



**HAL**  
open science

## Les données dans la recherche en sciences humaines et sociales

Bénédicte Pincemin

► **To cite this version:**

Bénédicte Pincemin. Les données dans la recherche en sciences humaines et sociales: Un partage d'expérience. XVIIIèmes Rencontres du Réseau international francophone de recherche en éducation et formation, Colloque scientifique " Abondance, pertinence, éthique : Questionnement sur les données en recherche, en enseignement-apprentissage, en formation et dans les politiques éducatives ", Réseau international francophone de recherche en éducation et formation; Université de Fribourg, Jul 2024, Fribourg, Suisse. hal-04676549

**HAL Id: hal-04676549**

**<https://hal.science/hal-04676549>**

Submitted on 23 Aug 2024

**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



*XVIII<sup>èmes</sup> Rencontres du Réseau international francophone de recherche en éducation et formation – Colloque scientifique*  
***Abondance, pertinence, éthique : Questionnement sur les données en recherche, en enseignement-apprentissage, en formation et dans les politiques éducatives***  
Fribourg, mercredi 3 juillet 2024

# Les données dans la recherche en sciences humaines et sociales : un partage d'expérience

Bénédicte PINCEMIN

CNRS, IHRIM UMR5317, ENS de Lyon



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

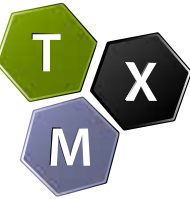


# Plan

1. Le rapport aux données dans la recherche : un contexte, une expérience
2. Illustration de l'approche textométrique sur 4 corpus
3. Données : objectives ou subjectives ?
4. (R)évolution scientifique : incidence globale de la place prise par les données sur le fonctionnement de la recherche

# Plan

1. **Le rapport aux données dans la recherche : un contexte, une expérience**
2. Illustration de l'approche textométrique sur 4 corpus
3. Données : objectives ou subjectives ?
4. (R)évolution scientifique : incidence globale de la place prise par les données sur le fonctionnement de la recherche

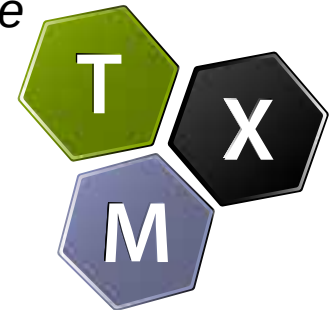


# Angle d'approche : Linguistique

- Textualité
  - « L'objet de la linguistique est le texte »
  - « Le global détermine le local »
  - Intertextualité : le texte fonctionne en corpus
- Sémantique
  - Distributionnelle : les contextes d'emploi sont déterminants
  - Interprétative : le sens est construit, une pluralité de lectures sont possibles
- cf. François Rastier (1987, 1994, 2001, 2011...)

# Sollicitation : Logiciel TXM

- Approche d'analyse des données textuelles née dans les années 1970-1980
  - École d'*Analyse des données* autour de Jean-Paul Benzécri (1973, 1981) : statistique exploratoire (
  - Application à la lexicologie, à l'étude des textes politiques, à la description de la langue à partir des textes littéraires du Trésor de la langue française (TLF)
  - Logiciels Lexico, SPAD-T, Hyperbase, Alceste, Weblex... puis 2<sup>e</sup> génération : Lexico 5 , DTM-Vic, Hyperbase Web, Lexico 5, IraMuTeQ, TXM, Trameur...
- Projet ANR Textométrie (2007-2010) : *Fédération des recherches et développements en textométrie autour de la création d'une plateforme logicielle ouverte*
  - Évolution des corpus : des textes numérisés enrichis, structurés
  - Open-source : gratuit mais aussi transparence, modularité
- Logiciel TXM :
  - Une interface utilisateur graphique
  - Multiplateforme : Windows, Mac OS, Linux ; et portail Web.



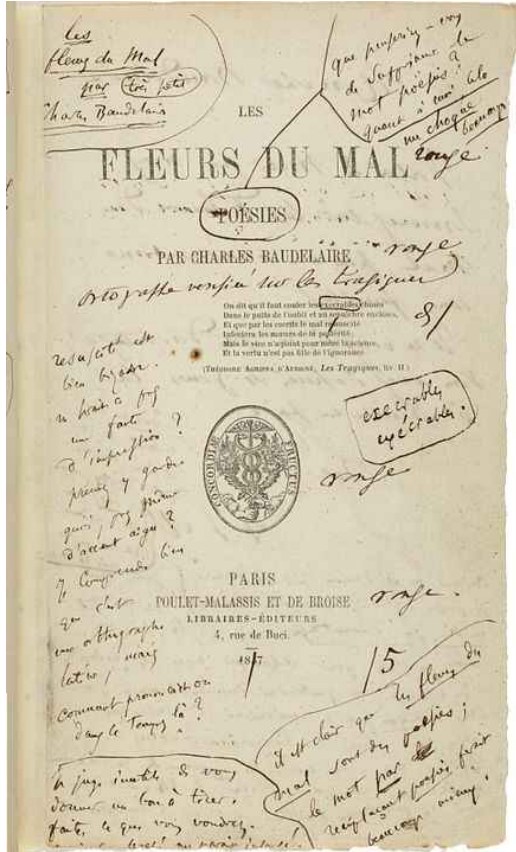
# Plan

1. Le rapport aux données dans la recherche : un contexte, une expérience
2. **Illustration de l'approche textométrique sur 4 corpus**

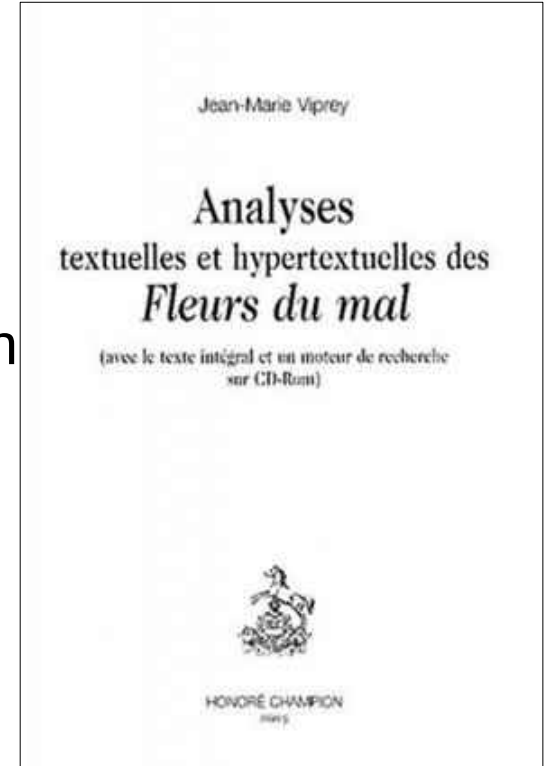


3. Données : objectives ou subjectives ?
4. (R)évolution scientifique : incidence globale de la place prise par les données sur le fonctionnement de la recherche

# Corpus FLEURS



- Le recueil poétique des *Fleurs du Mal* de Baudelaire (plusieurs éd. entre 1857 et 1868)
- Édition numérique préparée par Jean-Marie Viprey, univ. Besançon (thèse puis ouvrage en 2002)
- Corpus préparé pour TXM : téléchargeable depuis la [bibliothèque de fichiers pour TXM](#)





TXM

Fichier Outils Affichage Aide

Corpus

FLEURS: [pos="S."...] Requête: [pos="S."] Seuils: Fmin: 1 Fmax: 99999

word	Fréquence T=33233
cœur	110
yeux	92
âme	60
ciel	58
amour	46
œil	46
soleil	44
corps	41
air	36
soir	36
nuit	32
mer	30
beauté	29

FLEURS: [pos="J"]:lemme Requête: [pos="J"] Seuils: Fmin: 1 Fmax: 99999

lemme	Fréquence
grand	72
plein	63
vieux	51
doux	43
long	42
profond	37
autre	29
cher	26
éternel	25
lourd	23
charmant	21
étrange	20
pâle	20

FLEURS: [pos="V"]:le... Requête: [pos="V"] Seuils: Fmin: 1 Fmax: 99999

lemme	Fréquence
être	343
avoir	206
faire	124
voir	73
dire	55
pouvoir	55
vouloir	52
aimer	50
savoir	46
aller	43
prendre	43
venir	39
connaître	21

FLEURS: [pos="A"] Requête: [pos="A"] Seuils: Fmin: 1 Fmax: 99999

word	Fréquence
pas	60
toujours	36
jamais	33
bien	21
donc	19
parfois	15
très	14
encor	13
le plus	12
loin	12
trop	11
souvent	10
à travers	9

Ouverture du tableau d'index

**Vocabulaire dominant par catégorie grammaticale : relevés des noms communs, adjectifs qualificatifs, verbes et adverbes les plus fréquents dans les Fleurs du Mal.**

TXM

Fichier Outils Affichage Aide

Corpus

FLEURS: [pos="R" ...]

Requête: [pos="R" & lemme="noir"]

Seuils: Fmin: 1 Fmax: 99999

lemme	Fréquence T=33233
noir	61
blanc	17
bleu	14
rose	13
rouge	13
brun	10
jaune	5
vermeil	5
fauve	4
blond	2
violet	2
écarlate	1
roux	1

FLEURS:[lemme="vermeil"]

Afficher paramètres

poeme_titre_ε^	Contexte gauche	Pivot	Contexte droit
BÉNÉDICTION	Retrouve l'ambrosie et le nectar	vermeil	. Il joue avec le vent, ca
ENNEMI	mon jardin bien peu de fruits	vermeils	. Voilà que j'ai touché l'
AUBE	les débauchés l'aube blanche et	vermeille	Entre en société de l'id
PETITES VIEILLES	tombant Ensanglante le ciel de blessures	vermeilles	, Pensive, s'asseyait à l'
AME DU VIN	prison de verre et mes cires	vermeilles	, Un chant plein de lumi

FLEURS Edition - Page 11

file:///home/bpincemi/TXM/corpora/fleurs6/HTML/FLEURS/default

Ma jeunesse ne fut qu' un ténébreux orage,  
 Traversé ça et là par de brillants soleils ;  
 Le tonnerre et la pluie ont fait un tel ravage,  
 Qu' il reste en mon jardin bien peu de fruits **vermeils** .

