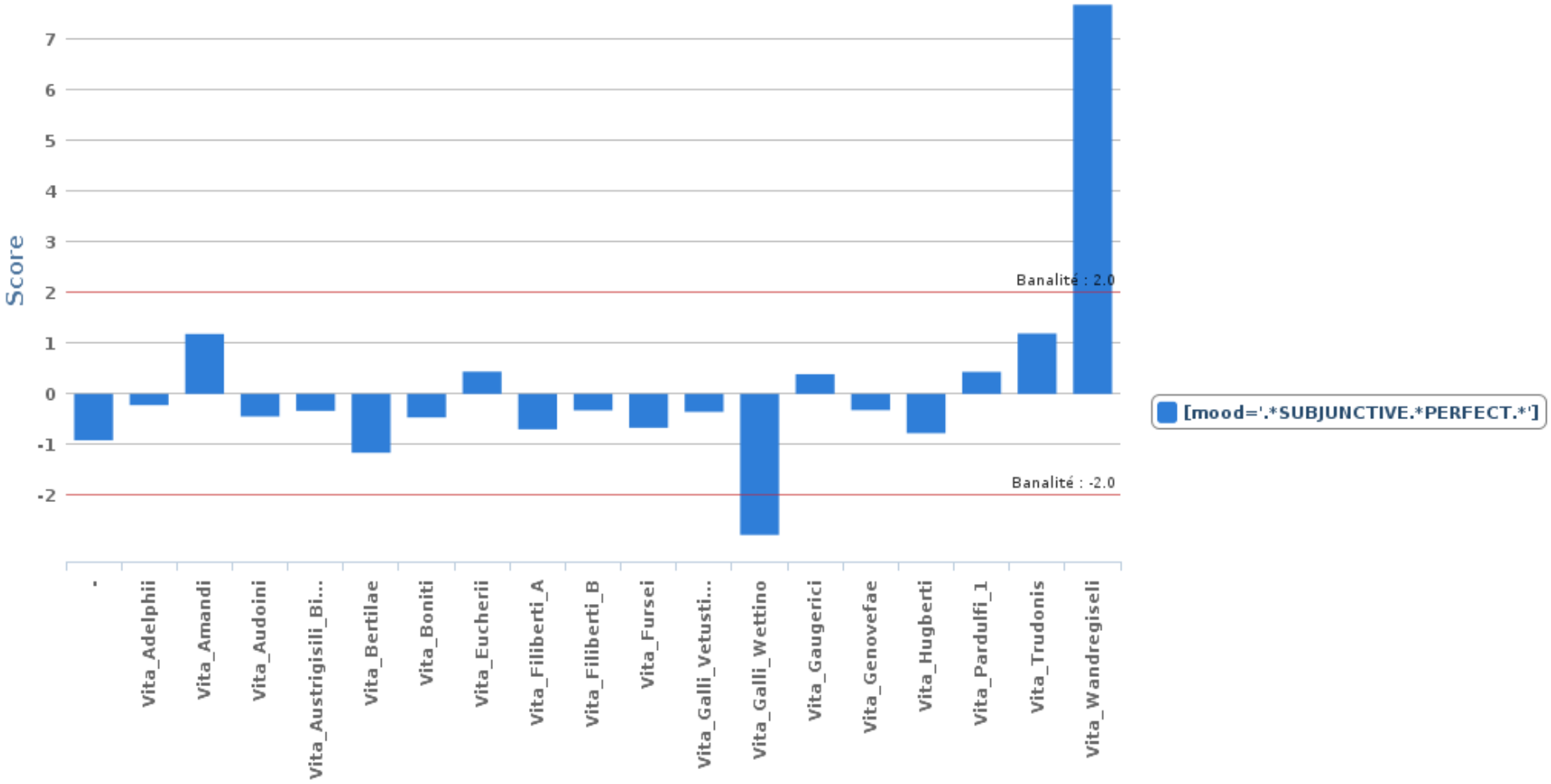
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

