Journée d'étude « Outils d'exploration de corpus »
Club Corpus de Sorbonne Université
Paris, vendredi 17 juin 2022

Présentation du logiciel TXM
Un point de vue utilisateur

Bénédicte PINCEMIN
IHRIM, université de Lyon, CNRS

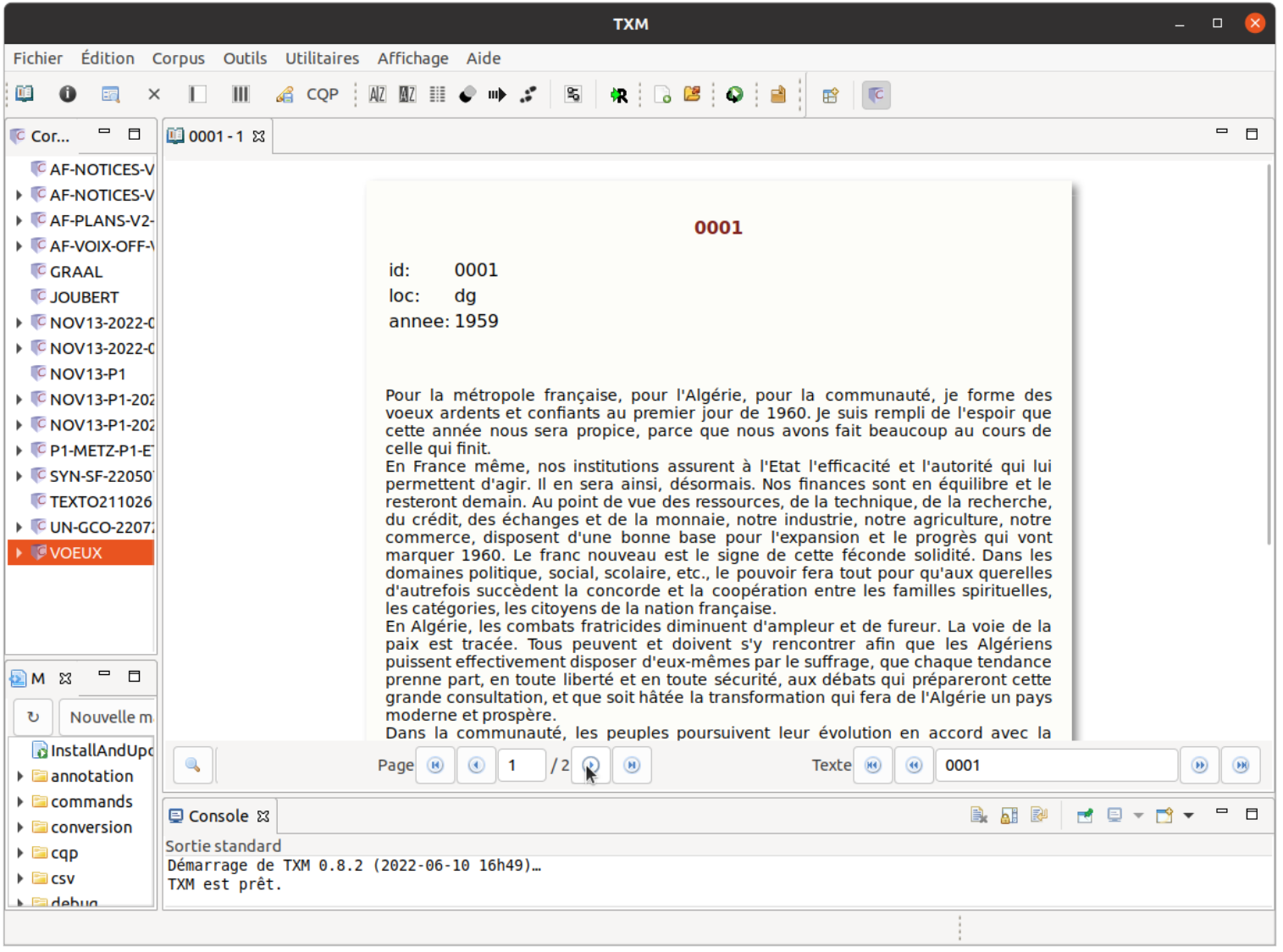
Plan

- Mini-démo : aperçu concret des types d’analyses disponibles dans TXM
- Positionnement dans le paysage des outils d’exploration de corpus (Textométrie, TXM)
- Mise en évidence de quelques caractéristiques fortes de TXM
- Évocation de développements récents
- TXM pour quels points de vue, pour quels usages ? Des forces et limites dessinant un profil

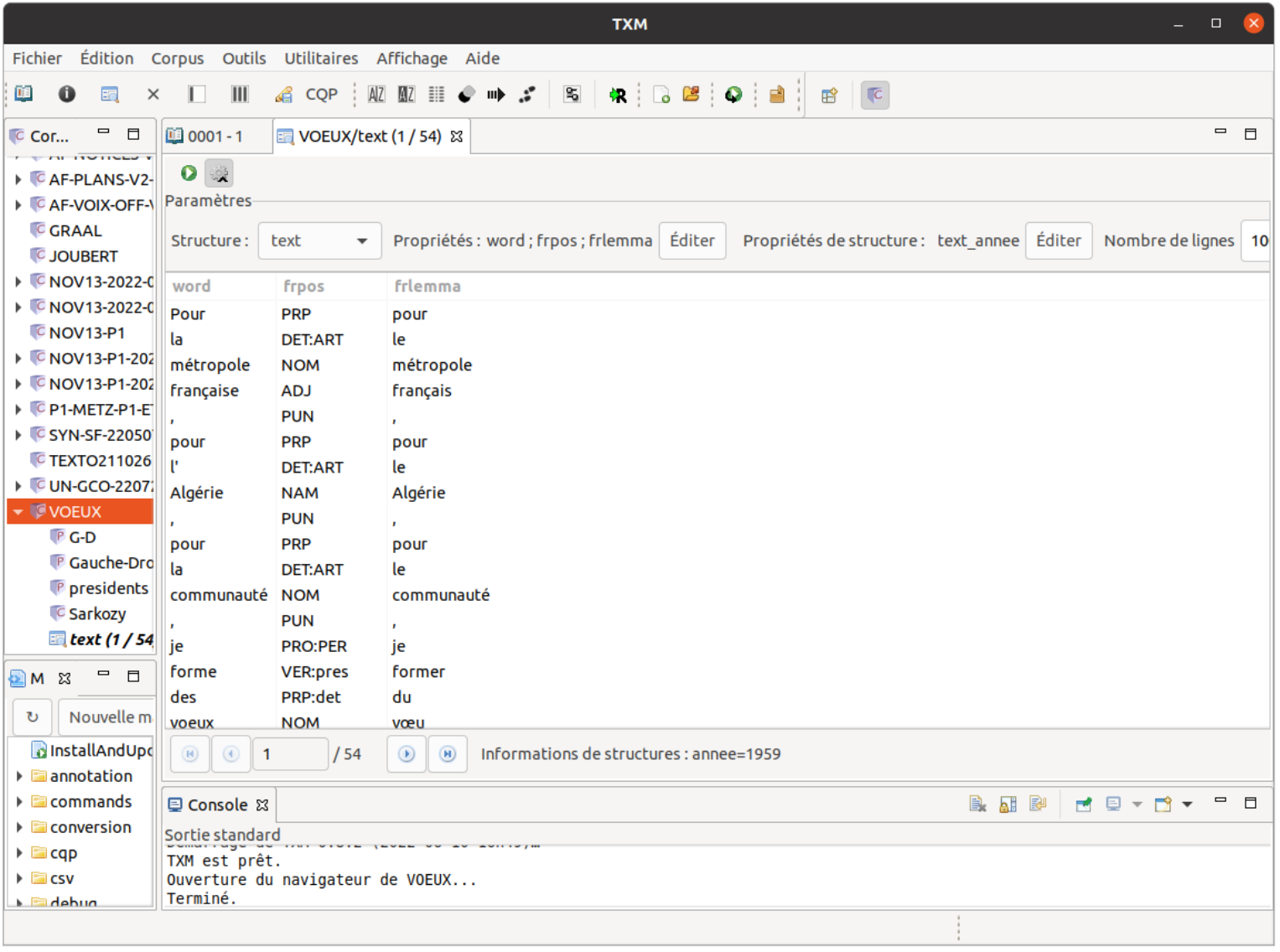
Mini-Démo TXM

- Corpus utilisé : VOEUX
  - Conçu et réalisé par Jean-Marc Leblanc (CEDITEC, université Paris Est Créteil)
  - Composé des discours de vœux des Présidents de la République en France, sous la V^{e} République, sur la période 1959-2012 (54 discours, 60 000 mots)
  - Intéressant comme support d'illustrations car
    - accessible à un large public,
    - possibilité d'analyses diachroniques,
    - possibilités de contrastes selon les présidents.
  - Disponible comme corpus exemple dès l'installation de TXM.

ÉDITION



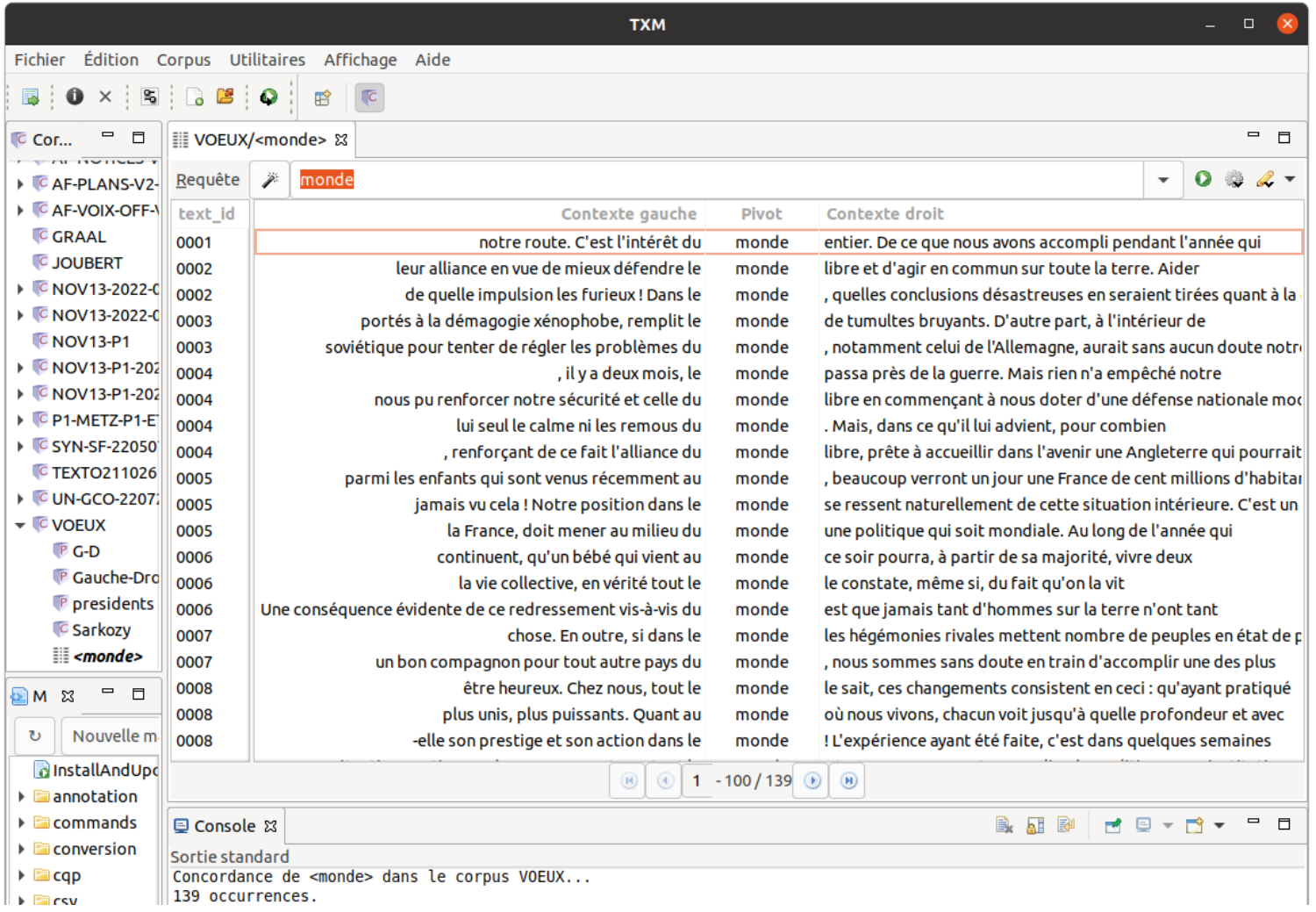
- Pour feuilleter les textes
- « Retour au texte » : toujours pouvoir voir les données dont vient un résultat.



NAVIGATEUR (ou VUE INTERNE)

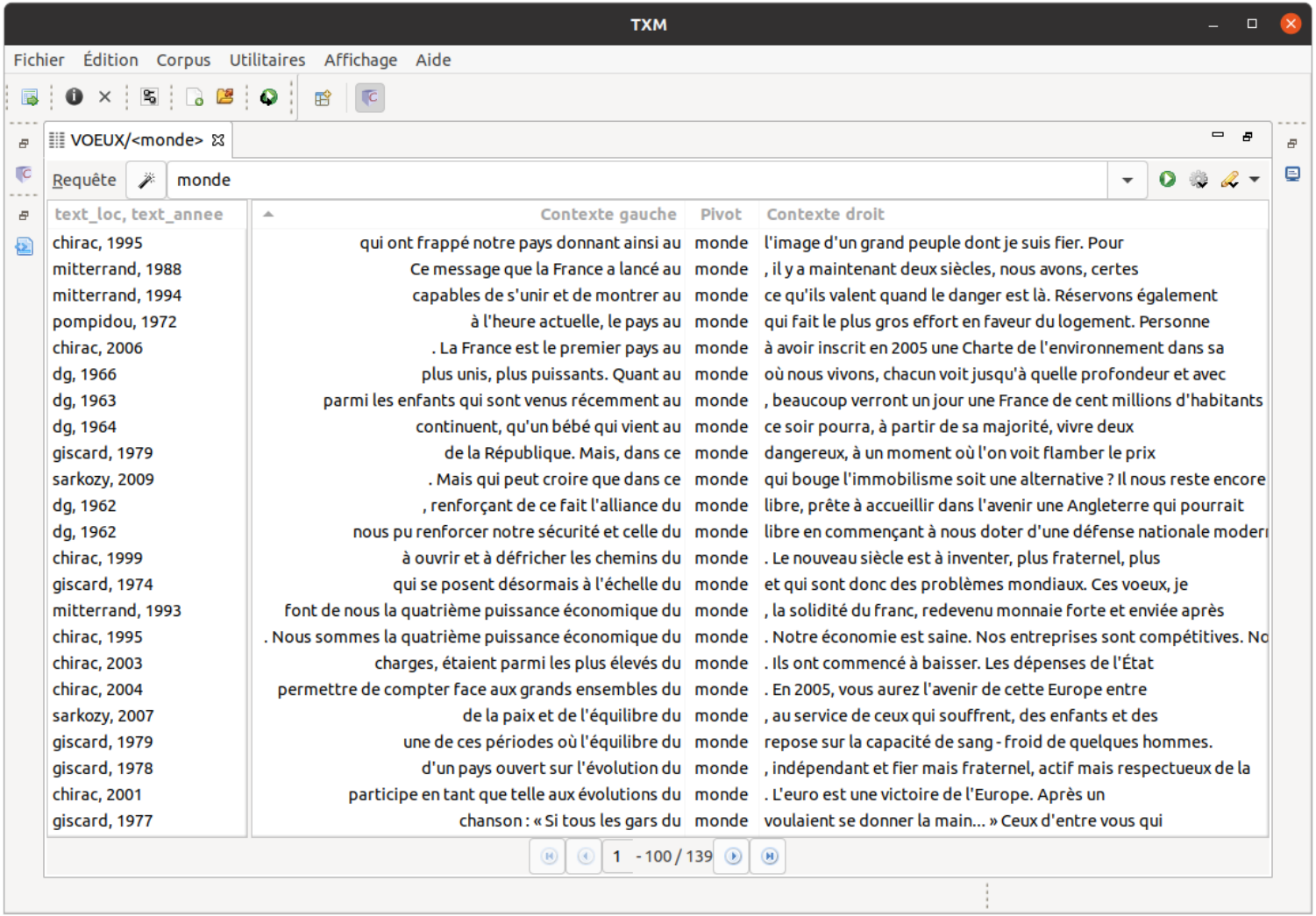
- Visualiser comment mes textes sont représentés dans TXM.

