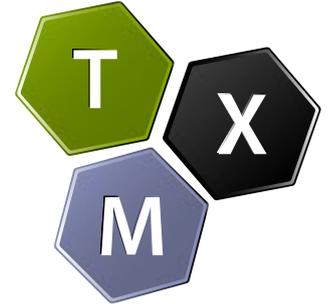


Présentation du logiciel TXM Un point de vue utilisateur



Bénédicte PINCEMIN

IHRIM, université de Lyon, CNRS



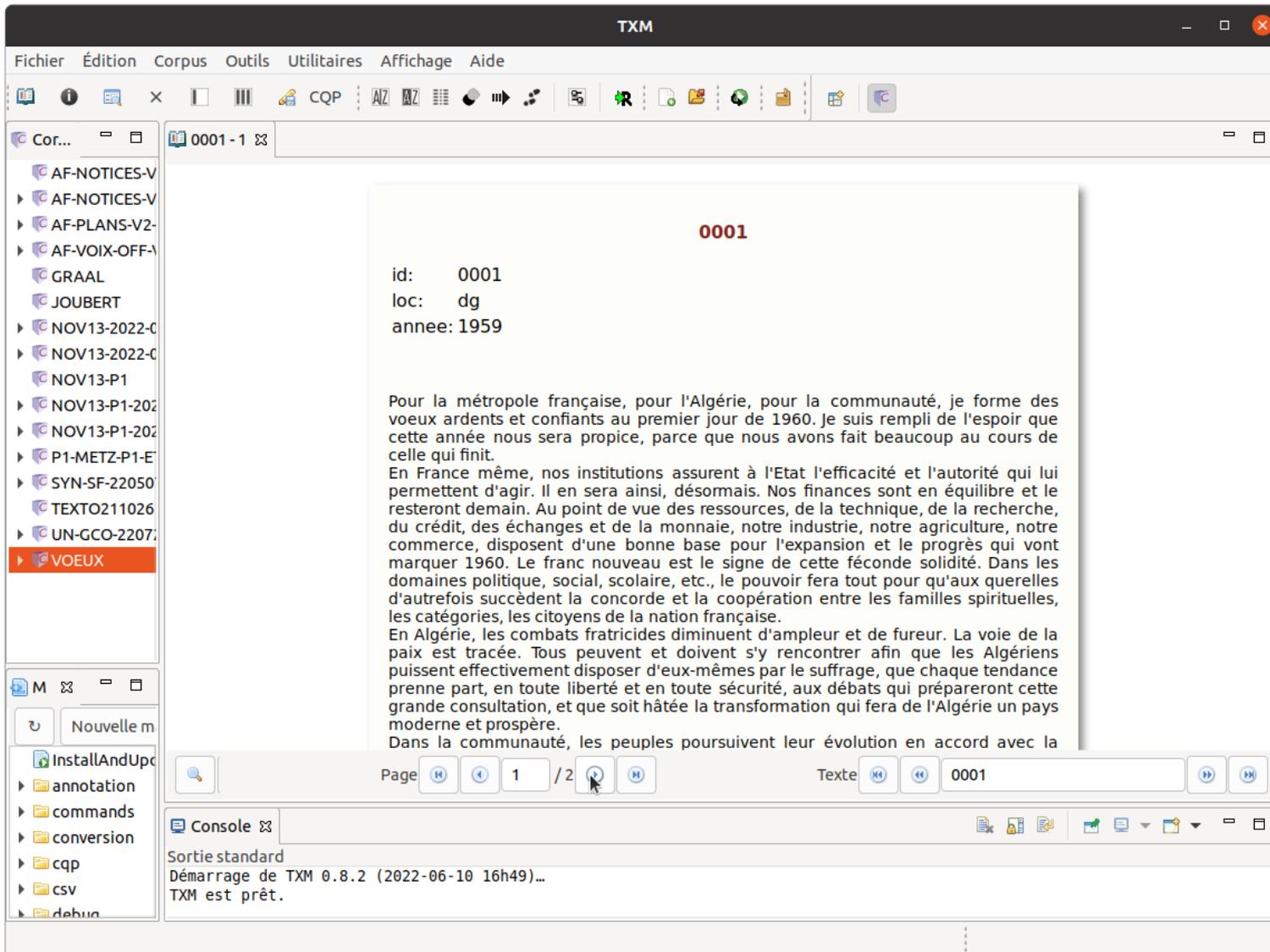
This work is licensed under the Creative Commons Attribution 4.0 International License.
<http://creativecommons.org/licenses/by/4.0/>

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.

TXM

Fichier Édition Corpus Outils Utilitaires Affichage Aide

Cor... 0001 - 1 VOEUX/text (1 / 54)

Paramètres

Structure : text Propriétés : word ; frpos ; frlemma Éditer Propriétés de structure : text_annee Éditer Nombre de lignes 10

word	frpos	frlemma
Pour	PRP	pour
la	DET:ART	le
métropole	NOM	métropole
française	ADJ	français
,	PUN	,
pour	PRP	pour
l'	DET:ART	le
Algérie	NAM	Algérie
,	PUN	,
pour	PRP	pour
la	DET:ART	le
communauté	NOM	communauté
,	PUN	,
je	PRO:PER	je
forme	VER:pres	former
des	PRP:det	du
voeux	NOM	vœu

1 / 54 Informations de structures : annee=1959

Console

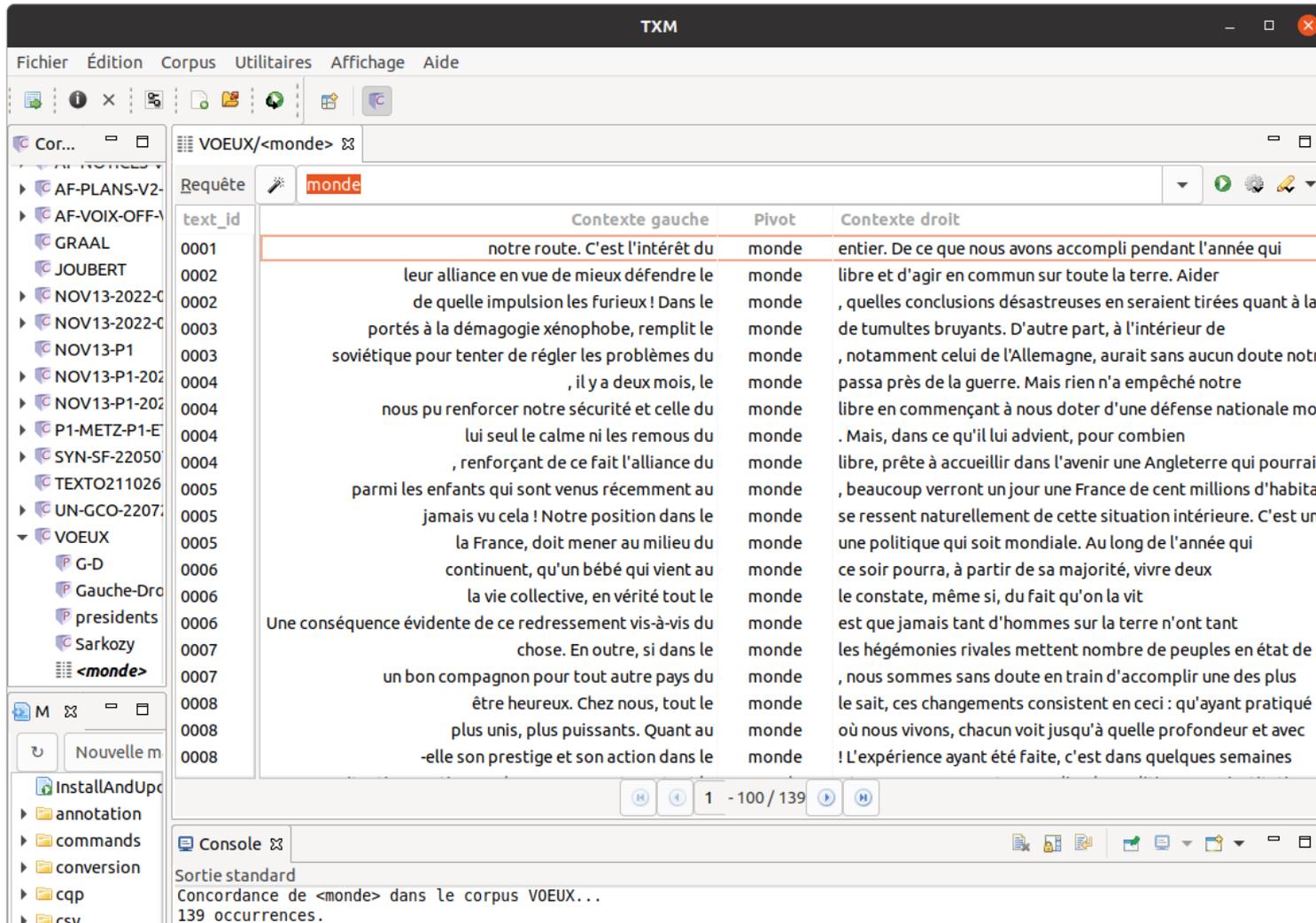
Sortie standard

TXM est prêt.
Ouverture du navigateur de VOEUX...
Terminé.

