

Formation Edition numérique

Exploration textométrique

Simon Gabay

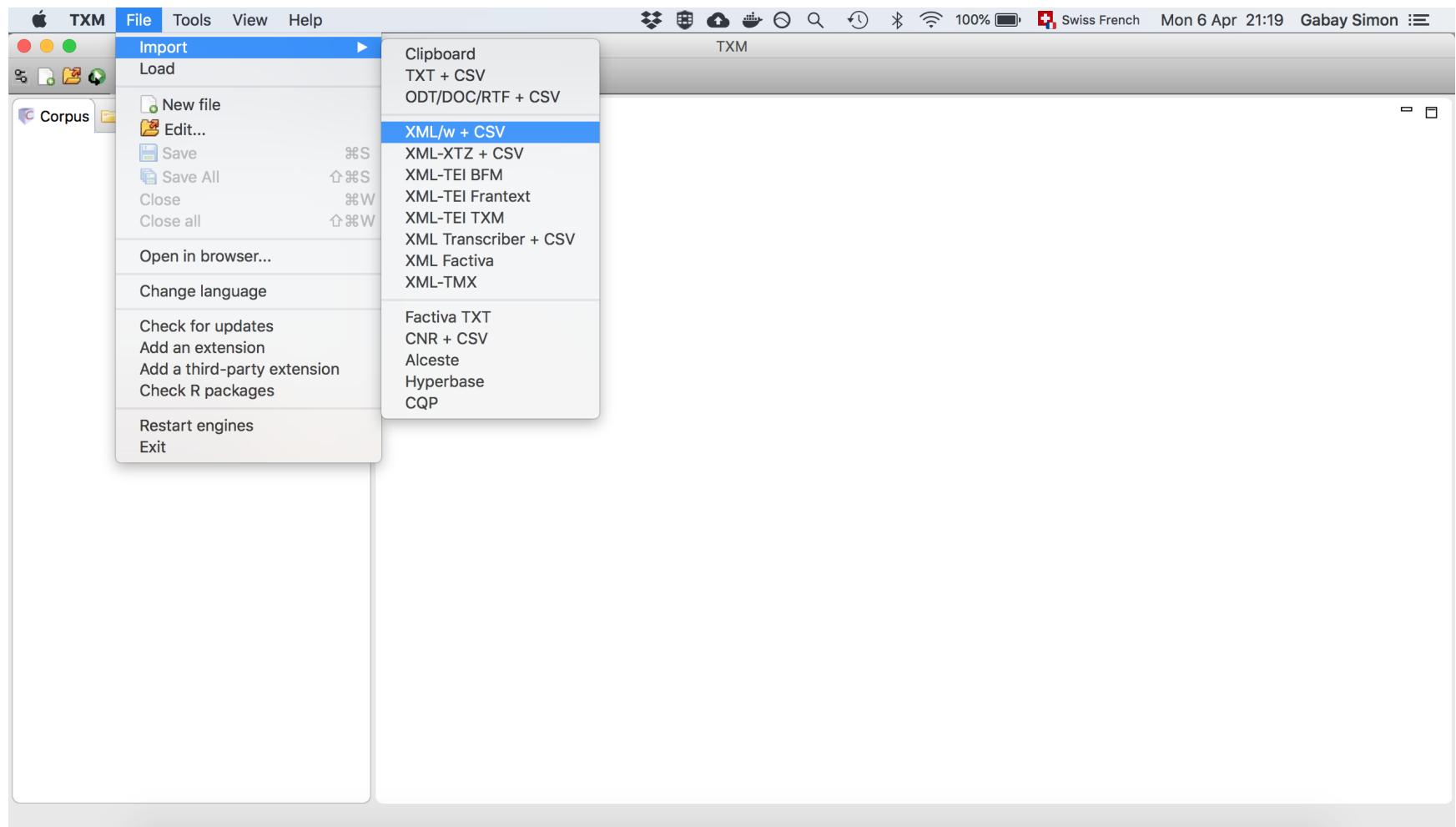


Premier corpus

Corpus en XML

```
598 ▾      <sp who="#oreste" xml:id="A3S1_16">
599        <speaker rend="center">ORESTE.</speaker>
600 ▾      <lg>
601        <l n="799" xml:id="v799">I'abuse, cher Amy, de ton trop d'amitié.</l>
602        <l n="800" xml:id="v800">Mais pardonne à des maux, dont toy feul as pitié.</l>
603        <l n="801" xml:id="v801">Excufe vn Malheureux, qui perd tout ce qu'il aime,</l>
604        <l n="802" xml:id="v802">Que tout le monde hait, &amp; qui fe hait luy-mefme.</l>
605        <l n="803" xml:id="v803">Que ne puis-je à mon tour, dans vn fort plus heu-<lb/>reux...</l>
606      </lg>
607    </sp>
608
609 ▾      <sp who="#pylade" xml:id="A3S1_17">
610        <speaker rend="center">PYLADE.</speaker>
611 ▾      <lg>
612        <l n="804" xml:id="v804">Diffimulez, Seigneur, c'eft tout ce que ie veux.</l>
613        <l n="805" xml:id="v805">Gardez qu'autant le coup vostre deffein n'éclate.</l>
614        <l n="806" xml:id="v806">Oubliez iufque-là qu'<persName ref="#hermione">Hermionne</persName>
615        <l n="807" xml:id="v807">Oubliez vostre amour. Elle vient, ie la voy.</l>
616      </lg>
617    </sp>
```

