

# Formation TXM

## Projet ANTRACT, 15 octobre 2018

Bénédicte Pincemin



Ce document est publié sous licence Creative Commons Attribution 4.0 International  
<https://creativecommons.org/licenses/by/4.0>

**Avertissement** : Dans ce document, les noms des corpus et la forme précise des requêtes CQL, en particulier celles utilisant la structuration des corpus, **correspondent à l'état des corpus en octobre 2018**. Ces éléments ne sont plus directement utilisables tels quels pour les dernières versions des corpus dont les structures ont été améliorées, **un travail d'adaptation est nécessaire pour la plupart des requêtes** complexes.

## 1. Introduction

Lexicométrie existe depuis les années ~80, logiciels disponibles depuis les années ~90  
ex. Lexico, Hyperbase, DtmVic, le Trameur ; Iramuteq

TXM = Logiciel open-source

- ouverture/transparence scientifique,
- mutualisation des développements,
- utilisation de composants (ex. R, CQP, et -moins intégré- TreeTagger)

+ capacité à travailler sur corpus structurés et enrichis (vs. minitel !).

Né +/- avec l'ANR.

≠ text (data) mining, fouille de textes (données) : le document comme source vs ressource (Valette JADT 2016)

## 2. Interface

3 zones : Corpus (navigation), Résultats, Console.

## 3. Charger

Corpus déjà préparé : format .txm.

(≠ IMPORTER ; importer + exporter → .txm)

## 4. T

sélection d'un corpus → on voit sa taille s'afficher en nb de mots en bas de la fenêtre

## 5. Édition

Feuilleter le corpus, voir les textes

3 façons de lancer les commandes

Navigation : par page (notice), par texte (année) ; 1er, dernier.

Expliciter la composition d'une notice, structurée dans TXM en métadonnées puis texte (avec différentes parties).

Affichage ≠ indexation

Escamots : Treetagger : frlemma, frpos

- frlemma = fr (français) + lemma (lemme = entrée du dictionnaire)
- frpos = fr + pos (= part-of-speech = partie du discours = catégorie grammaticale, nature du mot)

Treetagger :

- un module (expertise TAL) → peut être changé
- fonctionne par apprentissage → plusieurs modèles de langue (selon langue, époque, type de textes, écrit / oral, etc.)
- automatique → se méfier des erreurs

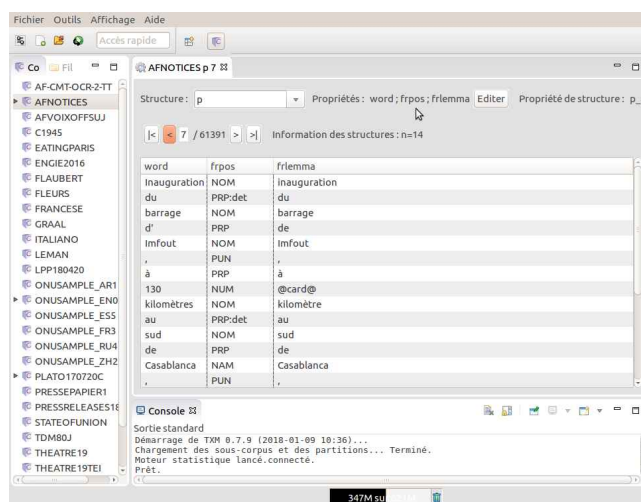
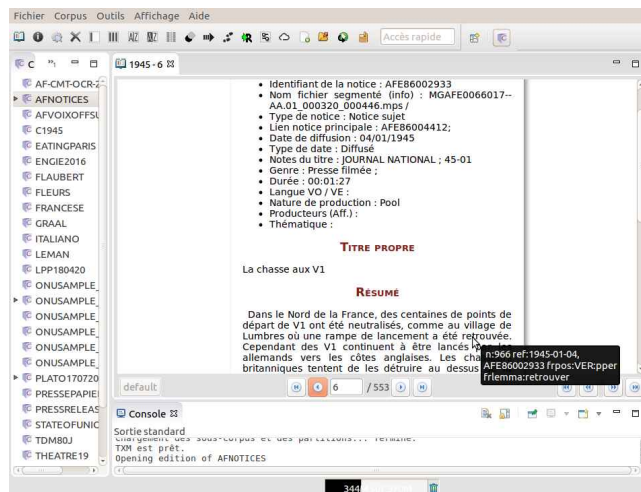
## 6. Vue interne

Comment TXM voit les mots,  
ce qu'il voit « derrière » les mots

Propriétés :

1. word = graphies = découpage en mots ; dépendance au moteur CQP
2. word, frpos, frlemma
3. n, word, frpos, frlemma, ref

On voit que les ponctuations sont prises en compte comme des mots (on peut s'en servir).