CONCORDANCE



- Une présentation des contextes d'un mot (ou motif plus complexe) « en tableau », efficace pour l'analyse :
  - parcours dense et synthétique des emplois
  - mise en évidence des phraséologies et formulations récurrentes

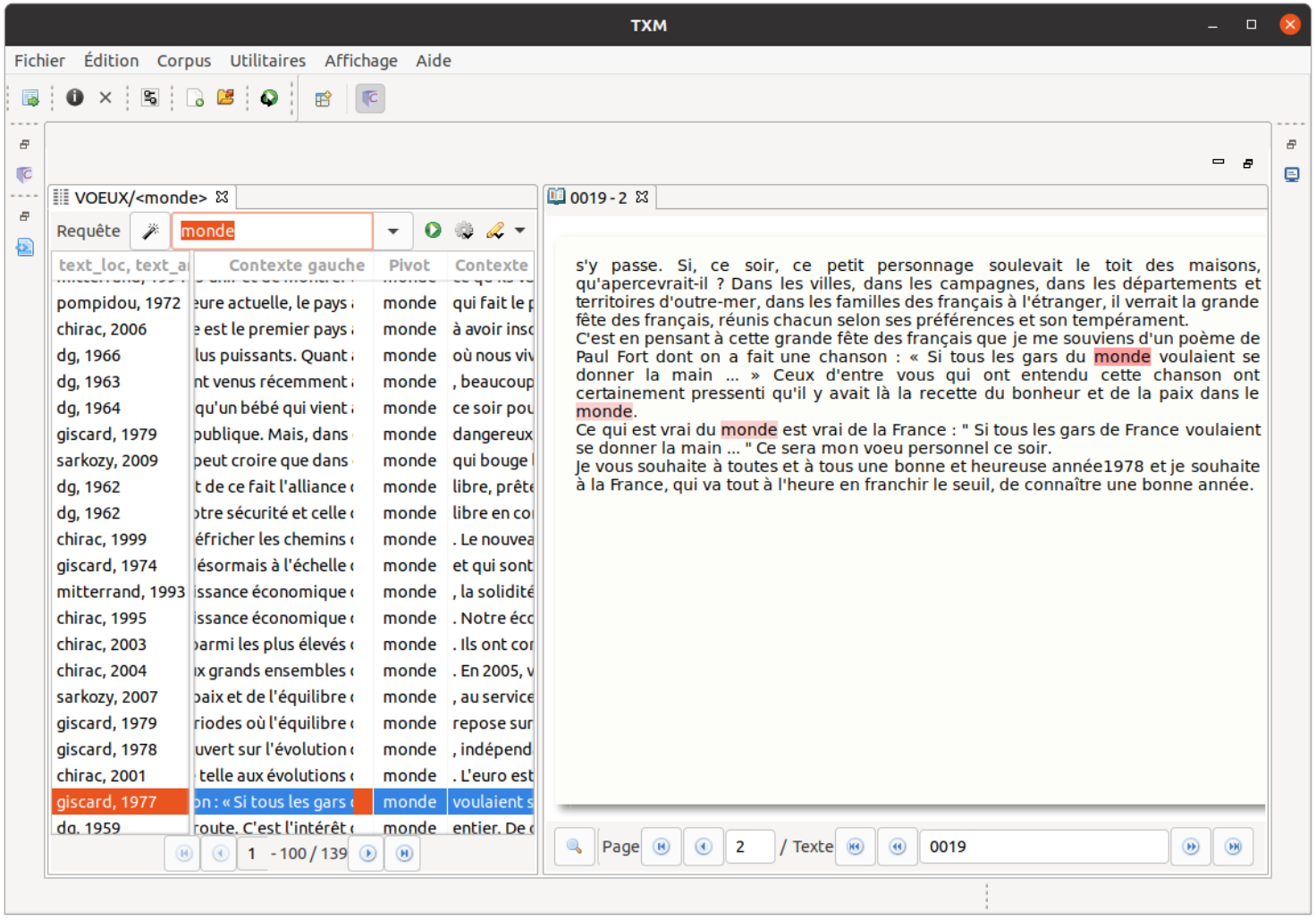
CONCORDANCE



text_loc, text_annee	Contexte gauche	Pivot	Contexte droit
chirac, 1995	qui ont frappé notre pays donnant ainsi au	monde	l'image d'un grand peuple dont je suis fier. Pour
mitterrand, 1988	Ce message que la France a lancé au	monde	, il y a maintenant deux siècles, nous avons, certes
mitterrand, 1994	capables de s'unir et de montrer au	monde	ce qu'ils valent quand le danger est là. Réservons également
pompidou, 1972	à l'heure actuelle, le pays au	monde	qui fait le plus gros effort en faveur du logement. Personne
chirac, 2006	. La France est le premier pays au	monde	à avoir inscrit en 2005 une Charte de l'environnement dans sa
dg, 1966	plus unis, plus puissants. Quant au	monde	où nous vivons, chacun voit jusqu'à quelle profondeur et avec
dg, 1963	parmi les enfants qui sont venus récemment au	monde	, beaucoup verront un jour une France de cent millions d'habitants
dg, 1964	continuent, qu'un bébé qui vient au	monde	ce soir pourra, à partir de sa majorité, vivre deux
giscard, 1979	de la République. Mais, dans ce	monde	dangereux, à un moment où l'on voit flamber le prix
sarkozy, 2009	. Mais qui peut croire que dans ce	monde	qui bouge l'immobilisme soit une alternative ? Il nous reste encore
dg, 1962	, renforçant de ce fait l'alliance du	monde	libre, prête à accueillir dans l'avenir une Angleterre qui pourrait
dg, 1962	nous pu renforcer notre sécurité et celle du	monde	libre en commençant à nous doter d'une défense nationale moderr
chirac, 1999	à ouvrir et à défricher les chemins du	monde	. Le nouveau siècle est à inventer, plus fraternel, plus
giscard, 1974	qui se posent désormais à l'échelle du	monde	et qui sont donc des problèmes mondiaux. Ces voeux, je
mitterrand, 1993	font de nous la quatrième puissance économique du	monde	, la solidité du franc, redevenu monnaie forte et enviée après
chirac, 1995	. Nous sommes la quatrième puissance économique du	monde	. Notre économie est saine. Nos entreprises sont compétitives. No
chirac, 2003	charges, étaient parmi les plus élevés du	monde	. Ils ont commencé à baisser. Les dépenses de l'État
chirac, 2004	permettre de compter face aux grands ensembles du	monde	. En 2005, vous aurez l'avenir de cette Europe entre
sarkozy, 2007	de la paix et de l'équilibre du	monde	, au service de ceux qui souffrent, des enfants et des
giscard, 1979	une de ces périodes où l'équilibre du	monde	repose sur la capacité de sang - froid de quelques hommes.
giscard, 1978	d'un pays ouvert sur l'évolution du	monde	, indépendant et fier mais fraternel, actif mais respectueux de la
chirac, 2001	participe en tant que telle aux évolutions du	monde	. L'euro est une victoire de l'Europe. Après un
giscard, 1977	chanson : « Si tous les gars du	monde	voulaient se donner la main... » Ceux d'entre vous qui

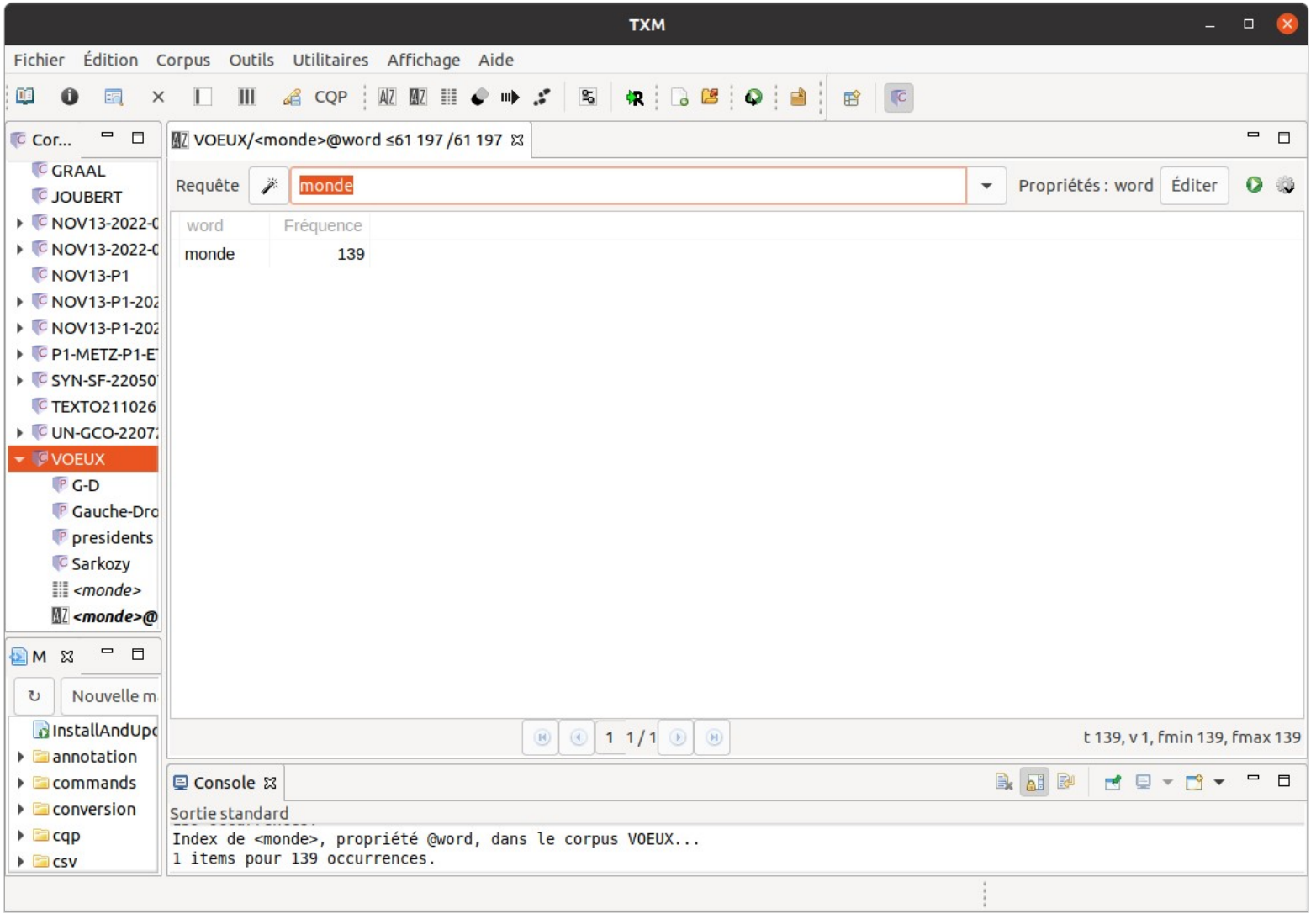
- Exemple avec :
  - Tri sur le contexte gauche
  - Composition personnalisée des références (colonne de gauche)

CONCORDANCE



- Retour à l'ÉDITION, pour un accès au contexte textuel, par double-clic sur une ligne de concordance.

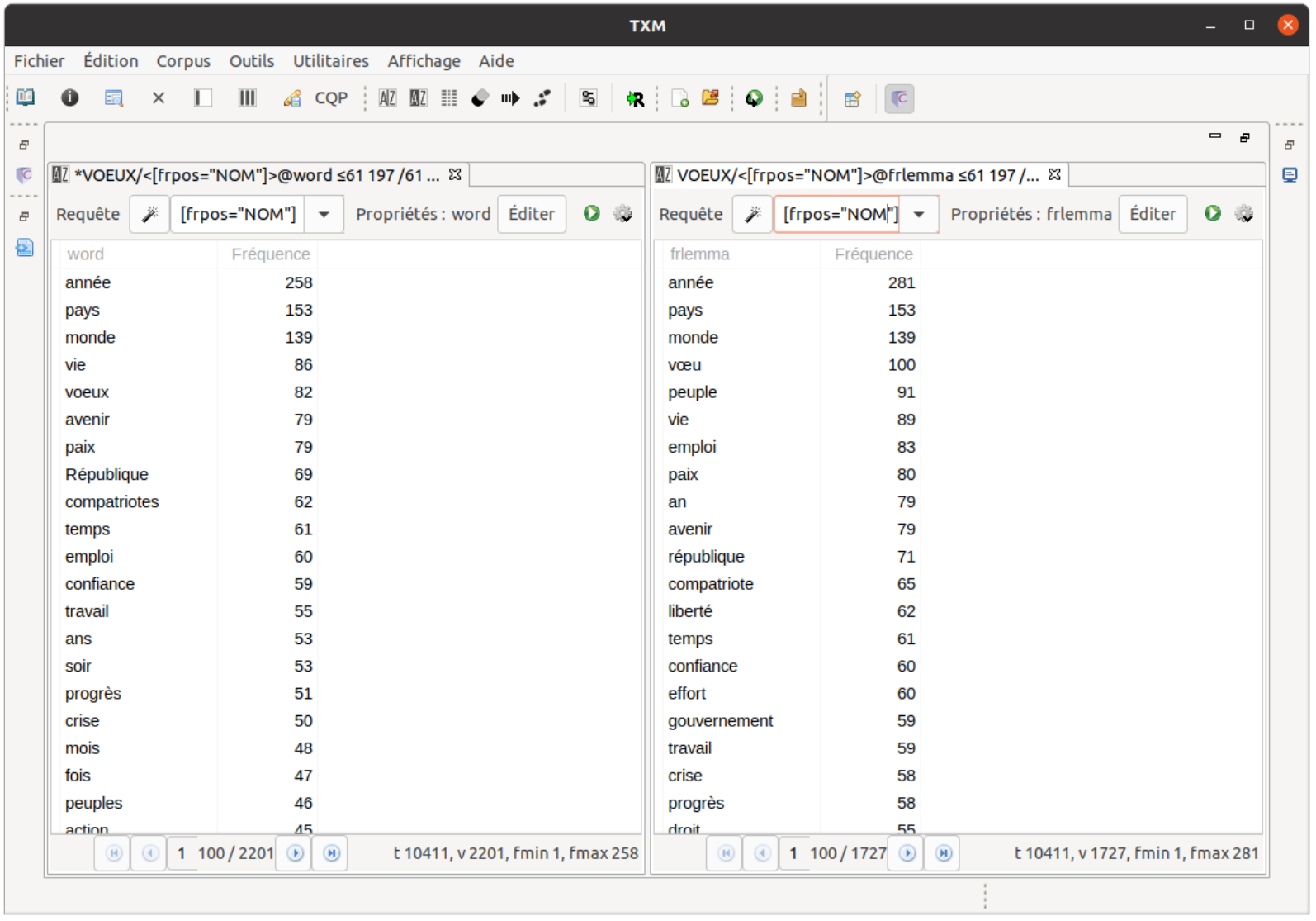
INDEX



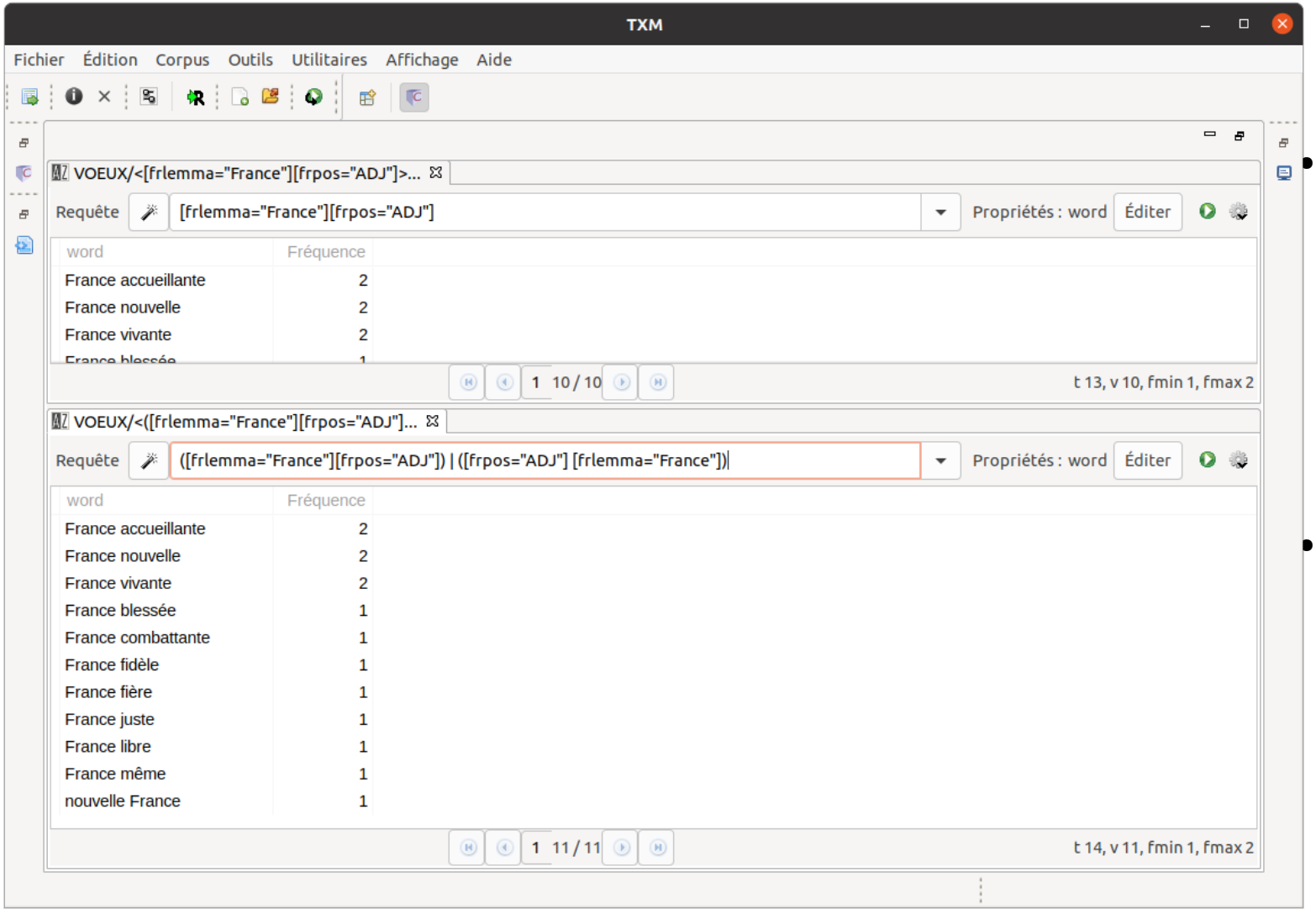
- L'INDEX fournit la liste des réalisations d'une requête, avec leur fréquence.
- Exemple pour un mot

