

Mini-ateliers TXM

Notes de 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 juillet 2019**. 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.

Recherche d'un mot

Scénario général :

- sélectionner le corpus AFNOTICES
- demander une concordance
- saisir le mot dans le champ requête et lancer la recherche
- observer les résultats dans l'ordre chronologique (cf. dates dans la colonne de gauche), ou trier sur les contextes (à droite ou à gauche).
- double-clic sur une ligne de concordance pour mieux voir les occurrences dans le contexte de la notice (cliquer sur l'onglet de l'édition et la tirer vers la droite de la fenêtre pour avoir un affichage plus confortable).

Exemples de requêtes :

musée
cinéma
"cinéma"%cd

Pour la **recherche d'une expression**, penser à utiliser l'assistant de requête (baguette magique à gauche du champ requête)

Selon le besoin, faire une recherche en **lemme** (frlemma = forme normalisée du mot qu'on chercherait dans le dictionnaire) ou en chaîne de caractères (« word », avec des opérateurs %cd, des troncatures, des caractères facultatifs notés avec ?, etc.)

Exemples de requêtes :

```
[frlemma = "prisonnier"] [frlemma = "de"] [frlemma = "guerre"]  
[frlemma="d(octeu)?r"%cd][word="Petiot"%cd]
```

La recherche d'« auto-gestion » dans AFNOTICES est infructueuse, aucune des requêtes suivantes de donne de résultats :

```
"autogestion"%cd  
"auto-?gestion"%cd  
".*gestion.*"
```

Le lemme n'est pas toujours avantageux, par ex. une recherche sur

```
[fr lemma="ménager"]
```

trouvera les "arts ménagers" (pluriel de l'adjectif) mais pas la "ménagère", car le nom "ménagère" n'est pas la forme féminine d'un mot "ménager". Pour trouver aussi la ménagère il faudra alors une requête du type :

```
"ménagere?s?"%cd
```

Une troncature est peut-être plus simple

```
"ménag.*"%cd
```

mais attrape aussi ménage, ménagerie, ménagement...

(On peut utiliser des troncatures à n'importe quel endroit du mot, pas seulement à la fin mais aussi au début ou/et au milieu, ex.

```
".*ménag.*"%cd
```

attrape aussi les déménagements, les aménagements, le surmenage, entre autres...

Attention, **particularités dans le corpus AFVOIXOFFV01** :

- frlemma devient frplemma, idem frpos -> frppos, ex.

```
[frplemma = "prisonnier"] [frplemma = "de"] [frplemma = "guerre"]
```

- il n'y a pas de majuscules, ainsi pour rechercher la ville de Cannes, on a cherché :
cannes

Observer dans un **INDEX** ce qu'attrape une **requête complexe**.

Par exemple la requête

```
"cinéma"%cd
```

permet de trouver les occurrences avec des variations d'écriture en casse (majuscule/minuscule, c'est le %c) et au niveau des accents (diacritiques, c'est le %d) :

cinéma	642
CINEMA	154
Cinéma	64
cinema	5
CINÉMA	4

C'est ainsi que la recherche sur le "yéyé" s'affine petit à petit, on commence par regarder en INDEX ce que rapporte

```
. *yé . *
```

puis

```
yé . *
```

puis on cible de plus en plus précisément les formes qui nous intéressent :

```
yé-?yé  
"yé(-?yé)?"%cd
```

Le langage de requête permet de **croiser la recherche avec une métadonnée**, par exemple on a fait une concordance de

```
[word="cinéma"%cd & _.notice_typededate="Non diffusé"]
```

dans le corpus AFNOTICES, pour trouver les mentions du cinéma dans des notices de sujets catalogués comme non diffusés, et on a obtenu 12 occurrences.

