



**HAL**  
open science

## Gérer le "bruit" dans les corpus en textométrie

Bénédicte Pincemin

► **To cite this version:**

Bénédicte Pincemin. Gérer le "bruit" dans les corpus en textométrie: Retour d'expérience et propositions. Journée d'étude "Bruit de fond ou valeur ajoutée? Gérer le bruit lors des traitements informatiques des corpus linguistiques", Université Grenoble Alpes; Università di Roma La Sapienza, Apr 2023, Grenoble, France. halshs-04127640

**HAL Id: halshs-04127640**

**<https://shs.hal.science/halshs-04127640>**

Submitted on 14 Jun 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

*Journée d'étude « Bruit de fond ou valeur ajoutée ?  
Gérer le bruit lors des traitements informatiques des corpus linguistiques »  
Grenoble, vendredi 28 avril 2023*

# Gérer le « bruit » dans les corpus en textométrie Retour d'expérience et propositions

Bénédicte PINCEMIN

Université de Lyon, CNRS, IHRIM UMR 5317



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

# Plan

- Introduction
- **Contexte 1 : Linguistique diachronique**
  - travailler avec un étiquetage morphosyntaxique automatique non exempt d'erreurs
  - l'analyse factorielle des correspondances, un outil justement fait pour décanter les données ?
- **Contexte 2 : Analyse historique d'archives audiovisuelles**
  - travailler avec une transcription automatique non exempte d'erreurs
  - lancer des analyses focalisées sur les mentions de plan sans les interférences avec les autres mots, par annotation et projection
- Éliminer le bruit en corrigeant ? L'annotation en question
- Apprivoiser le bruit



# S'entendre sur le bruit

- Cadre de l'exposé - mon expérience est celle du bruit :
  - dans une acception conceptuelle (bruit vs silence → relation aux erreurs)
  - plutôt qu'au sens courant (« bruit du ventilateur » → notamment dans les recueils de données audio)
- Trois caractéristiques donnant à réfléchir
  - Relativité à l'objet visé : /surplus/, /non-sens/, /entour/
  - Pas ce qu'on cherche : encombre et occulte, ou nouveau et inattendu (et du coup valeur ajoutée) ?
  - L'identifier et le délimiter : pas d'évidence, continuum ? Fait quand même partie des données ?

# BFM & Oral représenté



- Contexte de la recherche :
  - en équipe : Céline Guillot-Barbance, Alexei Lavrentiev, Serge Heiden, Bénédicte Pincemin
  - en lien avec la *Base de français médiéval*
    - corpus de textes structurés (TEI) : 170 textes, 5 millions de mots
    - initié en 1989, en science ouverte
    - <http://bfm-corpus.org>
  - sur la période 2011-2020
    - avec une évolution du corpus (quantitative et qualitative)
    - et l'apport successif de différents calculs (spécificités, AFC) réalisés avec le logiciel open-source TXM.

# BFM & Oral représenté



- Pour les derniers résultats sur le 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>
- Pour les derniers résultats sur le volet plus méthodologique et statistique :
  - 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://doi.org/10.1007/978-3-030-52680-1_11),  
<https://halshs.archives-ouvertes.fr/halshs-03070182>

# Une recherche avec la Base de français médiéval :

## L'oral représenté au Moyen Âge



- Si l'on s'intéresse à la variation linguistique, le caractère oral ou écrit ne serait-il pas l'une des principales dimensions de variation dans la langue ?
- Pour le Moyen Âge, pas de paroles enregistrées ! Mais :
- Des observables pour une approche contrastive :
  - De l'écrit qui se donne comme de l'oral : le *discours direct* (DD)
  - Des *genres textuels oralisés* (ex. chanson de geste) ou non
- 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)

# Balisage TEI du DD dans la BFM



Annotation semi-automatique, en se basant sur la typographie des éditeurs (guillemets – incises comprises).

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.)

Le « non discours direct » (non-DD) est le « reste », il n'est pas balisé.

=> Dans l'analyse, on le note « **z** » (donc trois modalités possibles : **q / sp / z**).



# BFM & DD : unités d'analyse

- En 2020, travail sur le corpus BFM016DD19 : 4,2 millions de mots, 137 textes.
- Définition de **59 « unités de discours » (DU)** croisant DD et genres textuels :
  - (q / sp / z) x 32 genres = 59 unités de discours (DU)
  - Notation : q\_rbreLfLn, sp\_dramatiqueR2, z\_lettreH1  
soit : {DD : q, sp ou z}\_{genre (28) : lettre, roman, chronique...}{domaine (7) : littéraire, religieux, juridique...}{nb textes : 1, 2 ou n c'est-à-dire >2}
- Description par leur profil morphosyntaxique : **33 étiquettes POS**
  - Information linguistique a priori pertinente
    - Formes graphiques : très dépendante des variations phonétiques et graphiques selon les régions et les époques
    - Lemmes : non disponibles
  - Granularité intermédiaire : jeu d'étiquettes **CATTEX**, 64 valeurs différentes dans notre corpus
  - Sur l'ensemble du corpus, étiquetage automatique avec un **modèle TreeTagger** spécialisé pour le français de cette époque réalisé par l'équipe BFM
    - étiquetage manuel vérifié pour un peu plus d'un million de mots (30 %) dans 38 textes.



# BFM & DD : données pour l'AFC

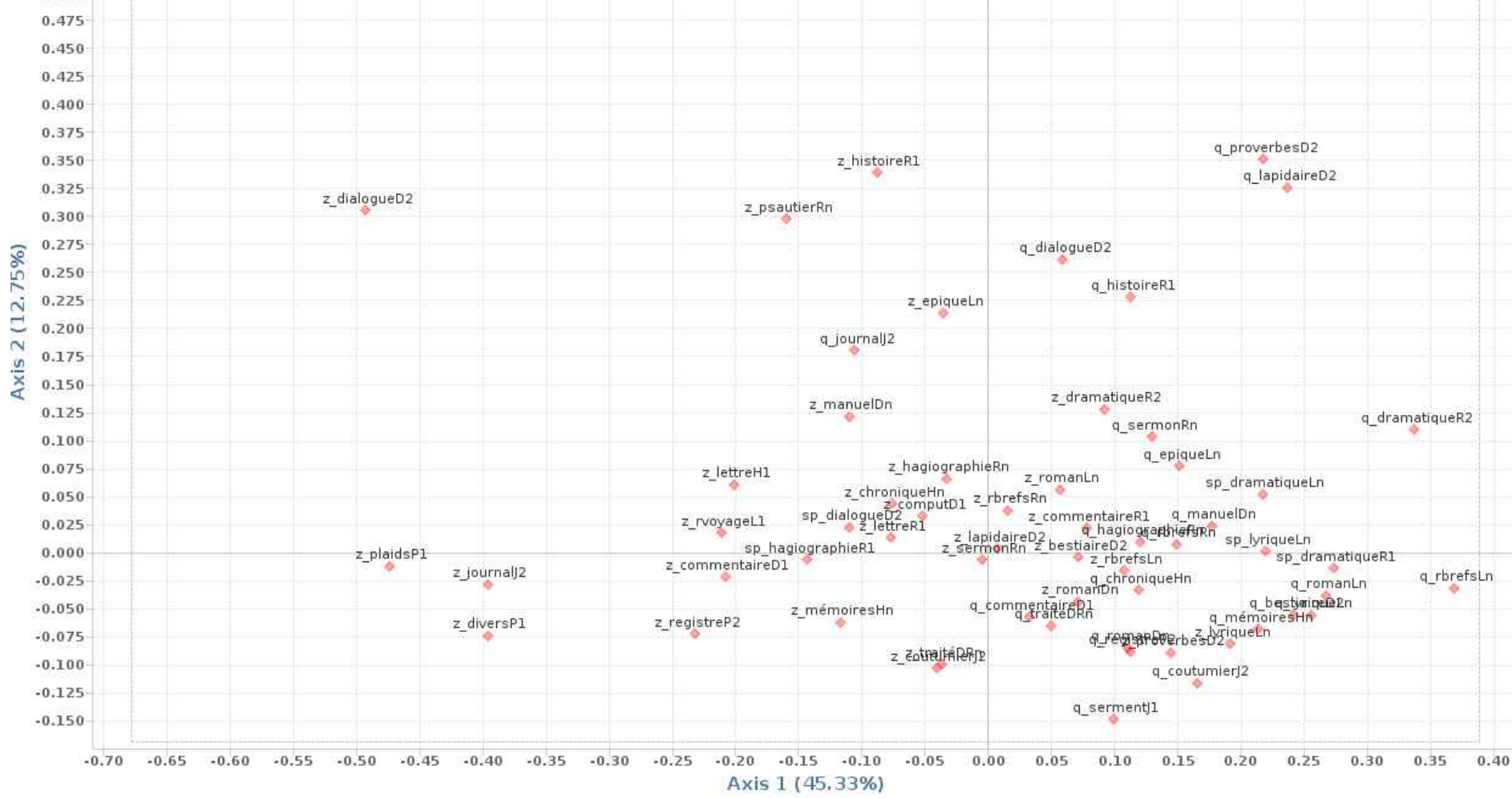


fropos	Frequi	q_bestiaireD2 t=1698	z_bestiaireD2 t=34705	q_chroniqueHn t=63436	z_chroniqueHn t=346440	q_commentaireD1 t=130	z_cc
NOMcom	655584	261	6289	10394	61254	22	
VERcjj	526369	272	5694	9065	46542	15	
PRE	385087	146	3318	6336	37924	19	
PROper	294126	211	2759	6928	23048	17	
ADVgen	235283	108	2208	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	

- Les DU sont décrites par leurs « ingrédients » (telle POS, dans telle proportion)
- Et réciproquement, les POS sont représentées par les types de discours qui les mobilisent plus ou moins fortement.