INDEX

- Exemple 2 : l'index d'une catégorie grammaticale, ici les noms communs



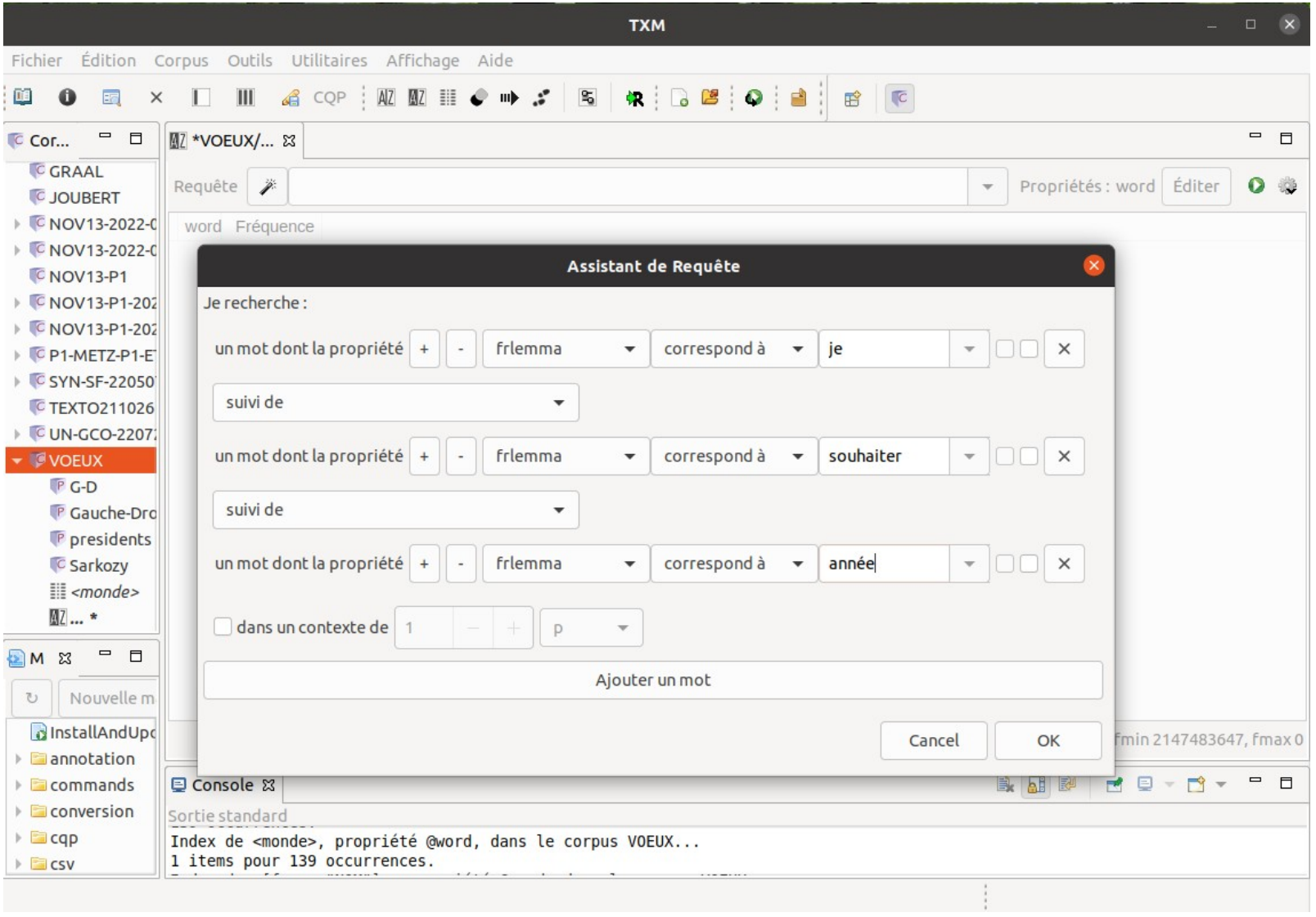
INDEX

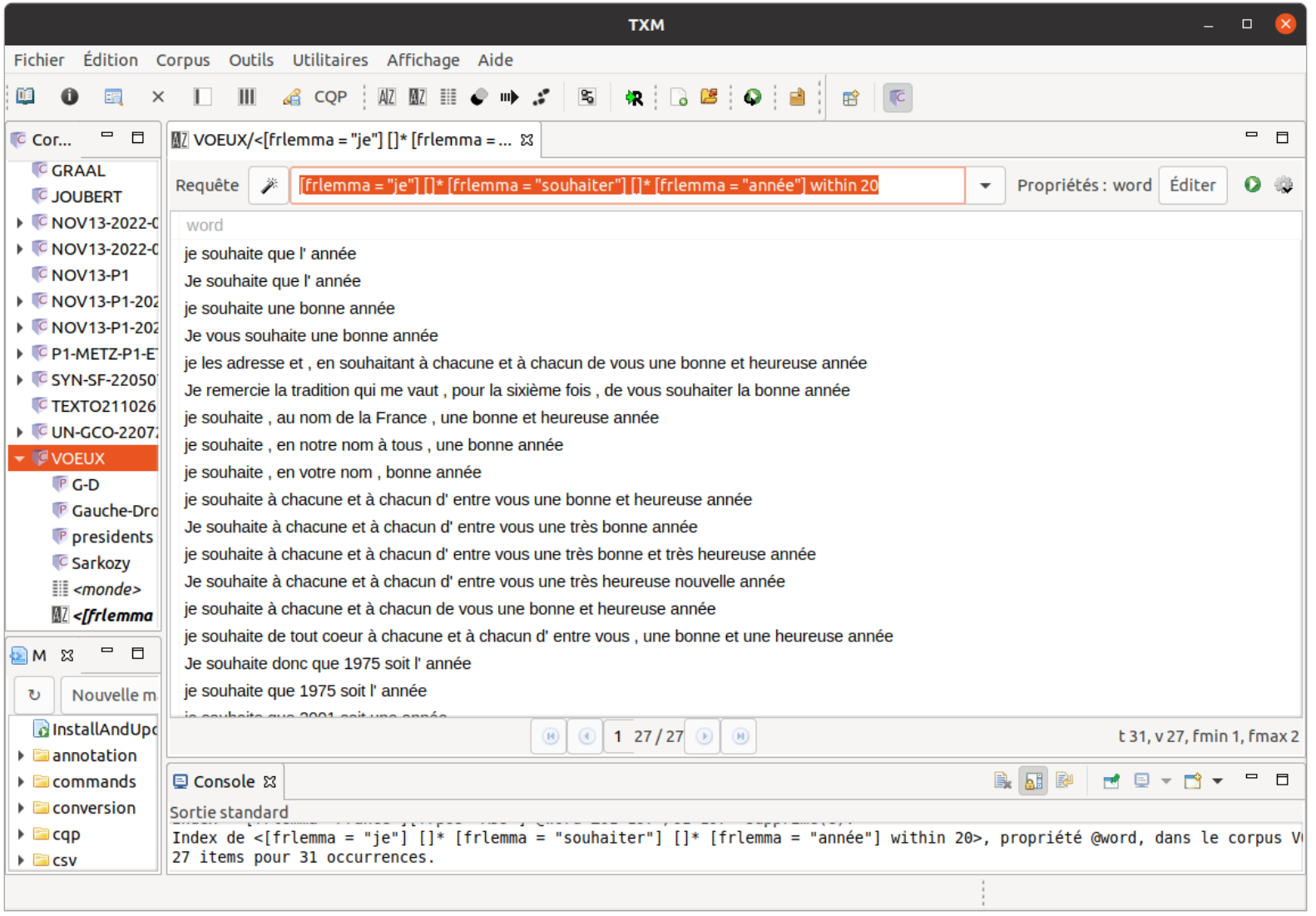


- Exemple 3 : analyse distributionnelle : dans mon corpus, quels adjectifs suivent le nom France ?
- Ou plus généralement, le qualifient ?

INDEX

- Un exemple illustrant la puissance du langage d'interrogation CQL : avec l'assistant de requête, on cherche le motif « je … souhaiter … année »

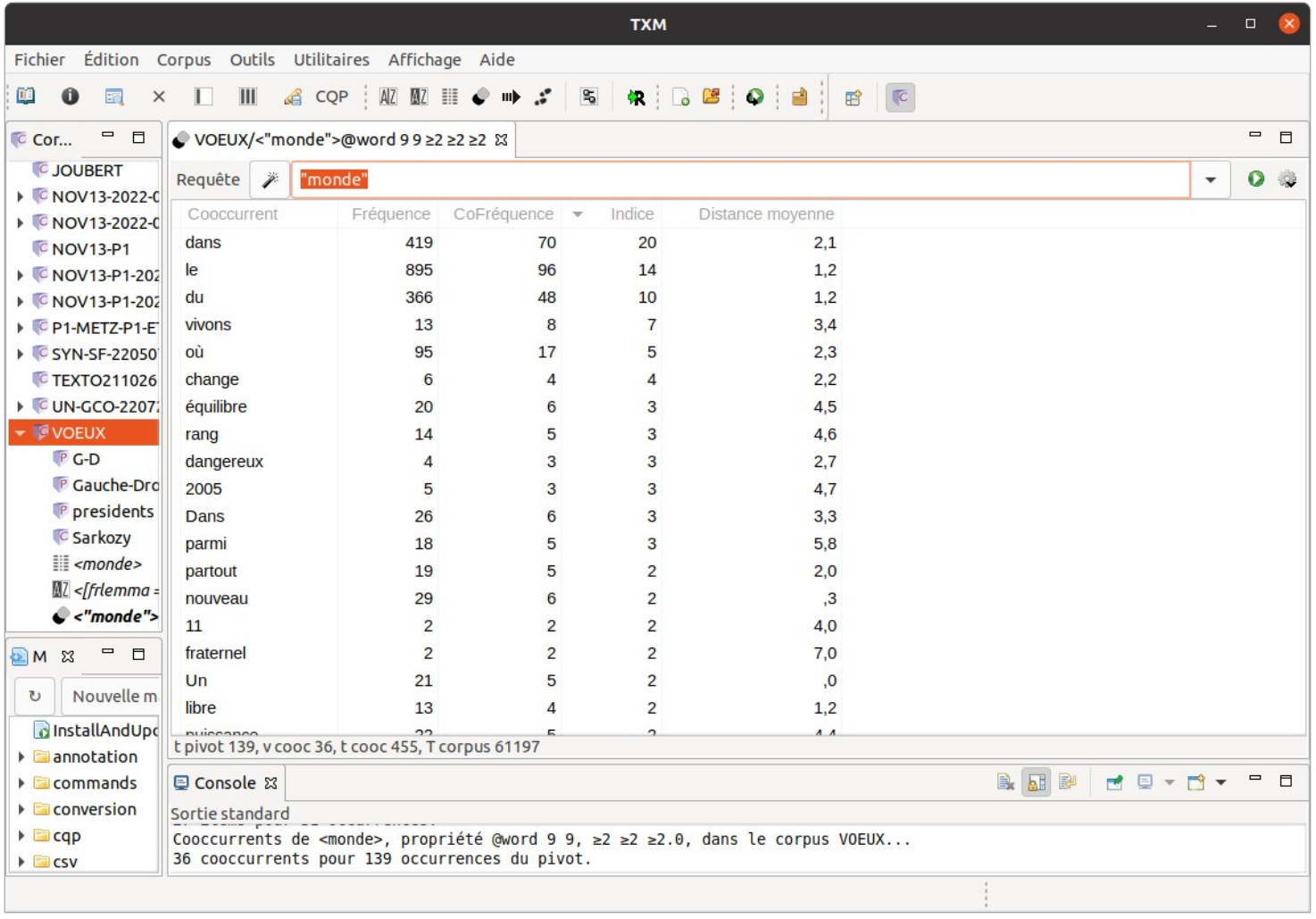




INDEX

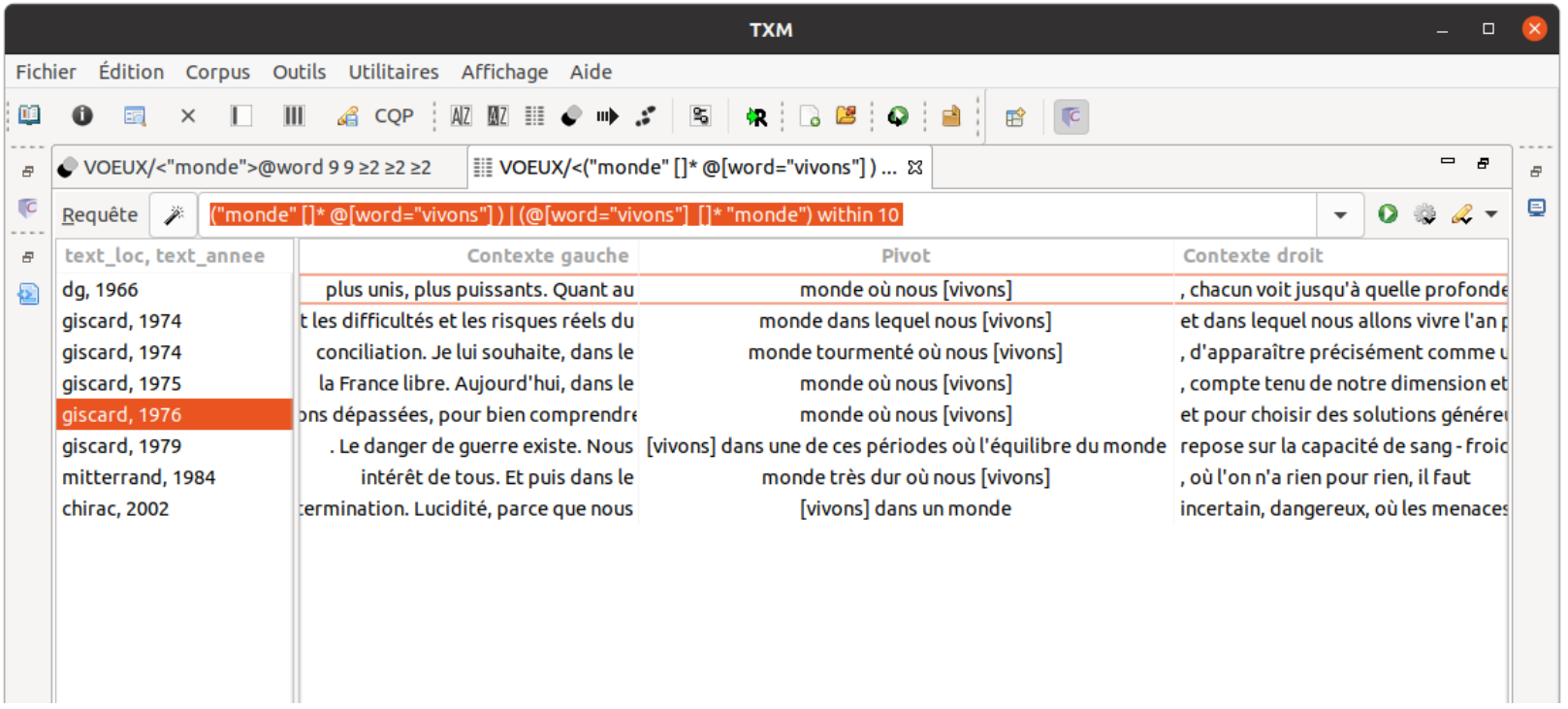
- Diversité des résultats trouvés pour le motif « je … souhaiter … année »

COOCCURRENCE



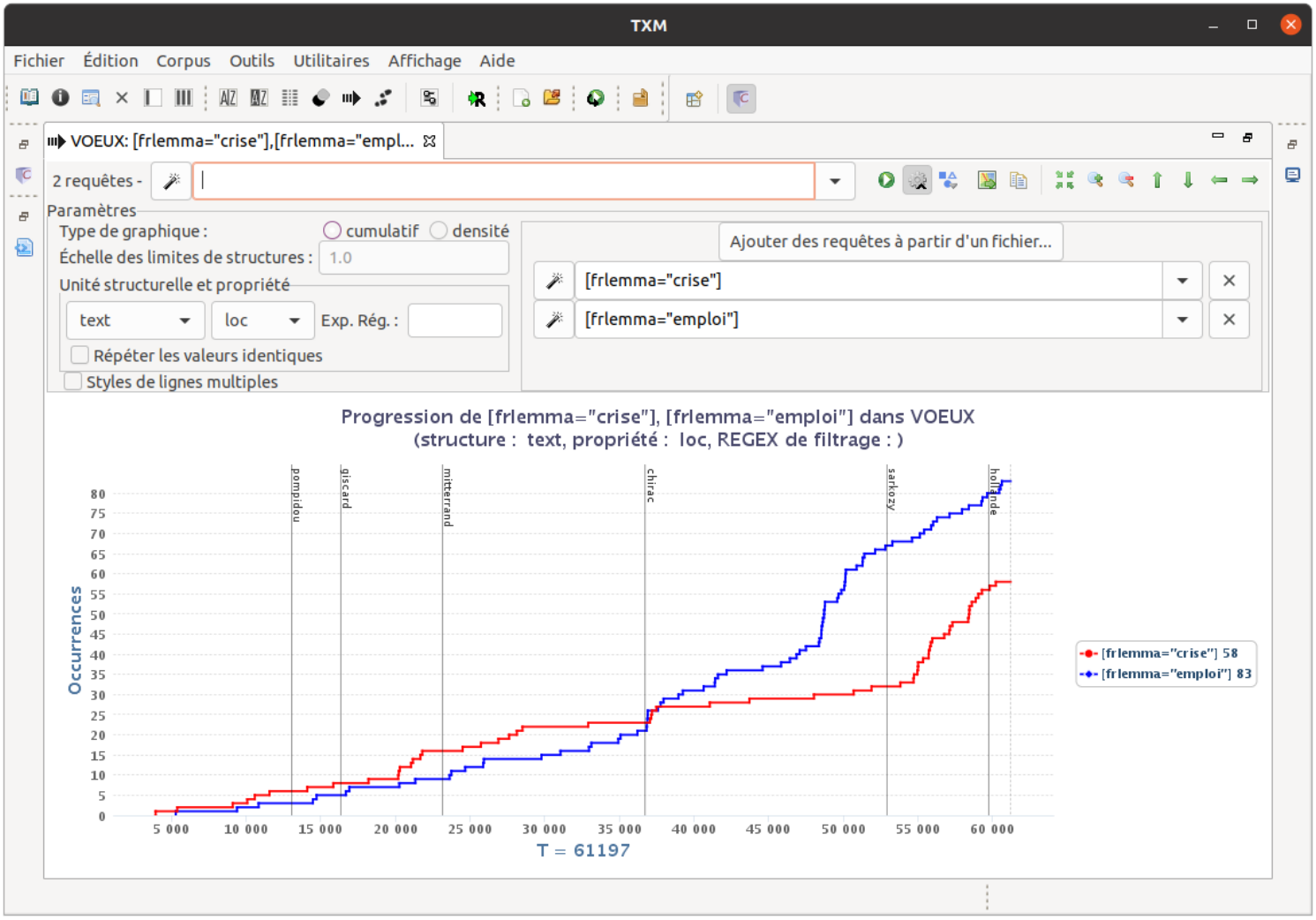
- Pour un mot ou motif donné, recherche des mots qu'il attire (mots statistiquement sur-représentés dans son voisinage)

