

TXM

École Normale Supérieure de Paris, 3 octobre 2019.

Le cours a été conçu par Simon Gabay (UniNe) et Giovanni Pietro Vitali (University of Cork).

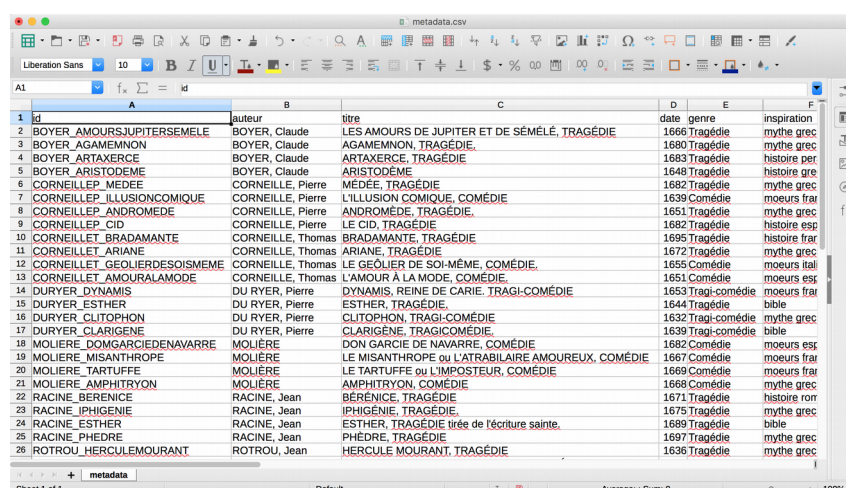
Le corpus a été réuni par Jean-Baptiste Camps (Ecole des Chartes) à partir de documents disponibles en ligne [www.theatre-classique.fr], et a été adapté pour le le présent cours.

Le présent cours est sous licence CC-BY 2.0 [<https://creativecommons.org/licenses/by/2.0/fr>]

Construction du corpus

Le corpus contient :

- 1) Des fichiers XML-TEI (attention à la construction des titres)
- 2) Un fichier “metada.csv”, avec un *id*, et différentes informations (auteur, titre, genre, *etc.*)



id	auteur	titre	date	genre	inspiration
1	BOYER	AMOURS JUPITERSEMELE	1666	Tragédie	mythe grec
2	BOYER	AGAMEMNON	1680	Tragédie	mythe grec
3	BOYER	ARTAXERCE	1683	Tragédie	histoire per
4	BOYER	ARISTODEME	1649	Tragédie	histoire gre
5	CORNEILLE	MEDEE	1682	Tragédie	mythe grec
6	CORNEILLE	L'ILLUSION COMIQUE, COMEDIE	1639	Comédie	moeurs fra
7	CORNEILLE	ANDROMÈDE, TRAGÉDIE.	1651	Tragédie	mythe grec
8	CORNEILLE	CID	1682	Tragédie	histoire esp
9	CORNEILLE	BRADAMANTE, TRAGÉDIE	1695	Tragédie	histoire fra
10	CORNEILLE	ARIANE, TRAGÉDIE	1672	Tragédie	mythe grec
11	CORNEILLE	LE GEOLIER DE SOI-MÊME, COMÉDIE.	1655	Comédie	moeurs itali
12	CORNEILLE	L'AMOUR À LA MODE, COMÉDIE.	1651	Comédie	moeurs esp
13	DURUYER	DYNAMIS	1653	Tragi-comédie	moeurs fra
14	DURUYER	ESTHER	1644	Tragédie	bible
15	DURUYER	CLITOPHON	1632	Tragi-comédie	mythe grec
16	DURUYER	CLARIGÈNE, TRAGICOMÉDIE.	1639	Tragi-comédie	bible
17	MOLIERE	DON GARCIE DE NAVARRE, COMÉDIE	1682	Comédie	moeurs esp
18	MOLIERE	LE MISANTHROPE ou L'ATRABILAIRE AMOUREUX, COMÉDIE	1667	Comédie	moeurs fra
19	MOLIERE	TARTUFFE	1669	Comédie	moeurs fra
20	MOLIERE	AMPHITRYON, COMÉDIE	1668	Comédie	mythe grec
21	RACINE	BÉRÉNICE, TRAGÉDIE	1671	Tragédie	histoire rom
22	RACINE	IPHIGÉNIE, TRAGÉDIE.	1675	Tragédie	mythe grec
23	RACINE	ESTHER, TRAGÉDIE tirée de l'écriture sainte.	1689	Tragédie	bible
24	RACINE	PHÈDRE, TRAGÉDIE	1697	Tragédie	mythe grec
25	ROTROU	HERCULE MOURANT, TRAGÉDIE	1636	Tragédie	mythe grec