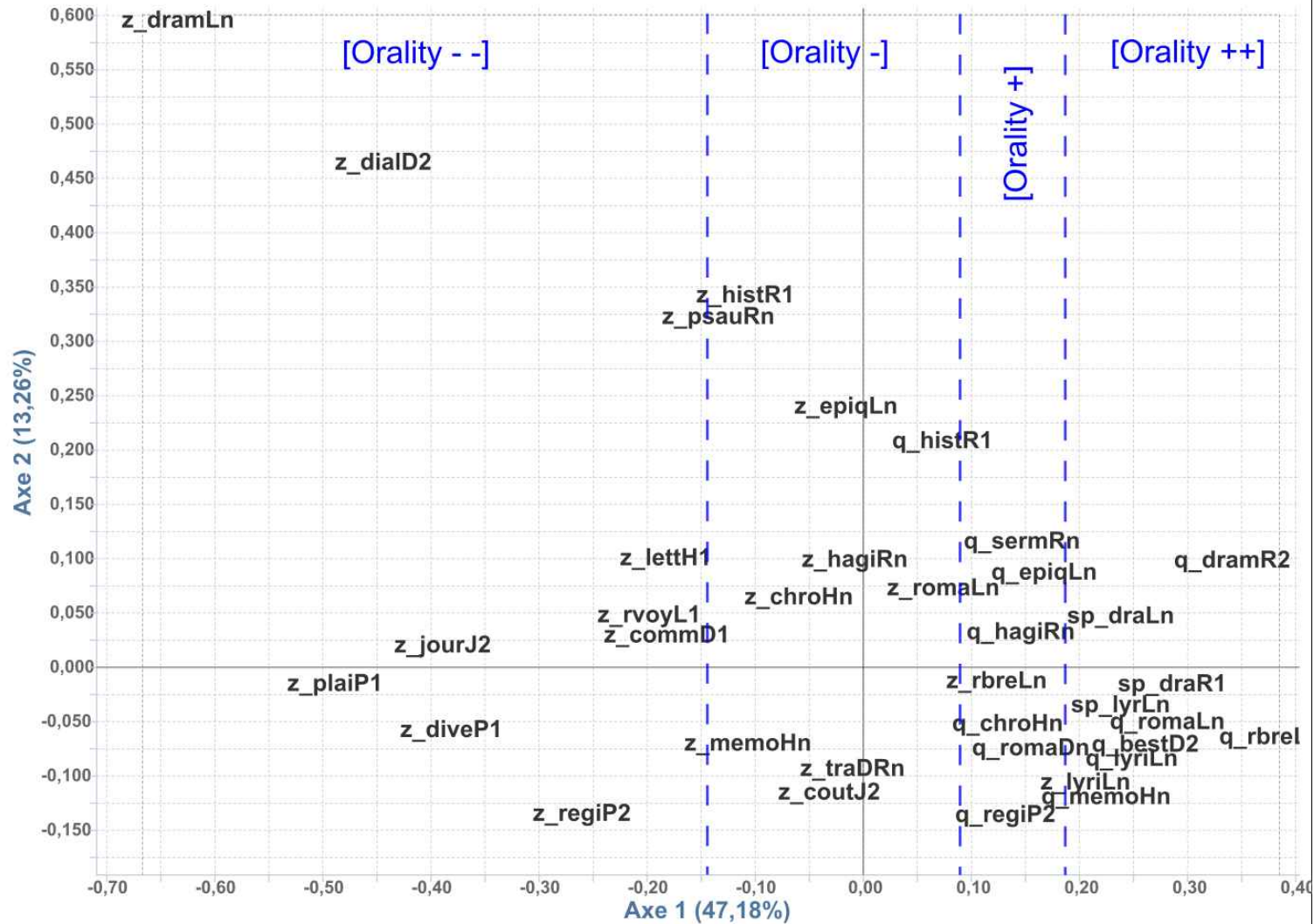


# BFM & DD : AFC 59 DU x 33 POS – Les DU





# BFM & DD : 4 zones







# BFM & DD : gradient d'oralité



Rank	DU	c1	Orb1	Cos <sup>2</sup> 1	Q12	Mass	Word Nb	Ellipse size
1	z_dramatiqueLn	-0,33	0,16	<b>0,31</b>	0,53	0,01	334	medium
2	z_dialogueD2	-0,49	0,07	<b>0,38</b>	0,52	0,01	261	medium
3	z_plaidsP1	-0,47	<b>2,46</b>	<b>0,72</b>	0,72	0,28	10153	small
4	z_diversP1	-0,40	<b>1,53</b>	<b>0,70</b>	0,72	0,25	9025	small
5	z_journalJ2	-0,40	<b>21,26</b>	<b>0,86</b>	0,86	<b>3,51</b>	125563	small
6	z_registreP2	-0,23	<b>18,16</b>	<b>0,69</b>	0,76	<b>8,73</b>	312528	small
7	z_rvoyageL1	-0,21	<b>1,13</b>	<b>0,36</b>	0,36	0,66	23469	small
8	z_commentaireD1	-0,21	0,35	<b>0,70</b>	0,71	0,21	7573	small
9	z_lettreH1	-0,20	0,10	<b>0,39</b>	0,42	0,06	2263	medium
10	z_psautierRn	-0,16	<b>1,27</b>	0,10	0,46	<b>1,29</b>	46058	small
11	sp_hagiographieRn	-0,14	0,35	0,14	0,14	0,44	15732	small
12	z_mémoiresHn	-0,12	<b>2,86</b>	<b>0,31</b>	0,40	<b>5,40</b>	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	<b>1,50</b>	53537	small
17	z_lettreR1	-0,08	0,19	0,13	0,13	0,84	30003	small
18	z_chroniqueHn	-0,08	<b>2,06</b>	<b>0,32</b>	0,42	<b>9,24</b>	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	<b>4,37</b>	156475	small
21	z_traitéDRn	-0,04	0,25	0,03	0,23	<b>4,91</b>	175799	small
22	z_epiqueLn	-0,04	0,06	0,02	0,58	<b>1,22</b>	43785	small
23	z_hagiographieRn	-0,03	0,15	0,08	0,42	<b>3,44</b>	123094	small
24	z_sermonRn	0,00	0,00	0,00	0,00	<b>3,26</b>	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	<b>2,29</b>	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	<b>1,91</b>	0,20	0,41	<b>15,25</b>	545635	small
30	q_dialogueD2	0,06	0,00	0,01	0,15	0,00	91	large
31	z_romanDn	0,07	0,81	<b>0,36</b>	0,50	<b>4,20</b>	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	<b>1,26</b>	45201	small
34	z_dramatiqueR2	0,09	0,02	0,10	0,31	0,06	2319	medium

[Orality - -]

[Orality +]

35	q_sermentJ1	0,10	0,00	0,05	0,16	0,00	101	large
36	z_rbrefsLn	0,11	0,37	<b>0,34</b>	0,35	0,83	29696	small
37	q_romanDn	0,11	<b>1,20</b>	<b>0,45</b>	0,72	<b>2,56</b>	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	<b>0,51</b>	0,55	<b>1,72</b>	61578	small
41	q_hagiographieRn	0,12	0,54	<b>0,48</b>	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	<b>0,36</b>	0,46	<b>1,00</b>	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	<b>3,54</b>	<b>0,54</b>	0,64	<b>2,49</b>	89262	small
49	q_mémoiresHn	0,21	0,11	<b>0,42</b>	0,46	0,06	2150	medium
50	sp_dramatiqueLn	0,22	<b>1,09</b>	<b>0,62</b>	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	<b>0,49</b>	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	<b>0,64</b>	0,67	0,05	1664	medium
55	q_lyriqueLn	0,26	0,20	<b>0,62</b>	0,65	0,08	2828	medium
56	q_romanLn	0,27	<b>28,80</b>	<b>0,91</b>	0,92	<b>10,44</b>	373570	small
57	sp_dramatiqueR1	0,27	0,92	<b>0,81</b>	0,81	0,32	11435	small
58	q_dramatiqueR2	0,34	0,05	<b>0,37</b>	0,41	0,01	409	medium
59	q_rbrefsLn	0,37	<b>2,37</b>	<b>0,78</b>	0,78	0,45	16115	small

[Orality -]

[Orality ++]

## **[+ oralité]**

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

## **[- oralité]**

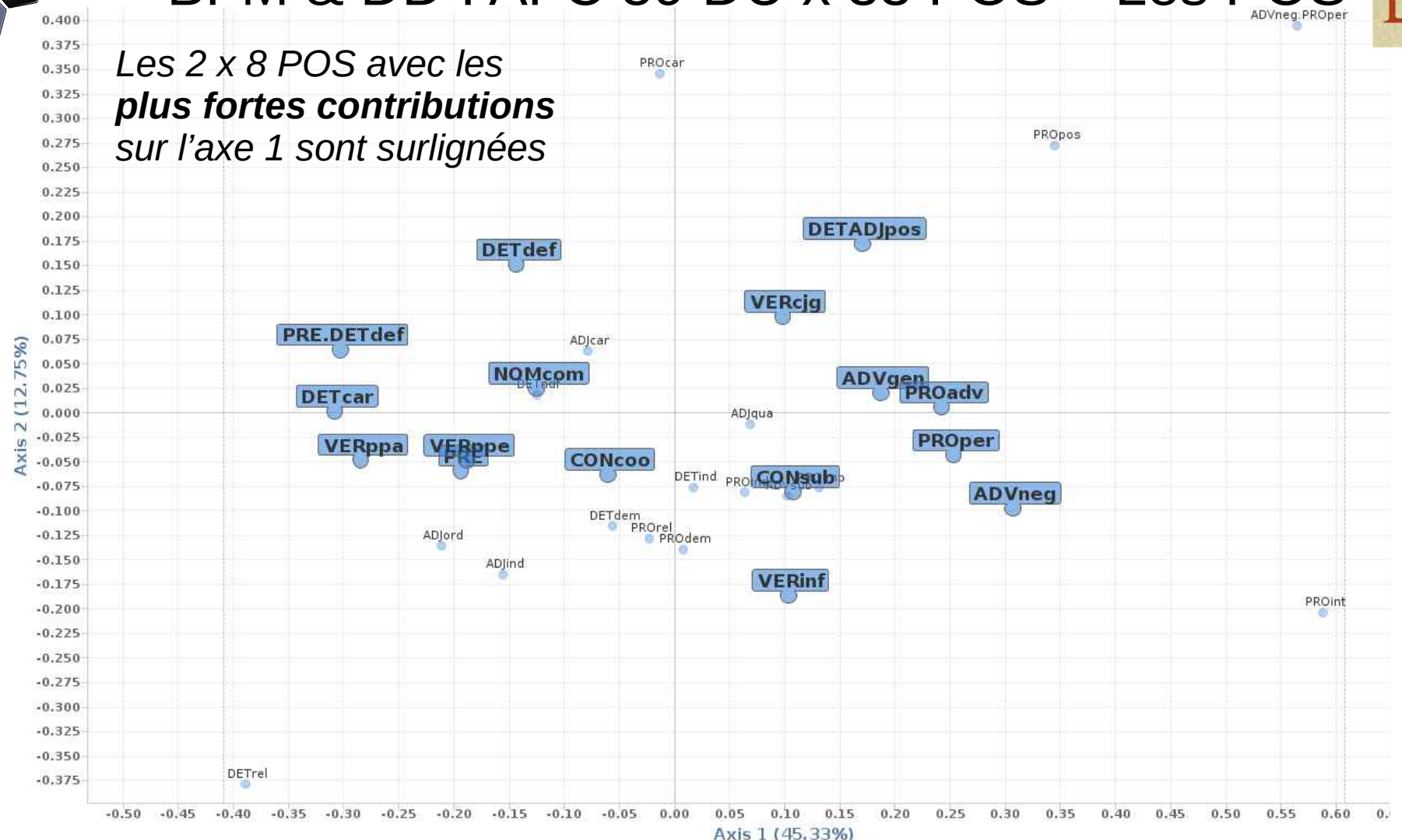
- Aucun genre oralisé ; pas d'oral représenté
- Genres des domaines juridiques et actes de la pratique (chartes, registres, comptes...)
- Apparaît plus hétérogène, plus dispersé.



# BFM & DD : AFC 59 DU x 33 POS – Les POS



Les 2 x 8 POS avec les plus fortes contributions sur l'axe 1 sont surlignées



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



## [- oralité]

Préposition  
Nom commun  
Préposition + Déterminant défini  
Participe passé  
Déterminant défini  
Déterminant cardinal  
Participe présent  
Conjonction de coordination

## [+ oralité]

Pronom personnel  
Adverbe général  
Adverbe de négation  
Verbe conjugué  
Pronom adverbial (*en, y*)  
Déterminant ou adjectif possessif  
Conjonction de subordination  
Verbe à l'infinitif



# BFM & DD : retour en détail sur la méthode

## Comment a-t-on géré le « bruit » ?



- **Étiquetage morphosyntaxique automatique**  
(étiquettes POS)
  - Ajuster la représentation des données en fonction du type d'analyse menée, avec deux critères : fiabilité et pertinence
- L'apport des outils statistiques
  - AFC et ordre des dimensions : fait pour décanter le bruit ?
  - La « loi des grands nombres » ?
  - Évaluer et gérer le manque de données avec les ellipses de confiance
  - Adopter les bons repères de lecture : les aides à l'interprétation de l'AFC

# Ajustement POS (1) : connaissance des données



- Une évaluation systématique des étiquettes est menée en comparant l'étiquette automatique (fropos) et l'étiquette manuelle (pos) sur les textes vérifiés qui n'ont pas fait partie du corpus d'apprentissage.

	A	B	C	D	E	F
1			match			
2	fropos	Données	n	o	Total Résultats	
3	ADJcar	Somme - adgar	10	10	20	50,00%
4		Somme - aucassin	1	0	1	0,00%
5		Somme - beroul	5	25	30	83,33%
6		Somme - brut2	1	10	11	90,91%
7		Somme - comput	12	17	29	58,62%
8		Somme - DialGreg2	12	3	15	20,00%
9		Somme - gcoin1	12	3	15	20,00%
10	ADJind	Somme - adgar	18	18	36	50,00%
11		Somme - aucassin	0	7	7	100,00%
12		Somme - beroul	4	16	20	80,00%
13		Somme - brut2	7	25	32	78,13%
14		Somme - comput	9	12	21	57,14%
15		Somme - DialGreg2	8	56	64	87,50%
16		Somme - gcoin1	4	12	16	75,00%
17	ADJord	Somme - adgar	1	5	6	83,33%

*Premières lignes du tableur d'évaluation  
(document de travail, Lavrentiev, 2012)*

*En rouge, les taux de réussite inférieurs à 75 %*

# Ajustement POS (2) : diagnostic

- On en déduit 3 groupes d'étiquettes :

## 1) Taux suffisant

- ADJord
- ADVgen
- ADVneg
- CONcoo
- CONsub
- DETdef
- DETdem
- DETind
- DETndf
- DETpos
- NOMcom
- PRE
- PRE.DETdef
- PROdem
- PROper
- PROper.PROper
- PROpos
- PROrel
- VERcjcj
- VERinf
- VERppa

## 2) Taux insuffisant ou fréquence trop faible + les ponctuations (à ne pas utiliser dans les calculs)

- ADJcar
- ADJpos
- ADVgen.PRO...
- ADVint
- ADVsub
- CONsub.PRO...
- DETcom
- DETint
- DETord
- DETrel
- ETR
- INJ
- OUT
- PON...
- PRE.DETcom
- PRE.PRO...
- PROcar
- PROimp
- PROint
- PROord
- PROrel.PRO...
- RED

## 3) Taux problématique (à discuter)

- ADJind (comput 57%, fréquence faible)
- ADJqua (DialGreg2 63% et comput 71%, fréquence haute)
- DETcar (DialGreg2 53% et comput 71%)
- NOMpro (adgar 59% et comput 64%)
- PROadv (DialGreg2 64%)
- PROind (comput 58% et adgar 73%)
- VERppe (comput 73% et DialGreg2 73%)



# Ajustement POS (3) : seuillage

- On remarque que les plus basses fréquences concentrent les étiquettes non fiables
  - Cela paraît d'ailleurs assez cohérent du fait du fonctionnement de l'étiqueteur, basé sur un apprentissage
- Les basses fréquences ont également un rôle très faible dans l'AFC (basée sur le « poids », l'inertie)  
=> filtrage des fréquences < 1700

Paramètres

Seuils

Fmin 1700 - + Fmax 4225629 - + Vmax 4225629 - +

Fusion ou Suppression de colonnes Fusion ou Suppression de lignes

Fropos	Fréquence	_bestiaireD2 t=1696	z_bestiaireD2 t=34710	q_chroniqueHn t=63462	z_chroniqueHn t=346247	q_commentaireD1 t=130	z_
PROimp	2155	4	36	26	184	0	
PROpos	1765	2	27	54	130	1	
INJ	1635	4	4	60	22	0	
ADVint	1230	3	11	29	31	0	
PRE.PROrel	837	0	0	2	39	0	
PROord	822	0	4	3	103	0	
ADVgen.PROoper	720	0	8	17	34	0	
DETint	640	0	4	15	44	0	
PROoper.PROoper	462	0	0	16	6	0	
PRE.DETrel	405	0	0	2	34	0	
PROrel.PROoper	252	1	7	5	28	0	
ADVgen.PROadv	74	0	4	2	4	0	
ADVing	48	0	0	3	4	0	
PRE.PROoper	46	0	0	2	2	0	
CONsub.PROoper	43	0	1	3	8	0	
PROcom	12	0	0	0	0	0	
PRE.PROcom	11	0	0	0	1	0	
PROoper.PROadv	10	0	0	4	1	0	
PROint.PROoper	8	0	0	0	0	0	
PROrel.PROadv	7	0	1	0	1	0	
ADVneg.PROadv	3	0	0	0	0	0	

T 3699406 V 56 Fmin 3 Fmax 644922

Console

Sortie standard

Table lexicale de l'index de partition BFM016DD19/genre2017xdd\_alpha/<[fropos!="PON.\*|ETR|OUT|RED|ABR"]>@fropos ≤4 225 629 /4 225 629...



