Les données dans la recherche en sciences humaines et sociales

Bénédicte Pincemin

XVIII[èmes] 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

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

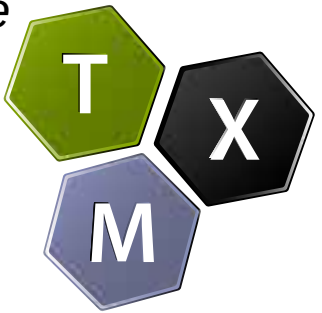
1. **Le rapport aux données dans la recherche : un contexte, une expérience**
2. Illustration de l’approche textométrique sur 4 corpus
3. Données : objectives ou subjectives ?
4. (R)évolution scientifique : incidence globale de la place prise par les données sur le fonctionnement de la recherche

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 2e 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.



https://textometrie.org

Plan

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

EVS

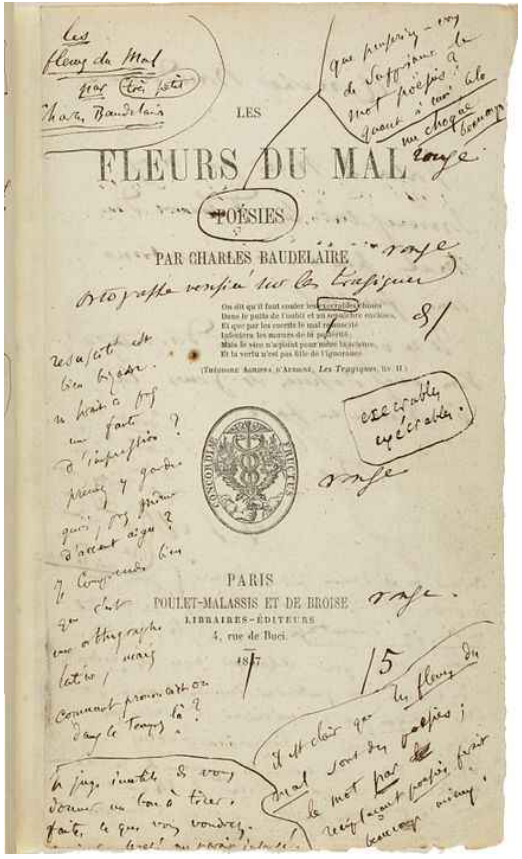
ANTRACT

LES ACTUALITÉS FRANÇAISES

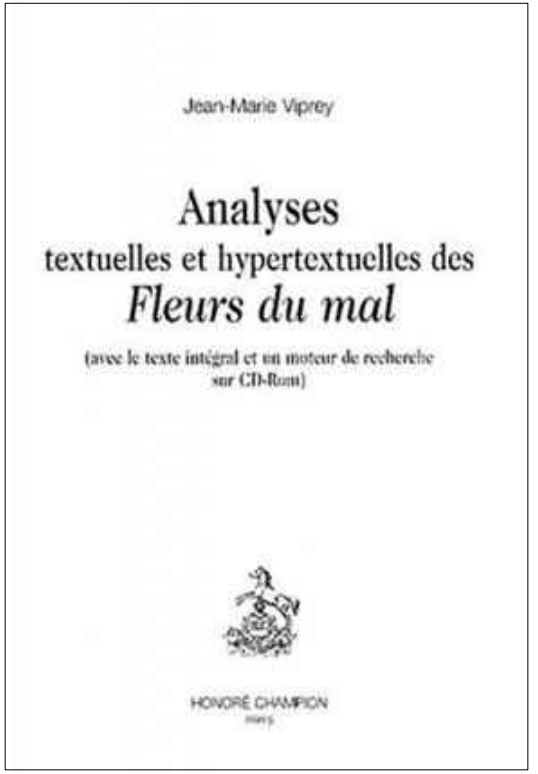
MATRICE

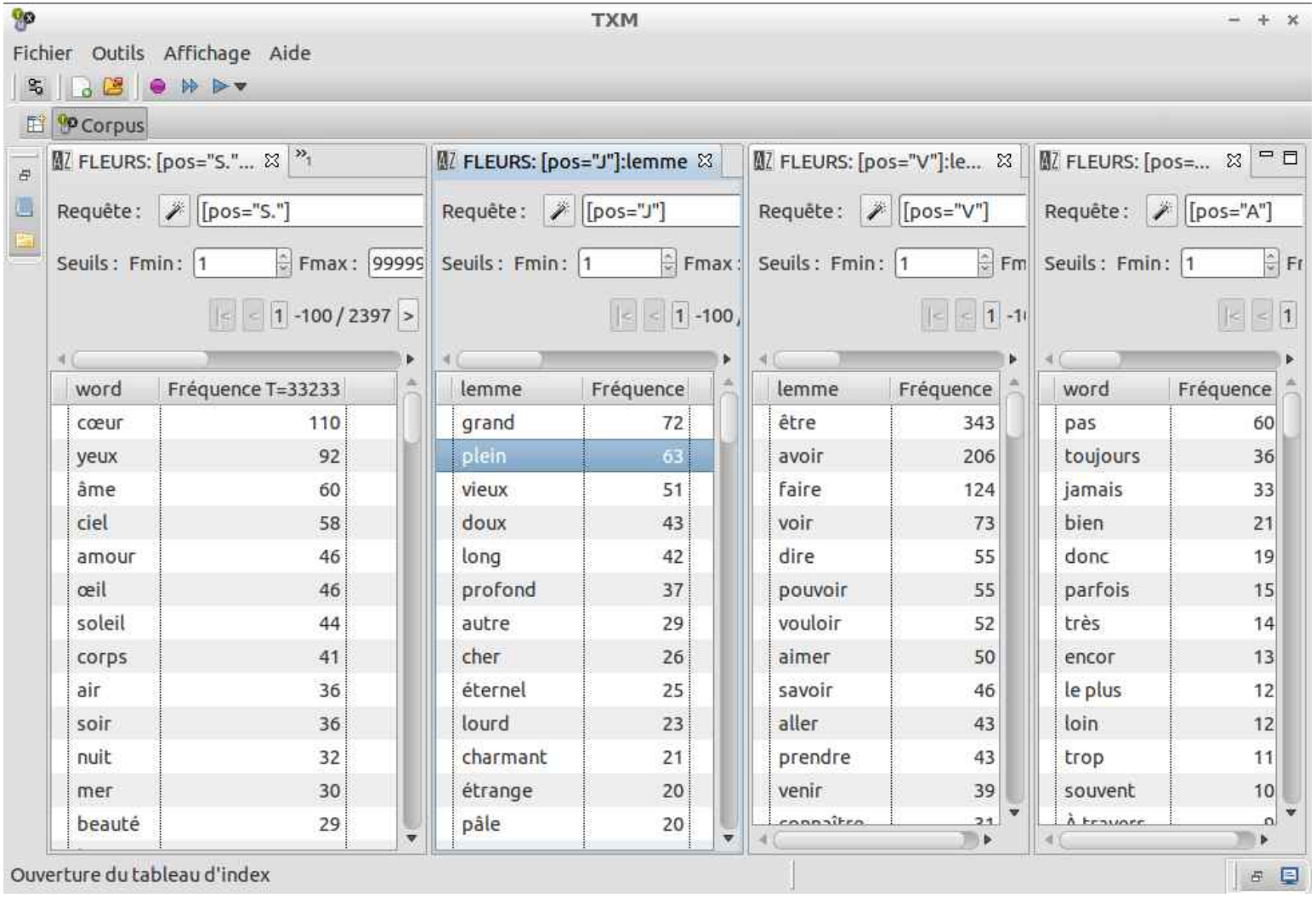
3. Données : objectives ou subjectives ?
4. (R)évolution scientifique : incidence globale de la place prise par les données sur le fonctionnement de la recherche