COOCCURRENCE



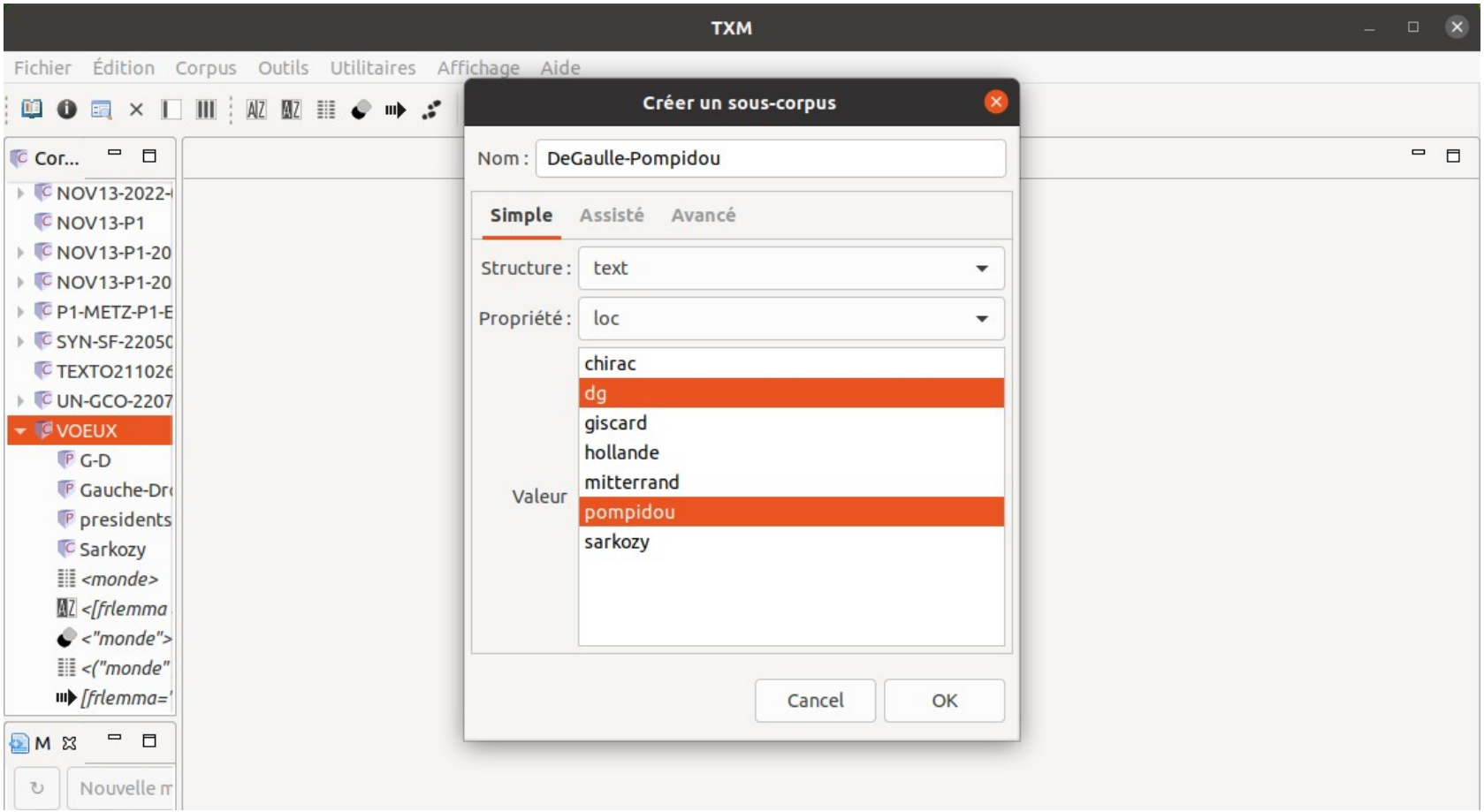
- Retour à l'ÉDITION, pour voir les cas de co-apparition, par double-clic sur une ligne du tableau des cooccurrents

PROGRESSION

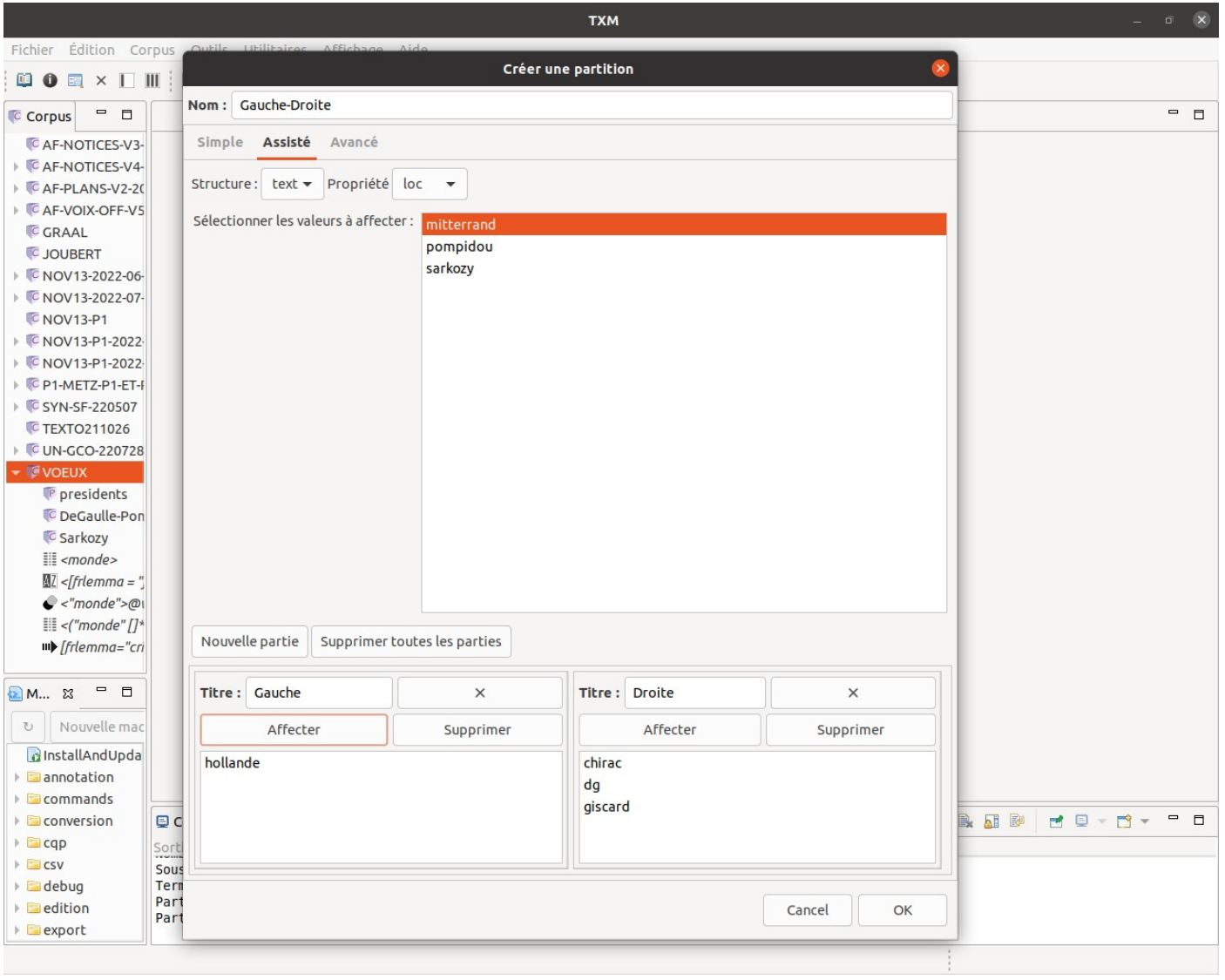


- On visualise la position des occurrences au fil du déroulement du corpus

SOUS-CORPUS



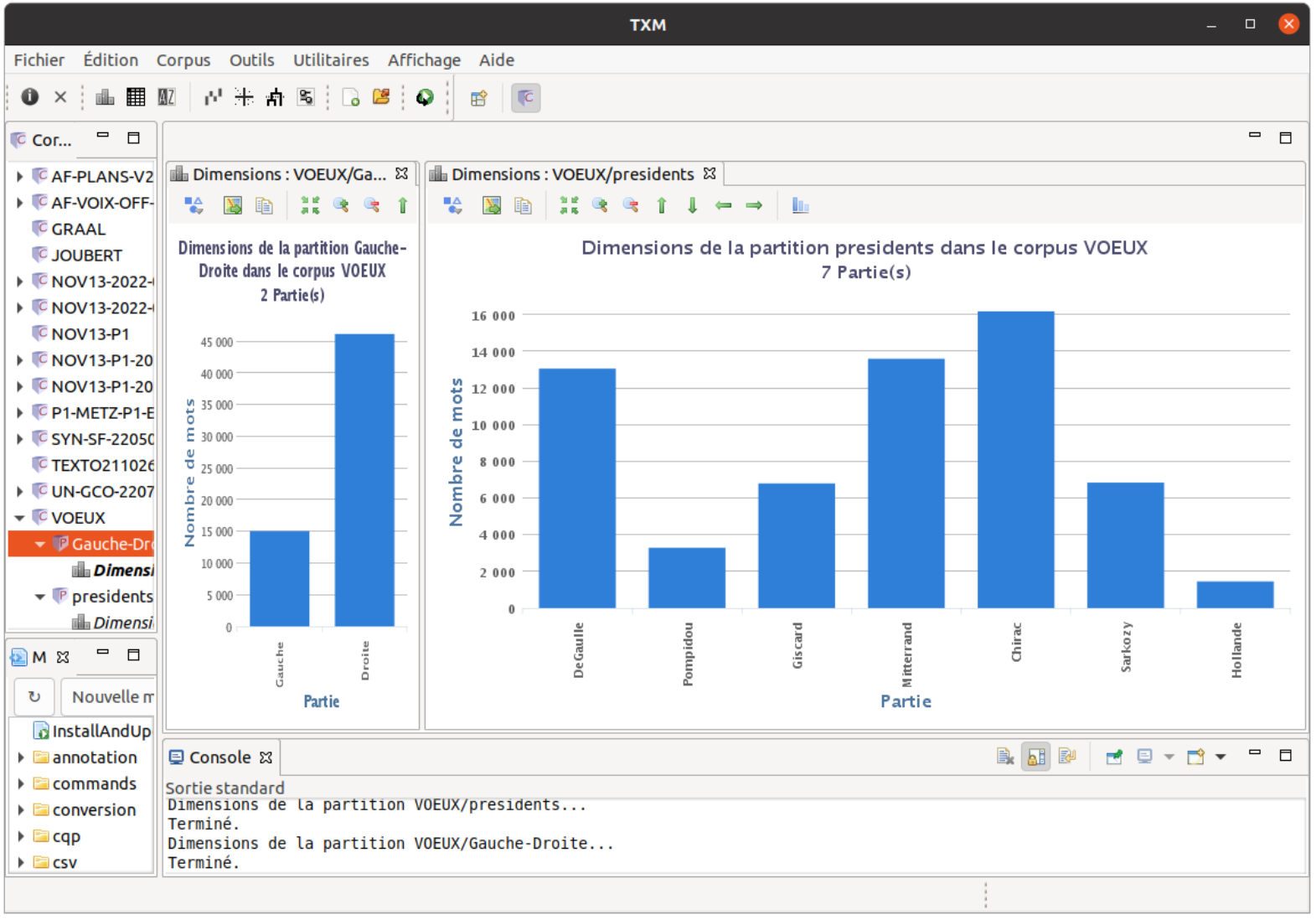
Possibilité de définir une sous-partie du corpus sur laquelle faire porter de nouvelles analyses



PARTITION

- Possibilité de diviser le corpus en un ensemble de parties :
par exemple les présidents, les années, les décennies, la gauche *vs* la droite, etc.

PARTITION Dimensions



Une fois la partition construite, on peut demander ses Propriétés et visualiser les tailles des différentes parties

INDEX sur une partition

- Les fréquences de chaque mot sont détaillées au niveau des parties
- Ici par exemple, le corpus est divisé en présidents et on considère les 300 noms les plus fréquents

