Séminaire Praxiling
Montpellier, lundi 9 mai 2022

Partage d'expérience en textométrie

Étude linguistique diachronique de l'oralité,
Textométrie sur corpus vidéo,
Réécriture analytique de corpus.

Bénédicte PINCEMIN
IHRIM, univ. Lyon, CNRS

This work is licensed under the Creative Commons Attribution 4.0 International License.
http://creativecommons.org/licenses/by/4.0/

Plan

Expérience dans le cadre de deux projets :

- **Base de Français Médiéval**
  - Étude de l’oral représenté, en trois grandes étapes
- **ANR Antract**, Analyse transdisciplinaire des *Actualités françaises* (1945-1968)
  - Textométrie sur corpus vidéo
  - Analyse par annotation puis projection

Point de vue

- Projets interdisciplinaires, en équipe
  - BFM : linguistique diachronique, ancien français, philologie numérique, traitement automatique des langues, textométrie…
  - Antract : histoire du XXe s., archives audiovisuelles, reconnaissance automatique de la parole, analyse d’image, traitement de données, textométrie…
- => ici : textométrie, méthodologie

La Base de Français Médiéval

- Textes intégraux du IXe au XVe s. : aujourd'hui 170 textes, près de 5 millions de mots.
- Développée à l'ENS Fontenay – Saint-Cloud puis Lyon depuis 1989 (Ch. Marchello-Nizia, C. Guillot-Barbance)
- Ressources pour l'enseignement et la recherche sur la langue, la littérature et la civilisation médiévale
- Philologie numérique et humanités numériques : édition XML-TEI, et interface d'analyse (Weblex puis portail TXM)
- Science ouverte : FAIR ; licences ouvertes (Etalab pour texte, CC BY-NC-SA pour l'apparat critique éventuel, GPL v.2 pour TXM)
- Adresse : http://bfm-corpus.org

L’oral représenté au Moyen Âge

BFM

- Pas de données de paroles enregistrées ! Mais :
- Pas de simple dichotomie oral / écrit
  - Continuum
  - Distinguer le canal et le mode de conception (Koch & Österreicher 1990 et 2001), « scripturalité à destin vocal » (Koch 1993)
- Des observables pour une approche contrastive :
  - De l’écrit qui se donne comme de l’oral : le discours direct
  - Des genres textuels oralisés (ex. chanson de geste) ou non

La BFM : Un corpus annoté et structuré

- Certaines informations vont nous intéresser plus spécifiquement :
  - Métadonnées sur les textes : date de composition (et date du manuscrit) (sous différentes formes plus ou moins précises), domaine, genre, forme, dialecte, etc.
  - Annotation morphosyntaxique (POS) sur les mots (58 étiquettes CATTEX 2009) :
    - automatique (TreeTagger avec un modèle entraîné sur la BFM)
    - vérifiée pour un sous-ensemble de textes
    - description linguistique plus unifiée et synthétique que les formes graphiques, moins dépendante des variations phonétiques et graphiques selon les régions et les époques
  - Annotation des passages au discours direct (DD) : semi-automatique, en se basant sur la typographie des éditeurs (guillemets – incises comprises).

BFM : ex. de balisage du DD (TEI)

Balise **<q>** de citation (cf. guillemets) :

Car il ont tozjorz esté sainz et haitiez. **<q>**« Certes, fet mes sires Gauvains, ce me plest mout. »**</q>** Granz est la joie que cil de la cort font a Boort et a Lion [...].

(qgraal_cm, p.160d-bis)

Balise **<sp>** de tour de parole (théâtre et « dialogues » pédagogiques) :

Connart, à l'extérieur du palais **<sp>**CONNARS Oiiés ! Oiiés ! Oiés, signeur ! Oiés vo preu et vo honneur ! [...]**</sp>**
Au palais : le Roi, Auberon **<sp>**LI ROIS A AUBERON Diva ! Iés tu chaiens, Auberons, mes courlieus ?**</sp>**

(bodelnic, p.74, v.225 sq.)

Du corpus TEI à la représentation TXM

- Un personnage peut citer les paroles d’un autre
  - Les structures <q> peuvent être imbriquées
  - Une propriété q est ajoutée à chaque mot et indique le niveau d’imbrication (de 0 à 3).
- Le « non discours direct » (non-DD) est le « reste »
  - Catégorie a priori hétérogène, à vocation d’abord contrastive
  - Il n’a pas de balisage propre (il est hors <sp> et <q>)
  - Dans les résultats, on le note « z » (donc trois modalités possibles : q / sp / z).

Outils textométriques

- Qualitatifs, d’observation des occurrences et de recherche
  - Concordances, listes avec fréquence, texte
  - Requêtes avancées avec le moteur CQP
- Quantitatifs et statistiques, principalement ici :
  - Spécificités (mots ou traits sur- ou sous-représentés)
  - Analyse factorielle des correspondances : visualisation cartographique et réduction synthétique

BFM & Oral représenté, Étape 1

- Céline GUILLOT, Alexei LAVRENTIEV, Bénédicte PINCEMIN, Serge HEIDEN (**2011**) - « Oral représenté au Moyen Âge : vers une définition et une méthodologie d'analyse », ***Représentation du sens linguistique V***, Université de Savoie, 25-27 mai 2011.
- Céline GUILLOT, Alexei LAVRENTIEV, Bénédicte PINCEMIN, Serge HEIDEN (**2013**) - « Le discours direct au Moyen Âge : vers une définition et une méthodologie d'analyse », *in* Dominique Lagorgette & Pierre Larrivée (dir.), *Représentation du sens linguistique 5*, Université de Savoie, Laboratoire Langages, Littératures, Sociétés, coll. Langages n°14, p. 17-41. https://halshs.archives-ouvertes.fr/halshs-00820262v1

BFM & DD (1) : corpus et méthode

- Corpus complet : 61 textes, 1,9 M mots → graphies
- Sous-corpus à morphosyntaxe vérifiée : 9 textes, 320 000 mots → POS
- D’abord spécificités DD vs non-DD sur les POS (sous-corpus vérifié)
- Puis examen plus approfondi de certains cas en modifiant la granularité :
  - Sur-catégories (ex. interrogatifs)
  - Graphies (ex. déictiques)
  - Sous-catégories par regroupement de graphies (ex. personnes des pronoms)
  - Différenciation de contextes ou constructions (ex. infinitif précédé d’un verbe modal ou du verbe *aller*, participe passé précédé d’un auxiliaire -temps composé ou passif-, participe présent précédé d’un verbe conjugué ou de *en*)

BFM & DD (1) : aperçu des observations

Pos	F totale	f dans le non DD	Spécif. pour le non DD	f dans le DD	Spécif. pour le DD
PONpxx	2432	4	*-1000*	2428	*1000*
PONpga	1091	83	-302,32	1008	*1000*
PONpdr	1085	84	-298,81	1001	*1000*
PROper	30414	16029	-195,37	14385	*1000*
PONfrt	13086	6945	-73,67	6141	*1000*
PROint	270	24	-71,1	246	*1000*
ADVneg	7141	3617	-68,07	3524	*1000*
INJ	232	30	-50,89	202	*1000*
NA	82	0	-33,34	82	*1000*
ADVint	106	11	-26,52	95	*1000*
ADJpos	225	70	-18,66	155	*1000*
PROper.PROper	55	3	-17,35	52	*1000*

BFM & DD (1) : aperçu des observations

pos	F totale	f dans le non DD	Spécif. pour le non DD	f dans le DD	Spécif. pour le DD
DETdef	16314	11861	*1000*	4453	-233,60
NOMcom	49225	31920	*1000*	17305	-88,44
CONcoo	18888	12715	*1000*	6173	-80,98
PRE	28687	18630	*1000*	10057	-51,39
DETndf	2531	1888	*1000*	643	-48,42
PRE.DETdef	4233	2996	*1000*	1237	-41,97
DETcom	166	163	*1000*	3	-30,59
VERppa	1017	773	*1000*	244	-24,29
DETcar	1187	870	*1000*	317	-19,12
PROind	3439	2326	*1000*	1113	-16,50
VERppe	10065	6505	15,48	3560	-15,45
NOMpro	8541	5544	15,05	2997	-15,06

BFM & DD (1) : résultats

- Discours direct
  - Ponctuations [tous types]
  - Interrogatifs directs
  - Interjections [sauf *He*, *ha* – apostrophe au lecteur]
  - Négation [toutes formes]
  - Pronoms [nombreux types]
  - Possessifs
  - Démonstratifs
  - Infinitif [modal + inf. mais pas *aller* + inf.]
  - Conjonctions de subordination

- Non Discours direct
  - Déterminants
  - Noms communs et propres
  - Conjonctions de coordination [ET, mais pas CAR, MES, NE]
  - Prépositions
  - Cardinaux
  - Participes (présent et passé) [y compris Vcjg + participe, *en* + participe]
  - Pronoms indéfinis
  - Adjectifs qualificatifs
  - Adverbes de subordination

BFM & DD (1) : résultats

- Déictiques (ÇA, CI, HUI, ICI, ILLUEC)
  - Fortement DD, mais :
  - Les formes à majuscule initiale (moins fréquentes) semblent équiréparties DD / non-DD (structuration des unités narratives)
  - ILLUEC est équiréparti
- Personnes des pronoms personnels et des possessifs (pro., adj., dét.)
  - 1ère et 2ème personne fortement DD
  - 1ère personne pas absente du non-DD en fréquence absolue (plusieurs centaines d’occurrences)
  - Il faudrait pouvoir prendre en compte les personnes verbales

BFM & DD (1) : aperçu des observations

Déictiques : Word	F totale	f dans les passages <q>	Spécif. pour les passages <q>	f dans les passages <sp>	Spécif. pour les passages <sp>	f dans le non DD	Spécif pour le non DD
ça	271	163	**10**	16	-0	92	-32
Ça	19	10	2	0	-0	9	-1
ci + cy	1072	627	**10**	84	2	361	-128
Ci + CI + Cy	189	54	0	3	-2	132	-0
hui + huy + ui + uy	223	133	**10**	42	10	48	-48
Hui + Ui	22	10	1	6	2	6	-4
ici + icy	280	192	**10**	19	0	69	-53
Ici + Icy + ICY	93	12	-2	5	0	76	2
iluec + ilueques + illuec + illueques	207	34	-2	18	1	155	1

BFM & DD (1) : aperçu des observations

Personne	F totale	f dans le non DD	Spécif. pour le non DD / pronoms personnels et possessifs	f dans le DD	Spécif. pour le DD / pronoms personnels et possessifs
1s	6015	502	*-1000*	5513	*1000*
2s	1265	17	*-1000*	1248	*1000*
1p	1259	243	-138,91	1016	*1000*
2p	3804	156	*-1000*	3648	*1000*
3	25853	19390	*1000*	6463	*-1000*
Total		20308		17888	

BFM & DD (1) : aperçu des observations

	F totale	f dans le non DD	Spécif pour le non DD	f dans le DD	Spécif pour le DD
Pooir + infinitif	1240	628	-13	612	13
Devoir + infinitif	521	179	-34	342	34
Voloir + infinitif	417	200	-8	217	8
Aller + infinitif	223	163	5	60	-8
Vb + infinitif	4267	2288	-22	1979	22
Taille (tous les infinitifs)	T=364967	t=221814		t=143153	

Requête-type utilisée pour gérer les insertions :

```
[X]([pos="DET.*"]?[pos="NOM.*"]|[pos="ADV.*|PRO.*"]){0,2}
[pos="VERinf" & q="0"]
```

BFM & DD (1) : bilan

- Éléments de caractérisation linguistique :
  - des sur-représentations attendues (personnes 1 et 2, interjections, interrogatifs, démonstratifs,…)
  - une opposition globale pronom (DD) *vs* nom (non-DD)
- Méthodologiquement
  - des catégories retravaillées : fusion, diffraction, contextualisation
  - des spécificités sur des constructions, par rapport à une référence ciblée pertinente (Weblex puis utilitaire PlotSpecif)
- Étendre le corpus :
  - Consolider les observations (quantitativement, qualitativement)
  - Distinguer et caractériser des types de textes

BFM & Oral représenté, Étape 2

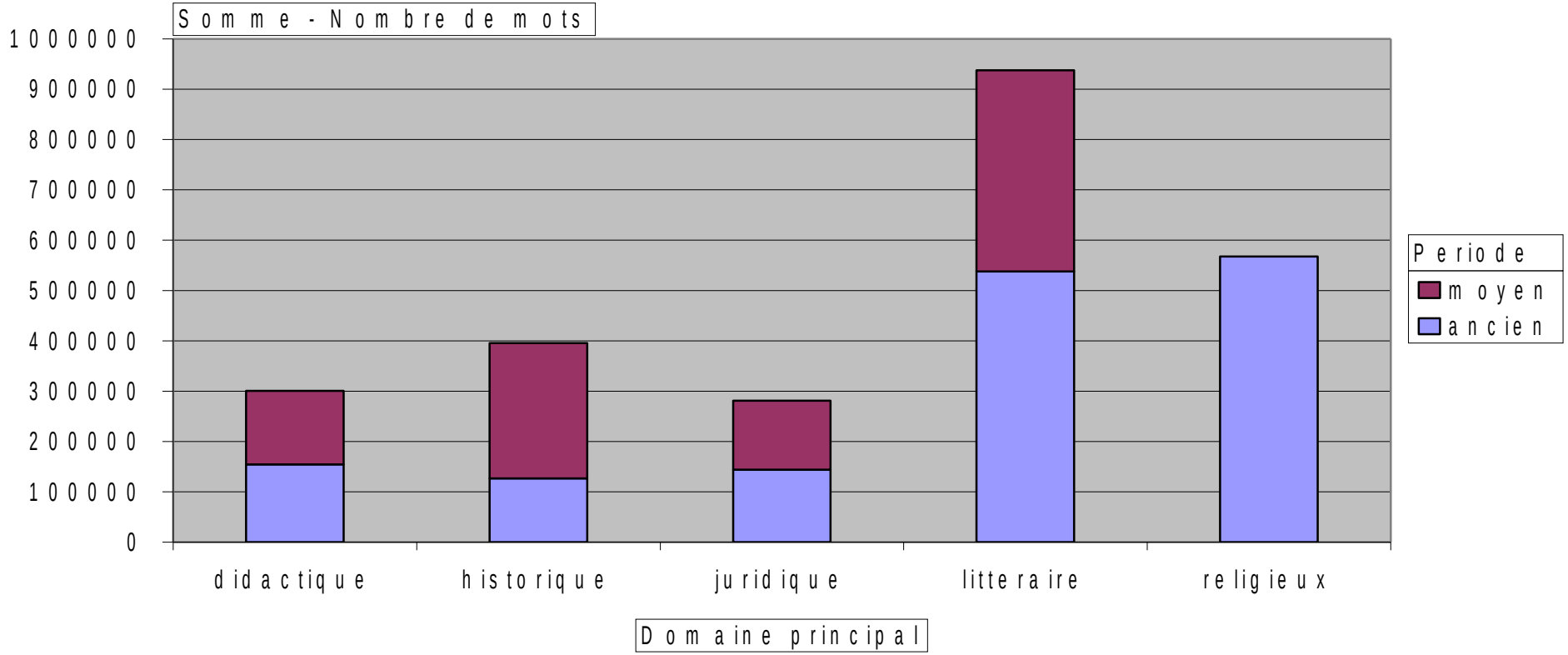
- Céline GUILLOT, Alexei LAVRENTIEV, Serge HEIDEN, Bénédicte PINCEMIN (**2012**) - « L'oral représenté dans un corpus de français médiéval (9e-15e) : paramètres de variation diamésique (discours direct vs récit), générique et diachronique », ***Colloque DIA II*** *Les variations diasystématiques et leurs interdépendances*, Copenhague, 19-21 novembre 2012.
- Céline GUILLOT, Serge HEIDEN, Alexei LAVRENTIEV, Bénédicte PINCEMIN (**2015**) - « L'oral représenté dans un corpus de français médiéval (9e-15e) : approche contrastive et outillée de la variation diasystémique », *in* Kirsten Jeppesen Kragh et Jan Lindschouw (éds), *Les variations diasystémiques et leurs interdépendances dans les langues romanes -Actes du Colloque DIA II*, Strasbourg : Éditions de linguistique et de philologie, p. 15-28. https://halshs.archives-ouvertes.fr/halshs-00760647v2

BFM & DD (2) : corpus et méthode

- Corpus complet : 78 textes, 2,5 M mots (+25%)
- Sous-corpus à morphosyntaxe vérifiée : 20 textes, 600 000 mots (x2)
- Croiser l’opposition DD / non-DD avec d’autres axes de variation : 2 périodes, 5 domaines, 20 textes vérifiés
- Caractériser avec les spécificités au niveau 3 DD x 2 périodes
- Observer les variations dominantes avec des AFC

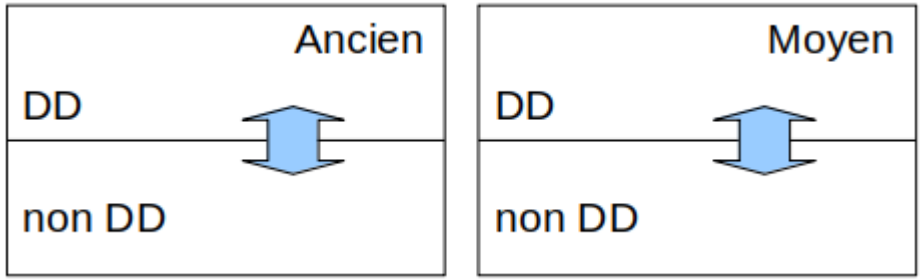
BFM & DD (2) : corpus

Dimensions du corpus complet par domaine et par période :

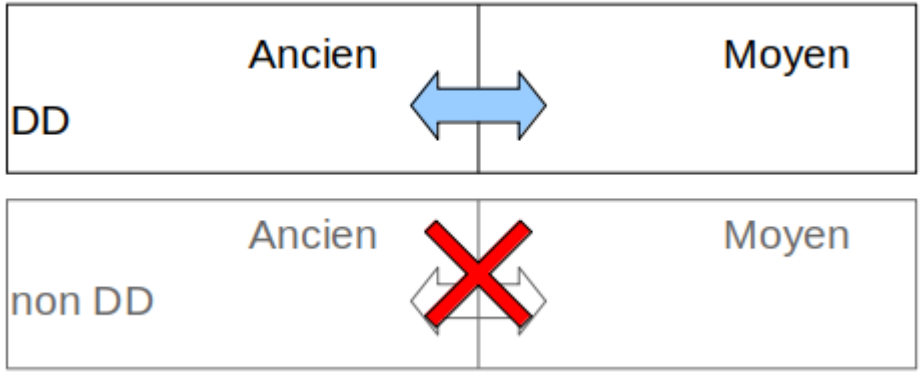


BFM & DD (2) : contrastes diachroniques

- Contraste DD / non DD à l'intérieur de chaque période



- Contraste des deux périodes pour le DD

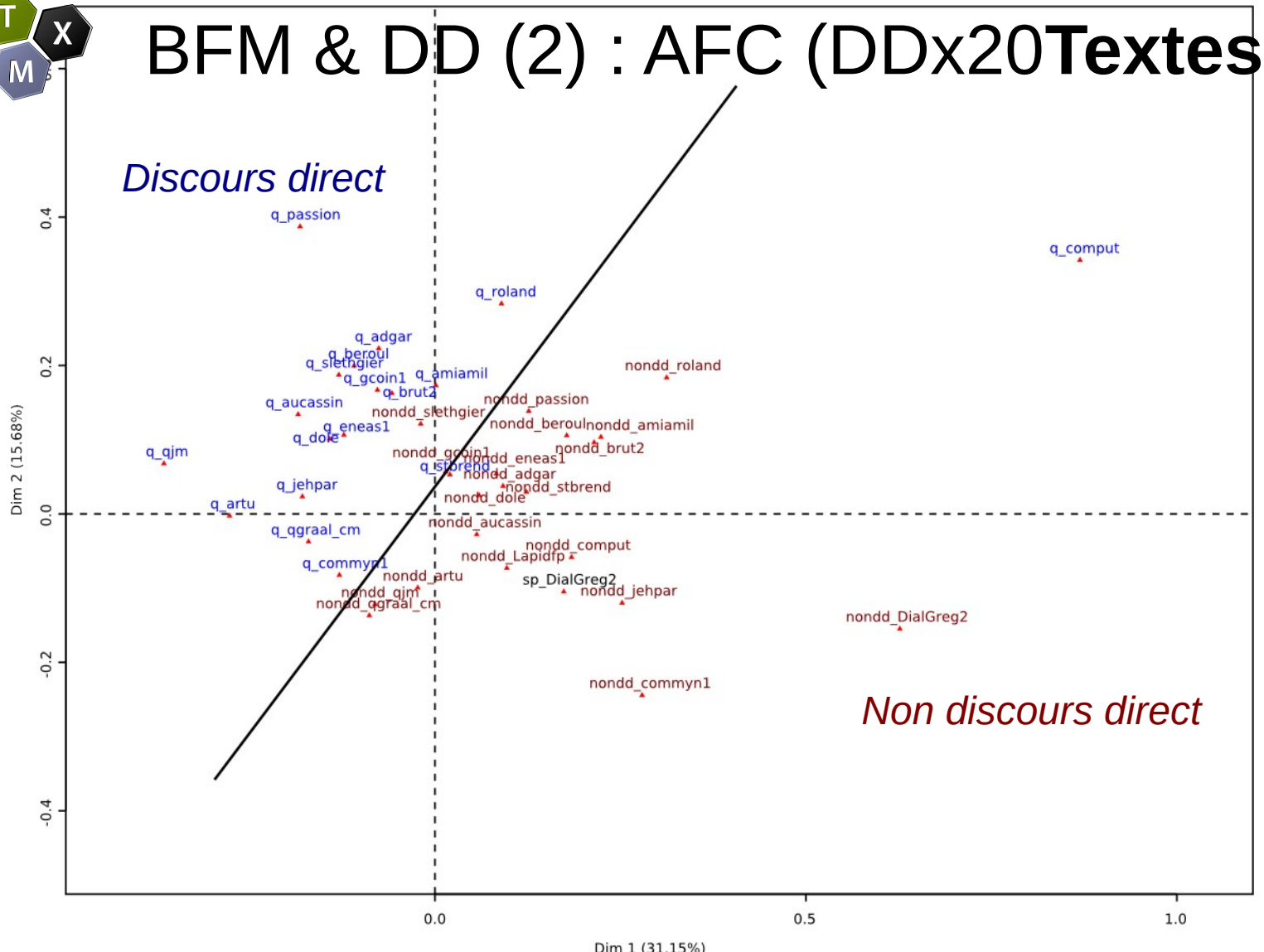


BFM & DD (2) : résultats pour la diachronie

- Scores de spécificité des étiquettes morphosyntaxiques par période dans le DD
  - Les caractéristiques majeures sont globalement stables (seules 1 ou 2 catégories ne sont sélectionnées que pour une période seulement)
  - Une grande partie des catégories spécifiques du moyen français sont spécifiques au DD dans les deux périodes (PROimp, PROper, DETpos, VERinf…)

    => Les traits caractéristiques du DD deviennent plus marqués en moyen français
  - Mais le moyen français n'est représenté que par 3 textes dans le sous-corpus vérifié, d'où influence possible du style de l'auteur, du genre, etc...