# Ajustement POS (4) : influence sur l'analyse



Lignes	Q12	Q13	Q23	Masse	Dist	▼ Cont1	Cos <sup>2</sup> 1	Cont2	Cos <sup>2</sup> 2
DETcom	0,83	0,92	0,17	0,25	3,77	17,09	0,79	1,98	0,04
NOMpro	0,58	0,79	0,22	3,03	0,36	14,67	0,58	0,14	0,00
PRE.DETcom	0,77	0,87	0,19	0,15	4,87	12,64	0,73	1,63	0,04
PROper	0,57	0,71	0,15	7,90	0,09	9,11	0,57	0,19	0,01
PRE	0,79	0,83	0,04	10,44	0,04	7,13	0,79	0,00	0,00
ADVgen	0,74	0,69	0,10	6,42	0,05	5,78	0,66	1,42	0,07
VERcjb	0,54	0,64	0,14	13,93	0,03	4,95	0,52	0,38	0,02
ADVneg	0,71	0,78	0,14	2,06	0,13	4,58	0,68	0,47	0,03

- On calcule une première AFC sur le tableau seuillé et on examine les POS qui contribuent de façon majeure à l'axe 1.
- DETcom et PRE.DETcom (*ledit*) introduisent un effet diachronique singulier (apparaît fin XIIIe s. : cas rare d'apparition d'une catégorie) : **on le note** et on retire ces catégories de l'analyse (*peeling*) pour pouvoir observer les contrastes liés aux genres et DD/nonDD.





# Ajustement POS (5) : influence sur l'analyse



- On recalcule l'AFC (tableau seuillé et sans DETcom/PRE.DETcom), et on examine à nouveau les POS qui contribuent à l'axe 1 de façon marquée (supérieure à la moyenne ~3%).
- On trouve une seule POS non fiable qu'il ne faut pas laisser influencer de façon majeure l'analyse : NOMpro  
→ à éliminer  
(et s'en rappeler pour l'interprétation des résultats)
- Pour les autres peu fiables, impact acceptable :
  - soit  $ctr1 < 1\%$ ,
  - soit  $< 3\%$  avec juste 1 texte sur 7 de qualité 50-70 (DETcar, PROadv, ADJqua),
  - soit  $< 5\%$  mais quasi pas de pb de qualité (tous les textes de qualité au-dessus de 70) (VERppe)

Lignes	Q12	Q13	Q23	Masse	Dist	Cont1	Cos²1
NOMpro	0,65	0,94	0,31	3,04	0,37	22,02	0,64
PROper	0,72	0,73	0,04	7,94	0,09	14,73	0,71
PRE	0,74	0,89	0,18	10,48	0,04	9,29	0,72
ADVgen	0,68	0,69	0,07	6,45	0,05	7,16	0,65
ADVneg	0,79	0,79	0,01	2,07	0,12	6,90	0,79
NOMcom	0,79	0,78	0,06	17,54	0,01	5,87	0,75
VERcjq	0,44	0,77	0,40	13,99	0,03	4,74	0,40
VERppe	0,65	0,58	0,08	3,46	0,06	4,20	0,57
PRE.DETdef	0,70	0,66	0,07	1,62	0,09	3,60	0,64
DETdef	0,34	0,27	0,07	5,37	0,06	2,87	0,27
VERinf	0,44	0,55	0,29	2,83	0,06	2,46	0,35
DETcar	0,58	0,53	0,09	0,37	0,33	2,46	0,51
PROadv	0,43	0,50	0,12	1,15	0,11	2,26	0,41
ADJqua	0,28	0,29	0,01	3,27	0,03	1,88	0,28
CONsub	0,61	0,61	0,20	3,58	0,03	1,86	0,51
DETpos	0,43	0,53	0,23	2,14	0,06	1,85	0,37
VERppa	0,35	0,39	0,06	0,37	0,15	1,09	0,34
DETrel	0,23	0,33	0,20	0,12	0,64	0,75	0,18
PROint	0,38	0,38	0,01	0,06	0,51	0,63	0,37
ADVneg.PROper	0,31	0,51	0,23	0,06	0,50	0,46	0,30
DETdem	0,21	0,17	0,15	0,69	0,06	0,43	0,12



# Ajustement POS (6) : influence sur l'analyse



- On recalcule l'AFC (tableau seuillé et sans DETcom/PRE.DETcom ni NOMpro).
- On observe que ADJpos s'accapare l'axe 2, en lien avec un seul genre, le psautier (usage spécifique : « la meie aneme », « la tue misericorde »... )
  - cela ne participe pas à une description mettant en relation les genres
  - l'ADJpos correspond à un usage proche du DETpos (vers lequel il va évoluer)

Lignes	Q12	Q13	Q23	Mass1	Dist	Cont1	Cos²1	Cont2	Cos²2
PROper	0,76	0,84	0,16	8,19	0,08	17,00	0,72	1,38	0,04
PRE	0,87	0,91	0,13	10,81	0,04	14,50	0,83	1,08	0,04
NOMcom	0,84	0,84	0,02	18,09	0,02	9,63	0,84	0,15	0,01
ADVgen	0,69	0,75	0,09	6,65	0,05	8,37	0,67	0,31	0,02
ADVneg	0,72	0,80	0,08	2,13	0,11	7,09	0,72	0,03	0,00
VERcvg	0,51	0,68	0,31	14,43	0,03	6,08	0,44	1,51	0,07
PRE.DETdef	0,74	0,75	0,06	1,67	0,10	5,19	0,71	0,25	0,02
VERppe	0,60	0,51	0,10	3,57	0,07	5,00	0,50	1,40	0,09
DETdef	0,39	0,73	0,43	5,54	0,06	4,73	0,34	0,93	0,05
PROadv	0,47	0,50	0,05	1,19	0,10	3,04	0,46	0,09	0,01
DETcar	0,50	0,41	0,09	0,38	0,37	2,53	0,41	0,79	0,09
DETpos	0,49	0,42	0,13	2,21	0,05	2,28	0,39	0,22	0,10
<b>ADJpos</b>	<b>0,98</b>	<b>0,05</b>	<b>0,96</b>	<b>0,10</b>	<b>16,27</b>	<b>1,89</b>	<b>0,0</b>	<b>84,48</b>	<b>0,95</b>
VERinf	0,34	0,45	0,31	2,92	0,05	1,87	0,24	1,10	0,10
ADJqua	0,21	0,27	0,06	3,38	0,03	1,62	0,21	0,00	0,00

```

Sortie standard
Analyse des correspondances de la table lexicale BFM016DD19/genre2017xdd_alpha/<[fropos!="PON.*|ETR|OUT|RED|ABR"]>@fropos :
Terminé.
Index de <[[fropos="ADJpos" & .text_genre="psautier"]]>, propriété @word, dans le corpus BFM016DD19...
738 items pour 1 743 occurrences.
  
```

=> on fusionne ADJpos et DETpos en DETADJpos.

# Ajustement POS (7) : récapitulatif



- Connaissance des données et diagnostic : repérage des étiquettes POS non fiables
- Concentration des POS non fiables en basse fréquence → **seuillage** ( $f > 1\,700$ )
- On écarte DETcom/PRE.DETcom (*ledit*), car cas isolé (majeur) de changement diachronique occultant les observations sur DD et genres (**peeling**)
- Il faut également **filtrer** NOMpro, car peu fiable et influence majeure sur l'axe 1 sur lequel se focalise l'analyse.
- On **fusionne** ADJpos à DETpos, dont l'usage est proche, pour éviter que l'association isolée de l'ADJpos et du psautier n'accapare l'axe 2, et avoir ainsi un plan factoriel plus équilibré.
- L'AFC repose au final sur 33 POS, couvrant 3 565 375 occurrences, soit plus de 96 % des mots (après exclusion des ponctuations, marques éditoriales, mots étrangers et abréviations).

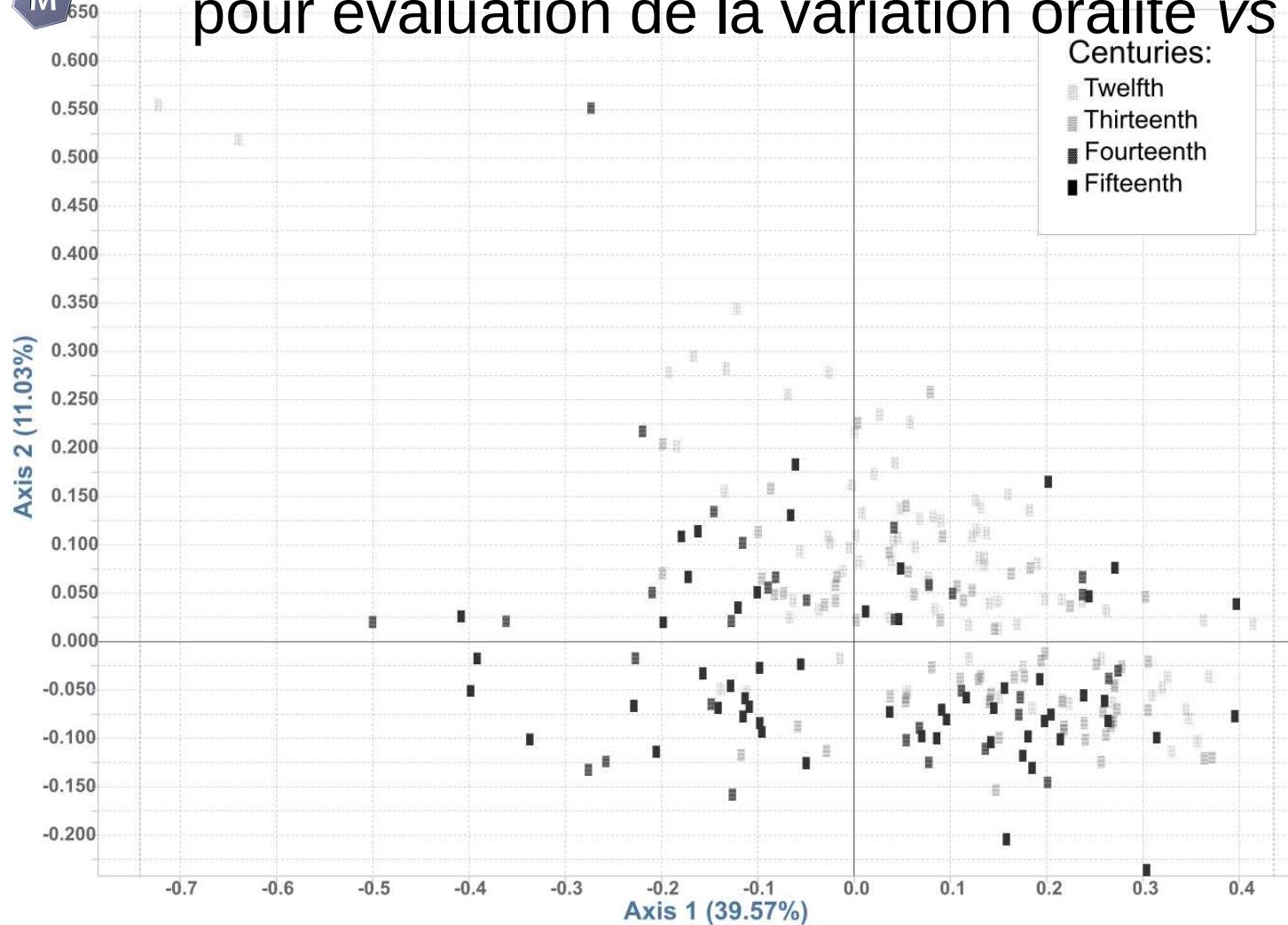


# Ajustement POS (8) : discussion

- Est-ce rigoureux ? Est-ce qu'on n'arrange pas les données pour leur faire dire ce qu'on veut ?
  - Le seuillage qui enlève les basses fréquences allège l'analyse **sans modifier les résultats** (poids trop faible) → cohérent avec l'outil d'analyse utilisé
  - Les opérations de *peeling*, par élimination mais aussi fusion (élimination d'une distinction) :
    - Elles consistent à écarter un point phénomène singulier qui accapare l'analyse, pour **pouvoir poursuivre** celle-ci.
    - Cas rencontrés :
      - DETcom/PREDETcom : apparition de *ledit* au XIIIe s., qui a sa propre catégorie grammaticale.
      - NOMpro : on ne veut pas laisser une étiquette insuffisamment fiable influencer fortement l'analyse (22%).
      - Association spécifique de l'ADJpos et du Psautier.
    - Attention, on ne fait pas disparaître sous le tapis, mais on note la première observation pour en **tenir compte dans l'analyse globale**.
    - Par ex., nous avons fait une AFC complémentaire qui montre que l'oralité est une dimension de contraste même plus forte que la diachronie, **en précisant** bien que ce résultat **ne tient pas compte** de l'apparition de *ledit* au XIIIe s.