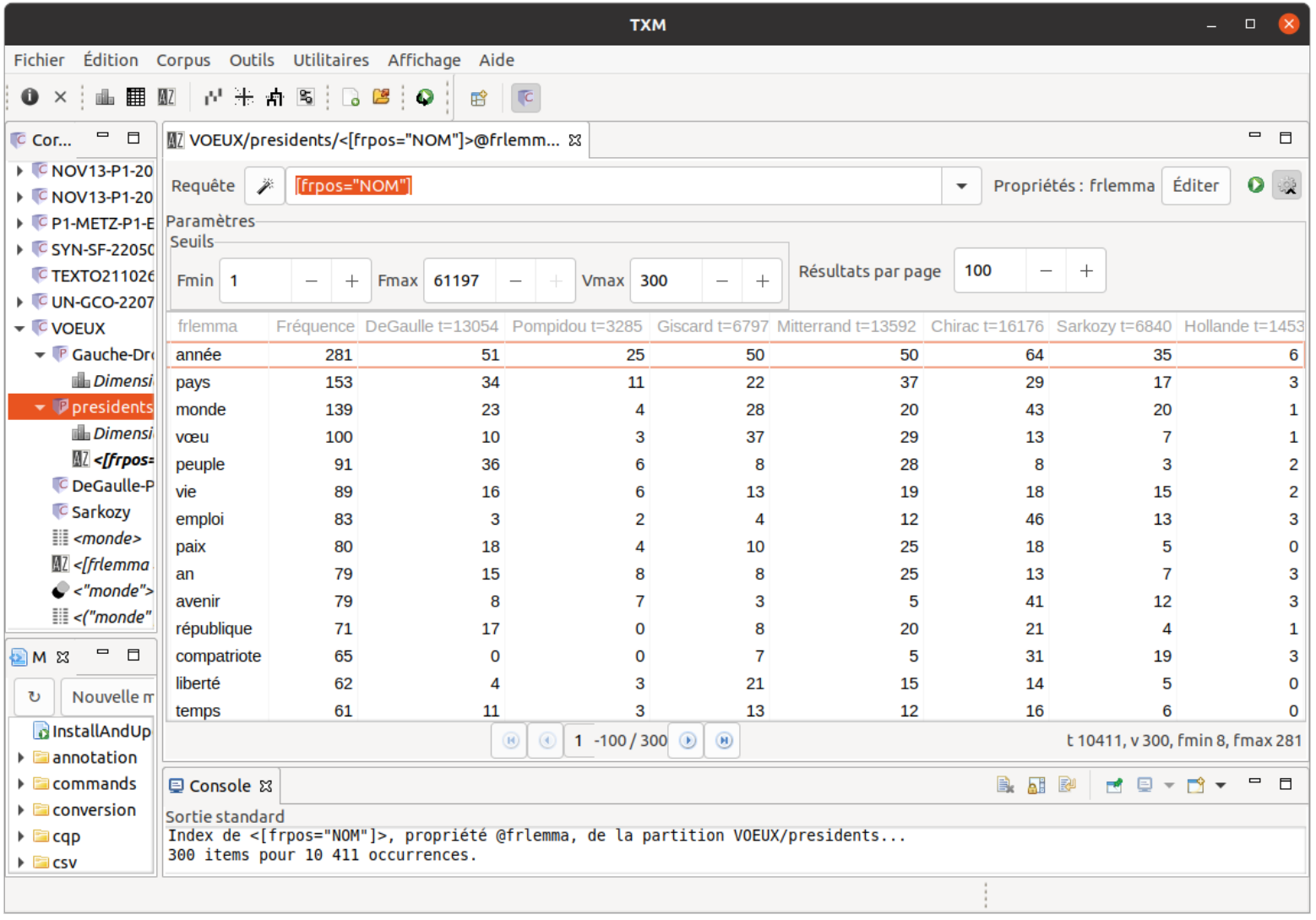
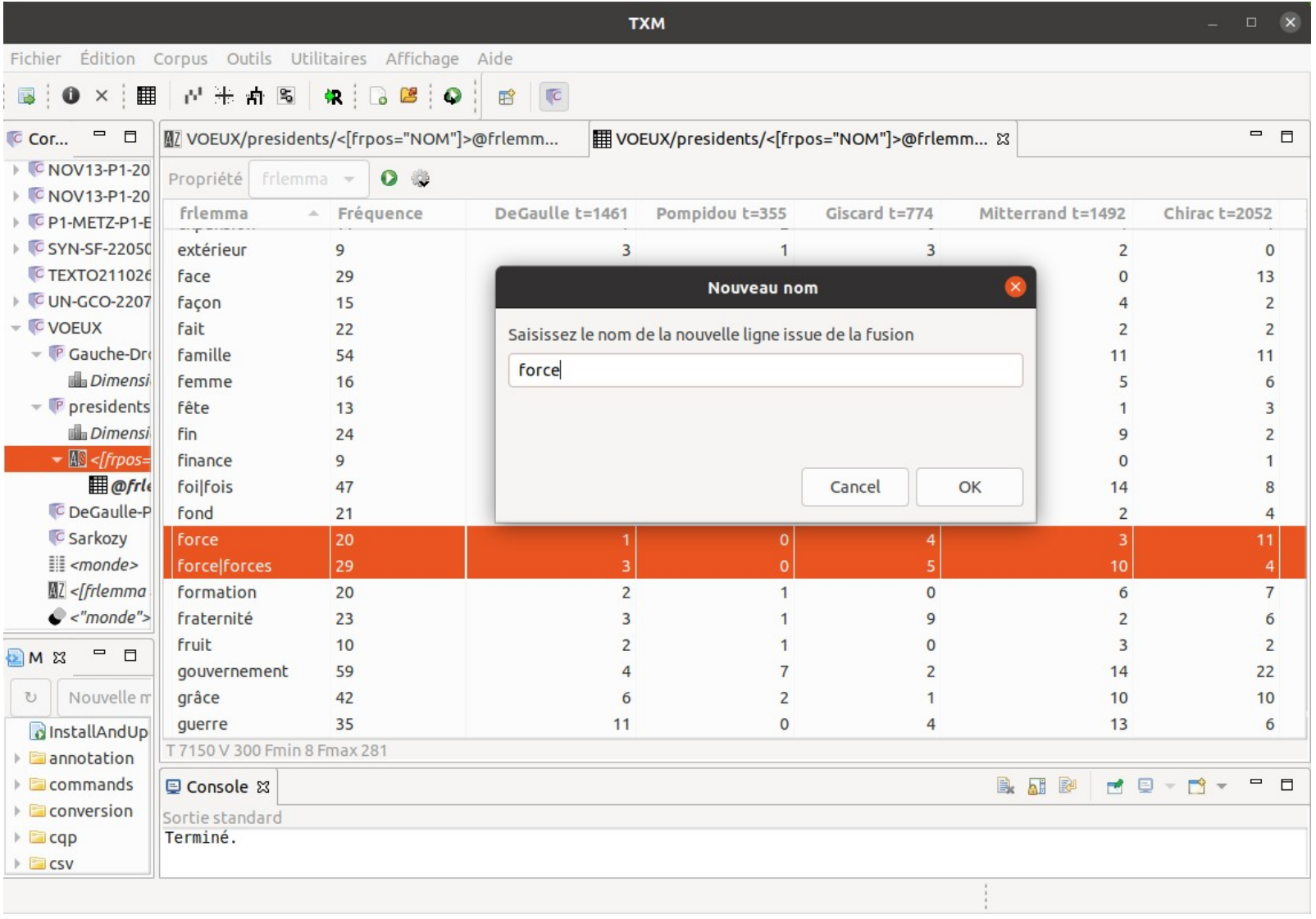
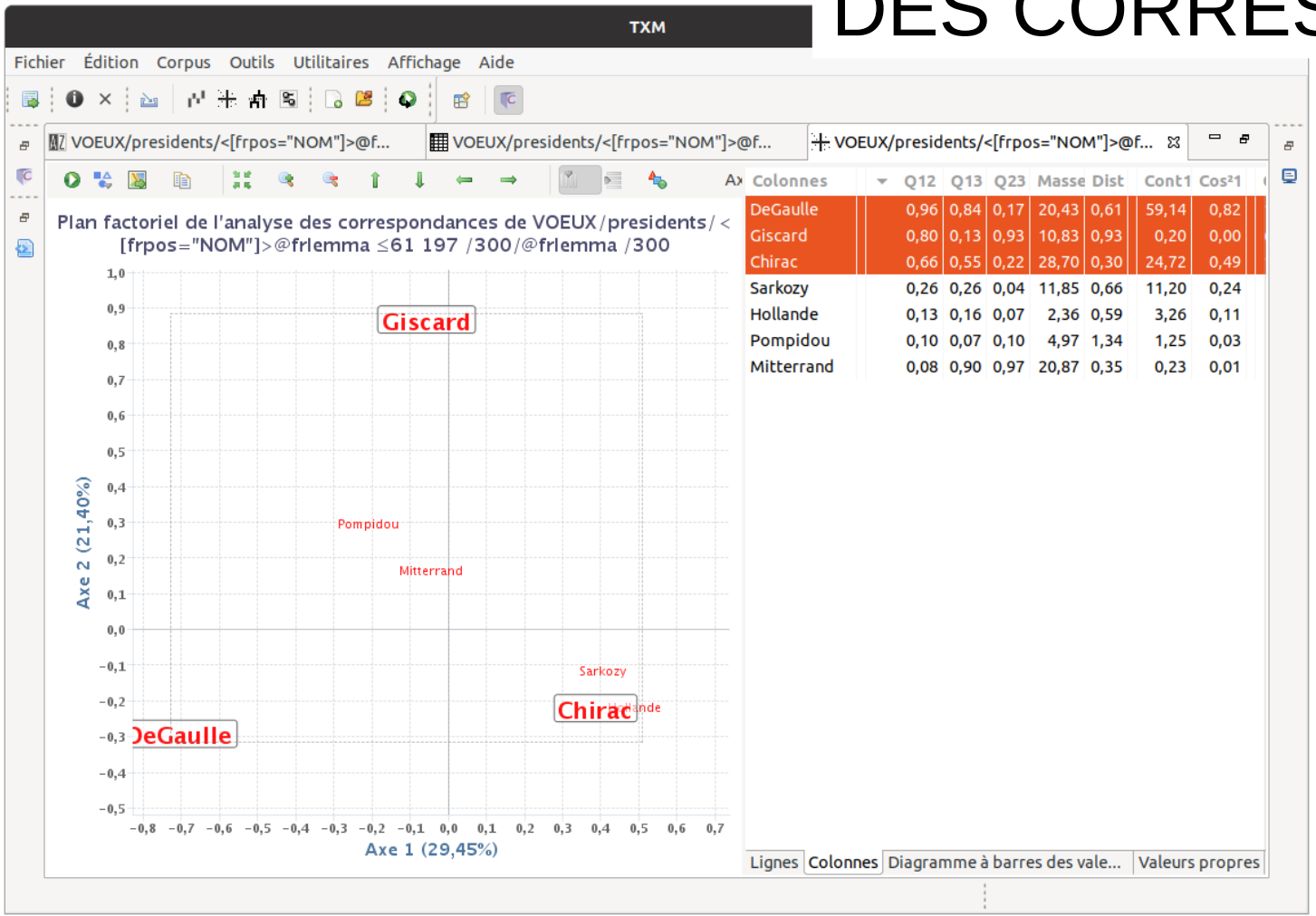


TABLE LEXICALE



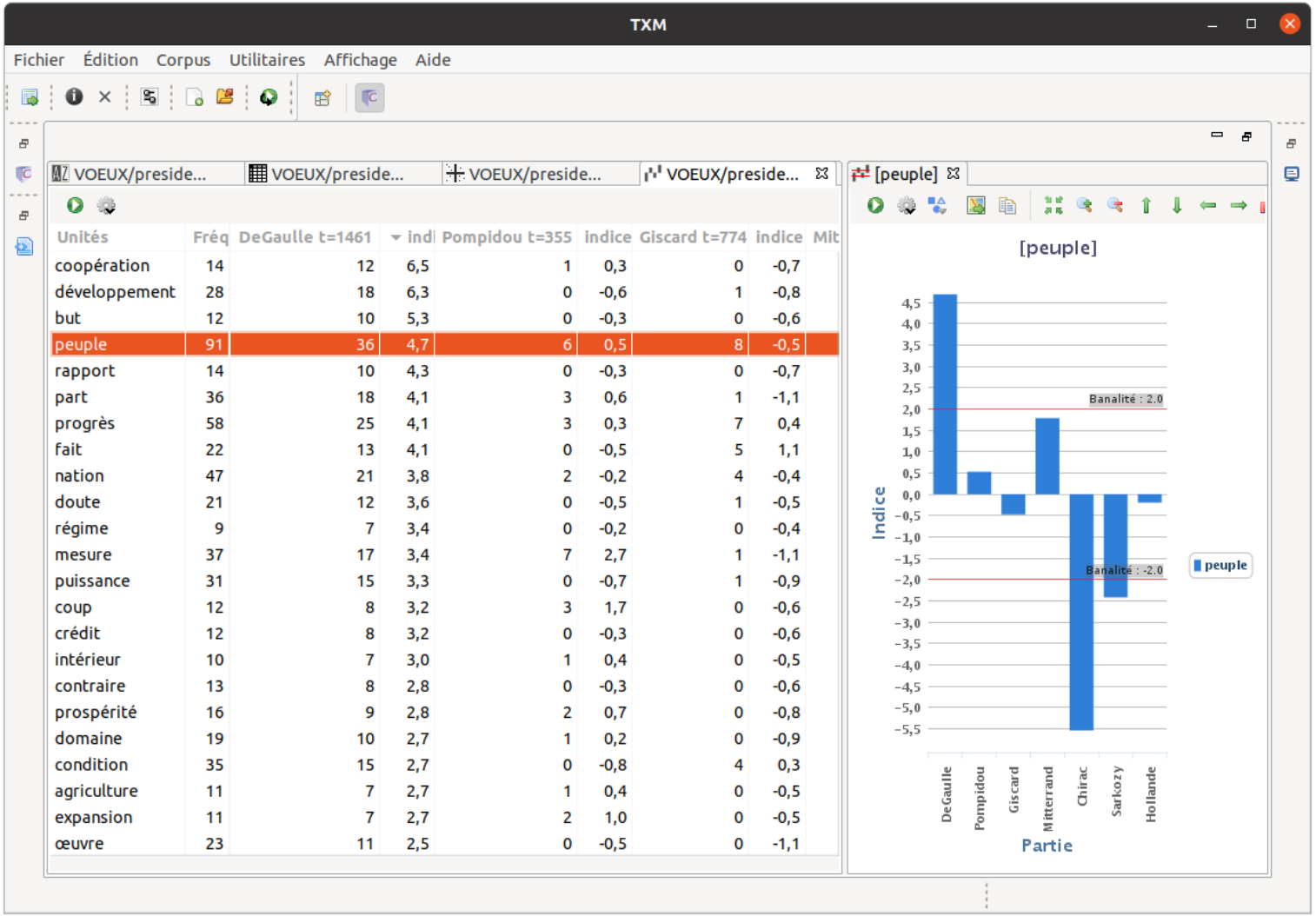
- Préparation d'un tableau croisant les mots et les parties pour des analyses statistiques
- On peut retoucher le tableau : supprimer ou fusionner des lignes

ANALYSE FACTORIELLE DES CORRESPONDANCES



- Pour une visualisation cartographique de la répartition des mots dans les parties, avec des indicateurs pour l'interprétation du graphique
- Le calcul trouve la représentation 2D qui préserve le mieux les contrastes

SPÉCIFICITÉS



Unités	Fréq	DeGaulle t=1461	indi	Pompidou t=355	indice	Giscard t=774	indice	Mit
coopération	14	12	6,5	1	0,3	0	-0,7	
développement	28	18	6,3	0	-0,6	1	-0,8	
but	12	10	5,3	0	-0,3	0	-0,6	
peuple	91	36	4,7	6	0,5	8	-0,5	
rapport	14	10	4,3	0	-0,3	0	-0,7	
part	36	18	4,1	3	0,6	1	-1,1	
progrès	58	25	4,1	3	0,3	7	0,4	
fait	22	13	4,1	0	-0,5	5	1,1	
nation	47	21	3,8	2	-0,2	4	-0,4	
doute	21	12	3,6	0	-0,5	1	-0,5	
régime	9	7	3,4	0	-0,2	0	-0,4	
mesure	37	17	3,4	7	2,7	1	-1,1	
puissance	31	15	3,3	0	-0,7	1	-0,9	
coup	12	8	3,2	3	1,7	0	-0,6	
crédit	12	8	3,2	0	-0,3	0	-0,6	
intérieur	10	7	3,0	1	0,4	0	-0,5	
contraire	13	8	2,8	0	-0,3	0	-0,6	
prospérité	16	9	2,8	2	0,7	0	-0,8	
domaine	19	10	2,7	1	0,2	0	-0,9	
condition	35	15	2,7	0	-0,8	4	0,3	
agriculture	11	7	2,7	1	0,4	0	-0,5	
expansion	11	7	2,7	2	1,0	0	-0,5	
œuvre	23	11	2,5	0	-0,5	0	-1,1	

Étude de la répartition des mots dans les parties :

- Repère les mots sur-employés ou sous-employés dans chaque partie (par rapport à une répartition au hasard)
- Ou réciproquement : diagramme des parties dans lesquelles un mot est sur- ou sous-représenté.

Plan

- Mini-démo : aperçu concret des types d’analyses disponibles dans TXM
- **Positionnement dans le paysage des outils d’exploration de corpus**
  - Origine et principes de conception de TXM
  - La textométrie par rapport à d’autres approches des corpus
  - Présentation générale de TXM
- Mise en évidence de quelques caractéristiques fortes de TXM
- Évocation de développements récents
- TXM pour quels points de vue, pour quels usages ? Des forces et limites dessinant un profil

Historique

- Projet ANR 2007-2010 « Fédération des recherches et développements en textométrie autour de la création d'une plateforme logicielle ouverte »
- Open-source :
  - mutualiser et partager le développement
  - transparence utile scientifiquement
- Corpus annotés et structurés, notamment TEI
  - pas que TXT - ne pas appauvrir/limiter les nouveaux corpus
  - pouvoir aller plus loin qu'un ajustement logiciel a posteriori aux corpus étiquetés

Textométrie

- Analyse quantitative ~ text mining, distant reading, avec des méthodes statistiques de référence :
  - Calcul de « keywords » avec les **Spécificités** (test exact de Fisher) : pas d'hypothèse de distribution, pas de limite de validité, modélisation compréhensible, notation efficace
  - **AFC** (analyse factorielle des correspondances) : adaptées aux tables de contingence
- Analyse qualitative ~ close reading
  - **Concordance** (KWIC) : avec références de localisation, une ligne par résultat, superposition des pivots, contextes triables
  - Le **retour au texte** : texte source central, accessible en lien avec les résultats de traitement
- La contruction d'un parcours (*vs*. automatisme) :
  - pas un ou quelques résultats tout faits, le chercheur est impliqué, on ne fait pas « tourner » le logiciel
  - pas d'évaluation mesurable (*vs*. banc d'essais TAL)
  - pas de "preuve" (pas de modélisation statistique de la langue)

Une typologie des approches logicielles pour l'analyse de données textuelles, inspirée de Lejeune (2010, 2017)

Textométrie,
Lexicométrie

EXPLORER
le matériau empirique,
les données source

Concordanciers enrichis
avec analyse quantitative

Concordanciers KWIC
et moteurs de recherche
linguistiques

CALCULER
« bavard »,
non focalisé

Textométrie
avec annotation
ou catégories en
cours d'analyse

Analyse par marqueurs
linguistiques ou rhétoriques
±généraux prédéfinis
(dictionnaires, grammaires, indicateurs)

MONTRER
Qualitatif,
« muet » (focalisé)