Import XML

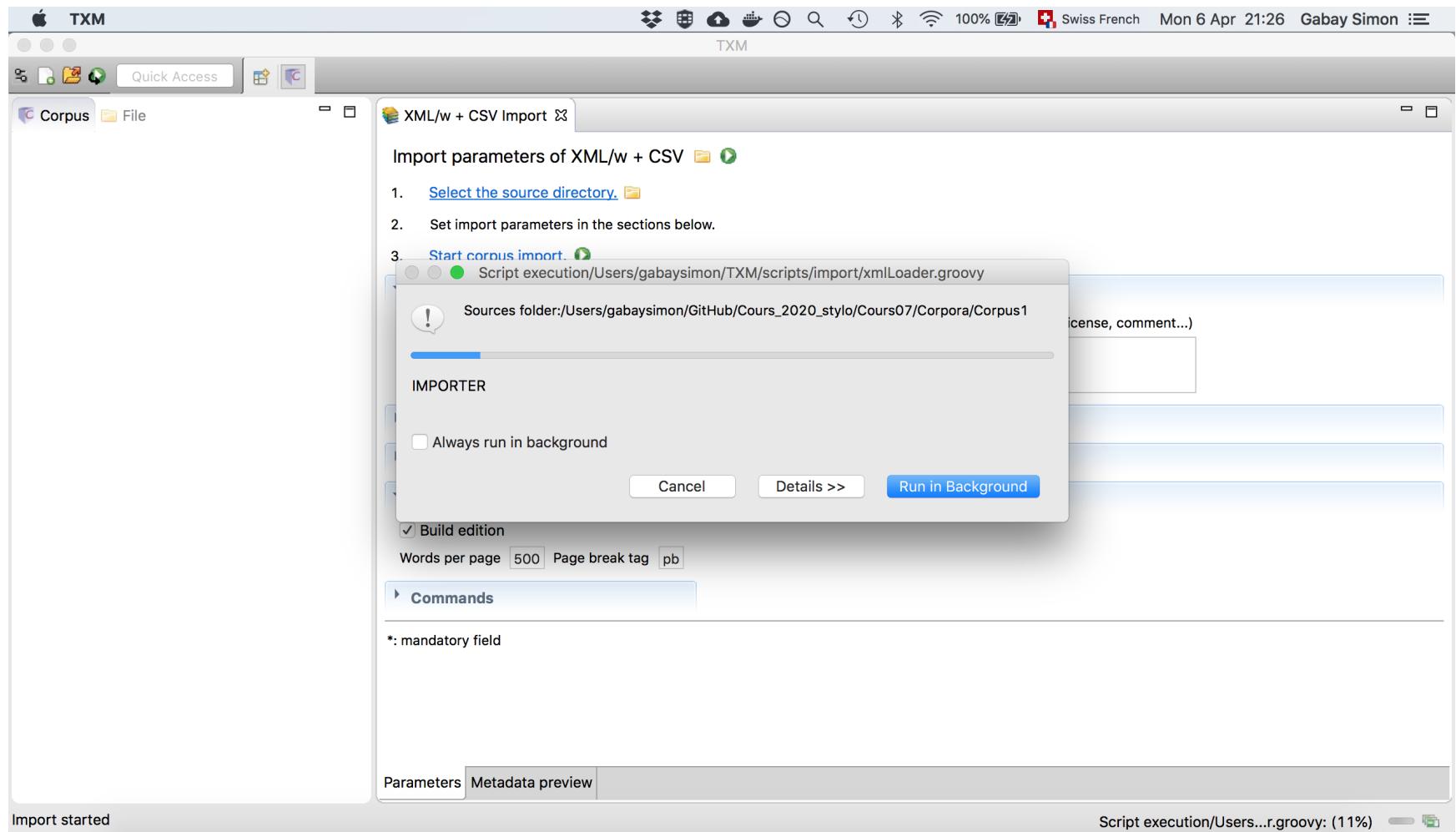


Paramètres

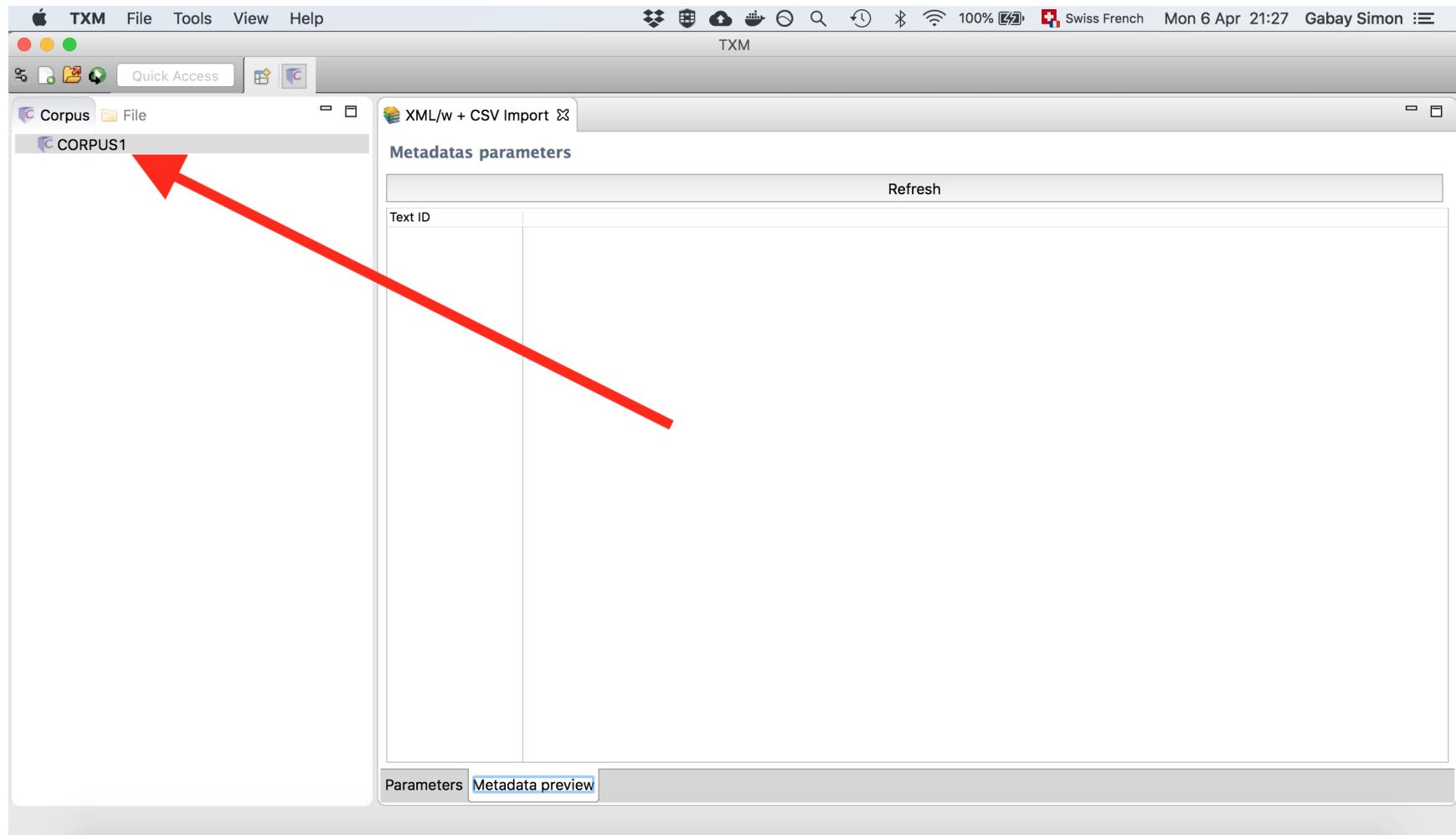
Effectuer les trois étapes:

1. Sélectionner le dossier CORPUS1
2. Paramétriser l'import (pour l'instant nous ne faisons rien)
3. Importer

Chargement



Chargé!

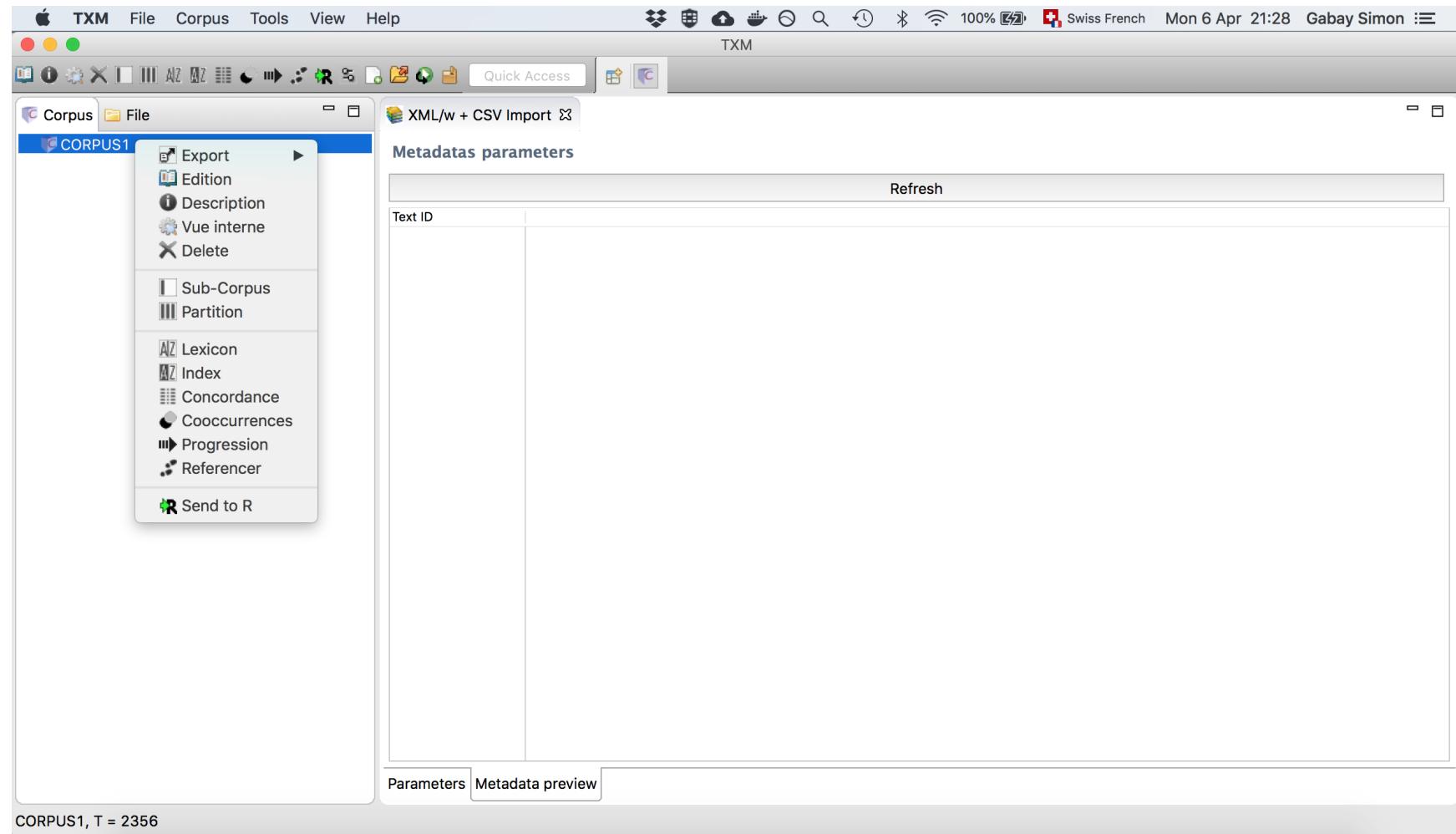


Options

Observez (clic droit sur le nom du corpus dans la colonne de gauche)

- Dimensions
- Edition
- Lexicon
- Index

Options



Un mot

- Dans l'index, cherchez *seigneur*
- Affichez la concordance (clic droit)
- Affichez l'occurrence dans l'édition (clic droit sur l'occurrence)

Revenez à l'index

- Affichez la cooccurrence (clic droit)

Revenez à l'index

- Affichez la progression (clic droit)

Vers la concordance

The screenshot shows the TXM (Text Mining) software interface. The main window displays a frequency table for the word "Seigneur". The table has two columns: "word" and "Frequency". The first row shows "Seigneur" with a frequency of 1. A context menu is open over this row, listing three options: "Send to concordance", "Send to cooccurrence", and "Send to progression".

TXM

File Tools View Help

Quick Access

Corpus File

CORPUS1

"Seigneur":word
word

XML/w + CSV Import andromaque - 20 CORPUS1: []:word CORPUS1 CORPUS1: "Seigneur":word

Query: Seigneur Properties: word Edit Search

Thresholds: Fmin: 1 Fmax: 9999999 Vmax: 9999999 Page size: 100

1 -1 / 1

Send to concordance
Send to cooccurrence
Send to progression

word	Frequency
Seigneur	1

Vers l'édition

The screenshot shows the TXM (Text Mining) application interface on a Mac OS X system. The window title is "TXM". The menu bar includes "File", "Tools", "View", "Help", and "Swiss French". The toolbar contains various icons for file operations and search. The main pane displays a search results table for the query "[word="Seigneur"]".

