

Corpus, ressources et linguistique outillée · M2SOL034

CM 3 : Textométrie avancée

Ljudmila PETKOVIĆ

Semestre 2, 2024-2025

14 février 2025

Sorbonne Université

Master « Langue et Informatique » (M1 ScLan)

UFR Sociologie et Informatique pour les Sciences Humaines

Cours adapté de FORT (*s.d.*), LEJEUNE (2023), HEIDEN (*s.d.*) et PINCEMIN (2012).

Types de corpus et d'import

Trois familles de corpus

1. corpus de textes écrits (presse-papier, TXT, XML, TEI)
 - éditions alignées avec images de facsimilés
2. corpus de transcriptions d'enregistrements (TRS)
 - éventuellement synchronisées avec le son ou la vidéo
3. corpus multilingues alignés (TMX)
 - au niveau d'une structure textuelle (phrase, paragraphe)

Formats enrichis :

- XML : avec métadonnées
- issus d'autres logiciels : Hyperbase, Alceste...
- XML-TEI : Frantext, Transcriber...

Import dans TXM

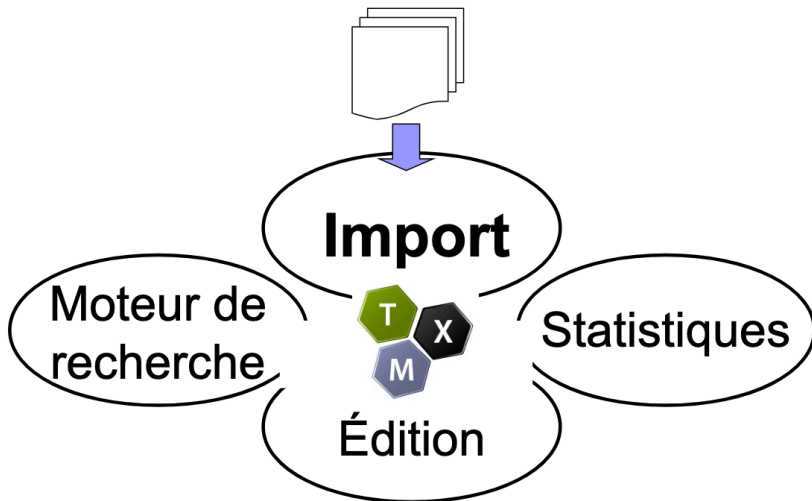


Figure 1 – Schéma basique du *workflow* dans TXM (HEIDEN, *s.d.*).

Import du corpus

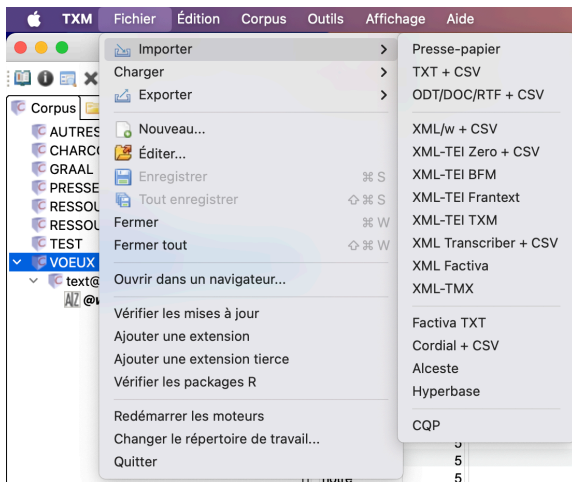


Figure 2 – Options d'import du corpus aux différents formats.

Niveaux d'import

	TXT	XML/w	XML-TEI
<i>Unités Textuelles</i>	fichiers	fichiers	fichiers
<i>Métadonnées</i>	CSV	CSV	teiHeader
<i>Mots</i>	brut	<w>?	<w>?
<i>Structures</i>	-	toutes	spécifique
<i>Plans</i>	-	XSL frontale	spécifique

Figure 3 – Carte des niveaux d'import TXM ([HEIDEN, s.d.](#)).

Paramètres d'importation du corpus

AZ VOEUX/text@annee=1959/@word

 Import XML-TEI Zero + CSV
 

Paramètres d'import du module XML-TEI Zero + CSV
 



- [Sélectionner le répertoire des fichiers sources et nommer le corpus](#) 
- Régler les paramètres d'import dans les sections ci-dessous.
- [Lancer l'import du corpus.](#) 

▼ Description

Nom complet, auteur, date, licence, commentaire...

