M2SOL034 Corpus, ressources et linguistique outillée

TD 3 : TXM II

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

Table des matières

1 Exercices : commandes avancées

Rappel : corpus VOEUX est un corpus de 54 discours de présidents français pour le Nouvel An (1959-2009).

À partir du corpus VOEUX, créer des partitions et répondre aux questions suivantes :

1. quelles expressions permettent de trouver *France*, *français* et *françaises* ? Vérifier si les résultats sont satisfaisants.
2. rechercher le terme *sécurité* dans les discours de tous les présidents. L'analyse des fréquences paraît-elle comme une méthode fiable dans ce cas ?
3. générer un graphique de spécificité de la lemme *Algérie* dans tous les discours et analyser le résultat.

4. générer un graphique de progression des termes *États-Unis* (attention aux diacritiques) et *Algérie*, en incluant la segmentation du graphique par les différentes parties (noms des présidents). Analyser le résultat. Que peut-on déduire du point de vue de :
   — fréquence globale ;
   — progression temporelle ?
5. générer un graphique d'analyse factorielle des correspondances en s'appuyant sur la propriété de lemme. Analyser le résultat. Quels présidents aborde le sujet de la sécurité ?
6. Travail sur un corpus externe
   (a) télécharger le corpus MPT (Mariage pour tous) : https://gitlab.huma-num.fr/txm/txm-ressources/-/blob/master/corpora/mpt/MPT-2023-06-23.txm
   (b) l'importer dans TXM (option `Charger > un corpus binaire (.txm)...`)
   (c) trouver une hypothèse (ex : les femmes subissent plus d'interruptions que les hommes)
   (d) la valider ou l'invalider par le corpus (argumenter)

   Rapport à rendre (PDF) + présentation (5 min + 1 min de questions), travail en binôme possible.

2 Solutions

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

1. `[word="(F|f)ran.*" & frlemma != "franc"]`
2. Les parties étant inégales, on ne peut pas comparer les fréquences brutes d'une partie par rapport à une autre. Le calcul des spécificités de Lafon corrige ce biais en ajustant les fréquences en fonction des tailles relatives des sous-corpus.
3. Après avoir sélectionné l'option
   `Calculer le diagramme en bâtons des lignes sélectionnées`,
   le graphique correspondant apparaît (Figure 1), où nous observons deux lignes rouges qui délimitent deux zones de banalité :
   — une entre -2.0 et 0 ;
   — une entre 0 et 2.0.
   Si l'indice est entre -2.0 et 2.0, le mot est considéré comme banal dans cette partie (pas de spécificité significative).

   Au-delà de ces seuils, le mot est soit sur-représenté (au-dessus de 2.0), soit sous-représenté (en dessous de -2.0).

   Le mot *Algérie* est très sur-représenté dans les discours associés à de Gaulle (ce qui est historiquement pertinent), contrairement à ceux de

Chirac, où ce mot est sous-représenté par rapport à ce qui serait attendu si les occurrences étaient réparties de manière aléatoire. Concernant les autres discours, le mot peut être considéré comme banal, sans variation significative.

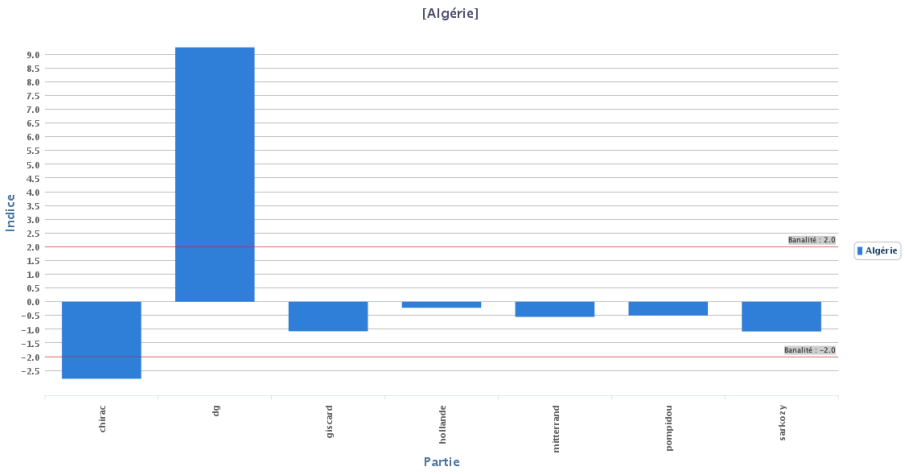

FIGURE 1 – Diagramme en bâtons indiquant les spécificités du mot `Algérie` dans les discours présidentiels.