Corpus FLEURS

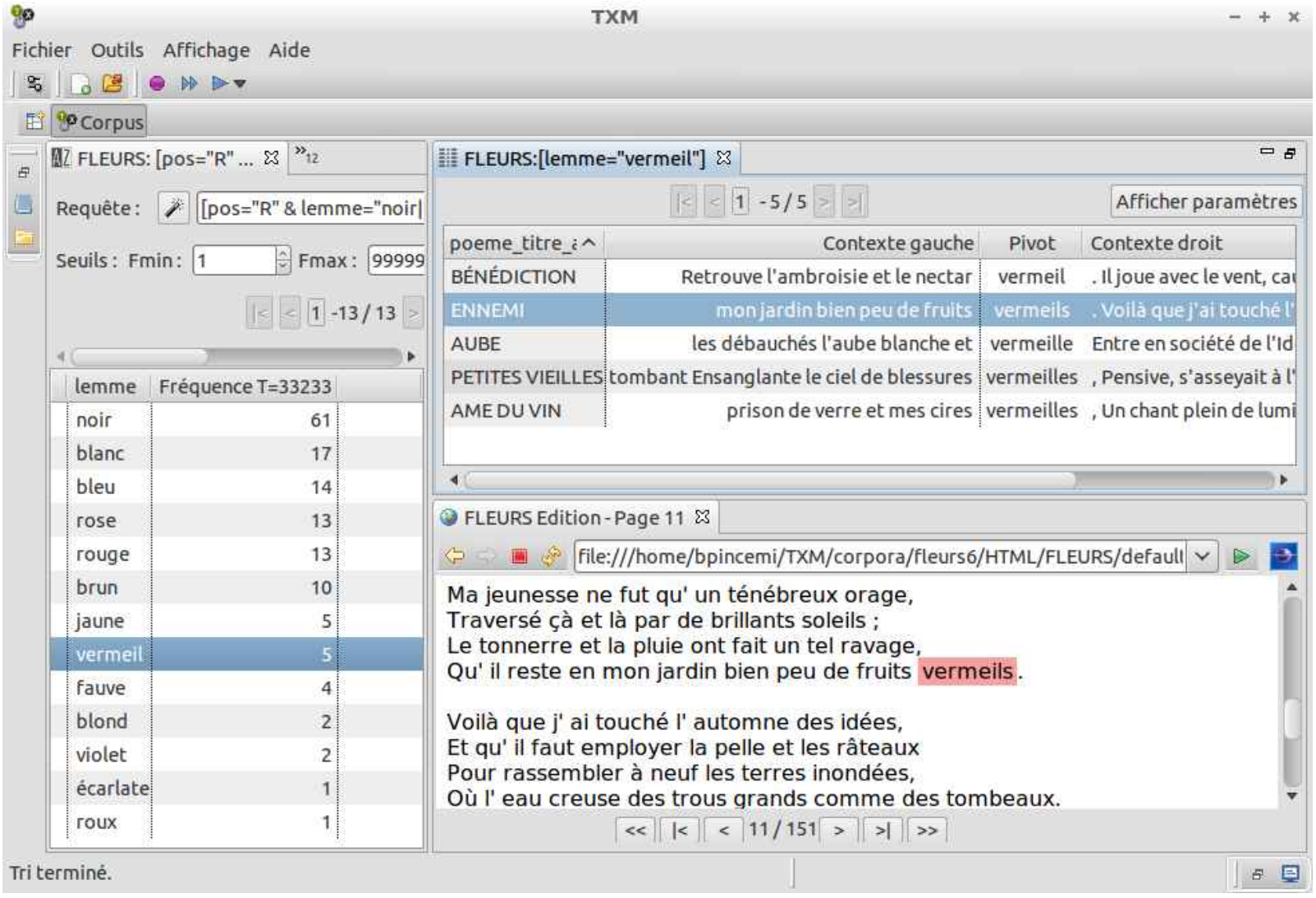


- Le recueil poétique des *Fleurs du Mal* de Baudelaire (plusieurs éd. entre 1857 et 1868)
- Édition numérique préparée par Jean-Marie Viprey, univ. Besançon (thèse puis ouvrage en 2002)
- Corpus préparé pour TXM : téléchargeable depuis la bibliothèque de fichiers pour TXM