▶ Langue principale	▶ Segmentation lexicale
▶ Éditions	▶ Police d'affichage
▶ Commandes	▶ Plans textuels
▶ Options	

* : champ obligatoire

Figure 4 – Préparation de l'importation du corpus.

Pourquoi structurer un document ?

La structuration permet de :

- expliciter pour la machine
- exploiter la dynamique interne du corpus

Contraignant mais permet d'utiliser les fonctions avancées qui tirent partie des sous-corpus (donc des meta-données)

Exemple de problème de structuration

Documents sauvegardés par Sorbonne Université

Judi 30 septembre 2021 à 14 h 37

LE FIGARO

Nom de la source
Le Figaro
Type de source
Presse • Journaux
Périodicité
Quotidien
Couverture géographique
Nationale
Provenance
France

Mardi 25 octobre 2011

Le Figaro • no. 20910 • p. 26 • 311 mots

La Digital Mum intègre Médiamétrie

La nouvelle ménagère pourra être prise en compte par les régies publicitaires.

Paule Gonzalès

INTERNET Elle existe, elle est identifiée et désormais elle sera prise en compte par les régies publicitaires. La Digital Mum, cette femme active qui surfe régulièrement sur le Web, découverte il y a un peu moins d'un an par WebMediaGroup et développée par l'agence médias KR Médias, fait son entrée chez Médiamétrie, le temple de la mesure d'audience.

« Si elle n'est pas encore une cible de

ernes aussi bien pour réserver les vacances sur le Web que pour effectuer les achats de Noël.

Active pour les fêtes de fin d'année

Une nouvelle étude constate que « 70 % des Digital Mums utiliseront Internet pour leurs achats de fin d'année et pour 19 % d'entre elles c'est même précisément durant cette période de l'année qu'elles prévoient d'acheter le plus sur la Toile ». Toutefois, si la Digital Mum

p. 26



Figure 5 – Problème de structuration (LEJEUNE, 2023).

Exemple de données structurées

```
<corpus>
<article titre="La « digital mum », nouvel eldorado des marques Un baromètre trimestriel créé par KR Media et WebMediaGroup
permettra de mieux comprendre ses comportements." date="2011 04 27" journal="Le Figaro, no. 20756">COMMUNICATION Le profil
de la « digital mum » s'affine. Avec la généralisation d'Internet et des nouvelles technologies, elle est même en passe
d'éclipser « la ménagère de moins de 50 ans », cible commerciale très convoitée née dans les années 1960. Pour la première
fois, l'impact des nouveaux médias dans la consommation et dans la vie des mères de famille est étudié. L'agence KR Media,
conseil en stratégie et achat d'espaces publicitaires, ne s'y trompe pas. Elle propose à ses clients « une vision de la
ménagère de moins de 50 ans qui soit plus en phase avec la réalité de notre société. La »digital mum* relie parfaitement les
mondes physiques et numériques dans lesquels nos annonceurs déploient leurs actions marketing. La »digital mum* est ainsi la
nouvelle cible universelle que nous devons mieux comprendre. » Aussi vient-elle de signer avec WebMediaGroup, inventeur de
la « digital mum », un partenariat pour mieux la définir et la suivre dans ses comportements médias et d'achat, via un
baromètre trimestriel. À terme, ce baromètre pourrait séduire d'autres acteurs dont Médiamétrie qui a du mal à vendre à
l'international le concept de ménagère de moins de 50 ans. La « digital mum » est « une femme ayant au moins un enfant à
charge et se connectant au moins une fois par semaine à Internet », explique Isabelle Bordry, PDG de WebMediaGroup. En
France, les « digital mums » représentent 17 % de la population des 15 ans et plus, soit 8,7 millions. C'est presque autant
que les ménagères de moins de 50 ans, qui sont 10,7 millions en France. D'ailleurs, elles se confondent un peu. Ainsi 80 %
des « digital mums » sont des ménagères de moins de 50 ans. Selon Isabelle Bordry, « la »digital mum* a en moyenne 40 ans
mais le sentiment d'en avoir 33 et déclare agir autant par intuition que par raison ». Enfin, 45 % de ces dernières ont un
revenu mensuel par foyer supérieur à 2 700 euros net contre 39 % pour la ménagère de moins de 50 ans. Cible non homogène
L'objectif de cette étude est de mieux cerner les caractéristiques de cette nouvelle génération de consommatrices. Cela «
devrait sensibiliser les marques non seulement à l'évolution de leur communication, mais aussi à celle des services et
```

Figure 6 – Problème de structuration (LEJEUNE, 2023).