NAVIGATEUR (ou VUE INTERNE)

- Visualiser comment mes textes sont représentés dans TXM.

CONCORDANCE



Requête: monde

text_id	Contexte gauche	Pivot	Contexte droit
0001	notre route. C'est l'intérêt du	monde	entier. De ce que nous avons accompli pendant l'année qui
0002	leur alliance en vue de mieux défendre le	monde	libre et d'agir en commun sur toute la terre. Aider
0002	de quelle impulsion les furieux ! Dans le	monde	, quelles conclusions désastreuses en seraient tirées quant à la
0003	portés à la démagogie xénophobe, remplit le	monde	de tumultes bruyants. D'autre part, à l'intérieur de
0003	soviétique pour tenter de régler les problèmes du	monde	, notamment celui de l'Allemagne, aurait sans aucun doute notr
0004	, il y a deux mois, le	monde	passa près de la guerre. Mais rien n'a empêché notre
0004	nous pu renforcer notre sécurité et celle du	monde	libre en commençant à nous doter d'une défense nationale moc
0004	lui seul le calme ni les remous du	monde	. Mais, dans ce qu'il lui advient, pour combien
0004	, renforçant de ce fait l'alliance du	monde	libre, prête à accueillir dans l'avenir une Angleterre qui pourrait
0005	parmi les enfants qui sont venus récemment au	monde	, beaucoup verront un jour une France de cent millions d'habita
0005	jamais vu cela ! Notre position dans le	monde	se ressent naturellement de cette situation intérieure. C'est un
0005	la France, doit mener au milieu du	monde	une politique qui soit mondiale. Au long de l'année qui
0006	continuent, qu'un bébé qui vient au	monde	ce soir pourra, à partir de sa majorité, vivre deux
0006	la vie collective, en vérité tout le	monde	le constate, même si, du fait qu'on la vit
0006	Une conséquence évidente de ce redressement vis-à-vis du	monde	est que jamais tant d'hommes sur la terre n'ont tant
0007	chose. En outre, si dans le	monde	les hégémonies rivales mettent nombre de peuples en état de p
0007	un bon compagnon pour tout autre pays du	monde	, nous sommes sans doute en train d'accomplir une des plus
0008	être heureux. Chez nous, tout le	monde	le sait, ces changements consistent en ceci : qu'ayant pratiqué
0008	plus unis, plus puissants. Quant au	monde	où nous vivons, chacun voit jusqu'à quelle profondeur et avec
0008	-elle son prestige et son action dans le	monde	! L'expérience ayant été faite, c'est dans quelques semaines

Sortie standard
Concordance de <monde> dans le corpus VOEUX...
139 occurrences.

- 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

The screenshot shows the TXM software interface. The title bar reads 'TXM'. The menu bar includes 'Fichier', 'Édition', 'Corpus', 'Utilitaires', 'Affichage', and 'Aide'. The toolbar contains various icons for file operations and search. The main window displays a concordance search for the word 'monde' in the corpus 'VOEUX/<monde>'. The search results are organized into three columns: 'Contexte gauche', 'Pivot', and 'Contexte droit'. The 'Contexte gauche' column lists various years and names (e.g., 'chirac, 1995', 'mitterrand, 1988'). The 'Pivot' column contains the word 'monde'. The 'Contexte droit' column contains the corresponding text snippets. At the bottom, there are navigation buttons and a page indicator '1 - 100 / 139'.

Contexte gauche	Pivot	Contexte droit
chirac, 1995	monde	l'image d'un grand peuple dont je suis fier. Pour
mitterrand, 1988	monde	, il y a maintenant deux siècles, nous avons, certes
mitterrand, 1994	monde	ce qu'ils valent quand le danger est là. Réserveons également
pompidou, 1972	monde	qui fait le plus gros effort en faveur du logement. Personne
chirac, 2006	monde	à avoir inscrit en 2005 une Charte de l'environnement dans sa
dg, 1966	monde	où nous vivons, chacun voit jusqu'à quelle profondeur et avec
dg, 1963	monde	, beaucoup verront un jour une France de cent millions d'habitants
dg, 1964	monde	ce soir pourra, à partir de sa majorité, vivre deux
giscard, 1979	monde	dangereux, à un moment où l'on voit flamber le prix
sarkozy, 2009	monde	qui bouge l'immobilisme soit une alternative ? Il nous reste encore
dg, 1962	monde	libre, prête à accueillir dans l'avenir une Angleterre qui pourrait
dg, 1962	monde	libre en commençant à nous doter d'une défense nationale moderne
chirac, 1999	monde	. Le nouveau siècle est à inventer, plus fraternel, plus
giscard, 1974	monde	et qui sont donc des problèmes mondiaux. Ces vœux, je
mitterrand, 1993	monde	, la solidité du franc, redevenu monnaie forte et enviée après
chirac, 1995	monde	. Notre économie est saine. Nos entreprises sont compétitives. No
chirac, 2003	monde	. Ils ont commencé à baisser. Les dépenses de l'État
chirac, 2004	monde	. En 2005, vous aurez l'avenir de cette Europe entre
sarkozy, 2007	monde	, au service de ceux qui souffrent, des enfants et des
giscard, 1979	monde	repose sur la capacité de sang - froid de quelques hommes.
giscard, 1978	monde	, indépendant et fier mais fraternel, actif mais respectueux de la
chirac, 2001	monde	. L'euro est une victoire de l'Europe. Après un
giscard, 1977	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

The screenshot shows the TXM software interface. The main window displays a concordance search for the word "monde". The search results are shown in a table with columns for "text_loc, text_a", "Contexte gauche", "Pivot", and "Contexte". The word "monde" is highlighted in red in the search results. The search results table is as follows:

text_loc, text_a	Contexte gauche	Pivot	Contexte
pompidou, 1972	ure actuelle, le pays	monde	qui fait le p
chirac, 2006	e est le premier pays	monde	à avoir insc
dg, 1966	lus puissants. Quant	monde	où nous viv
dg, 1963	nt venus récemment	monde	, beaucoup
dg, 1964	qu'un bébé qui vient	monde	ce soir pou
giscard, 1979	publique. Mais, dans	monde	dangereux
sarkozy, 2009	peut croire que dans	monde	qui bouge
dg, 1962	t de ce fait l'alliance	monde	libre, prête
dg, 1962	otre sécurité et celle	monde	libre en co
chirac, 1999	éfricher les chemins	monde	. Le nouvea
giscard, 1974	ésormais à l'échelle	monde	et qui sont
mitterrand, 1993	issance économique	monde	, la solidité
chirac, 1995	issance économique	monde	. Notre écc
chirac, 2003	armi les plus élevés	monde	. Ils ont cor
chirac, 2004	ix grands ensembles	monde	. En 2005, v
sarkozy, 2007	paix et de l'équilibre	monde	, au service
giscard, 1979	riodes où l'équilibre	monde	repose sur
giscard, 1978	ouvert sur l'évolution	monde	, indépend
chirac, 2001	telle aux évolutions	monde	. L'euro est
giscard, 1977	on : « Si tous les gars	monde	voulaient s
da. 1959	route. C'est l'intérêt	monde	entier. De c

The search results table is followed by a text preview window showing the context of the search results. The text preview window displays the following text:

s'y passe. Si, ce soir, ce petit personnage soulevait le toit des maisons, qu'apercevrait-il ? Dans les villes, dans les campagnes, dans les départements et territoires d'outre-mer, dans les familles des français à l'étranger, il verrait la grande fête des français, réunis chacun selon ses préférences et son tempérament. C'est en pensant à cette grande fête des français que je me souviens d'un poème de Paul Fort dont on a fait une chanson : « Si tous les gars du monde voulaient se donner la main ... » Ceux d'entre vous qui ont entendu cette chanson ont certainement pressenti qu'il y avait là la recette du bonheur et de la paix dans le monde.

Ce qui est vrai du monde est vrai de la France : " Si tous les gars de France voulaient se donner la main ... " Ce sera mon vœu personnel ce soir.

Je vous souhaite à toutes et à tous une bonne et heureuse année 1978 et je souhaite à la France, qui va tout à l'heure en franchir le seuil, de connaître une bonne année.

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

TXM

Fichier Édition Corpus Outils Utilitaires Affichage Aide

Cor... VOEUX/<monde>@word ≤61 197 /61 197

Requête monde Propriétés : word Éditer

word	Fréquence
monde	139

t 139, v 1, fmin 139, fmax 139

Sortie standard
Index de <monde>, propriété @word, dans le corpus VOEUX...
1 items pour 139 occurrences.