# 7. Description

Comment TXM voit le texte (ses mots mais aussi ses parties (structures) et les informations associées aux textes), ce qu'il sait sur lui, ce qu'on peut lui demander d'utiliser + savoir comment lui demander.

- Unités lexicales = mots
- Structures = délimitations (ex. une année, une notice, un résumé, un paragraphe)
- Propriétés = informations (catégorie, numéro, etc.) associée à une structure ou à un mot.

Dans les notices documentaires nous avons distingué :

- des métadonnées → deviennent des propriétés de la structure notice (ex. la date, la bobine vidéo, la durée, le fait que cela ait été diffusé ou non).
- des champs textuels, du contenu → structures, avec des mots dedans.



Les noms ont été typographiquement "réduits" (pas d'espace, pas de majuscule, pas d'accent, pas de ponctuation) → ce sont ces formes réduites qu'il faudra utiliser pour "parler" à TXM.

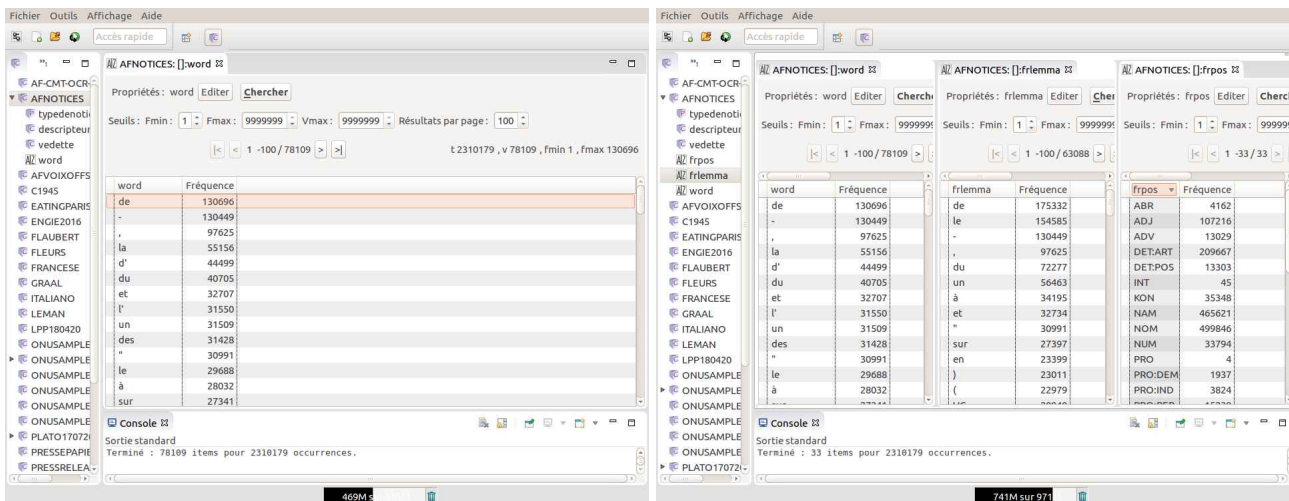
Attention, les énumérations inachevées peuvent se terminer par un point (au lieu de points de suspension).

p : englobe le contenu de chaque champ textuel (titrepropre, résumé, séquences, descripteursaffcol, generiqueaffcol) ; p1 : paragraphes (dans les résumés et les séquences)

Si on veut récupérer cette page : copier/coller ou bien voir le chemin du fichier.

# 8. Lexique

Une première fonction d'analyse : lister tous les mots du corpus avec leur fréquence



Au départ...

Différents lexiques, en variant la propriété...

Principe simple à comprendre, mais déjà apports non négligeables, notamment synthèse : T (longueur du corpus en mots) >> V (nombre de mots différents). Paramètres : expliquer aussi quantitatifs (seuils)

Usage :

1er mot peu indicatif.

Dominantes : prise de connaissance synthétique ; ici les abréviations de prise de vue et les "plans", la France et Paris, la foule... → points d'entrée ;  
recherche +/- exhaustive d'un thème (peu de mots très fréquents, beaucoup peu fréquents ; les hapax -fréq. 1- représentent ~la moitié du vocabulaire).

