

# Atelier TXM

Notes de Bénédicte Pincemin



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

**Remarque :** Cette formation fait suite à la journée d'initiation à TXM du 15 octobre 2018, et aux mini-ateliers de la réunion Historiens du 10 juillet 2019. Des documents supports ont été rassemblés sur le sharedocs du projet :  
sharedocs.huma-num.fr > ANR > ANTRACT > SP2 Corpus & Recherche technologique > TXM > formation\_et\_aide

Le présent document n'entend donc pas être un support d'initiation, mais un aide-mémoire pour les participants à l'atelier, et un document de travail complémentaire pour les autres membres du projet intéressés.

**Avertissement :** 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 janvier 2020**. 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**.

## Préparation de l'ordinateur : logiciel et données

### Installation / mise à jour de TXM, et récupération du corpus AFVOIXOFFV01

Sujet : TXM 16 janvier : préparez vos ordis dès que possible !

De : Pincemin, Bénédicte

Date : 12/01/2020 à 18:30

Pour : antract-all@lists.eurecom.fr

Copie à : équipe TXM

Chers tous,

Ceci est un message pour les participants à la session TXM de jeudi.

L'objectif de cette journée sera de :

- (re)prendre en main ensemble TXM,
- à partir de vos problématiques de recherche sur le corpus des Actualités françaises, et

- en découvrant le nouveau corpus "AFVOIXOFF", qui intègre à la fois les transcriptions automatiques du commentaire parlé, les métadonnées documentaires de l'INA, et l'accès aux vidéos sur le serveur Okapi.

**Pour travailler dans les meilleures conditions, merci de faire tout votre possible pour préparer votre ordinateur à l'avance** et vérifier ainsi que tout se passe bien. La démarche à suivre est décrite ici :

[https://groupes.renater.fr/wiki/txm-info/public/chantier\\_antract#recette\\_etape\\_22](https://groupes.renater.fr/wiki/txm-info/public/chantier_antract#recette_etape_22)

Suivre les "étapes initiales" de la recette 2.2 pour avoir la dernière version de TXM (celle qui permet l'accès aux vidéos Okapi) et le nouveau corpus AFVOIXOFFV02. Vous pouvez bien sûr ensuite, si vous le souhaitez, poursuivre avec les sections suivantes de la recette pour commencer à découvrir les nouvelles fonctionnalités et vous assurer qu'elles marchent bien pour vous (la "recette" est ce qui permet de tester un nouveau programme avec les variétés de cas de figures d'environnement informatique et d'usages).

Idéalement, **commencer à préparer son ordinateur avant mardi**, dans un lieu où vous bénéficiez d'une bonne connexion internet (toute l'installation se déroule en mode connecté) :

- si vous rencontrez une difficulté de compréhension des instructions, peut-être pourrez-vous prévoir un moment en marge de l'atelier Okapi du mardi pour vous entr'aider ?
- si l'installation ne se passe pas comme prévu, et qu'il semble y avoir un problème de fonctionnement, vous pourrez contacter l'équipe TXM à l'adresse (en copie de ce mail) :

textometrie à [groupes.renater.fr](mailto:groupes.renater.fr)

(de préférence avant mardi après-midi).

## Récupération du corpus AFNOTICES

De : Pincemin, Bénédicte

Date : 16/01/2020 à 09:36

Pour : [antract-all@lists.eurecom.fr](mailto:antract-all@lists.eurecom.fr)

Pour l'atelier d'aujourd'hui (et d'une façon générale !)

Voici un lien pour (re)télécharger le corpus des Notices documentaires, celui que nous avons utilisé à la formation du 15 octobre 2018.

Il peut être complémentaire à celui de la VoixOff :

<https://sharedocs.huma-num.fr/wl/?id=JEgMRMjyvBjwISFuyZ3UqKrJcZCMvtpi>

Une fois le fichier téléchargé, on le rentre dans TXM avec la commande Fichier > Charger > Un corpus binaire

## Découverte du corpus AFVOIXOFFV02 : édition et retour à la vidéo

La commande EDITION (sélectionner le cube du corpus dans la marge gauche, puis commande EDITION, dont l'icône est le petit livre) va nous servir à voir ce que contient le corpus.