BFM & DD (2) : AFC (DDx20**Textes**)xPOS



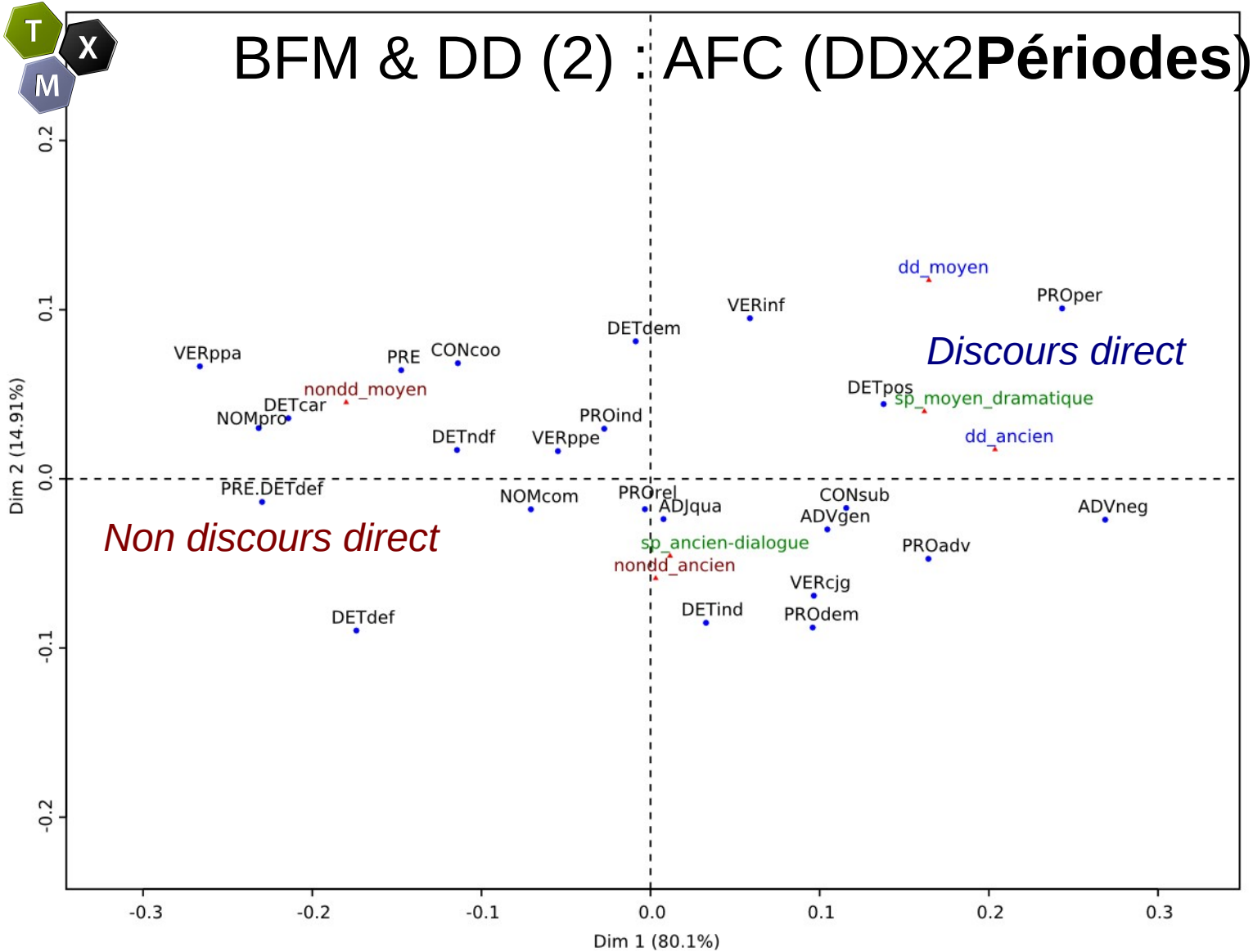
Les anomalies s'expliquent :

- q_comput (*Comput* de Philippe de Thaon) : guillemets pour citer des mots
- sp_DialGreg2 (*Vie de Saint Benoît*) : « faux » dialogue
- Autres points : mal représentés dans le plan (cos^2<0,15)

BFM & DD (2) : Gestion de la fiabilité de l’étiquetage POS sur le corpus complet

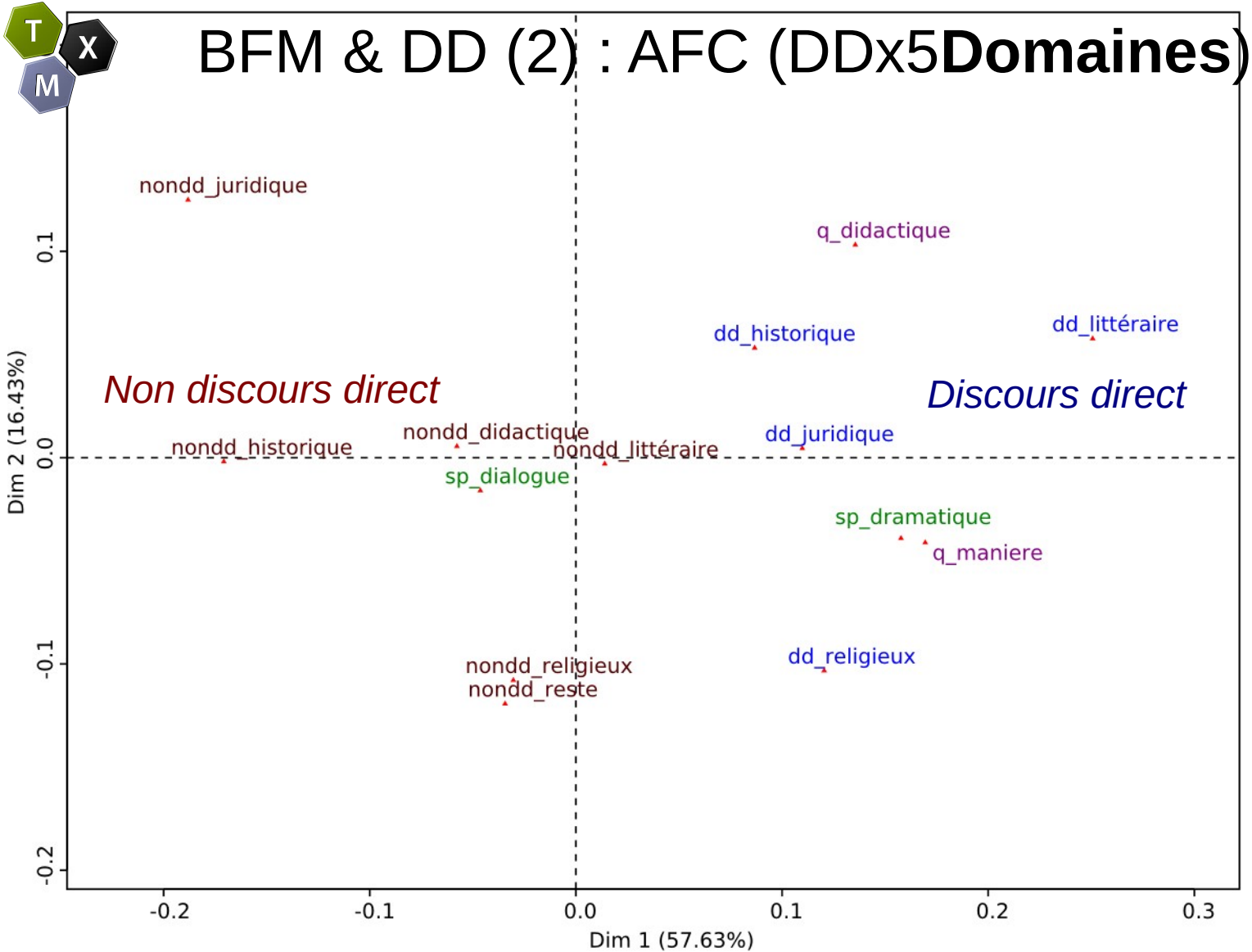
- Évaluation de la fiabilité
  - Pour chaque étiquette : taux de réussite sur les textes ne faisant pas partie du corpus d’apprentissage (vérifié)
- Étude de l’impact sur l’AFC
  - Les étiquettes « peu fiables » (taux de réussite < 75%) sont généralement peu fréquentes, elles ne pèsent quasiment pas dans la construction du plan factoriel
  - Les quelques étiquettes relativement peu fiables et fréquentes (NOMpro, VERppe, ADJqua) ont un faible impact sur la configuration des plans factoriels (si on les retire la configuration reste stable)
- Solution
  - On applique un seuil quantitatif pour l'analyse factorielle : étiquettes avec plus de 4000 occ., il reste 24 étiquettes.

BFM & DD (2) : AFC (DDx2**Périodes**)x24POS



- Le <sp> de l'ancien français contient une majorité de dialogues didactiques, et il se trouve mal représenté dans le plan (cos^2<0,13).
- Le <sp> du moyen français sont tous des textes dramatiques.

BFM & DD (2) : AFC (DDx5**Domaines**)x24POS



Pour avoir des regroupements de textes homogènes, deux genres ont été dissociés des domaines discursifs :

- les *dialogues didactiques* (sp_dialogue) et
- le *théâtre* (sp_dramatique).

BFM & DD (2) : bilan

- Linguistiquement :
  - Caractérisation morphosyntaxique globalement stable du DD
  - La variation DD / non-DD domine les autres (diachronie, domaines)

    => c'est un facteur de variation dont il faudrait davantage tenir compte dans les études linguistiques
  - Quelques textes ou genres ont des usages du DD qui s'écartent de l'oralité (les dialogues didactiques ; cas particulier des guillemets pour le *Comput*)

    => décrire et gérer des singularités
- Méthodologiquement
  - Gestion de la fiabilité des étiquettes pour travailler sur le corpus complet
  - L'AFC comme moyen de mettre en évidence des variations dominantes, en lien avec la représentation sous-jacente (POS)

BFM & Oral représenté, Étape 3

- Céline GUILLOT-BARBANCE, Bénédicte PINCEMIN, Alexei LAVRENTIEV, Serge HEIDEN (**2016**) - « Représentation de l'oral en français médiéval et genres textuels », *Journée d'études « Autour du clivage langue parlée / langue écrite du latin au français »*, Clermont-Ferrand, 25 mars 2016.

  => *Publication du volet plus linguistique :*

- Céline GUILLOT-BARBANCE, Bénédicte PINCEMIN, Alexei LAVRENTIEV (**2017**) - « Représentation de l'oral en français médiéval et genres textuels », *Langages*, 208 (4/2017), p. 53-68. https://halshs.archives-ouvertes.fr/halshs-01495132

  => *Publication du volet plus statistique :*

- Bénédicte PINCEMIN, Céline GUILLOT-BARBANCE, Alexei LAVRENTIEV (**2018**) - « Using the First Axis of a Correspondence Analysis as an Analytical Tool. Application to Establish and Define an Orality Gradient for Genres of Medieval French Texts », *in* D.-F. Iezzi, L. Celardo, M. Misuraca (eds), *JADT' 18. Proceedings of the 14th International Conference on Statistical Analysis of Textual Data,* Rome, 12-15 juin 2018, UniversItalia, Roma, vol. 2, p. 594-601. https://halshs.archives-ouvertes.fr/halshs-01759219

- Bénédicte PINCEMIN, Céline GUILLOT-BARBANCE, Alexei LAVRENTIEV (**2020**) - « Using the First Axis of a Correspondence Analysis as an Analytic Tool. Application to Establish and Define an Orality Gradient for Genres of Medieval French Texts », *in* D. F. Iezzi, D. Mayaffre, M. Misuraca (eds), *Text Analytics. Advances and Challenges*, Heidelberg : Springer, p. 127-143. https://doi.org/10.1007/978-3-030-52680-1_11, https://halshs.archives-ouvertes.fr/halshs-03070182

BFM & DD (3) : corpus et méthode

- Corpus complet : 137 textes, 4,2 M mots (+75%)
  - Pour 2020 : étiquetage morphosyntaxique automatique amélioré avec un nouveau modèle TreeTagger
- Sous-corpus à morphosyntaxe vérifiée :
  - de 26 textes, 745 000 mots (2016-2018) (+25%)
  - à 38 textes, 1,5 M mots (2020) (x2)
- Grâce à l'échelle plus large du corpus, croiser l'opposition DD / non-DD avec les genres textuels, pour observer l'oralité dans le détail de ces deux dimensions de variation
  - (q / sp / z) x 32 genres = 59 unités de discours (DU)
  - Notation : {DD : q, sp ou z}_{genre}{domaine}{nb textes : 1, 2 ou n c'est-à-dire >2}

    ex. : `q_rbrefLn`, `sp_dramatiqueR2`, `z_lettreH1`
- L'AFC comme premier moyen d'exploration
- Dans une démarche pleinement textométrique, trouver des façons de revenir aux caractéristiques linguistiques concrètes dans les textes

BFM & DD (3) : corpus (détail, 1/4)

Domaine	Genre	Siècle					Nb tokens (textes)	
		11	12	13	14	15	**Total**	% DD
acte de la pratique	divers					10746 (1)	**10746 (1)**	0 %
	plaids				13439 (1)		**13439 (1)**	0 %
	registre				390210 (2)		**390210 (2)**	1 %
	Total				**403649 (3)**	**10746 (1)**	**414395 (4)**	
didactique	bestiaire		15883 (1)	25422 (1)			**41305 (2)**	5 %
	commentaire					9554 (1)	**9554 (1)**	2 %
	comput		16897 (1)				**16897 (1)**	0 %
	dialogue		17297 (1)	10362 (1)			**27659 (2)**	99 %
	lapidaire		16228 (2)				**16228 (2)**	< 1 %
	manuel				24631 (2)	3818 (1)	**28449 (3)**	70 %
	proverbes		809 (1)		7507 (1)		**8316 (2)**	1 %
	roman			118222 (4)	21775 (1)	140226 (2)	**280223 (7)**	38 %
	traité			124527 (2)	94053 (2)		**218580 (4)**	10 %
	Total		**67114 (6)**	**278533 (8)**	**60340 (5)**	**153598 (4)**	**647211 (24)**	

BFM & DD (3) : corpus (détail, 2/4)

Domaine	Genre	Siècle					Nb tokens (textes)	
		11	12	13	14	15	**Total**	% DD
historique	chronique		98060 (6)	195495 (6)	135865 (3)	32756 (1)	**462176 (16)**	16 %
	lettre			2626 (1)			**2626 (1)**	0 %
	mémoires					228685 (8)	**228685 (8)**	1 %
	Total		**98060 (6)**	**198121 (7)**	**135865 (3)**	**261441 (9)**	**693487 (25)**	
juridique	coutumier			155291 (1)		17724 (1)	**173015 (2)**	2 %
	journal					155626 (2)	**155626 (2)**	1 %
	serment	131 (1)					**131 (1)**	100 %
	Total	**131 (1)**		**155291 (1)**		**173350 (3)**	**328772 (5)**	

BFM & DD (3) : corpus (détail, 3/4)

Domaine	Genre	Siècle					Nb tokens (textes)	
		11	12	13	14	15	**Total**	% DD
littéraire	dramatique		12551 (1)	14962 (2)			**27513 (3)**	98 %
	épique		71834 (3)	29942 (1)			**101776 (4)**	47 %
	lyrique		9085 (2)	33064 (2)		69447 (6)	**111596 (10)**	4 %
	récits brefs		2292 (2)	11675 (1)		39356 (1)	**53323 (4)**	38 %
	roman		568621 (14)	311639 (6)	171103 (2)	28965 (1)	**1080328 (23)**	42 %
	récit de voyage				27926 (1)		**27926 (1)**	1 %
	Total		**664383 (22)**	**401282 (12)**	**199029 (3)**	**137768 (8)**	**1402462 (45)**	

BFM & DD (3) : corpus (détail, 4/4)

Domaine	Genre	Siècle					Nb tokens (textes)	
		11	12	13	14	15	**Total**	% DD
religieux	commentaire		55900 (1)				**55900 (1)**	0 %
	dramatique	3414 (1)			14059 (1)		**17473 (2)**	84 %
	hagiographie	7692 (3)	76458 (3)	68802 (4)	31096 (3)	17595 (4)	**201643 (17)**	30 %
	histoire		103237 (1)				**103237 (1)**	37 %
	lettre		33378 (1)				**33378 (1)**	0 %
	récits brefs		57954 (2)	62211 (2)			**120165 (4)**	21 %
	psautier		56675 (3)				**56675 (3)**	0 %
	sermon	906 (1)	144090 (3)	5426 (1)			**150422 (5)**	10 %
	Total	**12012 (5)**	**527692 (14)**	**136439 (7)**	**132781 (5)**	**17595 (4)**	**826519 (35)**	
Total		**12143 (6)**	**1357249 (48)**	**1169666 (35)**	**931664 (19)**	**754498 (29)**	**4225220 (137)**	

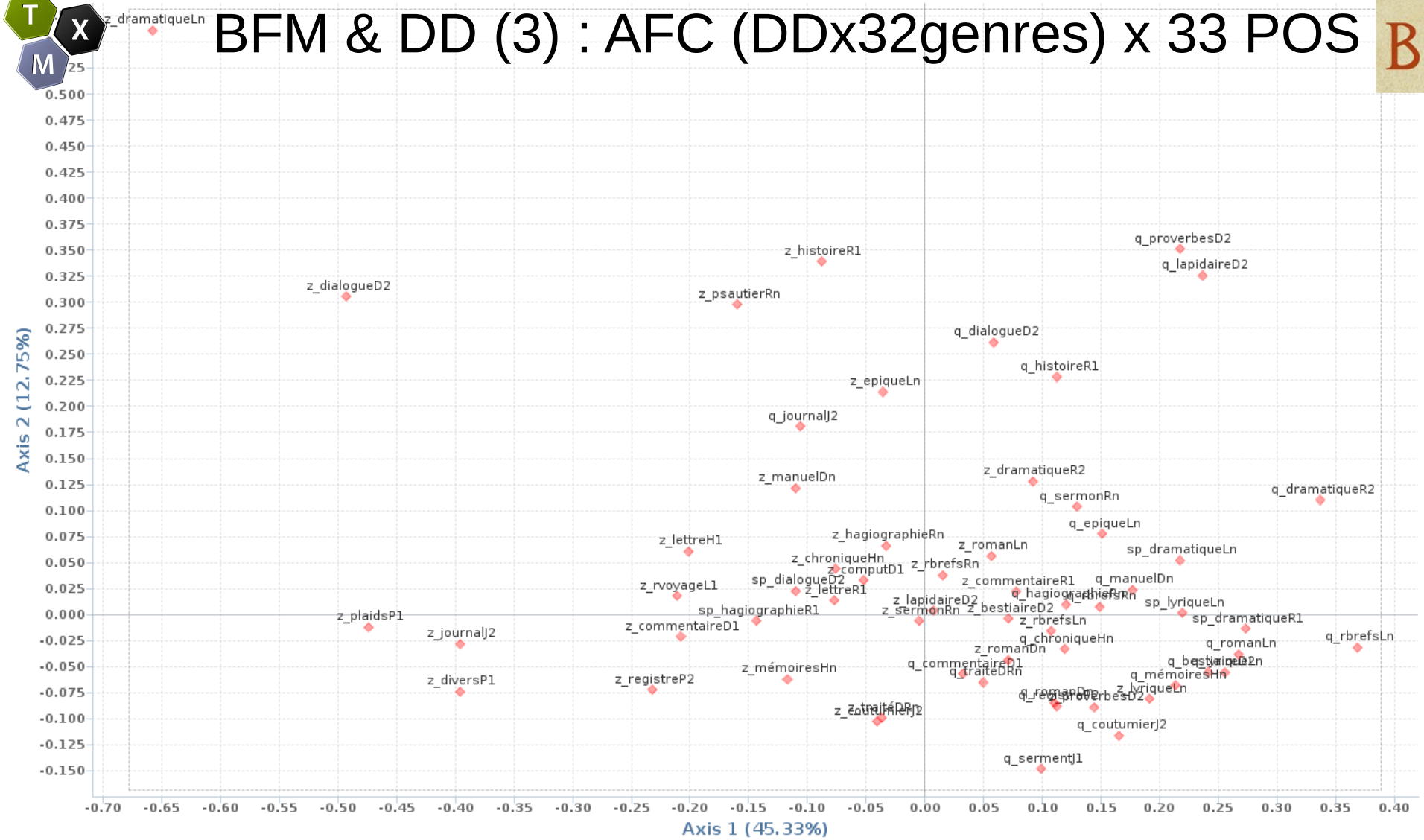
BFM & DD (3) : unités d’analyse

- Définition de 59 « unités de discours »
  - (q / sp / z) x 32 genres = 59 unités de discours (DU)
  - Notation : {DD : q, sp ou z}_{genre}{domaine : }{nb textes : 1, 2 ou n c’est-à-dire >2}
    ex. : `q_rbrefLn`, `sp_dramatiqueR2`, `z_lettreH1`
- Sélection des 33 POS descriptives
  - On écarte les mots étrangers, abréviations, ponctuations et marques éditoriales
    - `[fropos!="PON.*|ETR|OUT|RED|ABR"]`
  - Faible fiabilité et faible poids => Seuillage quantitatif (f > 1 700)
  - Non fiable (>25 % erreur) et Ctrb1 élevée (14,67%) => élimination de `NOMpro`
  - Neutralisation de catégories qui accaparent l’analyse :
    - LEDIT apparaissant fin XIIIe → fort effet sur axe 1 => élimination `DETcom` et `PRE.DETcom`
    - `ADJpos` fortement corrélé aux Psautiers → mobilise l’axe 2 => fusion `ADJpos` + `DETpos`

BFM & DD (3) : données pour l'AFC

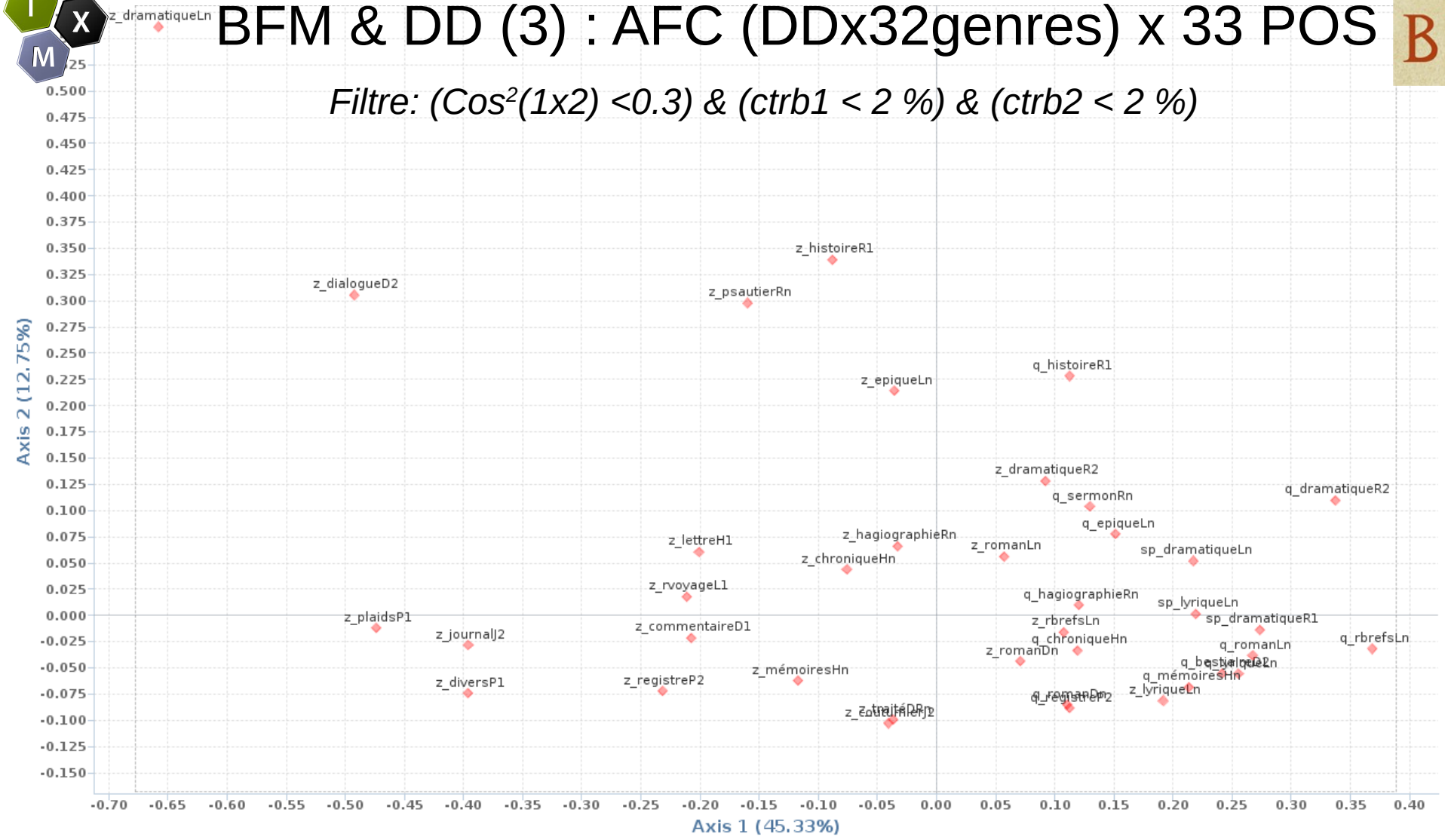
fropos	Frequ	q_bestiaireD2 t=1698	z_bestiaireD2 t=34705	q_chroniqueHn t=63436	z_chroniqueHn t=346440	q_commentaireD1 t=130	z_cc
NOMcom	655584	261	6289	10394	61254	22	
VERcjg	526369	272	5694	9065	46542	15	
PRE	385087	146	3318	6336	37924	19	
PROper	294126	211	2759	6928	23048	17	
ADVgen	235283	108	2208	4120	22906	2	
CONcoo	226973	111	2023	4289	25018	4	
DETdef	195153	57	1552	2686	23840	5	
CONsub	127472	72	1492	2496	12137	5	
VERppe	125911	33	903	1865	11267	4	
ADJqua	123262	62	1172	1965	9670	1	
VERinf	104566	68	1176	2068	8544	5	
PRorel	87659	53	912	1238	7266	1	
DETADJpos	82152	57	877	1888	7240	4	
ADVneg	75719	51	819	1493	4488	4	
PRE.DETdef	62212	12	414	679	6705	0	
PRodem	46559	17	559	707	3301	2	
PROadv	43411	17	313	697	3735	2	
DETind	39757	12	465	740	3738	2	
PROind	37841	15	422	587	3054	1	
DETndf	24413	6	156	196	2008	1	
DETdem	23628	7	173	388	1870	0	
VERppa	13012	3	132	158	1020	2	
DETcar	12933	0	83	167	1355	2	
ADJind	6290	1	53	96	649	0	

BFM & DD (3) : AFC (DDx32genres) x 33 POS



BFM & DD (3) : AFC (DDx32genres) x 33 POS

Filtre: (Cos²(1x2) <0.3) & (ctrb1 < 2 %) & (ctrb2 < 2 %)