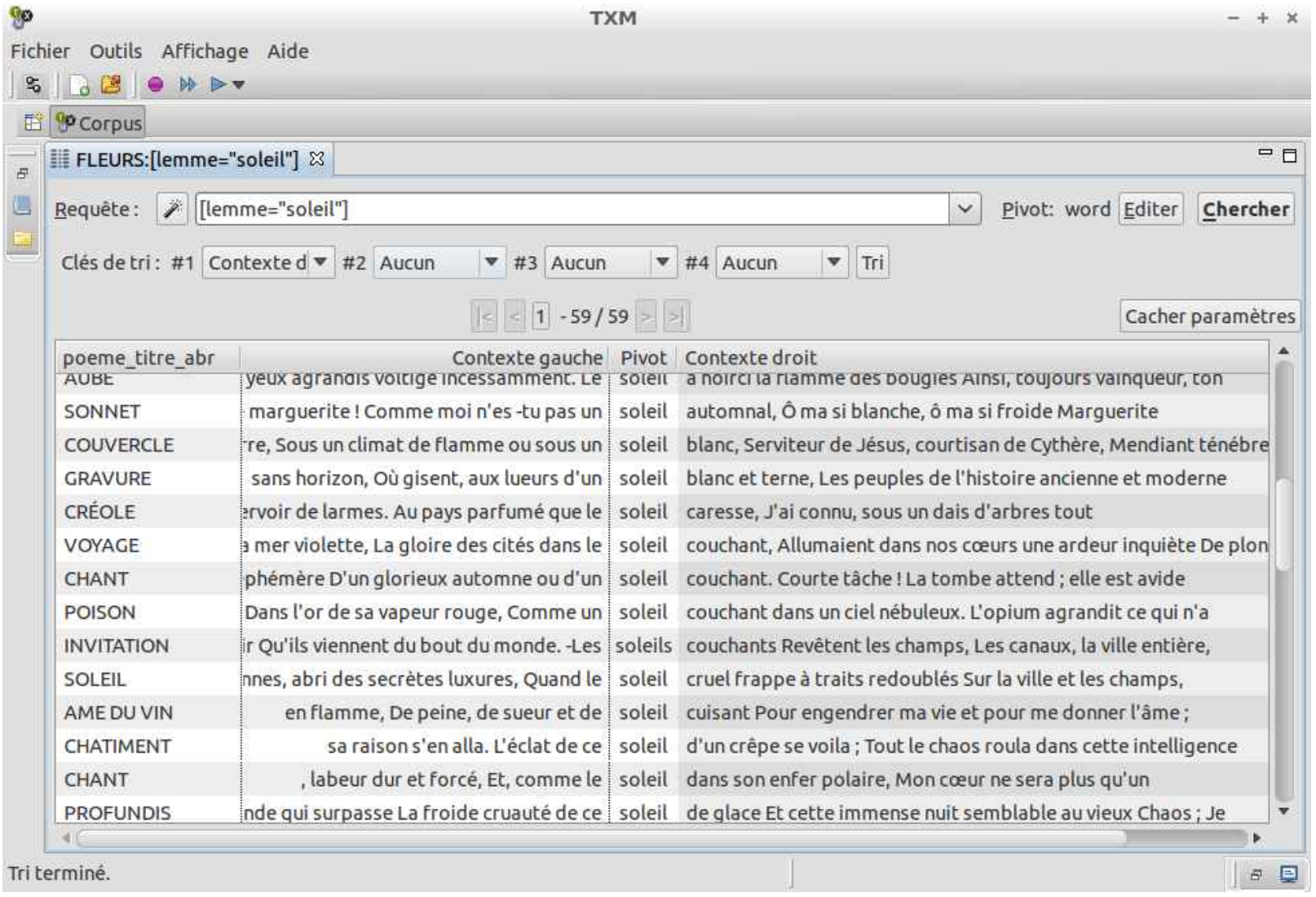




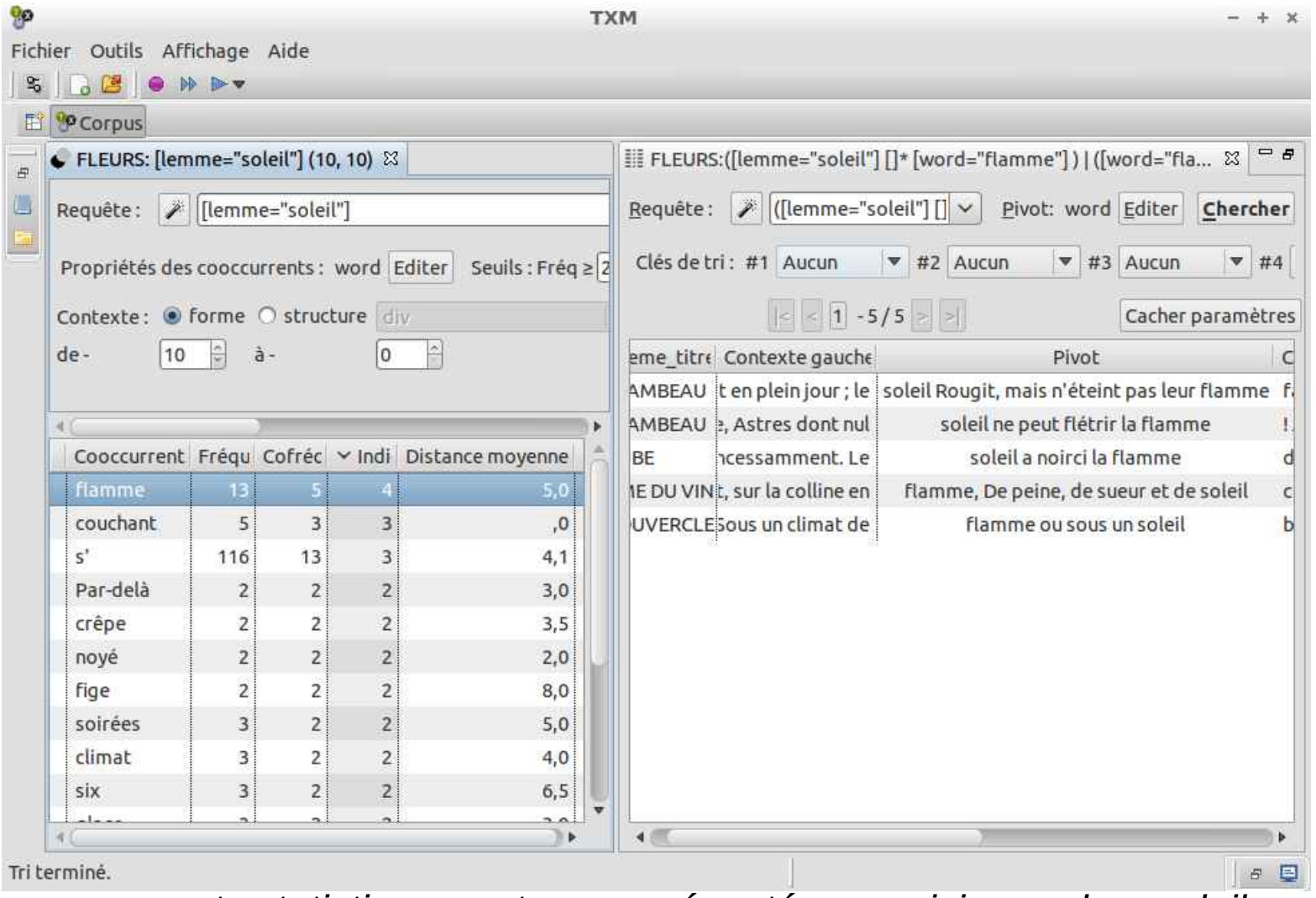
Vocabulaire dominant par catégorie grammaticale : *relevés des noms communs, adjectifs qualificatifs, verbes et adverbes les plus fréquents dans les* Fleurs du Mal.



Dominances au sein d'un paradigme : *emplois des adjectifs de couleur dans les* Fleurs du Mal*.*



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



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
beau métier	7
métier dur	6
autre métier	4
même métier	4
joli métier	3
métier physique	3
ancien métier	2
métier artisanal	2
métier complet	2
métier difficile	2
métier traditionnel	2
métier très individualiste	2
petit métier	2
autres métiers	1
beaux métiers	1
bon métier	1
dernier métier	1

Concordancier du syntagme « beau métier »

moi, c'est le plus	beau métier	du monde pêcheur. Alors en fait, quand
est vrai que c'est un	beau métier	que c'est une denrée qui est
joli métier, on a un	beau métier	. Bon, un métier de travail, faut pas rêver.
on a quand même un	beau métier	, on a quand même des bonnes choses.
même comme un	beau métier	quoi. Ouais il y a une bonne image, et ça
bah voilà c'est un	beau métier	, que quand même quelques inquiétudes
métier, qui est le plus	beau métier	du monde, pour rien au monde je ne ferais

Concordancier du syntagme « métier dur »

une image que c'est un	métier dur	parce qu'effectivement l'hiver
Non que c'est un	métier dur	, pas toujours évident. C'est bien
pêcheur professionnel, c'est un	métier dur	, difficile, parce que malgré tout
donc ça reste malgré tout un	métier dur	, un métier physique. Euh, il y a
,c'est malgré tout un	métier dur	. C'est un métier de passion si on
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
seul	185
individualiste	19
individualisme	1
solitaire	1

Syntagme (lemmes)	F
être tout seul	31
travailler seul	7
aller tout seul	3
apprendre tout seul	3
être seul	3
faire tout seul	3
travailler tout seul	3
aller seul	2
lancer tout seul	2
rester seul	2

Extrait du concordancier du lemme « seul »

d'ici et puis je suis tranquille : je suis	seul	, je vois pas pourquoi j'irais à Lugrin ou à
au ventre d'aller au milieu du lac, tout	seul	. Euh, je vous parle en tant que femme.
pêcheur, seul, qui va faire tout, tout	seul	, euh, alors à ce moment là, oui, il
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
voulez progressez, faut y aller tout	seul	, se faire son expérience euh, se jeter à
me suis démener pendant sept ans tout	seul	pour avoir tous ces clients là. Donc j'ai
encore euh, je continue d'apprendre tout	seul	, quoi, en fait euh, j'ai encore euh
quelque chose qu'on peut apprendre tout	seul	automatiquement quand on arrive sur
tout seul. Si, vous apprendrez tout	seul	. Si, ben je me suis chopé la grêle, euh
demi, une fois qu'il aura fait ça tout	seul	, euh, il peut se lancer y a pas de souci
voilà. Et puis après j'ai commencé tout	seul	. Alors on va prendre une journée d'été
bien passé, puis après j'ai continué tout	seul	. Non je m'étais jamais intéressé à,
Mais là oui, on s'est débrouillé tout	seul	. Si, même, vous prenez la maison
vieille camionnette, je me démerde tout	seul	, voilà hin, perso. C'est pas mon
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'établis-sement