Voilà que j' ai touché l' automne des idées,  
 Et qu' il faut employer la pelle et les râteaux  
 Pour rassembler à neuf les terres inondées,  
 Où l' eau creuse des trous grands comme des tombeaux.

Tri terminé.

TXM

Fichier Outils Affichage Aide

Corpus

FLEURS:[lemme="soleil"]

Requête : [lemme="soleil"] Pivot: word [Modifier] [Chercher]

Clés de tri : #1 Contexte d #2 Aucun #3 Aucun #4 Aucun Tri

1 - 59 / 59 Cacher paramètres

poeme_titre_abr	Contexte gauche	Pivot	Contexte droit
AUBE	yeux agrandis voitrige incessamment. Le	soleil	a noirci la riamme des bougies Ainsi, toujours vainqueur, ton
SONNET	marguerite ! Comme moi n'es -tu pas un	soleil	automnal, Ô ma si blanche, ô ma si froide Marguerite
COUVERCLE	re, Sous un climat de flamme ou sous un	soleil	blanc, Serviteur de Jésus, courtisan de Cythère, Mendiant ténébre
GRAVURE	sans horizon, Où gisent, aux lueurs d'un	soleil	blanc et terne, Les peuples de l'histoire ancienne et moderne
CRÉOLE	ervoir de larmes. Au pays parfumé que le	soleil	caresse, J'ai connu, sous un dais d'arbres tout
VOYAGE	la mer violette, La gloire des cités dans le	soleil	couchant, Allumaient dans nos cœurs une ardeur inquiète De plon
CHANT	phémère D'un glorieux automne ou d'un	soleil	couchant. Courte tâche ! La tombe attend ; elle est avide
POISON	Dans l'or de sa vapeur rouge, Comme un	soleil	couchant dans un ciel nébuleux. L'opium agrandit ce qui n'a
INVITATION	ir Qu'ils viennent du bout du monde. -Les	soleils	couchants Revêtent les champs, Les canaux, la ville entière,
SOLEIL	annes, abri des secrètes luxures, Quand le	soleil	cruel frappe à traits redoublés Sur la ville et les champs,
AME DU VIN	en flamme, De peine, de sueur et de	soleil	cuisant Pour engendrer ma vie et pour me donner l'âme ;
CHATIMENT	sa raison s'en alla. L'éclat de ce	soleil	d'un crêpe se voila ; Tout le chaos roula dans cette intelligence
CHANT	, labeur dur et forcé, Et, comme le	soleil	dans son enfer polaire, Mon cœur ne sera plus qu'un
PROFUNDIS	nde qui surpasse La froide cruauté de ce	soleil	de glace Et cette immense nuit semblable au vieux Chaos ; Je

Tri terminé.

**Concordance** : relevé des contextes d'emploi du mot « soleil » dans les Fleurs du Mal.

TXM

Fichier Outils Affichage Aide

Corpus

FLEURS: [lemme="soleil"] (10, 10)

Requête: [lemme="soleil"]

Propriétés des cooccurents: word Editer Seuils: Fréq ≥ 2

Contexte:  forme  structure d1v

de- 10 à- 0

Cooccurrent	Fréq	Cofréq	Indi	Distance moyenne
flamme	13	5	4	5,0
couchant	5	3	3	,0
s'	116	13	3	4,1
Par-delà	2	2	2	3,0
crêpe	2	2	2	3,5
noyé	2	2	2	2,0
fige	2	2	2	8,0
soirées	3	2	2	5,0
climat	3	2	2	4,0
six	3	2	2	6,5

Requête: ([lemme="soleil"] [\* [word="flamme"]]) | ([word="fla...]

Pivot: word Editer Chercher

Clés de tri: #1 Aucun #2 Aucun #3 Aucun #4

1 - 5 / 5

Cacher paramètres

eme_titre	Contexte gauche	Pivot	C
AMBEAU	t en plein jour ; le	soleil Rougit, mais n'éteint pas leur flamme	f
AMBEAU	z, Astres dont nul	soleil ne peut flétrir la flamme	!
BE	ncessamment. Le	soleil a noirci la flamme	d
ME DU VINT	t, sur la colline en	flamme, De peine, de sueur et de soleil	c
UVERCLE	sous un climat de	flamme ou sous un soleil	b

Tri terminé.

**Cooccurrences** : mots statistiquement sur-représentés au voisinage de « soleil » dans les Fleurs du Mal, et retour au texte.

# Master Géographie : Méthodes et outils



- Enseignants-chercheurs de l'[UMR 5600 Environnement Ville Société](#)
  - Emmanuelle Bonerandi-Richard †, Yves Le Lay, Luc Merchez
- Se donner des outils pour analyser méthodiquement aussi des données textuelles
- Enquête de terrain avec entretiens semi-directifs
  - Conception, recueil, transcription manuelle et analyse
  - Les étudiants travaillent en binôme
- Aspects de droits/propriété partiellement gérés
  - Anonymisation des locuteurs
  - Pas suffisant pour diffusion publique → objectif surtout pédagogique

# 2012 : Pêcheurs du lac Léman

- Relation des pêcheurs à leur environnement (naturel, économique, social...) et leur perception des évolutions
- 38 pêcheurs interviewés (sur les 46 pêcheurs côté français), 34 heures d'enregistrement audio, près de 500 000 mots.
- Le Lay Yves, Heiden Serge, Merchez Luc, Pincemin Bénédicte (2016) - « Retour de pêche. Le métier de pêcheur à travers le discours des professionnels français du lac Léman », in É. Comby, Y. Mosset et S. de Carrara (éds.), *Corpus de textes : composer, mesurer, interpréter*, Lyon : ENS éditions, collection « Sociétés, Espaces, Temps », p. 117-133.  
<https://halshs.archives-ouvertes.fr/halshs-01423605>

# Quelques observations sur le corpus LEMAN (Le Lay et al. 2016)



Syntagme (lemmes)	F	Concordancier du syntagme « beau métier »		
beau métier	7	moi, c'est le plus	beau métier	du monde pêcheur. Alors en fait, quand
métier dur	6	est vrai que c'est un	beau métier	que c'est une denrée qui est
autre métier	4	joli métier, on a un	beau métier	. Bon, un métier de travail, faut pas rêver.
même métier	4	on a quand même un	beau métier	, on a quand même des bonnes choses.
joli métier	3	même comme un	beau métier	quoi. Ouais il y a une bonne image, et ça
métier physique	3	bah voilà c'est un	beau métier	, que quand même quelques inquiétudes
ancien métier	2	métier, qui est le plus	beau métier	du monde, pour rien au monde je ne ferais
métier artisanal	2			
métier complet	2			
métier difficile	2			
métier traditionnel	2	Concordancier du syntagme « métier dur »		
métier très individualiste	2	une image que c'est un	métier dur	parce qu'effectivement l'hiver
petit métier	2	Non que c'est un	métier dur	, pas toujours évident. C'est bien
autres métiers	1	pêcheur professionnel, c'est un	métier dur	, difficile, parce que malgré tout
beaux métiers	1	donc ça reste malgré tout un	métier dur	, un métier physique. Euh, il y a
bon métier	1	,c'est malgré tout un	métier dur	. C'est un métier de passion si on
dernier métier	1	parce que c'est un	métier dur	quand même. Et puis quand on

Et aussi variantes : « tellement dur », « extrêmement dur », « beaucoup plus dur », « très très très très dur », en tout 11 occurrences chez sept pêcheurs différents.

# Quelques observations sur le corpus LEMAN (Le Lay et al. 2016)



## Le poids de la solitude

## Observations quantitatives et qualitatives, en contexte.

## Le tri des contextes à gauche fait ressortir la récurrence de « tout seul ».

Lemme	F	Extrait du concordancier du lemme « seul »	
seul	185	d'ici et puis je suis tranquille : je suis	seul   , je vois pas pourquoi j'irais à Lugrin ou à
individualiste	19	au ventre d'aller au milieu du lac, tout	seul   . Euh, je vous parle en tant que femme.
individualisme	1	pêcheur, seul, qui va faire tout, tout	seul   , euh, alors à ce moment là, oui, il
solitaire	1	Alors d'un côté je fais tout, tout	seul   , ou je suis tout seul et puis heu,
		sur le même bateau. Après j'allais tout	seul   avec mon bateau. Voilà oui. On va parler
Syntagme (lemmes)	F	voulez progressez, faut y aller tout	seul   , se faire son expérience euh, se jeter à
être tout seul	31	me suis démener pendant sept ans tout	seul   pour avoir tous ces clients là. Donc j'ai
travailler seul	7	encore euh, je continue d'apprendre tout	seul   , quoi, en fait euh, j'ai encore euh
aller tout seul	3	quelque chose qu'on peut apprendre tout	seul   automatiquement quand on arrive sur
apprendre tout seul	3	tout seul. Si, vous apprendrez tout	seul   . Si, ben je me suis chopé la grêle, euh
être seul	3	demi, une fois qu'il aura fait ça tout	seul   , euh, il peut se lancer y a pas de souci
faire tout seul	3	voilà. Et puis après j'ai commencé tout	seul   . Alors on va prendre une journée d'été
travailler tout seul	3	bien passé, puis après j'ai continué tout	seul   . Non je m'étais jamais intéressé à,
aller seul	2	Mais là oui, on s'est débrouillé tout	seul   . Si, même, vous prenez la maison
lancer tout seul	2	vieille camionnette, je me démerde tout	seul   , voilà hin, perso. C'est pas mon
rester seul	2	c'est mathématique ! Si t'es tout	seul   , et que tu pêches, tu pêches mal
		pas, c'est ton droit, tu es tout	seul   . C'est ton boulot. T'es, t'es ton
		Ouais. Que maintenant on est tout	seul   . Ouais. Les filets. Puisque c'est pus fin
		que bon on couvre, quand on est tout	seul   on est tout seul. On est tout seul on es
		employés, mais sinon, quand on est tout	seul   on peut pas. Et vous pensez que c'est un
		ça le problème. Puis quand on est tout	seul   , quand il faut faire tout seul ça