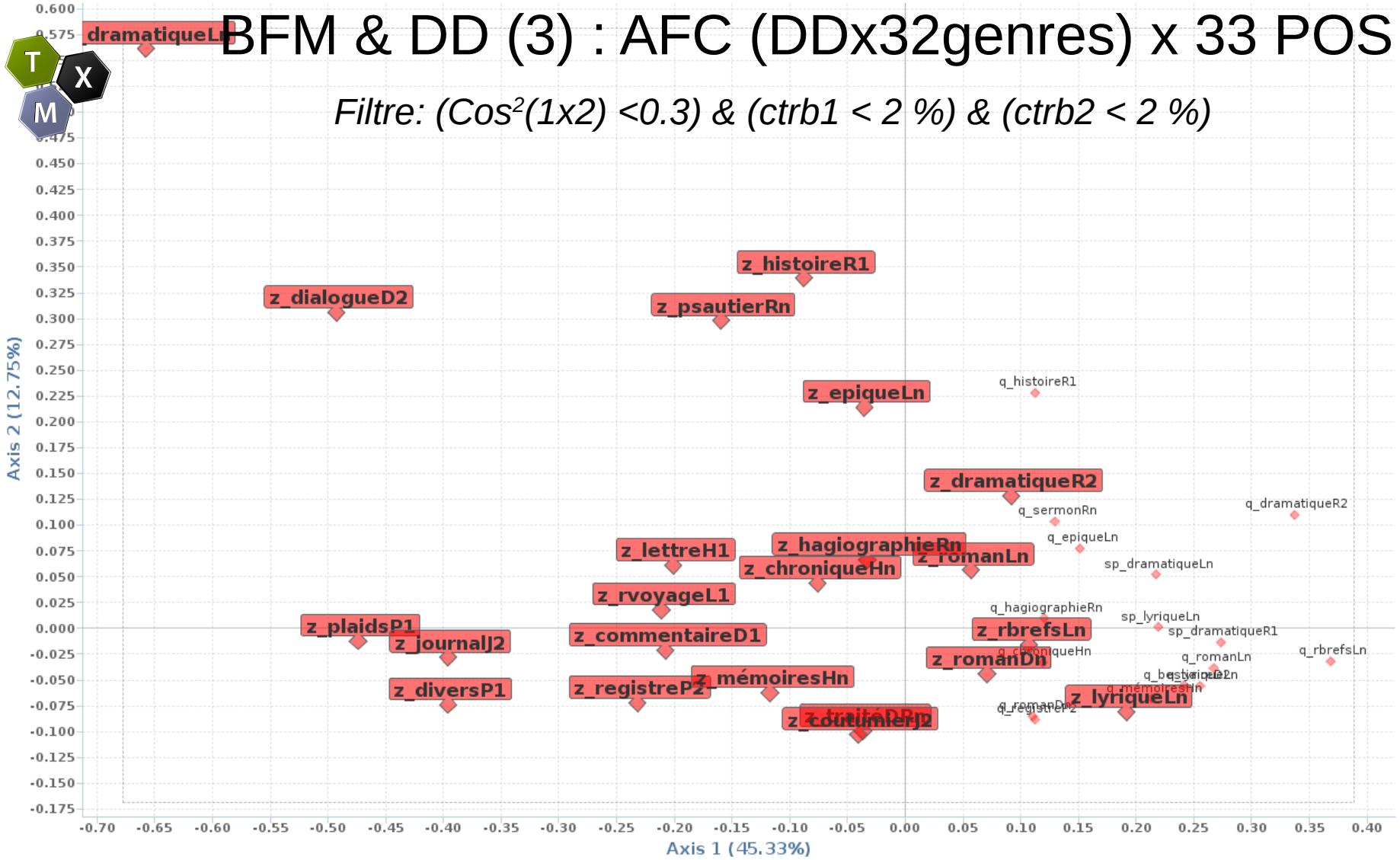
BFM & DD (3) : indicateurs de l'AFC

genresD_x_dd_alpha: [fropos!="PO | genresD_x_dd_alpha: fropos | genresD_x_dd_alpha:fropos

Colonnes	Q12	Q13	Q23	Mass	Dist	Cont1	Cos²1	Cont2	Cos²2	Cont3	Cos²3	c1	c2	c3
q_nouvelleL	,74	,77	,07	0,31	0,20	1,85	0,72	0,07	0,02	0,53	0,05	0,40	0,07	-0,11
q_fabliauL	,53	,52	,01	0,01	0,18	0,08	0,52	0,00	0,01	0,00	0,00	0,39	0,05	0,03
q_rbrefsL	,72	,71	,03	0,13	0,08	0,36	0,70	0,02	0,02	0,02	0,01	0,28	0,05	-0,03
q_dramatiqueR	,52	,21	,32	0,01	0,24	0,03	0,21	0,07	0,32	0,00	0,00	0,28	0,34	0,01
q_bestiaireamourD	,71	,79	,16	0,04	0,09	0,11	0,67	0,01	0,04	0,08	0,12	0,27	0,06	-0,11
sp_dramatiqueR	,80	,76	,06	0,32	0,08	0,83	0,75	0,09	0,05	0,04	0,01	0,26	0,07	-0,03
q_romanL	,90	,91	,06	10,44	0,08	26,82	0,88	1,18	0,03	4,41	0,04	0,26	0,04	-0,05
q_lyriqueL	,63	,70	,06	0,08	0,09	0,20	0,63	0,00	0,00	0,08	0,06	0,26	0,01	-0,08
q_narrationR	,58	,57	,02	0,11	0,09	0,25	0,57	0,01	0,01	0,01	0,01	0,25	0,04	-0,03
q_lapidaireD	,02	,11	,09	0,00	0,33	0,00	0,02	0,00	0,00	0,01	0,09	0,23	0,02	0,46
sp_lyriqueL	,49	,49	,03	0,02	0,06	0,05	0,48	0,00	0,01	0,01	0,02	0,22	0,03	-0,04
q_mémoiresH	,42	,51	,09	0,06	0,07	0,10	0,42	0,00	0,00	0,09	0,09	0,22	0,01	-0,10
sp_dramatiqueL	,63	,63	,09	0,60	0,05	0,99	0,59	0,10	0,04	0,32	0,05	0,21	0,05	0,06
z_lyriqueL	,55	,69	,14	2,49	0,05	3,52	0,55	0,00	0,00	3,61	0,14	0,19	0,00	-0,10
q_traitéR	,25	,30	,08	0,02	0,06	0,02	0,24	0,00	0,01	0,03	0,07	0,18	-0,04	-0,10
q_manuelD	,29	,28	,01	0,44	0,04	0,51	0,28	0,04	0,01	0,00	0,00	0,18	0,04	-0,01
q_chroniqueL	,36	,37	,04	0,08	0,04	0,10	0,34	0,01	0,01	0,03	0,03	0,18	0,03	0,05

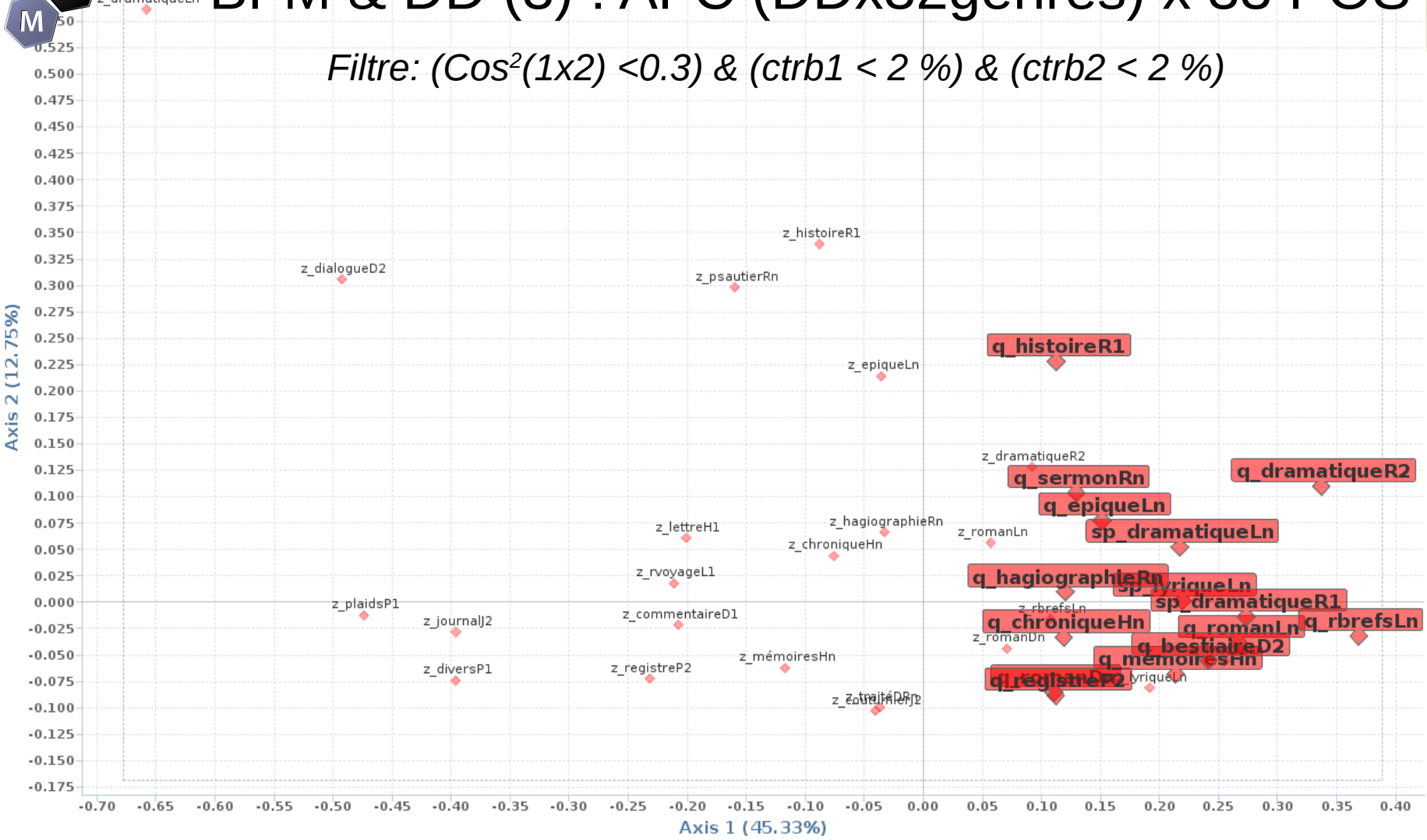
BFM & DD (3) : AFC (DDx32genres) x 33 POS

Filtre: (Cos²(1x2) <0.3) & (ctrb1 < 2 %) & (ctrb2 < 2 %)



BFM & DD (3) : AFC (DDx32genres) x 33 POS

Filtre: (Cos²(1x2) <0.3) & (ctrb1 < 2 %) & (ctrb2 < 2 %)



BFM & DD (3) : AFC (DDx32genres) x 33 POS

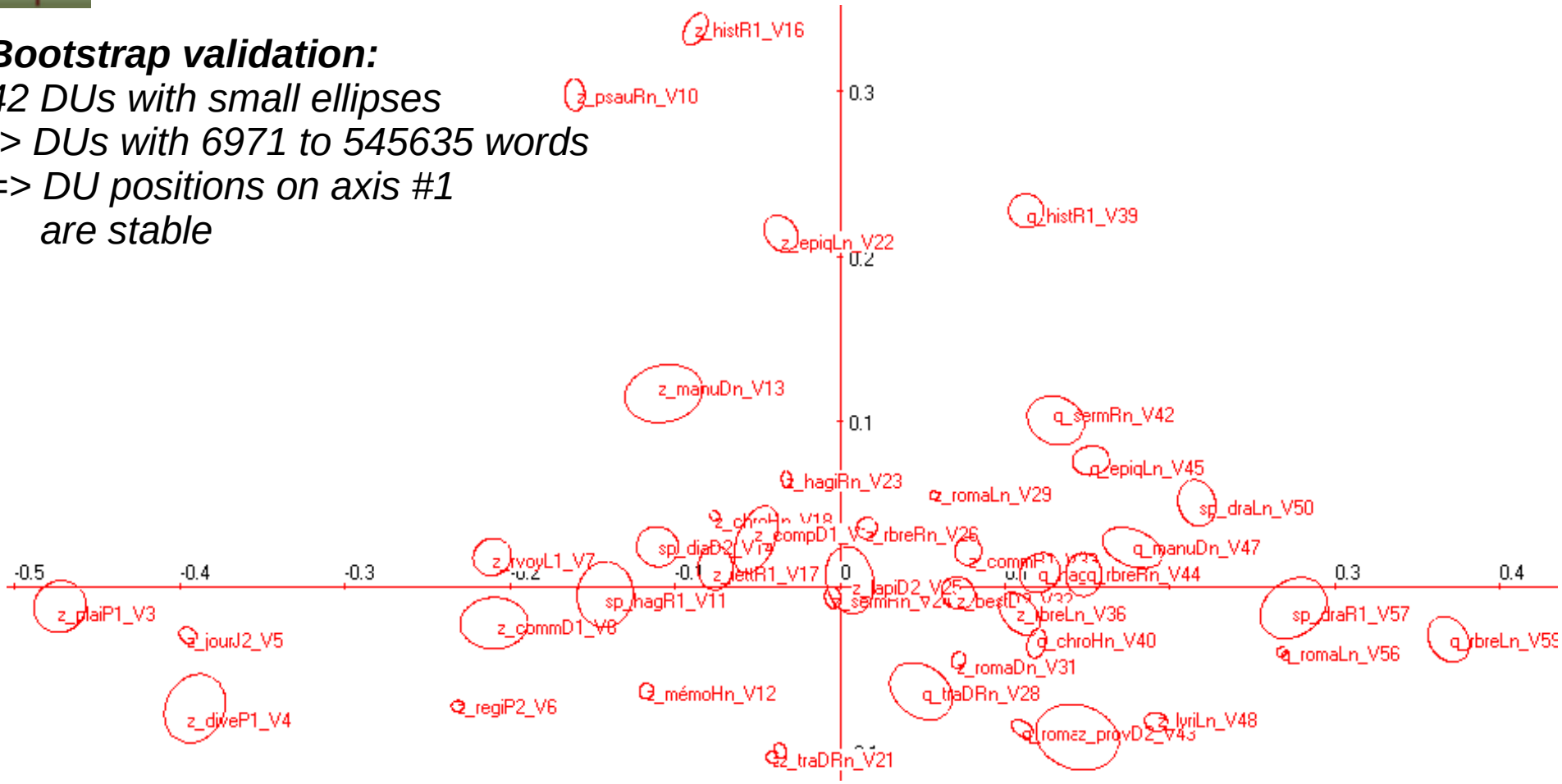
Bootstrap validation:

42 DUs with small ellipses

-> DUs with 6971 to 545635 words

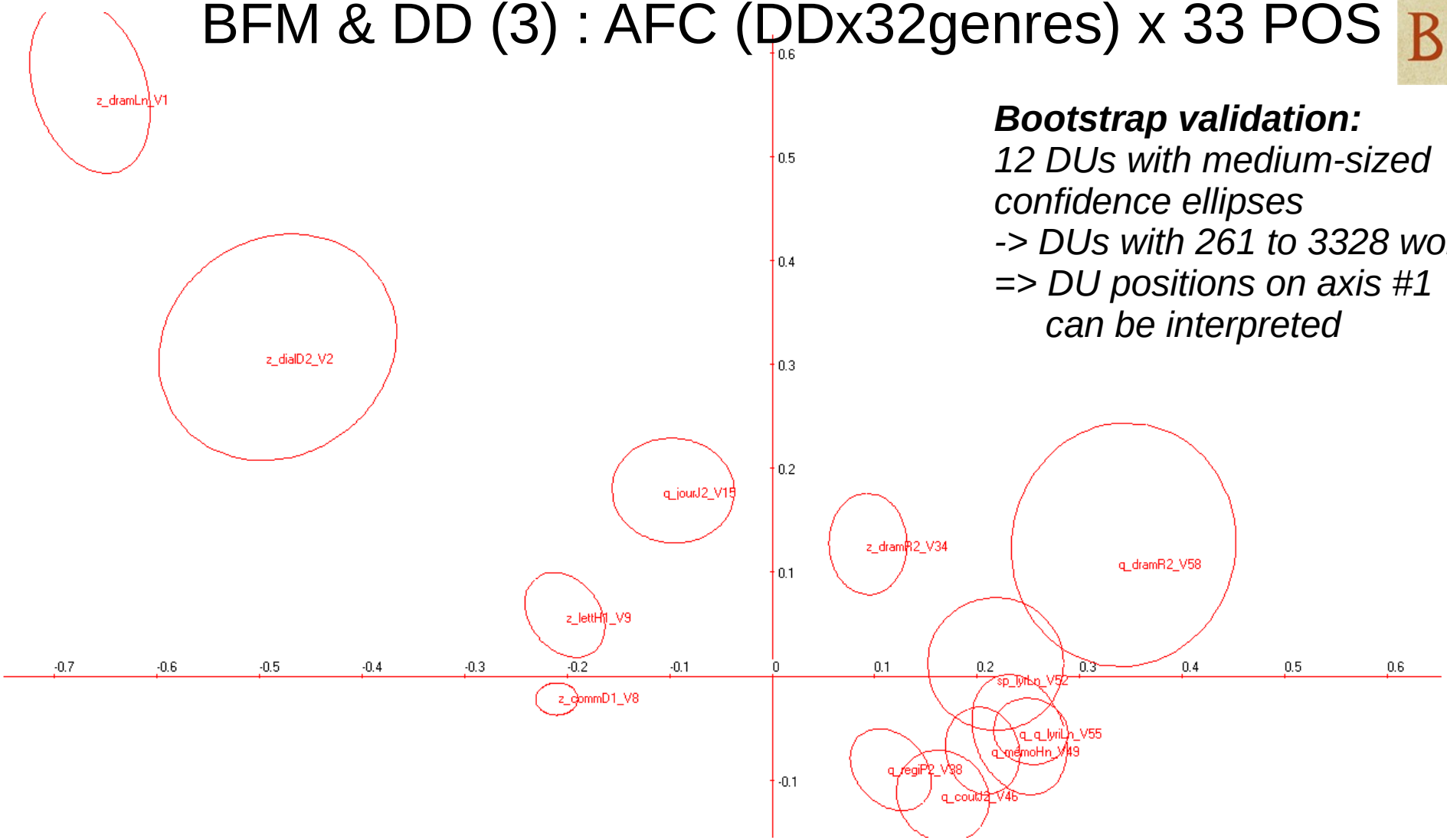
=> DU positions on axis #1

are stable



BFM & DD (3) : AFC (DDx32genres) x 33 POS

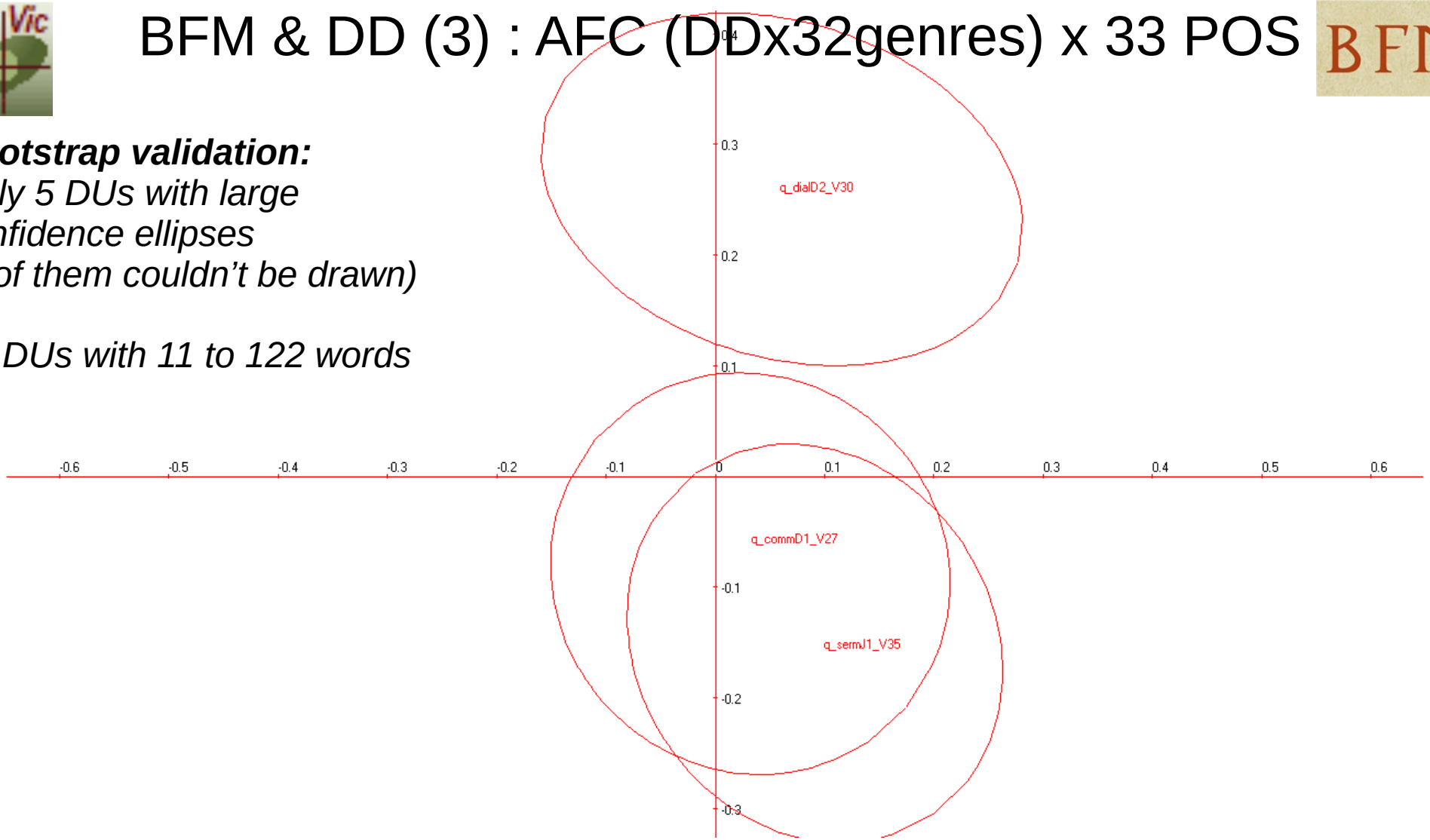
Bootstrap validation:
12 DUs with medium-sized confidence ellipses
-> DUs with 261 to 3328 words
=> DU positions on axis #1 can be interpreted



BFM & DD (3) : AFC (DDx32genres) x 33 POS

Bootstrap validation:
Only 5 DUs with large confidence ellipses
(2 of them couldn't be drawn)

=> DUs with 11 to 122 words



BFM & DD (3) : gradient d'oralité

rank	DU	c1	Ctrb1	Cos²1	Q12	Mass	Word Nb	Ellipse size
1	z_dramatiqueLn	-0,66	0,16	**0,31**	0,53	0,01	334	medium
2	z_dialogueD2	-0,49	0,07	**0,38**	0,52	0,01	261	medium
3	z_plaidsP1	-0,47	**2,46**	**0,72**	0,72	0,28	10153	small
4	z_diversP1	-0,40	**1,53**	**0,70**	0,72	0,25	9025	small
5	z_journalJ2	-0,40	**21,26**	**0,86**	0,86	**3,51**	125563	small
6	z_registreP2	-0,23	**18,16**	**0,69**	0,76	**8,73**	312528	small
7	z_rvoyageL1	-0,21	**1,13**	**0,36**	0,36	0,66	23469	small
8	z_commentaireD1	-0,21	0,35	**0,70**	0,71	0,21	7573	small
9	z_lettreH1	-0,20	0,10	**0,39**	0,42	0,06	2263	medium
10	z_psautierRn	-0,16	**1,27**	0,10	0,46	**1,29**	46058	small
11	sp_hagiographieR1	-0,14	0,35	0,14	0,14	0,44	15732	small
12	z_mémoiresHn	-0,12	**2,86**	**0,31**	0,40	**5,40**	193101	small
13	z_manuelDn	-0,11	0,09	0,07	0,16	0,19	6971	small
14	sp_dialogueD2	-0,11	0,31	0,17	0,18	0,65	23405	small
15	q_journalJ2	-0,11	0,01	0,04	0,14	0,03	1131	medium
16	z_histoireR1	-0,09	0,45	0,04	0,68	**1,50**	53537	small
17	z_lettreR1	-0,08	0,19	0,13	0,13	0,84	30003	small
18	z_chroniqueHn	-0,08	**2,06**	**0,32**	0,42	**9,24**	330617	small
19	z_computD1	-0,05	0,04	0,02	0,03	0,39	14055	small
20	z_coutumierJ2	-0,04	0,28	0,04	0,29	**4,37**	156475	small
21	z_traitéDRn	-0,04	0,25	0,03	0,23	**4,91**	175799	small
22	z_epiqueLn	-0,04	0,06	0,02	0,58	**1,22**	43785	small
23	z_hagiographieRn	-0,03	0,15	0,08	0,42	**3,44**	123094	small
24	z_sermonRn	0,00	0,00	0,00	0,00	**3,26**	116703	small
25	z_lapidaireD2	0,01	0,00	0,00	0,00	0,39	13781	small
26	z_rbrefsRn	0,02	0,02	0,01	0,08	**2,29**	81911	small
27	q_commentaireD1	0,03	0,00	0,00	0,01	0,00	122	**large**
28	q_traitéDRn	0,05	0,05	0,05	0,14	0,50	17960	small
29	z_romanLn	0,06	**1,91**	0,20	0,41	**15,25**	545635	small
30	q_dialogueD2	0,06	0,00	0,01	0,15	0,00	91	**large**
31	z_romanDn	0,07	0,81	**0,36**	0,50	**4,20**	150333	small
32	z_bestiaireD2	0,07	0,19	0,28	0,28	0,95	34160	small
33	z_commentaireR1	0,08	0,30	0,10	0,11	**1,26**	45201	small
34	z_dramatiqueR2	0,09	0,02	0,10	0,31	0,06	2319	medium

[Orality - -]

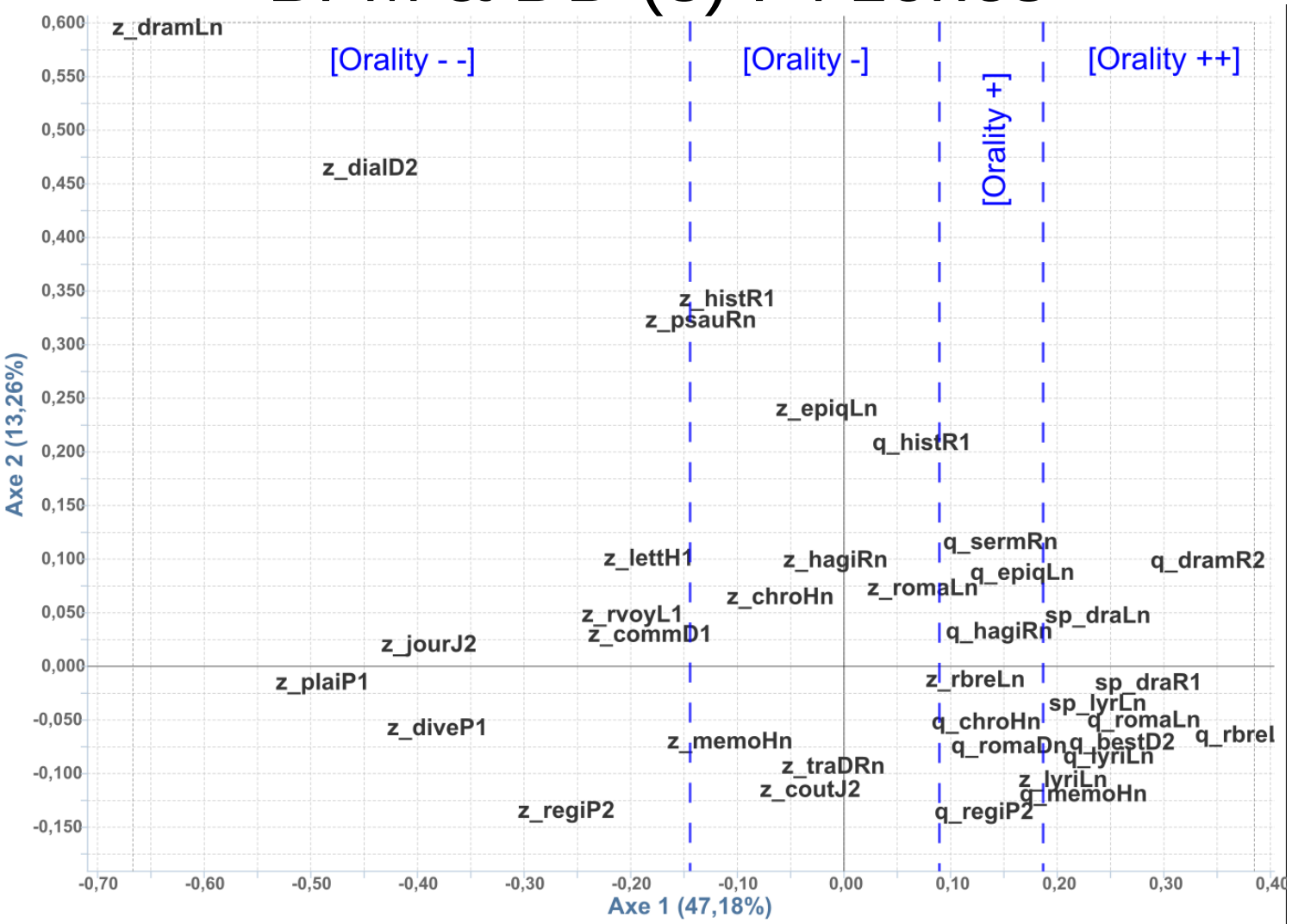
[Orality -]