UT et QUOD (conj.) dans 19 vies de saints



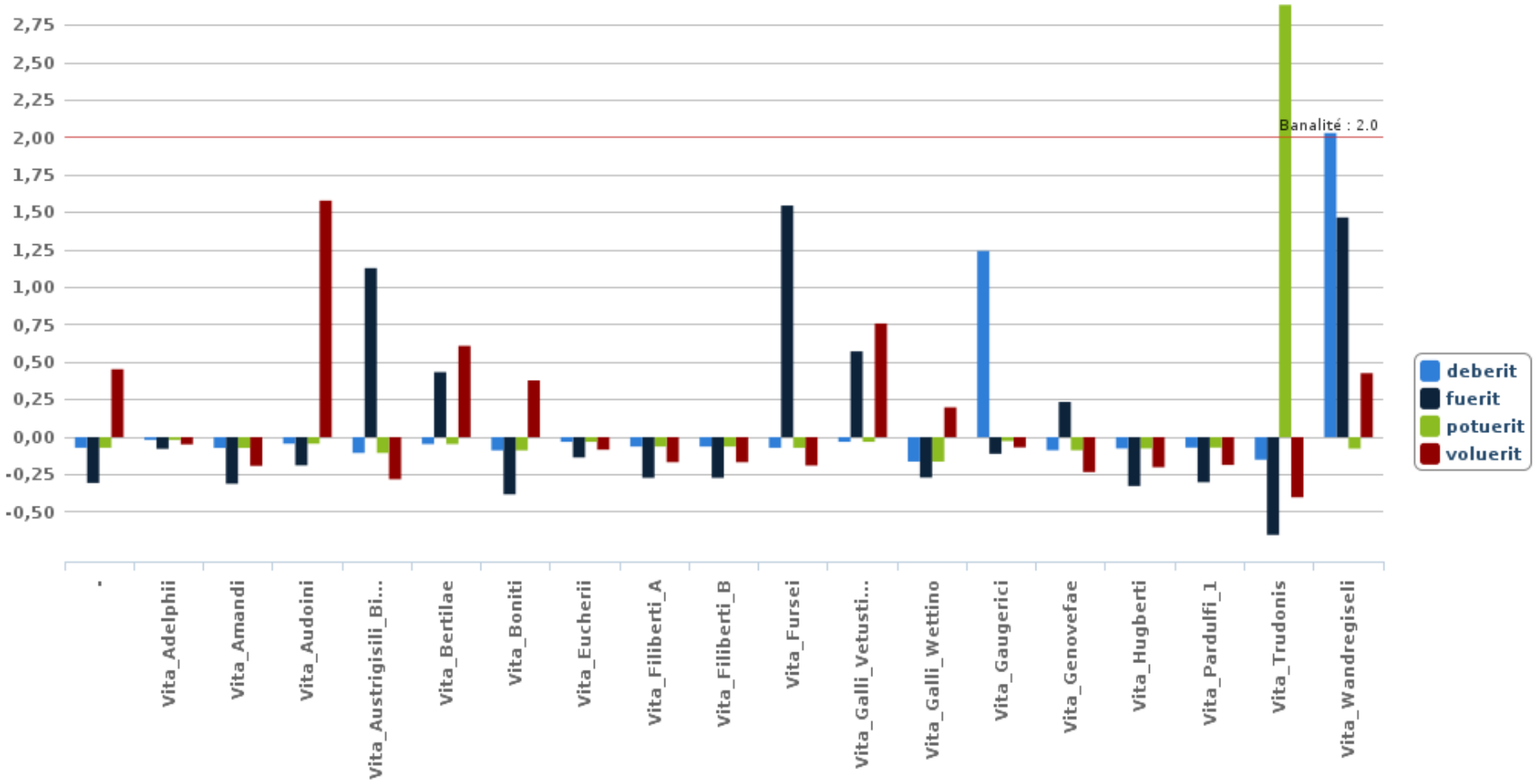
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