XML

angl. *eXtensible Markup Language*

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

Exemple du fichier XML

Cf. le tutoriel W3C¹

```
<?xml version="1.0" encoding="UTF-8"?>
<note>
  <to>Sapna</to>
  <from>Tom</from>
  <heading>Meeting</heading>
  <body>At 11 AM on Monday morning.</body>
</note>
```

Figure 7 – Exemple d'un fichier XML.

```
▼ object {1}
  ▼ note {4}
    to : Sapna
    from : Tom
    heading : Meeting
    body : At 11 AM on Monday morning.
```

Figure 8 – Exemple de la structure arborescente du fichier XML.

1. <https://www.w3schools.blog/xml-tutorial>.

TEI

angl. *Text Encoding Initiative*² (depuis 1987)

Consortium à but non lucratif :

- auto-financé
- institutions, projets de recherche et chercheurs
- **standard pour la représentation des textes numériques**
 - un format SGML au début, XML maintenant
- documentation, outils et formations

2. <http://www.tei-c.org>

Exemple du document XML-TEI

TEI

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-stylesheet type="text/xsl" href="../../../Teinte/tei2html.xsl"?>
3 <TEI xmlns="http://www.tei-c.org/ns/1.0">
4   <teiHeader xmlns="">
5     <fileDesc>
6       <titleStmt>
7         <title>Oeuvres complètes de J. M. Charcot. Tome 1. Leçons sur les maladies du système nerveux</title>
8         <author>Bourneville, Désiré Magloire</author>
9         <author>Charcot, Jean-Martin</author>
10      </titleStmt>
11    </fileDesc>
12    <profileDesc>
13      <creation>
14        <date when="1892" />
15      </creation>
16    </profileDesc>
17  </teiHeader>
18  <text xmlns="">
19    <body>
20      <div>
21        <p>
22          <s>ŒUVRES COMPLÈTES</s>
23        </p>
24        <p>
25          <s>DE</s>
26        </p>
27        <p>
28          <s>J.M. GH ARGOT</s>
29        </p>
30        <p>
31          <s>LEÇONS</s>
32        </p>
33        <p>
34          <s>sur les</s>
35        </p>

```

Figure 9 – Corpus CHARCOT au format XML-TEI.

Exemple du document XML-TEI : mode édition

CL_000001_001_TEXTE.TXT

ŒUVRES COMPLÈTES

DE

J. M. GH ARGOT

LEÇONS

sur les

Figure 10 – Corpus CHARCOT au format XML-TEI : mode édition.

Métadonnées

« des données sur les données »

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, corpus 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"
```

Figure 11 – Identifiants, locuteurs (ex. dg : De Gaulle), année.

Autres fonctionnalités TXM

Progression

Graphique cumulative de l'évolution d'un ou de plusieurs motifs au fil d'un corpus, exprimés par des requêtes CQL.

T : total général (nombre de mots dans le corpus)

Occurrences : fréquences générales dans le corpus

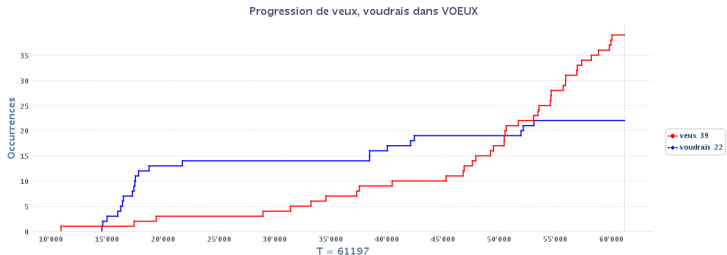


Figure 12 – Évolution des formes verbales, exemple de LEJEUNE (2023)