Search Query: [word="Seigneur"]

Sort Keys: #1 None, #2 None, #3 None, #4 None, Sort

Table Headers: text_id, Left context, Keyword, Right context

Table Data:

text_id	Left context	Keyword	Right context
andromaque	, PYLADE PYLADE. M oderez donc,	Seigneur	, cette fureur extrême. le ne vous connoy plus. Vou
andromaque	faites? TRAGEDIE. 39 Faites taire,	Seign...	, ce transport inquiet. Commandez à vos yeux de g
andromaque	du Barbare ... PYLADE. Vous l'accusez,	Seign...	, de ce deftin bizarre. Cependant tourmenté de fes
andromaque	trompeurs attraitz, Au lieu de l'enleuer,	Seign...	, ie la fuirais. Quoy? Vostre amour fe veut charger
andromaque	m'abandonne. Va-t'en. PYLADE. Allons,	Seign...	, enleuons Hermionne. Au trauers des perils vn gra
andromaque	plus heu- reux ... PYLADE. Diffimulez,	Seign...	, c'eft tout ce que ie veux. Gardez qu'auant le coup

A context menu is open over the rightmost column of the table, listing options such as "References display options", "Display options", "Sort options", "Contexts display options", "Pagination options", "Delete line", and "Display in full text".

Partitionner le corpus

Nous allons créer des partitions dans notre corpus.

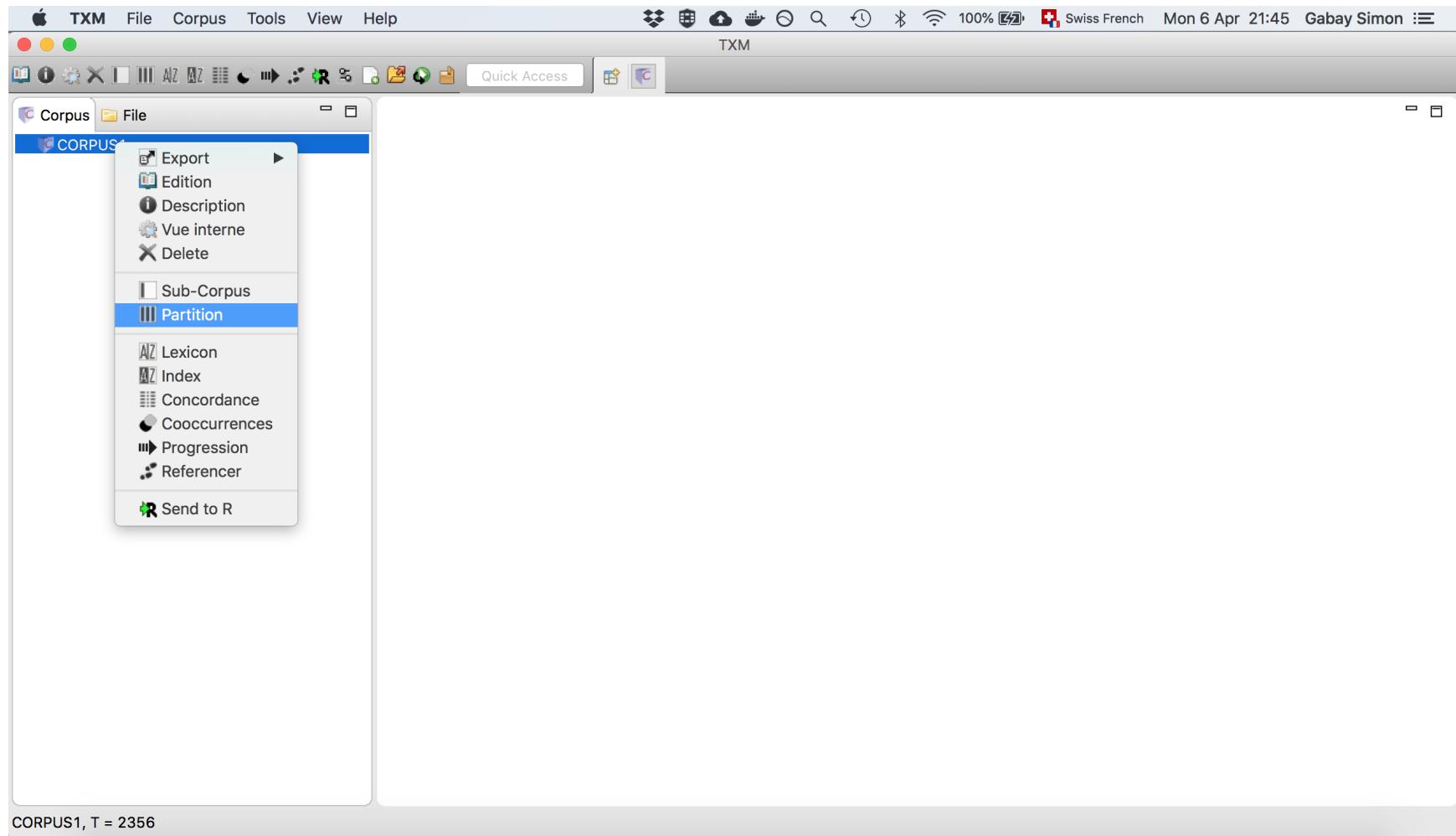
Rappelons que notre fichier original est en XML-TEI, soit un langage (XML) et un vocabulaire (TEI) qui servent de standard dans les humanités numériques. Si l'on est formé en XML-TEI, on sait que, dans les pièces de théâtre, les prises de paroles sont balisées avec l'élément `<sp>` et le nom de la personne qui prend la parole est indiqué sous une forme codifiée avec l'attribut `@who`.

```
<sp who="#codeNom">
    <speaker>NOM</speaker>
    <l n="versNumero1">un vers</l>
    <l n="versNumero2">un autre vers</l>
    <l n="versNumero3">encore un autre vers</l>
</sp>
```

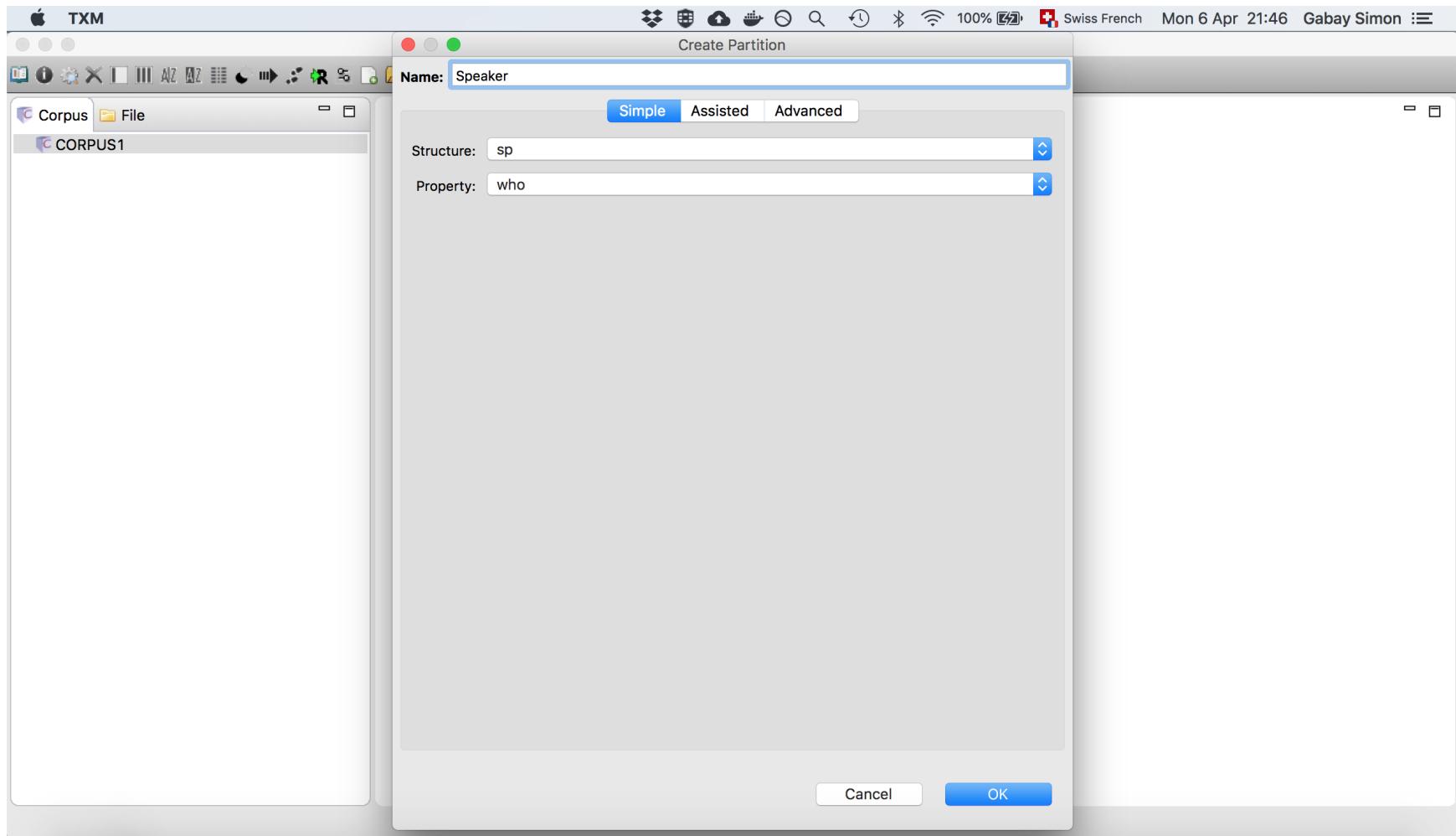
Corpus en XML

```
598 ▾ <sp who="#oreste" xml:id="A3S1_16">
599   <speaker rend="center">ORESTE.</speaker>
600 ▾   <lg>
601     <l n="799" xml:id="v799">I'abuse, cher Amy, de ton trop d'amitié.</l>
602     <l n="800" xml:id="v800">Mais pardonne à des maux, dont toy feul as pitié.</l>
603     <l n="801" xml:id="v801">Excufe vn Malheureux, qui perd tout ce qu'il aime,</l>
604     <l n="802" xml:id="v802">Que tout le monde hait, &amp; qui fe hait luy-mefme.</l>
605     <l n="803" xml:id="v803">Que ne puis-je à mon tour, dans vn fort plus heu-<lb/>reux...</l>
606   </lg>
607 </sp>
608
609 ▾ <sp who="#pylade" xml:id="A3S1_17">
610   <speaker rend="center">PYLADE.</speaker>
611 ▾   <lg>
612     <l n="804" xml:id="v804">Diffimulez, Seigneur, c'eft tout ce que ie veux.</l>
613     <l n="805" xml:id="v805">Gardez qu'autant le coup vostre deffein n'éclate.</l>
614     <l n="806" xml:id="v806">Oubliez iufque-là qu'<persName ref="#hermione">Hermionne</persName>
615     <l n="807" xml:id="v807">Oubliez vostre amour. Elle vient, ie la voy.</l>
616   </lg>
617 </sp>
```