INDEX

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

TXM

Fichier Édition Corpus Outils Utilitaires Affichage Aide

Requête [frpos="NOM"] Propriétés : word Éditer

word	Fréquence
année	258
pays	153
monde	139
vie	86
voeux	82
avenir	79
paix	79
République	69
compatriotes	62
temps	61
emploi	60
confiance	59
travail	55
ans	53
soir	53
progrès	51
crise	50
mois	48
fois	47
peuples	46
action	45

Requête [frpos="NOM"] Propriétés : frlemma Éditer

frlemma	Fréquence
année	281
pays	153
monde	139
vœu	100
peuple	91
vie	89
emploi	83
paix	80
an	79
avenir	79
république	71
compatriote	65
liberté	62
temps	61
confiance	60
effort	60
gouvernement	59
travail	59
crise	58
progrès	58
droit	55

1 100 / 2201 t 10411, v 2201, fmin 1, fmax 258

1 100 / 1727 t 10411, v 1727, fmin 1, fmax 281

INDEX

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

TXM

Fichier Édition Corpus Outils Utilitaires Affichage Aide

Requête [frlemma="France"][frpos="ADJ"] Propriétés : word Éditer

word	Fréquence
France accueillante	2
France nouvelle	2
France vivante	2
France blessée	1

t 13, v 10, fmin 1, fmax 2

Requête ([frlemma="France"][frpos="ADJ"]) | ([frpos="ADJ"] [frlemma="France"]) Propriétés : word Éditer

word	Fréquence
France accueillante	2
France nouvelle	2
France vivante	2
France blessée	1
France combattante	1
France fidèle	1
France frère	1
France juste	1
France libre	1
France même	1
nouvelle France	1

t 14, v 11, fmin 1, fmax 2

INDEX

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

INDEX

The screenshot shows the TXM software interface. The main window displays a list of corpora on the left, with 'VOEUX' selected. The central area shows a search query being built in the 'Assistant de Requête' dialog box. The query is: 'un mot dont la propriété + - frlemma correspond à je', followed by 'suivi de', 'un mot dont la propriété + - frlemma correspond à souhaiter', followed by 'suivi de', and 'un mot dont la propriété + - frlemma correspond à année'. The dialog also includes a section for 'dans un contexte de' with a value of '1' and a property 'p'. The 'Ajouter un mot' field is empty. The 'Cancel' and 'OK' buttons are visible at the bottom of the dialog.

Je recherche :

un mot dont la propriété + - frlemma correspond à je

suivi de

un mot dont la propriété + - frlemma correspond à souhaiter

suivi de

un mot dont la propriété + - frlemma correspond à année

dans un contexte de 1 - + p

Ajouter un mot

Cancel OK

Sortie standard
Index de <monde>, propriété @word, dans le corpus VOEUX...
1 items pour 139 occurrences.

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

TXM

Fichier Édition Corpus Outils Utilitaires Affichage Aide

Cor... VOEUX/<[frlemma = "je"] []* [frlemma = ...

Requête `[frlemma = "je"] []* [frlemma = "souhaiter"] []* [frlemma = "année"] within 20` Propriétés : word Éditer

word

je souhaite que l' année

Je souhaite que l' année

je souhaite une bonne année

Je vous souhaite une bonne année

je les adresse et , en souhaitant à chacune et à chacun de vous une bonne et heureuse année

Je remercie la tradition qui me vaut , pour la sixième fois , de vous souhaiter la bonne année

je souhaite , au nom de la France , une bonne et heureuse année

je souhaite , en notre nom à tous , une bonne année

je souhaite , en votre nom , bonne année

je souhaite à chacune et à chacun d' entre vous une bonne et heureuse année

Je souhaite à chacune et à chacun d' entre vous une très bonne année

je souhaite à chacune et à chacun d' entre vous une très bonne et très heureuse année

Je souhaite à chacune et à chacun d' entre vous une très heureuse nouvelle année

je souhaite à chacune et à chacun de vous une bonne et heureuse année

je souhaite de tout coeur à chacune et à chacun d' entre vous , une bonne et une heureuse année

Je souhaite donc que 1975 soit l' année

je souhaite que 1975 soit l' année

je souhaite que 2001 soit une année

1 27/27 t 31, v 27, fmin 1, fmax 2

Console

Sortie standard

Index de <[frlemma = "je"] []* [frlemma = "souhaiter"] []* [frlemma = "année"] within 20>, propriété @word, dans le corpus V
27 items pour 31 occurrences.

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)

The screenshot shows the TXM software interface. The main window displays a table of co-occurrences for the word "monde". The table has columns for Cooccurent, Fréquence, CoFréquence, Indice, and Distance moyenne. The search query is "monde". The table lists various words and their associated statistics. Below the table, there is a console window showing the output of the search: "Sortie standard" and "Cooccurents de <monde>, propriété @word 9 9, ≥2 ≥2 ≥2.0, dans le corpus VOEUX... 36 cooccurents pour 139 occurrences du pivot."

Cooccurent	Fréquence	CoFréquence	Indice	Distance moyenne
dans	419	70	20	2,1
le	895	96	14	1,2
du	366	48	10	1,2
vivons	13	8	7	3,4
où	95	17	5	2,3
change	6	4	4	2,2
équilibre	20	6	3	4,5
rang	14	5	3	4,6
dangereux	4	3	3	2,7
2005	5	3	3	4,7
Dans	26	6	3	3,3
parmi	18	5	3	5,8
partout	19	5	2	2,0
nouveau	29	6	2	,3
11	2	2	2	4,0
fraternel	2	2	2	7,0
Un	21	5	2	,0
libre	13	4	2	1,2
quiescence	22	5	2	4,4

t pivot 139, v cooc 36, t cooc 455, T corpus 61197

Sortie standard
Cooccurents de <monde>, propriété @word 9 9, ≥2 ≥2 ≥2.0, dans le corpus VOEUX...
36 cooccurents pour 139 occurrences du pivot.