Progression

■ Utilisation des étiquettes POS

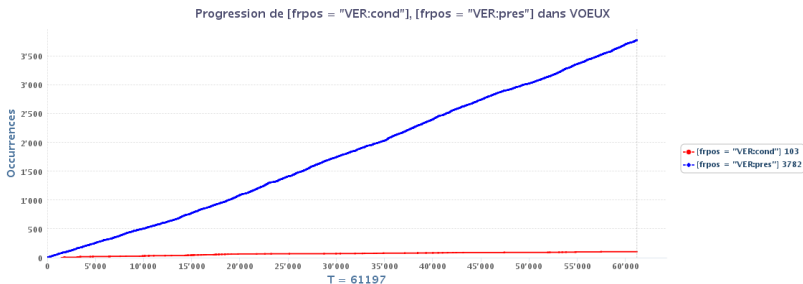


Figure 13 – Évolution des formes verbales *veux* (prés.) et *voudrais* (cond.) à l'aide des étiquettes POS, exemple de LEJEUNE (2023).

Comparer au sein d'un corpus

Création de sous-corpus et de partitions

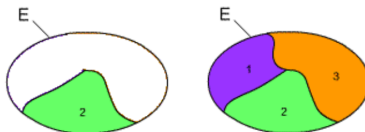


Figure 14 – Sous-corpus vs. partitions dans un ensemble E, adapté de FORT (*s.d.*).

- **sous-corpus** : regroupement « minimal » déterminé selon les métadonnées, sélection d'une partie de l'ensemble sans contrainte
- **partition** : un ensemble de sous-corpus, divise tout l'ensemble en parties disjointes et exhaustives

On peut ensuite « opposer » des partitions pour faire émerger des phénomènes par contraste.

Exemple de création de partition

Créer une partition du corpus VOEUX selon le locuteur MITTERRAND.

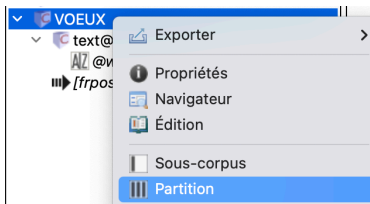


Figure 15 – Création de la partition.

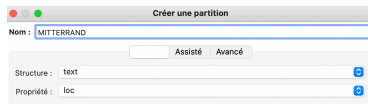


Figure 16 – Choix du locuteur
François Mitterrand

Dimensions de la partition

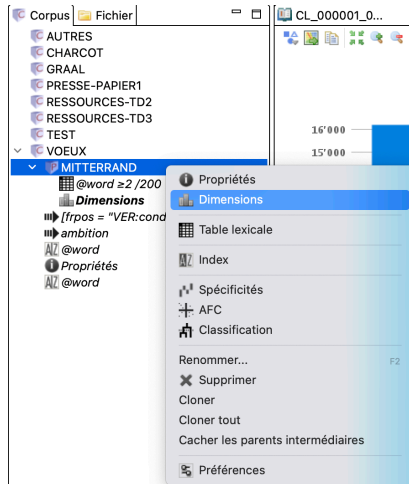


Figure 17 – Dimension d'une partition.

Dimensions de la partition

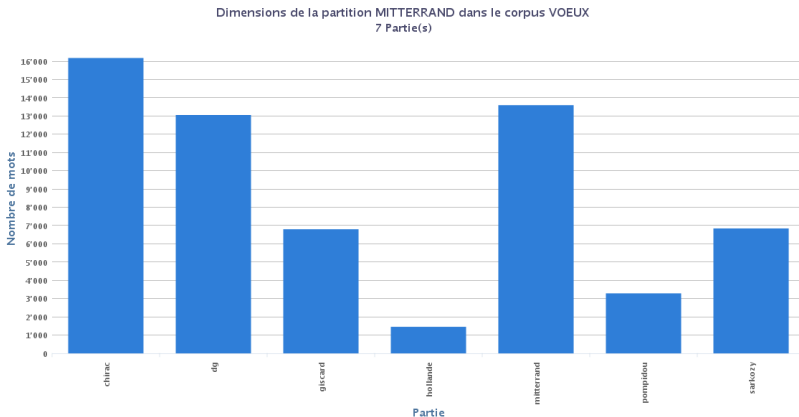


Figure 18 – Dimension de la partition MITTERRAND dans le corpus VOEUX.