formes du subjonctif parfait dans 19 vies de saints



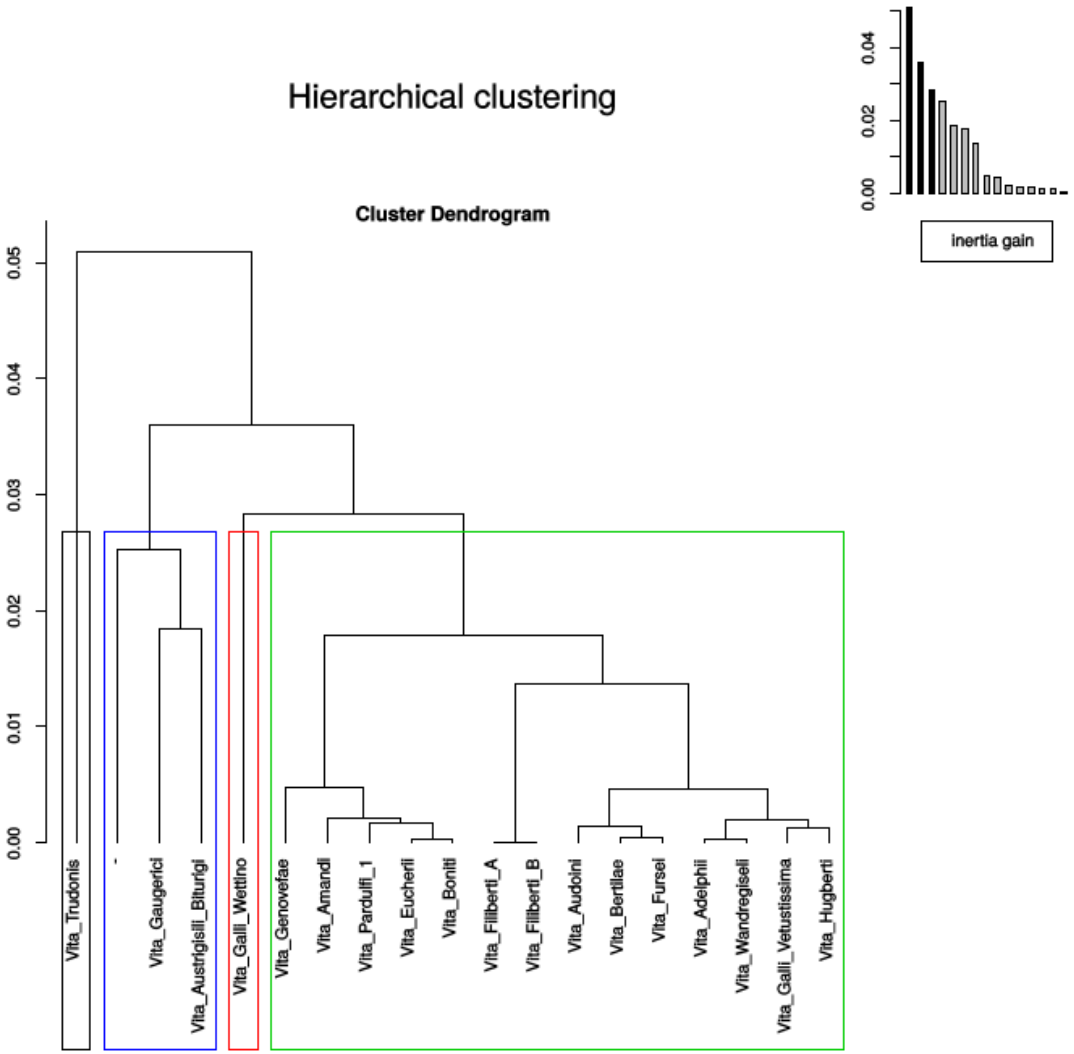
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

fuerit, voluerit, deberit, potuerit dans 19 vies de saints



Merci de votre attention!

classification hiérarchique ascendante des 19 vies



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Fiches descripteurs – base de données