Dans le cas de la recherche des "procès" qui **ne** soient **pas** liés à Pétain ou Nuremberg, on a d'abord expérimenté des requêtes du type :

```
[fr lemma="procès"] [][word!="Nuremberg"%d]  
[fr lemma="procès"] [word!="Pétain"] [word!="Nuremberg|Pétain"%d] [word!="Pétain"]  
[fr lemma="procès"] [word!="Pétain"%cd] [word!="Nuremberg|Pétain"%cd] [word!  
="Pétain"%cd]
```

mais il apparaît plus simple de construire un **sous-corpus** de notices qui ne contiennent ni Pétain ni Nuremberg, puis dans ce sous-corpus de faire une recherche sur les procès.

Si on fait un sous-corpus avancé avec la requête :

```
<descripteursaffcol>[word!="Pétain|Nuremberg"%cd]+</descripteursaffcol> expand  
to notice
```

on voit que les descripteurs ne sont pas assez complets car on retrouve des sujets sur le procès Pétain, il faut sans doute faire un sous-corpus plus restrictif comme :

```
<notice>[word!="Pétain|Nuremberg"%cd]+</notice>
```

Pour vérifier la présence dans le corpus AFNOTICES d'une **notice** dont on connaît l'identifiant, on a fait une requête comme ceci, dans une CONCORDANCE :

```
<notice>[_notice_identifiantdelanotice="AFE02000438"]
```

ou (autre exemple) :

```
<notice>[_notice_identifiantdelanotice="AFE85001251"]
```

La 1ère requête montre que la notice est absente, alors que pour la 2nde on trouve la notice correspondante : double-cliquer sur la ligne de concordance obtenue pour visualiser la fiche.

Synthèse de l'environnement d'un mot par ses cooccurrences

Scénario général :

- sélectionner le corpus AFNOTICES
- demander des cooccurrences
- saisir le mot (dit "pivot") dans le champ requête
- dans le corpus AFNOTICES, il est judicieux aussi de modifier le paramétrage du contexte : demander un contexte en "structure" (et non en "forme"), choisir la structure "notice" dans la liste déroulante, et inverser les coches des cases suivantes, à savoir : décocher les deux "Utiliser le contexte..." et cocher "Inclure la structure..."
- Pour lire les résultats : les chiffres sont, dans l'ordre : la fréquence du mot dans le corpus, le nombre de fois où il apparaît avec le mot pivot, le score statistique (ex. 5 → 1 chance sur 100 000 (1 suivi de 5 zéros)), la distance moyenne. Possibilité de voir les contextes où les mots apparaissent ensemble en double-cliquant sur la ligne.

Retour à la vidéo

Cela concerne le corpus des transcriptions AFVOIXOFFV01 : il n'est pas encore finalisé -bug dans les métadonnées, pb de doublons à régler-, mais on peut l'utiliser pour explorer les possibilités que cela va apporter.

Le retour à la vidéo se fait depuis une concordance, par un clic-droit sur une ligne de concordance, on accède à la commande "Jouer le média".

Cependant pour que cette commande soit disponible, il faut avoir installé l'extension MediaPlayer, disponible via le menu

Fichier > Ajouter une extension

(il faut être connecté au réseau internet) (voir manuel utilisateur TXM si besoin pour l'installation de l'extension, mais a priori c'est intuitif).

Par ailleurs, il faut avoir récupéré les fichiers de vidéos (par ftp sur le site de l'INA) et les avoir mis à l'endroit où TXM va les chercher :

- soit un répertoire "media" dans le répertoire du corpus :

<répertoire utilisateur>/TXM-0.8.0/corpora/AFVOIXOFFV01/media

(ou même chose avec TXM au lieu de TXM-0.8.0 si vous avez une ancienne version de TXM, la 0.7.9 que nous avons utilisée pour la formation en octobre 2018)

- soit un lien symbolique pour indiquer que le contenu de ce répertoire se trouve sur un DD externe.

Evidemment, tout ceci est un peu technique, et la perspective à moyen terme est que TXM puisse directement aller lire les vidéos sur le site de l'INA, une fois que l'utilisateur a entré un login/mot de passe permettant de sécuriser l'accès à ces données.