[Orality +]

35	q_sermentJ1	0,10	0,00	0,05	0,16	0,00	101	**large**
36	z_rbrefsLn	0,11	0,37	**0,34**	0,35	0,83	29696	small
37	q_romanDn	0,11	**1,20**	**0,45**	0,72	**2,56**	91489	small
38	q_registreP2	0,11	0,05	0,20	0,33	0,09	3328	medium
39	q_histoireR1	0,11	0,42	0,09	0,46	0,87	30976	small
40	q_chroniqueHn	0,12	0,94	**0,51**	0,55	**1,72**	61578	small
41	q_hagiographieRn	0,12	0,54	**0,48**	0,48	0,96	34236	small
42	q_sermonRn	0,13	0,23	0,21	0,35	0,35	12485	small
43	z_proverbesD2	0,14	0,16	0,15	0,21	0,20	7195	small
44	q_rbrefsRn	0,15	0,48	0,25	0,26	0,57	20254	small
45	q_epiqueLn	0,15	0,88	**0,36**	0,46	**1,00**	35811	small
46	q_coutumierJ2	0,17	0,07	0,16	0,24	0,07	2375	medium
47	q_manuelDn	0,18	0,53	0,28	0,28	0,44	15758	small
48	z_lyriqueLn	0,19	**3,54**	**0,54**	0,64	**2,49**	89262	small
49	q_mémoiresHn	0,21	0,11	**0,42**	0,46	0,06	2150	medium
50	sp_dramatiqueLn	0,22	**1,09**	**0,62**	0,66	0,60	21387	small
51	q_proverbesD2	0,22	0,00	0,07	0,27	0,00	25	**very large**
52	sp_lyriqueLn	0,22	0,05	**0,49**	0,49	0,02	892	medium
53	q_lapidaireD2	0,24	0,00	0,02	0,07	0,00	11	**very large**
54	q_bestiaireD2	0,24	0,10	**0,64**	0,67	0,05	1664	medium
55	q_lyriqueLn	0,26	0,20	**0,62**	0,65	0,08	2828	medium
56	q_romanLn	0,27	**28,80**	**0,91**	0,92	**10,44**	373570	small
57	sp_dramatiqueR1	0,27	0,92	**0,81**	0,81	0,32	11435	small
58	q_dramatiqueR2	0,34	0,05	**0,37**	0,41	0,01	409	medium
59	q_rbrefsLn	0,37	**2,37**	**0,78**	0,78	0,45	16115	small

[Orality + +]

BFM & DD (3) : 4 zones



BFM & DD (3) : genres et discours direct

BFM

[+ oralité]

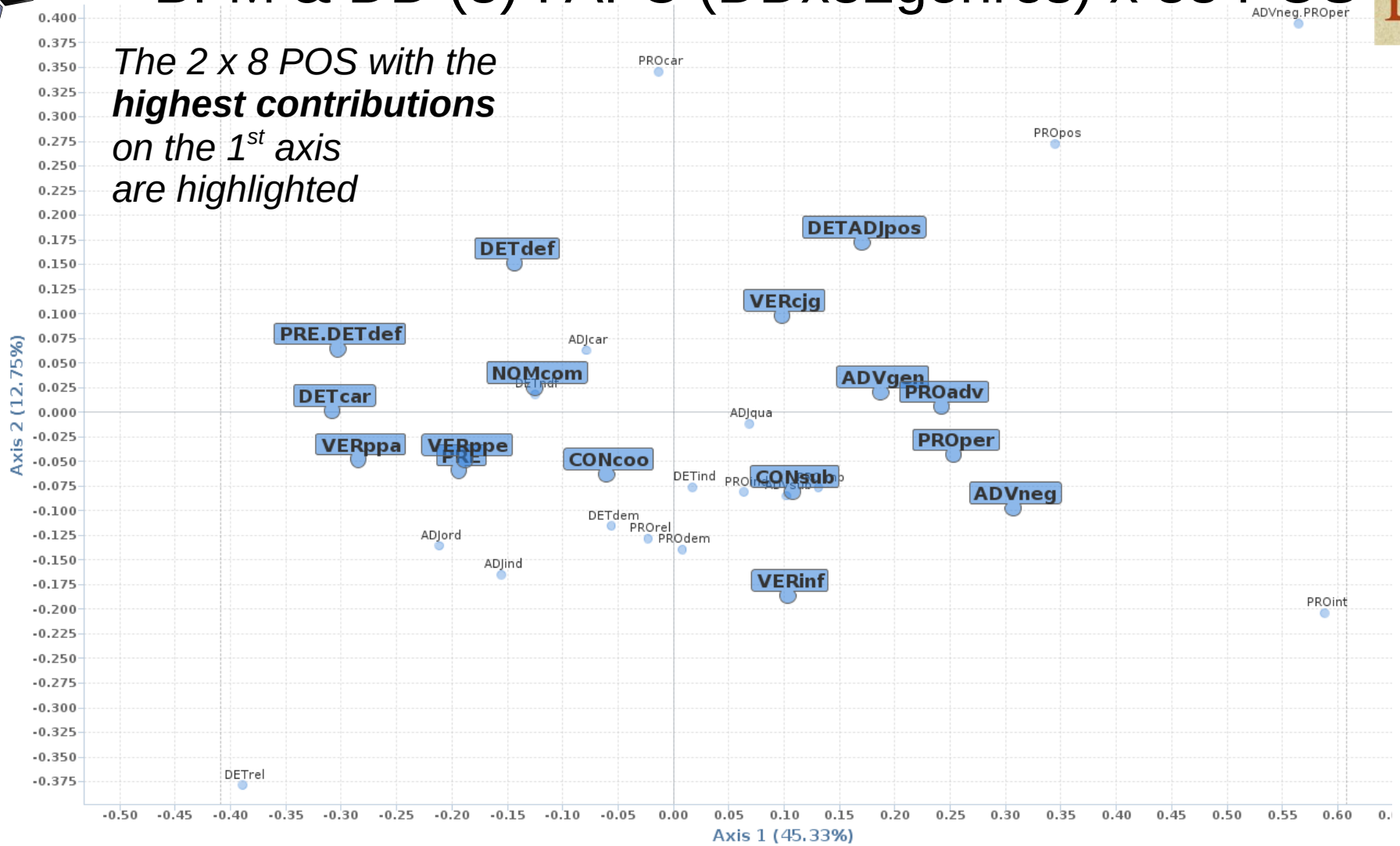
- Des genres oralisés : Théâtre, fabliau
- D'autres genres narratifs brefs (nouvelle, récit bref)
- Oral représenté
- Genres du domaine littéraire

[- oralité]

- Aucun genre oralisé ; pas d'oral représenté
- Genres des domaines juridiques et actes de la pratique (chartes, registres, comptes…)
- Apparaît plus hétérogène, plus dispersé.

BFM & DD (3) : AFC (DDx32genres) x 33 POS

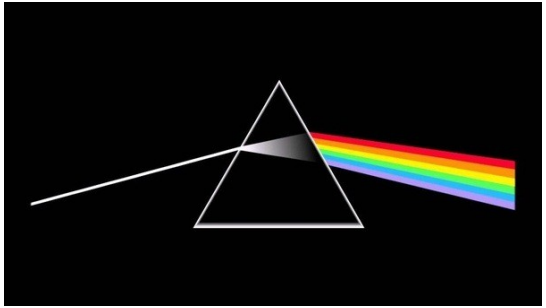
The 2 x 8 POS with the ***highest contributions*** *on the 1st axis are highlighted*



BFM & DD (3) : caractérisation par les POS contribuant le plus à l'axe 1

[- oralité]	[+ oralité]
Preposition	Personal pronoun
Common noun	General adverb
Preposition + definite determiner	Negative adverb
Past participle	Finite verb
Definite determiner	Adverbial pronoun (*en*, *y*)
Cardinal determiner	Possessive determiner or adjective
Present participle	Subordinating conjunction
Coordinating conjunction	Infinitive verb

BFM & DD (3) : l'AFC comme outil de décomposition de l'information



- L'AFC est souvent mobilisée comme outil de visualisation (2D ou plus)
- Mais c'est de fait aussi un outil d'analyse en composantes, en facteurs
- Le 1er facteur est mathématiquement celui qui représente la principale dimension de variation
- L'oralité apparaît comme cette dimension dominante
  - cf. les travaux de Biber, dès 1988, et notamment article de 1994 « The ubiquitous oral versus literate dimension: A survey of multidimensional studies ». *In* J. Connor-Linton & L.W. Amoroso (Eds), *Measured language: Quantitative studies of acquisition, assessment, and variation,* Washington DC: Georgetown University Press, p. 1-20.

BFM & DD (3) : confronter l'oralité à d'autres gradients possibles

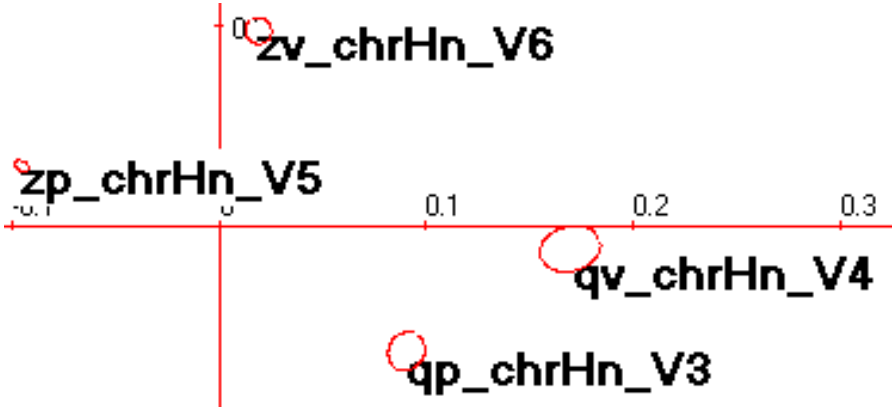
- Nos unités de discours ne peuvent exprimer que des variations sur le discours direct et les genres.
- Il faudrait construire des unités de discours avec d'autres « degrés de liberté », d'autres dimensions de variation possibles et pouvant être captées par un gradient (1D).
- Expériences / mises à l'épreuve :
  - **Forme** : vers (v) / prose (p) / mixte (m). 3DD x 32 genres x 3 formes = 76 DU (+ 4 de 18 à 27 mots)
  - **Textes** (non groupés par genre) : 3DD x 137 textes = 247 DU (+ 9 de 10 à 27 mots)
  - **Diachronie** : 238 DU colorées selon 4 siècles (on a retiré 9 DU des Xe et XIe s.)
  - **Échelle** : cf. (Mazziotta & Glickman, 2019). 3DD x 3 textes = 6DU *vs* 247 DU

=> On calcule l'AFC et on regarde si sur l'axe 1 les paires z/q et z/sp sont toujours dans le même sens, ce qui montre que l'axe 1 s'oriente toujours selon l'oralité.

BFM & DD (3) : oralité et forme

Méthode :

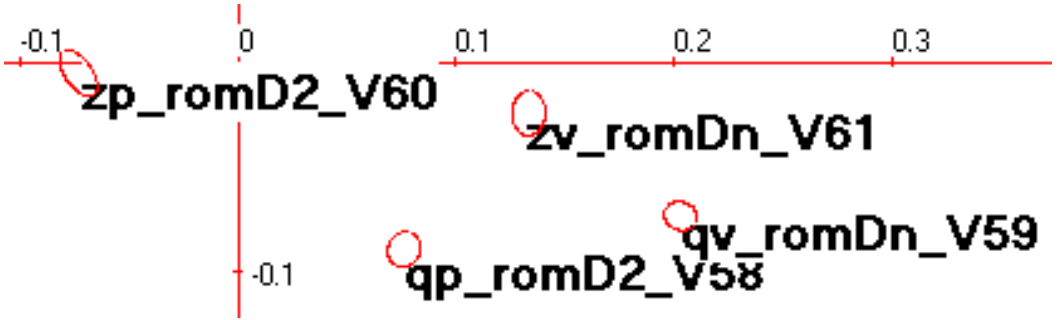
- 9 genres ont plusieurs formes, et 6 d'entre eux ont une variation DD / non-DD.
- Dans l'AFC générale on regarde la position relative des variations DD/non-DD et forme à l'intérieur d'un genre.



Genre Chroniques Historiques

Illustre la configuration typique :

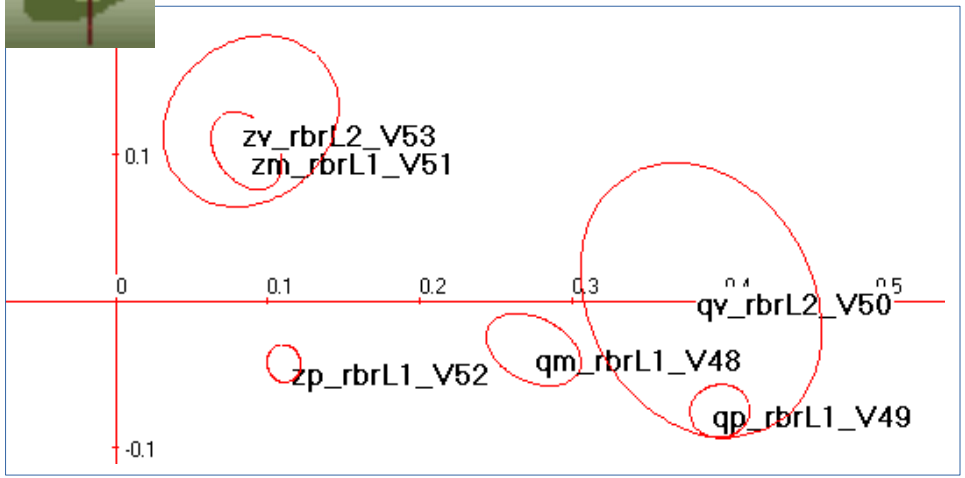
- D'abord z avant q,
- **Puis** généralement p (avant m) avant v.

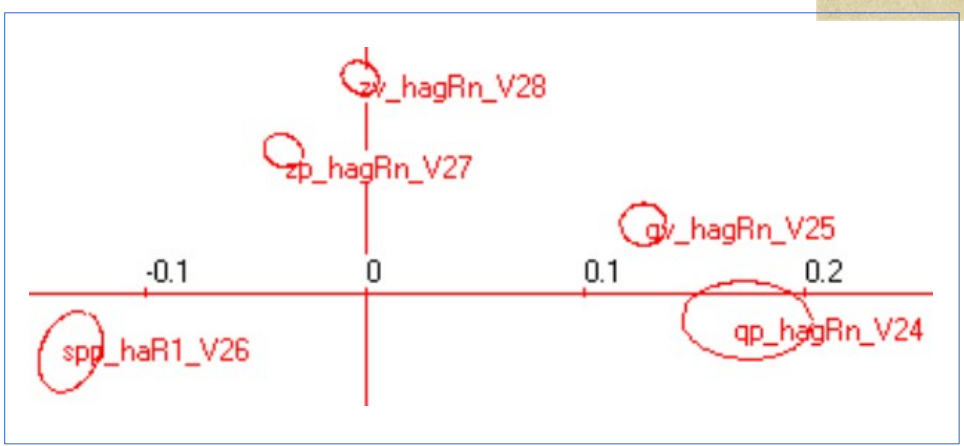


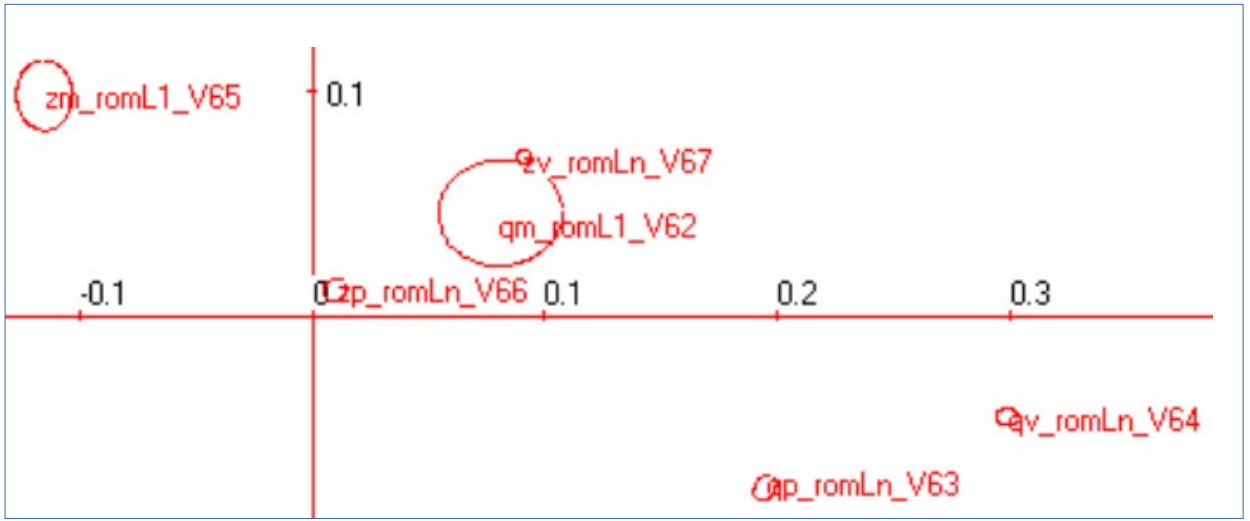
Genre Romans didactiques

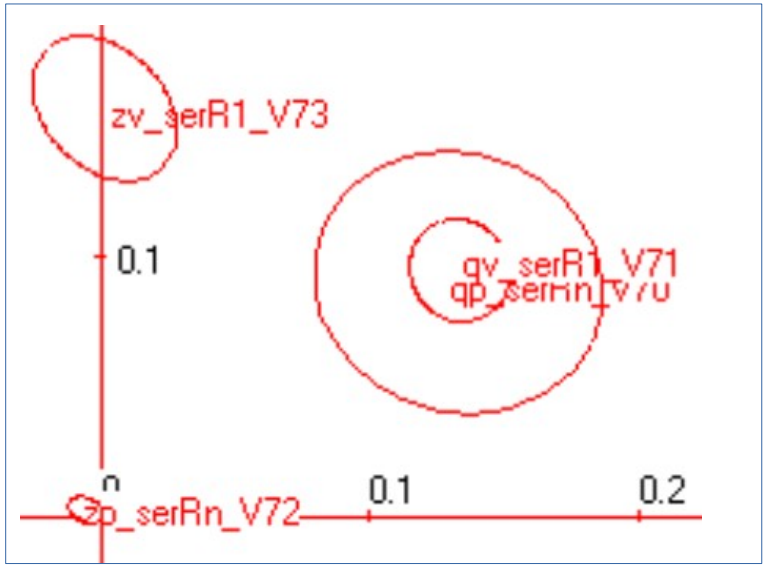
- Seul cas d'inversion significative (ellipses) entre z et q
- Le roman didactique en vers correspond aux différentes parties du *Roman de la Rose*, qui pourrait constituer un genre à lui seul (très différent de RomD en prose).