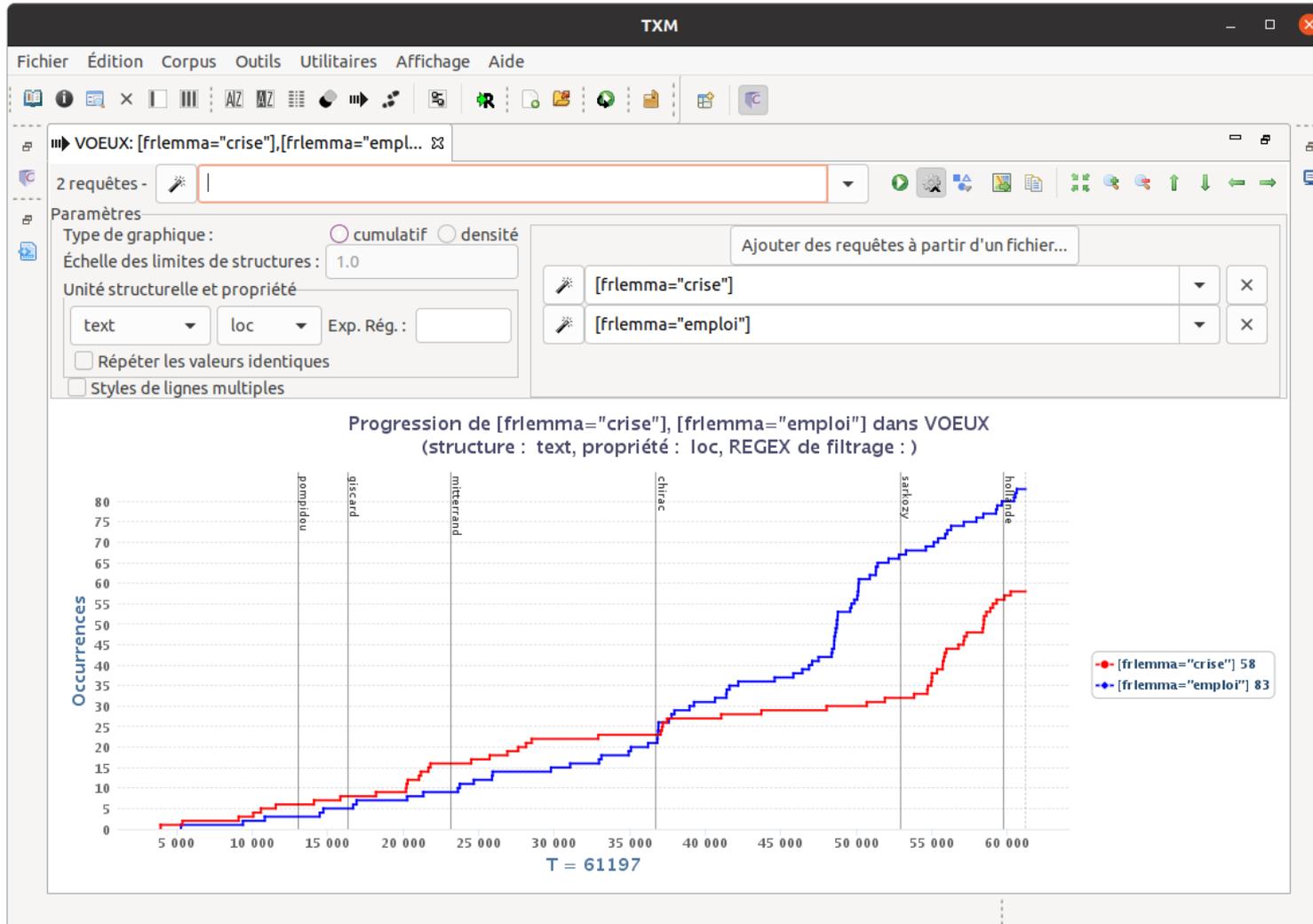
COOCCURRENCE

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

text_loc, text_annee	Contexte gauche	Pivot	Contexte droit
dg, 1966	plus unis, plus puissants. Quant au	monde où nous [vivons]	, chacun voit jusqu'à quelle profonde
giscard, 1974	t les difficultés et les risques réels du	monde dans lequel nous [vivons]	et dans lequel nous allons vivre l'an p
giscard, 1974	conciliation. Je lui souhaite, dans le	monde tourmenté où nous [vivons]	, d'apparaître précisément comme u
giscard, 1975	la France libre. Aujourd'hui, dans le	monde où nous [vivons]	, compte tenu de notre dimension et
giscard, 1976	ons dépassées, pour bien comprendre	monde où nous [vivons]	et pour choisir des solutions généra
giscard, 1979	. Le danger de guerre existe. Nous	[vivons] dans une de ces périodes où l'équilibre du monde	repose sur la capacité de sang - froic
mitterrand, 1984	intérêt de tous. Et puis dans le	monde très dur où nous [vivons]	, où l'on n'a rien pour rien, il faut
chirac, 2002	termination. Lucidité, parce que nous	[vivons] dans un monde	incertain, dangereux, où les menaces

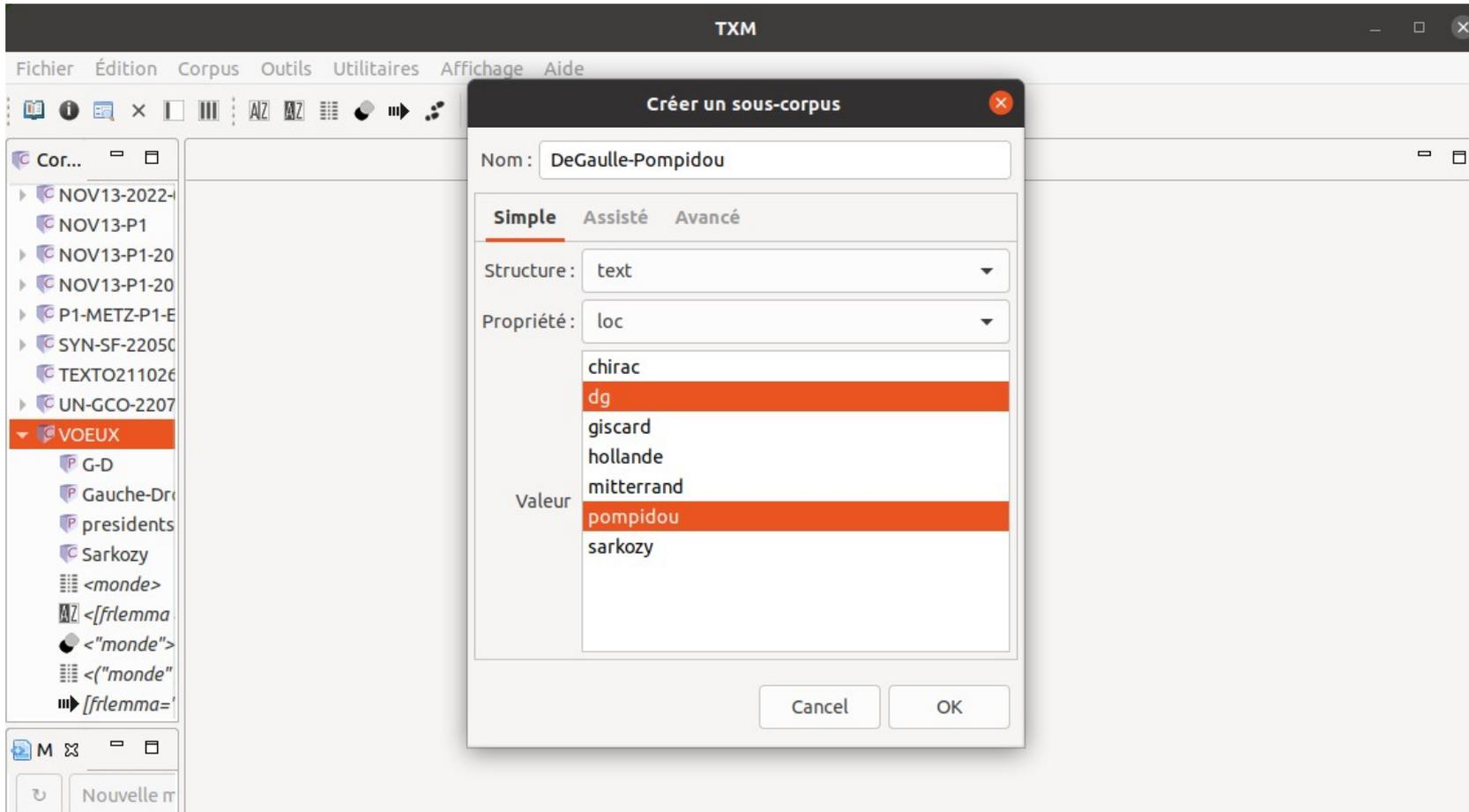
PROGRESSION

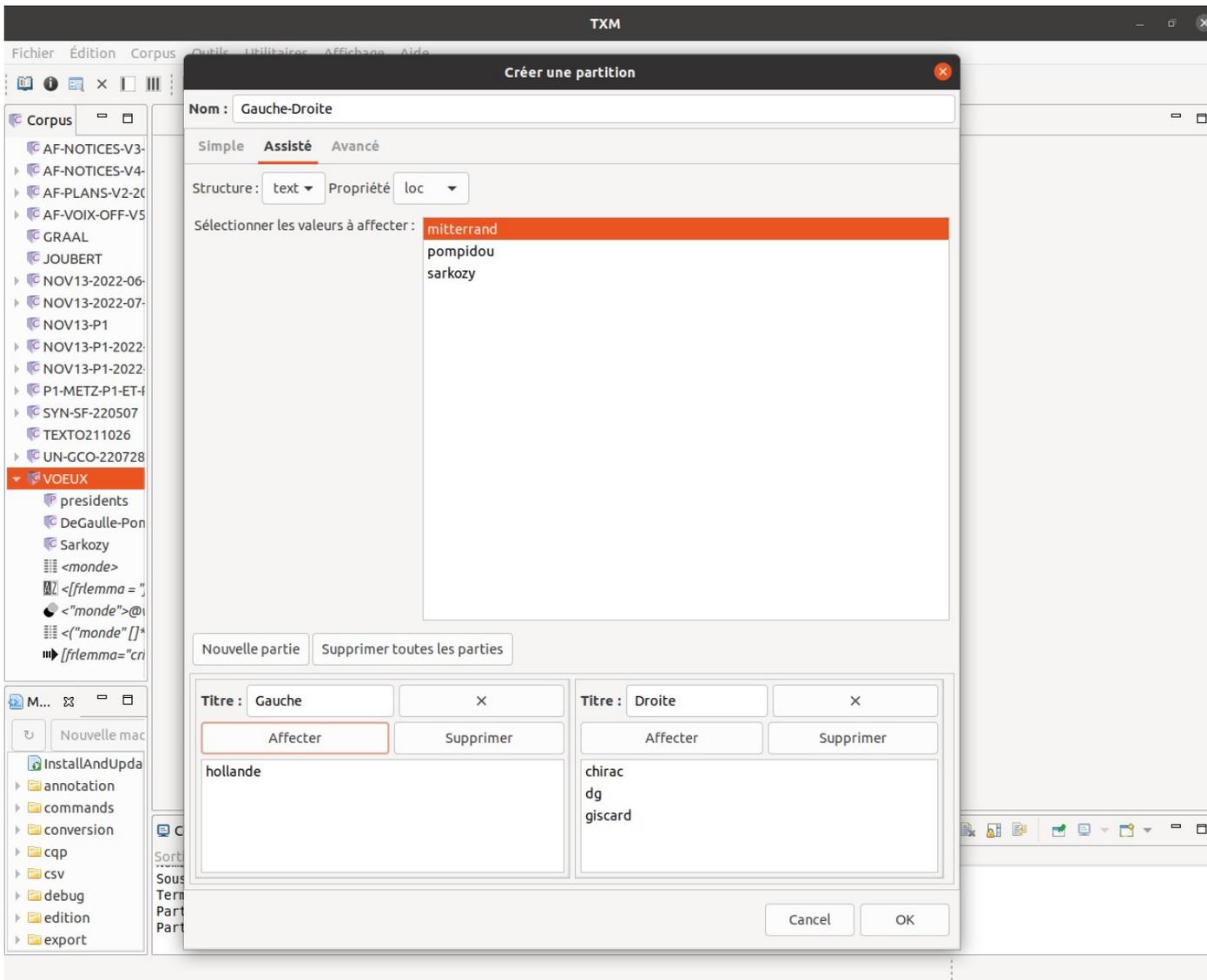
- On visualise la position des occurrences au fil du déroulement du corpus