Lemmes ; si on veut les comparer ? → affichage synoptique (expliquer le fonctionnement du gestionnaire de fenêtres)

Ce qui change : groupement qui donne du poids (mais perte de l'information réduite), désambiguïsation (ex. parti), formes artéfactuelles (ex. "surimpressionner" en fréquence 1142, p. 101-200).

On gagne en synthèse, ce qu'on perd c'est le contexte : d'où lien hypertexte vers la concordance (nécessité de revenir au contexte pour interpréter)

POS : on trie alphabétiquement en cliquant sur l'entête de colonne pour mieux voir la structure du jeu d'étiquettes

interpréter : retour au texte

## 9. Export

- pour utiliser un résultat en dehors de TXM (publication, autres traitements,...)
- pour conserver un résultat après fermeture de TXM (mais c'est en train de changer, pour TXM 0.8)

Export du "pauvre" : copier/coller par ctrl-c ctrl-v (ctrl devient pomme/commande chez les macs). On peut coller dans un traitement de texte ou un tableur. Permet un export sélectif, mais lourd/inadapté pour un export complet.

Export du "riche" :

réglages nécessaires du fait de la diversité des environnements logiciels. Principe : il faut que TXM et le tableur s'entendent sur les choix de codage. On va d'abord essayer de faire en sorte que TXM s'adapte.

pré-réglage des préférences :

- windows français : windows-1252
- mac : x-MacRoman (ou Romania ?)

Test sur un lexique avec diacritiques (pour vérifier codage des caractères)

Ouverture dans un éditeur de texte, pour voir à quoi ressemble un fichier .csv. Le .csv est au .xls ce qu'est le .txt au .doc : une forme simplifiée, mais lisible par quasiment tous les logiciels (d'édition de texte / de tableau).

Paramétrage de l'ouverture dans LibreOffice/OpenOffice : 1ère colonne = texte ; suivantes = US english.

Si problème :

- confusion des colonnes : changer le séparateur de colonnes.
- certains caractères illisibles : vérifier le paramétrage ; utiliser plutôt par UTF-8 et passer par une commande "import de données" (ex. Excel 2003 : Données > Données externes > Importer des données), en profiter pour typer les colonnes.

## 10. Concordance

Une autre lecture grâce à une disposition très particulière des contextes : vue « dense » des emplois d'un mot, dans ses contextes.

1. Requête simple pour avoir un exemple.  
énorme

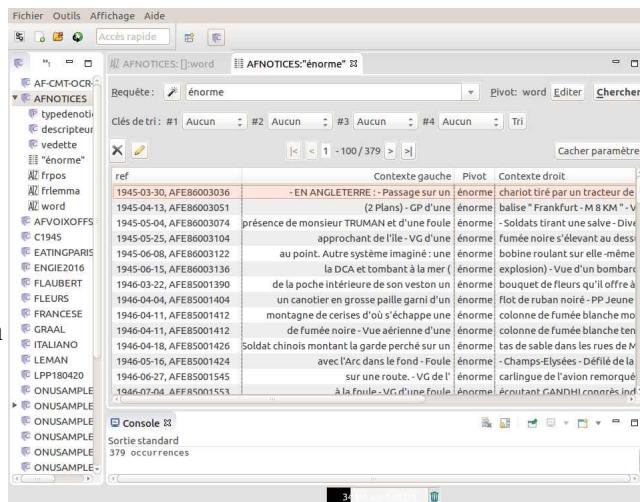
→ foule énorme

Présentation qui permet d'avoir un affichage dense des contextes d'emploi.

La concordance est efficace pour rendre compte de l'usage d'un terme.

2. Tri sur contexte. (Noter que le tri gauche est un tri "à l'envers", qu'on ne ferait pas dans un tableur.)

3. Localisations : Options d'affichage des références, par ex. ajout de notice:type de date.



4. Taille contexte :

3 modes de réglage de la taille (menu contextuel ; préférences ; lien hypertexte vers texte intégral).

retour au texte par double-clic sur la ligne ; rappel affichage synoptique (cette fois-ci haut/bas).

Complémentarité entre concordance et édition :

- concordance : contextes d'usage (local & global), vue dense et synoptique, régularités de voisinage immédiat
- édition : disposition dans le texte (paragraphe, début...), regroupements ; lecture (mise en page).

## 11. CQL

Le langage d'interrogation, pour exprimer des choses plus fines qu'un mot tel quel : une expression complexe, des conditions sur les catégories grammaticales, etc.