# Quelques observations sur le corpus LEMAN (Le Lay et al. 2016)



## Spécificités des verbes

selon la taille de l'établissement

Verbe	Gros établissements			Verbe	Moyens établissements			Verbe	Petits établissements		
	F	Fp	Sp		F	Fp	Sp		F	Fp	Sp
équiper	36	26	8,4	pêcher	718	463	7,0	apprendre	138	58	8,6
étudier	10	9	4,5	respecter	51	43	5,1	sauver	24	16	6,1
prévoir	23	15	4,2	concerner	19	19	5,0	espérer	29	17	5,3
surgeler	24	15	3,9	associer	22	21	4,5	lever	239	75	4,8
suivre	68	31	3,7	vendre	527	332	4,0	subir	11	9	4,7
placer	12	9	3,4	lier	31	27	3,8	travailler	416	118	4,7
HAPAX	580	180	3,0	savoir	887	539	3,7	former	53	24	4,6
confondre	5	5	3,0	parler	188	126	3,3	dépêcher	20	12	4,0
rééquilibrer	5	5	3,0	transformer	68	50	2,9	permettre	112	38	3,5
créer	33	17	3,0	payer	208	136	2,9	appréhender	7	6	3,4
diminuer	28	15	2,9	connaître	235	151	2,7	progresser	7	6	3,4
lancer	39	19	2,9	recupérer	46	35	2,6	enlever	48	20	3,3
adapter	34	17	2,8	marier	10	10	2,6	jeter	30	14	3,1
marquer	29	15	2,7	racheter	24	20	2,5	préférer	48	19	2,9
attaquer	14	9	2,6	attaquer	13	12	2,3	entourer	6	5	2,8
pousser	14	9	2,6	frayer	23	19	2,3	imprégner	4	4	2,8
développer	53	23	2,5	démarrer	29	23	2,2	essayer	153	46	2,8
embêter	22	12	2,5	perdre	62	44	2,2	préserver	11	7	2,7
donner	197	67	67	évaluer	12	11	2,2	gagner	134	41	2,7

/modernisation/

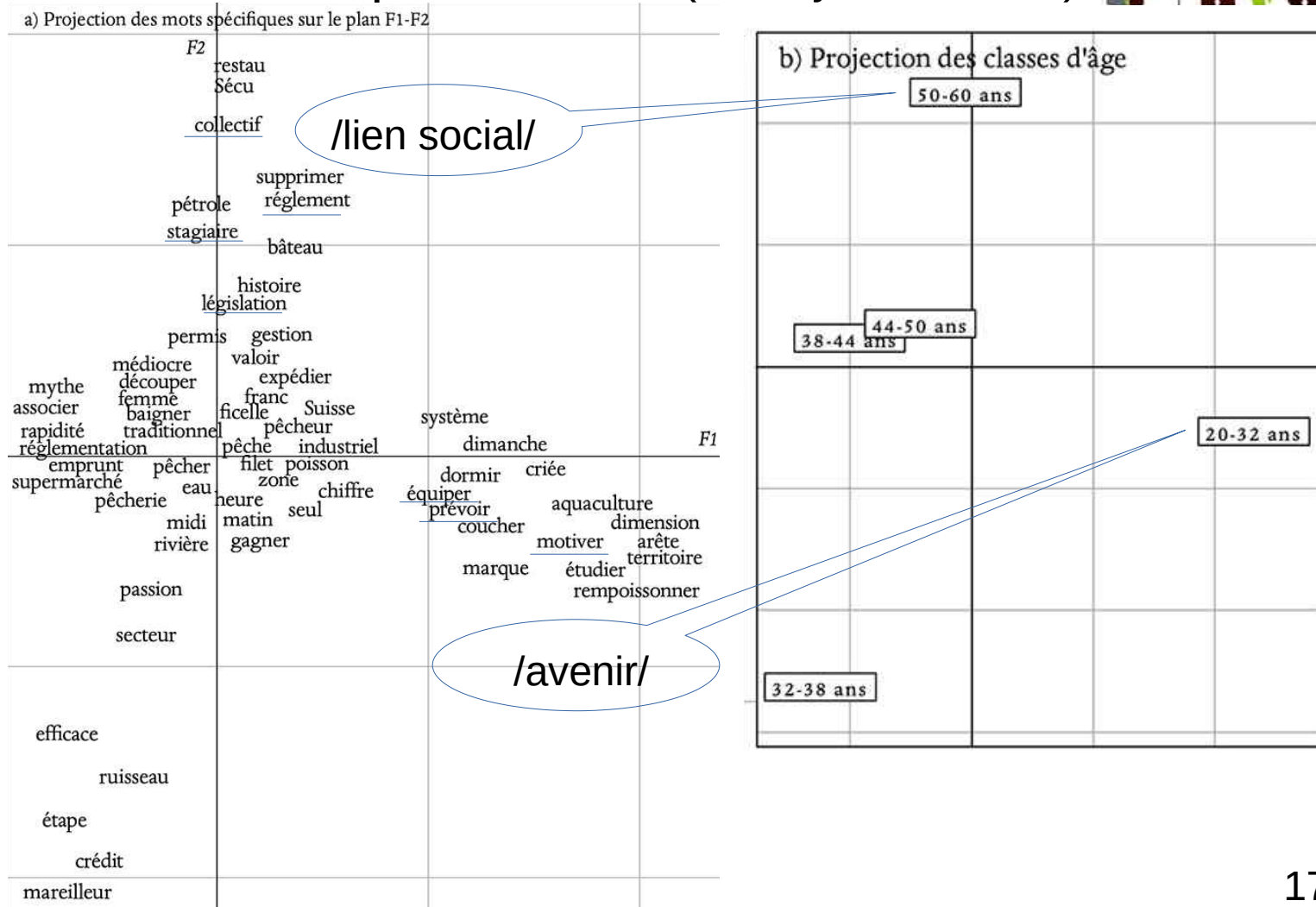
/métier/

/perspectives/

# Quelques observations sur le corpus LEMAN (Le Lay et al. 2016)

## Étude de l'influence de l'âge

- 1) Partition en 5 tranches d'âge
- 2) Calcul des spécificités sur les lemmes
- 3) Calcul d'une AFC sur le tableau croisant les 5 tranches d'âge et les lemmes les plus spécifiques



# Histoire : 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.



# 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**

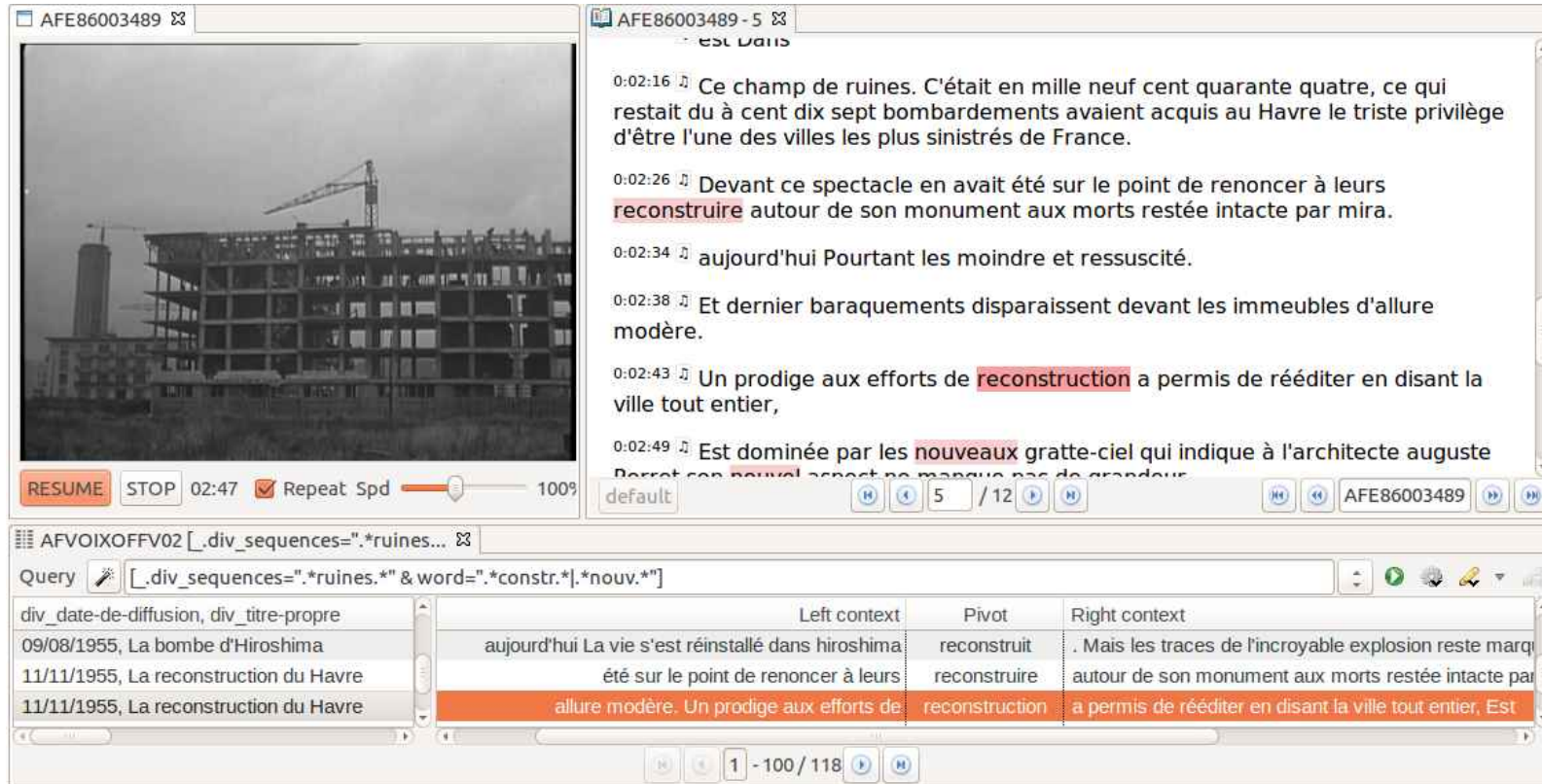
- PP du ratelier de queues de billard
- Monsieur Marcel Van LEEMPUT jouant au billard
- PP d'un point au cadre

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

AFE86003489 - 5

est Paris

0:02:16 Ce champ de ruines. C'était en mille neuf cent quarante quatre, ce qui restait du à cent dix sept bombardements avaient acquis au Havre le triste privilège d'être l'une des villes les plus sinistrés de France.