Dans notre cas, il s'agit d'un corpus de 36 pièces de théâtre du XVII^{ème} siècle français dont le texte a été modernisé.

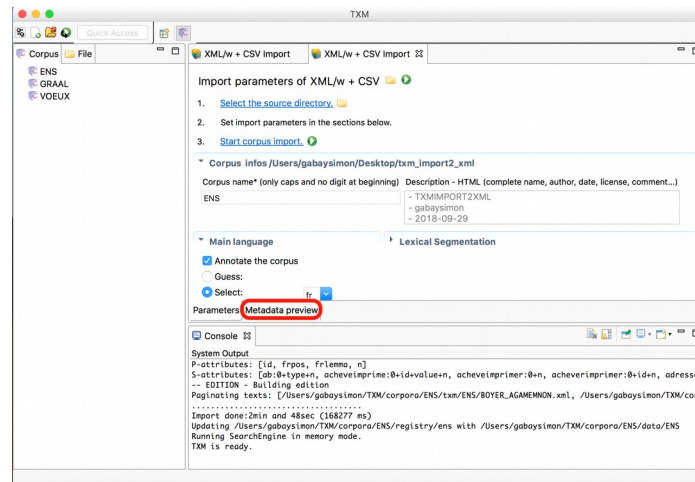
Attention, l'ouverture du fichier avec excel peut poser des problèmes : il faut importer le fichier CSV au format UTF-8 pour éviter de parler d' *IPHIG*√*â**NIE* ou de *TRAG*√*â**DIE*... Ou bien utiliser Libreoffice, qui a le double avantage d'être plus performant et... gratuit !

Importation du corpus

Ouvrir TXM, puis

- 1) File>Import>XML/w+CSV

- 2) Choisir le fichier où se trouve le corpus
- 3) Dans *corpus info*, donner un nom en majuscule uniquement, sans chiffre comme premier caractère.
- 4) Dans *main language*, sélectionner “Annotate the corpus” et sélectionner “fr” (si vous travaillez un texte en français contemporain). Cela va permettre l’étiquetage morpho-syntaxique (*POS tagging*, pour *Part Of Speech tagging*) avec TreeTagger.
- 5) Dans l’onglet en bas, au dessus de la console, sélectionner “metadata preview” pour voir si le fichier csv est bien lu, puis revenir à “parameters”.



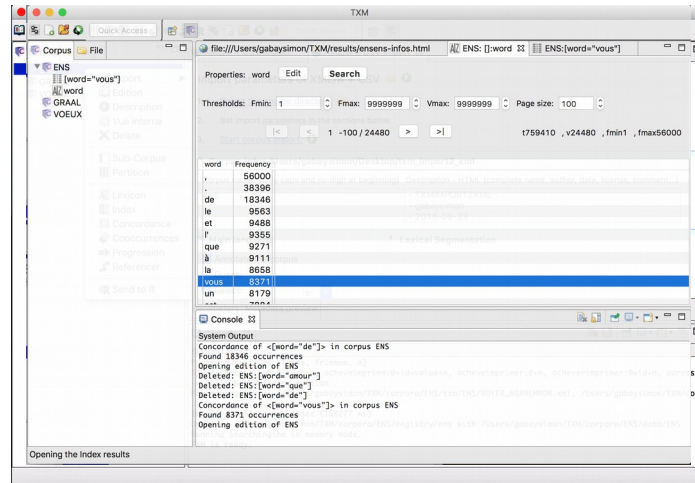
- 6) *Start corpus import*. Observer le log.