Il y a tt un langage (CQL = Corpus Query Language) qui permet de combiner des mots (alternatives, mais aussi suites de mots) comme de préciser ce que l'on veut pour un mot donné, par exemple

foyers? → plusieurs sens dont sens technique vidéo

étud.\* → étudiants-ouvriers -> étudiant.\*

"ouvrier.\*"%d

"ouvrier.\*"%cd

"étudiant.\*|ouvrier.\*"%cd → on attrape (~1. 50) des "ETUDIANTS"

Pour faire le | en mac : alt + shift + L

"étudiant.\*|ouvrier.\*|travailleur.\*"%cd

```
[frlemma="ouvrier"]
```

```
[frlemma="étudiant"]
```

```
[frlemma="étudiant|ouvrier"]
```

```
[frlemma="pouvoir"]
```

```
[frlemma="pouvoir" & frpos="NOM"]
```

```
[p1 & frlemma="télévision"]
```

```
[p1 & frlemma="télévis.*"]
```

```
[p1 & frlemma="télévision|téléviseur"]
```

```
[p1 & frlemma="télévis(ion|eur)"]
```

```
[p1 & word="Bardot"%c]
```

et pour pouvoir étudier les constructions linguistiques autour de BB (qualifications, etc.) on observe aussi les tris sur les contextes.

```
"arrière" "plan"
```

```
[word="arrière"] [word="plan"]
```

```
[word="arrière"%c] [word="plan"%c]
```

(en lemmes -> baisse fréquence -> erreur -> arriérer)

```
[word="Brigitte"%c] []* [word="Bardot"%c] within p1
```

```
[word="Bardot"%c] []* [word="Brigitte"%c] within p1
```

## 12. Index

*Listes synthétiques de mots (ou expressions, etc.) hors contexte, pour les mots qui vérifient certaines contraintes (de forme, de contexte).*

### Usage 1 : chercher la présence et fréquence d'un mot donné.

Si absence : autres manières de dire, voire variantes graphiques, voire erreurs (de transcription ou d'analyse). ex. :

bonheur  
portables?

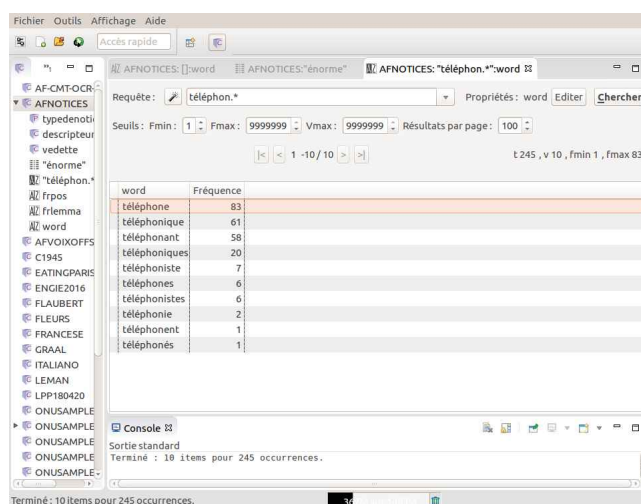
### Usage 2 : voir les formes de réalisations d'un mot, d'une famille de mots, voir dominances et creux

- radical par troncature : **téléphon.\***

retour au texte par double-clic sur le mot ;

noter que le sens ne passe pas que par les noms ;

essayer de bien différencier l'index (avec des fréquences, plus concis) et la concordance (avec des contextes).



word	Fréquence
téléphone	83
téléphonique	61
téléphonant	58
téléphoniques	20
téléphoniste	7
téléphones	6
téléphonistes	6
téléphonie	2
téléphonent	1
téléphonés	1

⇒ **1 ligne d'index = N lignes de concordance.** ⇐

```
[frlemma="anti.*"]
```

```
[frlemma=".*ette"]
```

```
[frlemma=".*ss.me"]
```

Rechercher les abréviations de prises de vue (1/2)

INDEX de :

```
" - " "[A-Z]{2,}"
```

Remarques :

- DG = De Gaulle
- GD : toutes les occ. sont avec PANO (ex. PANO horizontal GD)
- PC = abréviation (poste de commandement, parti communiste...)

