



# TXM - commandes avancées et import

Karën Fort

karen.fort@sorbonne-universite.fr / <https://members.loria.fr/KFort/>



# Quelques sources d'inspiration

par ordre d'importance décroissant

- ▶ Atelier TXM du 25 et 26 septembre 2014
- ▶ Manuel de TXM : <http://txm.sourceforge.net/doc/manual/manual1.xhtml>
- ▶ B. Pincemin (ICAR) et S. Heiden (ICAR)
- ▶ Vidéos de la formation de B. Pincemin (chapitrées) :  
[https://txm.sourceforge.net/enregistrement\\_atelier\\_initiation\\_TXM\\_fr.html](https://txm.sourceforge.net/enregistrement_atelier_initiation_TXM_fr.html)

Retours

Questions

Révisions

Comparer au sein d'un corpus

Import dans TXM

Pour finir

Bibliographie



# Expressions rationnelles (régulières)

Quelles expressions permettent de trouver :

- ▶ *patrie* et *patriote* ?
- ▶ *France*, *français* et *françaises* ?

# Expressions rationnelles (régulières)

Quelles expressions permettent de trouver :

- ▶ *patrie* et *patriote* ?

Exemple : `[word="patri.*"]`

- ▶ *France*, *français* et *françaises* ?

# Expressions rationnelles (régulières)

Quelles expressions permettent de trouver :

- ▶ *patrie* et *patriote* ?

Exemple : `[word="patri.*"]`

- ▶ *France*, *français* et *françaises* ?

Exemple : `[word="(F|f)ran.*"]`

# Expressions rationnelles (régulières) : surgénération

Est-ce que nos résultats sont satisfaisants ?

[word="patri.\*"] : trouve *patrimoine*

[word="(F|f)ran.\*"] : trouve *franchement*



# Petit rappel de CQL

## Corpus Query Language

- ▶ expressions régulières : *Europe*|*européen*.\*, [] (un mot), & et | (booléens)
- ▶ neutralisations (à ajouter **après** l'expression) :
  - ▶ %c pour neutraliser la casse ("europe"%c)
  - ▶ %d pour neutraliser les diacritiques (accents, cédille)
  - ▶ etc. (voir doc)
- ▶ assistant de requête

# CQL : utiliser la neutralisations

VOEUX/<[word="francais.\*|france"%cd]>@w... ☒

Requête 🔍 [word="francais.\*|france"%cd]

word	Fréquence
France	392
Français	147
Françaises	43
française	38
français	22
françaises	3
Française	2

Retours

## Comparer au sein d'un corpus

- Création de sous-corpus et de partition

- Spécificités

- Graphiques

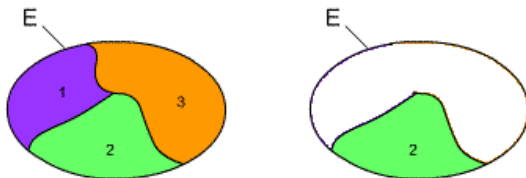
Import dans TXM

Pour finir

Bibliographie

# Partition vs sous-corpus

## Partition vs sous-ensemble



Créer une partition du corpus Vœux  
selon le locuteur

# Table lexicale

## Recherche et fréquence

Rechercher « ambition » chez tous les Présidents



Que peut-on en déduire ?

# Table lexicale

## Recherche et fréquence

Rechercher « ambition » chez tous les Présidents



Que peut-on en déduire ?

Les parties étant **inéga**les, on ne peut pas comparer les scores d'une partie par rapport à une autre

## Calcul de spécificités [Lafon, 1980]

Analogie de la boîte à œufs :

on a autant de boîtes à œufs que de présidents et on renverse les œufs n'importe comment

→ rare que 18 œufs tombent dans une **même** boîte

→ 1 chance sur 1, suivi de 5 zéros (exposant de la probabilité)

Calcul qui ne nécessite que 4 variables :

- ▶ T (total général)
- ▶ t (total dans la partie)
- ▶ F (fréquence générale)
- ▶ f (fréquence dans la partie)

# Calcul de spécificités

Expliqué par [B. Pincemin](#) (ICAR)

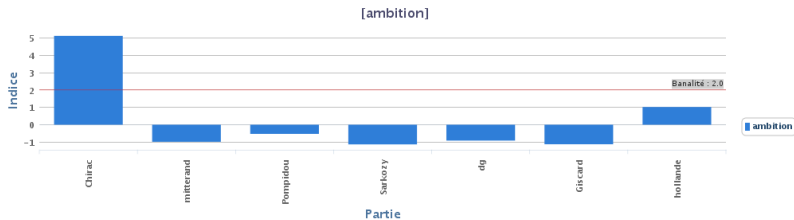


# Générer une vue graphique



Comment faire ?

