

M2SOL034 Corpus, ressources et linguistique outillée

TD 2 : TXM I

Ljudmila PETKOVIĆ

Sorbonne Université

Master « Langue et Informatique » (M1 ScLan)

UFR Sociologie et Informatique pour les Sciences Humaines

Semestre 2, 2024-2025, le 25 février 2025

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



Table des matières

1 Exercices : commandes de base	1
2 Solutions	2

1 Exercices : commandes de base

1. Après avoir installé TXM, chargez ¹ le corpus VOEUX déjà importé dans TXM, et accédez aux propriétés du corpus.
Indiquez (a) l'année la plus ancienne et (b) l'année la plus récente de l'édition.
2. Téléchargez le corpus sous format XML-TEI `charcot.xml` depuis Moodle (cf. le répertoire `ressources_TD2`), et importez-le dans TXM.
En une seule requête, affichez la liste des mots-pivots dérivés du mot `hystérie`.
3. Téléchargez le corpus *Du côté de chez Swann* de Marcel Proust sous format texte brut depuis le site du projet Gutenberg <https://www.gutenberg.org/files/12000/12000-h/12000-h.htm>.

1. charger : corpus déjà importé dans TXM auparavant ; importer : corpus brut (`txt`, `XML`, voire en provenance du presse-papier).

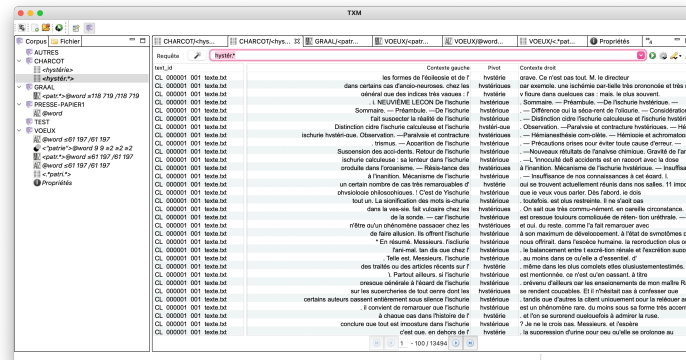
gutenberg.org/ebooks/2650, et importez-le dans TXM. Exportez le lexique du corpus importé dans un tableau. Que pouvez-vous constater concernant la répartition des fréquences de mots ?

4. Trouver en une seule recherche les des mots-pivots dérivés du mot « patrie » à partir du corpus VOEUX. Quels mots avez-vous extrait ?
 5. Trouver en une seule recherche le souhait de bonne année de chaque Président dans le corpus VOEUX.
 6. Affichez la liste des cooccurents du terme **hystérie**, avec leurs fréquences, leurs indices de spécificité et leurs distances moyennes à partir du corpus CHARCOT.
- Quel est l'indice de spécificité et la distance moyenne du cooccurrent **convulsive** ?
Comment interprétez-vous ces mesures ?
7. Construisez, à l'aide de TXM et sur un tableau un graphique illustrant la loi de Zipf sur le texte de votre choix. Analysez les résultats obtenus.

2 Solutions

NB : les calculs ont été effectués dans la version 0.8.1 du TXM sur Mac.

1. (a) l'année la plus ancienne : 1959
(b) l'année la plus récente : 2012
2. Nous obtenons une liste des mots-pivots dérivés du mot **hystérie** à partir du corpus CHARCOT :



3. Le lexique du corpus importé montre une distribution des fréquences lexicales évoquant la loi de Zipf, selon laquelle la fréquence d'un mot est inversement proportionnelle à son rang dans la liste globale des mots après le tri par ordre décroissant de fréquence. Autrement dit, peu de

mots apparaissent très souvent, tandis que la majorité des mots apparaissent rarement.

4. Afin de récupérer les fréquences des mots-pivots dérivés du mot « patrie », nous utilisons l'expression régulière `.*patri.*`. Cela nous permet d'extraire les mots *patrie*, *patriote*, *patriotisme*, *compatriotes* et *rapatriés*.
5. `[frlemma="je"] []*[frlemma="souhaiter"] []*[frlemma="année"] within s`
6. L'indice de spécificité du cooccurrent **convulsive** est 9 et la distance moyenne est 1.3.
Concernant l'indice de spécificité, plus il est élevé, plus la cooccurrence est remarquable.
Pour ce qui est de la distance moyenne, elle indique le nombre de mots entre le cooccurrent et le pivot quand ils sont dans le même voisinage.
7. Discussion en TD.