### Usage 3 : mettre au point une requête

```
[p1 & frlemma="télévis.*"]
```

```
[p1 & frlemma="télévision|téléviseur"]
```

```
[p1 & frlemma="télévis(ion|eur)"]
```

#### Usage 4 : construire un lexique ciblé, construire une liste

[frpos="NOM"] → foule  
[frpos="ADJ"], puis régler propriété=frlemma → jeune ; nouveau (neuf ?) ; noir, blanc (voir en concordance si locution).  
[frpos="VER.\*"] → + mouvement, - modaux (pouvoir, devoir, vouloir, falloir...) (cf. VOEUX, FLAUBERT, FLEURS pos="V" en lemme)

[frlemma="défiler"], en word, puis frpos  
Voir ici la nécessité de compléter la requête lors du retour au texte → se méfier du lien hypertexte de l'index.

Lister (hiérarchiquement) les/des mots-clés  
<item>[descripteursaffcol]</item>  
<item>[descripteursaffcol]+</item>  
<item\_type="DEL">[descripteursaffcol]+</item>  
<item>[descripteursaffcol & \_.item\_type="DEL"]+</item>

Vue sur les sujets dominants :  
[titrepropre]  
[titrepropre & frpos="NOM"]  
[frpos="NOM"] within titrepropre  
et titres répétés :  
<titrepropre>[]+</titrepropre>

#### Usage 5 : recherche distributionnelle :

[frlemma="jeune"][frpos="NOM"]  
[frlemma="vieux"][frpos="NOM"]

[frpos="NOM"][frlemma="allemand"]

[frlemma="armée|prisonnier|problème|peuple|jeune|question|chancelier|balle|réarmement|soldat|troupe|occupation|capitulation|bombardier|avion|aviation|débâcle|défaite|déroute|résistance|mine|mortier"][frpos="ADJ"]

[frlemma="armée|prisonnier|problème|peuple|jeune|question|chancelier|balle|réarmement|soldat|troupe|occupation|capitulation|bombardier|avion|aviation|débâcle|défaite|déroute|résistance|mine|mortier"]@[frpos="ADJ"]

Foule, adj. Antéposé :  
[frpos="ADV"]?[frpos="ADJ"] [(resume|sequences) & word="foules?"]

Foule, adj. Postposé :  
[(resume|sequences) & frlemma="foule"] [frpos="ADV"]?[frpos="ADJ|VER:pper"]

Foule, parasyonymes 1 :  
[frlemma="foule"][frlemma="de|du"] []  
[frlemma="foule"][frlemma="de|du"]@[]

Foule, parasyonymes 2 :  
@[frpos="NOM"][frlemma="énorme|massé|enthousiate|immense|applaudissant|nombreux|compact|dense|indigène|rassemblé|recueilli"]

On peut être intéressé par les résultats en graphies (on voit s'il s'agit d'un nom surtout utilisé au pluriel ou au singulier) ou en lemmes (pour que les formes singulier et pluriel d'un même nom soient comptabilisées ensemble pour le même mot).

## 13. Sous-corpus

Construire un sous-ensemble dans le corpus, que l'on utilise comme un plus petit corpus, plus spécialisé par exemple.

Cela fait un petit cube « [C] », sous le cube [C] du corpus.

Sous-corpus foule7docs :

```
<notice_identifiantdelanotice="AFE85001584|AFE02014928|AFE85009316|AFE85010014|AFE86000318|AFE86000603|AFE86001194">[] expand to notice
```

ou

```
<notice>[_notice_identifiantdelanotice="AFE85001584|AFE02014928|AFE85009316|AFE85010014|AFE86000318|AFE86000603|AFE86001194"] expand to notice
```

Sous-corpus foule2581docs :

```
[frlemma="foule"] expand to notice
```

ou

```
<notice>[*[frlemma="foule"]]*</notice>
```

(Sous-corpus bbardot4docs) :

```
<notice>[_notice_identifiantdelanotice="AFE85003888|AFE85007743|AFE85010261|AFE86000480"] expand to notice
```

Sous-corpus bbardot15docs

```
[descripteursaffcol & word="Bardot"][][word="Brigitte"] expand to notice
```

(Sous-corpus bbardot30docs)

```
<notice>(([*[word="Brigitte"%c][*][word="Bardot"%c][*]) | [*][word="Bardot"%c][*][word="Brigitte"%c][*])</notice>
```

Prisonniers (42 textes) :

Détail du cheminement :

```
<descripteursaffcol>([*][word="Seconde" | word="allema.*"%c][*][frlemma="prisonnier"]][* | [*][frlemma="prisonnier"]][*][word="Seconde" | word="allema.*"%c][*])</descripteursaffcol>
```