0:02:26 Devant ce spectacle en avait été sur le point de renoncer à leurs **reconstruire** autour de son monument aux morts restée intacte par mira.

0:02:34 aujourd'hui Pourtant les moindre et ressuscité.

0:02:38 Et dernier baraquements disparaissent devant les immeubles d'allure modère.

0:02:43 Un prodige aux efforts de **reconstruction** a permis de rééditer en disant la ville tout entier,

0:02:49 Est dominée par les **nouveaux** gratte-ciel qui indique à l'architecte auguste Berret son **nouvel** aspect ne manqua pas de grandeur.

RESUME STOP 02:47 Repeat Spd 100%

default 5 / 12 AFE86003489

AFVOIXOFFV02 [\_div\_sequences=".\*ruines...]

Query [\_div\_sequences=".\*ruines.\*" & word=".\*constr.\*.\*nou.v.\*"]

	Left context	Pivot	Right context
09/08/1955, La bombe d'Hiroshima	aujourd'hui La vie s'est réinstallé dans hiroshima	reconstruit	. Mais les traces de l'incroyable explosion reste marq
11/11/1955, La reconstruction du Havre	été sur le point de renoncer à leurs	reconstruire	autour de son monument aux morts restée intacte par
11/11/1955, La reconstruction du Havre	allure modère. Un prodige aux efforts de	reconstruction	a permis de rééditer en disant la ville tout entier, Est

1 - 100 / 118

- Source = Transcription automatique de la bande son
- 1260 émissions, au sein desquelles 10683 sujets synchronisés
- 1,5 millions de mots
- Retour à la vidéo en ligne depuis la concordance ou l'édition du texte

**Rq. :** Le retour au document source (vidéo, manuscrit, édition de référence...) généralise le retour au texte. Pour la transcription, cela permet un contrôle du traitement automatique, et un complément interprétatif face aux nécessaires réductions et choix opérés par la transcription.

Requête: [frlemma="foule"]

Requête: [frlemma="foule"] [frlemma="de/du"] Pivot: word Editer Chercher

Seuils: Fmin: 1 Fmax: 999

Clés de tri: #1 Aucun #2 Aucun #3 Aucun #4 Aucun Tri

1 - 100 / 738 Cacher paramètres

word	Fréquence
spectateurs	232
manifestants	63
gens	23
jeunes	22
pèlerins	16
curieux	15
voyageurs	14
étudiants	13
photographes	13
invités	12
fidèles	11
visiteurs	11
face	10
journalistes	8
la	8
piétons	8
grévistes	7
Hindous	7
ouvriers	7

ref	Contexte gauche	Pivot	Contexte droite
1945-02-02, AFE86002962	ant la mairie en présence d'une	foule de spectateurs	- Specta
1945-03-09, AFE86003007	vec l'officiant, en arrière plan et	foule de soldats	suivant
1945-04-13, AFE86003052	grande rue de FRANCFORT avec	foule d'Allemands	se press
1945-05-25, AFE86003109	le président HERRIOT saluant la	foule du balcon	de l'Hôl
1945-06-01, AFE86003115	ABETH et MARGARET saluant la	foule du Balcon	du Pala
1945-06-15, AFE86003135	ur et Madame BENES, saluant la	foule d'une	fenêtre
1945-07-27, AFE86003194	ne tête nue, visage triste dans le	foule de St	Briec-
1945-08-03, AFE86003201	olie petite couturière (2 plans) -	Foule de soldats	(4 plans
1945-08-10, AFE86003205	ce, côté des tribunes garnies de	foule de spectateurs	- PA de
1945-08-31, AFE86003233	ierre aux journalistes (2 plans) -	Foule de journalistes	à l'intér
1945-09-21, AFE86003251	plans) - [Vue générale] de	foule de mineurs	rassem
1945-09-28, AFE86003265	gués japonais quittent la table-	Foule de marins	améric
1945-10-05, AFE86003271	me-orient. ils sont salué par une	foule de chinois	, de Mal
1945-12-07, AFE86003360	, sortant de l'avion et saluant la	foule du haut	de la pa
1946-01-11, AFE85001264	ES du décret de nationalisation,	foules d'ouvriers	défilant
1946-01-11, AFE85001269	umentaire sur des images d'une	foule de Varsoviens	accueill
1946-01-25, AFE85001294	mbodgiens. - VG panoramique. -	Foule de jeunes	Cambo
1946-02-08, AFE85001312	et de la place Saint Pierre - La	foule des chrétiens	sur le p
1946-02-08, AFE85001312	Sedia Gestatoria " entouré de la	foule des fidèles	- PA des
1946-02-15, AFE85001310	esse ZUPPICHEN finit sur ski	Foule de spectateurs	assis su

- Nature de production : Production propre
- Producteurs (Aff.) : Producteur - Les Actualités Françaises (LAF) - Paris - 1945;
- Thématique :

### TITRE PROPRE

Visite du général de Gaulle à Boulogne Billancourt

### RÉSUMÉ

Un long reportage sur l'île de France et la Normandie dévastées par les bombardements et les récentes inondations est précédé de brèves images de la visite du général DE GAULLE à Boulogne-Billancourt.

### SÉQUENCES

Plusieurs plans de la visite de DE GAULLE à BOULOGNE BILLANCOURT

- Arrivée du cortège officiel devant la mairie en présence d'une **foule de spectateurs**

- Spectateurs sous la neige

- GP de DE GAULLE (de dos) prononçant un discours au balcon de l'hôtel de ville- Policiers sur le toit d'un immeuble voisin

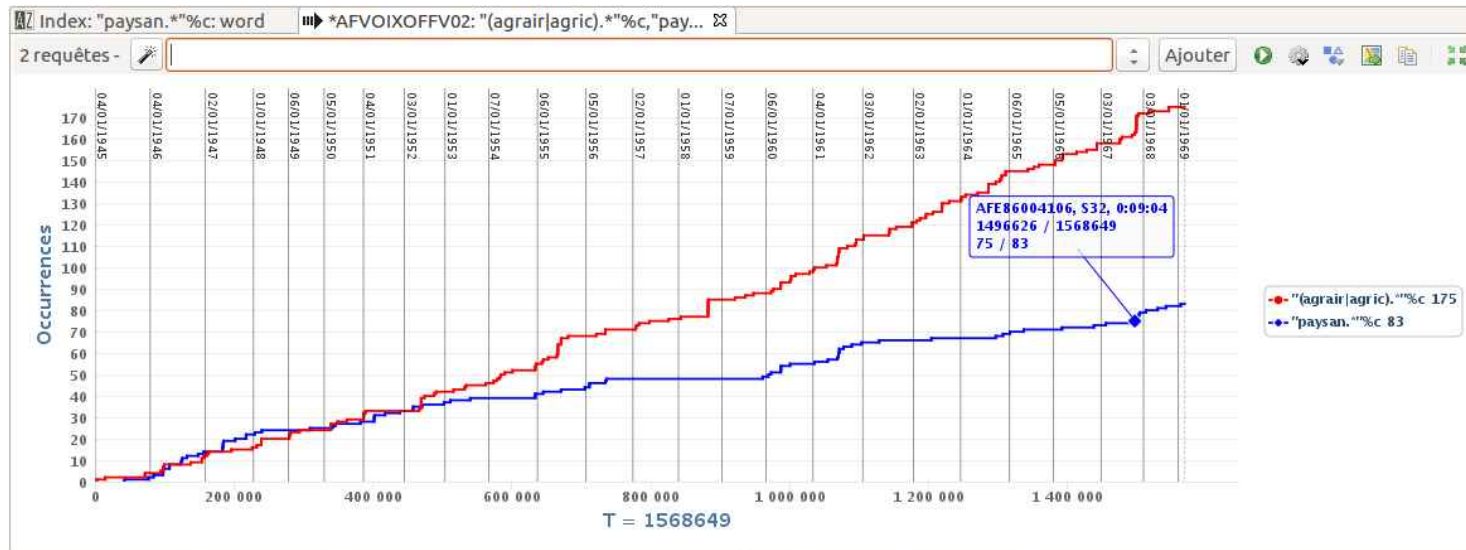
- 2 Vues aériennes de campagnes inondées

- Plusieurs plans de péniches bloquées par les glaces

- Route enneigée et verglacée

- Roues d'un camion patinant

*Inventaire chiffré des « foules de ... » dans le corpus des notices documentaires des Actualités françaises, et retour au contexte.*