On présente son organisation :

- en *journaux* ou *émissions* d'une date donnée (structure <text>),
- subdivisé en *sujets* (structure <div>).

On peut feuilleter de journal en journal avec les boutons en bas à droite, de sujet en sujet avec les boutons en bas au centre.

Une page par sujet ? En tout cas un nouveau sujet commence toujours sur une nouvelle page.

Cliquer sur les notes de musique dans l'édition pour revenir à la vidéo.

Entrer le login + mot de passe du compte Antract (celui de mardi 14, ou celui communiqué au projet).

On règle la position d'ouverture de la fenêtre de vidéo via les préférences :

Edition > Préférences > TXM > Utilisateur > MédiaPlayer

Sinon on peut toujours déplacer la fenêtre où on veut (cliquer sur l'onglet et le déplacer/tirer dans la direction où l'on veut aller), mais en réglant la préférence cela permet que la fenêtre s'ouvre dès le départ à l'endroit qui est le plus pratique habituellement pour soi.

La commande Affichage > Réinitialiser l'affichage peut être utile si on a déplacé des fenêtres et qu'on ne s'y retrouve plus.

Voir la recette pour différents types de retour à la vidéo, en la suivant vous testez le bon fonctionnement de ces développements "tout frais" et en même temps vous découvrez les différents accès possibles :

[https://groupes.renater.fr/wiki/txm-info/public/chantier\\_antract#recette\\_etape\\_22](https://groupes.renater.fr/wiki/txm-info/public/chantier_antract#recette_etape_22)

L'édition nous permet aussi de voir les propriétés, les informations/métadonnées associées à chaque journal et à chaque sujet (normalement ce serait plutôt avec la commande PROPRIÉTÉS du corpus, onglet Détails, mais cela ne fonctionne pas avec la version de TXM que nous avons).

Les propriétés sont plus ou moins intéressantes selon le point de vue. Par exemple, les propriétés "antract-..." sont plus techniques (pour vérifier des données et des traitements de préparation du corpus). Pour l'étude historique on sera notamment plus particulièrement intéressés par date-de-diffusion et titre-propre par exemple.

L'identifiant de la notice est `identifiant-de-la-notice` pour le sujet (`<div>`), et `id` pour le journal (`<text>`).

Dans l'état actuel du corpus, certaines propriétés (non montrées dans l'édition, mais apparaissant dans diverses commandes de sélection d'informations) sont plus techniques (liées notamment aux contraintes de construction du corpus), dont certaines qui dupliquent le contenu d'autres propriétés et ne seront donc pas utiles pour nous, par ex. :

pour `<text>`, `title = titre-propre`

`subtitle = notes-du-titre`

pour `<div>`, `type = titre-propre`

`topic = date-de-diffusion`

## Complémentarité avec les notices et croisement de la transcription avec les textes des notices documentaires

Complémentarité de AFNOTICES et AFVOIXOFF : dans AFNOTICES, les "textes" de la description documentaire (Résumé, Séquences) sont interrogeables de façon plus souple et précise ; dans AFVOIXOFF, ce sont comme des chaînes de caractères, il n'y a pas de mots, ni de lemmes ni de catégories grammaticales.

Dans AFVOIXOFF cependant on peut croiser des informations des notices et de la voixoff.

Exemple : je cherche les transcriptions (par leur 1er mot) telles qu'il y a des "ruines" mentionnées dans les séquences :

Dans AFVOIXOFFV02, CONCORDANCE de  
<div\_sequences=".\*ruines.\*">[ ]

**Remarque générale sur l'écriture des requêtes** : dans TXM le guillemet doit être droit (") et non français par exemple (« ). Attention donc aux requêtes notées dans un traitement de texte, souvent celui-ci transforme les guillemets, et alors la requête ne fonctionne plus... il faut rétablir les guillemets droits (ou régler l'option du traitement de texte qui permet qu'il ne fasse pas cette transformation ici gênante).