# Qui a de l'*ambition* ?



Exporter la vue : 4e bouton depuis la gauche

# Interprétation

- ▶ a du sens au-dessus de 3
- ▶ **zone de banalité** représentée sur le graphique pour éviter de surinterpréter
- ▶ une significativité négative peut avoir du sens : **nullax** ( $< -3$ )

Retours

Comparer au sein d'un corpus

Import dans TXM

Retour sur XML

Import(s)

Pour finir

Bibliographie

# Importer dans TXM

Nous avons vu lors du cours précédent comment importer du texte *via* le presse-papier (CTRL+C)

Il existe (évidemment) d'autres possibilités...

# Les imports dans TXM

Par Serge Heiden (concepteur)

## Carte des niveaux d'import TXM

	<b>TXT</b>	<b>XML/w</b>	<b>XML-TEI</b>
<i>Unités Textuelles</i>	fichiers	fichiers	fichiers
<i>Métadonnées</i>	CSV	CSV	teiHeader
<i>Mots</i>	brut	<w>?	<w>?

# *eXtensible Markup Language* (XML)

en 20 sec. (voir cours de P. Laublet)

- ▶ langage informatique de balisage (comme HTML ou SGML)
- ▶ ... textuel, structuré, et extensible car
- ▶ son « langage » (vocabulaire et grammaire) peut être redéfini (par exemple, *mabalise* peut être un nom de balise)
- ▶ syntaxe stricte, peut être validée par des outils automatiques

## XML : extrait de TCOF-POS

```
<?xml version="1.0" encoding="UTF-8"?>
<document>
<loc nb="L2">
<w lemme="L2" pos="LOC">L2</w>
<w lemme="ben" pos="INT">ben</w>
<w lemme="on" pos="PRO:cls">on</w>
<w lemme="attendre" pos="VER:pres">attend</w>
<w lemme="on" pos="PRO:cls">on</w>
<w lemme="attendre" pos="VER:pres">attend</w>
<w lemme="qui" pos="PRO:int">qui</w>
<w lemme="normalement" pos="ADV">normalement</w>
</loc>
```



# La *Text Encoding Initiative* (TEI)

en 20 sec. (voir <http://www.tei-c.org>)

Consortium à but non lucratif :

- ▶ auto-financé
- ▶ constitué d'institutions, de projets de recherche et de chercheurs (64 membres)
- ▶ qui existe depuis 1987
- ▶ qui développe et maintient un [standard pour la représentation des textes numériques](#) : un format SGML au début, XML maintenant
- ▶ outre la documentation du format, la TEI fournit des outils et des formations

# La TEI : exemple

Wikipédia, TEI, Le Cid

## Acte II, Scène 2

**DON RODRIGUE** À moi, Comte, deux mots.

**LE COMTE** Parle.

**DON RODRIGUE** Ôte-moi d'un doute.

Connais-tu bien Don Diègue ?

**LE COMTE** Oui.

**DON RODRIGUE** Parlons bas, écoute.

Sais-tu que ce vieillard fut la même vertu,

La vaillance et l'honneur de son temps ? Le sais-tu ?

# La TEI : exemple

## Wikipédia, TEI

```
<div type="Act" n="I"><head>Acte II</head>
  <div type="Scene" n="1"><head>Scène 2</head>
    <sp><speaker>Rodrigue</speaker>
      <l part="i">À moi, comte, deux mots.</l></sp>
    <sp><speaker>Comte</speaker>
      <l part="m">Parle</l></sp>
    <sp><speaker>Rodrique</speaker>
      <l part="f">Ôte-moi d'un doute</l></sp>
    <sp><speaker>Comte</speaker>
      <l part="i">Connais-tu bien Don Diègue ?</l></sp>
    <sp><speaker>Comte</speaker>
      <l part="m">Oui</l></sp>
    <sp><speaker>Rodrigue</speaker>
      <l part="f">Parlons bas, écoute.</l>
      <l>Sais-tu que ce vieillard fut la même vertu,</l>
      <l>La vaillance et l'honneur de son temps ? Le sais-tu ?</l></sp>
    ...
  </div>
  ...
</div>
```

## Exemple de fichier source

[Amblard et al., 2015]

Ps- Bien donc merci beaucoup monsieur d'être / d'être venu.