Créer une partition



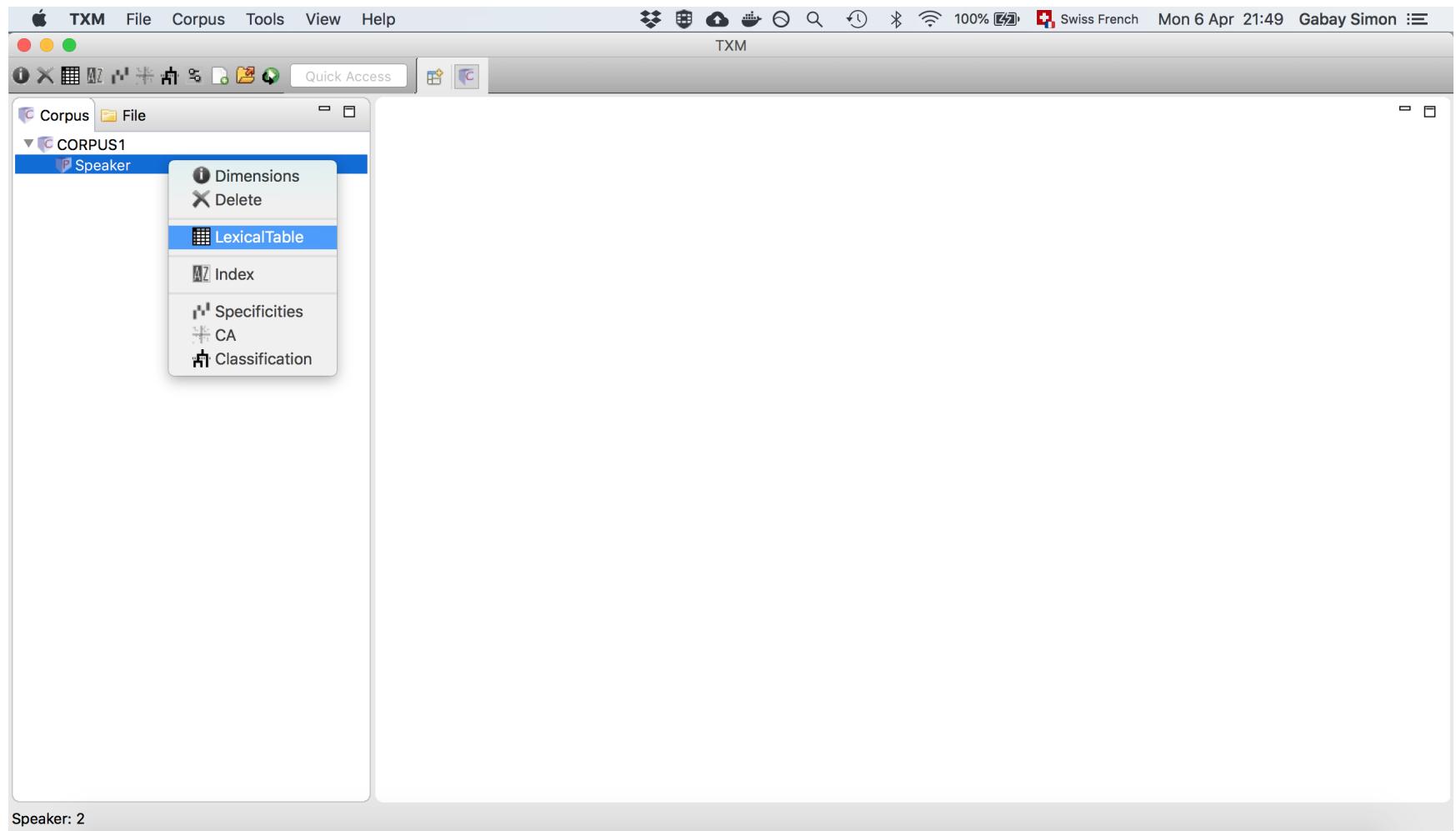
Créer une partition



Explorer la partition

- Créer une table lexicale
- Explorer la partition
- Supprimer la ponctuation
- Calculer les spécificités d'Oreste et de Pylade

Créer une table

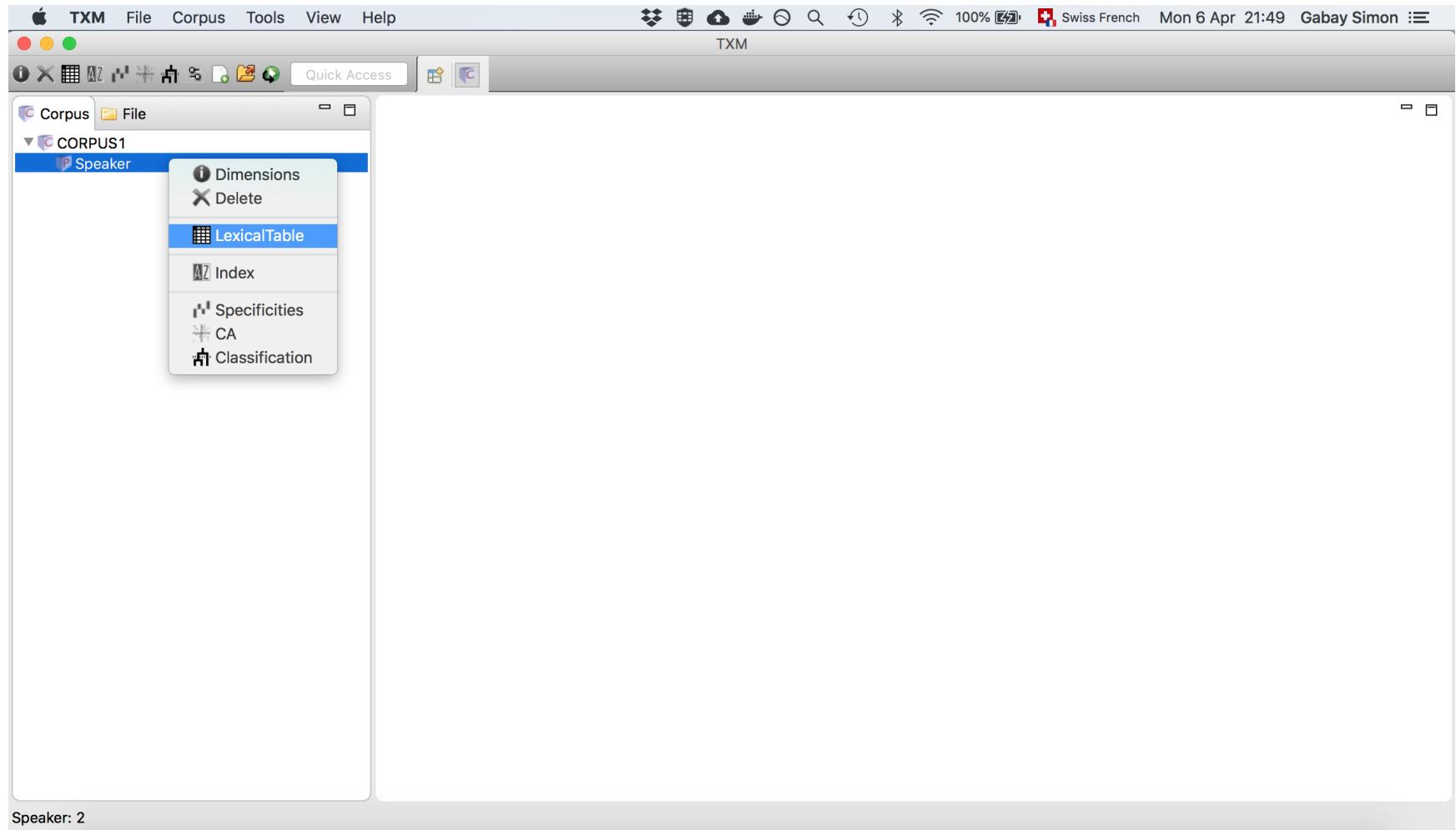


Supprimer des rangs

The screenshot shows the TXM software interface on a Mac OS X desktop. The main window displays a frequency table titled "Speaker: word" for the file "t599, v200, Fmin0, Fmax84". The table includes columns for word, Frequency, #oreste t=364, and #pylade t=235. The "Merge or Delete rows" tab is selected. To the right, a "Merge/Delete" dialog box is open, listing words for merging. The "merge" radio button is selected, and the "merge result name" field is empty. The "OK" button is highlighted.

word	Frequency	#oreste t=364	#pylade t=235
de	20	10	10
Racine	0	0	0
par	2	1	1
en	5	3	2
Andromaque	0	0	0
,	75	47	28
.	84	47	37
ANDROMAQVE	2	1	1
TRAGEDIE	2	1	1
A	1	0	1
dans	2	2	0
la	12	7	5
grand	1	0	1
Salle	0	0	0
du	2	2	0
Palais	1	0	1
Cour	1	0	1
des	3	2	1
à	11	7	4
M	1	0	1
Bandreau	0	0	0
La	1	1	0
est	0	0	0
le	21	14	7
n'eft	0	0	0
pas	1	0	1
r---	0	1	1

Supprimer des rangs



Calculer les spécificités

The screenshot shows the TXM (Text Mining) software interface on a Mac OS X system. The window title is "TXM". The menu bar includes "File", "Corpus", "Tools", "View", and "Help". The toolbar contains various icons for file operations like Open, Save, and Print.