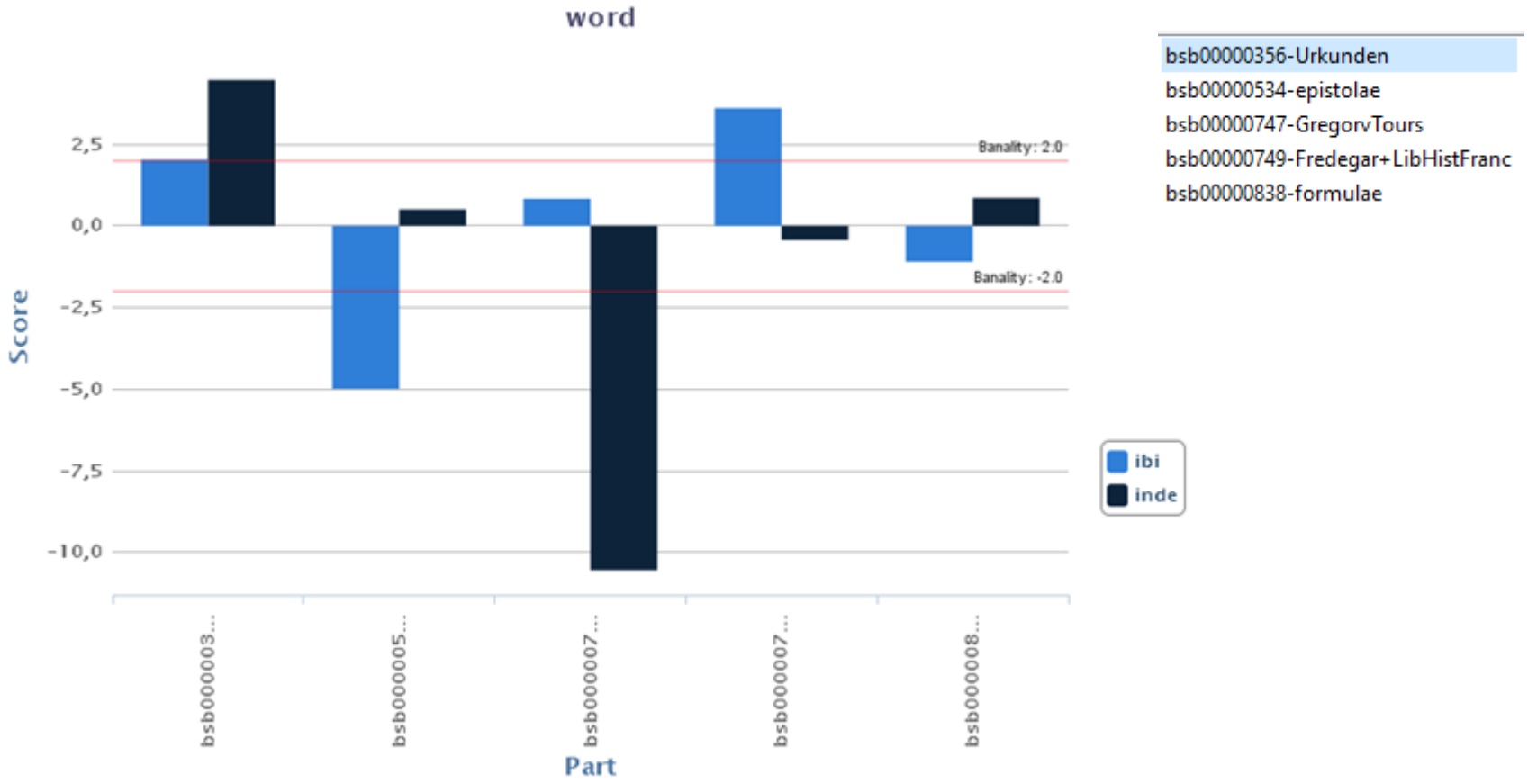
cp_lieu_CT	ms_date	ms_date_debut	ms_date_fin	ms_date_formelle	ms_date_CT	ms_lieu
NU	VIII/IXe siècl	0700-01-01	0899-01-01	0800-01-01	NULL	Freising
NULL	XIe siècle	1000-01-01	1099-01-01	1050-01-01	Samaran 1962:4	NULL
NULL	deuxième m	0950-01-01	0999-01-01	0975-01-01	cf. Aurelia bibli	NULL
origine probal	XIe siècle	NULL	NULL	NULL	Samaran 1962:4	NULL
NULL	X/XIe siècle	0900-01-01	1099-01-01	1000-01-01	selon Holveld 1	NULL
NULL	XIe siècle	1000-01-01	1099-01-01	1050-01-01	selon la BnF on	Saint-Pierre
NULL	deuxième q	0825-01-01	0849-01-01	0837-01-01	Datierung von E	NULL
NULL	première mo	0900-01-01	0949-01-01	0925-01-01	NULL	NULL
NULL	première mo	0800-01-01	0849-01-01	0825-01-01	Scarpatetti 200	NULL
NULL	Xe siècle	0900-01-01	0999-01-01	0950-01-01	selon Krusch	NULL