SOUS-CORPUS

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



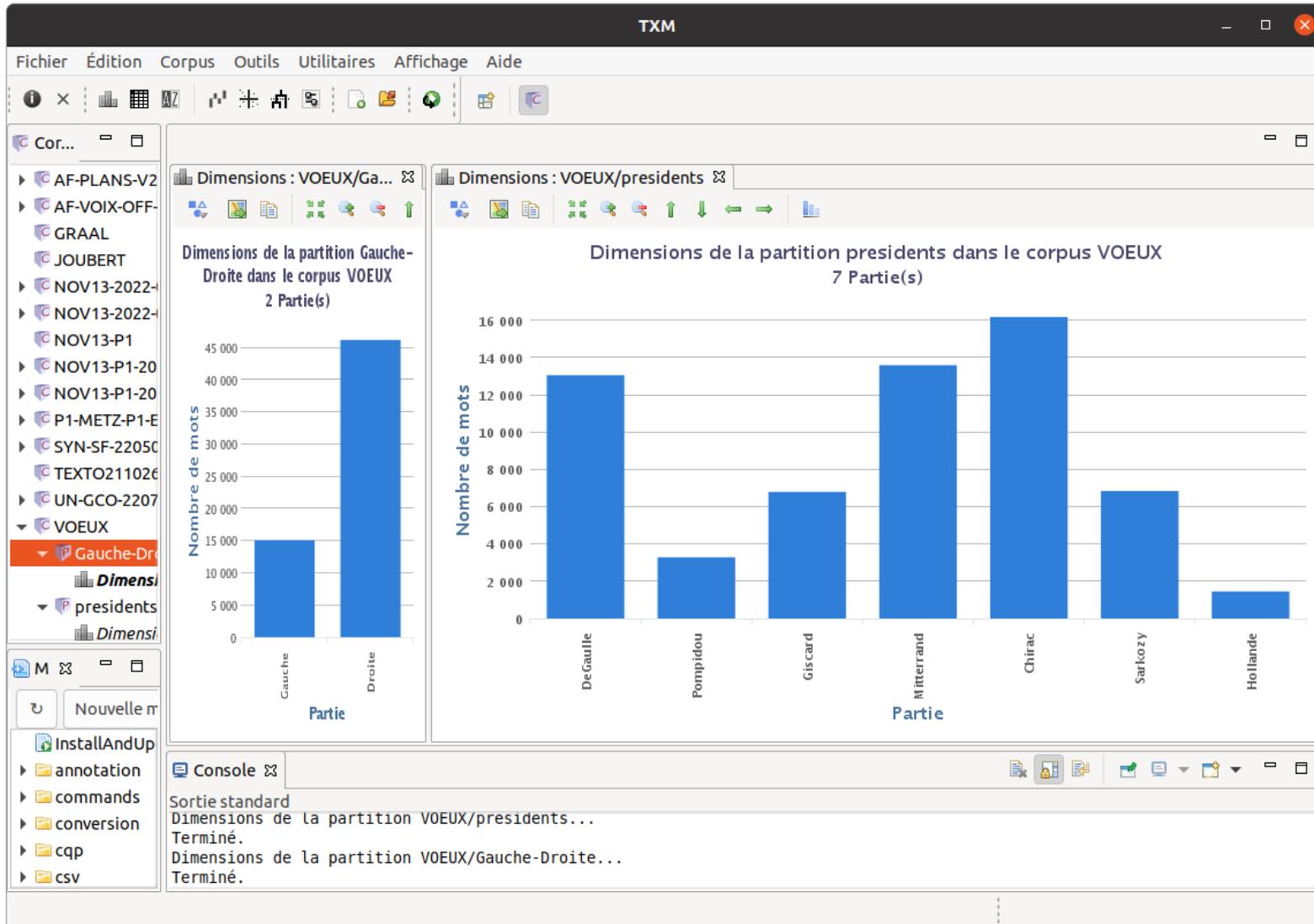


PARTITION

- 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

The screenshot shows the TXM software interface. The search query is `<[frpos="NOM"]>@frlemm...` with the property `frpos="NOM"` selected. The search parameters are set to Fmin: 1, Fmax: 61197, Vmax: 300, and Résultats par page: 100. The main window displays a table of word frequencies for the 'presidents' partition, sorted by the total frequency across all parts.

frlemma	Fréquence	DeGaulle t=13054	Pompidou t=3285	Giscard t=6797	Mitterrand t=13592	Chirac t=16176	Sarkozy t=6840	Hollande t=1453
année	281	51	25	50	50	64	35	6
pays	153	34	11	22	37	29	17	3
monde	139	23	4	28	20	43	20	1
vœu	100	10	3	37	29	13	7	1
peuple	91	36	6	8	28	8	3	2
vie	89	16	6	13	19	18	15	2
emploi	83	3	2	4	12	46	13	3
paix	80	18	4	10	25	18	5	0
an	79	15	8	8	25	13	7	3
avenir	79	8	7	3	5	41	12	3
république	71	17	0	8	20	21	4	1
compatriote	65	0	0	7	5	31	19	3
liberté	62	4	3	21	15	14	5	0
temps	61	11	3	13	12	16	6	0

The console output shows: `Sortie standard
Index de <[frpos="NOM"]>, propriété @frlemma, de la partition VOEUX/presidents...
300 items pour 10 411 occurrences.`

TXM

Fichier Édition Corpus Outils Utilitaires Affichage Aide

Cor... VOEUX/presidents/<[frpos="NOM"]>@frlemm... VOEUX/presidents/<[frpos="NOM"]>@frlemm...

Propriété frlemma

frlemma	Fréquence	DeGaulle t=1461	Pompidou t=355	Giscard t=774	Mitterrand t=1492	Chirac t=2052
extérieur	9	3	1	3	2	0
face	29				0	13
façon	15				4	2
fait	22				2	2
famille	54				11	11
femme	16				5	6
fête	13				1	3
fin	24				9	2
finance	9				0	1
foi fois	47				14	8
fond	21				2	4
force	20	1	0	4	3	11
force forces	29	3	0	5	10	4
formation	20	2	1	0	6	7
fraternité	23	3	1	9	2	6
fruit	10	2	1	0	3	2
gouvernement	59	4	7	2	14	22
grâce	42	6	2	1	10	10
guerre	35	11	0	4	13	6