The main workspace displays a lexical table titled "Speaker: word". The table has three columns: "wc~" (Word Category), "Frequency", and "#reste t=254 #pylade t=158". The table lists words and their frequencies, such as "O..." (9), "m..." (8), "me" (6), "moy" (6), "P..." (5), "p..." (4), "trop" (4), "mes" (3), "d..." (3), "ma" (3), "leur" (3), "q..." (3), "C'..." (3), "Non" (3), "d..." (2), "du" (2), "fuis" (2), "Le" (2), "non" (2), "las" (2), "ra..." (2), "lo..." (2), "La" (1), "e..." (1), "fa" (1), and "to..." (1). The total number of lines is 191, and the minimum frequency (Fmin) is 0.

A context menu is open over the table, with "Specificities" selected. Other options in the menu include "Export", "Delete", "CA", and "Classification".

The status bar at the bottom shows the text "org.txm.stat.engine.r.data.LexicalTableImpl@626c9fd2".

wc~	Frequency	#reste t=254 #pylade t=158
O...	9	9
m...	8	8
me	6	6
moy	6	0
P...	5	5
p...	4	4
trop	4	0
mes	3	3
d...	3	0
ma	3	0
leur	3	0
q...	3	0
C'...	3	0
Non	3	0
d...	2	0
du	2	0
fuis	2	0
Le	2	0
non	2	0
las	2	0
ra...	2	0
lo...	2	0
La	1	1
e...	1	0
fa	1	0
to...	1	0
...	1	0

Générer un histogramme

The screenshot shows the TXM (Text Mining) software interface. The menu bar includes Apple, TXM, File, Tools, View, Help, and a system status bar at the top right. The main window has a toolbar with icons for file operations like Open, Save, and Import. A "Quick Access" button is also present. On the left, a sidebar shows a tree structure of the corpus: CORPUS1 > Speaker > word. The main area displays a table titled "Speaker: word" with the following columns: Units, Frequency T 412, #oreste t=254 score, #pylade t=158 score, and a blue button labeled "Compute the histogram of the selected lines". The table lists various words and their counts, along with speaker scores.

Units	Frequency T 412	#oreste t=254 score	#pylade t=158 score
le	21	14 0.4	7 -0.4
de	20	10 -0.7	10 0.7
que	18	14 0.9	4 -0.9
ie	17	13 0.8	4 -0.8
ne	12	8 0.3	4 -0.3
la	12	7 -0.3	5 0.3
&	11	9 0.9	2 -0.9
à	11	7 0.2	4 -0.2
tout	10	5 -0.5	5 0.5
ORESTE	9	9 1.9	0 -1.9
PYLADE	9	0 -3.8	9 3.8
mon	8		
vn	8	0 -0.3	2 -0.3
vous	8	0 -3.4	8 3.4
le	7	5 0.3	2 -0.3
Et	7	5 0.3	2 -0.3
ce	7	2 -1.1	5 1.1
voftre	7	0 -3.0	7 3.0
moy	6	6 1.3	0 -1.3
me	6	6 1.3	0 -1.3
qui	6	4 0.2	2 -0.2
Seigneur	6	0 -2.5	6 2.5
Pylade	5	5 1.1	0 -1.1
les	5	3 -0.2	2 0.2
en	5	3 -0.2	2 0.2
plus	5	2 -0.5	3 0.5
trop	4	4 0.8	0 -0.8
pour	4	4 0.8	0 -0.8
Mais	4	3 0.3	1 -0.3
Il	4	3 0.3	1 -0.3
fur	4	2 -0.3	2 0.3
fe	4	2 -0.3	2 0.3

Second corpus

Description succincte

- **Groupe 1** (c. 1630-1650)
 - Pierre Du Ryer (fl. 1628-1655)
 - Georges de Scudéry (fl. 1631-1643)
 - Jean de Rotrou (fl. 1635-1649)
 - Paul Scarron (fl. 1648-1660)
- **Groupe 2**
 - Pierre Corneille (fl. 1629-1675)
- **Groupe 3** (c. 1650-1690)
 - Claude Boyer (fl. 1646-1697)
 - Thomas Corneille (fl. 1651-1696)
 - Molière (fl. 1655-1673)
 - Jean Racine (fl. 1664-1691)

On trouve des comédies, des tragédies, des tragi-comédies

Metadonnées

Les informations sur les textes, les *métadonnées*, sont importantes: il convient de les garder pour l'étude du corpus. Cette fois, nous allons donc les importer avec les textes. Pour cela il faut:

- Ajouter un fichier csv
- La première colonne, intitulée `id`, contient le nom du fichier (sans extension)
- Le contenu et le titre des autres colonnes sont libres: on met ce qu'on veut, comme le genre, la date...

Metadonnées

LibreOffice File Edit View Insert Format Styles Sheet Data Tools Window Help Swiss French Mon 6 Apr 22:10 Gabay Simon

