

Corpus, ressources et linguistique outillée · M2SOL034

CM 2 : Fondamentaux de la textométrie et TXM

Ljudmila PETKOVIĆ

Semestre 2, 2024-2025

7 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) et PINCEMIN et HEIDEN (2008)

Les « métriques »

De la lexicométrie à la textométrie

Analyse de données textuelles (ADT)

Application de calculs sur des données textuelles (grands corpus).

Développement des disciplines en France :

- **lexico**métrie (*circa* 1970) : application sur le lexique (mots)
 - statistique lexicale : évaluation de la richesse du vocabulaire
 - analyses factorielles, classifications : cartographies synthétiques
- **texto**métrie (*circa* 2004) : application sur le texte
- **logo**métrie (*circa* 2004) : application sur le discours

Communautés scientifiques concernées

Sciences humaines et sociales :

- corpus scientifiques
- archives historiques
- dépouillement d'enquêtes avec questions ouvertes
- œuvres littéraires
- ...

Enjeux textométriques

Développer des modèles statistiques pour rendre compte de caractéristiques significatives des données textuelles :

- attirances contextuelles des mots
 - phraséologie, champs thématiques. . .
- linéarité et organisation interne du texte
 - mots répartis au fil du texte ou apparaissant en « rafales »
- contrastes intertextuels
 - mesure statistique du sur-/sous-emploi d'un mot dans un texte
 - repérage des mots et des phrases caractéristiques d'un texte
- indicateurs d'évolution lexicale
 - période caractéristique d'un terme
 - détection des ruptures significative

La textométrie au service de la linguistique outillée

Calculs mathématiquement et **linguistiquement** significatifs :

≠ recherche d'information : focus sur les problématiques documentaires

- probabilités, statistiques, analyse des données
- expression et traduction mathématique d'hypothèses sur la langue et la textualité
- vue globale vs. consultation ciblée des contextes d'emploi

Retour au texte : prendre du recul pour interpréter des résultats.

« *L'outil dégrossit, l'humain interprète* » ([LEJEUNE, 2023](#))

≠ analyse sémantique latente : passage à d'autres disciplines

≠ TAL : calibrage des calculs

Modèle SÉMA

Synthèse : calculs statistiques → vues synthétiques significatives

- caractérisation des singularités d'un texte
- repérage des thèmes

Édition : présentation du texte (accès aux contextes)

Moteur de recherche : repérage des occurrences d'un motif donné

Annotation : enrichissement des corpus au fil des analyses

Au-delà des moteurs d'internet

Mettre en évidence des contrastes significatifs

- caractérisation et repérage des singularités

Expliciter les fonctionnalités à tous les niveaux

- théoriques, informatiques, méthodologiques...

Vision globale, qualitative, respectant une pluralité de réponses

≠ moteurs d'internet

- plus une page est citée, plus elle est mise en valeur
- critères de sélection opaques \approx « boîtes noires »
- conception « compétitive » : résultats classés par ordre de pertinence

(Ré)introduction à TXM

TXM³

- projet ANR « Textométrie »
- communauté d'utilisateurs et de développeurs active
- logiciel libre : pérenité → possibilité de faire évoluer le code
- multi-plateforme (Windows, Mac, Linux)
- portail en ligne txm-demo¹
- technologies de corpus supportées
 - Unicode, XML, TEI, outils de TAL, CQP, R
- analyse de grands corpus, structurés ou non
- intégration des outils externes
 - p. ex. TreeTagger² – étiquetage morphosyntaxique

1. <https://txm-demo.huma-num.fr/txm/>

2. <https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

3. <https://txm.gitpages.huma-num.fr/textometrie/>

Avantages de l'utilisation de TXM

- **interface** très complète
- **robustesse** : permet de traiter jusqu'à 10 millions de mots
- **puissance** : permet d'intégrer toutes sortes de traitement via le logiciel R (de statistiques)
- **rapidité** : permet d'interroger des millions de mots très efficacement via CQP (*Corpus Query Processor*)

(FORT, *s.d.*)

Fonctionnalités

Analyses statistiques basiques

- Index
- Concordances
- Cooccurrences

Analyses avancées

- Attirance contextuelle des mots et des expressions
- Spécificités lexicales
- Linéarité et organisation interne du texte
- Comparaisons de sous-corpus

Lexique

- liste de formes (ou de tokens)
- fréquence d'apparition
- lemmatisation / étiquetage (TreeTagger)
 - tâches maîtrisées mais non résolues
 - forme canonique (suppose corpus monolingue pour TXM)
 - étiquettes p. ex. NOM, ADJ, VER, ADV + morphologie
- revisiter le mot dans son contexte
- allers et retours entre le lexique et le corpus

Concordances

Vue synthétique des occurrences d'une forme (d'un motif) :