# BFM & DD : AFC 238 DU (textesxDD) x 33 POS pour évaluation de la variation oralité vs 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 XIII<sup>e</sup> s.

# BFM & DD : retour en détail sur la méthode

## Comment a-t-on géré le « bruit » ?

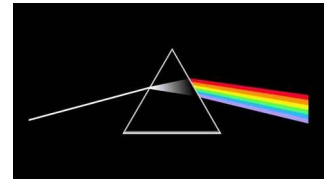


- Étiquetage morphosyntaxique automatique (étiquettes POS)
  - Ajuster la représentation des données en fonction du type d'analyse menée, avec deux critères : fiabilité et pertinence
- L'apport des outils statistiques
  - **AFC et ordre des dimensions : fait pour décanter le bruit ?**
  - La « loi des grands nombres » ?
  - Évaluer et gérer le manque de données avec les ellipses de confiance
  - Adopter les bons repères de lecture : les aides à l'interprétation de l'AFC

# Apport des outils statistiques (1) : l'AFC

## comme outil de hiérarchisation de l'information

- Par construction mathématique, l'axe 1 est choisi pour porter le maximum d'inertie (force de contraste & place prise dans le corpus)
  - Nous l'avons mobilisé comme outil analytique pour dé-composer l'information et mettre en évidence le caractère majeur de la variation observée
- Pour chaque entier  $n$ , l'axe  $n$  est construit pour porter le maximum d'inertie de l'espace restant (hors dimensions 1 à  $n-1$ ). Les dernières dimensions se partagent l'information restante : contrastes faibles, faiblement présents dans le corpus, dispersés/désordonnés (pas de renforcement collectif).
  - L'AFC est ainsi proposée comme moyen de « compression » des données : en réduisant les données aux  $n$  premiers axes, on dégage l'information la plus significative (en force et en présence), et **on élimine le bruit des petites variations isolées**.
  - Exemple d'application : la classification opérée sur les coordonnées factorielles limitées aux  $n$  premières dimensions (stratégie appliquée par la CAH de TXM, ne considérant pas plus que les 5 premières dimensions d'une AFC sous-jacente).



# BFM & DD : retour en détail sur la méthode

## Comment a-t-on géré le « bruit » ?



- Étiquetage morphosyntaxique automatique (étiquettes POS)
  - Ajuster la représentation des données en fonction du type d'analyse menée, avec deux critères : fiabilité et pertinence
- L'apport des outils statistiques
  - AFC et ordre des dimensions : fait pour décanter le bruit ?
  - **La « loi des grands nombres » ?**
  - Évaluer et gérer le manque de données avec les ellipses de confiance
  - Adopter les bons repères de lecture : les aides à l'interprétation de l'AFC

# Apport des outils statistiques (2) : vertus de la « loi des grands nombres » ?

- Étienne Brunet, pionnier de la textométrie et auteur du logiciel Hyperbase, invoquait volontiers la « loi des grands nombres ».
- [Wikipedia](#) en donne une présentation technique (mathématique) :
  - « En mathématiques, la loi des grands nombres permet d'interpréter la probabilité comme une fréquence de réalisation, justifiant ainsi le principe des sondages, et présente l'espérance comme une moyenne. Plus formellement, elle signifie que la moyenne empirique, calculée sur les valeurs d'un échantillon, converge vers l'espérance lorsque la taille de l'échantillon tend vers l'infini. »
- Autrement dit, de façon moins technique, c'est l'idée qu'avec un grand nombre d'observations (la grande fréquence d'un mot, la grande taille du corpus), les petites irrégularités et les écarts contingents, allant dans tous les sens, **se compensent** mutuellement, si bien que **la mesure synthétique obtenue peut être juste** alors même que le détail de quelques données ne permettait pas de se faire une bonne idée.

# Apport des outils statistiques (2) : vertus de la « loi des grands nombres » ?

- (Brunet, 1987, « Les noms propres chez Zola », rééd. 2016, *Écrits choisis*, tome III, p. 176)
  - « Il serait vain de nier ces difficultés théoriques et pratiques et, faute d'avoir désambiguïsé les quelque 72 000 occurrences de noms propres chez Zola, on ne saurait prétendre à l'exactitude du détail. Mais cette situation est assez répandue là où s'exerce la statistique, c'est-à-dire là où la mesure directe des phénomènes est peu praticable. Et l'on peut tirer des conclusions relativement précises et sûres à partir de données floues et incertaines, quand du moins la **loi des grands nombres** permet d'espérer la neutralisation des erreurs et des 'bruits' ».

## Apport des outils statistiques (2) : vertus de la « loi des grands nombres » ?

- Brunet, 2012, « Au fond du GOOFRE, un gisement de 44 milliards de mots », rééd. 2016, *Écrits choisis*, tome III :
  - Suite à l'initiative de numérisation de Google (*Google Books* et projet *Culturonomics*), É. Brunet étudie les possibilités d'exploration offertes par l'outil d'interrogation proposé par Google (*Ngram Viewer*) puis montre l'intérêt de pouvoir analyser les mêmes données avec son logiciel Hyperbase.



# Apport des outils statistiques (2) : vertus de la « loi des grands nombres » ?

- (Brunet, 2012, « Au fond du GOOFRE », rééd. 2016, p. 105)
  - « L'automate n'a pas su reconnaître les **s longs** de l'ancienne typographie (confondus avec des *f*). Et cette seule négligence a des effets incalculables dont témoigne la liste des spécificités de la première période (tableau 2). les éléments insolites qui se portent en tête de liste (*eft*, *fe*, *fur*, *fa*) sont des avatars mal repérés des formes régulières (*est*, *se*, *sur* et *sa*) et cela jette la suspicion sur tous les mots où l'on trouve un *s* ou un *f*, c'est-à-dire un tiers du lexique français. »

N°	écart	corpus	texte	mot
1	4253.21	3140768	2555713	étoit
1	4152.23	3949728	2848331	avoit
1	3960.18	195664144	35784324	:
1	3120.05	1280050	1178085	eft
1	2630.89	242179314	36059703	qu'
1	2522.23	1071243	883501	étoient
1	2279.89	1142929	838962	avoient
1	2059.68	411269418	52964418	que
1	2059.05	883838	663952	enfants
1	2022.51	837551	634116	fe
1	1998.37	53798781	9668157	vous
1	1861.42	175043732	24583789	on
1	1858.26	692574	529092	tems
1	1835.28	17471386	4007321	roi
1	1784.60	88742448	13752541	je
1	1734.17	379543570	47459314	il
1	1725.60	903139	577683	seroit
1	1686.97	687651	484637	habitans
1	1670.47	679676	477372	auroit
1	1609.60	546573	408475	pourvoit
1	1542.44	75773145	11466126	ils
1	1446.42	22902414	4319099	i
1	1410.82	323648405	39448288	qui
1	1393.91	687083	412015	ame
1	1384.09	198259170	25339482	ne
1	1358.64	32750023	5543476	point
1	1328.25	6755990	1694430	prince

# Apport des outils statistiques (2) : vertus de la « loi des grands nombres » ?

- Brunet, 2012, « Au fond du GOOFRE, un gisement de 44 milliards de mots », rééd. 2016, *Écrits choisis*, tome III, p. 116
  - Dans la conclusion,  
« La **loi des grands nombres** n'efface pas complètement les fautes de conception, les bévues des machines et les erreurs de traitement. Un corpus restreint mais plus pur vaut mieux qu'un teruil énorme, mais dégradé. »



=> Différencier l'erreur désordonnée, dispersée, aléatoire, et l'erreur orientée (toujours dans le même sens), répétitive.

# BFM & DD : retour en détail sur la méthode

## Comment a-t-on géré le « bruit » ?



- Étiquetage morphosyntaxique automatique (étiquettes POS)
  - Ajuster la représentation des données en fonction du type d'analyse menée, avec deux critères : fiabilité et pertinence
- L'apport des outils statistiques
  - AFC et ordre des dimensions : fait pour décanter le bruit ?
  - La « loi des grands nombres » ?
  - **Évaluer et gérer le manque de données avec les ellipses de confiance**
  - Adopter les bons repères de lecture : les aides à l'interprétation de l'AFC

# Apport des outils statistiques (3) : « petits nombres » et manque de données

- Par exemple : dans notre analyse sur le discours direct dans les genres, un genre avec peu de textes et peu de DD (ou peu de non-DD).
  - Une phrase de plus ou un mot de moins peuvent alors avoir un impact non négligeable sur le profil POS représentant la DU
- Une gestion qualitative : connaître son corpus ; et rappel, dans la notation, de la multiplicité des textes (1, 2 ou plus).
- Une gestion quantitative, statistique : validation par rééchantillonnage (*bootstrap*) et ellipses de confiance
  - On trace autour de chaque point une ellipse qui montre dans quelle mesure sa position est impactée s'il y a de petites variations dans les données – autrement dit, on introduit expérimentalement un « bruit » plausible, pour évaluer si les résultats se maintiennent
  - Petite ellipse => la position est stable, la représentation est fiable ;
  - Grande ellipse, ellipse traversant l'origine et s'étendant de part et d'autre => la position apparaît contingente, elle ne se prête pas à une interprétation générale
  - Confirmation de la quantité suffisante ou insuffisante de données.