4. Nous utilisons deux requêtes :
   — `[frlemma = "États-Unis"%d]` (avec la neutralisation des diacritiques)
   — `[frlemma = "Algérie"]`

   qui génèrent le graphique de progression cumulative des occurrences des lemmes `États-Unis` (en rouge) et `Algérie` (en bleu) sur la Figure 3.
   — au niveau de la fréquence globale, le lemme `Algérie` est mentionné deux fois plus souvent, surtout dans les discours associés à de Gaulle (21 occurrences) que `États-Unis` (11 occurrences) dans l'ensemble du corpus ;
   — le lemme `Algérie` est plus fréquemment mentionné et plus tôt dans le corpus (nous observons un saut de courbe, suivi d'un grand plateau), tandis que `États-Unis` apparaît de manière sporadique et plus tardive (la courbe est moins raide).

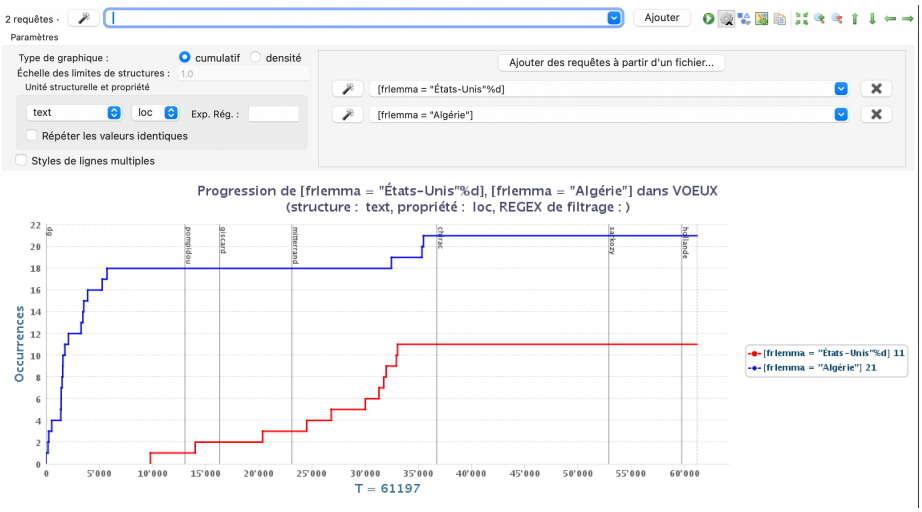

FIGURE 2 – Progression des lemmes `États-Unis` et `Algérie`.

5. Hollande, Chirac et Sarkozy sont situés à droite et associés à des mots comme `chômage`, `gouvernement`, et `sécurité`, reflétant des préoccupations modernes et sociales.

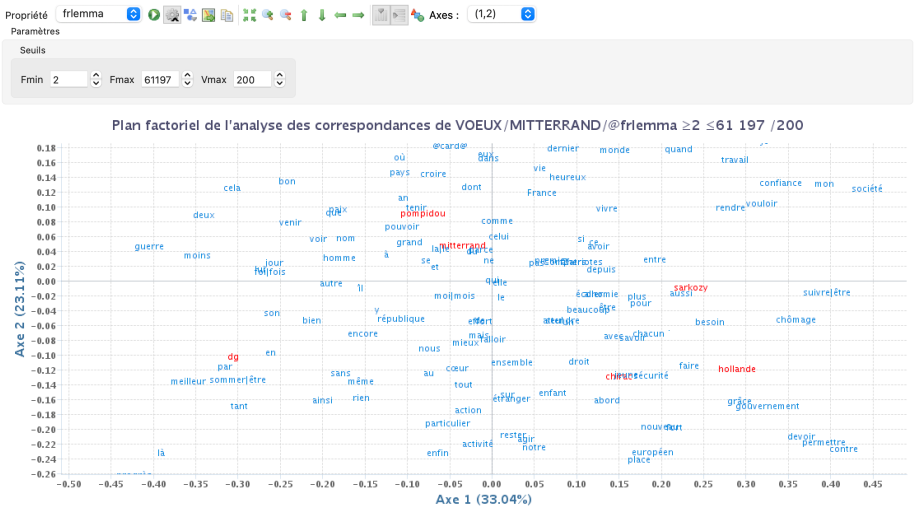

FIGURE 3 – Analyse factorielle des correspondances – corpus `VOEUX`.

6. Discussion en TD.