BFM & DD (3) : oralité et forme

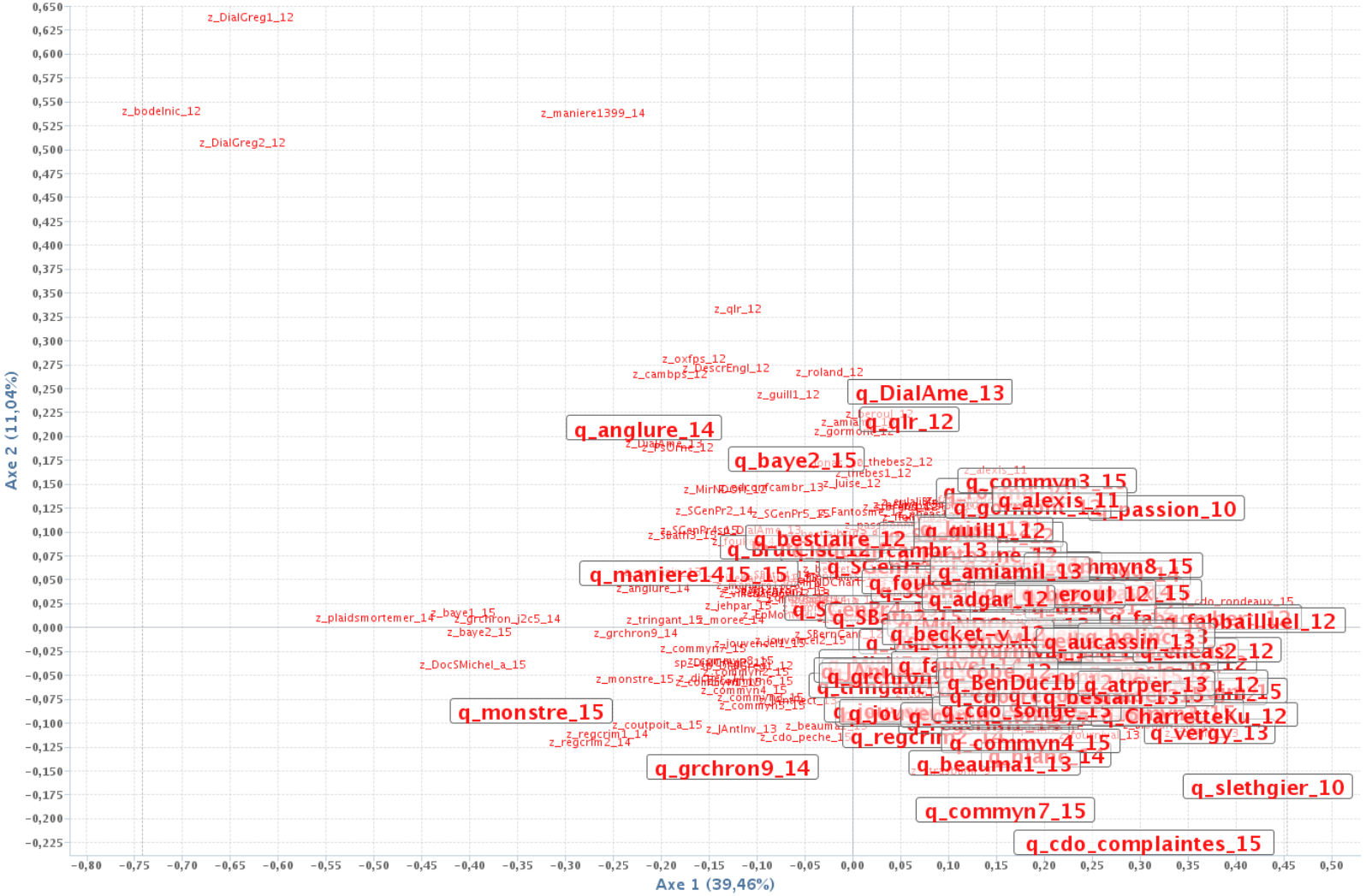




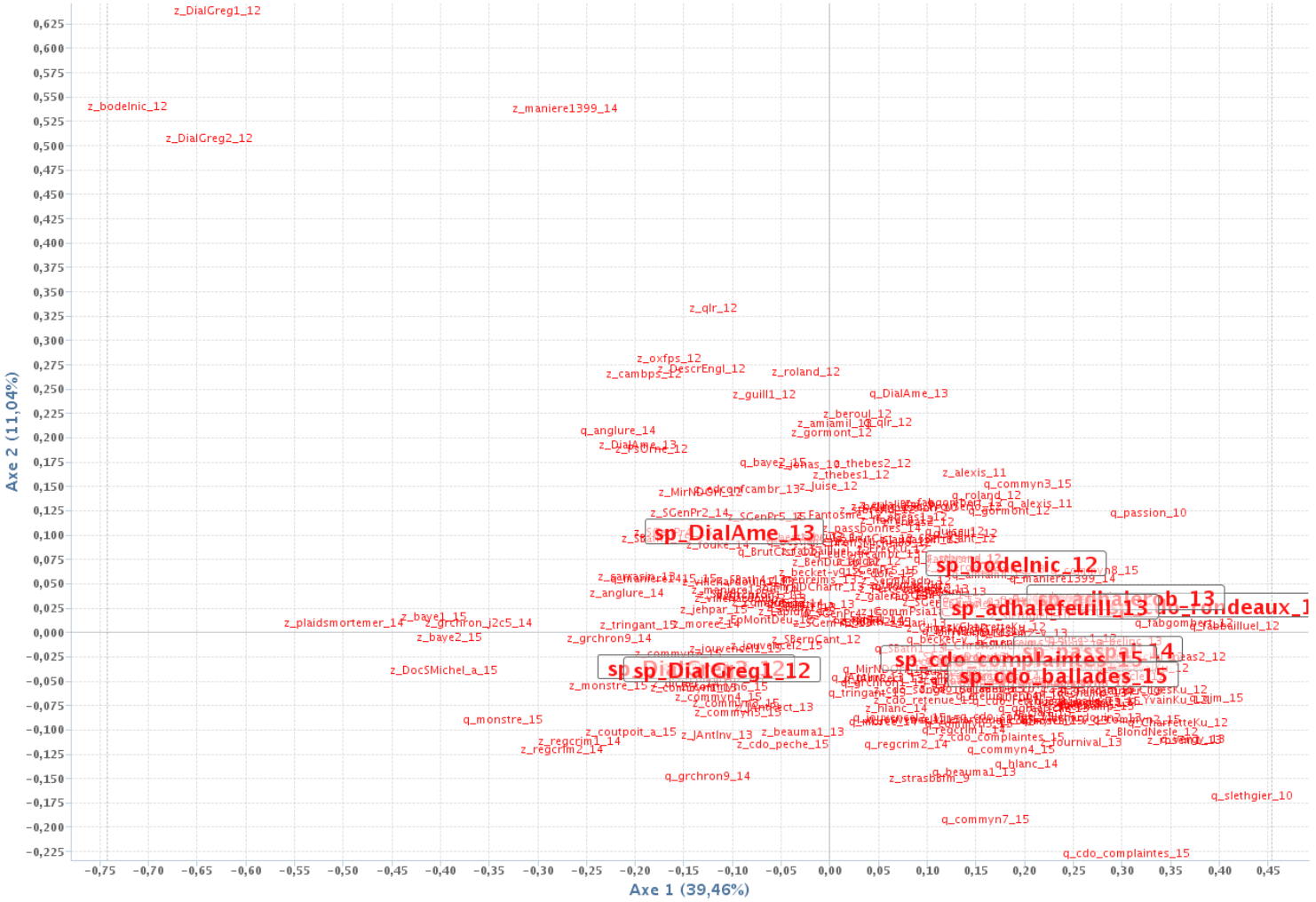




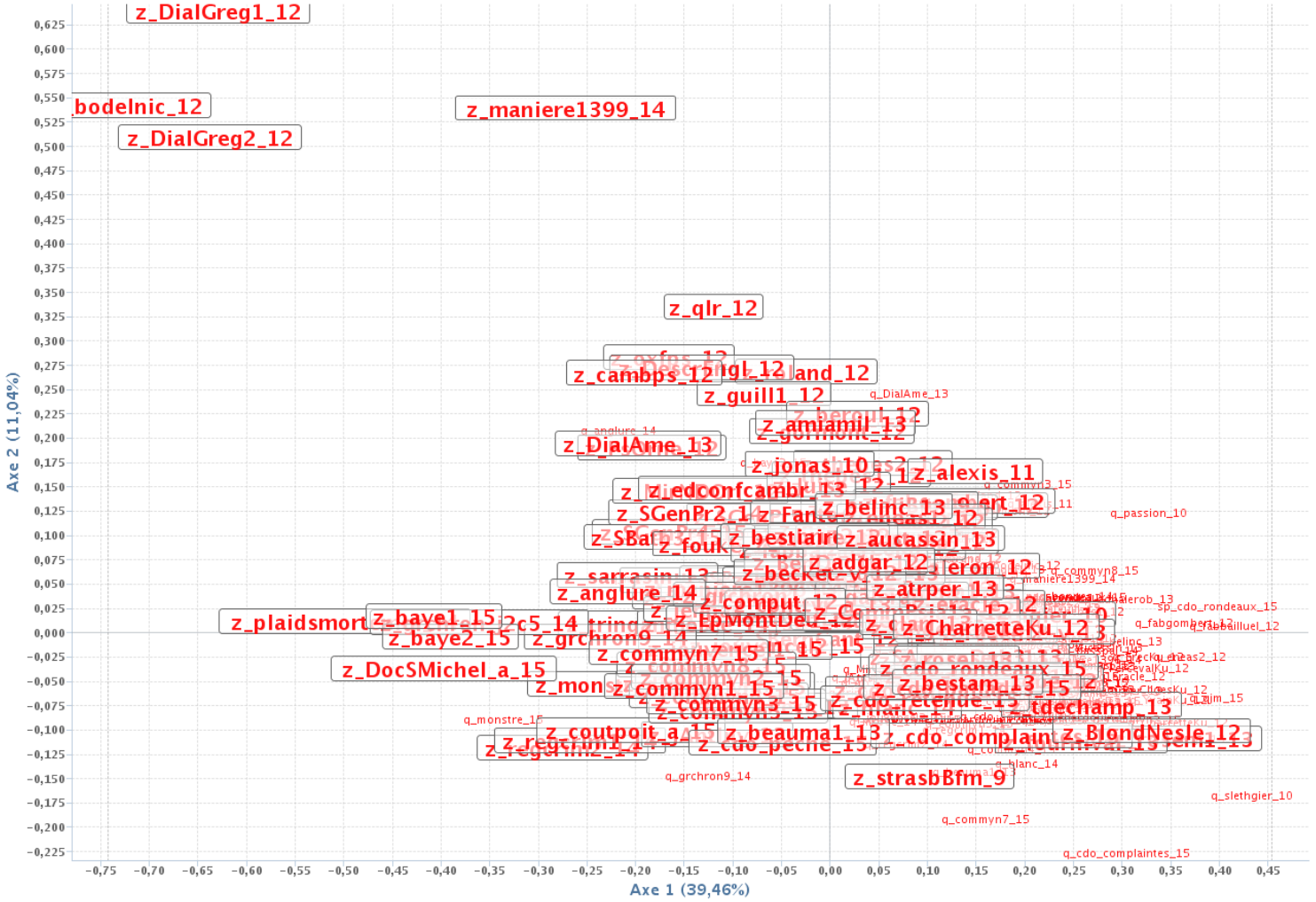
BFM & DD (3) : AFC (DDx137**Textes**) x 33POS


Axe 2 (11,04%)
Axe 1 (39,46%)
z_DialGreg1_12
z_bodelnic_12
z_maniere1399_14
z_DialGreg2_12
z_qlr_12
z_oxfps_12
z_roland_12
z_guill1_12
q_DialAme_13
q_anglure_14
q_qlr_12
q_baye2_15
q_commyn3_15
q_alexis_11
q_passion_10
q_bestiaire_12
q_maniere1415_15
q_amiamil_13
q_adgar_12
q_fabbailluel_12
q_becket-v_12
q_aucassin_133
q_BenDuc1b
q_atrper_13
q_monstre_15
q_cdo_songe_13
q_CharretteKu_12
q_vergy_13
q_commyn4_15
q_grchron9_14
q_beauma1_13
q_slethgier_10
q_commyn7_15
q_cdo_complaintes_15
z_plaidsmortemer_14
z_grchron_j2c5_14
z_bayel_15
z_bayel_2_15
z_grchron9_14
z_DocSMichel_a_15
z_monstre_15
z_coutpoit_a_15
z_regcrim1_14
z_regcrim2_14
z_JAntInv_13
z_tringant_15
z_moree_14
z_anglure_14
z_commyn7_15
z_commyn4_15
z_commyn5_15
z_commyn6_15
z_commyn2_15
z_commyn3_15
z_commyn8_15
z_commyn9_15
z_commyn1_15

BFM & DD (3) : AFC (DDx137**Textes**) x 33POS


Axe 2 (11,04%)
Axe 1 (39,46%)
sp_DialAme_13
sp_bodelnic_12
sp_adhalefeuill_13
sp_cdo_complaintes_15
sp_cdo_ballades_15
sp sp_DialGreg1_12
z_DialGreg1_12
z_bodelnic_12
z_maniere1399_14
z_DialGreg2_12
z_qlr_12
z_oxfps_12
z_roland_12
q_DialAme_13
q_anglure_14
q_passion_10
q_commyn3_15
q_commyn7_15
q_cdo_complaintes_15
q_slethgier_10
q_grchron9_14
q_monstre_15
z_DocSMichel_a_15
z_plaidsmortemer_14
z_grchron_j2c5_14
z_baye2_15
z_grchron9_14
z_monstre_15
z_coutpoit_a_15
z_regcrim1_14
z_regcrim2_14
z_JAntInv_13
z_beauma1_13
z_cdo_peche_15
q_regcrim2_14
q_commyn4_15
q_blanc_14
z_strasbrm_9
z_journival_13
z_BlondNesle_13
z_alexis_11
z_thebes1_12
z_guill1_12
z_beroul_12
z_gormont_12
z_amiamil_10
z_cambps_12
z_DescrEngl_12
z_SCenPr2_14
z_jehpar_15
z_tringant_12
z_EpMontDeu_15
z_SBernGant_12
z_commyn4_15
z_commyn5_13
z_commyn7_15
z_cdo_retenue_14
z_cdo_complaintes_15
z_commyn3_15
z_Fantosme_12
z_Juise_12
z_becket+vq_15
z_galeRad_13
z_passbonnes_14
z_ROUTE_14
z_BehDue_16
z_maniere1399_14
z_anglure_14
z_bayel_15
z_baye2_15
z_grchron9_14
z_monstre_15
z_commyn4_15
z_commyn5_13
z_commyn7_15
z_cdo_complaintes_15
z_cdo_peche_15
z_beauma1_13
z_JAntInv_13
z_regcrim1_14
z_regcrim2_14
z_coutpoit_a_15
z_monstre_15
z_DocSMichel_a_15
z_plaidsmortemer_14
z_grchron_j2c5_14
z_baye2_15
z_grchron9_14
z_tringant_12
z_jehpar_15
z_EpMontDeu_15
z_SBernGant_12
z_commyn4_15
z_commyn5_13
z_commyn7_15
z_cdo_complaintes_15
z_cdo_peche_15
z_beauma1_13
z_JAntInv_13
z_regcrim1_14
z_regcrim2_14
z_coutpoit_a_15
z_monstre_15
z_DocSMichel_a_15
z_plaidsmortemer_14
z_grchron_j2c5_14
z_baye2_15
z_grchron9_14
z_monstre_15

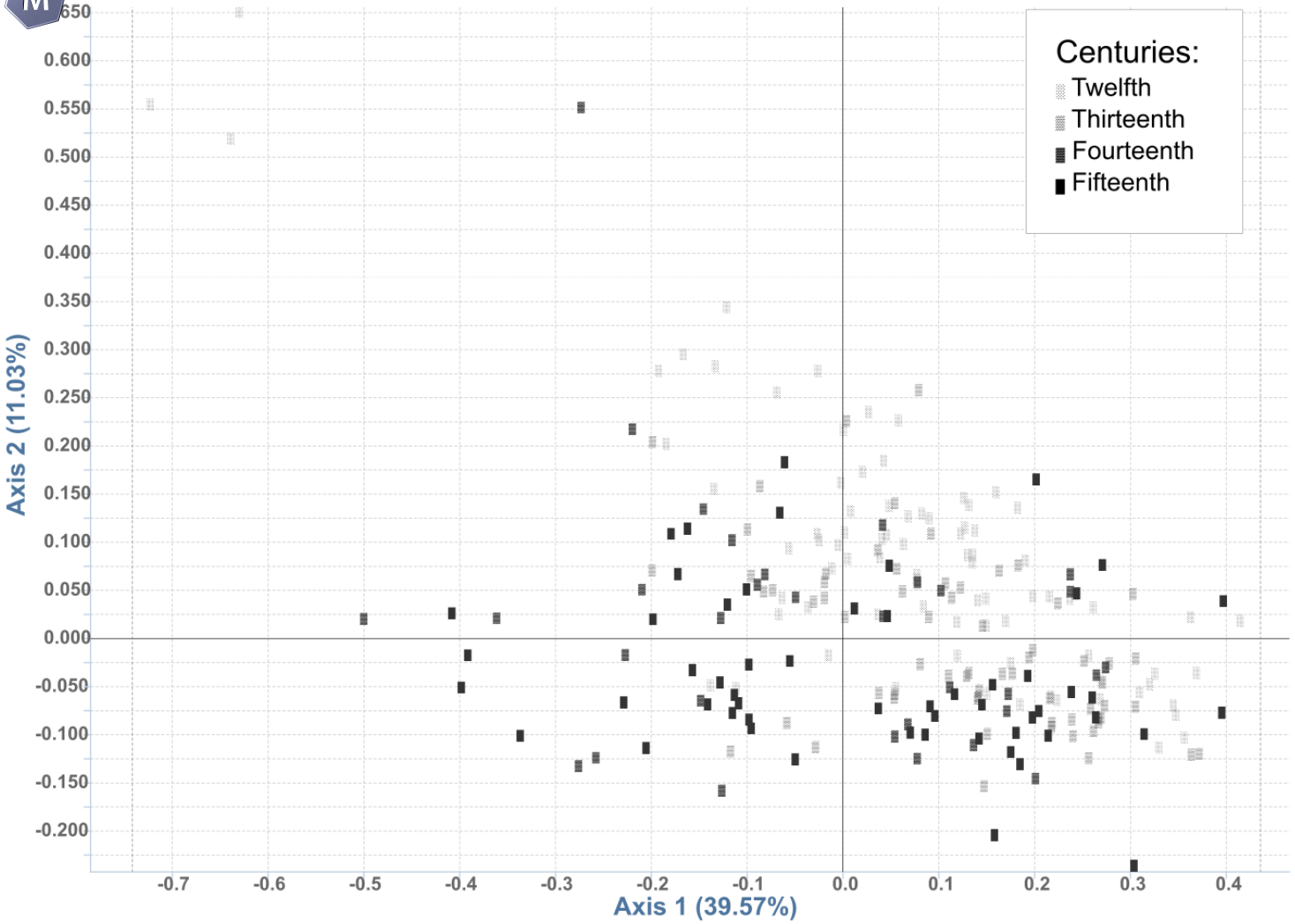
BFM & DD (3) : AFC (DDx137**Textes**) x 33POS



BFM & DD (3) : oralité sur les 247 DU

- Pour les points représentant des DU d'au moins 1000 mots (pour que la position dans l'AFC soit stable), on calcule la différence de coordonnées sur l'axe 1 entre le z et le q d'un même texte (pas de paire z/sp avec cette condition de taille)
- Il n'y a que 4 inversions sur 75 textes (paires de DU q et z) concernées, avec les écarts suivants :
  - z_rosem1_13 -0,12 → 1er volume du *Roman de la Rose*, déjà remarqué pour [+ oralité]
  - z_monstre_15 -0,11 → *Chronique* de Monstrelet, texte historique dont le DD a clairement des traits du pôle [- oralité]
  - z_rosem3_13 -0,03
  - z_tdechamp_13 -0,01

BFM & DD (3) : diachronie



- Diachronie plutôt sur l'axe 2
- Se rappeler cependant que nous avons écarté de l'analyse le déterminant composé *ledit* → cette analyse ne tient pas compte de son apparition au XIIIe s.

BFM & DD (3) : oralité et échelle

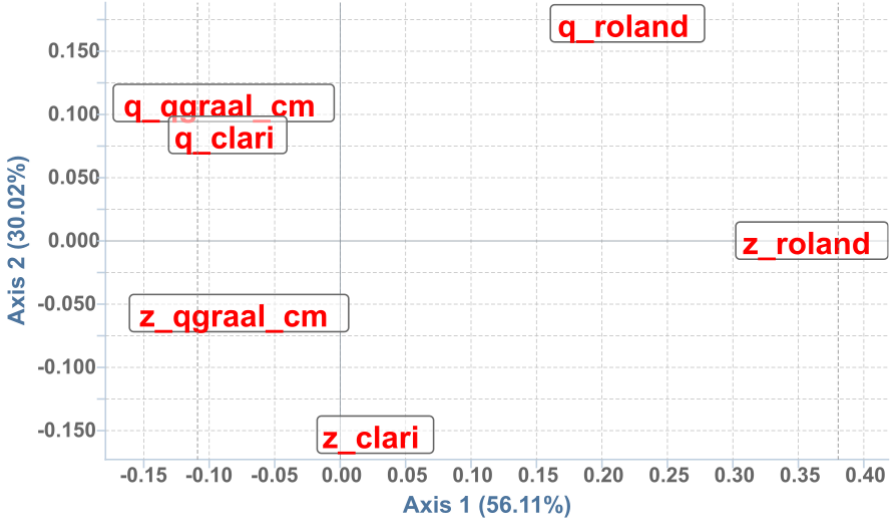
Mazziotta et Glickman (Linx, 78, 2019) étudient le contraste syntaxique entre oral représenté et narration dans 3 textes de genres différents :

- *Chanson de Roland* (`roland`),
- *Conquête de Constantinople* (`clari`),
- *Queste del Saint Graal* (`qgraal`).

Ils concluent que « L'opposition entre `roland` et les autres textes est plus forte que l'opposition entre oral représenté et narration ».

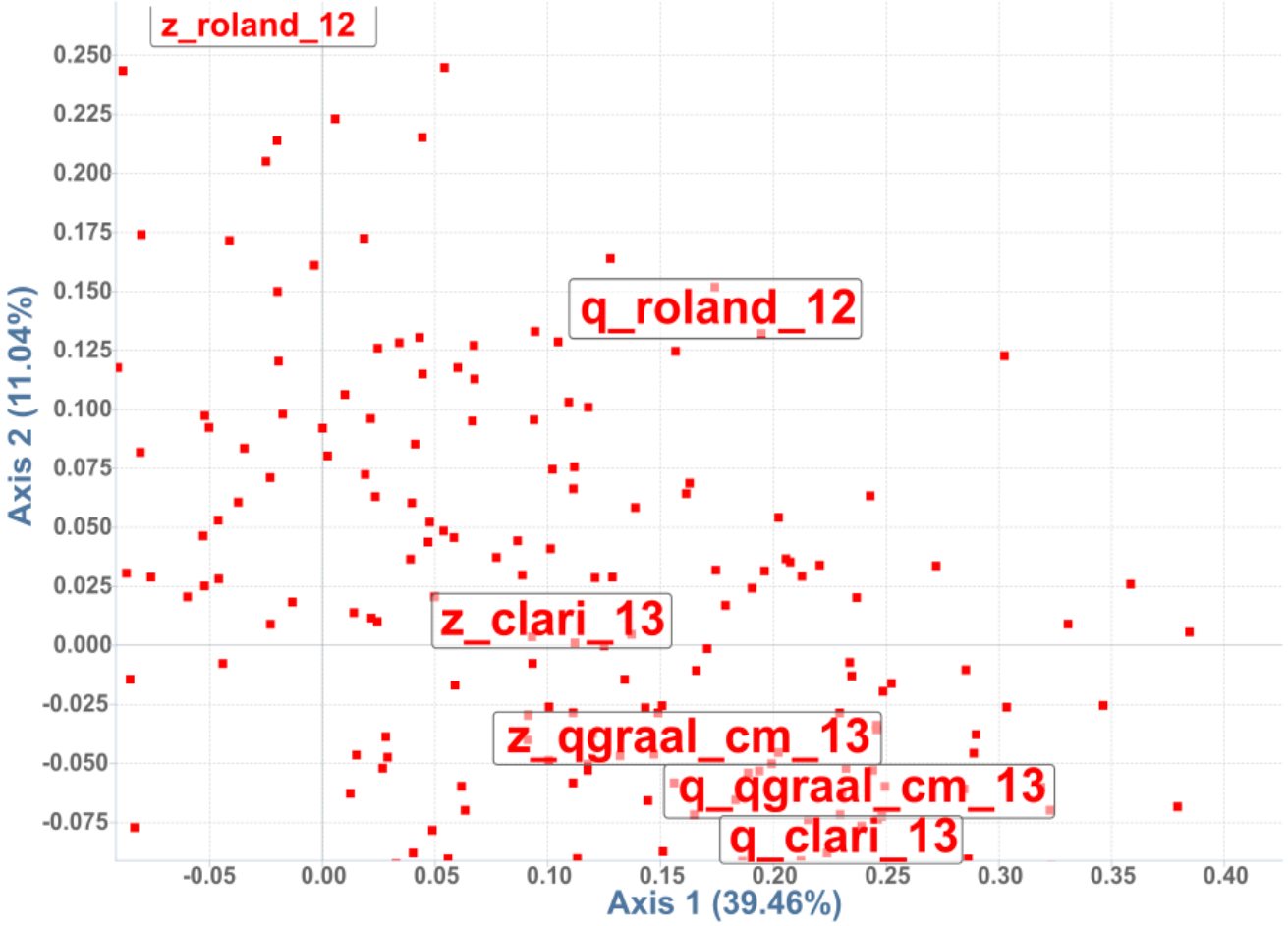
Les textes et leur annotation DD sont issus de la BFM, mais ils les caractérisent par une annotation syntaxique et non par des POS.

Nous calculons l'AFC (DDx3Textes) x 33 POS et nous observons de même que le contraste principal oppose `roland` aux autres textes.



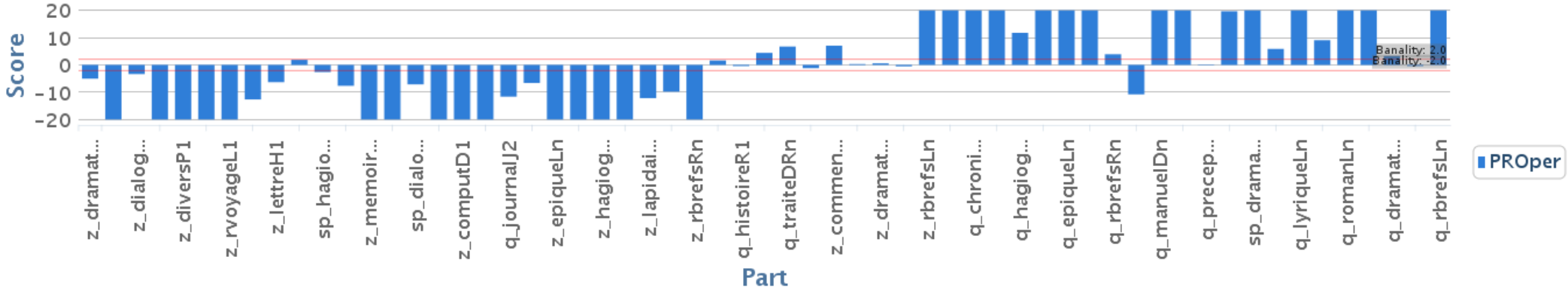
BFM & DD (3) : oralité et échelle

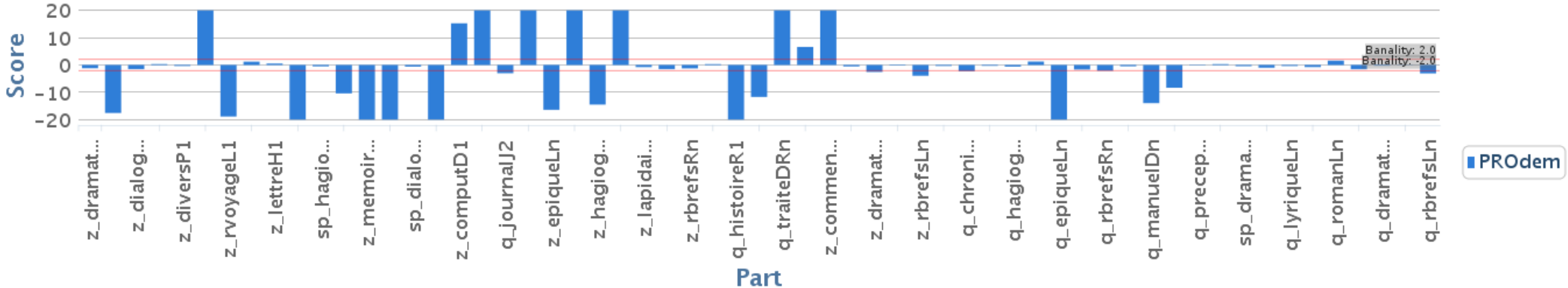
- Nous repérons les positions des 6 mêmes DU dans l'AFC calculée sur le corpus complet (DDx137Textes) x 33 POS : on a bien d'abord un gradient z / q sur l'axe 1
- L'oralité peut s'affirmer comme première dimension si le corpus est assez large et diversifié ; à l'échelle de quelques textes, les contrastes peuvent être dominés par des singularités individuelles fortes.



BFM & DD (3) : décrire un gradient

- Diagramme en bâtons de Spécificités sur les DU **triées selon leur coordonnée sur l'axe 1 :**
  - Le pronom personnel est un exemple de POS qui « accompagne » le continuum sur l'axe 1
  - Le pronom démonstratif est un exemple de POS qui concerne des genres dispersés sur l'axe 1.