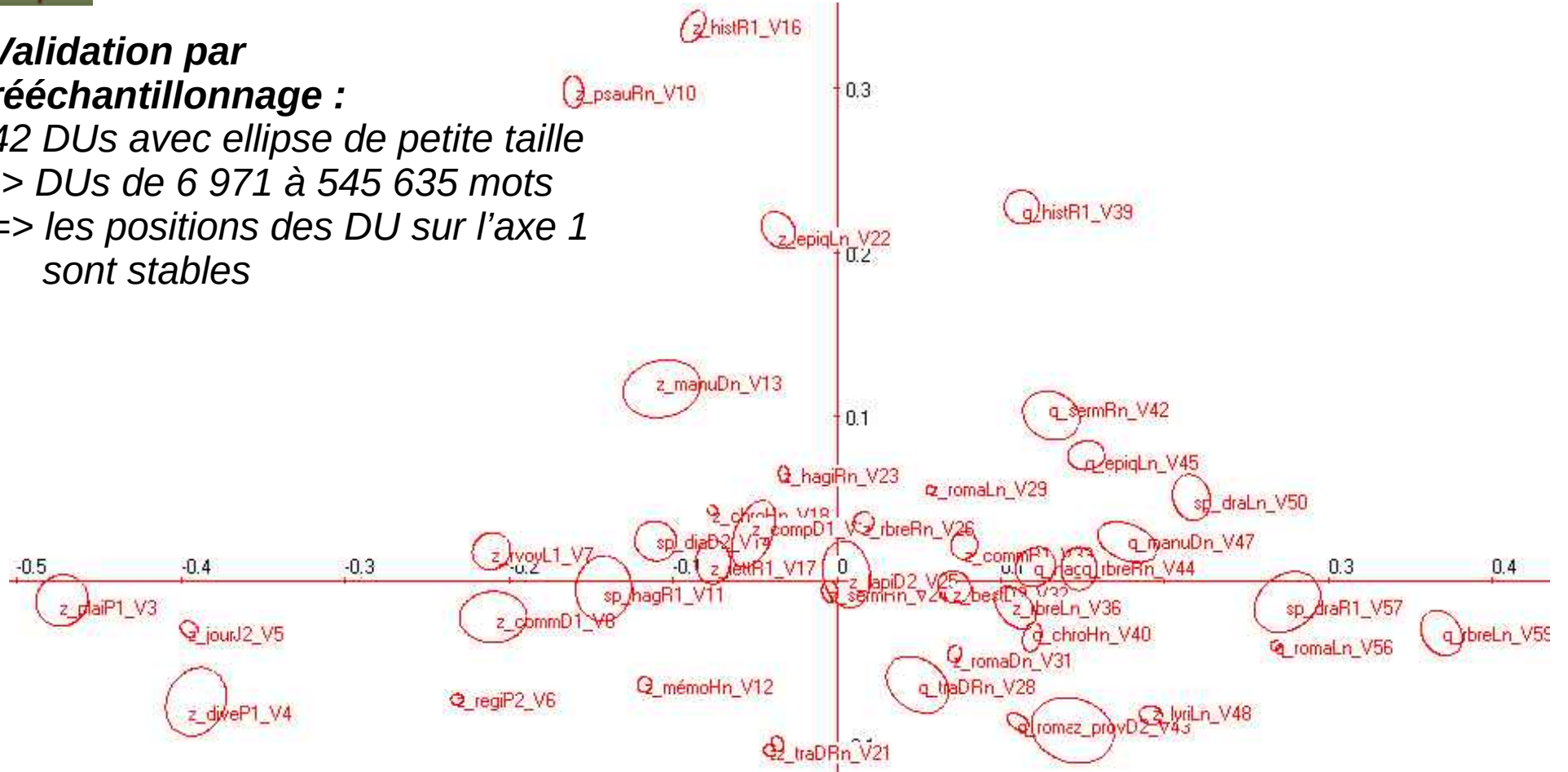


# BFM & DD : AFC 59 DU x 33 POS, petites ellipses



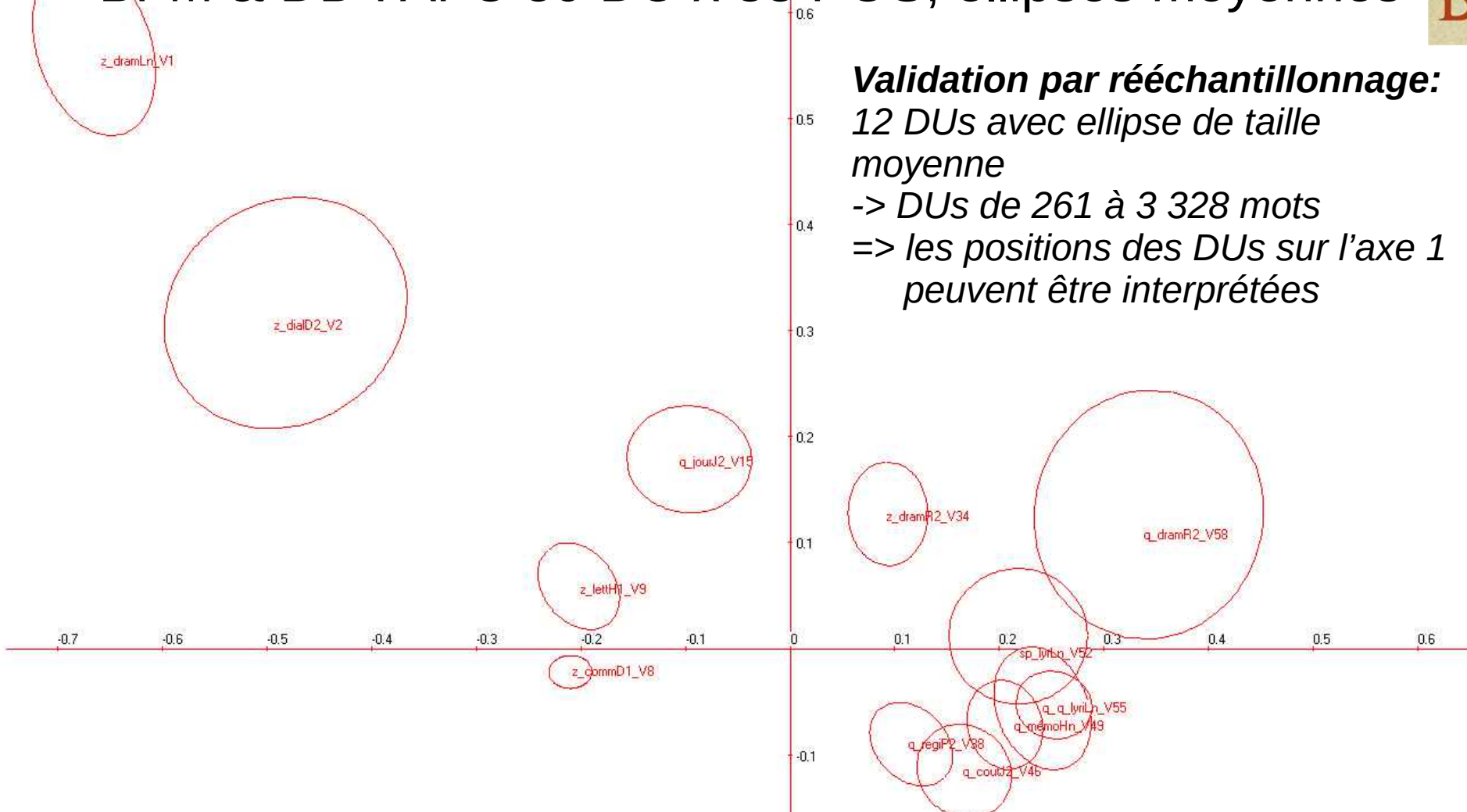
## Validation par rééchantillonnage :

42 DUs avec ellipse de petite taille  
 -> DUs de 6 971 à 545 635 mots  
 => les positions des DU sur l'axe 1 sont stables



# BFM & DD : AFC 59 DU x 33 POS, ellipses moyennes

BFM





# BFM & DD : AFC 59 DU x 33 POS, grandes ellipses

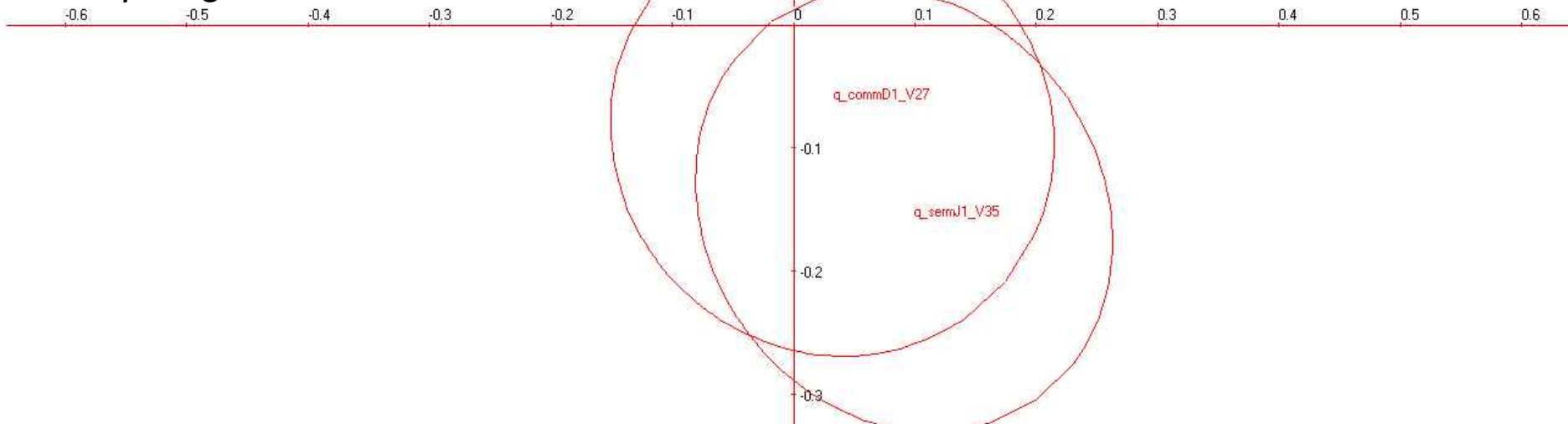


## **Validation par rééchantillonnage:**

*Seules 5 DUs avec ellipse de grande taille (2 ne pouvant être dessinées par le logiciel)*

*-> DUs de 11 à 122 mots*

*=> le champ des interprétations est trop large*



# BFM & DD : retour en détail sur la méthode

## Comment a-t-on géré le « bruit » ?



- Étiquetage morphosyntaxique automatique (étiquettes POS)
  - Ajuster la représentation des données en fonction du type d'analyse menée, avec deux critères : fiabilité et pertinence
- L'apport des outils statistiques
  - AFC et ordre des dimensions : fait pour décanter le bruit ?
  - La « loi des grands nombres » ?
  - Évaluer et gérer le manque de données avec les ellipses de confiance
  - **Adopter les bons repères de lecture : les aides à l'interprétation de l'AFC**



# Apport des outils statistiques (4) : visualisation plus intuitive

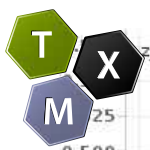
- Le tableau des « aides à l'interprétation » permet d'évaluer
  - Ce qui contribue de façon importante au contraste représenté (contribution à l'axe), et
  - Dans quelle mesure la projection dans le plan (2D) a déformé (aplatis) la position réelle du point dans l'espace complet (32 dimensions) (cosinus carré au plan)
- On focalise l'attention sur les points pertinents, pour leur importance ou/et la représentativité de leur position.



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



Colonnes	Q12	Q13	Q23	Mass	Dist	Cont1	Cos <sup>2</sup> 1	Cont2	Cos <sup>2</sup> 2	Cont3	Cos <sup>2</sup> 3	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_rbreffsL	,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