Progression

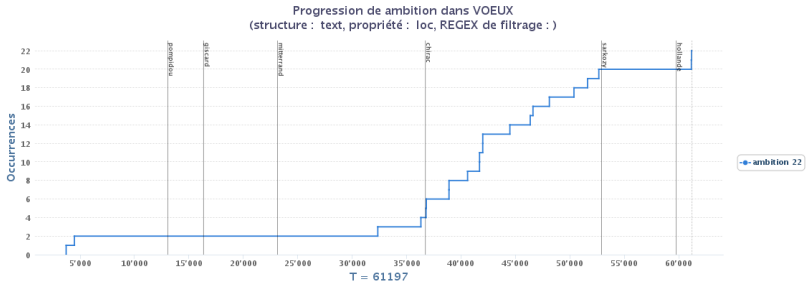


Figure 19 – Évolution du terme *ambition* dans les partitions du corpus VOEUX.

Table lexicale

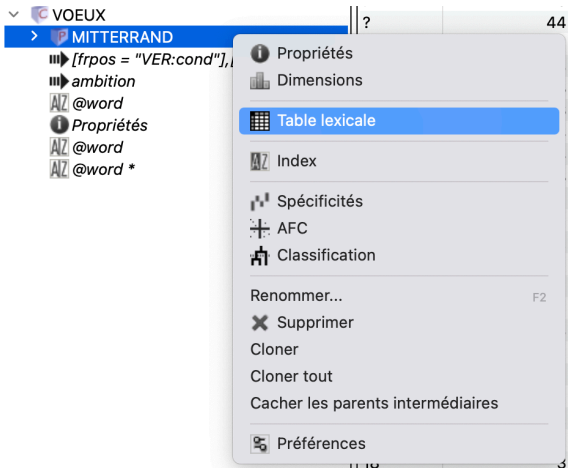


Figure 20 – Générer une table lexicale pour toutes les partitions.

Table lexicale

- les fréquences brutes sont biaisées par la taille des parties
 - les spécificités de Lafon permettent de rendre les comparaisons plus fiables en ajustant les fréquences en fonction des tailles relatives des sous-corpus
- T : total général (nombre de mots)
 - t : total dans la partie (nombre de mots)

*VOEUX/MITTE...	Dimensions ...	VOEUX/MITTE...	VOEUX/MITTE...	33	[ambition]	[ambition]	VOEUX/MITTE...	9					
Propriété	word												
Unités	^	Fréquence T 61197	chirac t=16176 indice dg t=13054 indice giscard t=6797 indice hollande t=1453 indice mitrerrand t=13592 indice pompidou t=3285 indice sarkozy t=6840 indice										
ambition	Calculer le diagramme en bâtons des lignes sélectionnées			0	-1.1	2	1.0	2	-1.0	0	-0.5	0	-1.1
ambitions				2	1.0	0	-0.1	0	-0.5	0	-0.1	0	-0.3
âme	4	2	0.5	1	0.2	0	-0.2	0	-0.0	0	-0.4	1	0.4
améliora...	3	0	-0.4	3	2.0	0	-0.2	0	-0.0	0	-0.3	0	-0.2
améliora...	1	0	-0.1	0	-0.1	0	-0.1	0	-0.0	0	-0.1	0	1.0

Figure 21 – Générer le diagramme en bâtons pour le terme *ambition* dans toutes les partitions.

Calcul de spécificités (Lafon, 1980)

- mesure le caractère attendu ou exceptionnel de la fréquence d'un mot (ou motif complexe, trait linguistique, etc.) dans une partie du corpus, au regard de sa fréquence dans l'ensemble du corpus et de la taille de la partie

Calcul qui ne nécessite que 4 variables :

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

Spécificités

Est-ce que la fréquence du mot est étonnante par rapport à la probabilité attendue ?

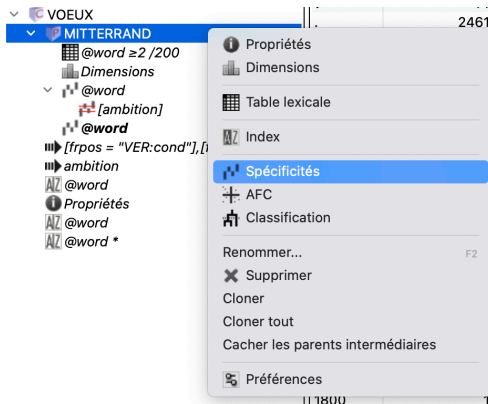


Figure 22 – Calcul des spécificités.