Registres
Annotation qualitative
assistée basée
sur le corpus

Text Mining

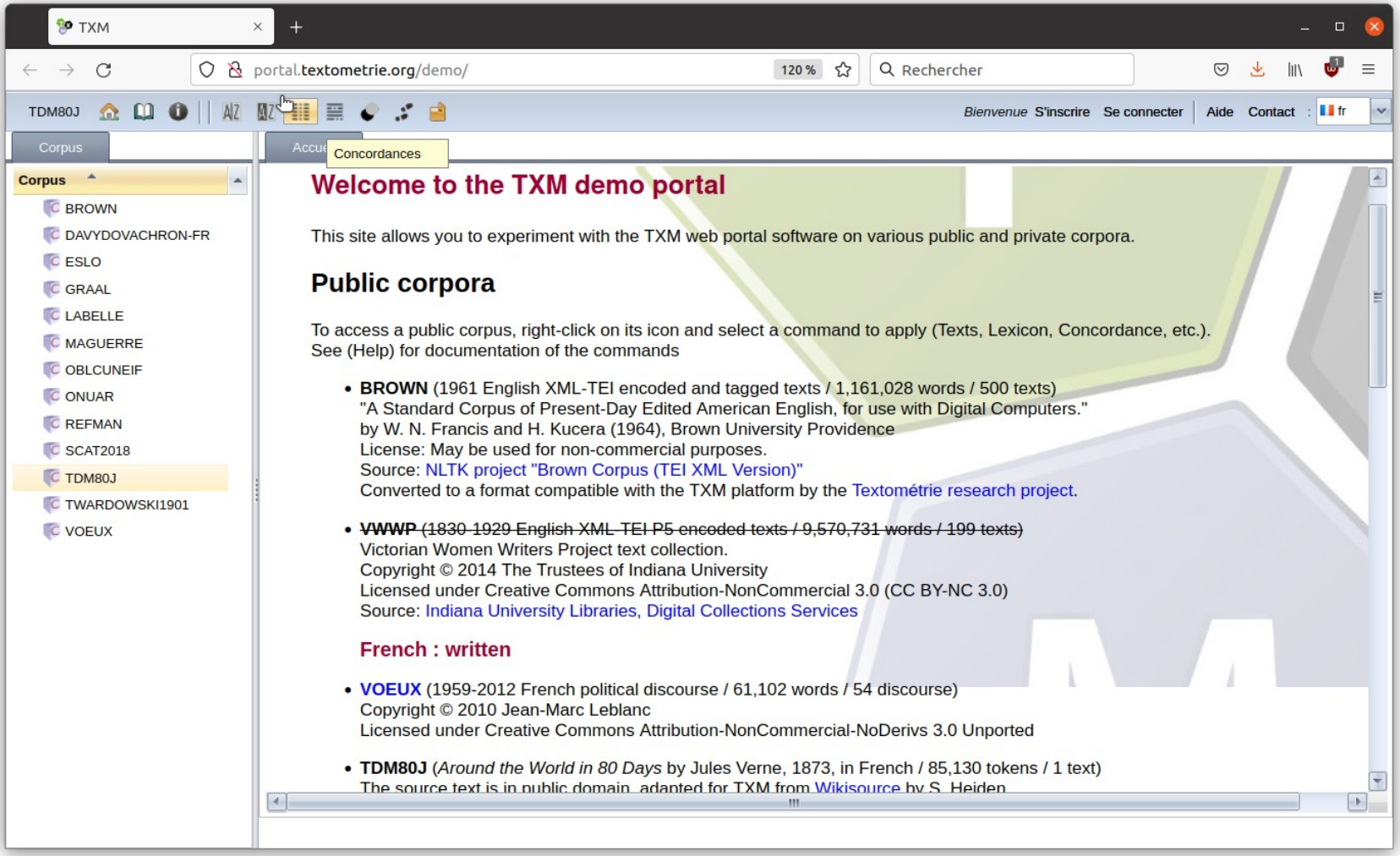
ANALYSER
En progressant à partir de
représentations à valeur-ajoutée
(catégories, annotation,
visualisations, synthèses)

CAQDAS
(Computer-assisted
qualitative data
analysis software)

Caractéristiques techniques de TXM (1)

- Version locale (Windows, Mac, Linux)
  - Plutôt plus de fonctionnalités (statistiques, assistant de requête, annotation…) → mais c'est en train d'évoluer
- Version en ligne (portail à installer, VM Huma-Num à administrer)
  - Portail Démo : des exemples de corpus, pas de corpus "de référence" général
  - Publication de corpus en ligne avec outils d'analyse (ex. BFM)
  - Possibilité de réglage fin des accès
  - Évite l'étape d'installation, notamment pour les cours

Portail Démo



Caractéristiques techniques de TXM (2)

- Open-source : gratuit mais aussi transparence, effort
- Composants spécialisés puissants : CQP (moteur de recherche), R (statistiques)
- Une interface qui permet d'enchaîner des calculs de façon interactive, de naviguer facilement entre eux
- Très grande souplesse et en même temps on ne peut pas tout faire : équilibre en terme de technicité/formalisme, ex. :
  - Interrogation : des mots sans syntaxe formelle, et de l’assistant de requête, au langage CQL
  - Import : du copier/coller ou des répertoires de TXT ou DOCX, au XML

Caractéristiques techniques de TXM (3)

- Du côté de l’ouverture encore :
  - Utilitaires (macros) pour des traitements complémentaires
  - Extensions
    - Ajout de fonctionnalités ± spécialisées ou contextualisées dans des projets (ex. annotation URS / ANR Democrat)
    - Installation et paramétrage de TreeTagger avec validation de la licence
    - Fonctionnalités prototypes (beta)
  - Exports dans des formats ouverts et standards (csv, svg, png…)
    - ex. Construction de table lexicale dans TXM et un calcul d’analyse ou de visualisation dans un autre outil
  - Articulation avec R

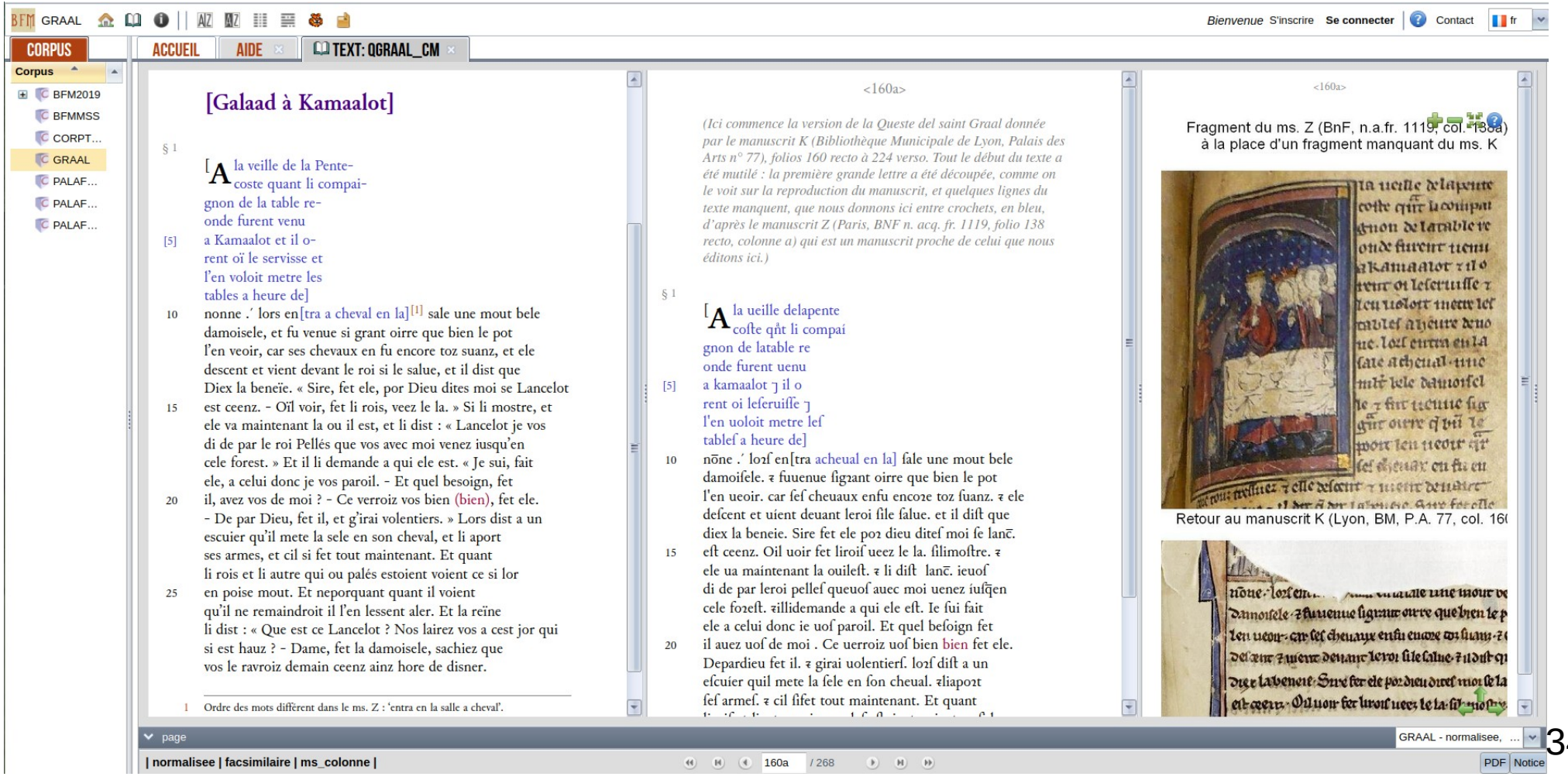
Plan

- Mini-démo : aperçu concret des types d’analyses disponibles dans TXM
- Positionnement dans le paysage des outils d’exploration de corpus (Textométrie, TXM)
- **Mise en évidence de quelques caractéristiques fortes de TXM**
  - Travailler sur des corpus riches ?
  - Fonctionnalités centrales & particularités/subtilités de TXM
- Évocation de développements récents
- TXM pour quels points de vue, pour quels usages ? Des forces et limites dessinant un profil

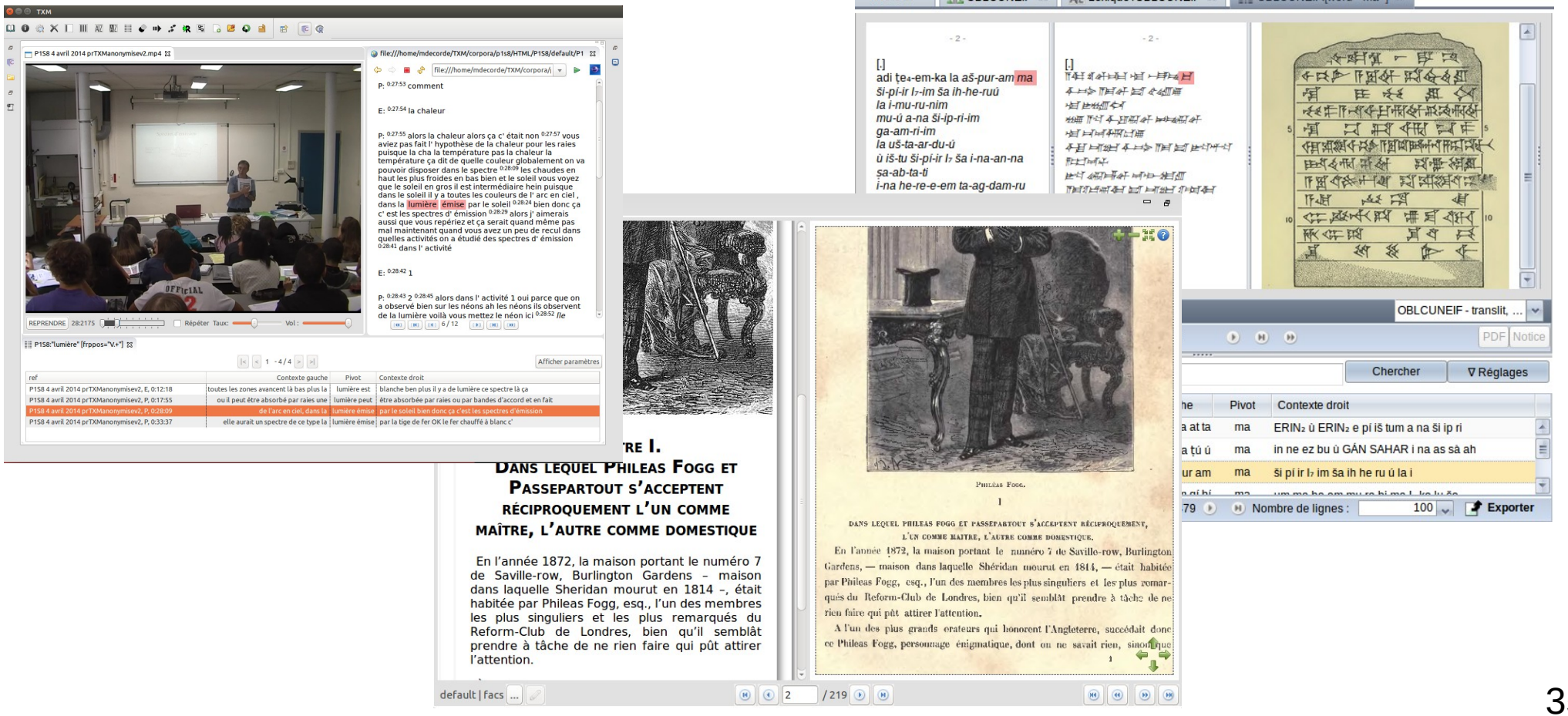
Travailler sur des corpus riches ?

- Interrogation et paramétrage des calculs
  - On peut utiliser les informations sur les mots ou sur les textes, les délimitations de parties ou de segments textuels (tours de parole, vers, etc.)
- Édition du texte
  - Mise en page HTML
  - Possibilité de texte affiché mais non indexé (n'entrant pas dans les recherches et les calculs)
- Construction dynamique des partitions
  - On peut même contraster des passages à l'intérieur des textes, ex. DD vs non DD
- Retour à la source : *facsimile* (ex. TDM80J) et multimedia

Exemple d'Édition avancée : le corpus GRAAL sur le portail BFM



D’autres exemples de retour au document source



Fonctionalités centrales & particularités/subtilités de TXM

- **Queries** : Recherche simple et puissance du langage d'interrogation CQL (une annotation morphosyntaxique peut être ajoutée à la volée au moment de l'import avec TreeTagger, ou livrée avec le corpus)
- ***word list*** : INDEX + CQL et propriété d'analyse = analyse distributionnelle
- ***KWIC*** : CONCORDANCE qui sont davantage qu'un relevé de contextes, du fait de l'effet visuel de superposition par alignement et tri ; retour au texte pour contexte élargi ; localisations souples
- ***Text visualization*** : EDITION avec retour au document source, édition synoptique
- ***collocations*** : COOCCURRENCES avec réglage des contextes en nombre de mots ou en structure (ex. phrase, paragraphe -en fonction des informations codées dans le corpus)
- ***keywords*** : SPÉCIFICITÉS : le calcul (basé sur un test exact de Fisher) ; les affichages par point de vue partie ou mot ; la possibilité de construire son tableau ; une macro pour calcul direct avec 4 paramètres, hors tableau.
- ***maps*** : AFC, avec graphique interactif et tableau d'aide à l'interprétation ; possibilité de construire son tableau.

Plan

- Mini-démo : aperçu concret des types d’analyses disponibles dans TXM
- Positionnement dans le paysage des outils d’exploration de corpus (Textométrie, TXM)
- Mise en évidence de quelques caractéristiques fortes de TXM
- **Évocation de développements récents**
  - L’annotation en cours d’analyse
  - Corpus d’entretiens et source multimédia
  - Illustration dans le contexte du projet ANR Antract
  - Mention : corpus annotés en syntaxe ; cas des corpus parallèles
- TXM pour quels points de vue, pour quels usages ? Des forces et limites dessinant un profil

Annotation en cours d’analyse

- Corriger ou ajouter des informations dans le corpus après l’import
- « Corpus mutable » - complexe car impacte toute la représentation du corpus et les index
- Plusieurs solutions ont été développées dans le contexte de projets avec des besoins différents
- Voir :
  - Serge Heiden, « Annotation-based Digital Text Corpora Analysis within the TXM Platform », JADT 2018, p. 367-374, hal-02015898.
  - Manuel de TXM 0.8 : https://pages.textometrie.org/txm-manual/

3 types d’annotation dans TXM

Type	***Projets***	***Interface***	***Nature technique***	***Intérêts***	***Limites***
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)

Notre exemple dans le cadre d’Antract portera sur une annotation de type « CQP sur Mot ».

Corpus d’entretiens et source multimédia

- Le corpus peut être saisi manuellement via une interface graphique spécialisée (ex. Transcriber) ou transcrit automatiquement par reconnaissance automatique de la parole
- Un import basé sur le format **XML TRS**, qui permet de noter de façon claire des **informations et structurations** utiles :
  - Sections d’entretien, tours de parole
  - Événements (ex. commentaire type didascalie) ou indication du transcripteur (ex. incertitude, trocature)
  - Repérage temporel (time-code)
  - Rq. des travaux aussi à partir d’autres formats
    - L’expérimentation précurseure de Flora Badin et al. (revue Corpus, 2021)
    - en cours à Praxiling sur la projection de lignes de transcription en CQP
- Dans TXM, le retour au texte s’enrichit d’un **retour au document vidéo source** (installer l’extension **MediaPlayer**)
  - Parce que toute transcription est obligée d’opérer des choix, on ne note pas tout : contexte pour l’interprétation ;
  - Pour pouvoir contrôler s’il y a un doute d’erreur, particulièrement dans le cas de transcription automatique.
- Voir :
  - 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. halshs-02779055.

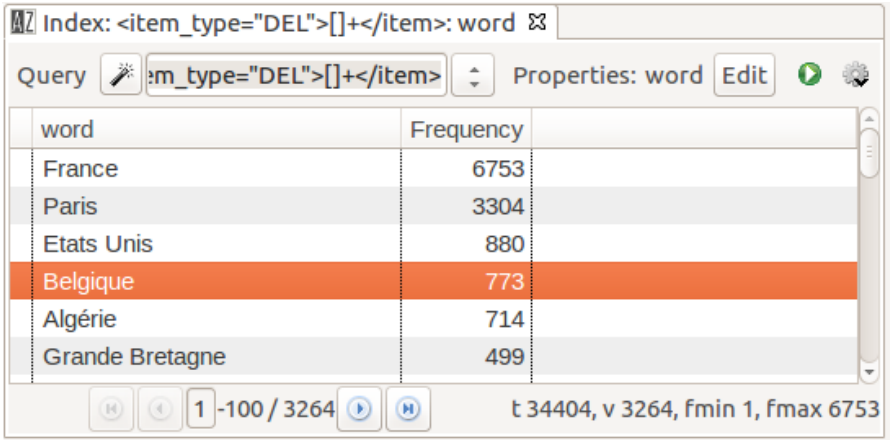
Illustration avec le projet ANR Antract

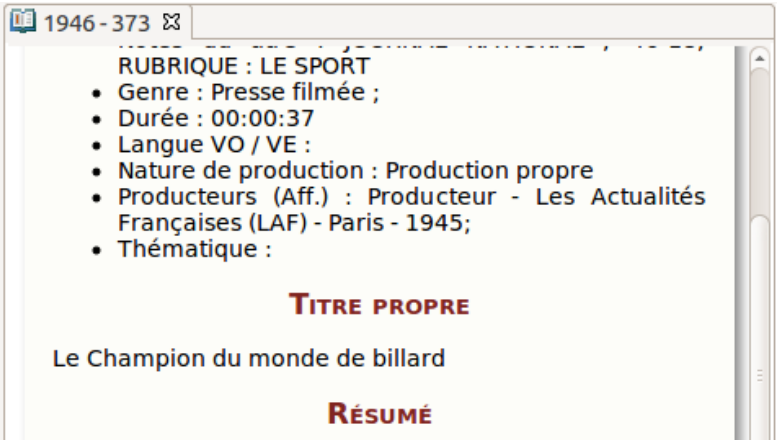
- Étude des *Actualités françaises* (1945-1968)
  - Actualités filmées diffusées dans les cinémas
  - Hebdomadaires, durée d'environ 10 mn
  - 1259 éditions, traitant en moyenne 8 sujets.