/modernisation/

/métier/

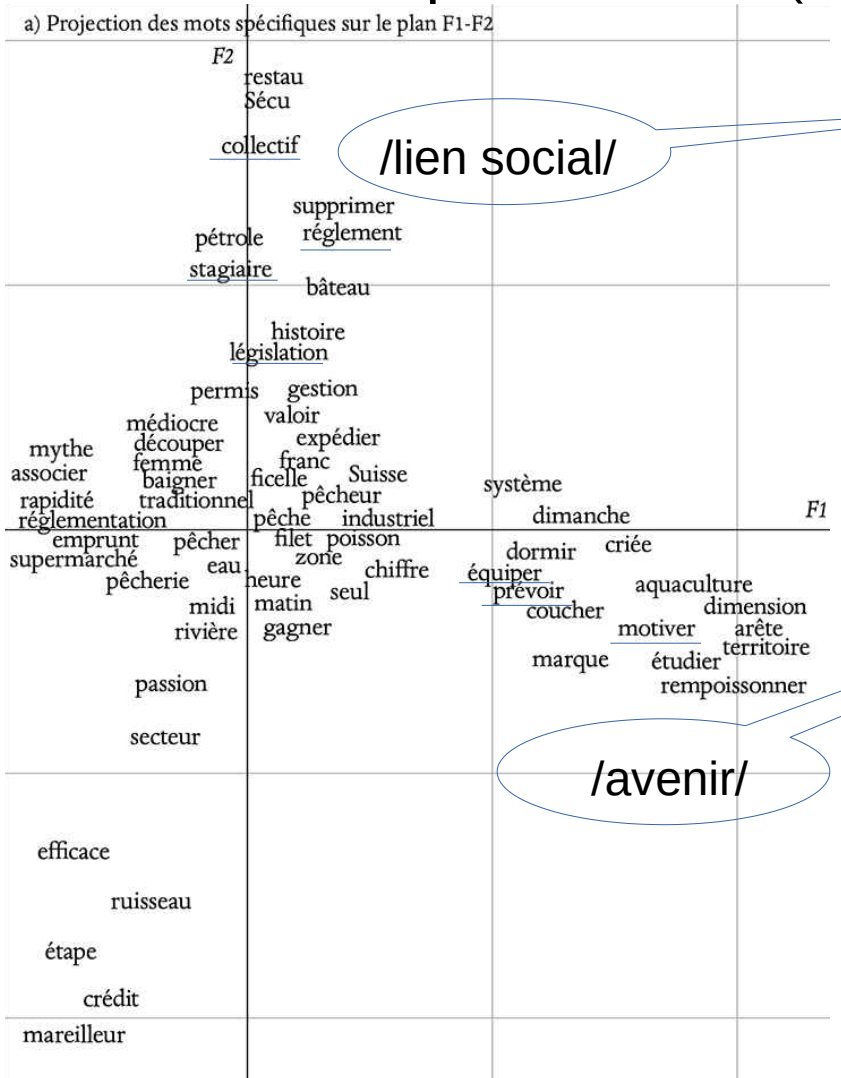
/perspectives/

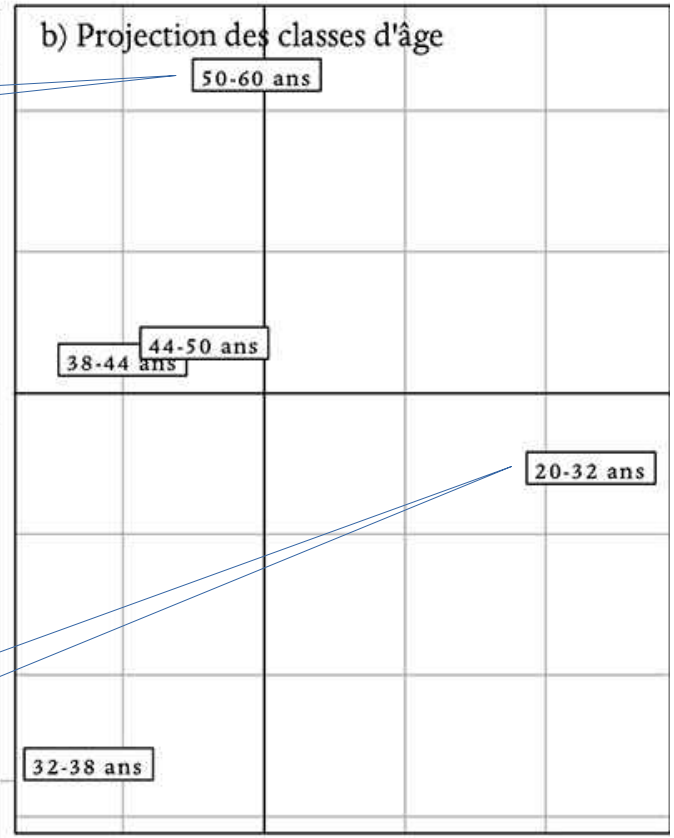
Gros établissements				Moyens établissements				Petits établissements			
Verbe	*F*	*Fp*	*Sp*	*Verbe*	*F*	*Fp*	*Sp*	*Verbe*	*F*	*Fp*	*Sp*
équiper	36	26	8,4	pêcher	718	463	7,0	apprendre	138	58	8,6
étudier	10	9	4,5	respecter	51	43	5,1	sauver	24	16	6,1
prévoir	23	15	4,2	concerner	19	19	5,0	espérer	29	17	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	récupérer	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

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

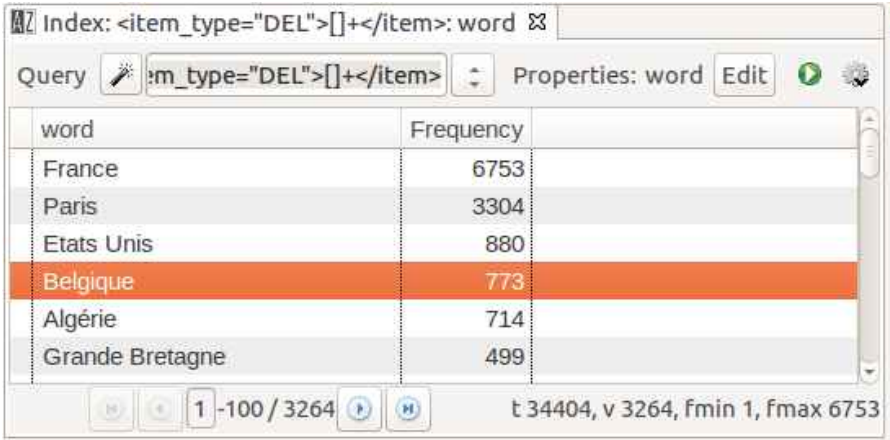


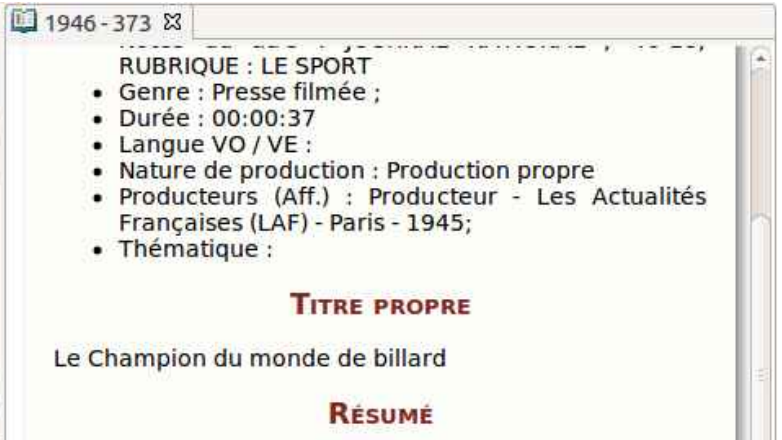


Histoire : Le projet ANR Antract

- 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.

