Découverte de Txm

Y. Dupont, G. Lejeune, L.Lefeuvre, K. Fort Sorbonne Université

Objectifs partie 1

- Prendre en main le logiciel Txm
- Comprendre les formats d'entrée et de sortie

Exercice 1 : Premiers pas

Démarrer avec un corpus

Nous allons utiliser le terme **Charger** lorsque le corpus est déjà étiqueté, segmenté en mots (*tokenisé*), ... mais **Importer** lorsque l'on a besoin que le logiciel se charge de découper/étiqueter/indexer le corpus. On importe un corpus quand il est "brut" (par opposition à "enrichi").

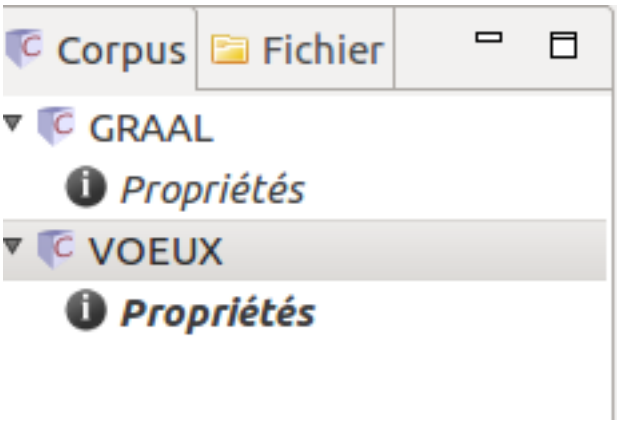

Figure 1: Fenêtre des corpus disponibles

- Le corpus "Vœux" devrait être déjà installé. S'il n'apparaît pas dans le menu de gauche (cf Figure 1):
  - Récupérez-le à partir du dossier `corpora` disponible ici : lien de téléchargement (ne pas décompresser l'archive)
  - Chargez le corpus : Fichier > charger

Quelques manipulations de base disponible avec un clic droit sur le corpus Vœux:

- Propriétés : vous aurez quelques détails sur le corpus
- Edition : en survolant les mots du corpus avec votre souris vous verrez . Quelles informations apparaissent ?
- Lexique : statistiques sur le vocabulaire du corpus, des mots mais pas que !

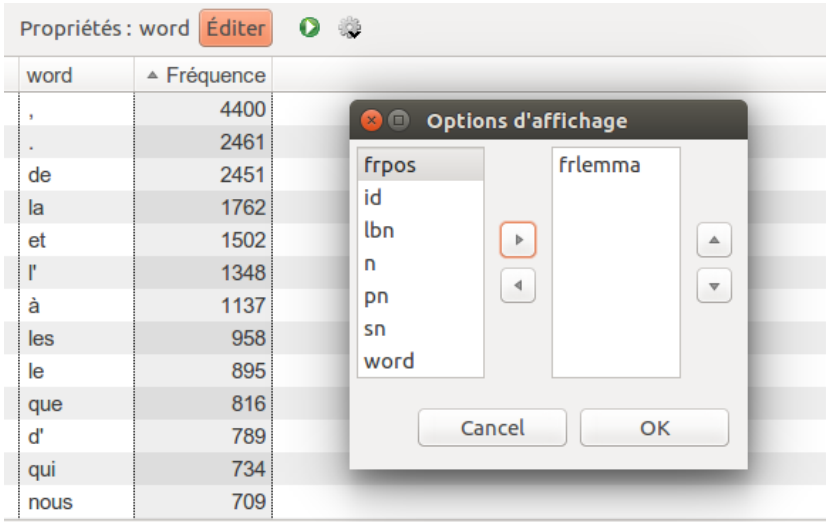

Figure 2: Modifier les informations affichées sur le lexique

Différentes catégories de mots

- Affichez l'ensemble des étiquettes POS du corpus (sélectionnez grâce au bouton "éditer" puis valider avec le petit bouton "play" juste à droite)
- Affichez l'ensemble des lemmes (formes dictionnairiques) du corpus
- Faites apparaître les hapax (mots d'une fréquence de 1) .

Exercice 2 : Utiliser un Index

L'index permet de " filtrer le lexique " à l'aide d'une requête (joue le rôle d'une condition), et donc de rechercher la fréquence d'une expression. Le langage de requête utilisé est le langage CQP. Jetez un coup d'œil sur la documentation de TXM sur le sujet et gardez la de côté, cela pourra vous être utile par la suite : Documentation.