Spécificités

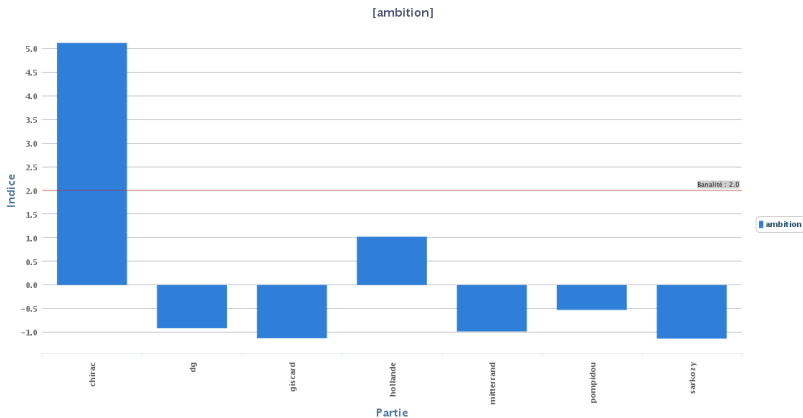


Figure 23 – Indice de spécificité pour le mot *ambition* dans les différentes partitions du corpus VOEUX.

Spécificités

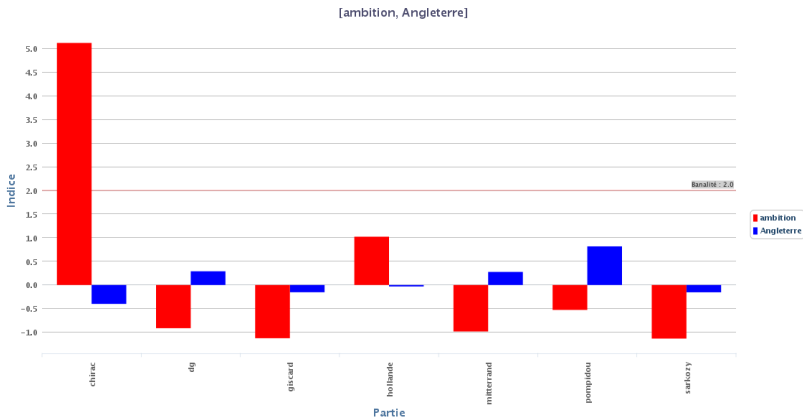


Figure 24 – Indices de spécificité pour les mots *ambition* et *Angleterre* dans les différentes partitions du corpus VOEUX.

Interprétation

- chaque partie est représentée par un ensemble de barres contiguës, classées dans le même ordre que dans le tableau
- chaque propriété de mot (forme graphique) sera représentée par une barre de la même couleur dans chaque partie
- les couleurs sont légendées dans le coin inférieur droit du graphique
- la ligne rouge délimite la **zone de banalité** autour de l'axe d'indice 2
 - les barres qui n'en sortent pas sont à considérer comme banales
 - banalité (forte probabilité) de l'apparition dans la partie
 - mots prévisibles d'après le modèle des spécificités

Interprétation

- zone de banalité (entre -2.0 et 2.0) représentée sur le graphique pour éviter de surinterpréter
- toute valeur en dehors de cet intervalle (en dessous de -2.0 ou au-dessus de 2.0) indique une spécificité statistique significative
 - valeurs positives (> 2.0) : le mot est sur-représenté ;
 - valeurs négatives (< -2.0) : le mot est sous-représenté
- une significativité négative peut avoir du sens : **nullax**
 - mots ayant un indice négatif mais aussi une fréquence nulle
 - mots dont l'absence dans la partie est statistiquement étonnante, compte tenu de sa fréquence en corpus et de la taille de la partie

Calcul des spécificités

Analogie de la boîte à œufs :

- on renverse aléatoirement les œufs (occurrences d'un mot) dans les boîtes à œufs (partitions du corpus)
- rare que beaucoup d'œufs tombent dans une même boîte
- si les œufs sont répartis au hasard, ils devraient plutôt se distribuer plus uniformément entre les boîtes

Analyse factorielle des correspondances (Benzécri, 1973)

- synthèse globale des relations entre mots (traits linguistiques, motifs) et textes (parties du corpus)
- les mots sont comparés les uns aux autres sur la base des textes qui les emploient
- et réciproquement les ressemblances et écarts entre textes sont évalués par le vocabulaire qu'ils mobilisent

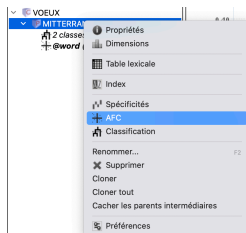


Figure 25 – Générer l'analyse factorielle des correspondances.