C’est fait, nous avons un corpus dans TXM !

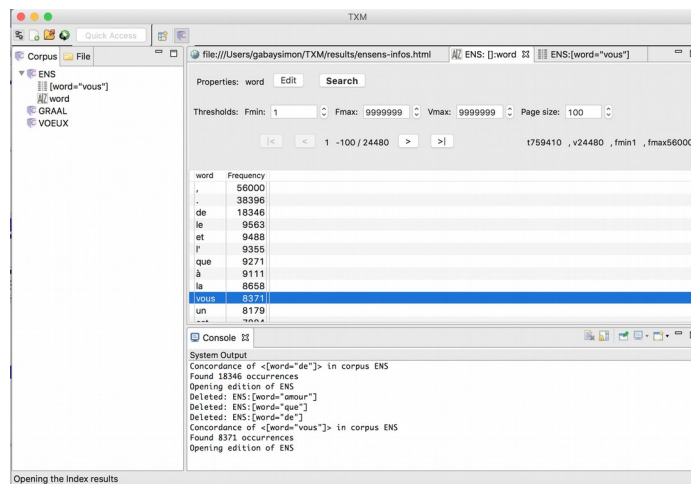
Première exploration du corpus : tour du propriétaire

Dans la liste de droite, notre corpus “ENS” est apparu.

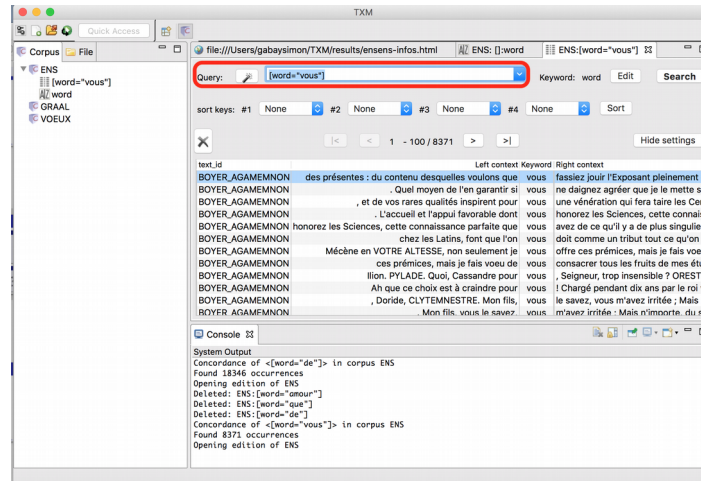
- 1) Allons sur le nom de notre corpus, et faisons un click droit dessus.



- 2) Sélectionner “Description”. Observez : vous y retrouvez des informations sur la taille du corpus, le nombre de words (en fait des tokens), et des informations sur le code XML (éléments et attributs).
- 3) Sélectionner “Lexicon”, observez la liste de fréquence, puis cliquez sur l’entrée “vous”.