- Sur le corpus Vœux, faites apparaître l'index des formes du verbe "souhaiter" avec la requête "souhaite.*". Le "." signifie "n'importe quel caractère" et l'opérateur "*" signifie "zéro ou plusieurs occurrences de ce qui précède". Ce sont des éléments du vocabulaire des expressions régulières[1]
- En une seule requête, trouvez les fréquences de la liste suivant : "patriotisme" , "compatriotes" "patrie" , "patriote" ,
- En une seule requête, trouvez le souhait de "bonne année" de chaque discours dans l'onglet concordance. Utilisez l'assistant de requête pour cela
- Retournez à l'index et cherchez quelles sont les unités composées de deux mots séparés par une espace, on parle aussi d'unités polylexicales.

Objectifs partie 2

- Utiliser de nouveaux corpus avec TXM
- Etiqueter morpho-syntaxiquement
- Faire des requêtes complexes

[1] Cous pouvez vous référer à la page Wikipedia sur le sujet (section "opérateurs")

Exercice 3 : Requêtage, tri et analyse de Concordances

Requête et tri

Toujours sur le corpus Vœux, observer les concordances du mot "temps".

- Triez selon le contexte droit pour faire apparaître les co-occurrents droits.
- Triez selon le contexte gauche : observez comment sont "rangés" les résultats
- Faites apparaître l'année et le locuteur(=méta-données) : clic droit sur text_id → options d'affichage des références

Contextes des occurrences

- Accédez au texte complet d'une occurrence de "temps" (clic droit puis afficher en plein texte) afin d'observer son contexte complet, on parle aussi de **retour au texte**.
- Observez que vous pouvez manipuler les différentes fenêtres. Mettez la fenêtre plein texte en regard de la liste des concordances pour faciliter les comparaisons.
- Modifiez la fenêtre d'affichage des concordances (taille du contexte gauche et du contexte droit).
- Affichez les concordances du nom "jeune" dans le corpus (singulier et pluriel) grâce aux requêtes CQL
- Modifiez l'affichage du pivot pour faire apparaître l'étiquette POS dans les résultats (cf. Figure 3)
- Affichez les formes verbales de "pouvoir" sans les formes nominales

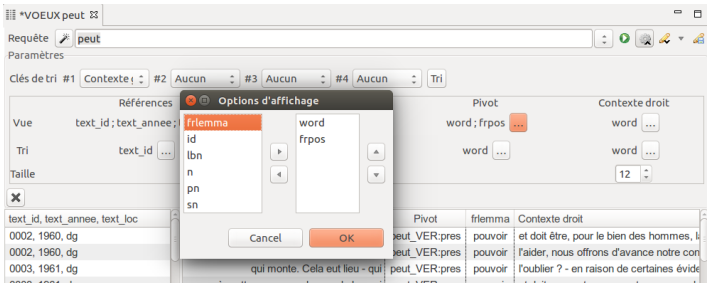

Figure 3: Ajout de propriétés dans l'affichage du pivot

Exercice 4 : Fonctionnalités supplémentaires

Import Xml simple dans Txm

Récupérez l'archive `https://lejeunegael.fr/resources/1999-05-17.zip` puis dézippez la. Placez le fichier correpsondant dans un dossier dédié (peu importe le nom)

- Menu Fichier puis importer, choisissez le format XML/w +csv.

- Sélectionnez le dossier créé précédemment (qui contient normalement un seul fichier XML)
- Dans la partie "Langue principale" (cf. Figure 4), sélectionnez "fr" (si vous ne le faites pas, rien n'apparaîtra dans le menu contextuel après l'import).
- Lancez l'import du Corpus

Figure 4: Le menu import

ANNEE1999 devrait apparaître dans le menu de gauche

Observez les propriétés de ce Corpus ANNEE1999. Sont-elles renseignées ? Pourquoi d'après vous ?

Le corpus n'est pas enrichi, TXM ne peut donc afficher que des choses assez simple. Observez dans le menu index via la baguette magique (Figure 5) que vous n'avez accès qu'à des informations limités sur les mots (Figure 6).

Figure 5: L'icône de la baguette magique

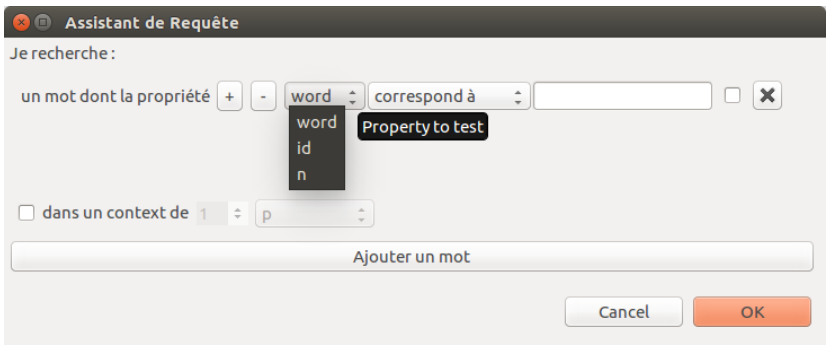

Figure 6: Propriétés des mots sans enrichissement

Vous vez vu que l'import ici se fait par format et **par dossier**.