Classification ascendante hiérarchique (Benzécri, 1979)

Méthode de rassemblement des éléments selon un critère de ressemblance défini au préalable qui s'exprimera sous la forme d'une matrice de distances.

- le calcul s'effectue à partir des colonnes ou des lignes d'une table lexicale ou d'une partition

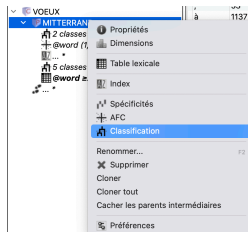


Figure 27 – Générer une classification ascendante hiérarchique.

Visualisation de la classification ascendante hiérarchique

Le dendrogramme avec des regroupements par classes d'éléments, composé de :

- de cadres de couleur correspondants aux regroupements par classes ;
- de l'échelle des indices de niveaux de regroupement située à gauche ;

En haut à droite : le diagramme des indices de niveaux (du nœud le plus haut au nœud le plus bas du dendrogramme).

Exemple de la CAH – corpus VOEUX

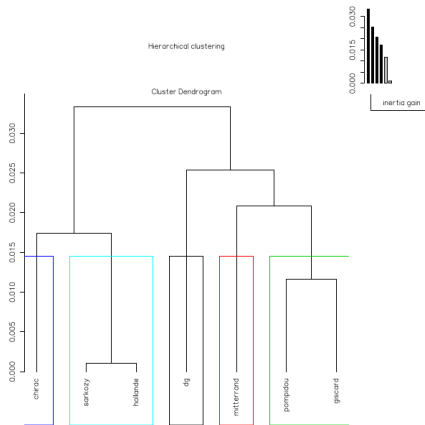










Figure 28 – La CAH représentée par le vecteur fréquence des formes graphiques en utilisant la méthode d'agrégation *ward* et une distance euclidienne.

Références

-  BENZÉCRI, J.-P. (1979). ***L'analyse Des Données : La Taxinomie (3^e, Vol. I)***. Paris, Dunod. (*voir p. 38*).
-  BENZÉCRI, J.-P. t. (1973). ***L'Analyse des données, Tome 1 : La taxinomie, Tome 2 : L'analyse des correspondances***. Paris, Dunod. (*voir p. 36*).
-  FORT, K. (s.d.). ***TXM : présentation et commandes de base***. Cours « Corpus, ressources et linguistique outillée », https://members.loria.fr/KFort/files/fichiers_cours/TXM_1.pdf. Consulté le 14 février 2025 (*voir pp. 1, 22*).
-  HEIDEN, S. (s.d.). ***Atelier préparation de corpus et import dans TXM***. Tutoriel, <https://txm.gitpages.huma-num.fr/textometrie/files/course%20materials/Diapositives%20-%20Atelier%20preparation%20de%20corpus%20et%20import%20dans%20TXM.pdf>. Consulté le 14 février 2025 (*voir pp. 1, 4, 6*).

-  LAFON, P. (1980). **Sur la variabilité de la fréquence des formes dans un corpus.** In : *Mots. Les langages du politique* 1.1. https://www.persee.fr/doc/mots_0243-6450_1980_num_1_1_1008, p. 127-165 (*voir p. 29*).
-  LEJEUNE, G. (2023). **TXM : la Textométrie à portée de clic.** Atelier TXM (Textométrie), <https://ceres.sorbonne-universite.fr/83ff891969d7e024646d832126d47f82/CERES-TXM.pdf>. Consulté le 14 février 2025 (*voir pp. 1, 9, 10, 19, 20*).
-  PINCEMIN, B. (2022). **Sémantique textométrique.** In : *La sémantique au pluriel. Théories et méthodes.* <https://shs.hal.science/halshs-03763801/>, p. 373-396 (*voir p. 34*).
-  PINCEMIN, B. (2012). **Atelier d'initiation à TXM de Bénédicte Pincemin du 27 septembre 2012.** Tutoriel, https://txm.gitpages.huma-num.fr/textometrie/html/enregistrement_atelier_initiation_TXM_fr.html. Consulté le 14 février 2025 (*voir p. 1*).

Licence

Le contenu de cette présentation est sous licence CC-BY-NC-SA 4.0
Utilisation non commerciale – Partage dans les mêmes conditions.