- ses contextes d'apparition
- triés de différentes façons

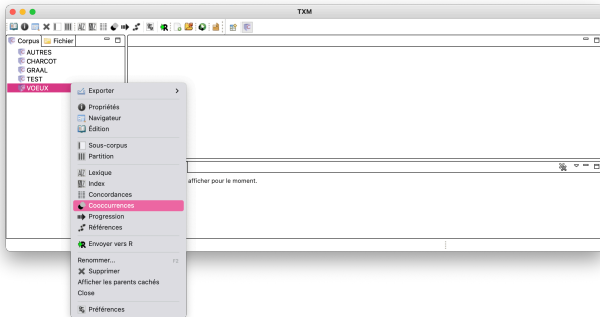
Utilisations :

- distribution dans le corpus
- expressions dérivées
- structures grammaticales

Concordances

Concordancier

Logiciel qui permet de faire un tri rapide de tous les mots d'un texte (ou d'un ensemble de textes), de situer des mots-pivot en contexte (KWIC – *Key Word in Context*), de compter le nombre d'occurrences, etc., à partir de chaînes de caractères.



Concordances

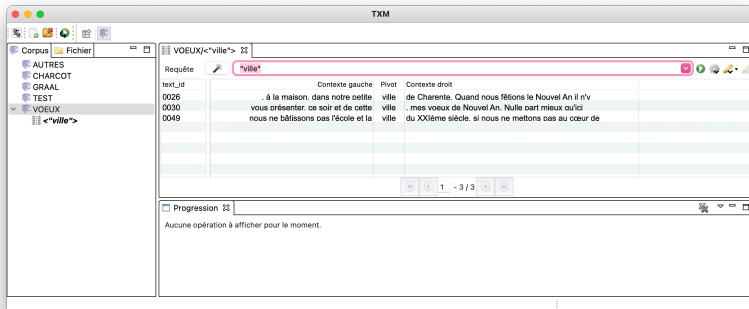


Figure 2 – Affichage du mot-pivot « ville » dans le corpus VOEUX.

Concordances : le pivot

Le pivot peut être :

- un mot (cas le plus simple)
- une séquence de mots
- un motif simple (détectable avec une expression régulière)
- un motif complexe (lexico-)syntaxique en langage CQL
(*Corpus Query Language*)

Propriétés du pivot

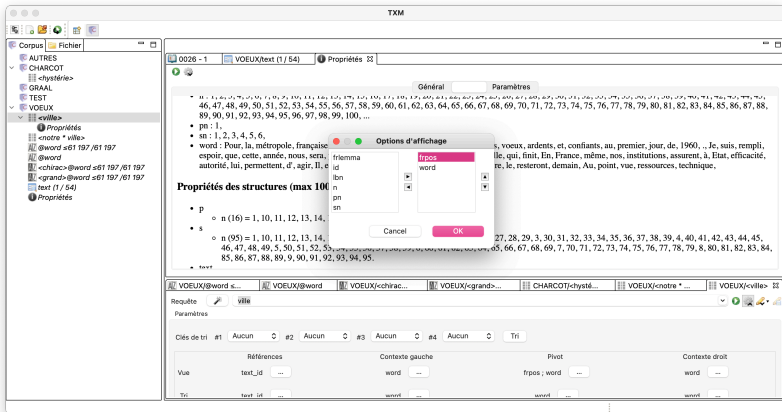


Figure 3 – Propriétés du pivot « ville » dans le corpus VOEUX.

Propriétés du pivot

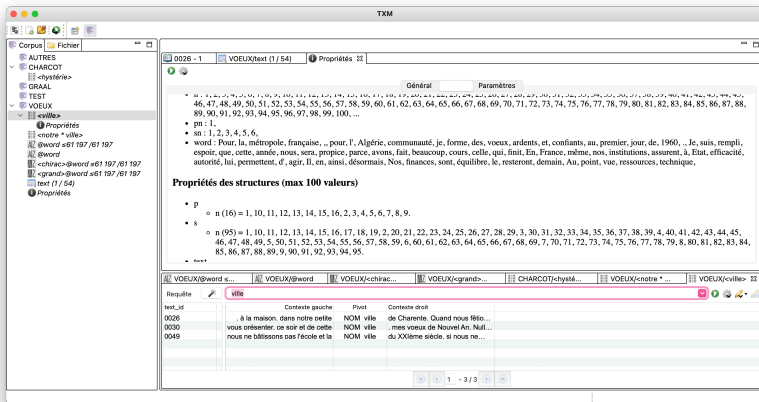


Figure 4 – Affichages des propriétés du pivot « ville » dans le corpus VOEUX.