Enrichir le Corpus

Avec le menu Fichier puis ajouter une extension, ajoutez le module Treetagger (deux éléments : base et modèles).

Vous allez ré-importer le corpus mais cette fois en activant l'option "Annoter le corpus" (cf. Figure 4).

Treetagger va avoir lemmatisé et étiqueté morpho-syntaxiquement le corpus, vous pouvez le voir en allant dans la fonction index (cf. Figure 7).

Figure 7: Propriétés des mots après enrichissement

La liste des étiquettes utilisées par Treetagger figure ici : `https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html`.

Utilisez l'index pour chercher les noms propres les plus fréquents (l'étiquette est NAM).

Objectifs partie 3

- rechercher des indices lexicaux
- comparer des sous-ensembles d'un corpus
- trouver des spécificités à des sous-ensembles d'un corpus

Exercice 5 : Progression dans le temps

Nous allons étudier ici la progression de l'utilisation d'une valeur (au sens de TXM) au cours du temps.

Retournons sur le corpus VŒUX. Nous allons évaluer l'évolution de certains usages au fil du temps et des présidents.

Pour cela, utilisez la fonction progression sur le corpus VŒUX.

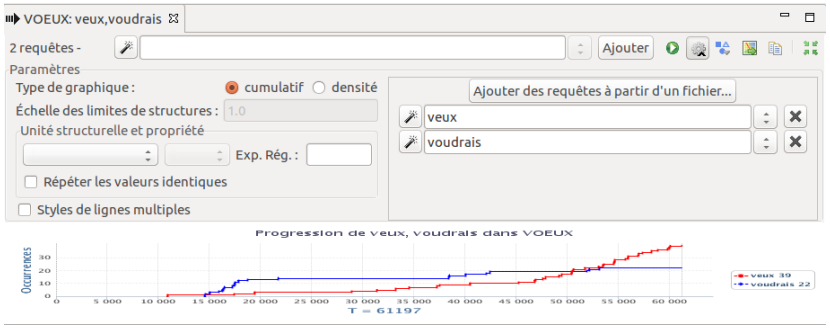

Figure 8: Les paramètres de l'onglet progression

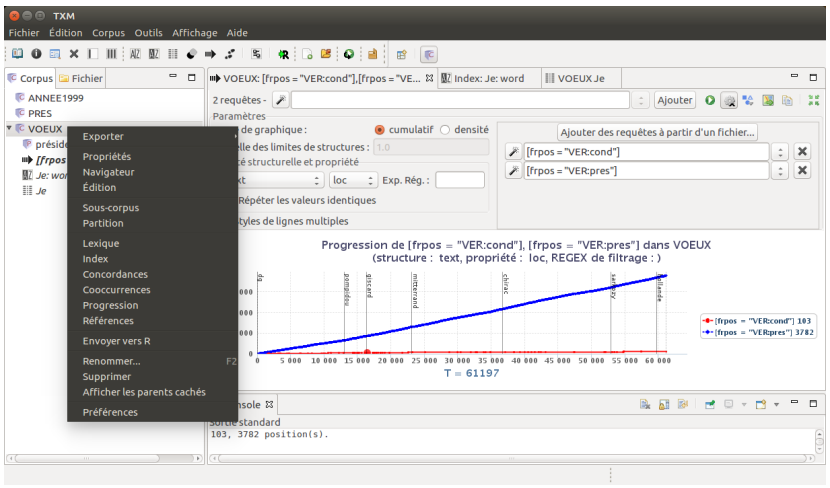

Figure 9: Menu contextuel du Corpus

1. Dans un premier temps, comparez l'évolution de l'usage des mots "veux" et "voudrais". Il est possible d'ajouter les mots un par un en les recherchant successivement. Notez-vous quelque chose d'intéressant ?

2. Affichez les paramètres de votre onglet progression (avec la roue dentée cf. Figure 8

3. Dans la partie "Unité structurelle et propriété", configurez TXM afin que les séparations (barres verticales) affichent les président (locuteurs). Choisissez dans le premier menu déroulant l'option "text" et dans le second l'option "loc".

4. Sous quel président l'usage de "voudrais" a-t-il connu la plus grande augmentation ? Sous quel président la tendance entre l'usage de "veux" et de "voudrais" a-t-il changé ? Peut-on supposer une future nouvelle inversion, et pourquoi ?