AFE85006346 "prisonniers" "de" "guerre" dans le titre

```
[frlemma="prisonnier"][frlemma="de"] [frlemma="guerre"] within titrepropre
```

→ ramène 4 : 1 déjà récupéré, 2 bons (AFE04014373, AFE85006346), 1 mauvais

Donc finalement la requête suivante qui sélectionne 42 notices :

```
<notice_identifiantdelanotice="AFE04014373|AFE85006346">[]+</notice> | (<notice>[*<descripteursaffcol>([*][word="Seconde" | word="allema.*"%c][*][frlemma="prisonnier"]][* | [*][frlemma="prisonnier"]][*][word="Seconde" | word="allema.*"%c][*])</descripteursaffcol>[*]</notice>)
```

ou (a priori plus robuste mais même résultat) :

```
<notice_identifiantdelanotice="AFE04014373|AFE85006346">[]+</notice> | (<notice>[*<descripteursaffcol>([*][word="allema.*|mondiale"%c]
```

```
[ ]*[frlemma="prisonnier"][]* | [ ]*[frlemma="prisonnier"]
[ ]*[word="allema.*|mondiale"%c][ ]*)</descripteursaffcol>[]*</notice>)
```

Commémoration (55 textes) :

```
<notice>[]*<descripteursaffcol>([ ]*[word="armistice|victoire|1914|1918|
1940|1945|11"%c][ ]*[frlemma="commémoration"][]* |
[ ]*[frlemma="commémoration"][]*[word="armistice|victoire|1914|1918|1940|
1945|11"%c][ ]*)</descripteursaffcol>[]*</notice>
```

intrus : AFE85005609 ?

On pourra alors lister les participants :

```
<item_type="PAR">[]+</item> | <item_type="DEI">[]*[word=", "][ ]*</item> |
<item_type="DEI">[word="[A-Z].+"][]*[frlemma="d[eu]"][]*</item>
```

(Attention seulement 22 textes sur 55 ont des "participants". Comment on peut le voir :

```
<notice>[]*<descripteursaffcol>([ ]*[word="armistice|victoire|1914|1918|
1940|1945|11"%c][ ]*[frlemma="commémoration"][]* |
[ ]*[frlemma="commémoration"][]*[word="armistice|victoire|1914|1918|1940|
1945|11"%c][ ]*)</
```

```
descripteursaffcol><generiqueaffcol>[]*[_ .item_type="PAR"][]*</
generiqueaffcol></notice>
```

```
<notice>[]*<descripteursaffcol>([ ]*[word="armistice|victoire|1914|1918|
1940|1945|11"%c][ ]*[frlemma="commémoration"][]* |
[ ]*[frlemma="commémoration"][]*[word="armistice|victoire|1914|1918|1940|
1945|11"%c][ ]*)</descripteursaffcol></notice>
```

)

Rq. on peut passer les sous-corpus à qqn d'autre par "exporter", comme les partitions.

## 14. Partition

Découper des parties dans le corpus pour les contraster entre elles : par exemple des types de notices, des périodes, etc.

1. Simple :

- par type\_de\_date

→ déséquilibre

On peut observer cela de plus près :

```
<notice_typededate="Non diffusé">[]
<titrepropre>[_ .notice_typededate="Non diffusé" &
word!="chutes?"]+</titrepropre>
```

2. Assistée :

(- par genre :