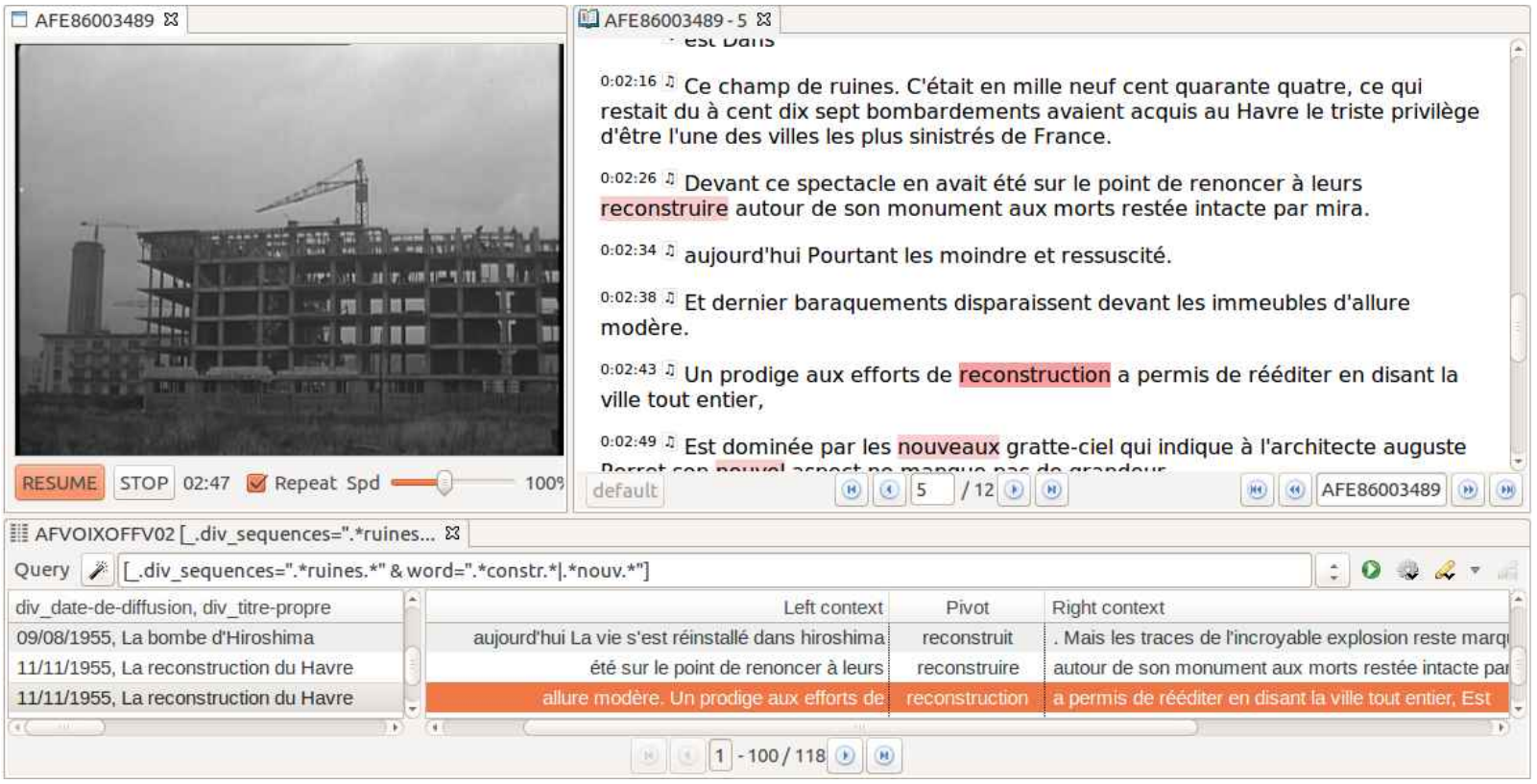
Corpus TXM AF-NOTICES





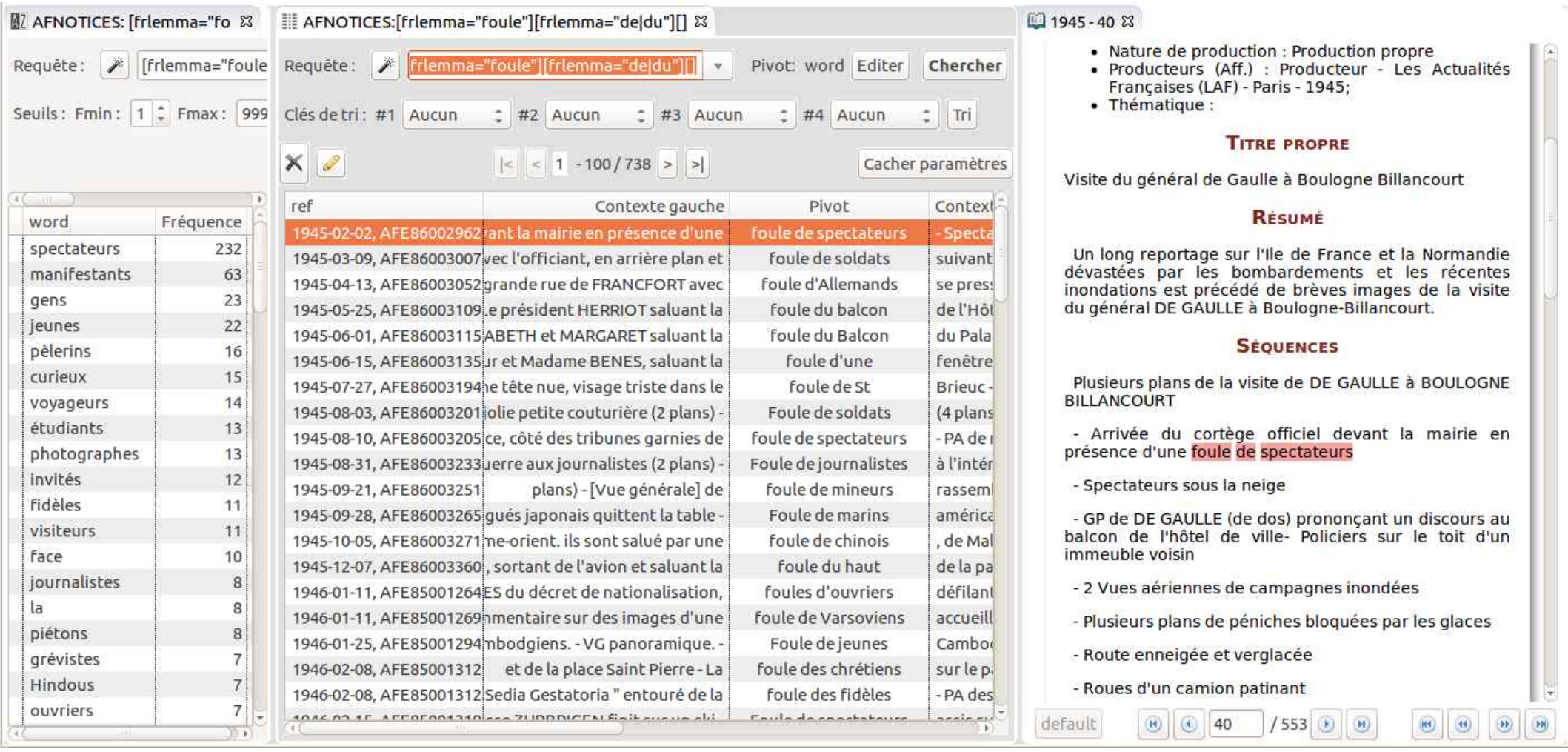
- Source = Base documentaire INA
- 10776 notices sujets de 1261 émissions
- 2,2 millions de mots
- Un corpus structuré (hors texte, listes de descripteurs typés, etc.)

