

# Outils informatiques d'analyse linguistique - TXM

## Commandes avancées et import

Karën Fort

karen.fort@paris-sorbonne.fr / <http://karenfort.org>

1er décembre 2016





# 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)

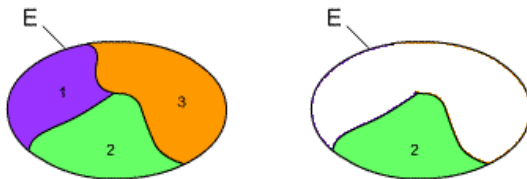


- 1 Sources
- 2 Comparer au sein d'un corpus
  - Création de sous-corpus et de partition
  - Spécificités
  - Graphiques
- 3 Import dans TXM
- 4 Pour finir
- 5 Bibliographie



# 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 ?



# 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** ( $< -3$ )



- 1 Sources
- 2 Comparer au sein d'un corpus
- 3 Import dans TXM**
  - Retour au contexte
  - Import(s)
- 4 Pour finir
- 5 Bibliographie



# Importer dans TXM

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



# *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="PR0:cls">on</w>
<w lemme="attendre" pos="VER:pres">attend</w>
<w lemme="on" pos="PR0:cls">on</w>
<w lemme="attendre" pos="VER:pres">attend</w>
<w lemme="qui" pos="PR0: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>

```



# Importés par S. Heiden

L'import dans TXM



- 1 Sources
- 2 Comparer au sein d'un corpus
- 3 Import dans TXM
- 4 Pour finir**
  - CQFR : Ce Qu'il Faut Retenir
  - TD à rendre
- 5 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



# Travail sur « vrai » corpus

## Exercice à faire pendant le prochain cours, noté

- télécharger le corpus MPT (Mariage pour tous) :  
<https://sourceforge.net/projects/txm/files/corpora/mpt>
- l'importer dans TXM
- trouver une hypothèse (ex : les femmes subissent plus d'interruptions que les hommes) ==> vous pouvez regarder dès maintenant le corpus
- la valider ou l'invalidier par le corpus (argumenter)
- rendre un document pdf décrivant succinctement votre démarche et les résultats obtenu (ou non), avec les schémas TXM appropriés





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.