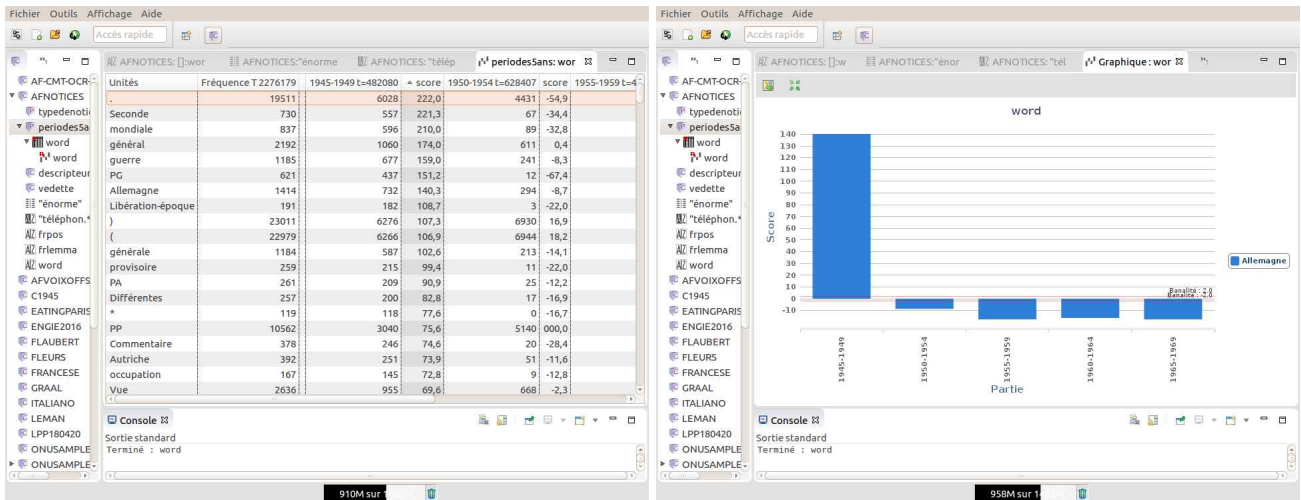
tout est "presse filmée", et une minorité de sujets sont quelque chose en plus (reportage, rétrospective, etc. -> 5 catégories mineures, certaines quasi vides -déclaration et rétrospectives). -> à observer en faisant une partition assistée sur sujet/genre et en ne créant que 5 classes en fonction du 2e genre.

Ce déséquilibre des genres fait que l'exploitation de la métadonnée sera sans doute très limitée pour les études contrastives.)

- periode5ans

# 15. Spécificités

Question de la répartition : quels sont les mots préférés -employés plus souvent qu'ailleurs- ou évités -employés moins souvent qu'ailleurs- dans chaque partie ?



**Tableau** : on trie sur le score d'une colonne pour voir les mots sur-représentés dans une **partie**

**Diagramme** : clic-droit sur une ligne (un mot) du tableau, pour voir le profil de répartition d'un mot sur l'ensemble des colonnes (parties).

Sur periodes5ans

générale (on lance directement le calcul depuis le cube [P] de la partition)

lecture des S+ d'une période en triant la colonne.

interprétation d'un score : probabilité, seuil à 3, + évaluation par rapport au corpus de référence.

lecture par ligne et affichage d'un diagramme en bâtons. Export.

(Intérêt (édition via InkScape) et limites (exploitation bureautique) de SVG, possibilité de paramétrage.)

Sur periodes5ans

[word=" - \* (VG | GP | PP | PM | PANO | Pano | PA | PG | GPP | PR | DP | VA | TRAV (EL) ? | VP | VSG | VE | ZAV) " ]

(sauvegarde de la table lexicale)

(Possibilité aussi de calcul des S+ d'un sous-corpus / corpus.)

Explications du calcul des spécificités : voir le Manuel utilisateur ; la FAQ sur le site de projet Textométrie (<http://textometrie.ens-lyon.fr>) ; l'article de référence de Pierre Lafon, voir site Textométrie, rubrique « Documents de référence ».

## 16. AFC

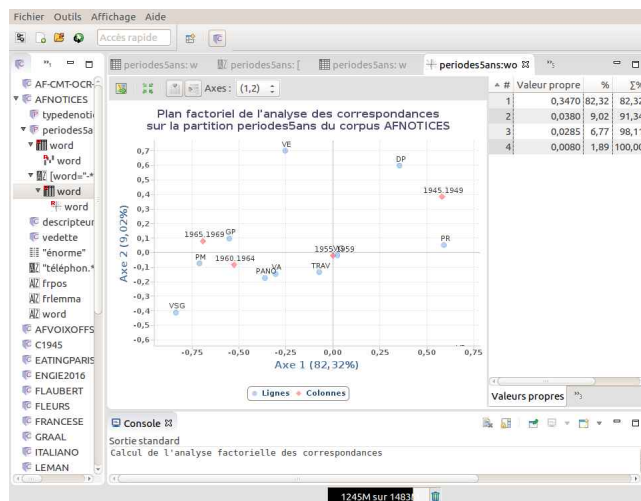
Une visualisation 2D, une sorte de cartographie, des ressemblances ou oppositions entre parties ou/et mots de ces parties.

