Plate-formes logicielles pour le TAL 3 : TXM - commandes avancées et import

Karën Fort

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

LETTRES SORBONNE UNIVERSITÉ

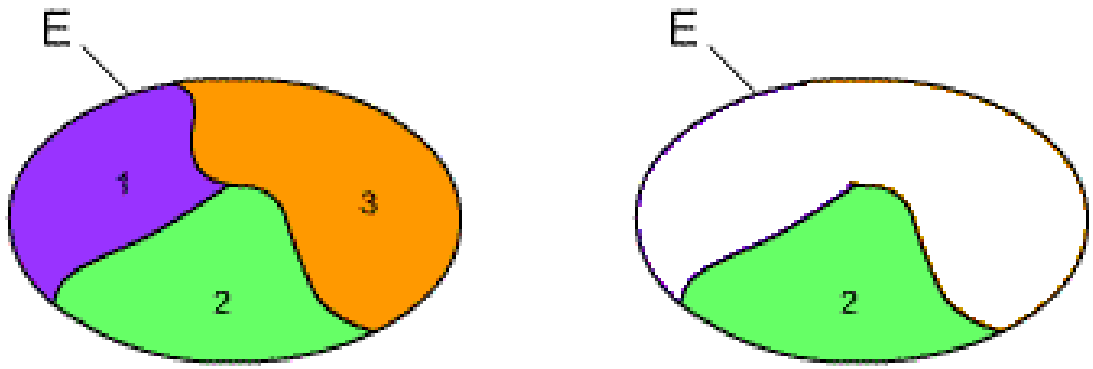
Quelques sources d'inspiration

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)

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 inégales, 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
$\rightarrow$ rare que 18 œufs tombent dans une **même** boîte
$\rightarrow$ 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)

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="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 maintien 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 Dièguе ?

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

Exemple de fichier source

```
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 ?
```

Exemple de fichier d'import (annoté par MElt)

```
<?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>
```

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"
```

Fonctionnalités avancées :

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

Import :

- types d'imports
- XML *vs* TEI
- balise $< w >$
- comment ajouter des informations dans le corpus

Travail sur « vrai » corpus

Exercice noté, à rendre par mail avant les vacances

- télécharger le corpus MPT (Mariage pour tous) : https://github.com/nlegrand/mariagepourtousInXML/blob/master/Readme.md
- l'importer dans TXM
- trouver une hypothèse (ex : les femmes subissent plus d'interruptions que les hommes)
- la valider ou l'invalider par le corpus (argumenter)

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.