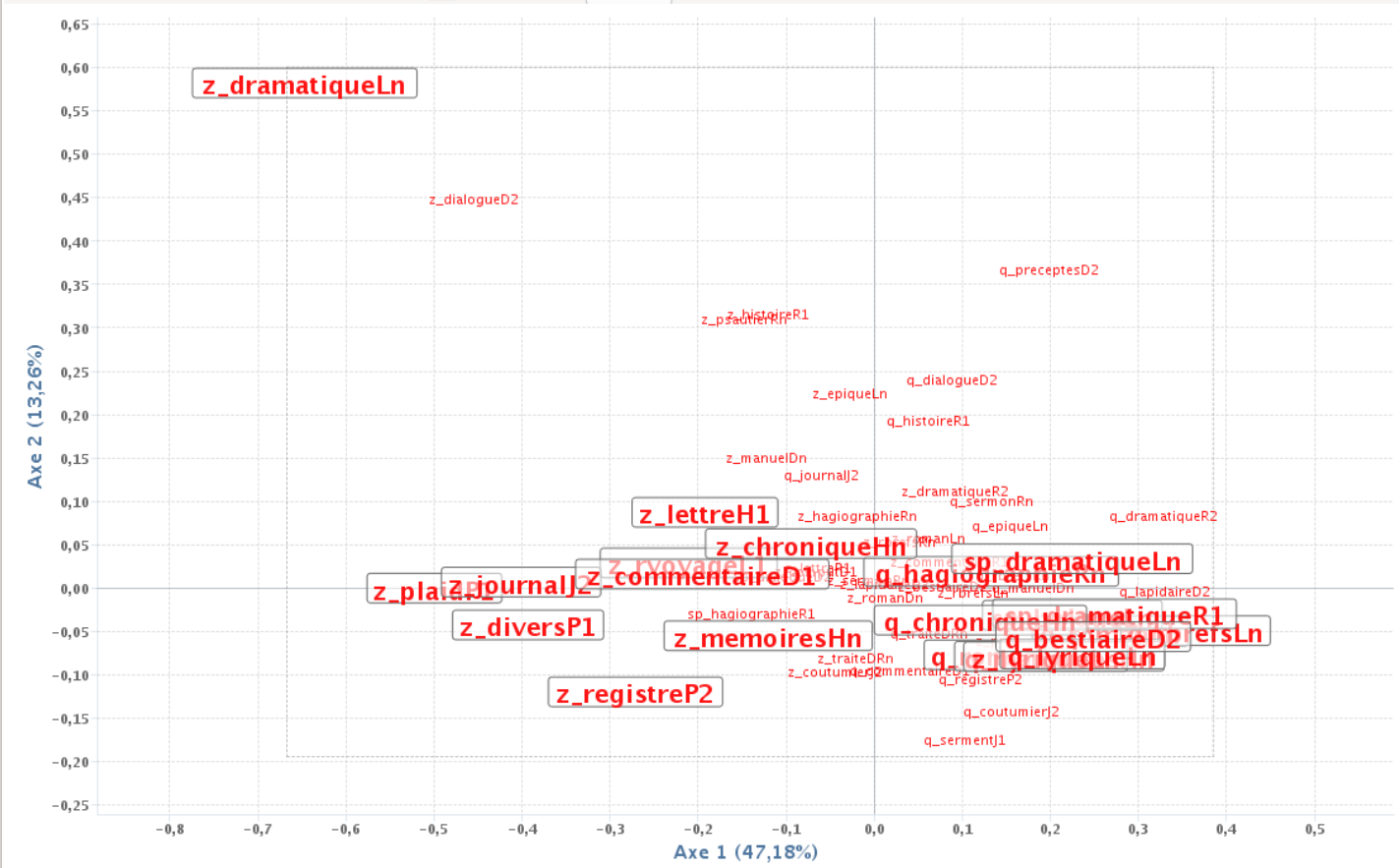




BFM & DD (3) : Sélection de DU caractérisant l'axe 1 (par cos²)

BFM016DD19/genre2017xdd_alpha/<[fropos!... | BFM016DD19/genre2017xdd_alpha/<[fropos!... | BFM016DD19/genre2017xdd_alpha/<[fropos!="PON.*|ETR|OUT|RED|ABR|ADVneg.PROadv|PROrel.PROadv|PROint.PROper...

Axes : (1,2)



Colonnes	Q12	Q13	Q23	Masse	Dist	Cont1	Cos²1
q_romanLn	0,92	0,94	0,09	10,47	0,08	26,58	0,89
z_journalJ2	0,86	0,86	0,00	3,48	0,17	19,81	0,86
sp_dramatiqueR1	0,79	0,86	0,09	0,32	0,09	0,84	0,78
q_rbrefsLn	0,77	0,85	0,11	0,45	0,16	2,13	0,76
z_plaidP1	0,72	0,77	0,05	0,28	0,27	2,48	0,72
z_diversP1	0,71	0,71	0,02	0,25	0,16	1,38	0,70
z_registreP2	0,83	0,78	0,23	8,60	0,10	22,38	0,69
q_lyriqueLn	0,74	0,74	0,14	0,28	0,07	0,55	0,67
q_romanDn	0,84	0,68	0,21	4,11	0,03	3,18	0,65
q_bestiaireD2	0,69	0,69	0,08	0,05	0,08	0,10	0,65
sp_dramatiqueLn	0,66	0,67	0,04	0,59	0,06	1,05	0,65
z_lyriqueLn	0,65	0,62	0,16	2,31	0,05	3,09	0,56
z_commentaireD1	0,54	0,56	0,03	0,21	0,05	0,28	0,54
q_hagiographieRn	0,54	0,61	0,09	0,96	0,03	0,60	0,53
q_chroniqueHn	0,58	0,60	0,13	1,73	0,02	0,89	0,52
sp_lyriqueLn	0,49	0,55	0,09	0,02	0,06	0,04	0,48
q_memoiresHn	0,50	0,52	0,15	0,06	0,07	0,09	0,44
z_memoiresHn	0,52	0,43	0,10	5,40	0,02	2,77	0,43
z_rbrefsLn	0,38	0,52	0,14	0,83	0,02	0,37	0,38
q_epiqueLn	0,46	0,50	0,21	1,00	0,04	0,85	0,38
z_lettreH1	0,43	0,51	0,23	0,06	0,07	0,08	0,35
z_rvoyageL1	0,34	0,40	0,07	0,66	0,08	1,03	0,34
q_dramatiqueR2	0,34	0,52	0,22	0,01	0,18	0,04	0,32
z_dramatiqueLn	0,54	0,35	0,29	0,01	0,89	0,14	0,30
z_bestiaireD2	0,29	0,29	0,00	0,96	0,01	0,21	0,29
q_manuelDn	0,29	0,37	0,08	0,44	0,04	0,51	0,29
z_chroniqueHn	0,43	0,49	0,34	9,23	0,01	1,68	0,29
q_rbrefsRn	0,28	0,32	0,05	0,57	0,07	0,50	0,27
z_romanLn	0,45	0,39	0,37	15,27	0,02	2,03	0,23
q_registreP2	0,38	0,47	0,41	0,09	0,05	0,05	0,22

Lignes | Colonnes | Diagramme à barres des valeurs propres | Valeurs propres

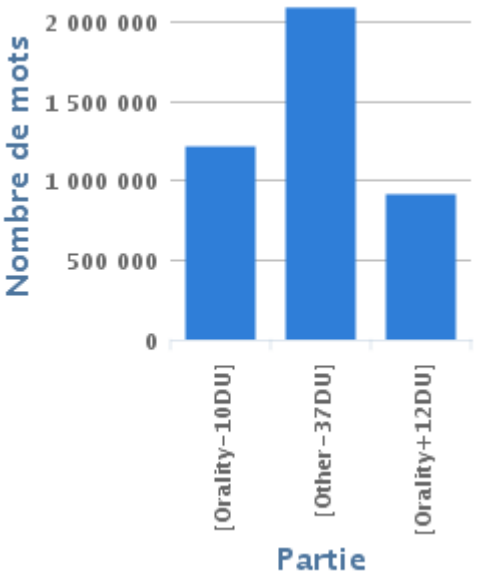
BFM & DD (3) : Composition des groupes de DU représentatives

[- Oralité]

- z_journalJ2
- z_plaidP1
- z_diversP1
- z_registreP2
- z_commentaireD1
- z_memoiresHn
- z_lettreH1
- z_rvoyageL1
- z_dramatiqueLn
- z_chroniqueHn

Cos^2 > 0,29
10 DU, mass = 28,18 %

Taille des groupes



Cos^2 > 0,4
12 DU, mass = 21,35 %

[+ Oralité]

- q_romanLn
- sp_dramatiqueR1
- q_rbrefsLn
- q_lyriqueLn
- q_romanDn
- q_bestiaireD2
- sp_dramatiqueLn
- z_lyriqueLn
- q_hagiographieRn
- q_chroniqueHn
- sp_lyriqueLn
- q_memoiresHn

BFM & DD (3) : détail des formes linguistiques en corpus

On étudie les formes de la catégorie **pronom démonstratif** qui était apparue peu liée à l'axe 1. L'influence de l'oralité apparaît en fait au niveau des formes :

- Le langage parlé recourt davantage au démonstratif *cist*, qui est plus lié au locuteur et davantage utilisé comme déictique ;
- Le démonstratif *cil* est davantage du côté [- Oralité] dans des usages anaphoriques et mémoriels (Guillot-Barbance 2017, 335-338).

[- Oralité]	**icellui** (141.4), ce (115.3), **icelle** (104.8), **iceulx** (84.8), **ceulx** (79.9), **icelles** (66.4), cen (28.3), **ycelle** (26.2), **ycellui** (23.6), CE (22.7), **chil** (22.5), **ceulz** (18.8), **yceulx** (16.3), chou (15.3), **ycelles** (15.0), **ciaus** (14.0), cecy (9.2), **ceaux** (8.5), Cen (6.8), **celuy** (6.5), **celles** (6.4), cela (5.9), **celx** (5.7), **chiax** (5.1)…
[+ Oralité]	Ce (84.8), C' (52.1), çou (36.0), c' (23.2), che (20.0), Che (18.5), Ch' (12.8), ice (7.7), **cist** (6.7), Ice (6.7), chou (6.1), **ceste** (5.9), **cestui** (5.8), **cesti** (5.4), ç' (5.3), cela (5.2)…

BFM & DD (3) : détail des formes linguistiques

Formes des ADJqua les plus caractéristiques du pôle **[– Oralité]**

fropos	word	F	f_redige	score
ADJqua	ladite	736	736	180,4
ADJqua	Saint	741	687	105,3
ADJqua	presens	434	432	101,6
ADJqua	Cedit	331	331	81,1
ADJqua	maistre	359	345	64,9
ADJqua	oudit	209	209	51,2
ADJqua	present	199	195	41,4
ADJqua	certeinnes	122	122	29,9
ADJqua	VIIJ	105	105	25,7
ADJqua	pluiseurs	102	102	25
ADJqua	feu	133	127	23,4
ADJqua	parisis	94	94	23
ADJqua	Grex	99	98	22,4
ADJqua	frans	244	210	22,1
ADJqua	grand	161	146	20,5
ADJqua	Petit	142	131	20,2
ADJqua	yceulx	82	82	20,1
ADJqua	dessusdit	81	81	19,8
ADJqua	sains	259	216	19,1
ADJqua	Sainte	121	112	17,6

Formes des ADJqua les plus caractéristiques du pôle **[+ Oralité]**

fropos	word	F	f_oralise	score
ADJqua	Biax	188	188	68,7
ADJqua	bele	428	360	68,1
ADJqua	biax	197	194	65,5
ADJqua	voir	235	211	50,4
ADJqua	douce	138	137	48,2
ADJqua	Biaus	116	116	42,4
ADJqua	grant	4404	2349	41,9
ADJqua	biau	125	121	38,2
ADJqua	Bele	104	103	35,9
ADJqua	boen	87	87	31,8
ADJqua	biaus	135	121	29
ADJqua	Biau	78	78	28,5
ADJqua	mal	182	152	28,5
ADJqua	douz	75	75	27,4
ADJqua	Lasse	73	73	26,7
ADJqua	fole	86	83	26
ADJqua	fol	99	92	25,1
ADJqua	las	111	100	24,5
ADJqua	chiere	72	71	24,3
ADJqua	chier	171	138	23,1

BFM & DD (3) : détail des formes linguistiques en corpus

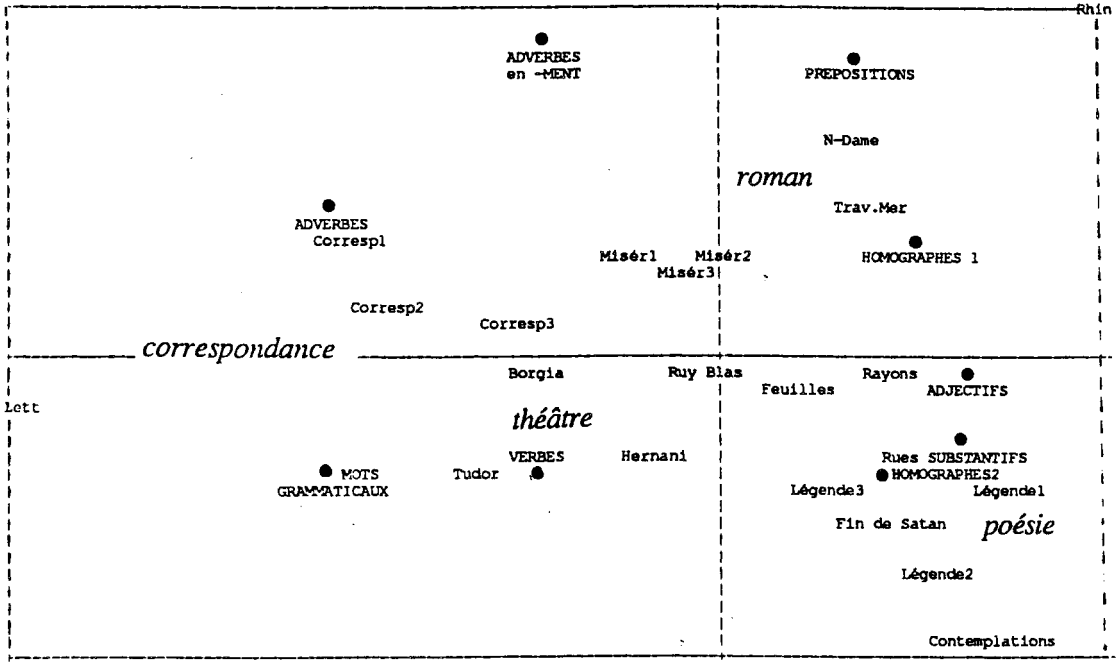
Étude de catégories [+ Oralité] :

Catégories	***Pôle [- oralité]***	***Pôle [+ oralité]***
ADJqua	+ diversité *present, saint, maistre* (2 textes)	TA : *bel, douz, chier* adj.évaluatifs : *grant, boen, mal, las, fol*
ADVgen	- diversité adv. spatio-temporels : *illec/lors* *ainsi/aussi* (2 textes)	adv. énonciatifs : *certes* adv. temporels/énonciatifs : *or/ja* adv. spatio-temporels : *mais/hui/onques/ci* intensif./quantificateurs : *mout/tant/trop*
PROind		*on*
PROadv		*en, i* → *il i a* (vs *il a*) ?
CONsub		*Que, Se, Quant* (ODM)

BFM & DD (3) : éléments méthodologiques

- Étude de contraste intertextuel et intratextuel
  - Les partitions peuvent aussi intervenir à l’intérieur des textes
  - Utilitaire TXM commands/AdvancedPartition
- Gestion de la fiabilité de l’étiquetage en AFC
  - Poids ; contribution aux dimensions concernées.
  - Et aussi *peeling*
- Variation de granularité des unités pour le calcul des Spécificités
  - Graphies, sous-groupes, sur-catégorie…
  - Utilitaire TXM r/PlotSpecif
- Stabilité des AFC sur les POS : un équilibre morphosyntaxique ?

AFC et opposition Verbe / Nom chez Brunet



Graphique 7b. Analyse factorielle des catégories grammaticales

In Étienne Brunet (1991a) « Les catégories grammaticales chez Hugo », republié dans ses *Écrits choisis*, tome III, *Tous comptes faits. Questions linguistiques*, Champion, 2016, chapitre numérique en ligne sur Texto! et HAL.

Opposition entre le « clan » du verbe et celui du nom, qu'Étienne Brunet remarque dans des corpus diversifiés :

- Giraudoux (1974a){I.11} p. 266
- Balzac (2003c){I.3} p.73-76
- Hugo (1988b){I.5} p.115 et (1991a) {III.11} p. 3K.11-13
- Flaubert (2003e){I.6} p. 138 sq. Et 146-147,
- Zola (1985e){I.7} p. 165
- Rimbaud (2006b){I.8} p. 186
- différentes traductions de l'Évangile (2002b){II.11} p. 218 et (2002a){III.20} p. 3T.16-17
- corpus étendu de littérature française (2002c){III.12} p. 3L.13-16, (2003d){III.6} p. 3F.15-16
- textes de genres et d'origines divers représentant la francophonie (2003d) {III.6} p. 3F.22 sq., etc.

Le projet ANR Antract

- Étude des *Actualités françaises* (1945-1968)
  - Actualités filmées diffusées dans les cinémas
  - Hebdomadaires, durée d’environ 10 mn
  - 1259 éditions, traitant en moyenne 8 sujets.

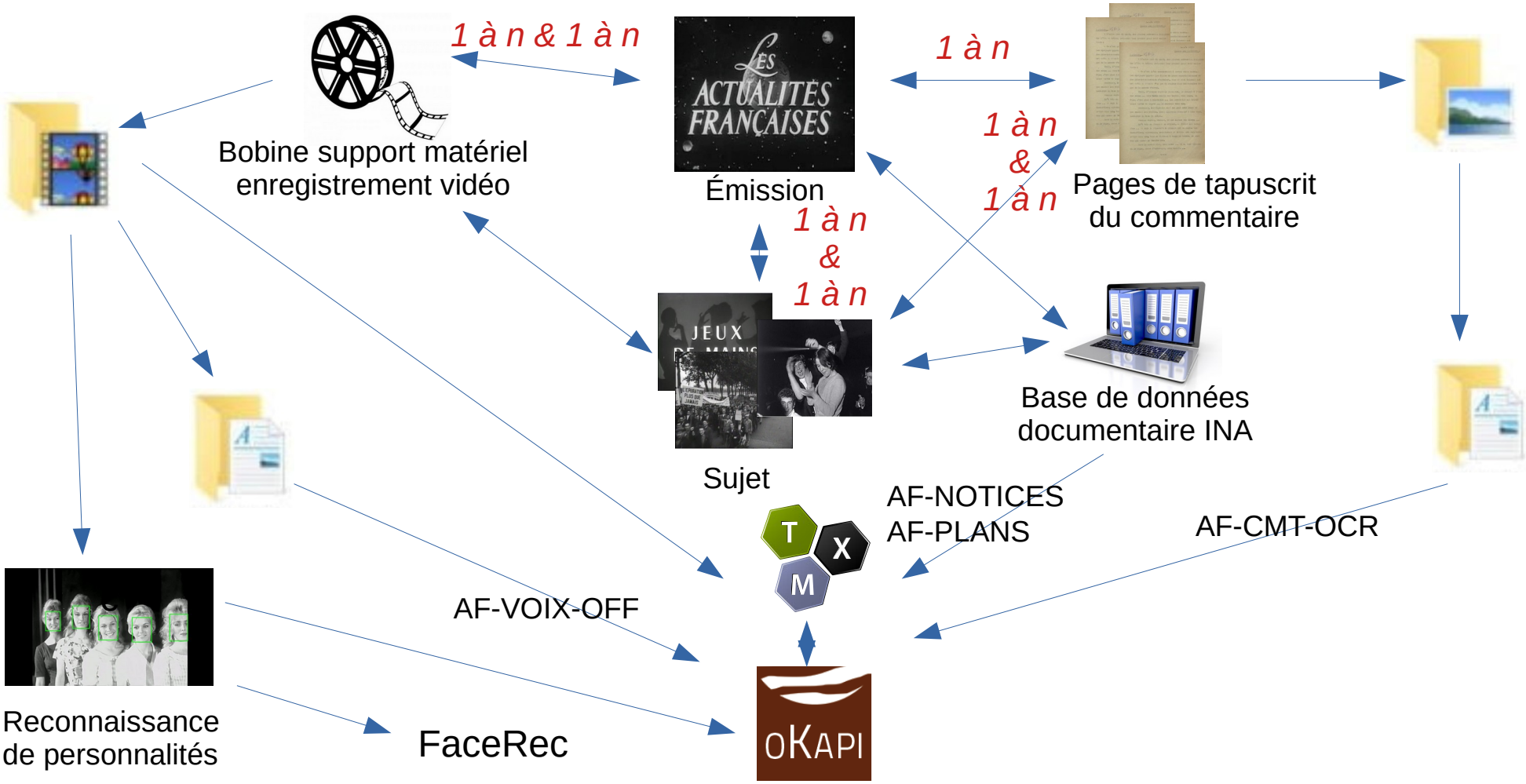
Le projet ANR Antract : partenaires

- Centre d’histoire sociale des mondes contemporains, Univ. Paris, UMR 8058
- Institut National de l’Audiovisuel, Bry-sur-Marne
- Laboratoire d’Informatique de l’Université du Maine, Univ. du Mans, EA 4023
- Département Data Science de l’École d’ingénieurs Eurecom, Sophia Antipolis
- Institut d’Histoire des Représentations et des Idées dans les Modernités, Univ. Lyon, UMR 5317

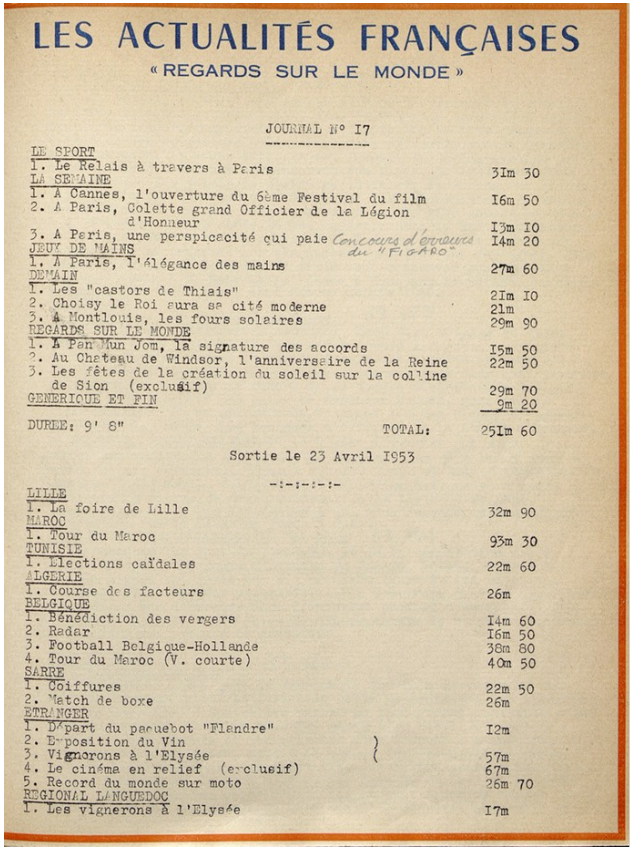
Textométrie sur corpus audiovisuel

- Retour au texte → retour au document source → vidéo
  - La transcription automatique doit pouvoir être contrôlée
  - La transcription manuelle elle-même impliquerait des choix de représentation (des interprétations, des réductions)
- Information de synchronisation
  - Manuellement : idéalement utiliser un logiciel de saisie de transcription comme Transcriber
  - Automatiquement : les time-codes peuvent être intégrés automatiquement dans un format XML.

Sources et unités multiples



Tapuscrit



LES ACTUALITÉS FRANÇAISES

« REGARDS SUR LE MONDE »

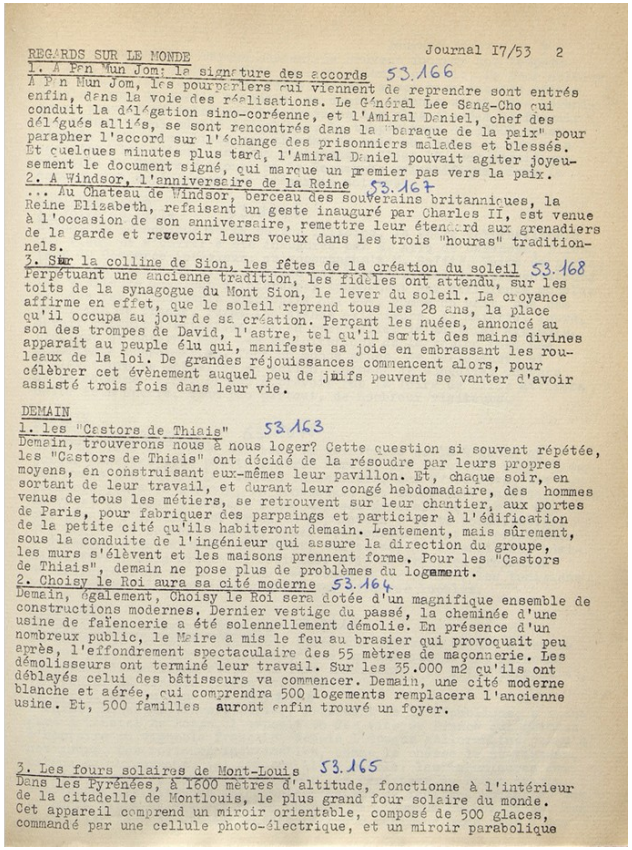
JOURNAL N° I7

LE SPORT	
1. Le Relais à travers à Paris	3Im 30
LA SEMAINE	
1. A Cannes, l'ouverture du 6ème Festival du film	I6m 50
2. A Paris, Colette grand Officier de la Légion d'Honneur	I3m I0
3. A Paris, une perspicacité qui paie Concours d'erreurs du "FIGARO"	I4m 20
JEUX DE MAINS	
1. A Paris, l'élégance des mains	27m 60
DEMAIN	
1. Les "castors de Thiais"	2Im I0
2. Choisy le Roi aura sa cité moderne	2Im
3. A Montlouis, les fours solaires	29m 90
REGARDS SUR LE MONDE	
1. A Pan Mun Jom, la signature des accords	I5m 50
2. Au Chateau de Windsor, l'anniversaire de la Reine	22m 50
3. Les fêtes de la création du soleil sur la colline de Sion (exclusif)	29m 70
GENERIQUE ET FIN	9m 20
DUREE: 9' 8" TOTAL:	25Im 60

Sortie le 23 Avril I953

-:-:-:-:-

LILLE	
1. La foire de Lille	32m 90
MAROC	
1. Tour du Maroc	93m 30
TUNISIE	
1. Elections caïdales	22m 60
ALGERIE	
1. Course des facteurs	26m
BELGIQUE	
1. Bénédiction des vergers	I4m 60
2. Radar	I6m 50
3. Football Belgique-Hollande	38m 80
4. Tour du Maroc (V. courte)	40m 50
SARRE	
1. Coiffures	22m 50
2. Match de boxe	26m
ETRANGER	
1. Départ du paquebot "Flandre"	I2m
2. Exposition du Vin }	
3. Vignerons à l'Elysée }	57m
4. Le cinéma en relief (exclusif)	67m
5. Record du monde sur moto	26m 70
REGIONAL LANGUEDOC	
1. Les vignerons à l'Elysée	I7m



Journal I7/53 2

REGARDS SUR LE MONDE

1. A Pan Mun Jom: la signature des accords 53.166

A Pan Mun Jom, les pourparlers qui viennent de reprendre sont entrés enfin, dans la voie des réalisations. Le Général Lee Sang-Cho qui conduit la délégation sino-coréenne, et l'Amiral Daniel, chef des délégués alliés, se sont rencontrés dans la "baraque de la paix" pour parapher l'accord sur l'échange des prisonniers malades et blessés. Et quelques minutes plus tard, l'Amiral Daniel pouvait agiter joyeusement le document signé, qui marque un premier pas vers la paix.