Nouveau nom

Saisissez le nom de la nouvelle ligne issue de la fusion

force|

Cancel OK

T 7150 V 300 Fmin 8 Fmax 281

Console

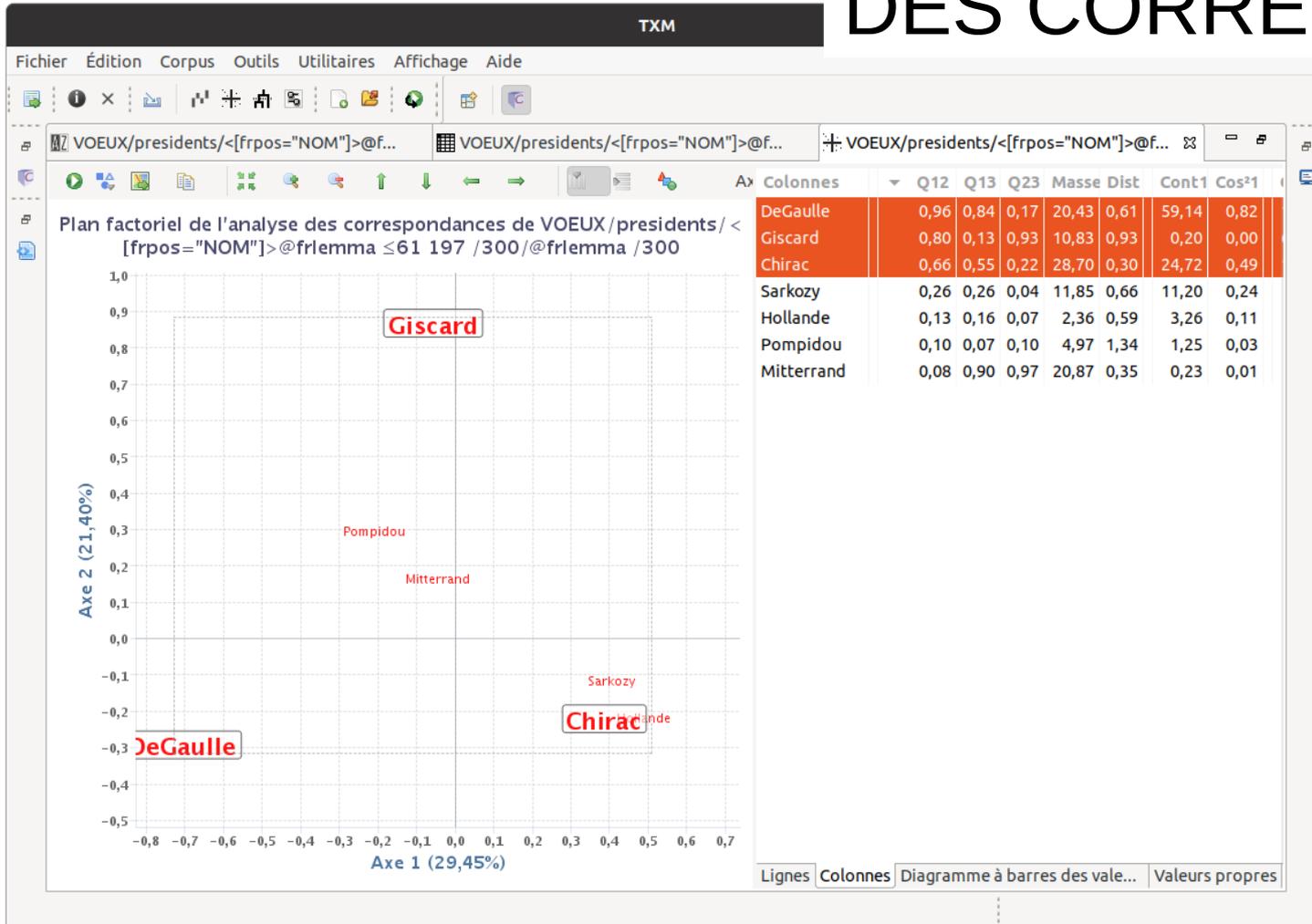
Sortie standard

Terminé.

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

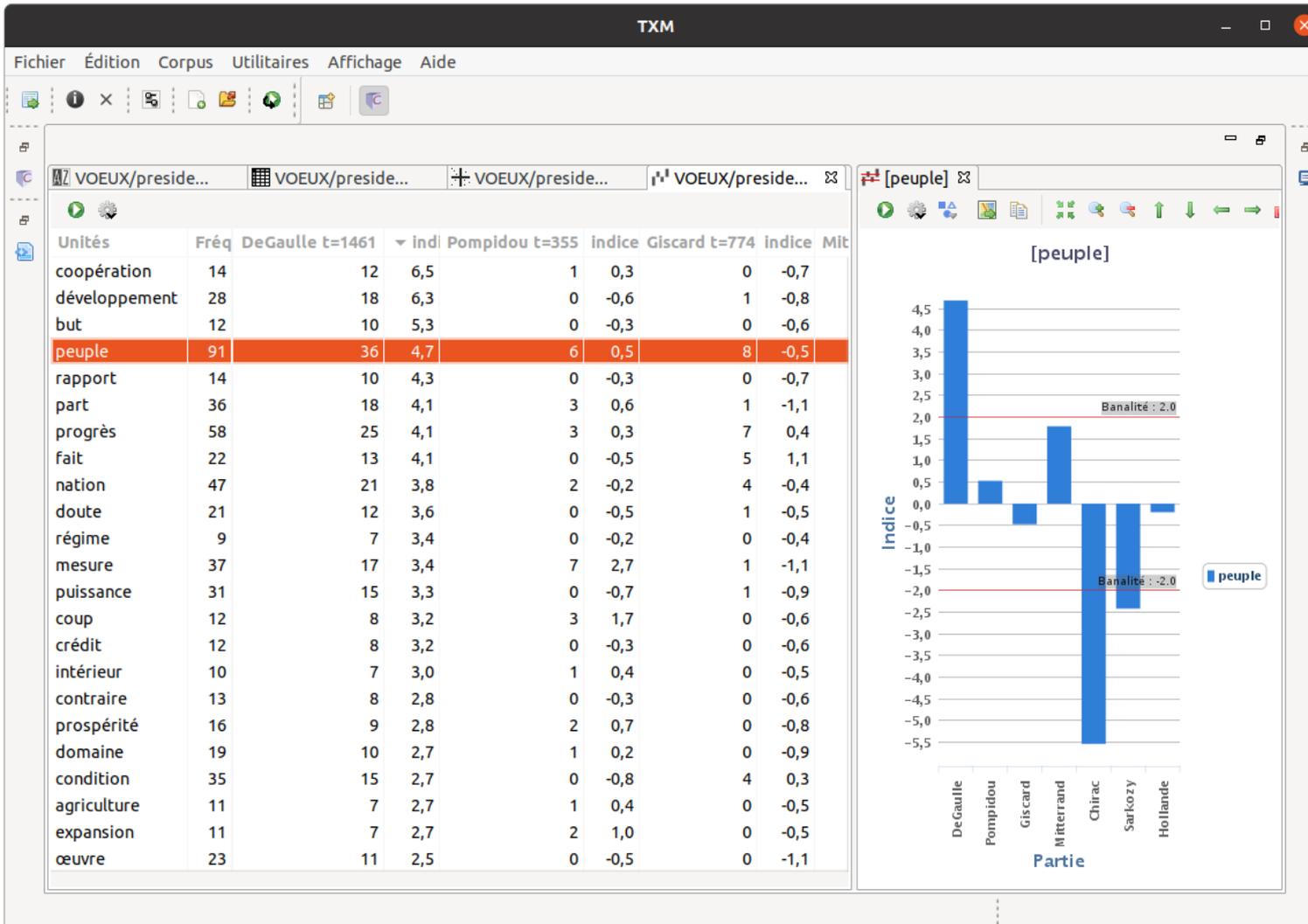


- Pour une visualisation cartographique de la répartition des mots dans les parties, avec des indicateurs pour l'interprétation du graphique
- Le calcul trouve la représentation 2D qui préserve le mieux les contrastes

SPÉCIFICITÉS

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 construction 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

Concordanciers enrichis
avec analyse quantitative

CALCULER
« bavard »,
non focalisé

Textométrie
avec annotation
ou catégories en
cours d'analyse

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

Text Mining

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

EXPLORER

le matériau empirique,
les données source

Concordanciers KWIC
et moteurs de recherche
linguistiques

Registres
Annotation qualitative
assistée basée
sur le corpus

CAQDAS
(Computer-assisted
qualitative data
analysis software)

MONTRER

Qualitatif,
« muet » (focalisé)

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

The screenshot shows a web browser window with the URL `portal.textometrie.org/demo/`. The page features a navigation bar with a search box and a sidebar menu titled "Corpus". The main content area displays a "Welcome to the TXM demo portal" message and a list of public corpora.

Welcome to the TXM demo portal

This site allows you to experiment with the TXM web portal software on various public and private corpora.

Public corpora

To access a public corpus, right-click on its icon and select a command to apply (Texts, Lexicon, Concordance, etc.). See (Help) for documentation of the commands

- **BROWN** (1961 English XML-TEI encoded and tagged texts / 1,161,028 words / 500 texts)
"A Standard Corpus of Present-Day Edited American English, for use with Digital Computers."
by W. N. Francis and H. Kucera (1964), Brown University Providence
License: May be used for non-commercial purposes.
Source: [NLTK project "Brown Corpus \(TEI XML Version\)"](#)
Converted to a format compatible with the TXM platform by the [Textométrie research project](#).
- **VWWP** (1830-1929 English XML-TEI P5 encoded texts / 9,570,731 words / 199 texts)
Victorian Women Writers Project text collection.
Copyright © 2014 The Trustees of Indiana University
Licensed under Creative Commons Attribution-NonCommercial 3.0 (CC BY-NC 3.0)
Source: [Indiana University Libraries, Digital Collections Services](#)

French : written

- **VOEUX** (1959-2012 French political discourse / 61,102 words / 54 discourse)
Copyright © 2010 Jean-Marc Leblanc
Licensed under Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported
- **TDM80J** (*Around the World in 80 Days* by Jules Verne, 1873, in French / 85,130 tokens / 1 text)
The source text is in public domain, adapted for TXM from [Wikisource](#) by S. Heiden

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 \pm 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

BFM GRAAL Accueil Aide TEXT: QGRAAL_CM

Bienvenue S'inscrire Se connecter Contact fr

CORPUS

Corpus

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

[Galaad à Kamaalot]

§ 1

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

[5]

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

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

<160a>

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

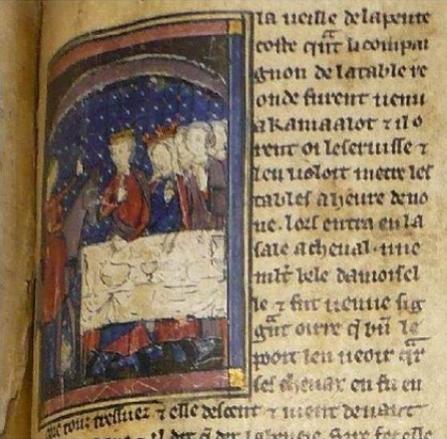
§ 1

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

[5]