AFVOIXOFFV02 "(agrair|agric).\*"%c 8

Requête "(agrair|agric).\*"%c

text_datedediffusion, text_id	Contexte gauche	Pivot	Contexte droit
30/05/1967, AFE86004091	marché commun	agricoles	en particulier a do
16/08/1967, AFE86004102	ux sous marin les	agriculteurs	de la mer des ver
12/09/1967, AFE86004106	un pays avant tout	agricole	et dans les villes a
17/10/1967, AFE86004111	a grande mutation	agricole	imposée par notre
17/10/1967, AFE86004111	ransformation de l'	agriculture	n'a pas abouti d'e
17/10/1967, AFE86004111	de la société civile	agricole	de la région de no
17/10/1967, AFE86004111	opéenne ouvre à l'	agriculture	française, un art c
17/10/1967, AFE86004111	tabilité du matériel	agricole	, deux exigeant. D
17/10/1967, AFE86004111	on irréversible de l'	agriculture	. Et elle vient de s
17/10/1967, AFE86004111	présent. Renover l'	agriculture	afin de la faire acc
17/10/1967, AFE86004111	a grande mutation	agricole	, C'De ça devrait e
30/10/1967, AFE86004113	itaires. La réforme	agraire	Pierre angulaire d
12/03/1968, AFE86004132	d'art du salon de l'	agriculture	après sa visite dé

AFVOIXOFFV02 "paysan.\*"%c 8

Requête "paysan.\*"%c

text_datedediffusion, text_id	Contexte gauche	Pivot	Contexte droit
05/05/1965, AFE86003983	s aérienne, euh Les	paysans	Nord vietnamiens e
09/03/1966, AFE86004027	faut apprendre aux	paysans	les techniques les p
09/11/1966, AFE86004062	ne Tous les mois de	paysans	qui lie les autres et
07/02/1967, AFE86004075	égard de la part des	paysans	soigneusement car
12/09/1967, AFE86004106	ertir Un peuple des	paysans	en ouvriers n'a pas
17/10/1967, AFE86004111	e les manifestations	paysannes	ont été les plus viol
17/10/1967, AFE86004111	s les manifestations	paysannes	à l'occasion d'une c
30/10/1967, AFE86004113	millions de familles	paysannes	en possession des
14/11/1967, AFE86004115	millions étaient des	paysans	Le trente cinquième
30/01/1968, AFE86004126	sous le costume du	paysan	indien qu'il ne quitte
29/05/1968, AFE86004143	ères manifestations	paysannes	. notamment En Br
28/08/1968, AFE86004156	il six trois cent mille	paysans	colombiens venus
22/01/1969, AFE86004177	ation des terres aux	paysans	seront remis sous

**Répartition au fil du temps des mots des familles de « agriculture » et « paysan » dans les transcriptions des Actualités françaises, et retour au contexte (modernisation vs tradition).**

Index: "(agrair|agri..."

Query Propertie

frlemma	Frequency
agricole	81
agriculture	62
agriculteur	20
agraire	8
Agricole	2
Agraire	1
Agricoles	1

t 175, v 7, fmin 1, fmax 81

AFVOIXOFFV02: "(agrair|agric).\*\*"%c (9, ...

Query "(agrair|agric).\*\*"%c

Parameters

Cooccurents properties: word Edit Thresholds: Fmin ≥

2 Cmin ≥ 2 Score ≥ 2,0

Context:  word  structure div Use the left context

from - 9 to - 0 and from 0 to 9

Cooccurrent	Frequency	CoFreq	▲ Score	Mean distance
agriculture	62	11	17	5.3
l'	27115	112	10	3.1
pays	1312	17	8	3.1
exploitations	11	4	8	2.0
marché	201	8	7	4.2
rendement	37	5	7	5.8
paysans	54	5	6	6.8
ministre	894	12	6	3.1
accroître	36	4	5	3.8
réforme	44	4	5	.0
machines	107	5	5	.0
collective	21	3	4	5.0
Tenguiz	3	2	4	3.5
internationalement	3	2	4	6.0
industrielle	68	4	4	3.2
potentiel	27	3	4	.7

t pivot 175, v cooc 78, t cooc 0, T corpus 1568649

AFVOIXOFFV02: "paysan.\*\*"%c (9, 9)

Query "paysan.\*\*"%c

Parameters

Cooccurents properties: word Edit Thresholds: Fmin ≥

2 Cmin ≥ 2 Score ≥ 2,0

Context:  word  structure div Use the left context

from - 9 to - 0 and from 0 to 9

Cooccurrent	Frequency	CoFreq	▲ Score	Mean distance
bretons	26	4	7	1.8
Bretagne	78	4	5	4.5
les	28383	56	5	2.7
ces	2637	12	4	3.4
manifestations	193	4	4	.0
mécontentement	11	2	4	1.5
des	20612	41	3	3.0
ouvriers	147	3	3	2.7
dirigeants	46	2	2	8.0
connaît	198	3	2	7.3
milieu	462	4	2	3.8
agricole	58	2	2	4.5
Déjà	60	2	2	4.0
agriculture	62	2	2	8.0
agitation	64	2	2	.0
dimanche	85	2	2	1.5

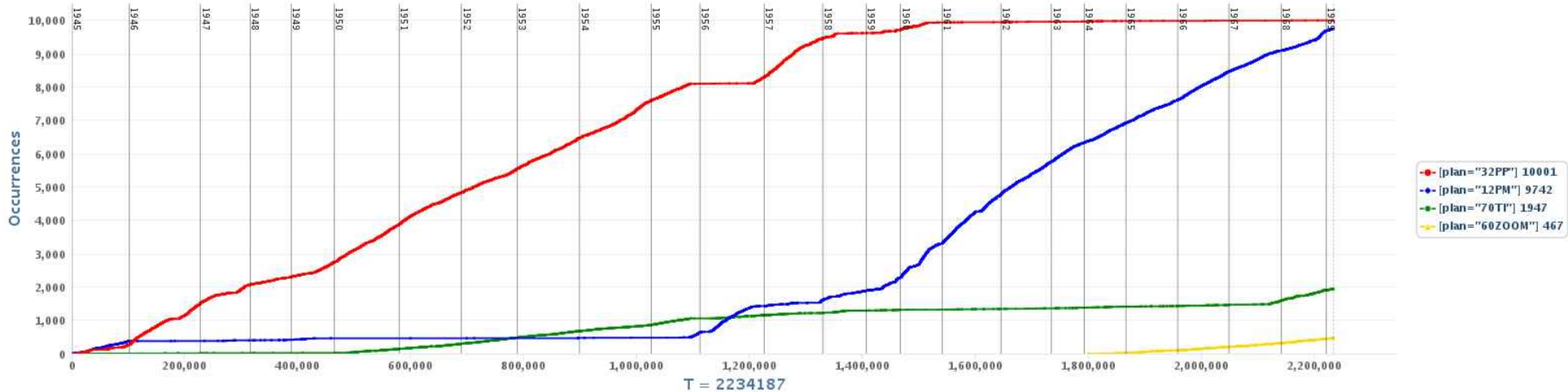
t pivot 83, v cooc 18, t cooc 0, T corpus 1568649

**Cooccurrence pour synthétiser les usages de deux familles de mots : dans les Actualités françaises, les mots « agriculteur » et « paysan » sont employés dans des contextes différents, le premier semble privilégié dans les descriptions économiques et conceptuelles, le second dans les descriptions plus politiques de manifestations avec des personnes.**





# Évolution dans le corpus des notices des Actualités françaises annoté en valeurs de plan



- **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

# Pour en savoir plus

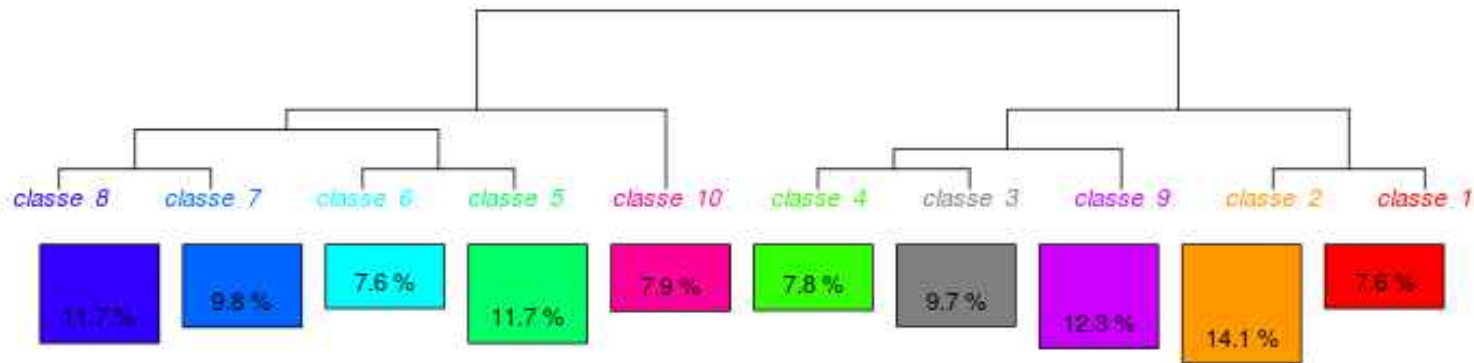
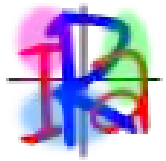
- Présentation générale du projet
  - Carrive Jean, Beloued Abdelkrim, Goetschel Pascale, Heiden Serge, Laurent Antoine, Lisena Pasquale, Mazuet Franck, Meignier Sylvain, Pincemin Bénédicte, Poels Géraldine, Troncy Raphaël (2021) - « Transdisciplinary Analysis of a Corpus of French Newsreels: The ANTRACT Project », *Digital Humanities Quarterly*, 15 (1), <http://digitalhumanities.org/dhq/vol/15/1/000523/000523.html>, <https://hal.archives-ouvertes.fr/hal-03166755>
- La textométrie sur corpus multimédia et le retour à la vidéo
  - Pincemin Bénédicte, Heiden Serge, Decorde Matthieu (2020) - « Textometry on Audiovisual Corpora: Experiments with TXM software », in *Proceedings of 15th International Conference on Statistical Analysis of Textual Data (JADT 2020)*, Université de Toulouse 3 Paul Sabatier, juin 2020. <https://halshs.archives-ouvertes.fr/halshs-02779055>
- Un ouvrage aux Éditions de l'INA réunissant les recherches des historiens sur le corpus → en préparation

# L'équipex Matrice

- MATRICE = Memory Analysis Tools for Research through International Collaboration and Experimentation
- Étudier l'articulation entre mémoire individuelle et mémoire sociale dans le contexte d'événements traumatiques
  - la Seconde Guerre mondiale et la Shoah
  - Les attentats du 11 septembre 2001 aux États-Unis
- « Transdisciplinarité », notamment histoire et psychologie ; des partenariats avec des Mémoriaux (Caen, Rivesaltes...)