2. A Windsor, l'anniversaire de la Reine 53.167

... Au Chateau de Windsor, berceau des souveraine britanniques, la Reine Elizabeth, refaisant un geste inauguré par Charles II, est venue à l'occasion de son anniversaire, remettre leur étendard aux grenadiers de la garde et recevoir leurs voeux dans les trois "houras" traditionnels.

3. Sur la colline de Sion, les fêtes de la création du soleil 53.168

Perpétuant une ancienne tradition, les fidèles ont attendu, sur les toits de la synagogue du Mont Sion, le lever du soleil. La croyance affirme en effet, que le soleil reprend tous les 28 ans, la place qu'il occupa au jour de sa création. Perçant les nuées, annoncé au son des trompes de David, l'astre, tel qu'il sortit des mains divines apparait au peuple élu qui, manifeste sa joie en embrassant les rouleaux de la loi. De grandes réjouissances commencent alors, pour célébrer cet évènement auquel peu de juifs peuvent se vanter d'avoir assisté trois fois dans leur vie.

DEMAIN

1. les "Castors de Thiais" 53.163

Demain, trouverons nous à nous loger? Cette question si souvent répétée, les "Castors de Thiais" ont décidé de la résoudre par leurs propres moyens, en construisant eux-mêmes leur pavillon. Et, chaque soir, en sortant de leur travail, et durant leur congé hebdomadaire, des hommes venus de tous les métiers, se retrouvent sur leur chantier, aux portes de Paris, pour fabriquer des parpaings et participer à l'édification de la petite cité qu'ils habiteront demain. Lentement, mais sûrement, sous la conduite de l'ingénieur qui assure la direction du groupe, les murs s'élèvent et les maisons prennent forme. Pour les "Castors de Thiais", demain ne pose plus de problèmes du logement.

2. Choisy le Roi aura sa cité moderne 53.164

Demain, également, Choisy le Roi sera dotée d'un magnifique ensemble de constructions modernes. Dernier vestige du passé, la cheminée d'une usine de faïencerie a été solennellement démolie. En présence d'un nombreux public, le Maire a mis le feu au brasier qui provoquait peu après, l'effondrement spectaculaire des 55 mètres de maçonnerie. Les démolisseurs ont terminé leur travail. Sur les 35.000 m2 qu'ils ont déblayés celui des bâtisseurs va commencer. Demain, une cité moderne blanche et aérée, qui comprendra 500 logements remplacera l'ancienne usine. Et, 500 familles auront enfin trouvé un foyer.

3. Les fours solaires de Mont-Louis 53.165

Dans les Pyrénées, à 1600 mètres d'altitude, fonctionne à l'intérieur de la citadelle de Montlouis, le plus grand four solaire du monde. Cet appareil comprend un miroir orientable, composé de 500 glaces, commandé par une cellule photo-électrique, et un miroir parabolique

FaceRec

FaceRec Visualizer | Insert URI of a video | GO | ANTRACT

LES ACTUALITES FRANCAISES :EDITION DU 22 JANVIER 1953

Date

1953-01-22T00:00:00

Segments

- La dernière séance du cabinet Truman
  00:00:13 - 00:00:49
- PREMIERE REUNION DE L'EQUIPE MINISTERIELLE D'EISENHOWER
  00:00:13 - 00:00:49
- Paris-Bogota par Air France
  00:00:49 - 00:01:21
- VOYAGE DE MONSIEUR LEONARD AU HOGGAR
  00:01:23 - 00:02:22
- TEMOIN DE LA PREHISTOIRE : LE COELACANTHE
  00:02:21 - 00:02:50
- RENTREE A LA SCENE D'HENRI GARAT
  00:02:51 - 00:04:03
- REMISE DE LA BARETTE A MONSEIGNEUR RONCALLI
  00:04:03 - 00:06:19
- Les 24 nouveaux cardinaux
  00:04:04 - 00:06:20
- Le Consistoire secret
  00:04:04 - 00:06:20
- Eisenhower Président
  00:06:20 - 00:08:16

Min confidence: 0.6

06:55 - 06:56 Dwight Eisenhower Confidence: 0.73

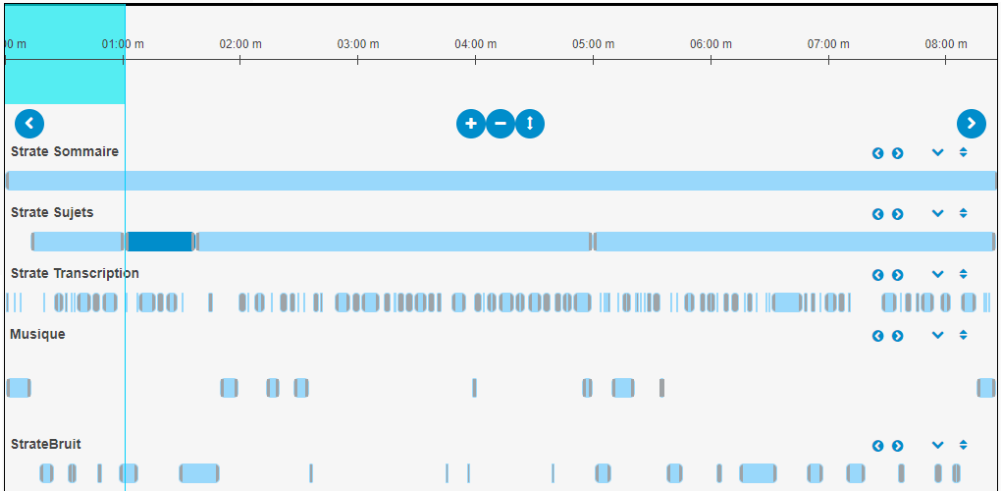
07:00 - 07:02 Dwight Eisenhower Confidence: 0.63

In this video:

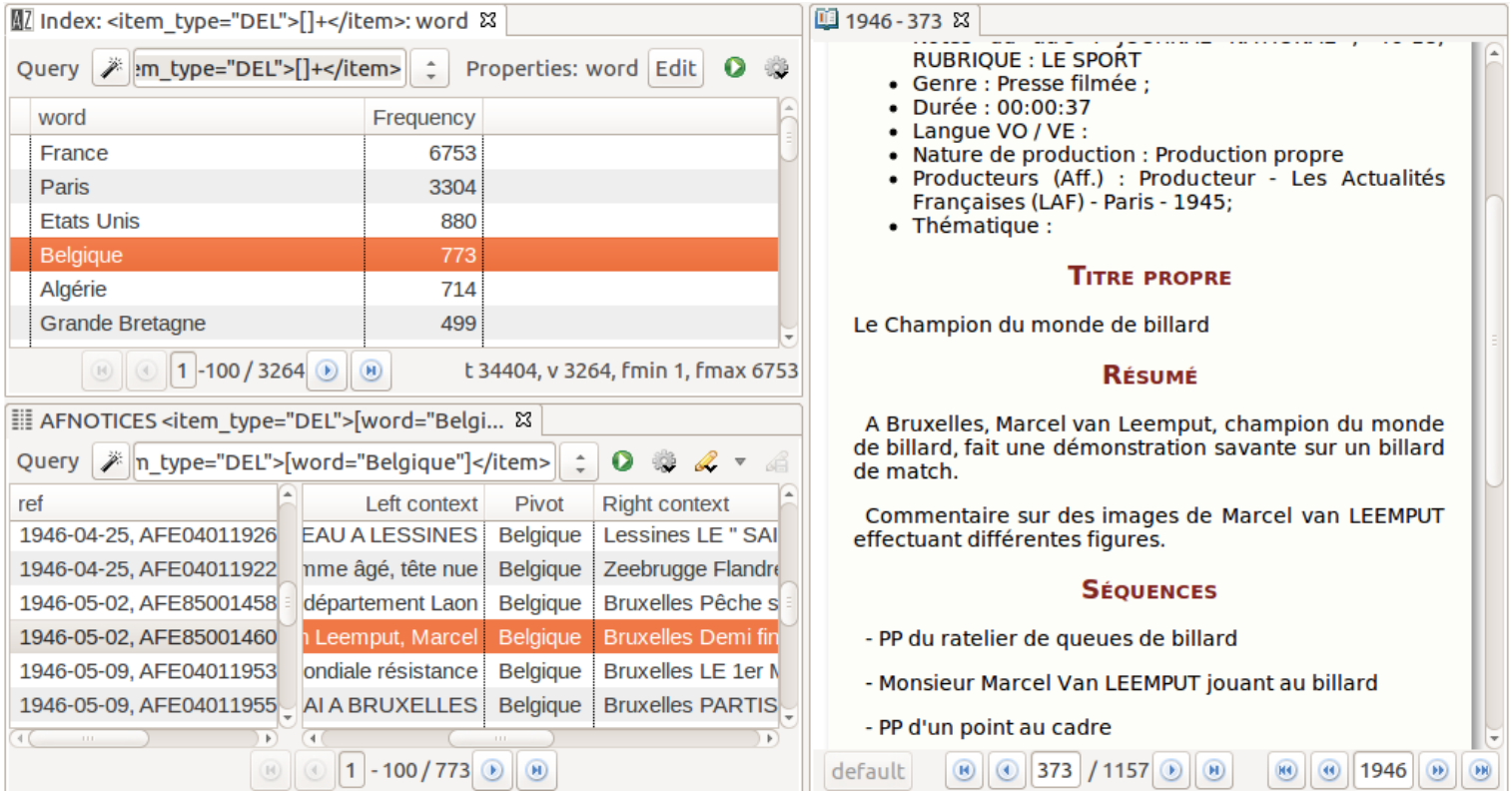
Dwight Eisenhower

Okapi

- Plateforme type web sémantique
- Annotation, ex. positionnement temporel de sujets

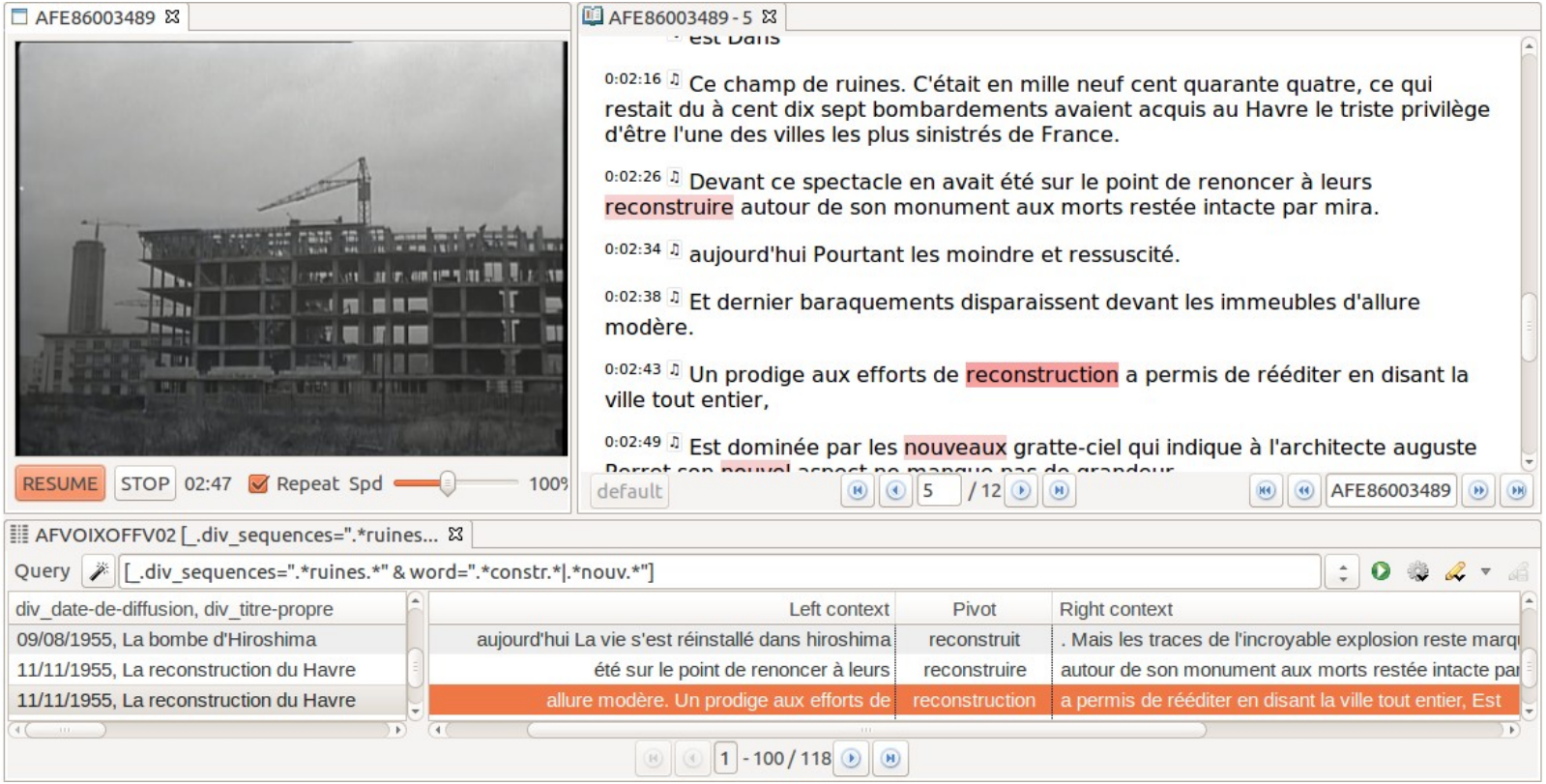


Corpus TXM AF-NOTICES



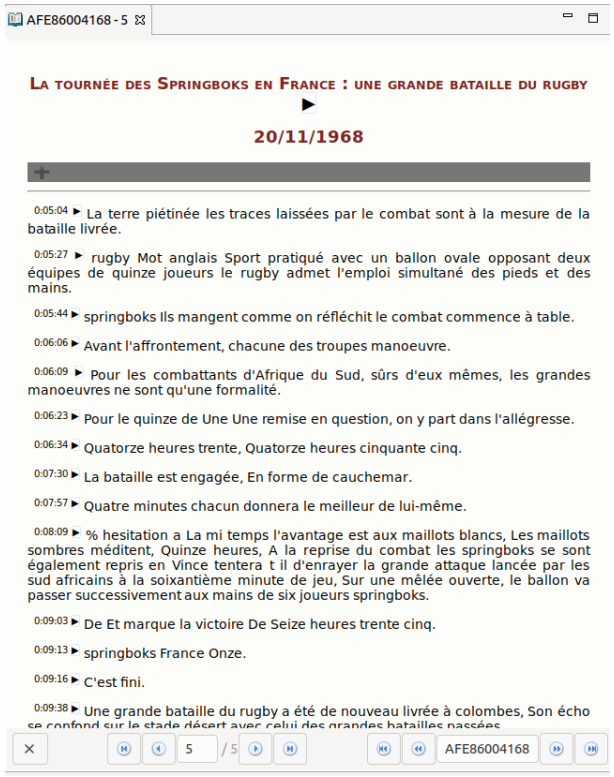
- Notices documentaires INA
- 1261 textes, 10776 sujets
- 2,2 millions de mots

Corpus TXM AF-VOIX-OFF

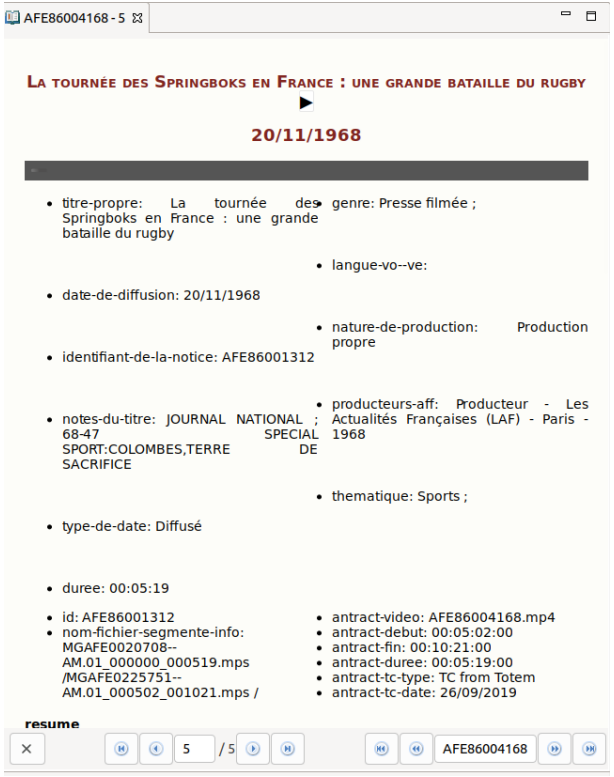


- Transcription automatique de la bande son
- 1260 textes, 10683 sujets synchronisés
- 1,5 millions de mots
- Retour à la vidéo depuis la concordance ou l’édition du texte

Un affichage structuré de la transcription



Transcription



Métadonnées dépliées

Sont visualisés sans faire partie des mots du corpus :

- les métadonnées,
- la mention du locuteur en début de tour de parole (option sans mention pour ce corpus monologue),
- les indications de temps,
- les titres de section,
- les commentaires type didascalies et annotations sur les mots (tels qu'encodés dans le format Transcriber), etc.

Possibilité d’accès distant aux vidéos

AFVOIXOFFV02 [frlemma = "foule"]

Query [frlemma = "foule"]

text_id, text_datedediffusion	Left context	Pivot	Right context
AFE86003736, 10/08/1960	siastiques du monde entier et de la	foule	des congressiste qui sera bientôt e
			des envoyés de la presse internati
			des étudiants en science se soit re
			des fidèles, à Immémorial. C'est là
			des fidèles à L'origine de ces poss
			des fidèles avec les présidents ser
			des fidèles se livre à des danses e
			des grands jours, monsieur Poujad
			des grands jours pour l'événement
			des journalistes le chef de l'état de
AFE86003728, 15/06/1960	voiture complètement bloqué par la	foule	des manifestants, il n'avait dû qu'u
AFE86003431, 07/10/1954	propriétaire Madame portrait de la	foule	des parieurs du gagnant anonyme
AFE86004792, 17/04/1952	donnant sa bénédiction Pascal à la	foule	des pèlerins. Hum. Pas de paix ég
AFE86004036, 11/05/1966	cardinal Wyszynski, on a évalué la	foule	des pèlerins à cent mille personne
AFE86003459, 21/04/1955	nné sa bénédiction urbi et orbi à la	foule	des pèlerins réunis place Saint-pie
AFE86003738, 24/08/1960	ard. aujourd'hui L'été il ramène des	foules	des peuples des collègues. touriste
AFE86004784, 21/02/1952	le prince de liège au milieu de la	foule	des représentants de tous les pays
AFE86003514, 09/05/1956	clubs sportifs de Moscou. Alors la	Foule	des travailleurs a clos les défilés q

You must authenticate to access the media file

Login to okapi.ina.fr

Login <my login>

Password ••••••••••

Cancel Reset All Connect

201 - 300 / 548

AFE86003459 - 2

DET: fête religieuse ; DET: Pâques ; DET: religion ; DET: catholicisme ; DEI: Pie XII ; DEI: bénédiction ; DEL: Italie ; DEL: Rome ; DEL: Vatican ;

generique_aff_lig

S 0: 0:00:03 euh

S 64: 0:00:06 à Rome à l'occasion des têtes de pas faire le pape qui n'était pas apparu en public depuis sa maladie a donné sa bénédiction urbi et orbi à la foule des pèlerins réunis place Saint-pierre.

0:00:18 Dans le message Pascal qu'il adresse traditionnellement à la chrétienté pie douze a supplié laisse avant d'orienter leurs recherches nucléaires vers début humanitaire.

default 2 / 1 AFE86003459

Croisement multi-sémiotique

word

Property word

Units	Frequency T 1568649	foule_in_documentary_desc t=396023	index
foule	515	353	93.6
président	1865	830	72.0
Gaulle	708	375	55.1
général	1750	731	50.8
république	782	344	29.4
la	44672	12268	26.9
accueil	194	119	25.5
cortège	150	99	25.0
enthousiasme	151	96	22.4
avait	2528	860	22.3
devant	1331	489	19.9
était	3789	1208	19.6
peuple	469	208	18.6
acclamations	67	52	18.4

word

Property word

Units	Frequency T 1568649	foule_in_documentary_desc ∩ voice_without_gaulle_president t=264308	index
foule	515	211	37.5
peloton	215	100	23.2
départ	719	223	20.1
minutes	612	182	14.6
étape	323	113	14.6
princesse	206	82	14.3
course	381	125	13.6
coureurs	129	58	13.0
roi	448	138	12.6
devant	1331	328	12.5
personnes	333	109	11.9
reine	354	114	11.9
carnaval	48	30	11.7
corrida	43	28	11.6

Calcul de ***résonance*** (au sens de Salem, JADT 2004) : mots spécifiques du <u>commentaire audio</u> pour les sujets dont la <u>notice documentaire</u> mentionne

- une « foule » (en haut),
- puis idem sans mention de « président » ou « [de] Gaulle », pour voir ce qui apparaît en dehors du cas le plus fréquent (en bas)

Utiliser la textométrie multimédia ?

- Transcriber : le logiciel… ou le format
  - Cas du retour à la vidéo AF-NOTICES
  - On n’a pas abordé le cas de transcriptions manuelles très riches et de la traduction entre un format de représentation initiale et une modélisation textométrique → apport des réflexions et expérimentations en cours à Praxiling (ex. JADT 2022)
- Extension MediaPlayer
  - Réglages de préférences de corpus
  - Documentation : dans le manuel utilisateur 0.8 en ligne
- Apports :
  - Complémentarité/contrôle pour la transcription
  - Contextualisation
  - Recherche multisémiotique
- Publication
  - Bénédicte PINCEMIN, Serge HEIDEN, Matthieu DECORDE (2020) - « Textometry on Audiovisual Corpora. Experiments with TXM software », *in* P. Marchand & P. Ratinaud (eds), *JADT’ 20. Proceedings of the 15th International Conference on Statistical Analysis of Textual Data*, Université de Toulouse, 16-19 juin 2020. https://halshs.archives-ouvertes.fr/halshs-02779055v1

Un thème de recherche : la grammaire cinématographique

- Façon dont les vidéos sont composées
  - types de plans
  - mouvements de caméra
- Analyse automatique de vidéo ?
- Dans TXM → textométrie, texte → passer par la description documentaire ?
  - Champ Séquences : description plan à plan

Grammaire cinématographique : pistes pour travailler avec TXM

- Difficultés
  - Variabilité des désignations
    - Description libre
    - Pratiques ± partagées
  - Mots qui s’intercalent
    - Distance variable
    - Pris dans la requête
- Solutions
  - Annotation par des catégories d’analyse (recodage)
  - Projection
    - Un nouveau corpus avec uniquement les catégories d’analyse

Vue du sujet « Rugby » dans TXM

AF-NOTICES-V3-2021-09-30/<[_.div_identi... | AFE86004168 - 5

TITRE PROPRE

La tournée des Springboks en France : une grande bataille du rugby

RÉSUMÉ

Résumé du second test-match opposant le XV de France à l'Afrique du Sud au stade Yves Manoir de Colombes. Victoire finale des Springboks (11-16).

SÉQUENCES

- VG en plongée une partie de la pelouse du stade de Colombes
- VG travées vides avec vieux journaux jonchant le sol (2 plans)
- GP d'un lustre éclairé, dans le couloir des vestiaires- TRAVEL dans les couloirs des vestiaires- TRAVEL le long de l'escalier menant des vestiaires au stade, arrivée sur le stade et PANO sur celui-ci

TITRE : " SPECIAL "

- GP publicité pour un ballon de rugby
- GP publicité pour des chaussures de rugby " La Chaussure de l'élite "
- BT catalogue de divers accessoires de rugbymen : bas, culottes, maillots

TITRE : " SPECIAL SPORT "

- GP, VG les SPRINGBOKS, prenant leur repas, le visage soucieux

TITRE : " COLOMBES TERRE DE SACRIFICE "

- GP de deux pieds, chaussés de chaussures de rugby, boueuses
- GP du visage soucieux des joueurs- plusieurs plans des joueurs sud africains puis français à l'entraînement ou effectuant des exercices d'assouplissement

TITRE : " SE

frpos=NOM, frlemma=plan, n=1014, ref=1968-11-20, AFE86001312, plan=00DP

- VG 2 plans
- PM, VG de la musique de la Garde républicaine défilant sur le terrain

5 / 5 AFE86004168

Annotation automatique par requêtes CQL

- Les requêtes pour la catégorie **Plan général** (10PG) :

```
[word="-*(VG|PG|GVG|CG)"%c]
[word="-*(plan|vues?)"%c][word="générale?s?"%c]
[word="-*PANORAMA"%c]
[word="-*(PE|VE)"%c]
[word="-*(plan|vue)s?"%c][][word="ensemble"%c]
```

Annotation automatique par requêtes CQL

cql_valeurs_de_plan_211007a.xlsx - LibreOffice Calc

	code	type	expressions	equation	commentaire
2	00DP	divers plans	DP, DV	[word="-*(DP|DV)"%c]	"divers" (?) ou "divers vues" (?), un seul résultat sur un sujet avec 3 plans successifs sur un même
3	00DP	divers plans	plan/vue(s)	[(resume|sequences) & word="-*(plan|vue)s?"%c]	17 894 occ. : attrape-tout des désignations floues (ou non).
4	01HS	hors sujet	en arrière plan, au premier plan, sur le deuxième plan, en avant plan, en A V plan, (Au) 2ème plan, Au/en 1er plan, sur le 1er plan, au/sur le dernier	([word="-*(premier|1er|avant|second|deuxième|2ème|arrière|dernier)"%cd] | ([word="A"][word="V"])) @[word="plan"%c])	1405 occ.
5	01HS	hors sujet	commissaire (général(e)) au plan, Secrétaire d'Etat ... au plan, et autres expressions	([word="au"%c] @[word="plan"%c])	11 occ.
6	01HS	hors sujet	sur le plan international, sur le plan	([word!="surimpressionnée?s?"%c][word="sur"%c][word="le"]@[word="plan"%c])	10 occ.
7	01HS	hors sujet	devant un/les plan(s)	([word="devant"%c][word="un|les|des"]@[word="plans?"%c])	6 occ.
8	01HS	hors sujet	plan Marshall,...	([(resume|sequences) & word="plan"%c][word="."]?[word="Monnet|Marshall?|Schuman|Courant|Pinay-Rueff|Challe"%c])	69 occ.
9	01HS	hors sujet	plan britanique, plan cadastral, plan mural, plan(s) incliné(s)...	([(resume|sequences) & word="plans?"%c][word="britanique|algérien|cadastra(l|ux)|architectura(l|ux)|mura(l|ux)|inclinés?|anciens?"%c])	16 occ. ; 2 occ. de "plan algérien" mais hors résumé/séquences
10	01HS	hors sujet	vues microscopiques, plans cinématographiques	([(resume|sequences) & word="-*(plan|vue)s?"%c][word="(microscopique|photographique|	6 occ. dont 1 qu'il n'aurait pas fallu sélectionner.
11	01HS	hors sujet	miroir plan	([word="miroirs?"%c][word="-"]? @[word="plans?"%c])	4 occ.