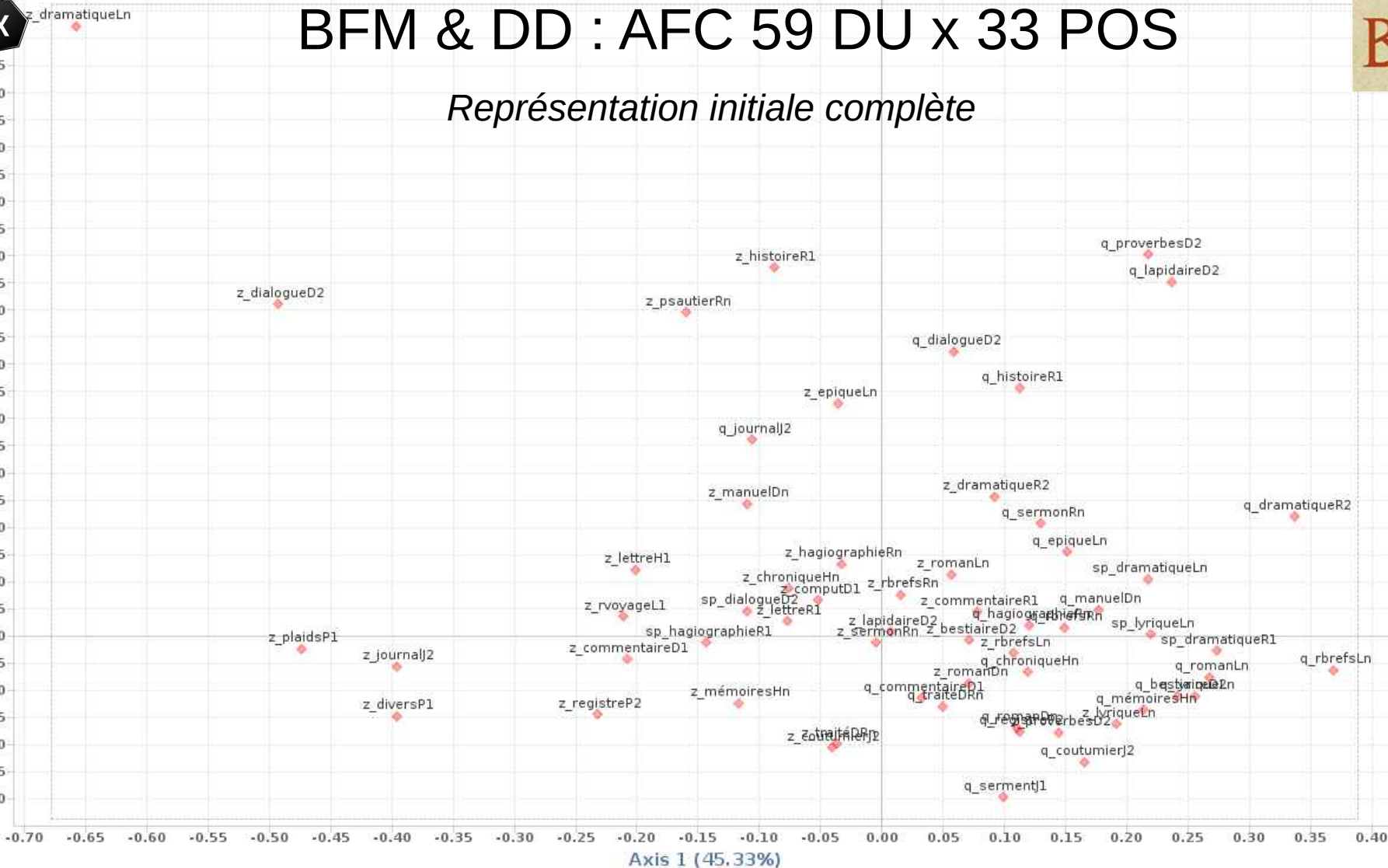


# BFM & DD : AFC 59 DU x 33 POS



*Représentation initiale complète*

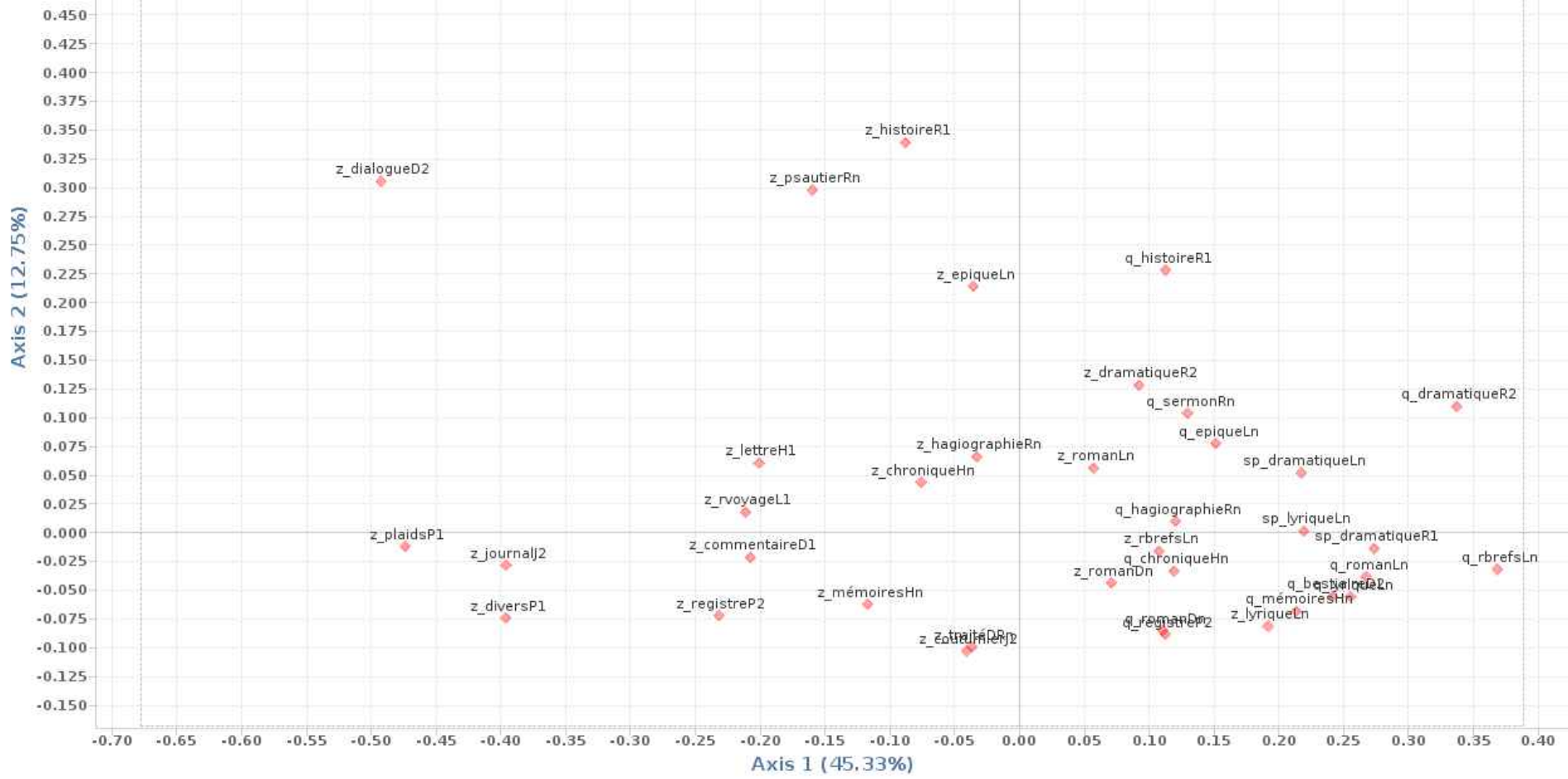
Axis 2 (12.75%)





# BFM & DD : AFC 59 DU x 33 POS

Filtre:  $(\text{Cos}^2(1x2) < 0.3)$  &  $(\text{ctrb1} < 2 \%)$  &  $(\text{ctrb2} < 2 \%)$







# BFM & DD : gradient d'oralité



Rank	DU	c1	Qtrb1	Cos²1	Q12	Mass	Word Nb	Ellipse size
1	z_dramatiqueLn	-0,33	0,16	<b>0,31</b>	0,53	0,01	334	medium
2	z_dialogueD2	-0,49	0,07	<b>0,38</b>	0,52	0,01	261	medium
3	z_plaidsP1	-0,47	<b>2,46</b>	<b>0,72</b>	0,72	0,28	10153	small
4	z_diversP1	-0,40	<b>1,53</b>	<b>0,70</b>	0,72	0,25	9025	small
5	z_journalJ2	-0,40	<b>21,26</b>	<b>0,86</b>	0,86	<b>3,51</b>	125563	small
6	z_registreP2	-0,23	<b>18,16</b>	<b>0,69</b>	0,76	<b>8,73</b>	312528	small
7	z_rvoyageL1	-0,21	<b>1,13</b>	<b>0,36</b>	0,36	0,66	23469	small
8	z_commentaireD1	-0,21	0,35	<b>0,70</b>	0,71	0,21	7573	small
9	z_lettreH1	-0,20	0,10	<b>0,39</b>	0,42	0,06	2263	medium
10	z_psautierRn	-0,16	<b>1,27</b>	0,10	0,46	<b>1,29</b>	46058	small
11	sp_hagiographieRn	-0,14	0,35	0,14	0,14	0,44	15732	small
12	z_mémoiresHn	-0,12	<b>2,86</b>	<b>0,31</b>	0,40	<b>5,40</b>	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	<b>1,50</b>	53537	small
17	z_lettreR1	-0,08	0,19	0,13	0,13	0,84	30003	small
18	z_chroniqueHn	-0,08	<b>2,06</b>	<b>0,32</b>	0,42	<b>9,24</b>	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	<b>4,37</b>	156475	small
21	z_traitéDRn	-0,04	0,25	0,03	0,23	<b>4,91</b>	175799	small
22	z_epiqueLn	-0,04	0,06	0,02	0,58	<b>1,22</b>	43785	small
23	z_hagiographieRn	-0,03	0,15	0,08	0,42	<b>3,44</b>	123094	small
24	z_sermonRn	0,00	0,00	0,00	0,00	<b>3,26</b>	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	<b>2,29</b>	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	<b>1,91</b>	0,20	0,41	<b>15,25</b>	545635	small
30	q_dialogueD2	0,06	0,00	0,01	0,15	0,00	91	large
31	z_romanDn	0,07	0,81	<b>0,36</b>	0,50	<b>4,20</b>	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	<b>1,26</b>	45201	small
34	z_dramatiqueR2	0,09	0,02	0,10	0,31	0,06	2319	medium

[Orality - -]

[Orality +]

35	q_sermentJ1	0,10	0,00	0,05	0,16	0,00	101	large
36	z_rbrefsLn	0,11	0,37	<b>0,34</b>	0,35	0,83	29696	small
37	q_romanDn	0,11	<b>1,20</b>	<b>0,45</b>	0,72	<b>2,56</b>	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	<b>0,51</b>	0,55	<b>1,72</b>	61578	small
41	q_hagiographieRn	0,12	0,54	<b>0,48</b>	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	<b>0,36</b>	0,46	<b>1,00</b>	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	<b>3,54</b>	<b>0,54</b>	0,64	<b>2,49</b>	89262	small
49	q_mémoiresHn	0,21	0,11	<b>0,42</b>	0,46	0,06	2150	medium
50	sp_dramatiqueLn	0,22	<b>1,09</b>	<b>0,62</b>	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	<b>0,49</b>	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	<b>0,64</b>	0,67	0,05	1664	medium
55	q_lyriqueLn	0,26	0,20	<b>0,62</b>	0,65	0,08	2828	medium
56	q_romanLn	0,27	<b>28,80</b>	<b>0,91</b>	0,92	<b>10,44</b>	373570	small
57	sp_dramatiqueR1	0,27	0,92	<b>0,81</b>	0,81	0,32	11435	small
58	q_dramatiqueR2	0,34	0,05	<b>0,37</b>	0,41	0,01	409	medium
59	q_rbrefsLn	0,37	<b>2,37</b>	<b>0,78</b>	0,78	0,45	16115	small

[Orality -]

[Orality ++]

# Plan

- Introduction
- **Contexte 1 : Linguistique diachronique**
  - travailler avec un étiquetage morphosyntaxique automatique non exempt d'erreurs
  - l'analyse factorielle des correspondances, un outil justement fait pour décanter les données ?
- **Contexte 2 : Analyse historique d'archives audiovisuelles**
  - travailler avec une transcription automatique non exempte d'erreurs
  - lancer des analyses focalisées sur les mentions de plan sans les interférences avec les autres mots, par annotation et projection
- Éliminer le bruit en corrigeant ? L'annotation en question
- Apprivoiser le bruit



# Le projet ANR Antract



- Étude des *Actualités françaises* (1945-1968)
  - Actualités filmées diffusées dans les cinémas
  - Hebdomadaires, durée d'environ 10 mn
  - 1261 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





# Corpus TXM AF-NOTICES



Index: <item\_type="DEL">[]+</item>: word

Query: :m\_type="DEL">[]+</item> Properties: word Edit

word	Frequency
France	6753
Paris	3304
Etats Unis	880
<b>Belgique</b>	<b>773</b>
Algérie	714
Grande Bretagne	499

1 -100 / 3264 t 34404, v 3264, fmin 1, fmax 6753

AFNOTICES <item\_type="DEL">[word="Belgi...]

Query: m\_type="DEL">[word="Belgique"]</item>

ref	Left context	Pivot	Right context
1946-04-25, AFE04011926	EAU A LESSINES	Belgique	Lessines LE " SAI
1946-04-25, AFE04011922	me âgé, tête nue	Belgique	Zeebrugge Flandre
1946-05-02, AFE85001458	département Laon	Belgique	Bruxelles Pêche s
1946-05-02, AFE85001460	<b>i Leemput, Marcel</b>	<b>Belgique</b>	<b>Bruxelles Demi fin</b>
1946-05-09, AFE04011953	ondiale résistance	Belgique	Bruxelles LE 1er M
1946-05-09, AFE04011955	AI A BRUXELLES	Belgique	Bruxelles PARTIS

1 -100 / 773

1946-373

RUBRIQUE : LE SPORT

- Genre : Presse filmée ;
- Durée : 00:00:37
- Langue VO / VE :
- Nature de production : Production propre
- Producteurs (Aff.) : Producteur - Les Actualités Françaises (LAF) - Paris - 1945;
- Thématique :

**TITRE PROPRE**

Le Champion du monde de billard

**RÉSUMÉ**

A Bruxelles, Marcel van Leemput, champion du monde de billard, fait une démonstration savante sur un billard de match.

Commentaire sur des images de Marcel van LEEMPUT effectuant différentes figures.

**SÉQUENCES**

- PP du ratelier de queues de billard
- Monsieur Marcel Van LEEMPUT jouant au billard
- PP d'un point au cadre

default 373 / 1157 1946

- Source = Base documentaire INA
- 10776 notices sujets de 1261 émissions
- 2,2 millions de mots
- Un corpus structuré (hors texte, listes de descripteurs typés, etc.)

# Corpus TXM AF-VOIX-OFF

AFE86003489 - 5 X

0:02:20 ▶ Cent dix sept bombardements avaient acquis au Havre le triste privilège d'être l'une des villes les plus sinistrées de France.

0:02:26 ▶ Devant ce spectacle, on avait été sur le point de renoncer à le reconstruire autour de son monument aux morts intact par miracle.

0:02:34 ▶ aujourd'hui Pourtant le Havre est ressuscité.

0:02:38 ▶ Les derniers baraquements disparaissent devant les immeubles d'allure moderne.

0:02:43 ▶ Un prodigieux effort de reconstruction a permis de rééditer en dix ans, la ville tout entière, est Par les gratte-ciel à l'architecte Auguste Perret.

0:02:53 ▶ Son nouvel aspect ne manque pas de grandeur.

0:03:03 ▶ Ainsi l'avenue foch bordée d'immeubles de sept étages est deux fois plus large que les Champs elysées.

0:03:14 ▶ Parallèlement à la ville le port du Havre est en pleine renaissance

Page 5 / 12 Texte AFE86003489

AF-VOIX-OFF-V5-2022-04-27/<[\_div\_seque... X

Requête `[_div_sequences="*.ruines.*" & word="*.constr.*|.nou.v.*"]`

date-de-diffusion, titre-propre	Contexte gauche	Pivot	Contexte droit
11/08/1955, La bombe d'Hiroshima	aujourd'hui La vie s'est réinstallée dans hiroshima	reconstruite	. Mais les traces de l'effroyable explosion restent marqu
17/11/1955, La reconstruction du Havre	été sur le point de renoncer à le	reconstruire	autour de son monument aux morts intact par miracle. a
17/11/1955, La reconstruction du Havre	d'allure moderne. Un prodigieux effort de	reconstruction	a permis de rééditer en dix ans, la ville tout entière

1 - 100 / 133

- Transcription automatique de la bande son des vidéos
- 1260 textes, 10683 sujets synchronisés
- 1,5 millions de mots
- Intégration de la description documentaire (notices) en métadonnées.



# Les (quasi)-doublons d'AF-VOIXOFF-V1



ref	Left context	Pivot	Right context
AFE86004649, S55, 0:06:31	j'un état martin etan il y a	foule	était de pour encourager les rescapés de ce tour d
AFE85003165, S8, 0:02:33	non quinzième étape martha état il y a	foule	des belges pour encourager les rescapés de ce tou
AFE86000584, S8, 0:00:20	de s'ouvrir à paris il y a	foule	depuis quelque temps au magique le saint des sair
AFE86004035, S3, 0:00:33	de s'ouvrir à paris il y a	foule	depuis quelques temps où magie club le saint des
AFE86000447, S11, 0:01:13	dimanche matin place louis quatorze il y a	foule	à l'église notre-dame des victoires riquier deux mi
AFE86004010, S51, 0:01:24	dimanche matin place louis quatorze il y a	foule	l'église notre-dame des victoires riquier deux milli
AFE85008648, S24, 0:00:30	le président de gaulle au milieu des acclamations	foule	généralement réservé jusqu'à buckingham palace
AFE86003718, S78, 0:08:36	le président de gaulle au milieu des acclamations	foule	généralement réservé jusqu'à buckingham palace
AFE86004447, S74, 0:08:18	taire c'était en dentelle débile une autre	foule	attendait le geste symbolique bon c'est oui et on r
AFE86003239, S40, 0:02:45	de devant l'hôtel de ville une autre	foule	attendait le geste symbolique oui et non et en rem
AFE86004608, S1, 0:00:06	découverte la terre dans la connaître d'autres	foule	et n'aura -t-on montre à travers paris centaines de
AFE85008338, S9, 0:00:05	il y avait	foule	à reims un jour de l'été mille neuf cents m le

**Rq.** : on a affaire ici aux premières transcriptions automatiques, avant que l'outil soit adapté/entraîné pour ces données d'archives. Il est mis en difficulté par la qualité ± dégradée de l'enregistrement, la façon de parler qui a beaucoup changé, l'évolution du vocabulaire, etc.