# Corpus SHOAH

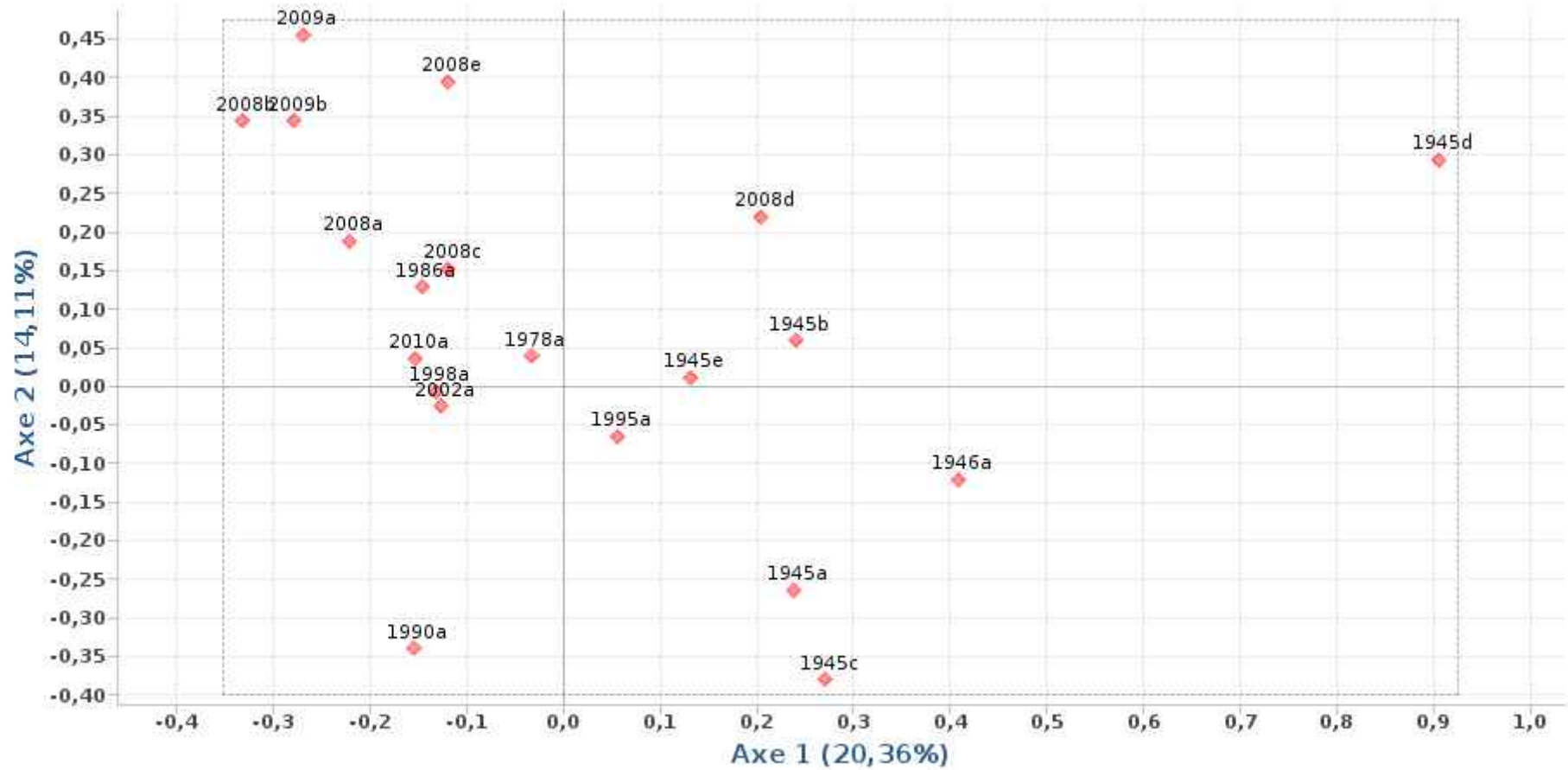
- Un ensemble de témoignages publiés par la Fondation pour la Mémoire de la Shoah (FMS)
  - 20 témoignages majeurs d'Auschwitz
  - trois moments d'écriture des témoins : l'immédiat après-guerre, les années 1970-1990 et les années 2000-2010



Logiciel IraMuTeQ  
<http://iramuteq.org>

wagon	travail	froid	coup	pain	rue	enfant	adèle	vie	juif
porte	kommand	neige	oeil	soupe	pari	mère	simon	situation	pays
bâtiment	block	fatigue	tête	ration	hôtel	soeur	répondre	mémoire	guerre
salle	crématoire	sommeil	regarder	manger	raffe	frère	revoir	sentiment	allemagne
mètre	camp	jambe	crier	pomme	argent	famille	file	condition	politique
étage	four	douleur	main	morceau	appartemen	père	embrasser	survie	france
monter	gaz	soif	bras	margarine	juillet	école	jeune	avenir	nazi
lit	ss	piéd	frapper	lait	carte	parent	siwek	vivre	europe
cour	kapo	allonger	visage	distribution	police	oncle	question	réalité	gouvememen
train	chantier	réveiller	poing	gamelle	ville	ainé	inquiéter	récit	communiste
descendre	appel	glacial	jetter	distribuer	habiter	fiis	zelow	monde	mondial
barbelé	travailler	force	hurler	sucré	lyon	tante	demandeur	espoir	population
escalier	détenu	fièvre	cri	nourriture	faux	naître	aimer	physique	hitler
bois	chef	corps	nie	cigarette	identité	régina	questionner	survivre	allemand
châlit	équipe	endormie	lever	beurre	arondissement	slatka	remercier	circonstance	hitlènen
couloir	créuser	pièce	sang	café	visa	âge	jacob	événement	résistant
immense	orchestre	domme	regard	confiture	quartier	régine	véronique	détail	pétain
camion	vorarbeiter	mort	voir	litre	métier	garçon	visite	comprendre	antisémitie
fenêtre	rang	gêler	homme	tranche	rome	marj	saluer	conscience	citoyen
mur	commando	bovaz	taper	thé	franc	grand_mère	rachel	existence	loi
traverser	sélection	vené	ouvrir	saucisson	tailleur	léon	rencontrer	respect	résistance
route	chambre	na	bâton	boite	patron	armée	rassurer	histore	propagande
planche	daw	sofiak	baisser	gâteau	lettre	filie	hélène	hispes	hongrie
long	ordre	cauchemar	revolver	chocolat	liv	cousin	heureux	témoin	populaire
arriver	kommandos	tomber	cravache	chaud	diancy	nancy	rendre	cama	régime
pièce	hufa	typhus	truser	tabac	vél	classe	sourire	contraire	journal
heure	rassemblement	réchauffer	timber	fromage	boulevard	manier	adresser	agr	français
fer	baraque	mout	genou	gramme	papier	habiter	aller	question	statut
	hórkamari	nuit	apenevoir	assiette		ané	amoureux	oublier	déporté
			inet				poser	esprit	antisémitisme
			hórkamari					deu	

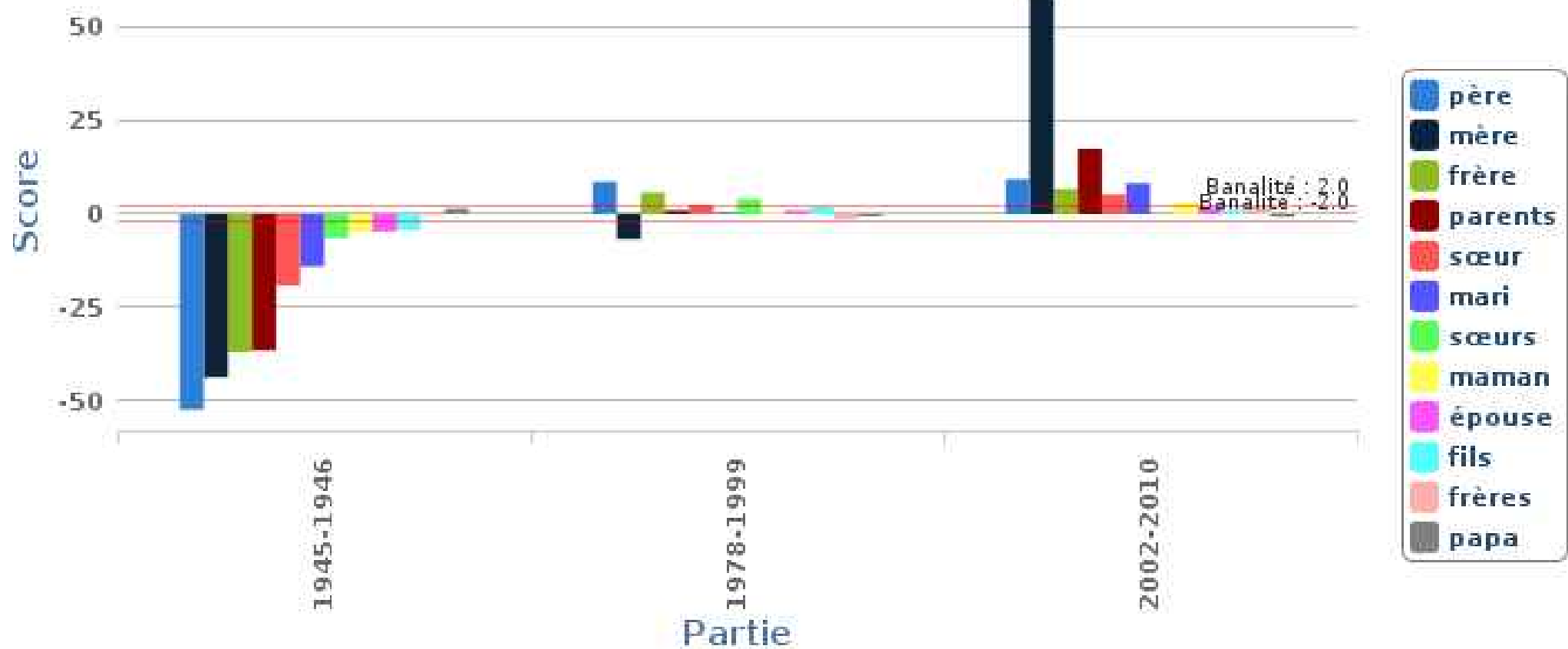
Mise en évidence de champs lexicaux par classification Reinert (méthode ALCESTE).