Corpus TXM AF-VOIX-OFF

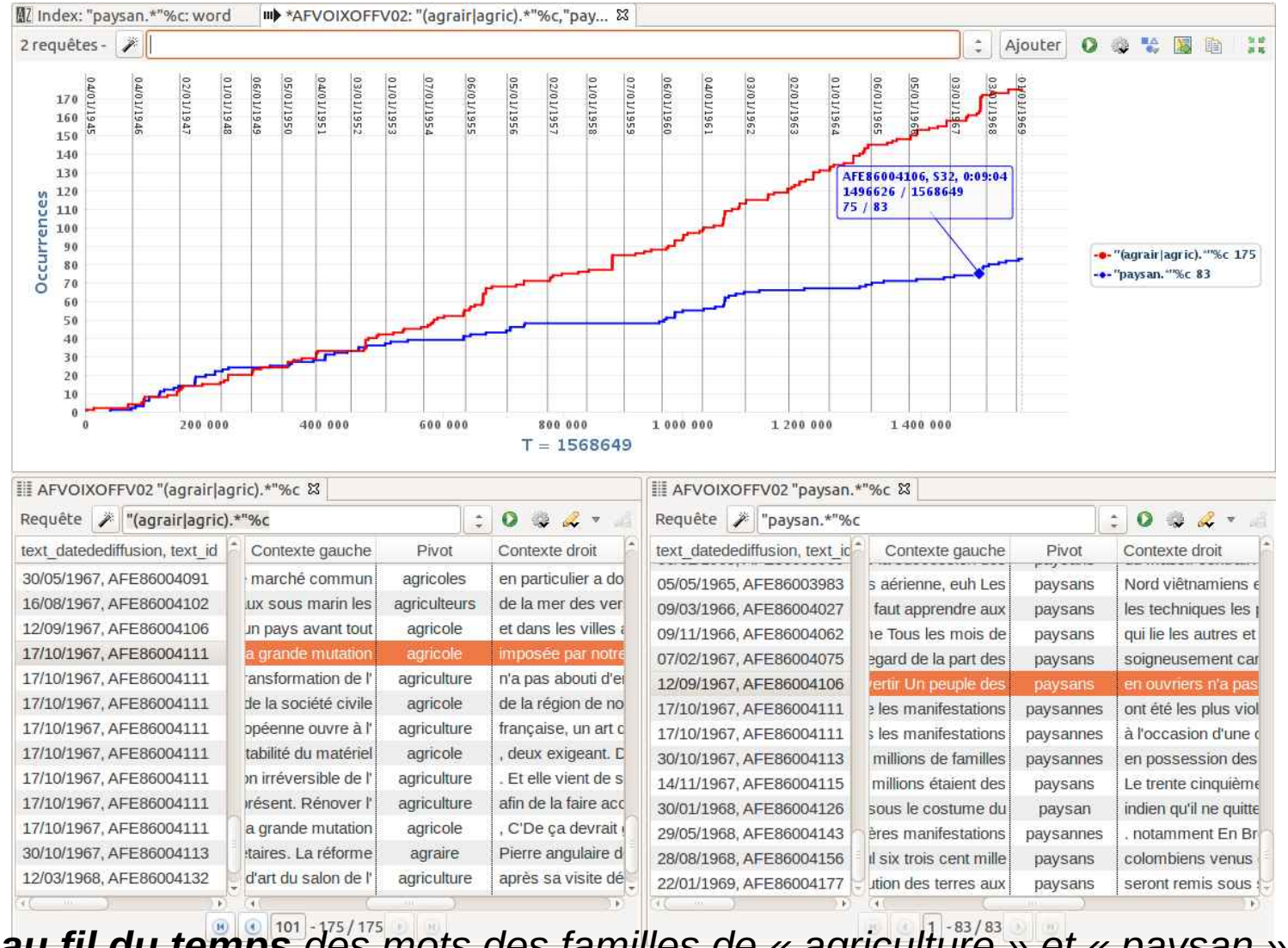


- Source = Transcription automatique de la bande son
- 1260 émissions, au sein desquelles 10683 sujets synchronisés
- 1,5 millions de mots
- Retour à la vidéo en ligne depuis la concordance ou l'édition du texte

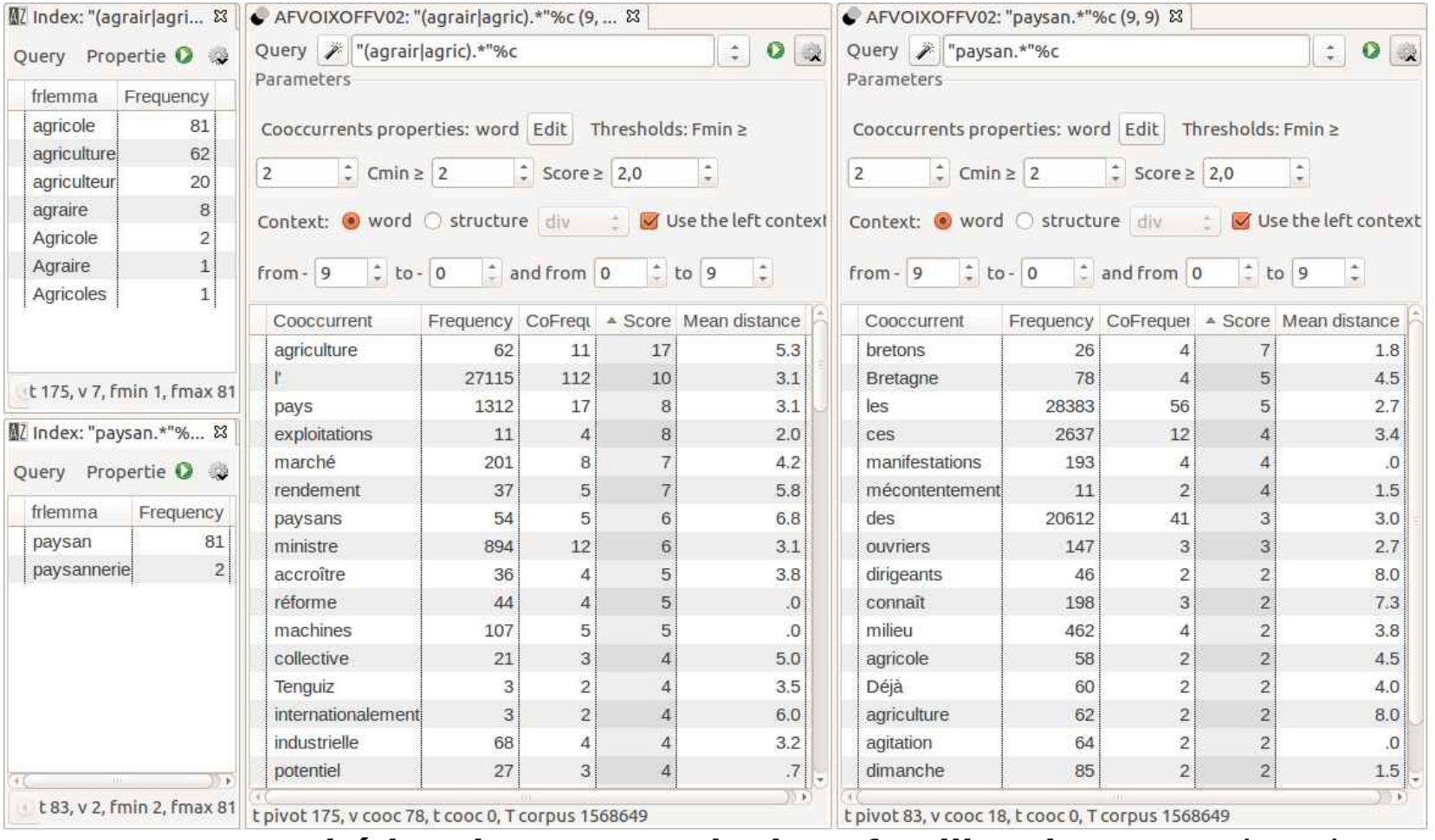
Rq. : Le **retour au document source** (vidéo, manuscrit, édition de référence…) généralise le retour au texte. Pour la transcription, cela permet un contrôle du traitement automatique, et un complément interprétatif face aux nécessaires réductions et choix opérés par la transcription.



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

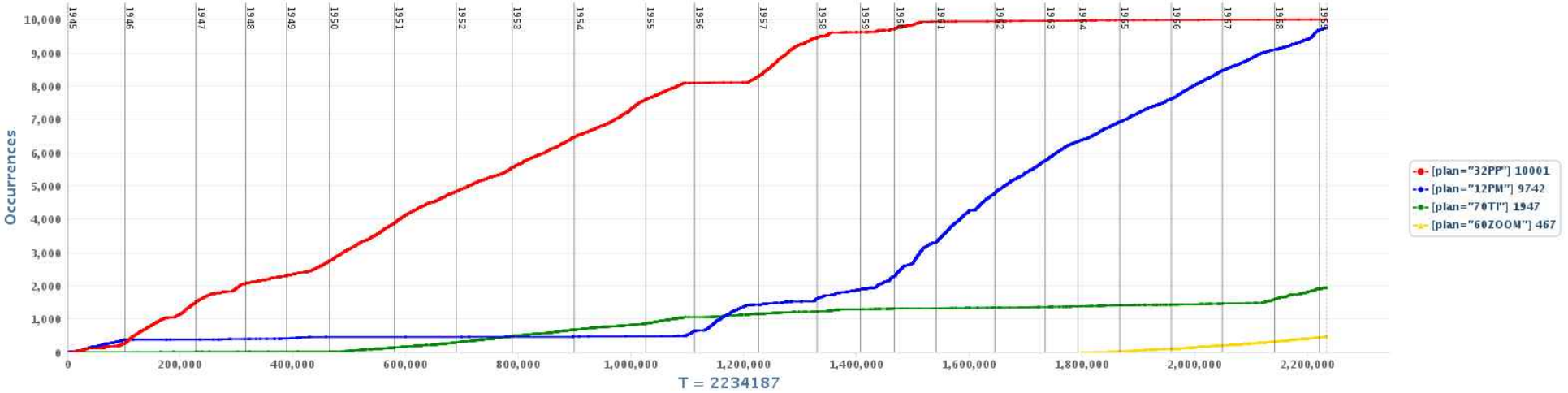


Répartition au fil du temps *des mots des familles de « agriculture » et « paysan » dans les transcriptions des* Actualités françaises*, et retour au contexte (modernisation* vs *tradition).*