# L'amélioration des corpus :

## cas d'AF-VOIX-OFF

- 6 versions !
- Parmi les actions :
  - Transcription issue de la dernière version d'ASR (*automatic speech recognition*), après adaptation à ces données d'archive
  - Réorganisation du corpus pour éviter les doublons : passage d'une logique de **sujet** documentaire à une logique d'**émission** métropolitaine diffusée.
  - Données de contexte (métadonnées)
    - Mise en relation transcription ↔ vidéo ↔ sujet ↔ informations documentaires : synchronisation **automatique**, synchronisation **semi-automatique**, **traitement robuste** des petites incohérences de synchronisation (chevauchement/trou au niveau des raccords)
  - Retour à la vidéo (et nécessité d'un accès en ligne).



# Importance du retour à la vidéo

- [Pourquoi ?] 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)
- [Comment ?] 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.
- Pour en savoir plus :
  - 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>



# Possibilité d'accès distant aux vidéos



AF-VOIX-OFF-V5-2022-04-27/<[\_div\_seque... X

Requête 

date-de-diffusion, titre-propre	Contexte gauche	Pivot	Contexte droit
29/09/1949, Le probleme de la bomb	avec laquelle est du	construire	les immenses labor
29/09/1949, Le probleme de la bomb			paraissent et c'é
12/11/1953, Le président du conseil à	le bon voisinage. Le	nouveau	combats contre les
04/08/1953, Opération "Bretagne"	one du delta est un	nouveau	ons réfugiés d'E
15/01/1953, Ouverture du procès d'O	mais C'est aussi un	nouveau	énagements éco
12/11/1953, Le président du conseil à	le bon voisinage. Le	nouveau	ne de défense ét.
28/01/1954, Indochine : à Dien Bien P	upé. Qui donne une	nouvelle	ondissement de
11/11/1954, Les événements d'Algérie	urité amenaient de	nouveaux	nche de castille.
06/01/1955, Ce qu'a été, dans le mon	plement connu un	nouveau	reconquise (%h
11/08/1955, La bombe d'Hiroshima	llée dans hiroshima	reconstruite	ns un style qui lui
17/11/1955, La reconstruction du Havre	int de renoncer à le	reconstruire	
17/11/1955, La reconstruction du Havre	prodigieux effort de	reconstruction	
17/11/1955, La reconstruction du Havre	Auguste Perret. Son	nouvel	
23/05/1956, En Algérie, scènes de la p	ie. Avec l'arrivée de	nouveaux	
03/10/1956, Récolte du riz en Camarg	ale de cette culture	nouvelle	

1 -100 / 133

You must authenticate to access the media file

Connexion à okapi.ina.fr

Identifiant

Mot de passe

Cancel    Remise à zéro    **Se connecter**

AFE86003489 - 5 X

**sequences**

PANO sur un champ de ruines. Ouvriers travaillant à déblayer des ruines à l'aide de brouettes. DP de quartiers du Havre en 1944. Hotel de ville détruit DP du Monument aux Morts resté intact. PANO baraquements provisoires avec en arrière plan les immeubles modernes. DP ville en construction. DP avenue Foch et Hotel de ville. DP immeubles modernes. DP radars du port du Havre. DP port et de son trafic. Départ d'un paquebot.

**descripteurs-aff-lig**

DET: Seconde Guerre mondiale ; DET: après guerre ; DET: reconstruction ; DET: architecture ; DET: Perret, Auguste ; DET: logement ; DET: béton armé ; DEI: immeuble ; DEI: tour-architecture ; DEI: chantier de construction ; DEI: port ; DEI: monument aux morts ; DEI: mairie ; DEL: France ; DEL: Seine Maritime ; DEL: Le Havre ;

**generique-aff-lig**

0:02:17  C'était en mille neuf cent quarante quatre, ce qui restait du

0:02:20  Cent dix sept bombardements avaient acquis au Havre le triste privilège d'être l'une des villes les plus sinistrées de France.

0:02:26  Devant ce spectacle, on avait été sur le point de renoncer à le **reconstruire** autour de son monument aux morts intact par miracle.

0:02:34  aujourd'hui Pourtant le Havre est ressuscité.

0:02:38  Les derniers baraquements disparaissent devant les immeubles d'allure moderne.

0:02:43  Un prodigieux effort de **reconstruction** a permis de rééditer en dix ans, la ville tout entière, est Par les gratte-ciel à l'architecte Auguste Perret.

Page 5 / 12 Texte AFE86003489



# Retour à la vidéo

AFE86003489 X



0:02:20 ▶ Cent dix sept bombardements avaient acquis au Havre le triste privilège d'être l'une des villes les plus sinistrées de France.

0:02:26 ▶ Devant ce spectacle, on avait été sur le point de renoncer à le **reconstruire** autour de son monument aux morts intact par miracle.

0:02:34 ▶ aujourd'hui Pourtant le Havre est ressuscité.

0:02:38 ▶ Les derniers baraquements disparaissent devant les immeubles d'allure moderne.

0:02:43 ▶ Un prodigieux effort de **reconstruction** a permis de rééditer en dix ans, la ville tout entière, est Par les gratte-ciel à l'architecte Auguste Perret.

0:02:53 ▶ Son **nouvel** aspect ne manque pas de grandeur.

0:03:03 ▶ Ainsi l'avenue foch bordée d'immeubles de sept étages est deux fois plus large que les Champs elysées.

0:03:14 ▶ Parallèlement à la ville le port du Havre est en pleine renaissance

Page 5 / 12 Texte AFE86003489

AFE86003489 - 5 X

AF-VOIX-OFF-V5-2022-04-27/<[\_div\_seque... X

Requête

date-de-diffusion, titre-propre	Contexte gauche	Pivot	Contexte droit
11/08/1955, La bombe d'Hiroshima	aujourd'hui La vie s'est réinstallée dans hiroshima	reconstruite	. Mais les traces de l'effroyable explosion restent marqu
17/11/1955, La reconstruction du Havre	été sur le point de renoncer à le	reconstruire	autour de son monument aux morts intact par miracle. a
17/11/1955, La reconstruction du Havre	d'allure moderne. Un prodigieux effort de	reconstruction	a permis de rééditer en dix ans, la ville tout entière

1 - 100 / 133



# Fonctionnement du retour à la vidéo

- Lien hypertexte depuis :
  - La CONCORDANCE (pivot)
  - L'ÉDITION : au niveau du texte (journal), de la section (sujet), du tour de parole, du mot (fenêtre)
- Possibilité d'adapter la position de la fenêtre vidéo
  - Par glisser-déplacer (drag & drop)
  - Par réglage d'une préférence
- Curseurs pour naviguer à l'intérieur de la vidéo





# Généralisation : le retour au document source

TXM

P158 4 avril 2014 prTXManonymisev2.mp4 02

file:///home/mdecorde/TXM/corpora/p158/HTML/P158/default/P1

p. 02753 comment

E: 02754 la chaleur

p. 02755 alors la chaleur alors ça c' était non 02757 vous aviez pas fait l' hypothèse de la chaleur pour les raies puisque la cha la température pas la chaleur la température ça dit de quelle couleur globalement on va pouvoir disposer dans le spectre 02809 les chaudes en haut les plus froides en bas bien et le soleil vous voyez que le soleil en gros il est intermédiaire hein puisque dans le soleil il y a toutes les couleurs de l' arc en ciel , dans la **lumière émise** par le soleil 02824 bien donc ça c' est les spectres d' émission 02829 alors j' aimerais aussi que vous repérez et ça serait quand même pas mal maintenant quand vous avez un peu de recul dans quelles activités on a étudié des spectres d' émission 02841 dans l' activité

E: 02842 1

p. 02843 2 02845 alors dans l' activité 1 oui parce que on a observé bien sur les néons ah les néons ils observent de la lumière voilà vous mettez le néon ici 02852 file

REPRENDRE | 28:2175 | Répéter Taux: | Vol: |

P158:"lumière" [rposv="v+"] 02

ref	Contexte gauche	Pivot	Contexte droit
P158 4 avril 2014 prTXManonymisev2, E, 0:12:18	toutes les zones avancent là bas plus la	lumière est	blanche ben plus il y a de lumière ce spectre là ça
P158 4 avril 2014 prTXManonymisev2, P, 0:17:55	ou il peut être absorbé par raies une	lumière peut	être absorbée par raies ou par bandes d'accord et en fait
P158 4 avril 2014 prTXManonymisev2, P, 0:28:09	de l'arc en ciel, dans la	lumière émise	par le soleil bien donc ça c' est les spectres d'émission
P158 4 avril 2014 prTXManonymisev2, P, 0:33:37	elle aura l'un spectre de ce type là	lumière émise	par la tige de fer OK le fer chauffé à blanc c'

Accueil | OBLCUNEIF | Lexique /OBLCUNEIF | OBLCUNEIF:[word="ma"]

-2-

[ ]

adi tea-em-ka la a5-pur-am **ma**

Si-pi-ir l-i-im Sa ih-he-ruù

la l-mu-ru-nim

mu-ù a-na si-ip-ri-im

ga-am-ri-im

la us-ta-ar-du-ù

ù is-tu si-pi-ir l-i Sa l-na-an-na

sa-ab-la-ti

l-na he-re-e-em la-ag-dam-ru

[ ]

Handwritten annotations in red and blue on the OBLCUNEIF interface.



PHILEAS FOGG.

## RE I. DANS LEQUEL PHILEAS FOGG ET PASSEPARTOUT S'ACCEPTENT RÉCIPROQUEMENT L'UN COMME MAÎTRE, L'AUTRE COMME DOMESTIQUE

En l'année 1872, la maison portant le numéro 7 de Saville-row, Burlington Gardens - maison dans laquelle Sheridan mourut en 1814 -, était habitée par Phileas Fogg, esq., l'un des membres les plus singuliers et les plus remarqués du Reform-Club de Londres, bien qu'il semblât prendre à tâche de ne rien faire qui pût attirer l'attention.

### DANS LEQUEL PHILEAS FOGG ET PASSEPARTOUT S'ACCEPTENT RÉCIPROQUEMENT, L'UN COMME MAÎTRE, L'AUTRE COMME DOMESTIQUE.

En l'année 1872, la maison portant le numéro 7 de Saville-row, Burlington Gardens, — maison dans laquelle Sheridan mourut en 1814, — était habitée par Phileas Fogg, esq., l'un des membres les plus singuliers et les plus remarqués du Reform-Club de Londres, bien qu'il semblât prendre à tâche de ne rien faire qui pût attirer l'attention.

A l'un des plus grands orateurs qui honorent l'Angleterre, succédait donc ce Phileas Fogg, personnage énigmatique, dont on ne savait rien, sinon que

OBLCUNEIF - translit, ...

PDF Notice

Chercher | Reqlages

he	Pivot	Contexte droit
a at ta	ma	ERIN: è ù ERIN: e pi is tum a na si ip ri
a tù ù	ma	in ne ez bu ù GÁN SAHAR i na as sa ah
ur am	ma	si pi ir l-i-im Sa ih he ru ù la i
o ai bi	ma	um ma ba em mu ca bi me l ka lu è

79 | Nombre de lignes : 100 | Exporter



# Exemple d'Édition synoptique : le corpus GRAAL sur le portail BFM

BFM

- Corpus
- BFM2019
- BFMSS
- CORPT...
- GRAAL
- PALAF...
- PALAF...
- PALAF...

**[Galaad à Kamaalot]**

§ 1

[A la veille de la Pente-  
coste quant li compai-  
gnon de la table re-  
onde furent venu  
a Kamaalot et il o-  
rent oï le servise et  
l'en voloit metre les  
tables a heure de]

10 nonne . ' lors en[tra a cheval en la]<sup>[1]</sup> sale une mout bele  
damoisele, et fu venue si grant oïrre que bien le pot  
l'en veoir, car ses chevaux en fu encore toz suanz, et ele  
descent et vient devant le roi si le salue, et il dist que  
Diex la benëie. « Sire, fet ele, por Dieu dites moi se Lancelot  
15 est ceenz. - Oïl voir, fet li rois, veez le la. » Si li mostre, et  
ele va maintenant la ou il est, et li dist : « Lancelot je vos  
di de par le roi Pellés que vos avec moi venez iusqu'en  
cele forest. » Et li li demande a qui ele est. « Je sui, fait  
ele, a celui donc je vos paroil. - Et quel besoign, fet  
il, avez vos de moi ? - Ce verroiz vos bien (bien), fet ele.  
20 - De par Dieu, fet il, et g'irai volentiers. » Lors dist a un  
escuier qu'il mete la sele en son cheval, et li apor  
ses armes, et cil si fet tout maintenant. Et quant  
li rois et li autre qui ou palés estoient voient ce si lor  
25 en poise mout. Et neporquant quant il voient  
qu'il ne remaindroit il l'en lessent aler. Et la reine  
li dist : « Que est ce Lancelot ? Nos lairez vos a cest jor qui  
si est hauz ? - Dame, fet la damoisele, sachiez que  
vos le ravroiz demain ceenz ainz hore de disner.