**Cartographie par analyse factorielle des correspondances (AFC) des 20 témoignages de la Shoah** : pour le calcul les témoignages sont représentés ici par leur fréquence d'emploi des 400 noms communs les plus fréquents. Il ressort que, dès au niveau de ce vocabulaire, les témoignages rédigés à la sortie des camps se démarquent de ceux rédigés ultérieurement (Mayaffre et al. 2018).



**Cartographie AFC des 20 témoignages de la Shoah : même analyse, en affichant les 97 noms structurant le plan (contribution > 0,5 % ou  $\cos^2$  au plan > 0,5). Le vocabulaire sensoriel de la **douleur** et du **quotidien** des camps contraste avec un vocabulaire **familial** ou plus **abstrait** (Mayaffre et al. 2018).**



**Diagramme de spécificité mettant en évidence le suremploi du vocabulaire familial dans les témoignages tardifs.**



# En contexte pédagogique ?

- Voir par exemple :
  - le **logiciel PISTES** (*Pour une investigation systématique des textes*) et sa documentation (livrets de propositions pédagogiques), publié par **Pierre Muller** en 1989 aux éditions du CNDP-INRP.
  - la **thèse d'Isabelle Monin**, soutenue en décembre 2023 : *L'épistolaire éducatif : spécificités grammaticales et génériques des bulletins scolaires et autres écrits de la communication Ecole-familles : des ingrédients linguistiques pour la formation des enseignants*. Doctorat en sciences du langage de l'université de Bourgogne Franche-Comté.

# Vue d'ensemble

<i>Fonctionnalités</i>	<i>Fleurs</i>	<i>Léman</i>	<i>Actualités Fr.</i>	<i>Schoah</i>
<b>Concordance</b> : examen systématique de contextes d'emploi	soleil	seul		
<b>Index</b> : inventaires et décomptes de mots	Noms, Adj, Vb, Adv ; Couleurs	métier ADJ (beau métier, métier dur)	foule de NOM	
<b>Cooccurrences</b> : attirances entre mots, synthèse statistique des contextes	soleil → flamme		agriculture / paysan	
<b>Spécificités</b> : détection et mesure de sur- ou sous-emploi		verbes selon la taille de l'entreprise		vocabulaire de la famille
<b>Progression</b> : répartition des mots au fil du texte, du corpus			agric. / paysan plans PP / PM	
<b>AFC</b> : visualisation spatiale, cartographique, des proximités ou des oppositions entre textes et en vocabulaire		vocabulaire caractéristique suivant l'âge		vocabulaire et distance temporelle
<b>Reinert</b> : synthèse thématique par mise en évidence de champs lexicaux				thématiques des témoignages

# Traits de l'approche textométrique

- Quantitatif et qualitatif
  - Cela a du sens de compter (sans course à la complexité)
  - Centralité du retour au texte
- Endogène, contexte
- Pas automatique
  - Pas de calculer du sens :
    - ordinateur → mémoriser et calculer
    - chercheur → conduire l'analyse
  - observer et explorer pour construire une interprétation ; comprendre ;
  - Enjeu de découverte
- Accéder à une autre échelle (limites cognitives)

# Plan

1. Le rapport aux données dans la recherche : un contexte, une expérience
2. Illustration de l'approche textométrique sur 4 corpus
3. **Données : objectives ou subjectives ?**  
→ Philologie et herméneutique
4. (R)évolution scientifique : incidence globale de la place prise par les données sur le fonctionnement de la recherche

# Les données en textométrie

- Corpus = texte + métadonnées (descripteurs, annotations...)
  - Numérique
- Enjeux scientifiques :
  - Rapport au réel, attestations
  - Nouveaux observables

=> linguistique de corpus, humanités numériques

# Les données ne sont pas données

« un corpus n'est pas un ensemble de données : comme toujours dans les sciences de la culture, **les données sont faites de ce que l'on se donne** (cf. n. 2 p. 96) et le point de vue qui préside à la constitution d'un corpus conditionne naturellement les recherches ultérieures. »

F. Rastier, *Arts et sciences du texte* (2001), chapitre III Philologie numérique, p. 86

=> Pas d'évidence ni de neutralité.

# Les données ne sont pas données

## 2. Rien ne nous est donné

[...] Trois hypothèses ont [...] cours.

(i) **Les données sont ce qu'on vous donne** : [...] les fournisseurs d'accès [...] mettent à disposition des données : mais [...] ce dont on ne connaît pas la source et le mode d'acquisition est ininterprétable.

(ii) **Les données sont ce qu'on vous prend**. [...] Mais en général nous l'ignorons et elles sont données à d'autres (pour des profilages marchands, policiers ou politiques).

(iii) **Les données sont ce qu'on se donne** – ce devrait du moins être le cas en linguistique de corpus et dans l'ensemble des sciences de la culture. En effet, le travail d'analyse commence toujours par la qualification des données comme telles, ne serait-ce que par la qualification et la délimitation du corpus d'étude.

F. Rastier, « Data vs corpora », in D. Mayaffre & L. Vanni (dir.), *L'intelligence artificielle des textes*, Champion, 2021, p. 211-212.

# Construire son corpus

- **Composer** : qu'est-ce qui répond à la problématique ? (cibler, délimiter, composer)
  - pas de langue générale
  - approche contrastive : données = univers !
  - un point d'entrée vers l'absence, l'implicite
- **Établir** : sources à choisir, caractériser → une version vs le texte
  - différences entre éditions, variabilité de transcription
- **Représenter** : numérisation, format
  - potentialités et limites (on n'a pas tout)
- En somme :
  - Philologie numérique
  - Assumer que les données ne soient pas parfaites : interprétabilité



# Exigences scientifiques

- Préparer les données  $\neq$  arranger les données
  - Faut-il s'interdire de retoucher le corpus ?
- Robustesse
  - C'est plutôt au logiciel de s'adapter aux données que l'inverse
- Vigilance critique
  - *A priori*, il peut y avoir des erreurs
    - enquêter en cas d'observation (trop) surprenante.
    - c'est le chercheur qui pilote, pas le logiciel !
  - Toujours revenir aux données, au texte.

# Différents rapports aux données textuelles

<i>Point de vue Computationnel</i>	<i>Point de vue Herméneutique</i>
du texte (matériau)	des textes (identifiés, choisis)
ressource	source
<i>big data</i> (quantité suffisante)	intégrité (robustesse du logiciel qui s'adapte)
éviter lecture	relire
connaissance (langue comme véhicule contingent et imparfait)	formulation (langue comme partie prenante du sens)
le contenu	des interprétations

(Péry-Woodley  
1995)

(Valette 2016)

# Exemples sur l'attention à la formulation

- Lebart & Salem 1994
  - Question ouverte : « Quelles sont les raisons qui, selon vous, peuvent faire hésiter une femme ou un couple à avoir un enfant ? »
  - Réponses : « manque d'argent » ≠ « raisons financières »
- Monin 2023
  - « Balance son sac » ≠ « A balancé son sac »
    - Présent : suggère durée, itération potentielle
    - Passé : accompli, fait
  - Scalarité des usages, asymétrie, ex. :
    - « Trimestre satisfaisant » → pas avec « insatisfaisant »
    - « Résultats corrects » → pas avec « incorrect »

# Plan

1. Le rapport aux données dans la recherche : un contexte, une expérience
2. Illustration de l'approche textométrique sur 4 corpus
3. Données : objectives ou subjectives ?
4. **(R)évolution scientifique** : incidence globale de la place prise par les données sur le fonctionnement de la recherche  
→ Science ouverte

# Avènement de la science ouverte

- Dynamique internationale, qui s'est particulièrement développée à partir des années 2000
  - accessibilité à tous des fruits de la recherche publique
  - mutualisation et capitalisation des efforts
  - transparence, reproductibilité, comparabilité, réutilisabilité/interopérabilité
- En France, le [Plan national pour la science ouverte](#) (Ministère de l'Enseignement supérieur et de la recherche, 2e éd., 2021)



« La science ouverte est la **diffusion** sans entrave des résultats, des méthodes et des produits de la recherche scientifique. Elle s'appuie sur l'opportunité que représente la **mutation numérique** pour développer l'accès ouvert aux **publications** et – autant que possible – aux **données**, aux **codes sources** et aux **méthodes** de la recherche. »

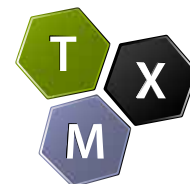
# La science ouverte : un état d'esprit

The screenshot shows the txm.bfm-corpus.org web application in a Mozilla Firefox browser. The interface includes a search bar, navigation tabs for 'Corpus' and 'Accueil', and a main content area displaying a medieval manuscript. The manuscript text is in Old French, with a large initial 'A' in red and blue. The interface also features a sidebar with corpus selection options (BFM2, BFM, COR, GRAAL, PALA) and a bottom navigation bar with 'courante', 'fac-similaire', and 'ms-colonne' options. The page number '160a / 268' is visible at the bottom.