metadata.csv

Liberation Sans 10 B I U T A1

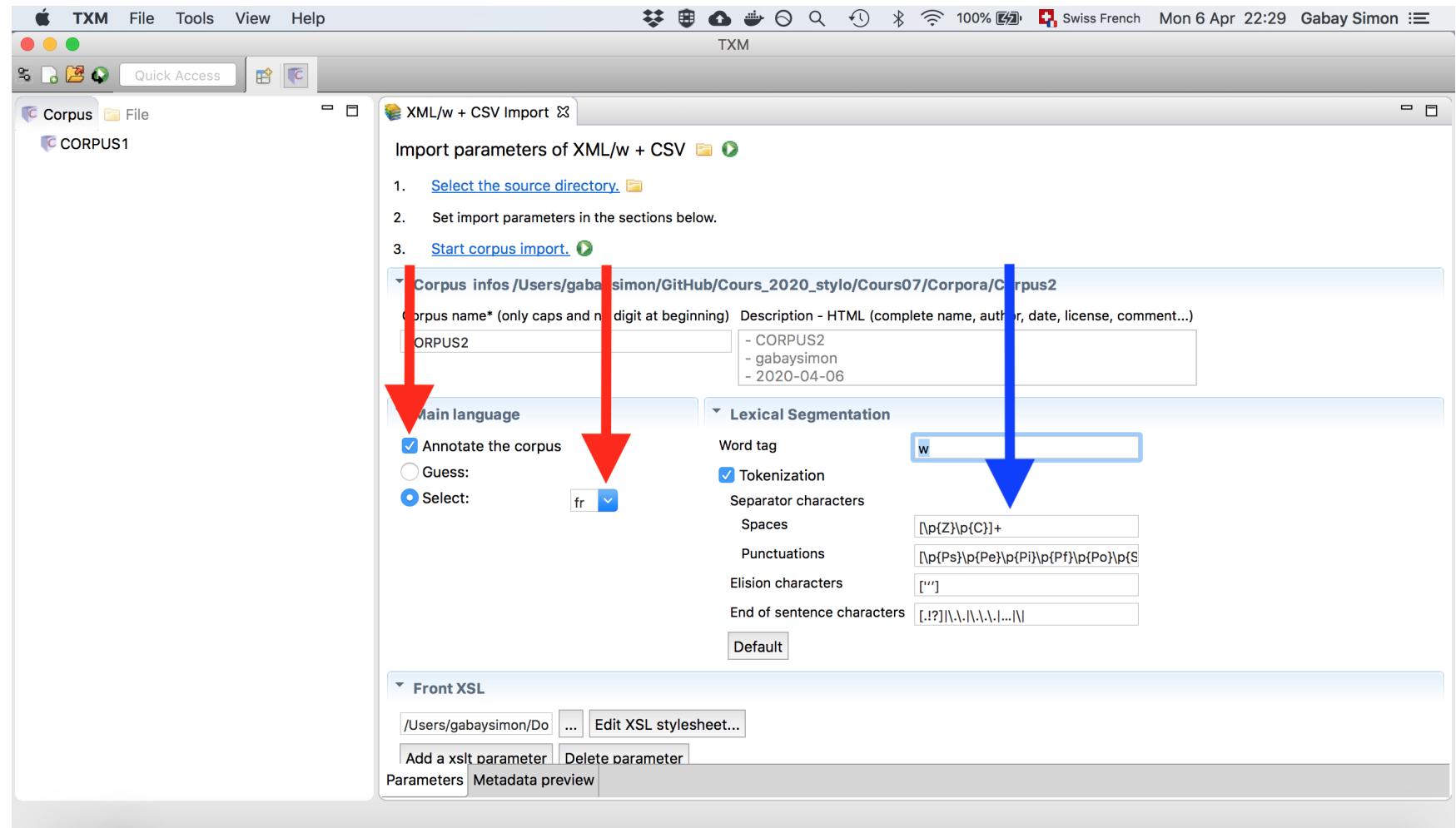
A	B	C	D	E	F	G	H	I
1	<u>auteur</u>	<u>titre</u>	<u>date</u>	<u>genre</u>	<u>inspiration</u>	<u>structure</u>	<u>type</u>	<u>periode</u>
2	BOYER_AMOURSJUPITERSEMELE	BOYER, Claude LES AMOURS DE JUPITER ET DE SÉMÉLÉ, TRAGÉDIE	1666	Tragédie	mythe grec	Cinq actes, un prologue	vers	1661-1670 2000
3	BOYER_AGAMEMNON	BOYER, Claude AGAMEMNON, TRAGÉDIE	1680	Tragédie	mythe grec	Cinq actes	vers	1671-1680 1500
4	BOYER_ARTAXERCE	BOYER, Claude ARTAXERCE, TRAGÉDIE	1683	Tragédie	histoire perse	Cinq actes	vers	1681-1690 1500
5	BOYER_ARISTODEME	BOYER, Claude ARISTODEME	1648	Tragédie	histoire grecque	Cinq actes	vers	1641-1650 1250
6	CORNEILLEP_MEDEE	CORNEILLE, Pierre MÉDEE, TRAGÉDIE	1682	Tragédie	mythe grec	Cinq actes	vers	1681-1690 1500
7	CORNEILLEP_ILLUSIONCOMIQUE	CORNEILLE, Pierre L'ILLUSION COMIQUE, COMÉDIE	1639	Comédie	moeurs françaises	Cinq actes	vers	1631-1640 1500
8	CORNEILLEP_ANDROMEDA	CORNEILLE, Pierre ANDROMÈDE, TRAGÉDIE	1651	Tragédie	mythe grec	Cinq actes, un prologue	vers	1641-1650 1500
9	CORNEILLEP_CID	CORNEILLE, Pierre LE CID, TRAGÉDIE	1682	Tragédie	histoire espagnole	Cinq actes	vers	1681-1690 1750
10	CORNEILLET_BRADAMANTE	CORNEILLE, Thomas BRADAMANTE, TRAGÉDIE	1695	Tragédie	histoire française	Cinq actes	vers	1691-1700 1750
11	CORNEILLET_ARIANE	CORNEILLE, Thomas ARIANE, TRAGÉDIE	1672	Tragédie	mythe grec	Cinq actes	vers	1671-1680 1750
12	CORNEILLET_GEOPLIERDESISMEME	CORNEILLE, Thomas LE GÉOLIER DE SOI-MÊME, COMÉDIE	1655	Comédie	moeurs italiennes	Cinq actes	vers	1651-1660 1750
13	CORNEILLET_AMOURALAMODE	CORNEILLE, Thomas L'AMOUR À LA MODE, COMÉDIE	1651	Comédie	moeurs espagnoles	Cinq actes	vers	1651-1660 1750
14	DURYER_DYNAMIS	DU RYER, Pierre DYNAMIS, REINE DE CARIE, TRAGI-COMÉDIE	1653	Tragi-comédie	moeurs françaises	Cinq actes	vers	1651-1660 1750
15	DURYER_ESTHER	DU RYER, Pierre ESTHER, TRAGÉDIE	1644	Tragédie	bible	Cinq actes	vers	1641-1650 1750
16	DURYER_CLITOPHON	DU RYER, Pierre CLITOPHON, TRAGI-COMÉDIE	1632	Tragi-comédie	mythe grec	Cinq actes	vers	1631-1640 1750
17	DURYER_CLARIGENE	DU RYER, Pierre CLARIGÈNE, TRAGICOMÉDIE	1639	Tragi-comédie	bible	Cinq actes	vers	1631-1640 1750
18	MOLIERE_DOMGARGIEDENAVARRE	MOLIÈRE DON GARCIE DE NAVARRE, COMÉDIE	1682	Comédie	moeurs espagnoles	Cinq actes	vers	1661-1670 1750
19	MOLIERE_MISANTHROPE	MOLIÈRE LE MISANTHROPE ou L'ATRABILAIRE AMOUREUX, COMÉDIE	1667	Comédie	moeurs françaises	Cinq actes	vers	1661-1670 1750
20	MOLIERE_TARTUFFE	MOLIÈRE LE TARTUFFE ou L'IMPOSTEUR, COMÉDIE	1669	Comédie	moeurs françaises	Cinq actes	vers	1661-1670 1750
21	MOLIERE_AMPHITRYON	MOLIÈRE AMPHITRYON, COMÉDIE	1668	Comédie	mythe grec	Trois actes, un prologue	vers	1661-1670 1750
22	RACINE_BERENICE	RACINE, Jean BÉRÉNICE, TRAGÉDIE	1671	Tragédie	histoire romaine	Cinq actes	vers	1671-1680 1500
23	RACINE_IPHIGENIE	RACINE, Jean IPHIGÉNIE, TRAGÉDIE	1675	Tragédie	mythe grec	Cinq actes	vers	1671-1680 1750
24	RACINE_ESTHER	RACINE, Jean ESTHER, TRAGÉDIE tirée de l'écriture sainte.	1689	Tragédie	bible	Trois actes, un prologue	vers	1681-1690 1250
25	RACINE_PHEDRE	RACINE, Jean PHÉDRE, TRAGÉDIE	1697	Tragédie	mythe grec	Cinq actes	vers	1691-1700 1500
26	ROTROU_HERCULEMOURANT	ROTROU, Jean HERCULE MOURANT, TRAGÉDIE	1636	Tragédie	mythe grec	Cinq actes	vers	1621-1630 1250
27	ROTROU_DOMBERNARDDECABRERE	ROTROU, Jean DON BERNARD DE CABRERE, TRAGI-COMÉDIE	1647	Tragi-comédie	histoire espagnole	Cinq actes	vers	1641-1650 1750
28	ROTROU_COSROES	ROTROU, Jean COSROËS, TRAGÉDIE	1649	Tragédie	histoire orientale	Cinq actes	vers	1641-1650 1500
29	ROTROU_DEUXPUCELLES	ROTROU, Jean LES DEUX PUCELLES, TRAGI-COMÉDIE	1639	Tragi-comédie	histoire romaine	Cinq actes	vers	1631-1640 1750
30	SCARRON_FAUSSEAPPARENCE	SCARRON, Paul LA FAUSSE APPARENCE COMÉDIE	1663	Tragi-comédie	moeurs espagnoles	Cinq actes	vers	1661-1670 1500
31	SCARRON_GARDIENDESOIMEME	SCARRON, Paul LE GARDIEN DE SOI-MÊME	1654	Comédie	moeurs italiennes	Cinq actes	vers	1651-1660 1750
32	SCARRON_DOMJAPHETDARMENIE	SCARRON, Paul DON JAPHET D'ARMÉNIE, COMÉDIE	1653	Comédie	moeurs françaises	Cinq actes	vers	1651-1660 1500
33	SCARRON_ECOLIERDESLAMANQUE	SCARRON, Paul L'ESCOLIER DE SALAMANQUE OU LES GÉNÉREUX ENNEMIS TRAGI-COMÉDIE	1655	Tragi-comédie	moeurs espagnoles	Cinq actes	vers	1651-1660 1500
34	SCUDERY_PRINCEDEGUISE	SCUDERY, Georges de LE PRINCE DÉGUISÉ, TRAGI-COMÉDIE	1636	Tragi-comédie	histoire médiévale	Cinq actes	vers	1631-1640 1500
35	SCUDERY_MORTDECESAR	SCUDERY, Georges de LA MORT DE CÉSAR, TRAGÉDIE	1637	Tragédie	histoire romaine	Cinq actes	vers	1631-1640 1250
36	SCUDERY_VASSALGENEREUX	SCUDERY, Georges de LE VASSAL GÉNÉREUX - TRAGI-COMÉDIE	1636	Tragi-comédie	moeurs françaises	Cinq actes	vers	1631-1640 1250
37	SCUDERY_ORANTE	SCUDERY, Georges de ORANTE, TRAGI-COMÉDIE	1636	Tragi-comédie	histoire médiévale	Cinq actes	vers	1631-1640 1250

Import XML

Recommençons l'import XML

- Activez la lemmatisation
- Regardez un peu les autres options au passage, comme la tokenisation
- (En cas de souci: Load>CORPORA/TXM)
- Allez directement à l'index

Lemmatisation



Exploration

Nous allons utiliser CQP (*Corpus Query Processor*): c'est un moteur de recherche puissant dans un corpus textuel