ed_scie_prenom	ed_scie_nom	ed_co_maison	ed_co_date	ed_co_lieu	ed_co_bib	e
Wilhelm	Levison	Monumenta Gern	1910	Hannover, Lei	MGMer. V p.3	
Bruno	Krusch	Monumenta Gern	1902	Hannover, Lei	MGMer. IV p.:	
Bruno	Krusch	Monumenta Gern	1910	Hannover, Lei	MGMer. V p.3	
Bruno	Krusch	Monumenta Gern	1902	Hannover, Lei	MGMer. IV, p.	
Wilhelm	Levison	Monumenta Gern	1910	Hannover, Lei	MGMer. V p.5	
Bruno	Krusch	Monumenta Gern	1902	Hannover, Lei	MGMer. IV, p.	
Bruno	Krusch	Monumenta Gern	1902	Hannover, Lei	MGMer. IV p.(	
Bruno	Krusch	Monumenta Gern	1902	Hannover, Lei	MGMer. IV p.:	
Bruno	Krusch	Monumenta Gern	1902	Hannover, Lei	MGMer. IV p.:	
Bruno	Krusch	Monumenta Gern	1902	Hannover, Lei	MGMer. IV p.:	

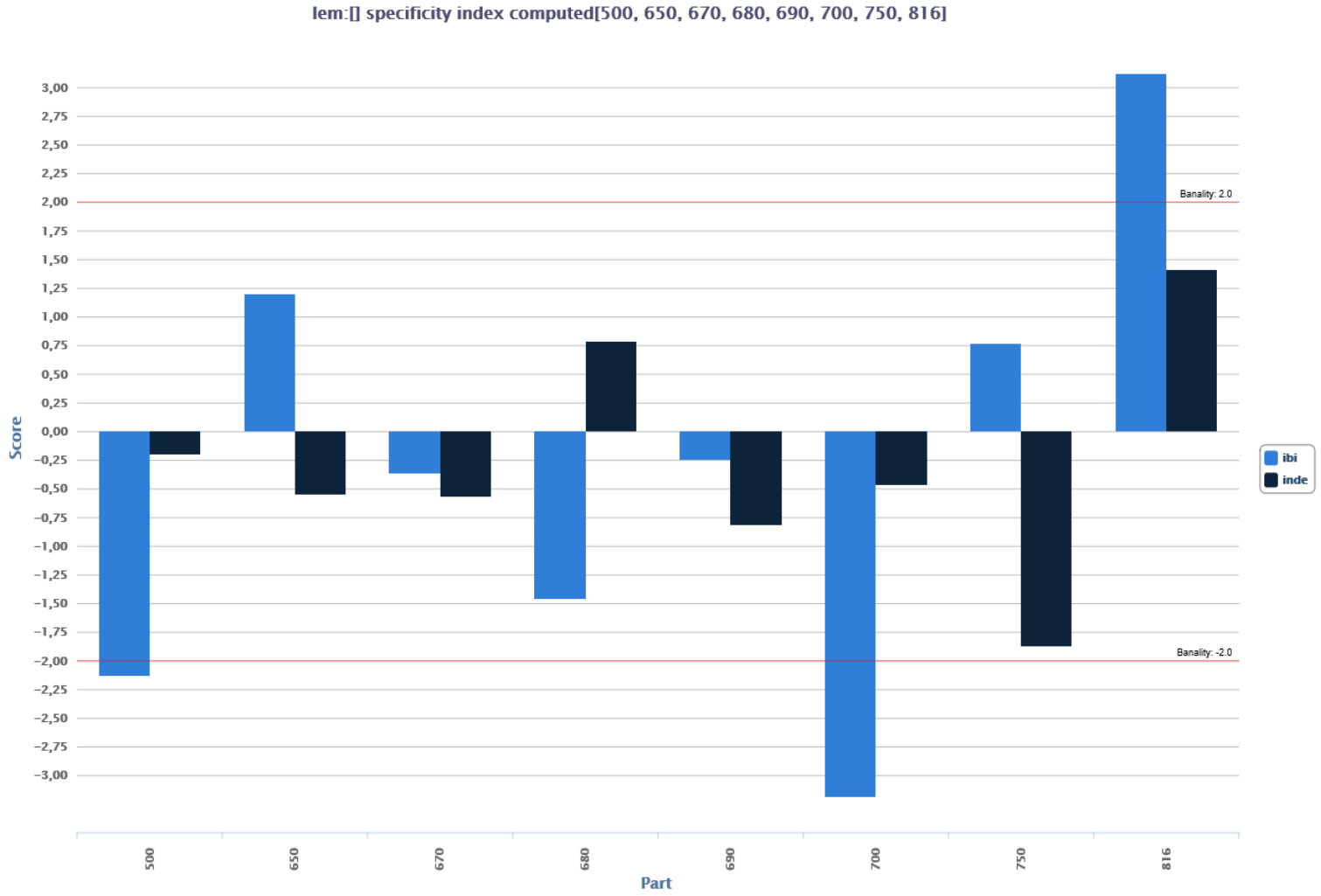
Fiches descripteurs – base de données

ident_ms_ville	ident_ms_bibliothèque	ident_ms_folio_debut	ident_ms_folio_fin	ident_ms_cote	ident_ms_CT
München	Bayerische Staatsbibliothe	136	147'	Clm 6293	NULL
Paris	Bibliothèque nationale de	NULL	NULL	lat. 5294	trois manuscrits
Orléans	Bibliothèque municipale	NULL	NULL	n. 331	NULL
Paris	Bibliothèque nationale de	95'	104'	lat. 5294	Colbertinus 250
Wien	Österreichische Nationalbi	NULL	NULL	Cod. 420 (Salisb. 3	Salzburg, Domka
Paris	Bibliothèque nationale de	207	217'	lat. 17002	Saint-Pierre de
Bruxelles	Bibliothèque royale de Bel	NULL	NULL	ms. 5374-75	NULL
St. Gallen	Stiftsbibliothek St. Gallen	NULL	NULL	Cod. Sang. 2106	depuis 2006 (cf.
St. Gallen	Sanktgallener Stiftsbiblioth	166	227	cod. 553	BHL 3246; seuler
Einsiedeln	Stiftsbibliothek Einsiedeln	297	371	Cod. 257	selon Krusch