10 nōne .’ lozf en[tra acheual en la] fale une mout bele
damoifele. ⁊ fuuene figzant oïrre que bien le pot
l'en ueoir. car fef cheuau enfu encōre toz suanz. ⁊ ele
descent et uient deuant le roi si le salue. et il dist que
diex la beneie. Sire fet ele por dieu ditef moi se lanç.
15 est ceenz. Oil uoir fet li roif ueez le la. filimofte. ⁊
ele ua maintenant la ouilef. ⁊ li dist lanç. ieuof
di de par le roi pellef queuof avec moi uenez iusfen
cele fozeft. ⁊ ilidemande a qui ele est. Je sui fait
ele a celui donc ie uof paroil. Et quel beoifn fet
il auez uof de moi . Ce uerroiz uof bien bien fet ele.
20 Depardieu fet il. ⁊ girai uolentierf. lozf dist a un
escuier quil mete la sele en son cheual. ⁊ liaport
fef armez. ⁊ cil sifet tout maintenant. Et quant

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



Retour au manuscrit K (Lyon, BM, P.A. 77, col. 16)

ndue. lozf en... une moure de
damoisele ⁊ fuuene figzant oïrre que bien le p
len ueor. car fef cheuau enfu encōre toz suanz. ⁊
de lant ⁊ uient deuant le roi si le salue ⁊ il dist qu
die ⁊ la beneie. Sire fet ele por dieu ditef moi se la
est ceenz. Oil uoir fet li roif ueez le la. filimofte.

page

GRAAL - normalisee, ...

normalisee | facsimilaire | ms_colonne |

160a / 268

PDF Notice

34

Fonctionnalités 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

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

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 préceuseure 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

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

Query: m_type="DEL">[]+</item> Properties: word Edit

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

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

AFNOTICES <item_type="DEL">[word="Belgi...]

Query: n_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	omme âgé, tête nue	Belgique	Zeebrugge Flandre
1946-05-02, AFE85001458	département Laon	Belgique	Bruxelles Pêche s
1946-05-02, AFE85001460	n Leemput, Marcel	Belgique	Bruxelles Demi fin
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.

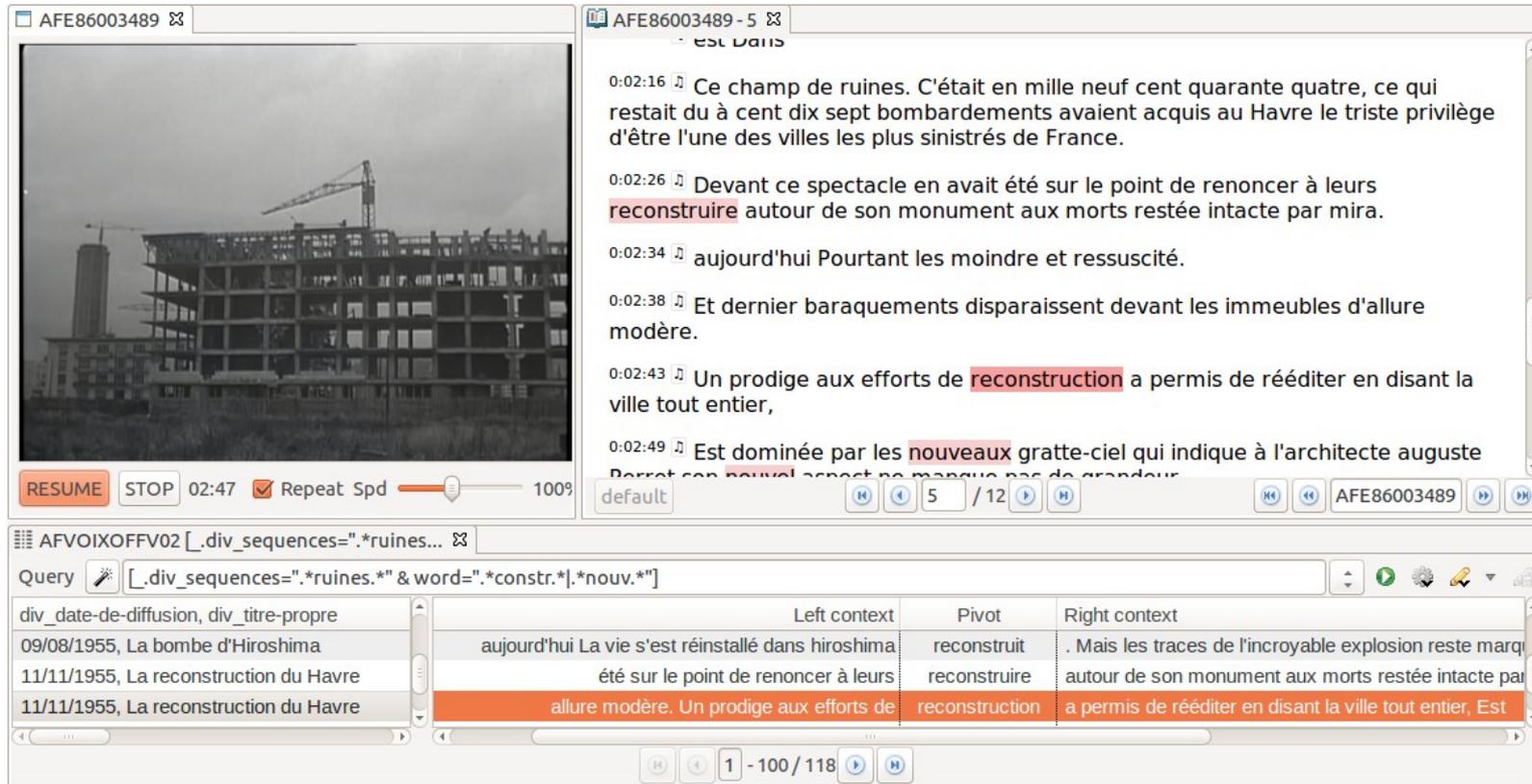
SÉQUENCES

- 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 Berré 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.*']

div_date-de-diffusion, div_titre-propre	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 opérées par la transcription.

Possibilité d'accès distant aux vidéos



The screenshot displays a video player interface with two main panes. The left pane shows a list of video entries with columns for text ID, date, left context, pivot, and right context. The right pane shows a detailed view of a video entry, including a list of tags, a title, and a transcript.

Left Pane (List of Videos):

text_id	text_datedediffusion	Left context	Pivot	Right context
AFE86003736	10/08/1960	siasuques du monde entier et de la	foule	ues congressiste qui sera bientot e
AFE86003728	15/06/1960	voiture complètement bloqué par la	foule	des manifestants, il n'avait dû qu'u
AFE86003431	07/10/1954	propriétaire Madame portrait de la	foule	des parieurs du gagnant anonyme
AFE86004792	17/04/1952	donnant sa bénédiction Pascal à la	foule	des pèlerins. Hum. Pas de paix ég
AFE86004036	11/05/1966	cardinal Wyszynski, on a évalué la	foule	des pèlerins à cent mille personne
AFE86003459	21/04/1955	onné sa bénédiction urbi et orbi à la	foule	des pèlerins réunis place Saint-pie
AFE86003738	24/08/1960	ard. aujourd'hui L'été il ramène des	foules	des peuples des collègues. touriste
AFE86004784	21/02/1952	le prince de liège au milieu de la	foule	des représentants de tous les pays
AFE86003514	09/05/1956	clubs sportifs de Moscou. Alors la	Foule	des travailleurs a clos les défilés q

Right Pane (Detailed View):

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

generique_aff_lig

S 0: 0:00:03 euh

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

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

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



AF-NOTICES-V3-2021-09-30/<[_div_identi... AFE86004168 - 5

TITRE PROPRE

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

RÉSUMÉ

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

SÉQUENCES

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

TITRE : " SPECIAL "

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

TITRE : " SPECIAL SPORT "

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

TITRE : " COLOMBES TERRE DE SACRIFICE "

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

TITRE : " SE "

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

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

X [Navigation] 5 / 5 [Navigation] AFE86004168 [Navigation]



Grammaire cinématographique : pistes pour travailler avec TXM



- Difficultés

- Variabilité des désignations
 - Description libre
 - Pratiques \pm partagées
- Mots qui s'intercalent
 - Distance variable
 - Pris dans la requête

- Solutions

- Annotation par des catégories d'analyse (recodage)
- Projection
 - Un nouveau corpus avec uniquement les catégories d'analyse



Annotation automatique par *ANTRACT* requêtes CQL

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

```
[word=" - * (VG | PG | GVG | CG) "%c ]
```

```
[word=" - * (plan | vues?) "%c ] [word="générale?s?" "%c ]
```

```
[word=" - * PANORAMA "%c ]
```

```
[word=" - * (PE | VE) "%c ]
```

```
[word=" - * (plan | vue)s?" "%c ] [ ] [word="ensemble" "%c ]
```



Annotation automatique par requêtes CQL



cql_valeurs_de_plan_211007a.xlsx - LibreOffice Calc

Fichier Édition Affichage Insertion Format Styles Feuille Données Outils Fenêtre Aide