5. Comparez à présent les verbes au présent et ceux au conditionnel grâce à l'assistant de requêtes (propriété frpos) et un utilisant le jeu d'étiquettes donné au TD précédent[2]. La courbe semble-t-elle indiquer la même évolution que ce que vous avez vu précédemment? Pourquoi ?

Exercice 6 : Analyse Factorielle des Correspondances (AFC)

Créez une partition du corpus VŒUX (clic droit sur le corpus dans la fenêtre de gauche cf. Figure 9) sur la structure "text" et la propriété "loc" que vous appellerez PRÉSIDENTS. La partition que vous venez de créer s'est affichée dans la fenêtre gauche. Sur cette partition que vous venez de créer, afficher les options disponibles via le clic droit et choisissez AFC (Analyse Factorielle des Correspondances)

1. Cachez les mots en cliquant dans la barre d'outils (au-dessus du graphique) sur "Afficher/Masquer les lignes", vous verrez uniquement les noms des présidents. Qui sont les deux présidents les plus éloignés des autres ?

2. Sélectionnez la propriété "frlemma", cliquez sur le triangle plan sur fond rouge juste à droite pour mettre à jour le graphique. Ré-affichez les lignes pour voir de nouveaux les mots dans le graphique. Cherchez les 3 mots qui rapprochent

[2] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html

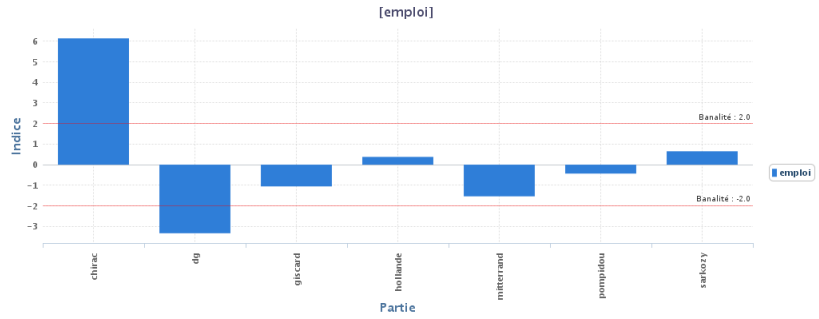

Figure 10: Spécificité du mot Emploi (typique de Jacques Chirac) dans les différentes partitions

Mitterrand et Pompidou, Pompidou et Giscard, Sarkozy et Chirac (Vous pouvez zoomer sur le graphique).

3. Utilisez maintenant la propriété "frpos". Quelles sont les 3 catégories les plus centrales ?

4. Toujours avec la propriété "frpos". Deux présidents usent d'une catégorie grammaticale bien plus que les autres. Donnez ces deux présidents et la catégorie associée.

Exercice 7 : Spécificités

Nous allons ici regarder les usages spécifiques à chaque président.

Pour le corpus PRÉSIDENTS, utilisez la fonction Spécificités afin d'accéder aux valeurs spécifiques de chaque président.

1. Regardons la propriété frpos pour voir les plus spécifiques pour chaque président. Avez-vous des résultats similaires avec ceux que vous avez eu en utilisant l'AFC ?

2. Donnez, pour chaque président, les 5 lemmes (propriété frlemma) qui lui sont les plus spécifiques (le plus grand indice). Dans quelle mesure ces lemmes sont-ils similaires avec ceux que vous avez trouvé avec l'outil AFC ?

3. Pour chaque *nom commun* le plus spécifique de chaque président, comparez son usage avec celui fait par les autres présidents (clic-droit sur le mot, puis "calculer le diagramme en bâtons..."),ceci va ouvrir un nouvel onglet comme dans la Figure 10.

4. Signalez, pour chaque président, le nom commun et les présidents pour lesquels la valeur de "banalité" est dépassée.

5. Certains mots ne sont pas utilisés par un président spécifique, on les appelle les "nullax". Donnez 5 "nullax" de chaque président.

Exercice 8 : Sous-corpus et partition

L'utilisation d'un sours-corpus permet de se concentrer sur une partie précise du corpus global.

1. Créez le sous-corpus chirac-pred en utilisant la fonction "sous-corpus" et en sélectionnant la propriété "loc". Ce corpus doit contenir les vœux de Chirac et de ses prédécesseurs. Comparez le vocabulaire spécifique de Chirac (frlemma) avec celui des autres présidents. Quelle différence voyez-vous par rapport à la question précédente ?

2. Même question en créant un sous-corpus chirac-succ, qui contiendra les vœux de Chirac et de ses succésseurs.