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

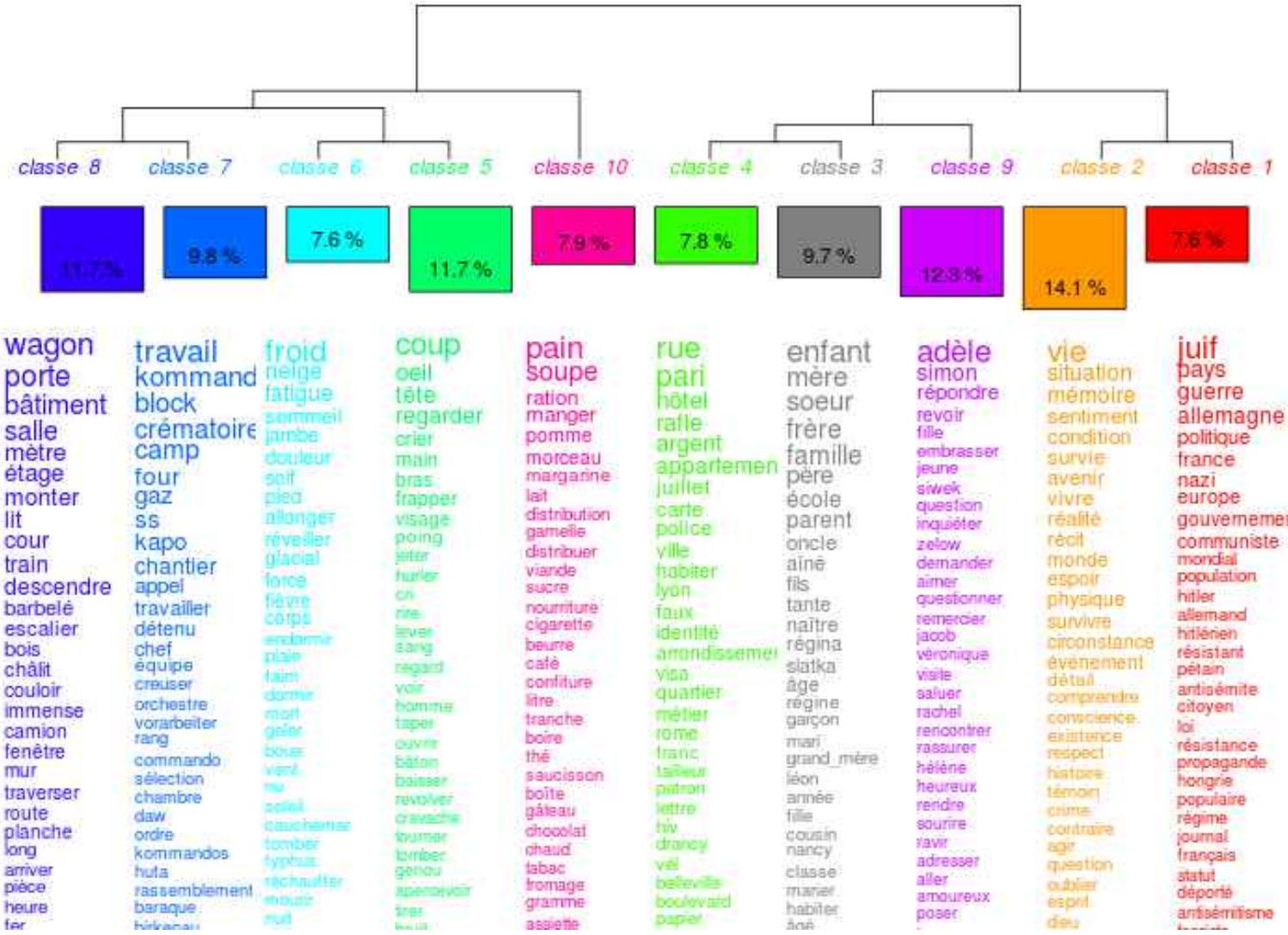
- Présentation générale du projet
  - Carrive Jean, Beloued Abdelkrim, Goetschel Pascale, Heiden Serge, Laurent Antoine, Lisena Pasquale, Mazuet Franck, Meignier Sylvain, Pincemin Bénédicte, Poels Géraldine, Troncy Raphaël (2021) - « Transdiciplinary Analysis of a Corpus of French Newsreels: The ANTRACT 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...)

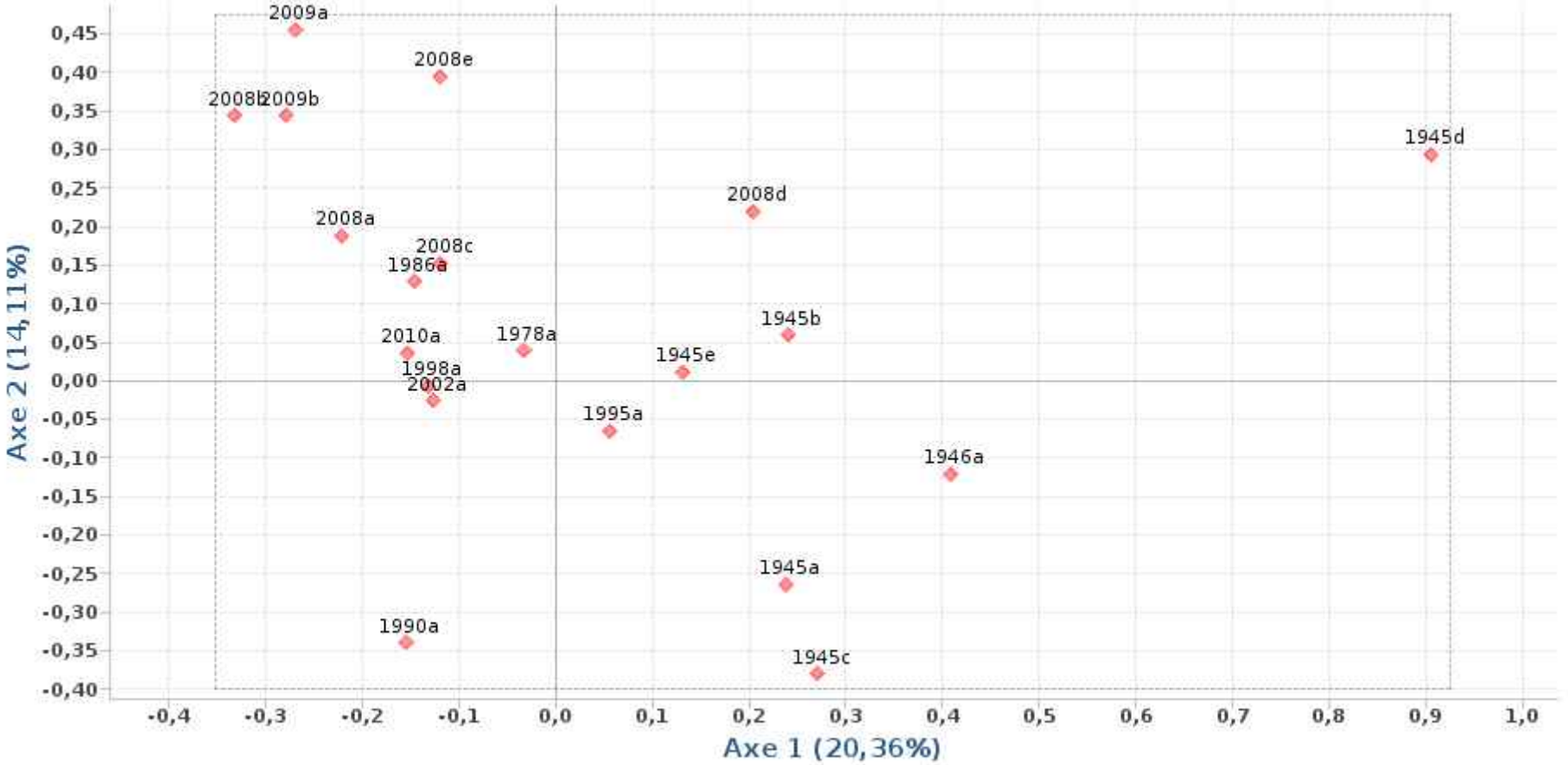
Corpus SHOAH

- 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

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 cos² 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).*