Le projet ANR Antract : partenaires

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

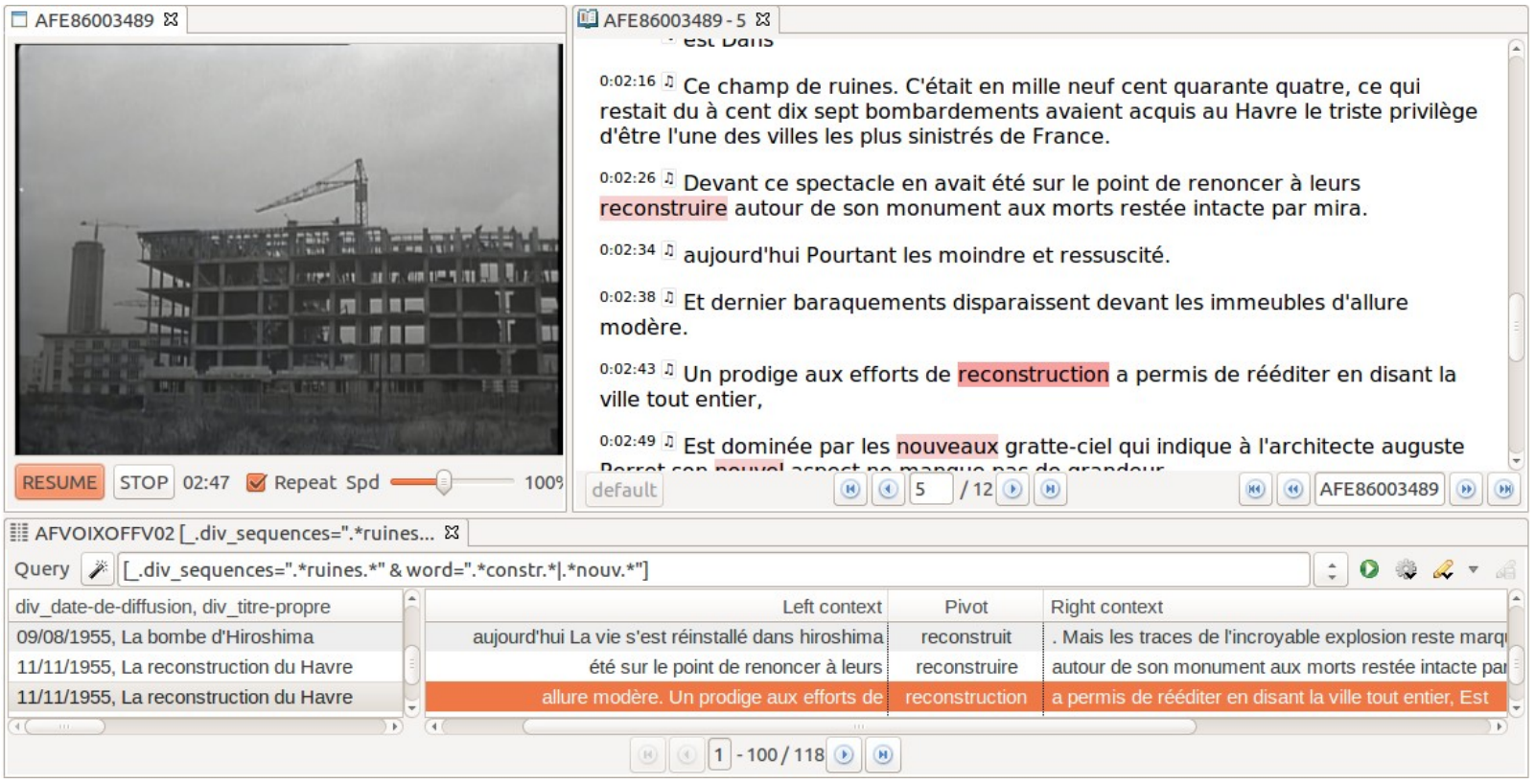
Corpus TXM AF-NOTICES





- 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.)

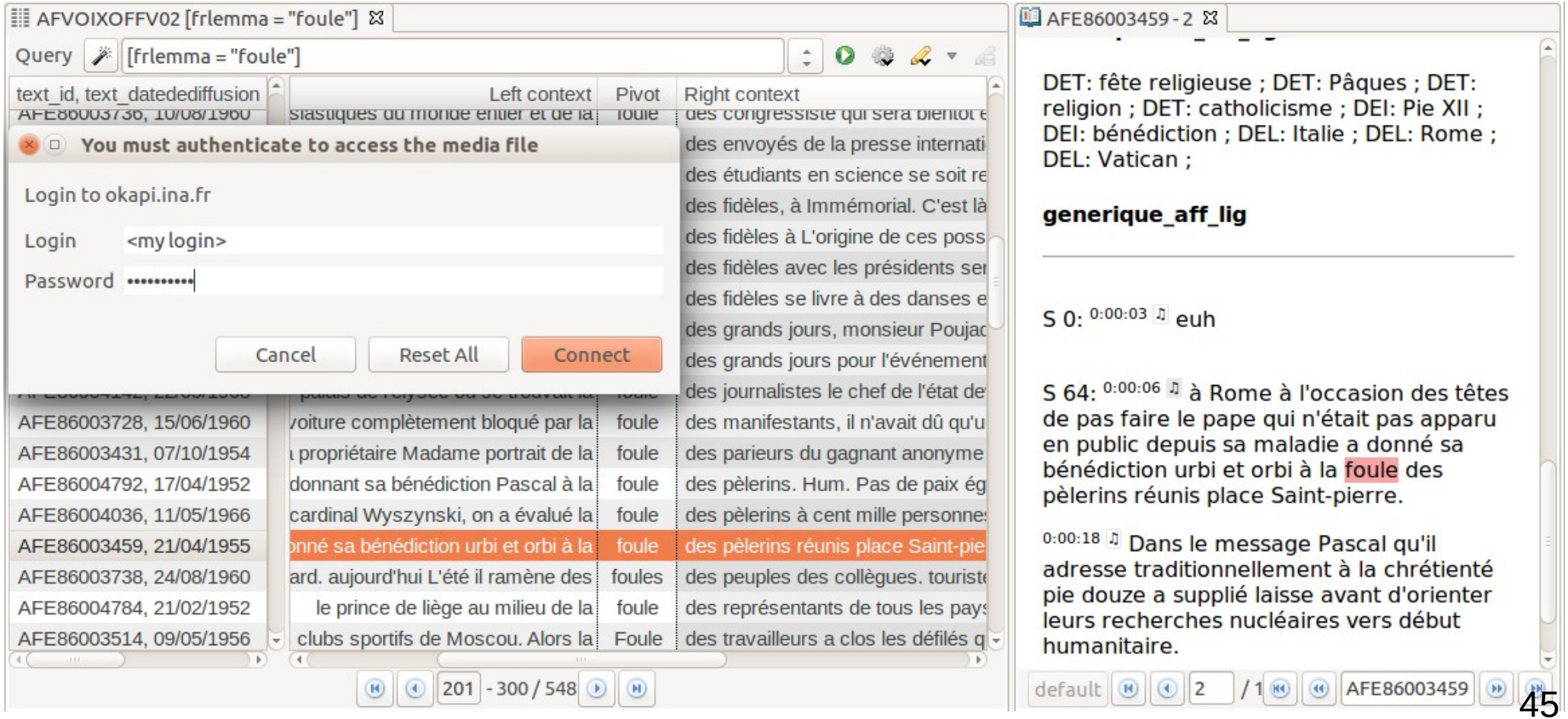
Corpus TXM AF-VOIX-OFF



- 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 opérées par la transcription.

Possibilité d’accès distant aux vidéos



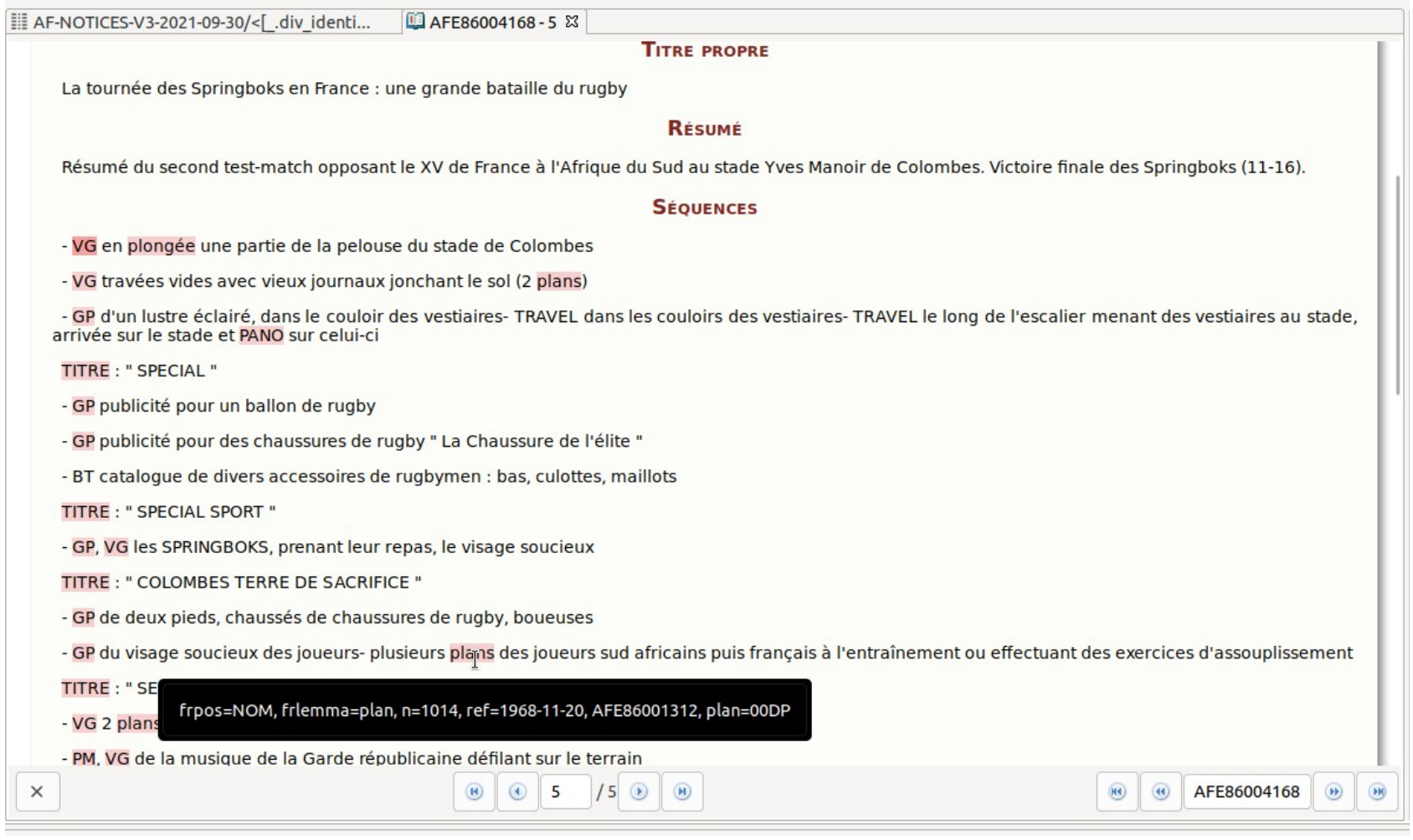
Fonctionnement du retour à la vidéo

- 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

Exemple pour l'annotation : Étude de la grammaire cinématographique

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

Vue du sujet « Rugby » dans TXM



Grammaire cinématographique : pistes pour travailler avec TXM

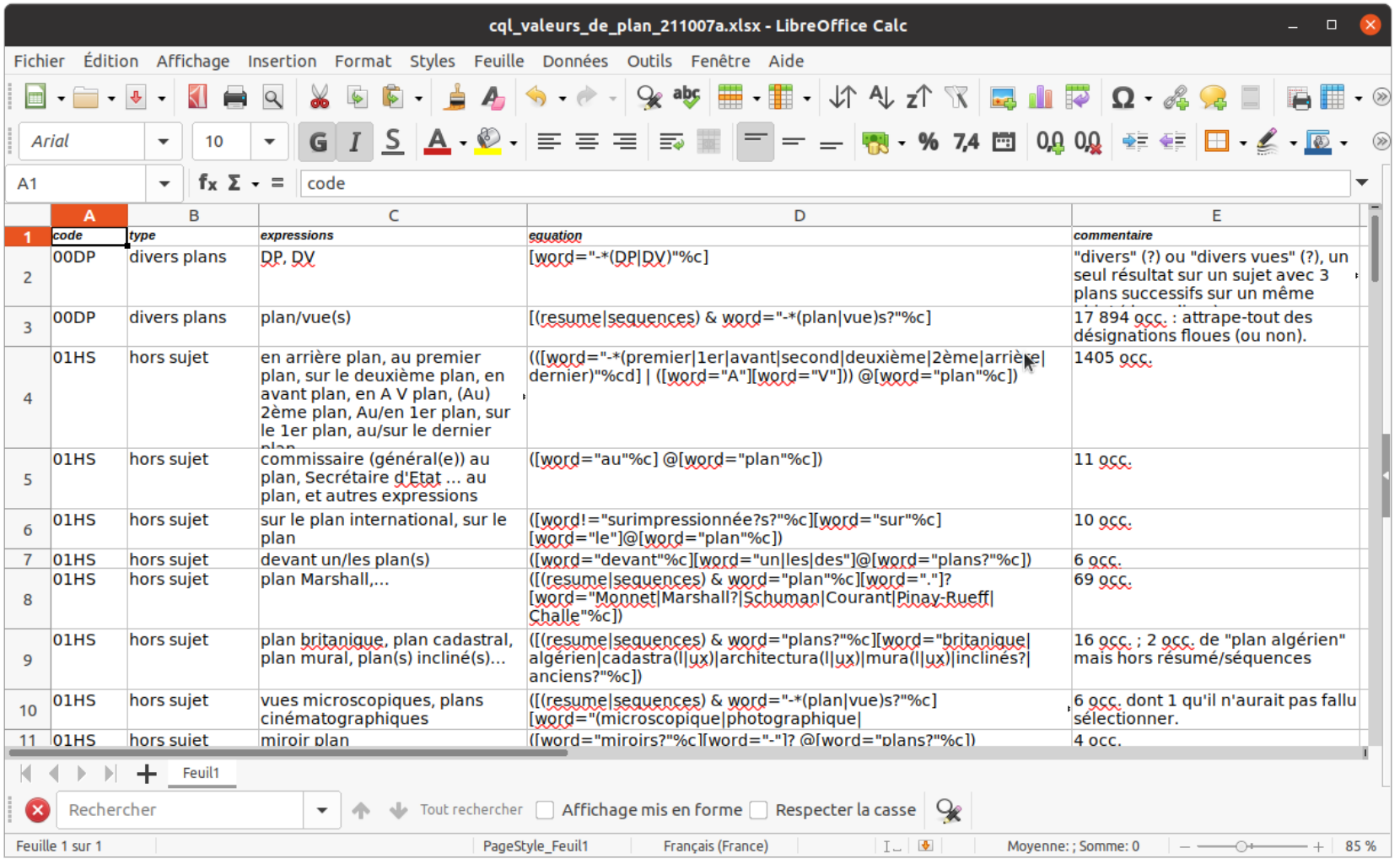
- Difficultés
  - Variabilité des désignations
    - Description libre
    - Pratiques ± partagées
  - Mots qui s’intercalent
    - Distance variable
    - Pris dans la requête
- Solutions
  - Annotation par des catégories d’analyse (recodage)
  - Projection
    - Un nouveau corpus avec uniquement les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