Double-clic sur la ligne de concordance pour consulter les contenus des sujets ainsi repérés. Mais ensuite on n'a pas d'aide pour retrouver "ruines" à l'intérieur de la métadonnée sequences (le surlignage porte sur le 1er mot de la transcription, et pas de ctrl-f).

Remarque : ici cela marche bien parce que la chaîne de caractères "ruines" a peu d'ambiguïté, en tout cas dans le corpus.

Pour une recherche plus fine sur "ruine(s)", utiliser AFNOTICES, mais alors on n'a plus le commentaire transcrit :

Dans AFNOTICES, CONCORDANCE de  
[sequences & frlemma="ruine"]

Si on veut une seule ligne de résultat par notice, on peut écrire une équation qui cherche la première occurrence du mot "ruine" dans une séquence :

Dans AFNOTICES, CONCORDANCE de

<sequences> [frlemma != "ruine"]\* [frlemma="ruine"] within sequences

Cependant l'affichage est peu lisible (le pivot de la concordance contient tout le début de la séquence, ce qui est souvent long et pas toujours de la même longueur) ; on s'en sert en double-cliquant sur chaque ligne, ce qui donne un accès organisé à l'édition de chaque texte concerné.

Suite de la remarque précédente : on peut aussi tester dans AFNOTICES que "ruines" ne correspond pas à d'autres mots plus longs qui l'incluraient :

Dans AFNOTICES, INDEX en word de  
[sequences & word=".\*ruines.\*"]

Résultat :

ruines	360
ruines-	1

Ici tout se passe bien mais dans un cas plus défavorable, comme arme(s), il faudrait contraindre davantage la requête pour mieux marquer les frontières du mot, par ex. :

<div\_sequences=".\* armee?s? .\*"%cd>[ ]

(bien noter le blanc avant le « a » de « armee?s? » ; pour avoir les sujets dont la séquence parle d'arme(s) ou d'armée(s), sans avoir le "charme", les "larmes", les "arméniens", etc.)

Je voudrais n'avoir qu'une occurrence par sujet ? Alors je peux chercher la première :

Dans AFNOTICES, CONCORDANCE de

<sequences> [frlemma != "ruine"]\* [frlemma="ruine"] within sequences

Comme cela ramène un segment long et de longueur irrégulière, la concordance est peu lisible, mais elle sert de sommaire pour consulter les éditions des sujets correspondant (en double-cliquant sur les lignes les unes après les autres).

La structure du corpus AFNOTICES est différente :

<text> = une année complète, qui contient une suite de

<notice> (avec différentes propriétés intéressantes : datedediffusion,...), qui contient plusieurs parties plus textuelles successives :

<titrepropre>

<resume>

<sequences>

<descripteursaffcol>

<generique>

Ces deux derniers champs sont structurés en <list> avec des <item>.

Une page par notice dans l'ÉDITION.

## Concordance avec dates dans AFVOIXOFFV02

Dans AFVOIXOFF, si je fais une CONCORDANCE, j'aurais besoin de voir la date en référence :

Clic-droit dans la colonne de gauche "ref",

"Options d'affichage des références", et choisir dans la boîte de gauche les informations qui nous intéressent, par exemple :

text: date-de-diffusion

div: titre-propre

On peut aussi repasser "ref" de la boîte de droite à celle de gauche, pour être moins encombré. Puis valider.

L'autre commande disponible par clic-droit sur la 1ère colonne de la concordance, "Option de tri des références", peut être utilisée pour faire que le tri des références soit chronologique : choisir

text: textorder

Mais on pourrait aussi choisir de trier alphabétiquement par titre de sujet si on préfère, avec

div: titre-propre

Ces réglages sont valables pour l'onglet de Concordance courant, ils ne sont pas appliqués si on ouvre une nouvelle concordance, ni à la fin de la session quand on quitte TXM.