1 Ordre des mots différent dans le ms. Z : 'entra en la salle a cheval'.

<160a>

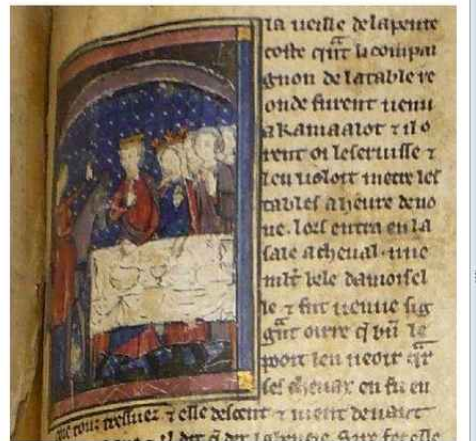
(Ici commence la version de la Queste del saint Graal donnée  
par le manuscrit K (Bibliothèque Municipale de Lyon, Palais des  
Arts n° 77), folios 160 recto à 224 verso. Tout le début du texte a  
été mutilé : la première grande lettre a été découpée, comme on  
le voit sur la reproduction du manuscrit, et quelques lignes du  
texte manquent, que nous donnons ici entre crochets, en bleu,  
d'après le manuscrit Z (Paris, BNF n. acq. fr. 1119, folio 138  
recto, colonne a) qui est un manuscrit proche de celui que nous  
éditons ici.)

§ 1

[A la ueille delapente  
coste qnt li compai  
gnon de latable re  
onde furent uenu  
5 a kamaalot ] il o  
rent oi leseruiffe ]  
l'en voloit metre lef  
tablef a heure de]

10 nōne . ' lozf en[tra acheual en la] fale une mout bele  
damoifele. ⁊ fuuene figzant oïrre que bien le pot  
l'en ueoir. car fef cheuau enfu encōre toz suanz. ⁊ ele  
descent et uient deuant le roi si le salue. et il dist que  
15 diex la benëie. Sire fet ele por dieu dites moi se lanç.  
est ceenz. Oil uoir fet liroif ueez le la. filimofre. ⁊  
ele ua maintenant la ouilest. ⁊ li dist lanç. ieuof  
di de par le roi pelles que uos avec moi uenez iusqen  
cele forest. ⁊ il li demande a qui ele est. Je sui fait  
20 ele a celui donc ie uos paroil. Et quel beoign fet  
il auez uof de moi . Ce uerroiz uof bien bien fet ele.  
Depardieu fet il. ⁊ girai uolentier. lozf dist a un  
escuier quil mete la sele en son cheual. ⁊ li apor  
ses armes. ⁊ cil si fet tout maintenant. Et quant

Fragment du ms. Z (BnF, n.a.fr. 1119, col. 138a) à la place d'un fragment manquant du ms. K



Retour au manuscrit K (Lyon, BM, P.A. 77, col. 160)



# Plan

- Introduction
- **Contexte 1 : Linguistique diachronique**
  - travailler avec un étiquetage morphosyntaxique automatique non exempt d'erreurs
  - l'analyse factorielle des correspondances, un outil justement fait pour décanter les données ?
- **Contexte 2 : Analyse historique d'archives audiovisuelles**
  - travailler avec une transcription automatique non exempte d'erreurs
  - lancer des analyses focalisées sur les mentions de plan sans les interférences avec les autres mots, par annotation et projection
- Éliminer le bruit en corrigeant ? L'annotation en question
- Apprivoiser le bruit



# Un thème de recherche : la grammaire cinématographique

- Contexte : thèse de Franck Mazuet
  - *Le cinéma de l'événement. Histoire de la société de presse filmée Les Actualités Françaises (1945-1969)*. Université Paris 1 Panthéon Sorbonne, 19 avril 2023.
- 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





# 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 "**

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

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

X [H] [L] 5 / 5 [R] [H] [H] [L] AFE86004168 [R] [H]



# Grammaire cinématographique : pistes pour travailler avec TXM



- Difficultés (~ bruits ?)
  - Variabilité des désignations
    - Description libre
    - Pratiques  $\pm$  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



# Annotation automatique par *ANTRACT* requêtes CQL

- ex. : 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

- Un tableau (xlsx/ods) avec
  - 19 valeurs de plan + Divers plans + Hors-sujet
  - 60 requêtes
    - Une même valeur de plan peut avoir plusieurs requêtes
    - 1 ligne = 1 requête pour 1 valeur de plan
- 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-V3-2021-10-11/<[\_div\_id="AF... AFE86004168 - 5

**TITRE PROPRE**

La toumée 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

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

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

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

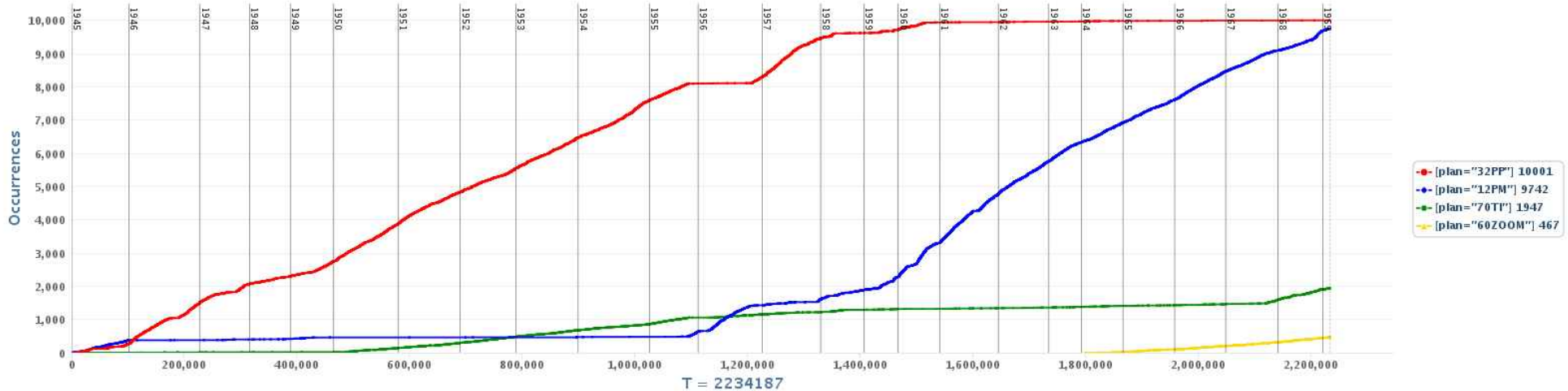
t93553, v21, fmin 10, fma: Page 5 / 5 Text AFE86004168 1247 / 1: Text AFE86004168

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...



# Exemple d'analyse sur le corpus annoté *INTRACT*



- **32PP** (Plan porté, 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



The screenshot shows a software interface with several windows displaying word frequency data. The main window on the left shows a list of words and their frequencies. The right window shows a similar list with a 'Units' column. The bottom window is a console window showing search results for specific sequences.

**Word Frequency Tables (Left and Middle Windows):**

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

**Word Frequency Tables (Right Windows):**

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

**Units and Frequency Table (Far Right Window):**

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
73FLOU	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 Output:**

```

System output
68 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.
68 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 (1<sup>ère</sup> 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)
- Un corpus « réécrit » pour l'analyse : normalisation des variations, neutralisation de l'entour gênant
- Pour en savoir plus :
  - 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>

# Plan

- Introduction
- **Contexte 1 : Linguistique diachronique**
  - travailler avec un étiquetage morphosyntaxique automatique non exempt d'erreurs
  - l'analyse factorielle des correspondances, un outil justement fait pour décanter les données ?
- **Contexte 2 : Analyse historique d'archives audiovisuelles**
  - travailler avec une transcription automatique non exempte d'erreurs
  - lancer des analyses focalisées sur les mentions de plan sans les interférences avec les autres mots, par annotation et projection
- **Éliminer le bruit en corrigeant ? L'annotation en question**
- Apprivoiser le bruit





# Corriger ? L'annotation en question(s)

- *Workflow* : dynamique du corpus et capitalisation
- Données originales et traçabilité
- Conception globale
- Équilibre annotation / analyse

# Corriger ? L'annotation en question(s)

- *Workflow* : dynamique du corpus et capitalisation
  - corpus final / sources (ex. nouvelle édition, enrichissement, ajout de textes)
  - prise en compte par les outils (entraînement, référence)
  - manuel ("en extension") / calculé ("en intensification")
  - travail collectif : partage et collaboration
- Données originales et traçabilité
- Conception globale
- Équilibre annotation / analyse

# Corriger ? L'annotation en question(s)

- *Workflow* : dynamique du corpus et capitalisation
- Données originales et traçabilité
  - ajouter vs remplacer
  - travailler au plus près des données
  - multiplicité de points de vue
- Conception globale
- Équilibre annotation / analyse

# Corriger ? L'annotation en question(s)

- *Workflow* : dynamique du corpus et capitalisation
- Données originales et traçabilité
- Conception globale
  - cohérence, efficacité
  - risque que la facilité technique d'ajout d'annotation conduise à les multiplier sans profit
- Équilibre annotation / analyse

# Corriger ? L'annotation en question(s)

- *Workflow* : dynamique du corpus et capitalisation
- Données originales et traçabilité
- Conception globale
- Équilibre annotation / analyse
  - ne pas attendre des données parfaites pour commencer l'analyse mais procéder par étapes (en connaissant et gérant les limites des données effectives)
  - s'interroger sur la nécessité de certains préalables, savoir dans quelles analyses on utilisera quel codage (ex. : entités nommées, analyse de sentiments, chaînes de référence)
  - possibilité de mener une analyse fine en restant très proche des données originales, sans l'intermédiaire d'unités prédéfinies ?
  - Si le global détermine le local (cf. François Rastier), cercle herméneutique analyse/édition en commençant par l'analyse ?

# Plan

- Introduction
- **Contexte 1 : Linguistique diachronique**
  - travailler avec un étiquetage morphosyntaxique automatique non exempt d'erreurs
  - l'analyse factorielle des correspondances, un outil justement fait pour décanter les données ?
- **Contexte 2 : Analyse historique d'archives audiovisuelles**
  - travailler avec une transcription automatique non exempte d'erreurs
  - lancer des analyses focalisées sur les mentions de plan sans les interférences avec les autres mots, par annotation et projection
- Éliminer le bruit en corrigeant ? L'annotation en question
- **Apprivoiser le bruit**





# Apprivoiser le bruit

- Bruit plutôt que silence : cf. stratégie générale de construction de requête
- Le critère d'**interprétabilité** du corpus
  - Par delà les critères de constitution de corpus et règles formelles (cohérence, pertinence, représentativité, homogénéité, etc.)
  - Faire un usage juste et approprié du corpus par la bonne connaissance de son contenu et de ses limites
  - Ni corpus parfait, idéal, ni données au kilomètre...
  - cf. Bénédicte Pincemin, 2012, « **Hétérogénéité des corpus et textométrie** », *Langages*, 187, 13-26.

# Apprivoiser le bruit

- Approche proposée :
  - Connaître (décrire, contextualiser) ses données
  - Garder une attention critique (contrôler)
  - Ajuster/adapter les traitements (capacité à rester au plus près des données originales)
  - Contextualiser l'interprétation

*Merci de votre attention !*

# Annexes et compléments

- Caractéristiques de la Base de français médiéval
- Les différents types d'annotation dans le logiciel TXM

# 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>



# Annotation par requêtes : tableau définissant les annotations à ajouter



cql\_valeurs\_de\_plan\_211007a.xlsx - LibreOffice Calc

Fichier Édition Affichage Insertion Format Styles Feuille Données Outils Fenêtre Aide

Arial 10 G I S A % 7,4 0,0 0,0

A1 fx Σ = code

	A	B	C	D	E
1	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 plan	(((word="-(premier 1er avant second deuxième 2ème arrière dernier)"%c)   ((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?"%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 britannique, plan cadastral, plan mural, plan(s) incliné(s)...	(((resume sequences) & word="plans?" %c) [word="britannique algérien cadastra( lux) architectura( lux) mura( lux) 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.

Rechercher Tout rechercher Affichage mis en forme Respecter la casse

Feuille 1 sur 1 PageStyle\_Feuil1 Français (France) Moyenne: ; Somme: 0 85 %

# 3 types d'annotation dans TXM



<i>Type</i>	<i>Projets</i>	<i>Interface</i>	<i>Nature technique</i>	<i>Intérêts</i>	<i>Limites</i>
CQP sur Mot	PaLaFra, BFM	Concordance	Propriétés lexicales	Simple à comprendre, peut répondre à beaucoup de besoins	Unités définies par la tokenisation
CQP sur Séquence de mots	BHE / SyMoGIH	Concordance	Structures et propriété	Délimitation de l'unité	Une seule annotation par structure (ref) ; complexité de la gestion des chevauchements
URS (Unité-Relation-Schéma)	Democrat , DTH	Edition, puis Concordance	Annotation déportée	Délimitation de l'unité, pas de contraintes structurelles, et modèle d'annotation structuré	Pas d'exploitation directe en CQP donc macros dédiées (ou projection vers CQP)

L'annotation par requêtes pour la grammaire cinématographique dans le projet Antract est une annotation de type « CQP sur Mot ».