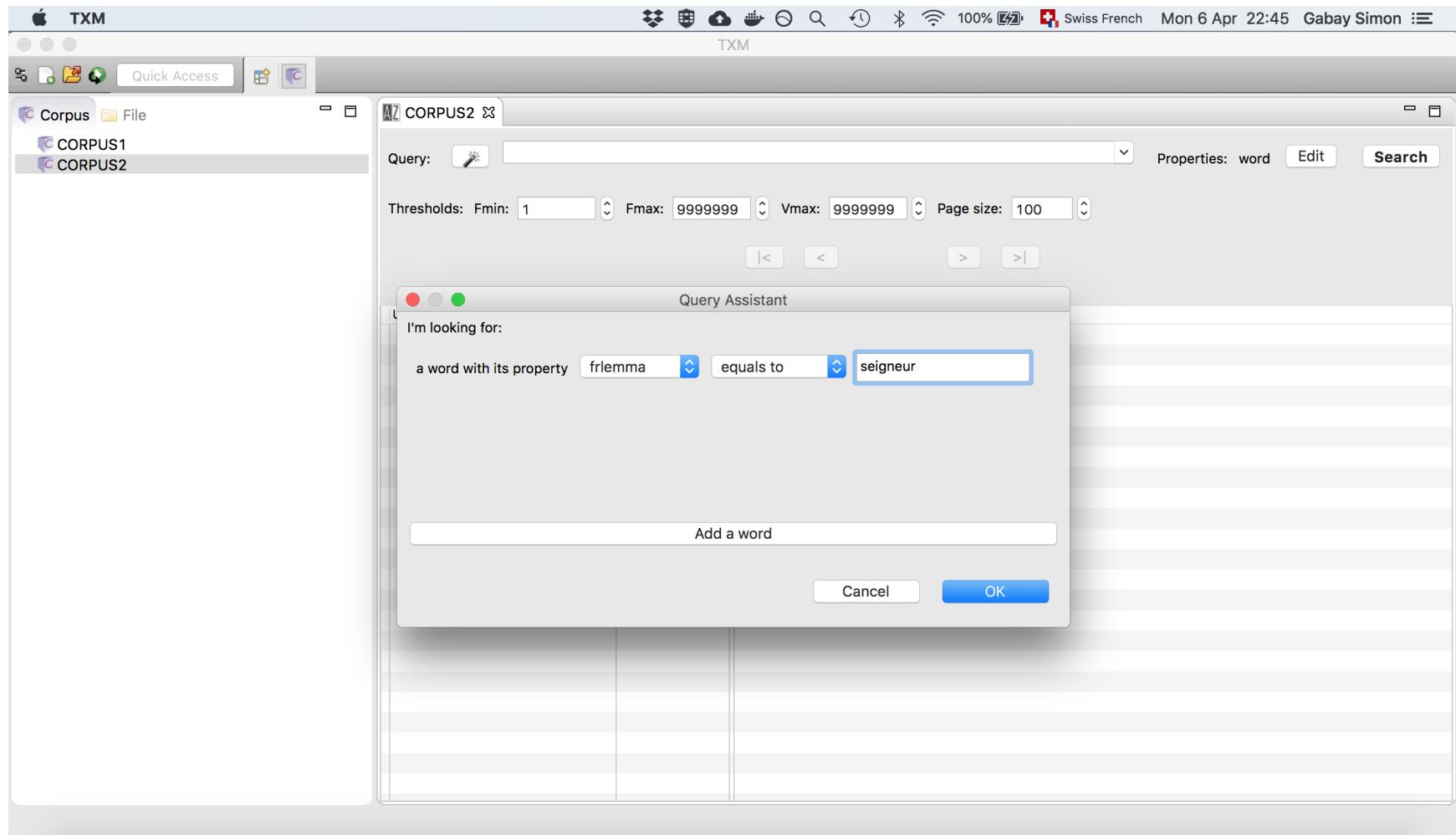
Pour utiliser CQP, il faut un langage: CQL (*Corpus Query Language*)

Une expression (ou équation) CQL est une chaîne de caractères exprimant un motif linguistique (un mot, ou une suite de mots) à partir des valeurs de leurs propriétés (comme la catégorie grammaticale, le lemme, la forme graphique).

Manuel de TXM [en ligne](#)

On avait déjà fait une première requête dans l'Index (*Seigneur*): allons plus loin, et cherchez le lemme *Seigneur* (indice: suivez la baguette magique).

Requête (lemme)



Exploration 2

Allons encore plus loin, et construisons une requête complexe...

- Regardez le jeu d'étiquette et fourni en annexe et cherchez des catégories grammaticales précises
1. Arriverez-vous à chercher tous les pronoms?
 2. Tous les pronoms suivis du verbe être?
 3. Tous les pronoms suivis du verbe être (peu importe la forme)?
 4. Tous les pronoms suivis d'un mot quelconque, suivi du verbe être (peu importe la forme)?

Requête complexe

The screenshot shows the TXM (Text Mining) application interface. The menu bar includes Apple, TXM, File, Tools, View, Help, and a Quick Access toolbar. The main window displays a search results table for the query [frpos="PRO*"][]{1}[frlemma="être"]:

Query: [frpos="PRO*"][]{1}[frlemma="être"]

Properties: word

Thresholds: Fmin: 1, Fmax: 9999999, Vmax: 9999999, Page size: 100

Results: t2543 , v881 , fmin1 , fmax248

word	Frequency
ce n' est	248
il n' est	210
qu' il est	75
s' il est	73
c' en est	64
qu' il soit	39
il s' est	33
qui n' est	31
il m' est	26
qui m' est	26
qu' elle est	25
il en est	24
tu n' es	23
elle n' est	21
qui s' est	20
vous n' êtes	20
dont il est	18
il vous est	17
s' il était	15
il en soit	14
qui vous est	14
il doit être	13
ce n' était	12

Done: 881 items for 2543 occurrences

Solutions

1. [frpos="PR0.*"]
2. [frpos="PR0.*"] [word="être"]
3. [frpos="PR0.*"] [frlemma="être"]
4. [frpos="PR0.*"] [] [frlemma="être"]

CQL advanced

- Ignorer:
 - %c casse, ex. [word="état">%c]
 - %d diacritiques, ex. [word="état">%d]
 - %d les deux, ex., [word="état">%cd]
- Opérateurs
 - = égal
 - != différent
 - | ou
 - & et
 - () priorité des opérations
- Quantificateurs
 - {1} une fois, {1,2} une ou deux fois
 - ? zéro ou une fois
 - + une seule fois ou plus

- Pour des raisons informatiques, il faut échapper certains caractères (ceux utilisés par CQL) en plaçant une barre oblique inverse avant:

- `? , * , + , | , &`
- `? , * , + , | , &`

- On peut enfin utiliser des expressions régulières:

- `.` n'importe quel caractère
- `[uv]` u ou v
- `[uvij]` u ou v ou i ou j
- `[a-z]` n'importe quelle minuscule non accentuée
- `[A-Z]` n'importe quelle majuscule non accentuée
- `\d` un chiffre
- `\s` un caractère d'espacement
- `\w` un caractère de mot
- Et beaucoup d'autres...

- Il est possible d'utiliser des propriétés de structures, c'est-à-dire d'utiliser les balises XML dans les requêtes:
 - <l> [* [frlemma="manger"] [* </l>

Exercices

Cherchez:

1. *monseigneur* ou *monsieur* (essayez d'autres termes équivalents)
2. *dame* ou *seigneur* précédé d'un pronom possessif
3. *dame* ou *seigneur* précédé d'un pronom possessif ou d'un article
4. Les pronoms possessifs ou articles, un mot quelconque, puis *dame* ou *seigneur*
5. Les pronoms possessifs ou articles, un ou deux mot quelconques, puis *dame* ou *seigneur*
6. Les mots finissant par *er*
7. Les interjections finissant par *h*
8. Les mots avec *ette* à la rime
9. Les vers composés de cinq ou six tokens

1. [word="monseigneur" | word="monsieur" | word="seigneur"]
2. [frpos="DET.POS"] [word="seigneur" | word="dame"]
3. [frpos="DET.*"] [word="seigneur" | word="dame"]
4. [frpos="DET.POS"] [] [word="seigneur" | word="dame"]
5. [frpos="DET.POS"] [] {1,2} [word="seigneur" | word="dame"]
6. [word=".*er"]
7. [word=" [aeiou]h"]
8. <l> [] * [word=".*ette"] </l>
9. <l> [] {5,6} </l>

Exploration

Vous savez désormais faire des partitions. À l'aide des métadonnées, vous pouvez créer des partitions riches pour l'exploration et l'analyse.

Essayez d'en faire:

- Par genre
- Par période
- Par auteur

Metadonnées

LibreOffice - metadata.csv

The screenshot shows a LibreOffice Calc spreadsheet titled "metadata.csv". The spreadsheet has 10 columns labeled A through I. Column A contains the primary key "id". Columns B and C contain the main title and subtitle information. Columns D and E contain the date and genre. Columns F and G contain the source and structure. Columns H and I contain the type and period. A green arrow points to column B, a red arrow points to column E, and a blue arrow points to column I.