Cependant, avec TXM 0.8, on peut conserver un résultat pour le retrouver après avoir quitté TXM, à la session suivante : clic-droit sur le résultat dans la vue Corpus (marge de gauche) et commande « conserver ». Le nom du résultat passe en caractères droits (au lieu des italiques).

On découvre que c'est une astuce pour garder un onglet concordance avec le réglage des références qu'on a redéfini, sans être obligé de recommencer la manoeuvre de réglage à chaque nouvelle concordance (onglet ou session).

La commande "Conserver" enregistre le résultat par ses paramètres (par ex. par sa requête), elle ne retient pas les retouches manuelles faites ensuite, par exemple la suppression de certaines lignes en concordance, ou la fusion/suppression de lignes/colonnes dans une table lexicale. Dans ce dernier cas, la conservation de la table après retouches se fera par les commandes « exporter » > Données ; la table ne sera pas gardée dans TXM à la fermeture de la session, mais elle pourra être rechargée en l'état. Se reporter au Manuel utilisateur, § 8.10.1 Sauvegarde d'une table lexicale.

## Question de Fabien sur la représentation graphique de la répartition quantitative des mots

Pour l'INDEX, si on veut calculer des % d'occurrences représentées ou cumulées, et construire un diagramme en bâtons, exporter le résultat et le retravailler dans un tableur (comme Calc ou Excel).

## Recherche de Pascale autour de la cooccurrence de « foule » et « émotion »

Je veux rechercher ce qui concerne (mentionne) émotion(s)

et "foule" dans le même sujet.

Je mets au point la requête pour les émotions :

Dans AFVOIXOFFV02, INDEX en word de

[word = "émoti.\*"%cd]

trouve :

émotion	91
émotions	27
émotif	2

Dans AFNOTICES, la même commande donne :

émotion	6
EMOTIONS	2
émotionné	1

On voit l'intérêt de la troncature (. \* : n'importe quel caractère, n'importe quel nombre de fois) et de la neutralisation de la casse (%c) et des diacritiques (%d), (en combinaison : %cd).

Je construis la requête pour rechercher les deux mots progressivement. D'abord j'utilise l'assistant de requête pour en écrire le squelette (dans AFNOTICES) :

[word = "émoti.\*"%cd][frlemma = "foule"]

Je commence à la structurer en ajoutant un OU (le baton vertical, AltGr-6 en PC, Alt-L en Mac) et des parenthèses extérieures :

([word = "émoti.\*"%cd][frlemma = "foule"] | [frlemma = "foule"][word = "émoti.\*"%cd])

J'ajoute la possibilité de mots qui s'insèrent (**attention, ne pas lancer la requête en l'état, à ce stade**) :

([word = "émoti.\*"%cd][ ]\* [frlemma = "foule"] | [frlemma = "foule"] [ ]\*[word = "émoti.\*"%cd])

et surtout, **avant de lancer la requête**, j'ajoute une contrainte pour que la requête ne s'étende pas indéfiniment à cause des [ ]\*, en indiquant qu'on doit rester dans la notice :

([word = "émoti.\*"%cd][ ]\* [frlemma = "foule"] | [frlemma = "foule"] [ ]\*[word = "émoti.\*"%cd]) **within notice**

Transposition au corpus AFVOIXOFFV02 :

CONCORDANCE de

([word = "émoti.\*"%cd][ ]\* [frlemma = "foule"] | [frlemma = "foule"] [ ]\*[word = "émoti.\*"%cd]) **within div**

C'est seulement la toute fin de la requête qui change, les sujets ne sont plus des <notice> mais des <div>.

Il n'y a pas tant que ça d'occurrences pour les émotions, il faudra sans doute chercher d'autres mots liés à ce thème.