Annotation automatique par requêtes CQL

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

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

Annotation automatique par requêtes CQL



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

Annotation automatique par requêtes CQL

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

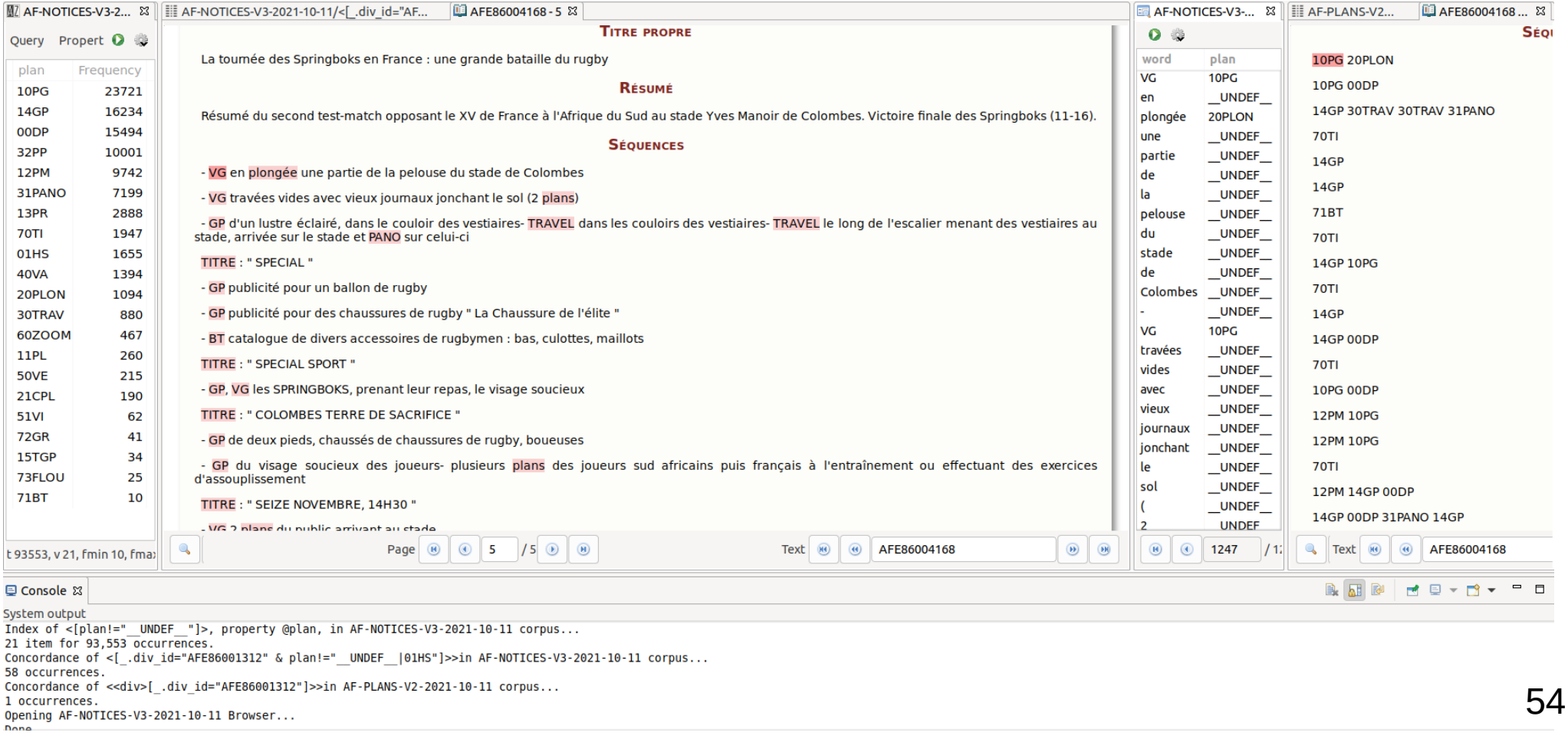
```
[frlemma="consulter|déplier|discuter|étudier|examiner|montrer|
regarder|tracer"%c]([]{0,3} [frpos=".*DET.*"%c])?
@[word="plans?"%c]
```

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

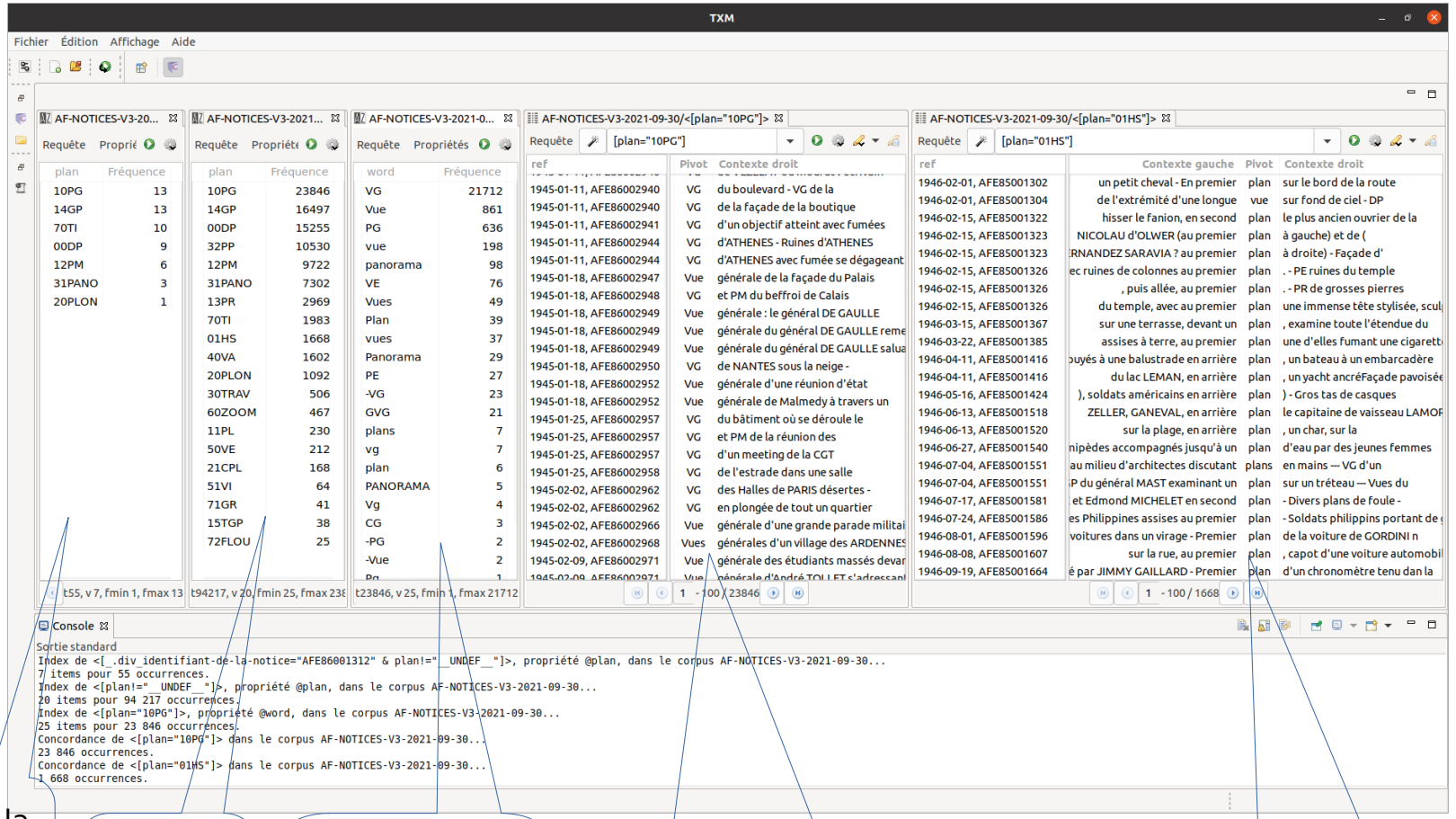
Annotation automatique par requêtes CQL

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

Corpus annoté



Vue des annotations



Dans la notice Rugby

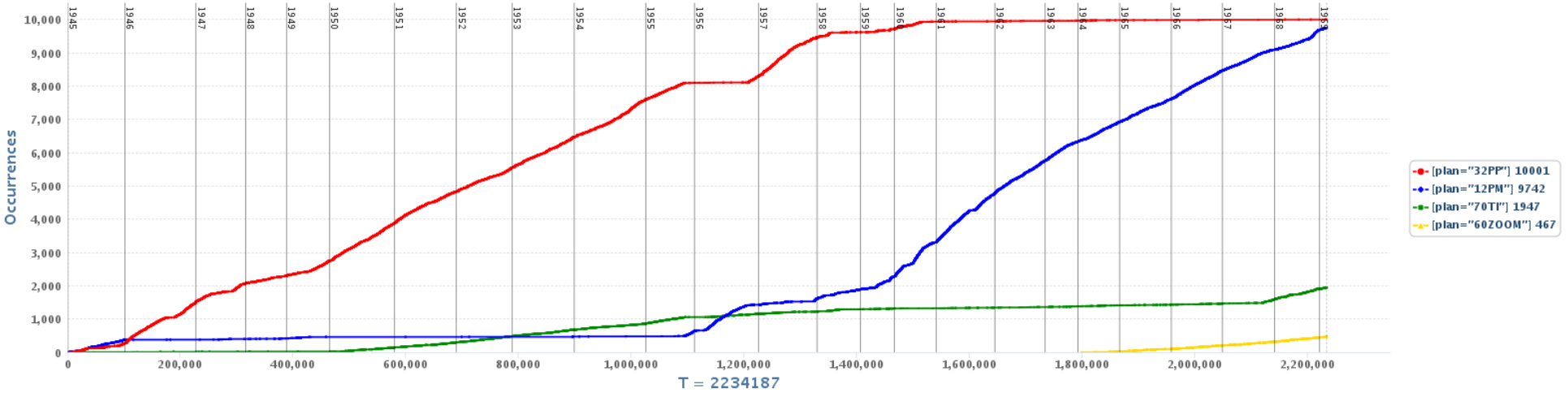
Dans tout le corpus

Les mots sous la catégorie « Plan général »…

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

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

Exemple d'analyse sur le corpus annoté

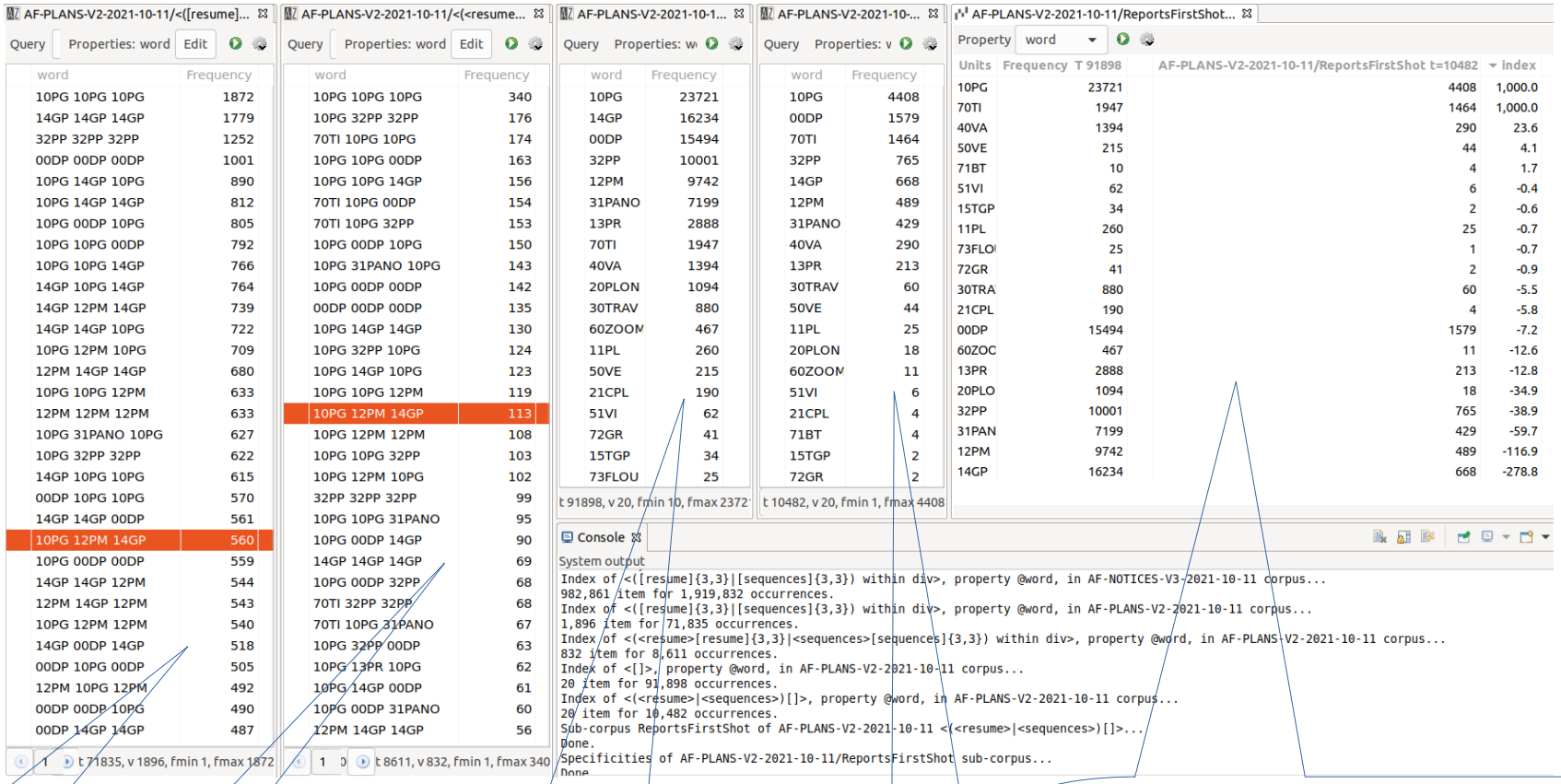


- 32PP (Plan porté, rouge) et 12PM (Plan moyen, bleu) montrent des profils complémentaires : Remplacement ? Équivalence ? Consigne de catalogage ?...
- 70TI (Titres, vert) : des périodes « magazine » où les titres semblent plus à la mode ?
- Apparition de 60ZOOM (Zoom, jaune) : nouveauté technique

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



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

SR de 3 plans en début de sujet

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

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

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

Annotation ou/et Projection : un outil méthodologique

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

TXM et les corpus annotés en syntaxe

- Import : pas d'analyse syntaxique à la volée
  - Tiger Search
  - annotations Tiger Search (a servi pour l'ajout de ponctuations)
  - Universal Dependancies
- Utilisation de l'annotation UD comme toute annotation CQP dans TXM (toutes fonctionnalités courantes)
- Pour TigerSearch :
  - affichage d'arbres syntaxiques
  - requêtes en INDEX et CONCORDANCES, concordances KNIC (Key Node in Context)
  - utilitaires : TIGER Summary, TIGER Index, TIGER Ratio, TIGER SVO Summary [cf. ANR Profiterole]
  - concordance KNIC dans le portail

TXM et les corpus parallèles alignés

- Import TMX
  - Format XML standard pour bases de traductions
- Analyse
  - Interrogations CQL croisées, ex.
    - [lemme="HIC"] :CorpusFRO [lemme="CIST"]
      - Occurrences du lemme HIC pour lesquelles on trouve le lemme CIST dans le passage aligné en ancien français.
    - [lemme="HIC"] :CorpusFRO ! [lemme="CIST"]
      - Occurrences du lemme HIC pour lesquelles on ne trouve pas le lemme CIST dans le passage aligné en ancien français.
- Chantier pas très actif actuellement (projets sur d’autres questions/données)
  - Mais des projets en cours, cf. présentation de Daniel Henkel cet après-midi
- Des outils plus spécialisés sur cette question :
  - cf. MkAlign/iTrameur cet après-midi

Plan

- Mini-démo : aperçu concret des types d’analyses disponibles dans TXM
- Positionnement dans le paysage des outils d’exploration de corpus (Textométrie, TXM)
- Mise en évidence de quelques caractéristiques fortes de TXM
- Évocation de développements récents
- **TXM pour quels points de vue, pour quels usages ? Des forces et limites dessinant un profil**

Caractéristiques d’une approche TXM

- Interface graphique
  - Moins souple que des briques logicielles ? (R en ligne de commande, notebook...)
  - Accompagne un parcours articulé, avec retour au texte
- Articulation possible avec d’autres outils
  - Excel, dtm-vic, IRaMuTeQ...

Caractéristiques d’une approche TXM

- Considérer des textes
  - Avec une structuration interne, des éléments de mise en forme, de contexte
  - Plus adapté à l’édition d’une œuvre qu’à une ressource massive de phrases ou de trigrammes
- Intertextualité et corpus de référence
  - Approche endogène plutôt que modèle externe (type Word2Vec)
  - Sensibilité aux doublons (bases de presse,…)

Caractéristiques d'une approche TXM

- Et… (cf. les quatre facettes de la textualité ci-dessous)
  - Langue : CQL (dont sensibilité à l'ordre des mots)
  - Construction d'un parcours interprétatif *vs* résultat

	Vision interne : texte *objet unique (objectivité relative)* *paradigme logico-grammatical*	*Vision externe : document* *contextes pluriels (subjectivité)* *paradigme rhétorico-herméneutique*
Domaine : *situé,* *cotexte (culturel,situationnel...)*	1. MATIÈRE LINGUISTIQUE	4. RÔLE CONSTITUTIF DE LA LECTURE
Domaine : *système,* *contexte (textuel)*	2. ORGANISATION INTERNE clôture et autonomie, linéarité, hiérarchie, orientation	3. INTERTEXTUALITÉ

Bénédicte Bommier-Pincemin, Thèse de doctorat de l'université Paris IV, 1999, chap. IV, p. 152.

Pour aller plus loin

- Site de TXM : https://textometrie.org
  - Point d'entrée principal, avec principaux accès aux autres sites liés dans l'encadré en haut à droite de l'accueil
- Communauté des utilisateurs : txm-users
  - Wiki : https://groupes.renater.fr/wiki/txm-users/index
  - Liste : https://groupes.renater.fr/sympa//info/txm-users/
- Adresse de contact de l'équipe TXM :
  - textometrie *à* groupes.renater.fr
  - Pour une question qui peut intéresser d'autres utilisateurs, préférer la liste txm-users