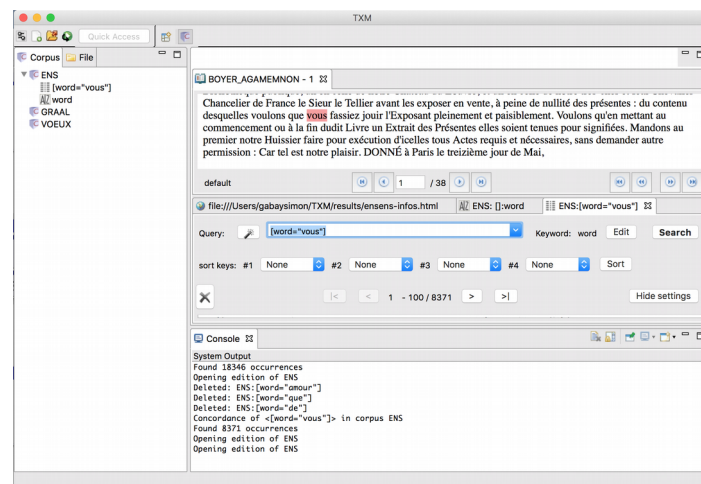


- 4) Nous obtenons la concordance du mot. Observez le champ “Query” : C’est une requête en CQL correspondant au résultat donné *infra*. Nous y reviendrons.



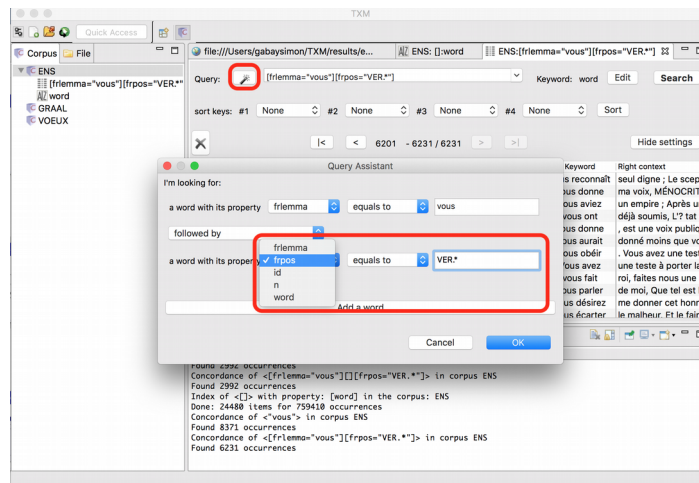
Note : Nous pouvons obtenir directement ce résultat en allant allant sur le nom du corpus, en cliquant sur “concordance” et en faisant directement la requête en CQL.

- 5) Double-cliquez sur la première entrée de la concordance : la transcription apparaît ! Nous sommes retourné à notre corpus, la (première) boucle est bouclée.



Seconde exploration du corpus : CQL

- 1) Repartons du “Lexicon” de notre corpus.
- 2) Tapez “vous” dans le champ “Query”. Observez.
- 3) Cliquez sur la petite baguette magique. Vous pouvez sélectionner si vous cherchez un mot (*word*), une catégorie grammaticale (*frpos*), un lemme (*frlemma*), et même combiner différents types de requêtes. Ce *Query Assistant* vous évite d’écrire vous même la requête en CQL (qui apparaîtra à côté de la baguette magique). N’oubliez pas de cliquer sur “Search”.

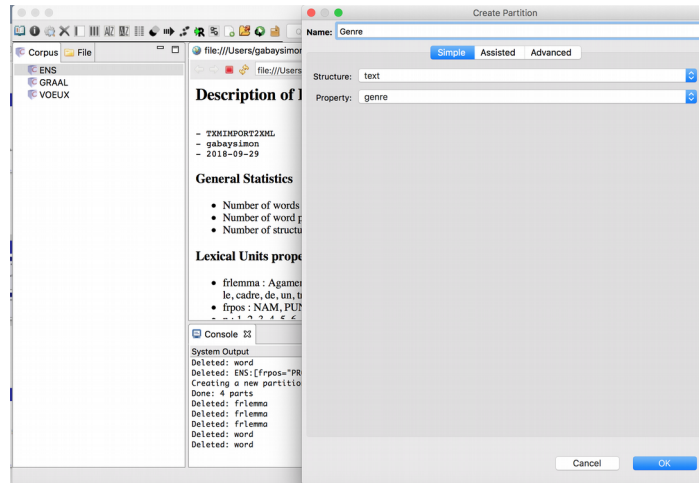


Cherchez avec le *Query Assistant* (en observant la requête en CQL) :