TXM, p.ex. IBI et INDE dans une partie du corpus actuel

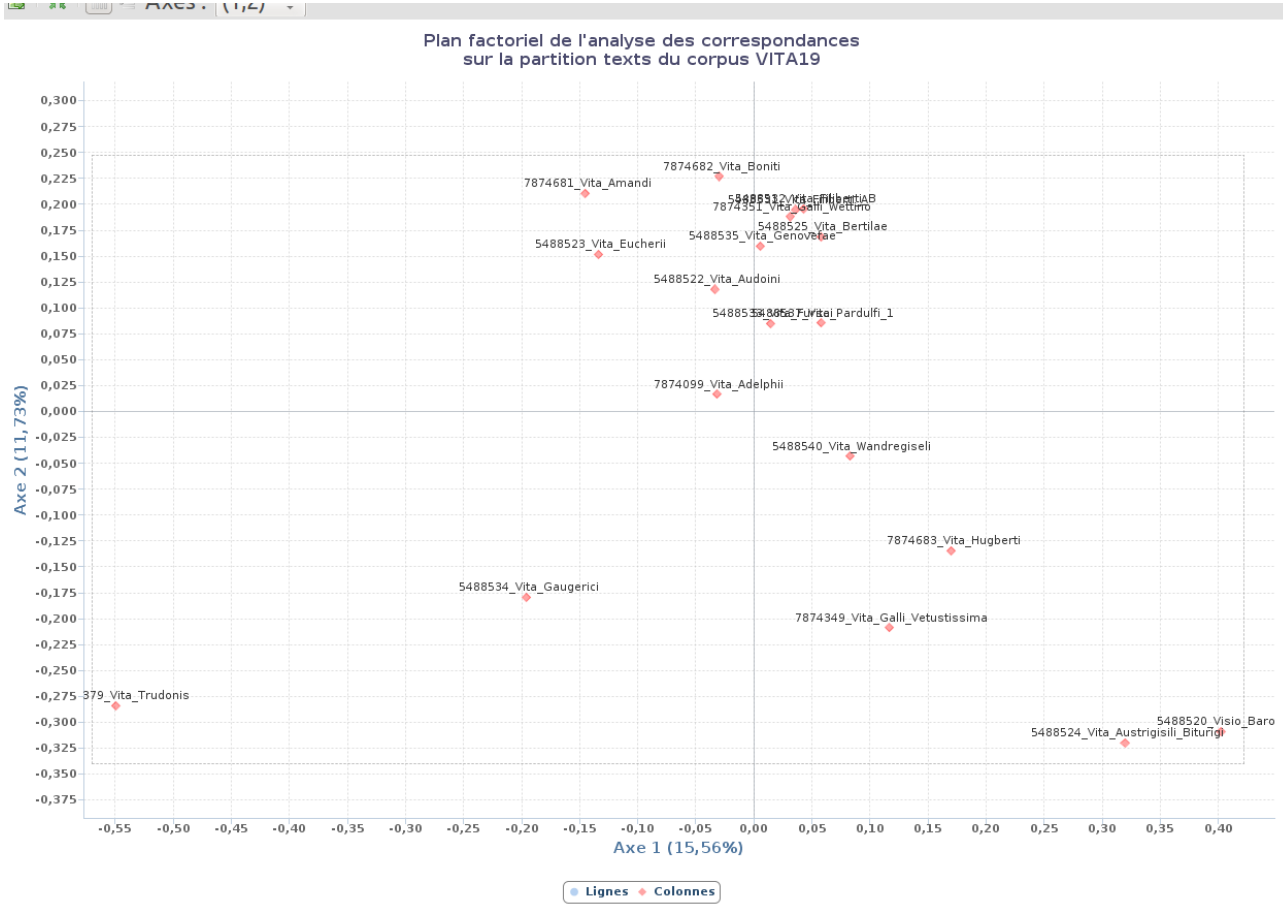


TXM, p.ex. IBI et INDE dans une partie du corpus actuel



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

TXM, plan factoriel de 19 vies de saint latines représentées par leurs lemmes



#	Valeur propre	%	Σ%
1	0,0572	15,56	15,56
2	0,0431	11,73	27,30
3	0,0396	10,77	38,07
4	0,0310	8,42	46,49
5	0,0278	7,56	54,05
6	0,0226	6,15	60,20
7	0,0209	5,69	65,89
8	0,0196	5,32	71,21
9	0,0173	4,72	75,93
10	0,0161	4,39	80,32
11	0,0142	3,86	84,18
12	0,0120	3,26	87,43
13	0,0113	3,07	90,50
14	0,0103	2,81	93,31
15	0,0091	2,47	95,78
16	0,0080	2,17	97,95
17	0,0073	1,99	99,95
18	0,0002	0,05	100,00

Valeurs propres | Info lignes | Info colonnes