Plusieurs fonctions de TXM peuvent aider :

- on peut faire une liste des mots du corpus par fréquence décroissante, et parcourir ainsi l'essentiel du vocabulaire présent en corpus. Par exemple en se focalisant sur un type de mots (nom ici, mais cela pourrait être Adjectif, verbe, adverbe...)

Sur le corpus AFVOIXOFFV02, INDEX en frlemma de [frpos="NOM"]

- 2e aide possible, les COOCCURRENCES : quand on parle d'émotions, y a-t-il d'autres mots qui viennent avec ?

La récolte ici n'est pas très bonne (peu de scores à 3 ou plus), on est un peu aux limites de la méthode avec un pivot de fréquence 120. En dessous de 100, on utilise la CONCORDANCE pour étudier les voisinages des mots.

Rq. A posteriori, je m'aperçois qu'on aurait pu déjà creuser davantage la famille de mots autour d'"émotion" en s'aidant de l'INDEX :

Sur AFVOIXOFFV02, INDEX sur word de "ém.\*"

(autrement dit, un bout de radical ouvrant une recherche très large)

On trie alphabétiquement la colonne des mots trouvés (en cliquant sur son en-tête).

On voit que pourraient nous intéresser aussi variations autour de "émouv-" et de "ému".

On construit alors une requête qui cible les différentes "branches" de la famille :

```
"émoti.*|émouv.*|émoi|émue?s?"
```

et on remet pour finir le %cd pour attraper quelques occurrences supplémentaires typographiées autrement :

```
"émoti.*|émouv.*|émoi|émue?s?"%cd
```

On obtient 298 résultats, au lieu des 120 avec le seul

```
"émoti.*"%cd
```

En fait il semble y avoir un **bug avec le %d** (il y a des intrus dans les mots trouvés : émoluments, émoussé, émousser) qui de plus n'apporte rien dans ce cas précis, donc on va s'en tenir à la requête :

```
"émoti.*|émouv.*|émoi|émue?s?"%c
```

(289 résultats)

## Recherche de Franck, sur les enchaînements de plans cinématographiques

Pour la recherche de Franck, petite expérimentation de l'annotation : on ouvre une concordance dans AFNOTICES, et on veut annoter les GP

avec une nouvelle propriété :

Concordance de <GP> dans le corpus AFNOTICES...

12660 occurrence(s).

Affectation de la valeur gros à la propriété valeurdeplan pour 76 occurrences.

Fail to save annotations of the corpus.

→ en fait l'annotation est finalement sauvée (après un certain délai), a posteriori on la retrouve et elle est interrogeable, ex.

Dans AFNOTICES, CONCORDANCE de

```
[valeurdeplan = "gros"]
```

Cependant les éditions de certaines années ont perdu toutes leurs indications de métadonnées et de structures.

À creuser, mais il n'est pas sûr que l'on va suivre cette voie par l'annotation.

On peut sinon de toutes façons rechercher par exemple les séquences de mots les plus répétées dans les séquences, par exemple les suites de 4 mots :

Dans le corpus AFNOTICES, INDEX de

```
[(resume|sequences) & word!="\p{P}+"][word!="\p{P}+"][word!="\p{P}+"][word!="\p{P}+"][(resume|sequences) & word!="\p{P}+"]
```

Cela donne :

d' un groupe de	567
VG de la foule	439
dans les rues de	390
l' Arc de Triomphe	344
dans une rue de	296

etc.

On peut mettre au point des requêtes qui vont chercher des formes variées d'appellation d'un plan, par exemple on peut chercher des plans désignés par les appellations "panoramique(s)" ou "plan(s) large(s)" ou "gros plan(s)" etc. (on peut allonger la requête en ajoutant d'autres désignations) :

Dans le corpus AFNOTICES, CONCORDANCE de  
`[frlemma = "panoramique"] | ([frlemma = "plan"] [frlemma = "large"]) | ([frlemma = "gros"] [frlemma = "plan"])`

## Études chronologiques

On veut diviser le corpus AFVOIXOFFV02 en périodes.