A	B	C	D	E	F	G	H	I
id	auteur	titre	date	genre	inspiration	structure	type	periode
1	BOYER_AMOURSJUPITERSEMELE	LES AMOURS DE JUPITER ET DE SÉMÉLÉ, TRAGÉDIE	166	Tragédie	mythe grec	Cinq actes, un prologue	vers	1661-1670
2	BOYER_AMAGEMNON	AGAMEMNON, TRAGÉDIE	168	Tragédie	mythe grec	Cinq actes	vers	1671-1680
3	BOYER_ARTAXERCE	ARTAXERCE, TRAGÉDIE	168	Tragédie	histoire perse	Cinq actes	vers	1681-1690
4	BOYER_ARISTODEME	ARISTODEME	164	Tragédie	histoire grecque	Cinq actes	vers	1641-1650
5	CORNEILLE_MEDEE	MÉDEE, TRAGÉDIE	168	Tragédie	mythe grec	Cinq actes	vers	1681-1690
6	CORNEILLELLP_ILLUSIONCOMIQUE	ILLUSION COMIQUE, COMÉDIE	163	Comédie	moeurs françaises	Cinq actes	vers	1631-1640
7	CORNEILLELP_ANDROMEDA	ANDROMÈDE, TRAGÉDIE	165	Tragédie	mythe grec	Cinq actes, un prologue	vers	1641-1650
8	CORNEILLELP_CID	CID, TRAGÉDIE	168	Tragédie	histoire espagnole	Cinq actes	vers	1681-1690
9	CORNEILLELT_BRADAMANTE	BRADAMANTE, TRAGÉDIE	169	Tragédie	histoire française	Cinq actes	vers	1691-1700
10	CORNEILLELT_ARIANE	ARIANE, TRAGÉDIE	167	Tragédie	mythe grec	Cinq actes	vers	1671-1680
11	CORNEILLELT_GEO	GEÖLIER DES OISMEME	166	Comédie	moeurs italiennes	Cinq actes	vers	1651-1660
12	CORNEILLELT_AMOURALAMODE	AMOUR À LA MODE, COMÉDIE	165	Comédie	moeurs espagnoles	Cinq actes	vers	1651-1660
13	DURYER_DYNAMIS	DYNAMIS, REINE DE CARIE, TRAGI-COMÉDIE	165	Tragi-comédie	moeurs françaises	Cinq actes	vers	1651-1660
14	DURYER_ESTHER	ESTHER, TRAGÉDIE	164	Tragédie	bible	Cinq actes	vers	1641-1650
15	DURYER_CLITOPHON	CLITOPHON, TRAGI-COMÉDIE	163	Tragi-comédie	mythe grec	Cinq actes	vers	1631-1640
16	DURYER_CLARIGENE	CLARIGÈNE, TRAGICOMÉDIE	163	Tragi-comédie	bible	Cinq actes	vers	1631-1640
17	MOLIERE_DOMGARGOUCHE	DOM GARCIE DE NAVARRE, COMÉDIE	168	Comédie	moeurs espagnoles	Cinq actes	vers	1661-1670
18	MOLIERE_MISANTHROPE	LE MISANTHROPE ou L'ATRABILAIRE AMOUREUX, COMÉDIE	166	Comédie	moeurs françaises	Cinq actes	vers	1661-1670
19	MOLIERE_TARTUFFE	LE TARTUFFE ou L'IMPOSTEUR, COMÉDIE	166	Comédie	mythe grec	Trois actes, un prologue	vers	1661-1670
20	MOLIERE_AMPHITRYON	AMPHITRYON, COMÉDIE	166	Comédie	histoire romaine	Cinq actes	vers	1671-1680
21	RACINE_BERENICE	BERÉNICE, TRAGÉDIE	167	Tragédie	mythe grec	Cinq actes	vers	1671-1680
22	RACINE_IPHIGENIE	IPHIGÉNIE, TRAGÉDIE	167	Tragédie	bible	Trois actes, un prologue	vers	1681-1690
23	RACINE_ESTHER	ESTHER, TRAGÉDIE tirée de l'écriture sainte.	168	Tragédie	mythe grec	Cinq actes	vers	1681-1690
24	RACINE_PHEdre	PHÉDRE, TRAGÉDIE	169	Tragédie	mythe grec	Cinq actes	vers	1691-1700
25	ROTROU_HERCULEMOURANT	HERCULE MOURANT, TRAGÉDIE	163	Tragédie	mythe grec	Cinq actes	vers	1621-1630
26	ROTROU_DOMBERNARDDECABRERE	DON BERNARD DE CABRERE, TRAGI-COMÉDIE	164	Tragi-comédie	histoire espagnole	Cinq actes	vers	1641-1650
27	ROTROU_COSROES	COSROËS, TRAGÉDIE	164	Tragédie	histoire orientale	Cinq actes	vers	1641-1650
28	ROTROU_DEUXPUCELLES	LES DEUX PUCELLES, TRAGI-COMÉDIE	163	Tragi-comédie	histoire romaine	Cinq actes	vers	1631-1640
29	SCARRON_FAUSSEAPPARENCE	LA FAUSSE APPARENCE COMÉDIE	166	Tragi-comédie	moeurs espagnoles	Cinq actes	vers	1661-1670
30	SCARRON_GARDIENDESOIMEME	LE GARDIEN DE SOI-MÊME	165	Comédie	moeurs italiennes	Cinq actes	vers	1651-1660
31	SCARRON_DOMJAPHETDARMENIE	DON JAPHET D'ARMÉNIE, COMÉDIE	165	Comédie	moeurs françaises	Cinq actes	vers	1651-1660
32	SCARRON_ECOLIERDESLAMANQUE	LE SCOLIER DE SALAMANQUE OU LES GÉNÉREUX ENNEMIS TRAGI-COMÉDIE	165	Tragi-comédie	moeurs espagnoles	Cinq actes	vers	1651-1660
33	SCUDERY_PRINCEDEGUISE	LE PRINCE DÉGUISÉ, TRAGI-COMÉDIE	163	Tragi-comédie	histoire médiévale	Cinq actes	vers	1631-1640
34	SCUDERY_MORTDECESAR	A MORT DE CÉSAR, TRAGÉDIE	163	Tragédie	histoire romaine	Cinq actes	vers	1631-1640
35	SCUDERY_VASSALGENEREUX	LE VASSAL GÉNÉREUX - TRAGI-COMÉDIE	163	Tragi-comédie	moeurs françaises	Cinq actes	vers	1631-1640
36	SCUDERY_ORANTE	ORANTE, TRAGI-COMÉDIE	163	Tragi-comédie	histoire médiévale	Cinq actes	vers	1631-1640

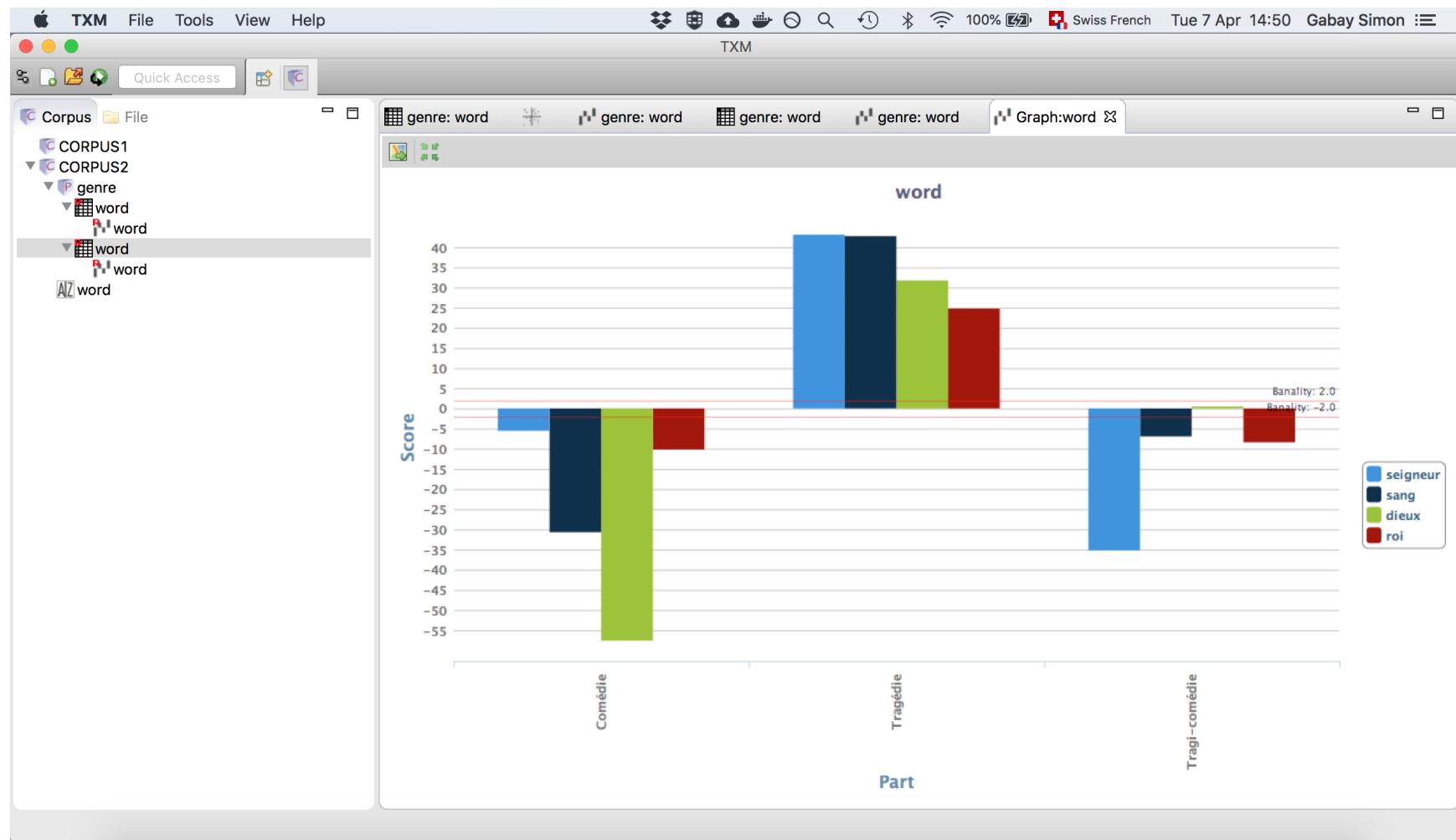
Pour chaque partition, partez à la découverte. Un processus d'exploration simple est:

1. création d'une table lexicale
2. Nettoyage de ladite table
3. Calculer les spécificités
4. Calculer la classification
5. Calculer l'AFC

Gestion de la table lexicale

La table lexicale permet notamment de supprimer des tokens. En fonction de ce que nous allons supprimer, nous allons créer une table plus ou moins grande:

- Si nous nous apprêtons à faire une analyse de type stylométrique, et donc à ne conserver que les mots vides, il est inutile (la plupart du temps) de faire une table contenant plus de 200 mots
- Si nous nous apprêtons à faire une analyse de type textométrique, et donc à ne conserver que les mots thématiques , il va falloir faire une table un peu plus grande, probablement d'environ 500 mots.



Sources

Ce cours reprend un cours donné avec Jean-Baptiste Camps lors de la formation Fophil de Neuchâtel.