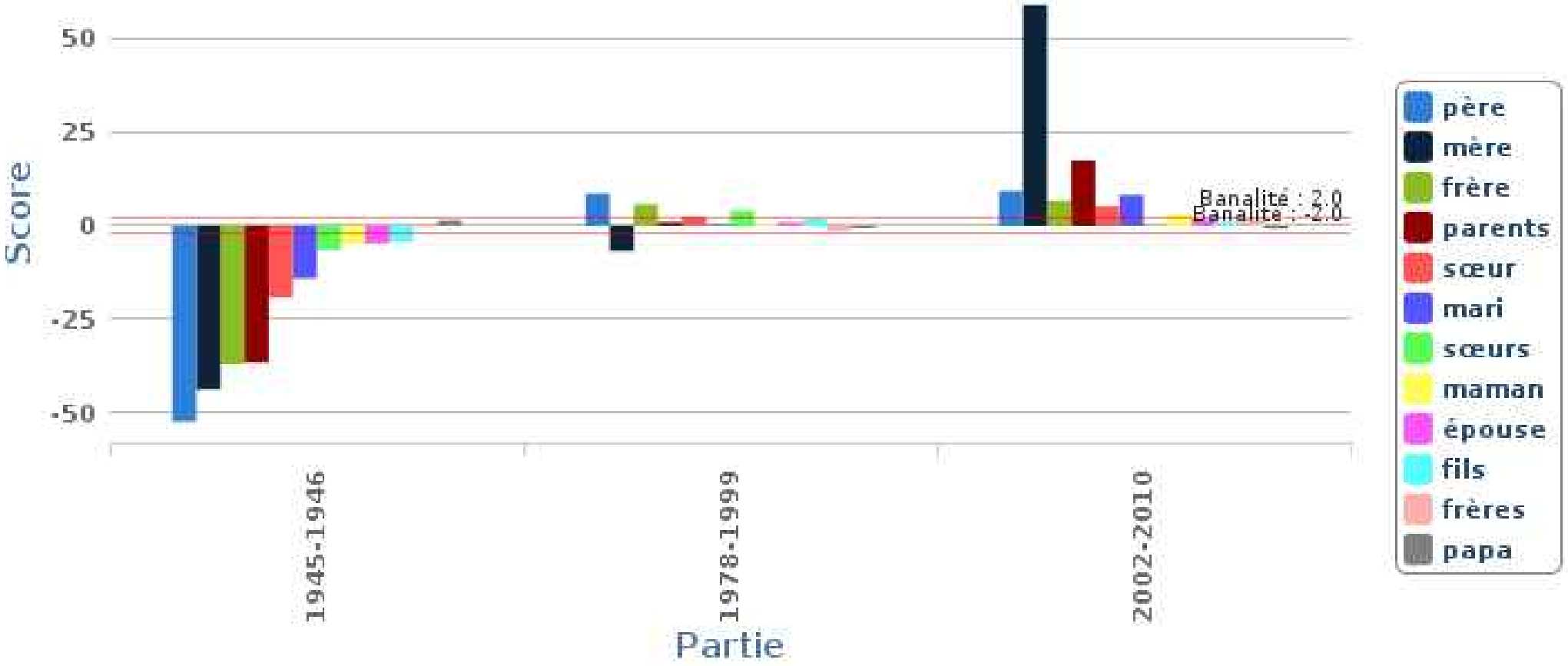


Diagramme de spécificité mettant en évidence le suremploi *du vocabulaire familial dans les témoignages tardifs.*

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

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

Point de vue Computationnel	***Point de vue Herméneutique***
du texte (matériau)	des textes (identifiés, choisis)
ressource	source
big data (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

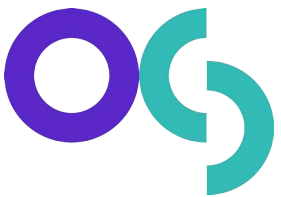
- Lebart & Salem 1994
  - Question ouverte : « Quelles sont les raisons qui, selon vous, peuvent faire hésiter une femme ou un couple à avoir un enfant ? »
  - Réponses : « manque d'argent » ≠ « raisons financières »
- Monin 2023
  - « Balance son sac » ≠ « A balancé son sac »
    - Présent : suggère 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

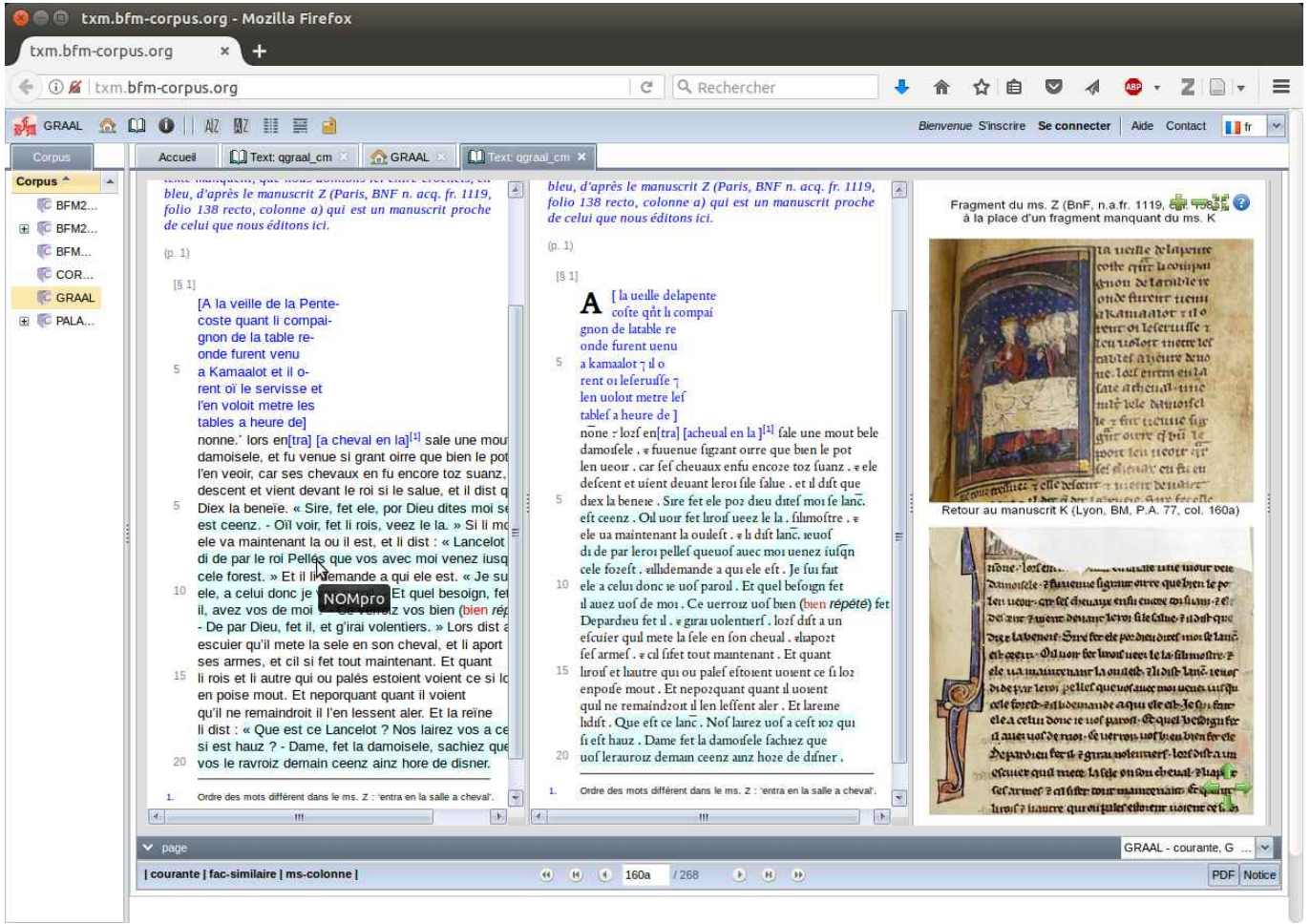
Avènement de la science ouverte

- 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



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

BFM

qu'au niveau de l'outil d'analyse

TXM

(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'initative 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) - « Transdiciplinary 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
- 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 Texual 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.