On sélectionne le corpus et on lui applique la commande PARTITION. Comme il n'y a pas de métadonnée donnant directement les années par exemple, on va construire les parties via des requêtes sur la métadonnée date-de-diffusion.

On va dans l'onglet "Avancé" :

Saisir le nom de la partition dans le champ "Nom", par exemple  
 Décennies

Saisir le nom de la première partie dans le champ où il ya "Partie 1", on remplace "Partie 1" par "années40" par exemple.

Dans le long champ qui suit mettre la requête qui définit ce qui est dans la partie, donc ici ce qui sélectionne les années 40 :

`[ _ .text_date-de-diffusion="..../194." ] expand to text`

Puis le petit "+" en bas à gauche de la boîte de dialogue permet d'ajouter de quoi définir une 2e partie, puis une 3e, etc.

`[ _ .text_date-de-diffusion="..../195." ] expand to text`

`[ _ .text_date-de-diffusion="..../196." ] expand to text`

On peut adapter les équations pour définir des périodes plus finement, par exemple pour des tranches de 5 ans :

partie 1 = 1945-1949

requête 1 = `<text_date-de-diffusion="..../194[5-9]">[ ] expand to text`

partie 2 = 1950-1954

requête 2 = `<text_date-de-diffusion="..../195[0-4]">[ ] expand to text`

...

partie 5 = 1965-1969

requête 5 = `<text_date-de-diffusion="..../196[5-9]">[ ] expand to text`

Si on a besoin de d'énumérer une à une des années, on peut le faire en s'inspirant du modèle suivant :

`<text_date-de-diffusion="..../19(49|50|51)">[ ] expand to text`

Attention à bien mettre chaque année dans une requête de partie et une seule (si possible pas d'oubli, et surtout pas de recouvrements).

Pour une partition par mois (tous les mois de janvier ensemble dans la même partie, etc.), pour voir ce qui est saisonnier, la partition n'est pas non plus difficile à construire :

partie 1 = janvier

requête 1 = `<text_date-de-diffusion="../01/...">[ ] expand to text`

partie 2 = février

requête 2 = `<text_date-de-diffusion="../02/...">[ ] expand to text`

...

partie 12 = décembre  
requête 12 = <text\_date-de-diffusion="../12/....">[] expand to text

Ensuite, on peut exploiter la partition Décennies par exemple pour voir si la mention d'émotion est uniforme ou évolue au fil des décennies.

Dans le corpus AFVOIXOFFV02, sur la partition Décennies,

INDEX en frlemma de

"émoti.\*|émouv.\*|émoi|émue?s?"%c

Sélectionner le résultat dans la marge gauche et lancer la commande TABLE LEXICALE sur cet index, en laissant la sélection par défaut "Calculer les marges à partir des fréquences de tous les mots du corpus"

Sélectionner toutes les lignes (avec la touche ctrl appuyée) concernant l'émotion et les fusionner en une seule qu'on peut appeler EMOTION par exemple

Sur la table qui ne compte plus maintenant que deux lignes (EMOTION et #RESTE#), demander un calcul de SPÉCIFICITÉS.

Dans le tableau des résultats des spécificités, sélectionner la ligne EMOTION et faire un clic-droit pour lancer la commande pour afficher le diagramme correspondant :

on obtient une visualisation de l'évolution fréquentielle du mot selon les périodes.

Rq. attention à l'échelle (axe vertical à gauche) et aux lignes rouges à -2 et 2 : entre ces lignes, le score n'est pas vraiment significatif. Donc ici, bien que les bâtons "montent" et "descendent", le calcul ne montre pas de variation pertinente au fil des 3 décennies.

Pour mémoire, dans l'atelier du 15 octobre 2018, nous avons vu encore une autre visualisation, par l'AFC (analyse factorielle des correspondances) ; mais il faut un peu de temps pour bien l'expliquer et la prendre en main (l'interprétation doit se baser notamment sur les tableaux d'indicateurs à droite du graphique).