Dans notre équipe,  
concevoir l'ouverture  
tant au niveau des  
données



qu'au niveau de  
l'outil d'analyse

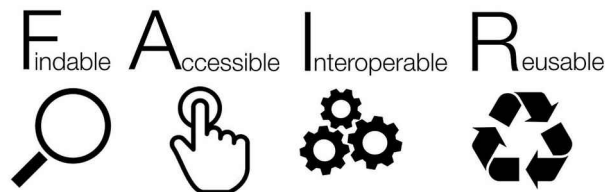


(Guillot et al. 2017)

# Science ouverte : des aspects techniques

- Les principes **FAIR**

- **F**indable (*facile à trouver*)
- **A**ccessible
- **I**nteroperable
- **R**eusable



- Importance des standards

- ex. XML pour les données structurées
- **TEI** (Text Encoding Initiative) pour la représentation des textes numériques



- Les Plans de gestion de données (PGD)

- notamment, que deviennent les données à l'issue d'un projet ?

- Infrastructures, ex. en France:

- les archives ouvertes **HAL** (Hyper Articles en Ligne)
- en sciences humaines et sociales, l'infrastructure de recherche **Huma-Num**



# Science ouverte : aspects juridiques

- Les licences, notamment CC : Creative Commons
  - explicite des possibilités de réemploi (attribution à l'auteur, possibilité de modification, rapport à l'argent, au secret)
- En France, la loi pour une République numérique (2016)
  - [Guide d'application pour les écrits scientifiques](#) (2018)
  - [Guide d'application pour les données de la recherche](#) (2022)
  - Voir aussi la [documentation de HAL](#)





# Science ouverte :

## aspects économiques / systémiques

- Évaluation
  - pour financement, recrutement, promotion...
  - Déclaration de San Francisco (*San Francisco Declaration on Research Assessment - DORA*, 2013), à l'initiative de la Société Américaine pour la Biologie Cellulaire (ASCB)
  - pour une évaluation moins bibliométrique, plus qualitative et diversifiée
  - ex. un nouveau type d'article scientifique, le *data paper*
- Citer ses sources : bibliographie, mais aussi
  - Image
  - Logiciel
  - Jeux de données

# Pour conclure : quels échos en matière pédagogique ?

- Deux clés « classiques » plus que jamais pertinentes à l'ère des données numériques :
  - Philologie : est-ce que mes données sont solides ? (sans doute pas parfaites, mais je connais leurs limites ?)
  - Herméneutique : est-ce que mes données font sens, comment est-ce que je les comprends ?
- Enjeu évaluatif
  - Les données ne s'apprécient pas au poids ! (De même qu'une copie ne s'apprécie pas au remplissage.) Reconnaître et valoriser le travail qualitatif :
  - Pour constituer un jeu de données
  - Pour le rendre utilisable par d'autres
  - Et d'une façon générale pour contextualiser, situer, citer, mettre en relation

# Bibliographie

- Benzécri Jean-Paul et al. (1973) – *L'analyse des données*. Tome 1 : *La taxinomie*, Tome 2 : *L'analyse des correspondances*. Paris, Dunod.
- Benzécri Jean-Paul et al. (1981) – *Pratique de l'analyse des données*. Tome 3 : *Linguistique et lexicologie*. Paris, Dunod.
- Bernard Michel et Bohet Baptiste (2017) – *Littérométrie. Outils numériques pour l'analyse des textes littéraires*. Paris, Presses de la Sorbonne nouvelle.
- Carrive Jean, Beloued Abdelkrim, Goetschel Pascale, Heiden Serge, Laurent Antoine, Lisena Pasquale, Mazuet Franck, Meignier Sylvain, Pincemin Bénédicte, Poels Géraldine, Troncy Raphaël (2021) - « Transdisciplinary Analysis of a Corpus of French Newsreels: The ANTRACT Project », *Digital Humanities Quarterly*, 15 (1), <http://digitalhumanities.org/dhq/vol/15/1/000523/000523.html>
- Guillot Céline, Heiden Serge, Lavrentiev Alexei (2017) - « Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique », *Diachroniques. Revue de Linguistique française diachronique*, 7, 168-184. <https://halshs.archives-ouvertes.fr/halshs-01809581>

# Bibliographie

- Heiden Serge, Magué Jean-Philippe, Pincemin Bénédicte (2010) – « TXM : une plateforme logicielle open-source pour la textométrie – conception et développement ». In Bolasco Sergio, Chiari Isabella, Giuliano Luca (eds), *Statistical Analysis of Textual Data. Proc. of JADT 2010*. Rome, Edizioni Universitarie di Lettere Economia Diritto, 1021-1032.
- Le Lay Yves, Heiden Serge, Merchez Luc, Pincemin Bénédicte (2016) - « Retour de pêche. Le métier de pêcheur à travers le discours des professionnels français du lac Léman », in Émeline Comby, Yannick Mosset et Stéphanie de Carrara (éds.), *Corpus de textes : composer, mesurer, interpréter*, Lyon : ENS éditions, collection « Sociétés, Espaces, Temps », ISBN 978-2-84788-827-0, p. 117-133.  
<https://halshs.archives-ouvertes.fr/halshs-01423605>
- Lebart Ludovic, Pincemin Bénédicte, Poudat Céline (2019). *Analyse des données textuelles*. Québec, Presses de l'Université du Québec.
- Lebart Ludovic, Salem André (1994). *Statistique textuelle*. Paris, Dunod.

# Bibliographie

- Mayaffre Damon, Pincemin Bénédicte, Heiden Serge, Weyl Philippe (2018) - « L'évolution de la mémoire de la Shoah au prisme de la statistique textuelle », *in* Denis Peschanski & Brigitte Sion (dir.), *La vérité du témoin*. Mémoire et mémorialisation, vol. 2, collection Mémoire(s), Hermann Éditeurs, Paris & Institut National de l'Audiovisuel, Bry-sur-Marne, chapitre VI, p. 93-124.  
<https://halshs.archives-ouvertes.fr/hal-01890536v1>
- Mayaffre Damon, Vanni Laurent (dir.) (2021) – *L'intelligence artificielle des textes*. Paris, Champion.
- Monin Isabelle (2023) – *L'épistolaire éducatif : spécificités grammaticales et génériques des bulletins scolaires et autres écrits de la communication Ecole-familles : des ingrédients linguistiques pour la formation des enseignants*. Thèse de doctorat en sciences du langage de l'université de Bourgogne Franche-Comté. <https://theses.fr/2023UBFCH024>
- Née Émilie (dir;) (2017) – *Méthodes et outils informatiques pour l'analyse des discours*. Presses universitaires de Rennes.
- Péry-Woodley Marie-Paule (1995) – « Quels corpus pour quels traitements automatiques ? », *Traitement automatique des langues*, 36 (1-2), p. 213-232.

# Bibliographie

- Pincemin Bénédicte (2012) – « Sémantique interprétative et textométrie », Christoph Cusimano (dir.), *Texto! Textes & Cultures*, 17 (3), <http://www.revue-texto.net/index.php?id=3049>.
- Pincemin Bénédicte (2018) – *Sept logiciels de textométrie*. Document de travail, juillet 2018, 11 pages. <https://halshs.archives-ouvertes.fr/halshs-01843695>
- Pincemin Bénédicte, Heiden Serge, Decorde Matthieu (2020) - « Textometry on Audiovisual Corpora: Experiments with TXM software », in Proceedings of 15th International Conference on Statistical Analysis of Textual Data (JADT 2020), Université de Toulouse 3 Paul Sabatier, juin 2020. <https://halshs.archives-ouvertes.fr/halshs-02779055>
- Pincemin Bénédicte (2017) - « Outils logiciels pour l'analyse de corpus textuels ». In Schnedecker Catherine, Aleksandrova Angelina (éds), *Le doctorat en France : mode(s) d'emploi*, Bruxelles, Peter Lang, ISBN 978-2-8076-0623-4, 173-203.
- Poudat Céline, Landragin Frédéric (2017) – *Explorer un corpus textuel*. Louvain-la-Neuve, De Boeck.

# Bibliographie

- Rastier François (1987) – *Sémantique interprétative*. Paris, Presses universitaires de France.
- Rastier François (2001) – *Arts et sciences du texte*. Paris, Presses universitaires de France.
- Rastier François, Cavazza Marc, Abeillé Anne (1994) – *Sémantique pour l'analyse*. Paris, Dunod.
- Rastier François (2011) – *La mesure et le grain. Sémantique de corpus*. Paris, Champion.
- Rastier François (2021) – « Data vs corpora ». In D. Mayaffre & L. Vanni (dir.), *L'intelligence artificielle des textes*. Paris, Champion.

# Bibliographie

- Ratinaud Pierre, Déjean Sébastien (2009) – « IraMuTeQ : implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre ». Colloque *Modélisation Appliquée aux Sciences Humaines et Sociales* (MASHS2009), Toulouse, [http://repere.no-ip.org/Members/pratinaud/mes-documents/articles-et-presentations/presentation\\_mashs2009.pdf/view](http://repere.no-ip.org/Members/pratinaud/mes-documents/articles-et-presentations/presentation_mashs2009.pdf/view)
- Reinert Max (1990) – « ALCESTE, une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval ». *Bulletin de méthodologie sociologique*, 26, mars 1990, 24-54.
- Valette Mathieu (2016) – « Analyse statistique des données textuelles et traitement automatique des langues. Une étude comparée », in Damon Mayaffre, Céline Poudat, Laurent Vanni, Véronique Magri, Peter Follette (éds), *Statistical Analysis of Textual Data. JADT 2016. Proceedings of 13th International Conference on Statistical Analysis of Textual Data*, vol. II, 697-706. <http://lexicometrica.univ-paris3.fr/jadt/jadt2016/01-ACTES/84134/84134.pdf>
- Viprey Jean-Marie (2002) – *Analyses textuelles et hypertextuelles des Fleurs du mal*. Paris, Champion.