Sur les valeurs de plan par période de 5 ans.  
Requête cf. § précédent

Sur le vocabulaire des titres par année:  
INDEX de  
[frpos="NOM"] within titrepropre  
avec Fmin=20  
TABLE LEXICALE marges = index  
- suppression des lignes des mots grammaticaux, M  
Monsieur MONSIEUR,...  
- fusion des lignes égales modulo la casse  
AFC : effet diachronique

retours aux S+ sur le même tableau pour  
l'interprétation

Attention c'est une projection, on perd des dimensions (vue « écrasée ») donc on ne peut pas se contenter de voir ce qui est proche/loin.  
Biblio pour en savoir plus : cf. site Textométrie (adresse au § précédent).



## 17. Cooccurrence

Quels mots sont « attirés » par un mot donné ?

1. Caractérisation des usages d'un mot  
[frlemma="foule"]  
contexte = notice, Fmin=Cmin=10  
Expliquer les paramètres et les colonnes.

2. Evaluer l'attraction possible entre deux mots :  
[frlemma="noir"]  
[frlemma="blanc"]  
(Possibilité de comparer avec FLEURS)

3. Aide à la construction d'un thème :  
"économi.\*"%cd  
"économi.\*|redressement|financ.\*"%cd  
etc.

Cooccurrent	Fréquence	Cofréquence	Indice	Distance moyenne
foule	4387	3170	1000	104,7
GAULLE	1984	1209	234	165,9
la	55156	17305	180	128,4
cortège	1015	678	162	125,5
Foule	1245	729	128	105,5
défilé	1262	689	102	116,7
acclamant	244	222	101	112,1
massée	207	197	100	75,0
saluant	829	498	94	114,4
drapeaux	873	508	88	131,2
de	130696	37058	88	142,2
annaludicant	1058	568	80	114,9

## 18. Progression

Un graphique (cumulatif) qui représente l'évolution au fil du corpus.

```
structure=text, propriété=annee  
[frlemma="foule"]  
"Bardot"%c
```

Etude diachronique des OPV :

on les liste d'abord en Index :

```
<item_type="OPV">[generiqueaffcol]  
+</item>
```

Petiot, Lucien	115
Becognee, Claude	84
Codur, Gilbert	83
Bakaes, Jacques	73

Du coup en progression on fait les requêtes :

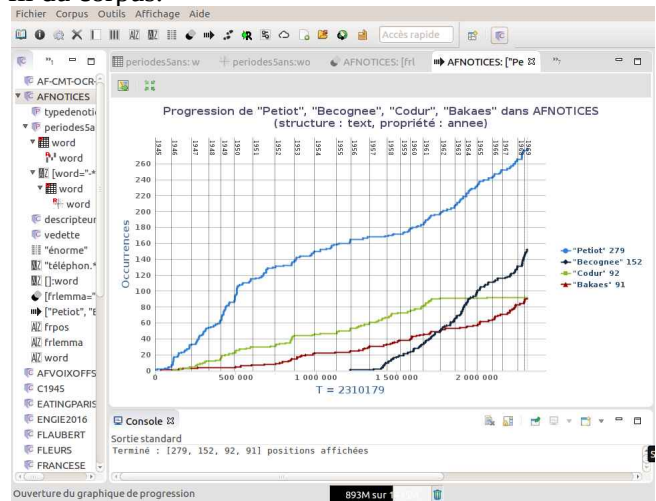
Petiot

Becognee

etc.

Etude de la diachronie du champ jour de semaine :

```
<notice_datedediffusionjoursemaine="lundi">[]
```



## 19. Divers

### Tokenisation

Gestion +/- bonne des tirets :

```
[word="- [^- ]+"]
```

```
[word="[^- ]+-"]
```

### CQL avancé

mode Greedy : SetMatchingStrategy → longest

Foule, complément du nom :

```
[(resume|sequences) & frlemma="foule"] ([frpos="ADV"]?[frpos="ADJ"]  
VER:pper"])? ([frpos="PRP.*"] [frpos="DET.*|ADV|ADJ|NOM|NAM|KON"] {1,10})
```

@

Foule, Ciblage des verbes conjugués associés :

```
[(resume|sequences) & frlemma="foule"][frpos!="VER.*" | frpos="VER:infi|  
VER:pper" | frlemma=".*être.*|avoir"]{0,10} @[frpos="VER.*" & frpos!  
="VER:infi|VER:pper" & frlemma!=".*être.*|avoir"]
```

## 20. Adresses

Taper textométrie dans un moteur de recherche → normalement on tombe sur le site du projet

<http://textometrie.ens-lyon.fr>

En haut à droite de la page d'accueil, lien vers tous les autres sites/ressources complémentaires :

- la liste des utilisateurs txm-users
- le wiki des utilisateurs,
- etc.