Sh4- ... de rien.

Ps- Alors dites-moi pourquoi est-ce que vous êtes ici ?

Quelles sont les informations dont je vais avoir besoin pour mon  
analyse ?

# Faire des choix : exemple de fichier d'import

annoté par MElt [Denis and Sagot, 2010] (pas TreeTagger)

```
<?xml version="1.0" encoding="UTF-8"?>
<document>
<u who="Ps">
<w cat="ADV" lemme="bien" >Bien</w>
<w cat="ADV" lemme="donc" >donc</w>
<w cat="FNO" lemme="merci" >merci</w>
<w cat="ADV" lemme="beaucoup" >beaucoup</w>
<w cat="NOM" lemme="monsieur" >monsieur</w>
<w cat="PRP" lemme="un" >d'</w>
<w cat="VER:infi" lemme="être" >être</w>
<w cat="MLT" lemme="*/" >/</w>
<w cat="PRP" lemme="un" >d'</w>
<w cat="VER:infi" lemme="être" >être</w>
<w cat="VER:pper" lemme="venir" >venu</w>
<w cat="ADV" lemme="*." >.</w>
</u>
```

# Est-ce suffisant ?

De quelles autres informations vais-je avoir besoin ?

**0001**

id: 0001

loc: dg

annee: 1959

Pour la métropole française, pour l'Algérie, pour la communauté, je forme des vœux ardents et confiants au premier jour de 1960. Je suis rempli de l'espoir que cette année nous sera propice, parce que nous avons fait beaucoup au cours de celle qui finit.

En France même, nos institutions assurent à l'Etat l'efficacité et l'autorité qui lui permettent d'agir. Il en sera ainsi, désormais. Nos finances sont en équilibre et le resteront demain. Au point de vue des ressources, de la technique, de la recherche, du crédit, des échanges et de la monnaie, notre industrie, notre agriculture, notre commerce, disposent d'une bonne base pour l'expansion et le progrès qui vont marquer 1960. Le franc nouveau est le signe de cette féconde solidité. Dans les domaines politique, social, scolaire, etc., le pouvoir fera tout pour qu'aux querelles d'autrefois succèdent la concorde et la coopération entre les familles spirituelles, les catégories, les citoyens de la nation française.

# Métadonnées : des données sur les données

Il n'y a pas de magie : les métadonnées doivent être décrites quelque part

- ▶ fichier [metadata.csv](#) (dans le répertoire du corpus)
- ▶ à l'import, TXM associe chaque texte du corpus à ses métadonnées

## Exemple de fichier metadata.csv (pour Voeux)

```
"id","loc","annee"  
"t0001","dg","1959"  
"t0002","dg","1960"  
"t0003","dg","1961"  
"t0004","dg","1962"  
"t0005","dg","1963"  
"t0006","dg","1964"  
"t0007","dg","1965"  
"t0008","dg","1966"  
"t0009","dg","1967"  
"t0010","dg","1968"  
"t0011","pompidou","1969"  
"t0012","pompidou","1970"  
"t0013","pompidou","1971"  
"t0014","pompidou","1972"  
"t0015","pompidou","1973"
```



Retours

Comparer au sein d'un corpus

Import dans TXM

**Pour finir**

CQFR : Ce Qu'il Faut Retenir

Bibliographie



Fonctionnalités avancées :

- ▶ partition de corpus
- ▶ table lexicale
- ▶ calcul de spécificité (Interprétation)

Import :

- ▶ types d'imports
- ▶ XML vs TEI
- ▶ balise  $\langle w \rangle$
- ▶ comment ajouter des informations dans le corpus



Amblard, M., Fort, K., Demily, C., Franck, N., and Musiol, M. (2015).

Analyse lexicale outillée de la parole transcrite de patients schizophrènes.

Revue TAL, 55(3) :91 – 115.



Denis, P. and Sagot, B. (2010).

Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français.

In

Traitement Automatique des Langues Naturelles : TALN 2010, Montréal, Canada.



Lafon, P. (1980).

Sur la variabilité de la fréquence des formes dans un corpus.

Mots : Saussure, Zipf, Lagado, des méthodes, des calculs, des doutes et le vocabulaire de quelques textes politiques,  
(1) :127–165.