Fréquences lexicales

Vue des types (mots uniques) et des tokens (formes de mots)

Table des fréquences : distribution par type (y compris les étiquettes POS)

Selon la loi de Zipf, on retrouve :

- en premières positions des mots grammaticaux
- en positions inférieures : mots sémantiquement chargés ou du genre textuel (si corpus homogène)

word	Fréquence ▾
.	4400
,	2461
de	2451
la	1762
et	1502
"	1010

Figure 5 – Tableau des fréquences (extrait).

Lexique vs. index : deux fonctionnalités différentes

Lexique

- calcule la fréquence pour une propriété de mot donné
 - forme, lemme... mais pas d'expressions complexes
- première visualisation du corpus : thèmes, *hapax*

Index

- calcule la fréquence d'une expression (mot unique ou non)
- agit comme un filtre sur le lexique
- adapté à la recherche à tâtons dans le corpus

Lexique vs. index

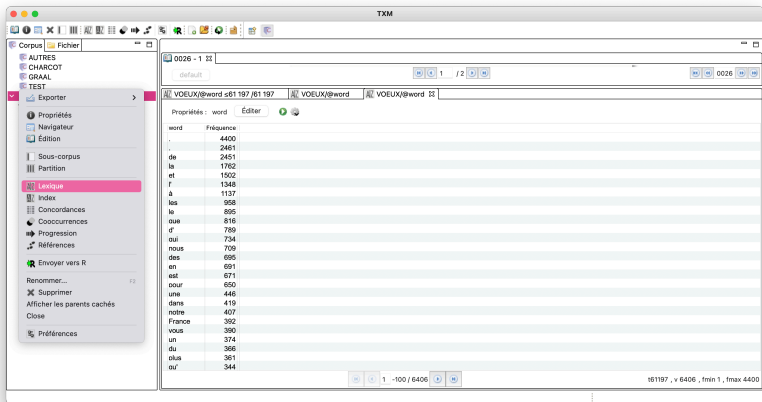


Figure 6 – Lexique (extrait).

Lexique vs. index

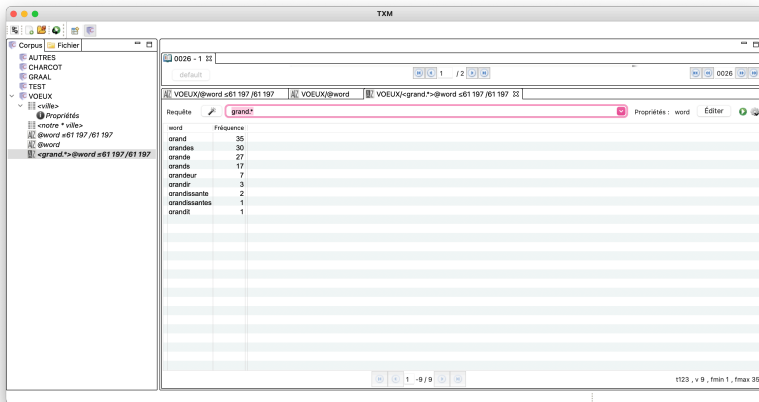


Figure 7 – Index (extrait).

CQL – *Corpus Query Language*

- expressions régulières :
 - `Europe|européen.*`, `[]` (un mot)
 - `&` et `|` (opérateurs booléens)
- neutralisations (à ajouter après l'expression) :
 - `%c` pour neutraliser la casse (`"europe"%c`)
 - `%d` pour neutraliser les diacritiques (accents, cédille)
 - `within s` : dans l'espace de la phrase
 - `within [X]` : dans l'espace de X mots
- assistant de requête
- tri du contexte droit et du contexte gauche

Relation entre TXM et TAL

TXM n'est pas un outil de TAL en tant que tel, mais

- il intègre des fonctionnalités de TAL, *via* TreeTagger
- il permet d'explorer les corpus et de les analyser manuellement (préalable au TAL)




Bien démarrer avec TXM

Installation

Téléchargement du logiciel + extension TreeTagger et prérequis :

<https://txm.gitpages.huma-num.fr/textometrie/files/software/TXM/0.8.3/>

Références

-  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 7 février 2025 (*voir pp. 1, 11*).
-  LEJEUNE, G. (2023). ***TXM : la Textométrie à portée de clic.*** Atelier TXM (Textométrie), <https://ceres.sorbonne-universite.fr/83ff891969d7e024646d832126d47f82/CERES-TXM.pdf>. Consulté le 7 février 2025 (*voir pp. 1, 6*).
-  PINCEMIN, B. et S. HEIDEN (2008). ***Qu'est-ce que la textométrie ? Présentation.*** Site du projet Textométrie, <https://txm.gitpages.huma-num.fr/textometrie/Introduction/>. Consulté le 7 février 2025 (*voir p. 1*).