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

Rechercher Tout rechercher Affichage mis en forme Respecter la casse

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



Annotation automatique par *ANTRACT* requêtes CQL

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

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

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



Annotation automatique par requêtes CQL

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



Corpus annoté



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

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

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

SÉQUENCES

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

TITRE : " SPECIAL "

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

TITRE : " SPECIAL SPORT "

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

TITRE : " COLOMBES TERRE DE SACRIFICE "

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

TITRE : " SEIZE NOVEMBRE, 14H30 "

- VG ? plans du public arrivant au stade

plan Frequency

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

word plan

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

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

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

Console

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



Vue des annotations



TXM

Fichier Édition Affichage Aide

Requête Propriété Requête Propriété Requête Propriétés Requête [plan="10PG"] Requête [plan="01HS"]

plan	Fréquence	plan	Fréquence	word	Fréquence	ref	Pivot	Contexte droit	ref	Contexte gauche	Pivot	Contexte droit
10PG	13	10PG	23846	VG	21712	1945-01-11, AFE86002940	VG	du boulevard - VG de la	1946-02-01, AFE85001302	un petit cheval - En premier	plan	sur le bord de la route
14GP	13	14GP	16497	Vue	861	1945-01-11, AFE86002940	VG	de la façade de la boutique	1946-02-01, AFE85001304	de l'extrémité d'une longue	vue	sur fond de ciel - DP
70TI	10	00DP	15255	PG	636	1945-01-11, AFE86002941	VG	d'un objectif atteint avec fumées	1946-02-15, AFE85001322	hisser le fanion, en second	plan	le plus ancien ouvrier de la
00DP	9	32PP	10530	vue	198	1945-01-11, AFE86002944	VG	d'ATHENES - Ruines d'ATHENES	1946-02-15, AFE85001323	NICOLAU d'OLWER (au premier	plan	à gauche) et de (
12PM	6	12PM	9722	panorama	98	1945-01-11, AFE86002944	VG	d'ATHENES avec fumée se dégageant	1946-02-15, AFE85001323	RNANDEZ SARAVIA ? au premier	plan	à droite) - Façade d'
31PANO	3	31PANO	7302	VE	76	1945-01-18, AFE86002947	VG	générale de la façade du Palais	1946-02-15, AFE85001326	ec ruines de colonnes au premier	plan	- PE ruines du temple
20PLO	1	13PR	2969	Vues	49	1945-01-18, AFE86002948	VG	et PM du beffroi de Calais	1946-02-15, AFE85001326	, puis allée, au premier	plan	- PR de grosses pierres
		70TI	1983	Plan	39	1945-01-18, AFE86002949	Vue	générale: le général DE GAULLE	1946-02-15, AFE85001326	du temple, avec au premier	plan	une immense tête stylisée, scul
		01HS	1668	vues	37	1945-01-18, AFE86002949	Vue	générale du général DE GAULLE reme	1946-03-15, AFE85001367	sur une terrasse, devant un	plan	, examine toute l'étendue du
		40VA	1602	Panorama	29	1945-01-18, AFE86002949	Vue	générale du général DE GAULLE salua	1946-03-22, AFE85001385	assises à terre, au premier	plan	une d'elles fumant une cigarett
		20PLO	1092	PE	27	1945-01-18, AFE86002950	VG	de NANTES sous la neige -	1946-04-11, AFE85001416	ouyés à une balustrade en arrière	plan	, un bateau à un embarcadère
		30TRAV	506	-VG	23	1945-01-18, AFE86002952	Vue	générale d'une réunion d'état	1946-04-11, AFE85001416	du lac LEMAN, en arrière	plan	, un yacht ancré façade pavoiée
		60ZOOM	467	GVG	21	1945-01-18, AFE86002952	Vue	générale de Malmedy à travers un	1946-05-16, AFE85001424), soldats américains en arrière	plan) - Gros tas de casques
		11PL	230	plans	7	1945-01-25, AFE86002957	VG	du bâtiment où se déroule le	1946-06-13, AFE85001518	ZELLER, GANEVAL, en arrière	plan	le capitaine de vaisseau LAMOF
		50VE	212	vg	7	1945-01-25, AFE86002957	VG	et PM de la réunion des	1946-06-13, AFE85001520	sur la plage, en arrière	plan	, un char, sur la
		21CPL	168	plan	6	1945-01-25, AFE86002957	VG	d'un meeting de la CGT	1946-06-27, AFE85001540	nipèdes accompagnés jusqu'à un	plan	d'eau par des jeunes femmes
		51VI	64	PANORAMA	5	1945-01-25, AFE86002958	VG	de l'estrade dans une salle	1946-07-04, AFE85001551	au milieu d'architectes discutant	plans	en mains - VG d'un
		71GR	41	Vg	4	1945-02-02, AFE86002962	VG	des Halles de PARIS désertes -	1946-07-04, AFE85001551	P du général MAST examinant un	plan	sur un tréteau - Vues du
		15TGP	38	CG	3	1945-02-02, AFE86002966	Vue	en plongée de tout un quartier	1946-07-17, AFE85001581	et Edmond MICHELET en second	plan	- Divers plans de foule -
		72FLOU	25	-PG	2	1945-02-02, AFE86002968	Vues	générale d'une grande parade militai	1946-07-24, AFE85001586	es Philippines assises au premier	plan	- Soldats philippins portant de
				-Vue	2	1945-02-02, AFE86002968	Vues	générales d'un village des ARDENNES	1946-08-01, AFE85001596	voitures dans un virage - Premier	plan	de la voiture de GORDINI n
				-Vue	2	1945-02-09, AFE86002971	Vue	générale des étudiants massés devar	1946-08-08, AFE85001607	sur la rue, au premier	plan	plan , capot d'une voiture automobili
				Pa	1	1945-02-09, AFE86002971	Vue	générale d'André TOLLET, l'adressan	1946-09-19, AFE85001664	é par JIMMY GAILLARD - Premier	plan	d'un chronomètre tenu dan la

ts5, v, 7, fmin 1, fmax 13 t94217, v, 20, fmin 25, fmax 238 t23846, v, 25, fmin 1, fmax 21712

1 - 100 / 23846

1 - 100 / 1668

Console

Sortie standard

```

Index de <[.div identifiant-de-la-notice="AFE86001312" & plan!="_UNDEF_"], propriété @plan, dans le corpus AF-NOTICES-V3-2021-09-30...
7 items pour 55 occurrences.
Index de <[plan!="_UNDEF_"], propriété @plan, dans le corpus AF-NOTICES-V3-2021-09-30...
70 items pour 94 217 occurrences.
Index de <[plan="10PG"], propriété @word, dans le corpus AF-NOTICES-V3-2021-09-30...
25 items pour 23 846 occurrences.
Concordance de <[plan="10PG"] dans le corpus AF-NOTICES-V3-2021-09-30...
23 846 occurrences.
Concordance de <[plan="01HS"] dans le corpus AF-NOTICES-V3-2021-09-30...
1 668 occurrences.

```

Dans la notice Rugby

Dans tout le corpus

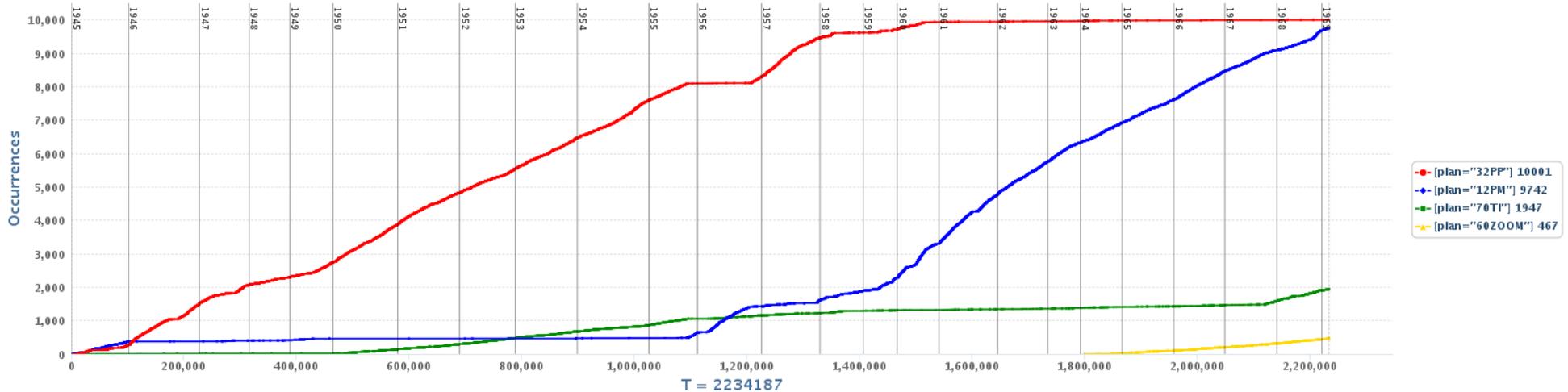
Les mots sous la catégorie « Plan général »...

... 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 *INTRACT*



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



Projection

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



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

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

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