Feuil1

Rechercher | Tout rechercher | Affichage mis en forme | Respecter la casse

Feuille 1 sur 1 | PageStyle_Feuil1 | Français (France) | Moyenne: ; Somme: 0 | 85 %

Annotation automatique par requêtes CQL

- Un exemple pris dans la catégorie Hors-sujet (01HS), illustrant davantage de possibilités du langage :

```
[frlemma="consulter|déplier|discuter|étudier|examiner|montrer|
regarder|tracer"%c]([]{0,3} [frpos=".*DET.*"%c])?
@[word="plans?"%c]
```

  - « Le général MAST au milieu d'architectes discutant plans en mains »
  - « GP du général MAST examinant un plan sur un tréteau »
  - « Salle avec techniciens du repérage, écouteurs aux oreilles, assis autour d'une table, consultant des cartes et des plans »
  - « GP du vieux monsieur qui trace un plan »
  - « PM du sculpteur traçant des traits sur le plan d'une maison »

Annotation automatique par requêtes CQL

- Un tableau avec
  - 19 valeurs de plan + Divers plans + Hors-sujet
  - 60 requêtes
    - Une même valeur de plan peut avoir plusieurs requêtes
- Utilitaire CQLList2WordProperties
  - Ajoute une propriété de mot
  - Les requêtes sont appliquées dans l'ordre → possibilité d'affinement de l'étiquetage par le contexte

Corpus annoté

AF-NOTICES-V3-2... | AF-NOTICES-V3-2021-10-11/<[_.div_id="AF... | AFE86004168 - 5

Query Propert

plan	Frequency
10PG	23721
14GP	16234
00DP	15494
32PP	10001
12PM	9742
31PANO	7199
13PR	2888
70TI	1947
01HS	1655
40VA	1394
20PLON	1094
30TRAV	880
60ZOOM	467
11PL	260
50VE	215
21CPL	190
51VI	62
72GR	41
15TGP	34
73FLOU	25
71BT	10

t 93553, v 21, fmin 10, fma

Titre propre

La tournée des Springboks en France : une grande bataille du rugby

Résumé

Résumé du second test-match opposant le XV de France à l'Afrique du Sud au stade Yves Manoir de Colombes. Victoire finale des Springboks (11-16).

Séquences

- VG en plongée une partie de la pelouse du stade de Colombes

- VG travées vides avec vieux journaux jonchant le sol (2 plans)

- GP d'un lustre éclairé, dans le couloir des vestiaires- TRAVEL dans les couloirs des vestiaires- TRAVEL le long de l'escalier menant des vestiaires au stade, arrivée sur le stade et PANO sur celui-ci

TITRE : " SPECIAL "

- GP publicité pour un ballon de rugby

- GP publicité pour des chaussures de rugby " La Chaussure de l'élite "

- BT catalogue de divers accessoires de rugbymen : bas, culottes, maillots

TITRE : " SPECIAL SPORT "

- GP, VG les SPRINGBOKS, prenant leur repas, le visage soucieux

TITRE : " COLOMBES TERRE DE SACRIFICE "

- GP de deux pieds, chaussés de chaussures de rugby, boueuses

- GP du visage soucieux des joueurs- plusieurs plans des joueurs sud africains puis français à l'entraînement ou effectuant des exercices d'assouplissement

TITRE : " SEIZE NOVEMBRE, 14H30 "

- VG 2 plans du public arrivant au stade

Page 5 / 5 Text AFE86004168

AF-NOTICES-V3-... | AF-PLANS-V2... | AFE86004168 ...

word	plan
VG	10PG
en	__UNDEF__
plongée	20PLON
une	__UNDEF__
partie	__UNDEF__
de	__UNDEF__
la	__UNDEF__
pelouse	__UNDEF__
du	__UNDEF__
stade	__UNDEF__
de	__UNDEF__
Colombes	__UNDEF__
-	__UNDEF__
VG	10PG
travées	__UNDEF__
vides	__UNDEF__
avec	__UNDEF__
vieux	__UNDEF__
journaux	__UNDEF__
jonchant	__UNDEF__
le	__UNDEF__
sol	__UNDEF__
(	__UNDEF__
2	__UNDEF

1247 / 1

Séq

10PG 20PLON
10PG 00DP
14GP 30TRAV 30TRAV 31PANO
70TI
14GP
14GP
71BT
70TI
14GP 10PG
70TI
14GP
14GP 00DP
70TI
10PG 00DP
12PM 10PG
12PM 10PG
70TI
12PM 14GP 00DP
14GP 00DP 31PANO 14GP

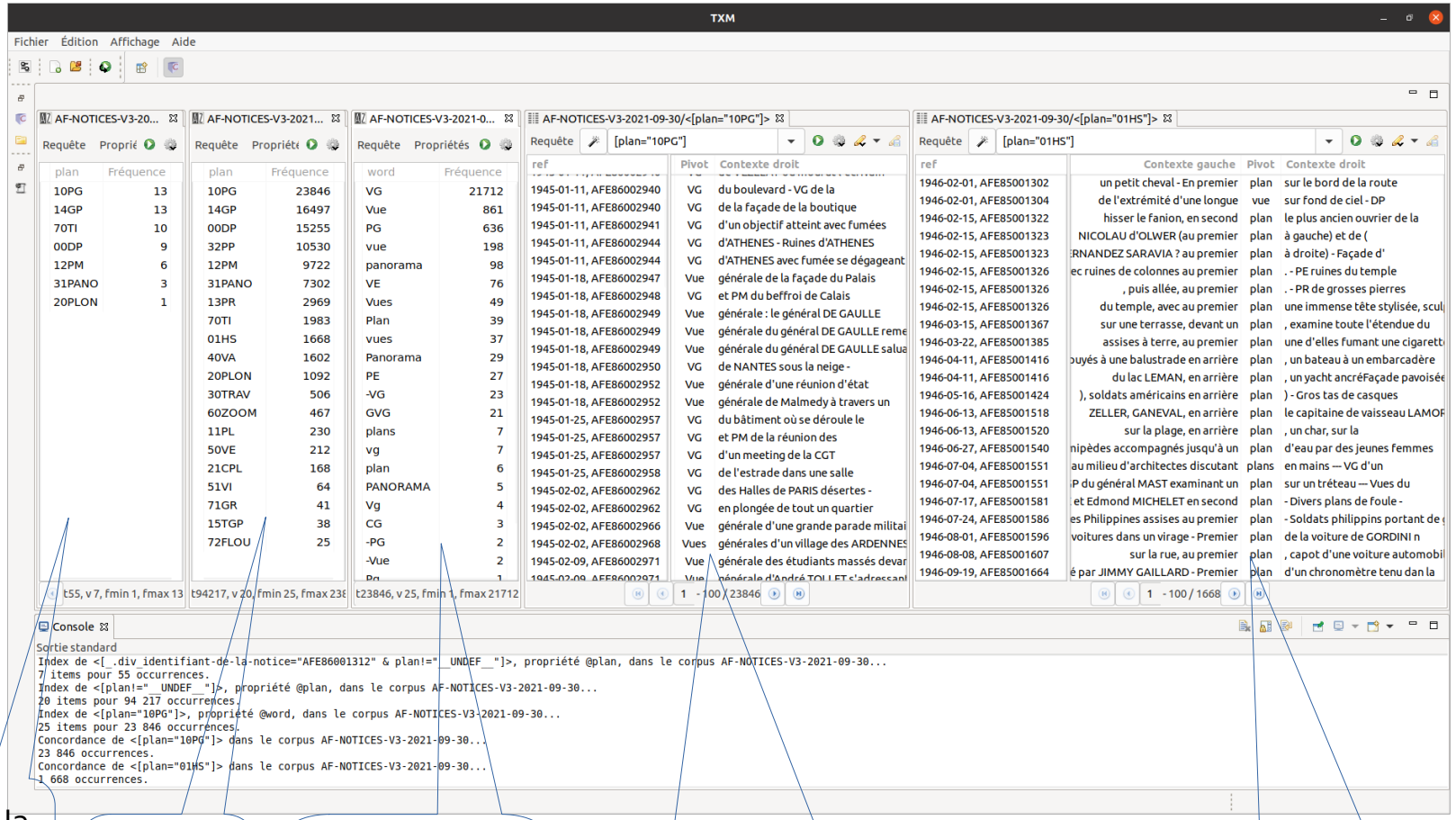
Text AFE86004168

Console

System output

```
Index of <[plan!="__UNDEF__"]>, property @plan, in AF-NOTICES-V3-2021-10-11 corpus...
21 item for 93,553 occurrences.
Concordance of <[_.div_id="AFE86001312" & plan!="__UNDEF__|01HS"]>>in AF-NOTICES-V3-2021-10-11 corpus...
58 occurrences.
Concordance of <<div>[_.div_id="AFE86001312"]>>in AF-PLANS-V2-2021-10-11 corpus...
1 occurrences.
Opening AF-NOTICES-V3-2021-10-11 Browser...
Done.
```

Vue des annotations



Dans la notice Rugby

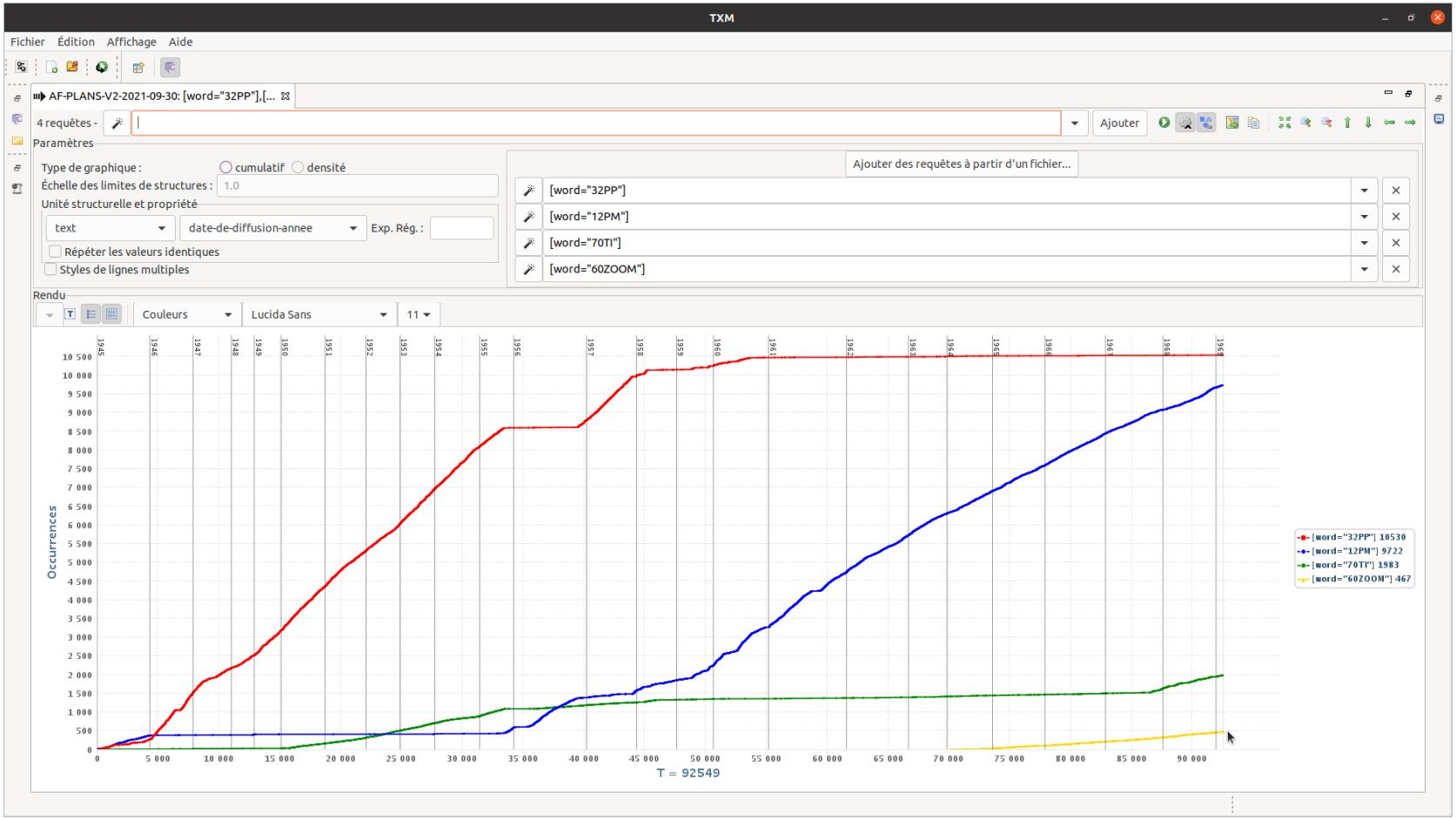
Dans tout le corpus

Les mots sous la catégorie « Plan général »…

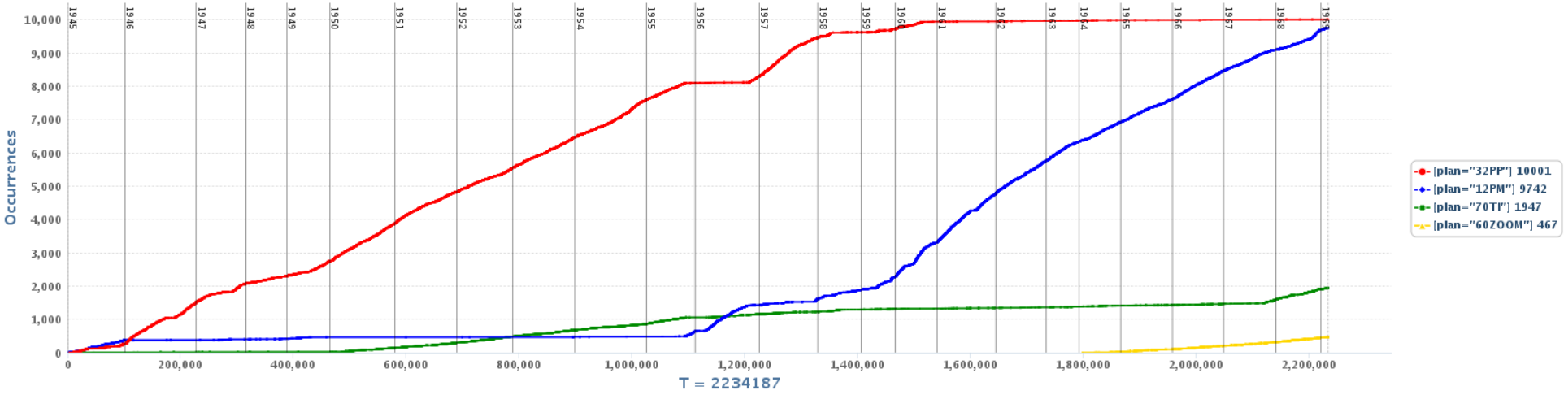
… avec leur contexte (on voit toute l'expression qui justifie l'annotation)

Cas de la catégorie « Hors sujet » (exclue des analyses)

Exemple d'analyse sur le corpus annoté



Exemple d'analyse sur le corpus annoté



- 32PP (Plan proche, rouge) et 12PM (Plan moyen, bleu) montrent des profils complémentaires : Remplacement ? Équivalence ? Consigne de catalogage ?...
- 70TI (Titres, vert) : des périodes « magazine » où les titres semblent plus à la mode ?
- Apparition de 60ZOOM (Zoom, jaune) : nouveauté technique

Projection

- Utilitaire WordProperty2Word
  - Génère un nouveau corpus
    - Il est utile de garder les deux corpus car ils se complètent (le premier est important pour le retour au texte)
  - On peut choisir de ne pas projeter certaines valeurs
    - Ici : pas les « hors-sujet »

Exemple d'analyse sur les plans de début de sujet

AF-PLANS-V2-2021-10-11/<([resume]...

Query Properties: word Edit

word	Frequency
10PG 10PG 10PG	1872
14GP 14GP 14GP	1779
32PP 32PP 32PP	1252
00DP 00DP 00DP	1001
10PG 14GP 10PG	890
10PG 14GP 14GP	812
10PG 00DP 10PG	805
10PG 10PG 00DP	792
10PG 10PG 14GP	766
14GP 10PG 14GP	764
14GP 12PM 14GP	739
14GP 14GP 10PG	722
10PG 12PM 10PG	709
12PM 14GP 14GP	680
10PG 10PG 12PM	633
12PM 12PM 12PM	633
10PG 31PANO 10PG	627
10PG 32PP 32PP	622
14GP 10PG 10PG	615
00DP 10PG 10PG	570
14GP 14GP 00DP	561
10PG 12PM 14GP	560
10PG 00DP 00DP	559
14GP 14GP 12PM	544
12PM 14GP 12PM	543
10PG 12PM 12PM	540
14GP 00DP 14GP	518
00DP 10PG 00DP	505
12PM 10PG 12PM	492
00DP 00DP 10PG	490
00DP 14GP 14GP	487

1 t 71835, v 1896, fmin 1, fmax 1872

AF-PLANS-V2-2021-10-11/<(<resume...

Query Properties: word Edit

word	Frequency
10PG 10PG 10PG	340
10PG 32PP 32PP	176
70TI 10PG 10PG	174
10PG 10PG 00DP	163
10PG 10PG 14GP	156
70TI 10PG 00DP	154
70TI 10PG 32PP	153
10PG 00DP 10PG	150
10PG 31PANO 10PG	143
10PG 00DP 00DP	142
00DP 00DP 00DP	135
10PG 14GP 14GP	130
10PG 32PP 10PG	124
10PG 14GP 10PG	123
10PG 10PG 12PM	119
10PG 12PM 14GP	113
10PG 12PM 12PM	108
10PG 10PG 32PP	103
10PG 12PM 10PG	102
32PP 32PP 32PP	99
10PG 10PG 31PANO	95
10PG 00DP 14GP	90
14GP 14GP 14GP	69
10PG 00DP 32PP	68
70TI 32PP 32PP	68
70TI 10PG 31PANO	67
10PG 32PP 00DP	63
10PG 13PR 10PG	62
10PG 14GP 00DP	61
10PG 00DP 31PANO	60
12PM 14GP 14GP	56

1 t 8611, v 832, fmin 1, fmax 340

AF-PLANS-V2-2021-10-1...

Query Properties: w

word	Frequency
10PG	23721
14GP	16234
00DP	15494
32PP	10001
12PM	9742
31PANO	7199
13PR	2888
70TI	1947
40VA	1394
20PLON	1094
30TRAV	880
60ZOOM	467
11PL	260
50VE	215
21CPL	190
51VI	62
72GR	41
15TGP	34
73FLOU	25

t 91898, v 20, fmin 10, fmax 2372

AF-PLANS-V2-2021-10-...

Query Properties: v

word	Frequency
10PG	4408
00DP	1579
70TI	1464
32PP	765
14GP	668
12PM	489
31PANO	429
40VA	290
13PR	213
30TRAV	60
50VE	44
11PL	25
20PLON	18
60ZOOM	11
51VI	6
21CPL	4
71BT	4
15TGP	2
72GR	2

t 10482, v 20, fmin 1, fmax 4408

AF-PLANS-V2-2021-10-11/ReportsFirstShot...

Property word

Units	Frequency T 91898	AF-PLANS-V2-2021-10-11/ReportsFirstShot t=10482	index
10PG	23721	4408	1,000.0
70TI	1947	1464	1,000.0
40VA	1394	290	23.6
50VE	215	44	4.1
71BT	10	4	1.7
51VI	62	6	-0.4
15TGP	34	2	-0.6
11PL	260	25	-0.7
73FLO	25	1	-0.7
72GR	41	2	-0.9
30TRA	880	60	-5.5
21CPL	190	4	-5.8
00DP	15494	1579	-7.2
60ZOC	467	11	-12.6
13PR	2888	213	-12.8
20PLO	1094	18	-34.9
32PP	10001	765	-38.9
31PAN	7199	429	-59.7
12PM	9742	489	-116.9
14GP	16234	668	-278.8

Console

System output

```
Index of <([resume]{3,3}|[sequences]{3,3}) within div>, property @word, in AF-NOTICES-V3-2021-10-11 corpus...
982,861 item for 1,919,832 occurrences.
Index of <([resume]{3,3}|[sequences]{3,3}) within div>, property @word, in AF-PLANS-V2-2021-10-11 corpus...
1,896 item for 71,835 occurrences.
Index of <(<resume>[resume]{3,3}|<sequences>[sequences]{3,3}) within div>, property @word, in AF-PLANS-V2-2021-10-11 corpus...
832 item for 8,611 occurrences.
Index of <[]>, property @word, in AF-PLANS-V2-2021-10-11 corpus...
20 item for 91,898 occurrences.
Index of <(<resume>|<sequences>)[]>, property @word, in AF-PLANS-V2-2021-10-11 corpus...
20 item for 10,482 occurrences.
Sub-corpus ReportsFirstShot of AF-PLANS-V2-2021-10-11 <(<resume>|<sequences>)[]>...
Done.
Specificities of AF-PLANS-V2-2021-10-11/ReportsFirstShot sub-corpus...
Done.
```

Les segments répétés (SR) de 3 plans pour tout le corpus

SR de 3 plans en début de sujet

L'ensemble des catégories de plan mentionnées (n'importe où : au début ou ailleurs)

Les catégories de plan en début de sujet (1ère mentionnée)

Le calcul statistique des SPÉCIFICITÉS sur les plans en 1ère position dans le sujet met en évidence les sur- et sous-emplois qui peuvent vraiment attirer notre attention

Annotation ou/et Projection : un outil méthodologique

- Pas spécifique aux corpus multimédia
- Annotation semi-automatique par requêtes
  - D’autres modes d’annotation sont disponibles dans TXM
  - Celui-ci nous intéresse pour :
    - la documentation systématique de l’annotation
    - sa compatibilité avec l’évolution du corpus (versions successives)
- Projection comme réalisation pragmatique du « corpus actif »
  - Voir Bénédicte PINCEMIN, Serge HEIDEN, Franck MAZUET (2022) - « The Textometric Concept of Active Corpus. Illustration by an Analysis Scenario based on Annotation then Projection », *in* M. Misuraca *et al*. (eds), *JADT’ 22. Proceedings of the 16th International Conference on Statistical Analysis of Textual Data*, VADISTAT - Per Simona Balbi, Univ. of Naples Federico II, July 6-8 2022. https://halshs.archives-ouvertes.fr/halshs-03667319

Place aux échanges !

+ Bonus post-présentation :
3 diapos de synthèse en conclusion →

- Expérimenter de nouvelles possibilités de la **méthode textométrique**
- Apprivoiser les limites de l’**analyse de corpus**
- Questionner la **linguistique** avec le gradient d’oralité et les polarités Nom/Verbe

Expérimenter de nouvelles possibilités de la méthode textométrique

- Calculer des spécificités à façon
  - Construire l’unité linguistique pertinente (qui n’est pas forcément le lemme ou la catégorie grammaticale tels quels : fusion, diffraction, contextualisation)
  - Et définir aussi le domaine de contraste
- L’AFC comme outil non seulement de visualisation mais aussi d’analyse
  - Mise en évidence et description de la principale variation
- Retour au document source
  - Mots du texte, vidéo…
- Construire une représentation analytique, médiatrice entre texte et traitements
- Explorer ou confirmer ?
  - Une vigilance critique à garder vis-à-vis des a priori

Apprivoiser les limites de l’analyse de corpus

- Quantité/diversité des données et possibilités d’observation
  - Détermination des catégories d’analyse
  - Ellipses de confiance
- Gérer l’imperfection des données et les limites de fiabilité de l’étiquetage
  - Une heureuse corrélation quantité-qualité pour l’étiquetage morphosyntaxique automatique
- Décrire les singularités et accéder aux régularités
  - Un *peeling* bien compris : non pas élagage de cas particuliers perturbants mais étapes successives dans l’analyse
- Ajuster les « degrés de liberté » des données
  - Pour concevoir et interpréter les analyses
- Visualisations adaptées et calculs de synthèses
  - Des outils pour lire et analyser les visualisations elles-mêmes

Questionner la linguistique
avec le gradient d’oralité et les polarités Nom/Verbe

- Un équilibre profond entre un pôle nominal et un pôle pronominal/verbal
  - Peut-on lui donner une explication essentiellement syntaxique ?
  - Un phénomène traversant la diversité des langues (cf. travaux de Biber)
- Affinités à différentes échelles, différentes unités linguistiques :
  - Au niveau infratextuel du passage, avec le discours direct
  - Au niveau intertextuel des genres textuels
- Cet axe de variation se manifeste avec un « poids » (ou « inertie », au sens physique) dominant les autres variations
  - Diachronie, vers/prose...
- Dissymétrie : le pôle pronominal/verbal apparaît plus consistant
  - Effet particulier de notre corpus ou structure fondamentalement dissymétrique ? L’aventure scientifique n’est pas finie...