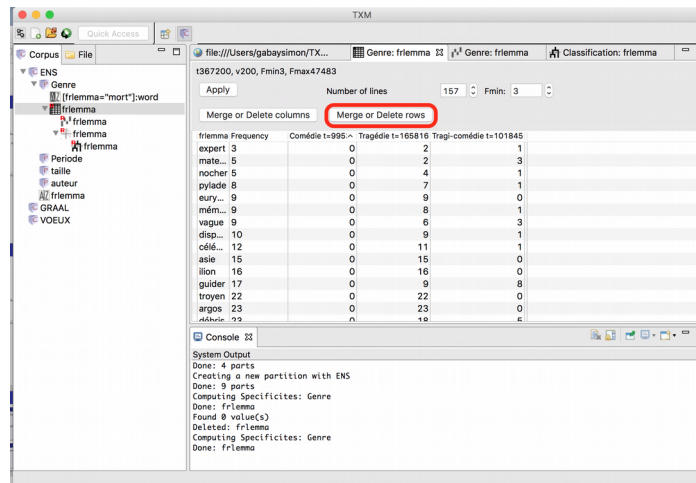
- a) Le pronom *vous* suivi d'un verbe
- Note: TXM utilisant Treetagger, nous sommes dépendant de son jeu d'étiquettes et de son lexique de lemmes pour le français. Le jeu d'étiquettes est fourni avec le cours.
- b) Le pronom *vous* suivi d'une virgule, puis d'un verbe
 - c) Un pronom suivi d'une virgule ou d'un point virgule, puis d'un verbe
 - d) Un pronom suivi d'une marque de ponctuation quelconque, puis d'un verbe

Troisième exploration du corpus : textométrie

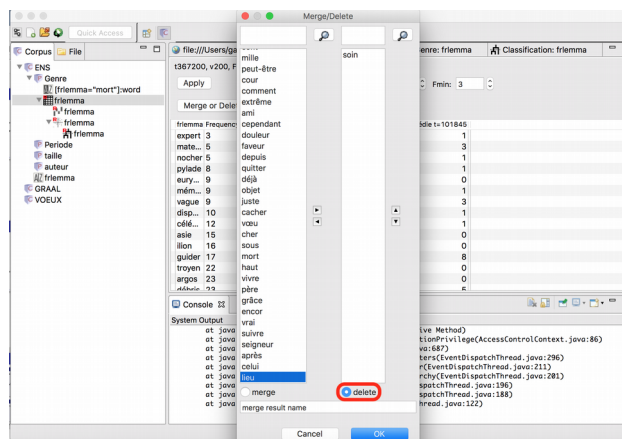
- 1) Rappelons-nous que nous avons renseigné des métadonnées : utilisons-les !
- 2) Cliquez sur le nom du corpus, et sélectionner "partition" (remarquez que vous pouvez aussi créer des sous-corpus).
 - a) Donnez un nom à votre partition (par exemple "Genre")
 - b) Sélectionner "Text" pour la "Structure" (qui renvoie à la structure XML).
 - c) Sélectionner "Genre" pour la "Property" (qui renvoie aux métadonnées du csv).



- d) Recommencer avec “période”, “auteur” comme “Property”.
- 3) Observez les dimensions de chaque partition pour contrôler votre corpus.
- 4) Explorez l’index des partitions (par exemple “mort” ou “sang” pour la partition “genre”).
- 5) Créez une table lexicale, et supprimez les *function words*
 - a) Cliquez sur “Merge and delete rows”



b) Cochez “Delete” et double-cliquez sur les mots que vous souhaitez retirer (*de, à, un, ou, et, le, la...*).



- 6) Recommencer en ne gardant que les *function words* (*de, à, un, ou, et, le, la...*) – c’est le principe de la stylométrie.
- 7) Calculez les “Specificities”, puis une “Classification” à partir du nom du corpus.
- 8) Essayez de repérer des “tics” de langue de certains auteurs (un indice, regardez les interjections comme “ah”).