

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 7 février 2025

Table des matières

1 Exercices : commandes de base	1
2 Solutions	2

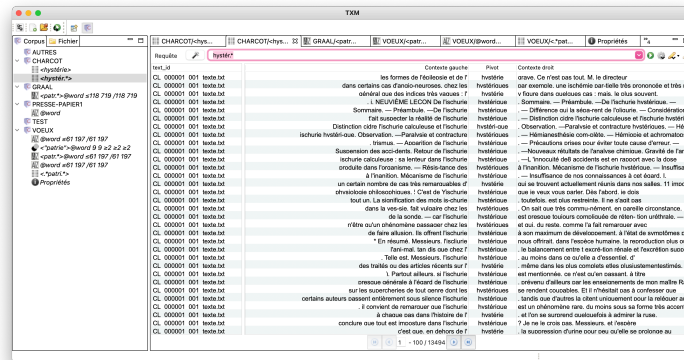
1 Exercices : commandes de base

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

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

- ## 2 Solutions

- l'année la plus ancienne : 1959
 - l'année la plus récente : 2012
- Nous obtenons une liste des mots-pivots dérivés du mot **hystérie** à partir du corpus **CHARCOT** :



- 2

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 statistique du cooccurrent `convulsive` est 43 et la distance moyenne est 2.1.
 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.