(suite du TD 2)

On travaille sur le corpus des romans de Zola (corpus de travail → multicritères), qu'on compare au corpus des romans de Balzac. Quel auteur **utilise-t-il le plus** verbe *chercher* (conjugué) ? **Justifiez**.

Frantext catégorisé

Lors d'une recherche sur corpus, deux types de biais sont très fréquents. On qualifie de **bruit** les résultats retournés par une requête, mais non pertinents, et de **silence** les résultats pertinents, mais non retournés par une requête.

Familiarisation avec le corpus

Sélectionnez « tous les textes» comme corpus de travail. Allez dans « Stats ». Combien cette base textuelle comporte-t-elle de textes, de mots, et sur quelle période ? Quelle est la différence avec Frantext intégral ?

Frantext catégorisé se distingue aussi par la présence d'**annotations métalinguistiques** : les catégories grammaticales des mots (on parle aussi souvent de **parties du discours** − en anglais *part of speech*, POS).

Recherchez les flexions du verbe *taire*, et visualisez les résultats.

Dans la fenêtre de visualisation, affichez les « Codes ». Quelle information apportent ces codes ?

Recherchez la séquence *en même temps*. Pour Frantext, combien de mots comporte cette locution ?

Parties du discours

Codes des parties du discours : http://cid.ens-lyon.fr/aide/ac_article.asp?fic=frantext_categgram.asp

Rechercher une partie du discours : &e(g=*code*), par exemple &e(g=V) pour les verbes.

Recherche un mot en précisant la partie du discours : &e(g=*code* c=*mot*), par exemple &e(g=S c=test)

Recherche d'un mot fléchi en précisant la partie du discours : &e(g=*code* c=&m*verbe*)

Recherchez les adverbes, puis, dans un second temps, la locution adverbiale *en même temps*.

Recherchez le verbe *taire* (mot fléchi) au participe passé (partie du discours). Combien d'occurrences trouvez-vous ? Citez un biais qui cause du **bruit** si on n'utilise pas la catégorisation.

Quels substantifs utilise-t-on le plus souvent directement (sans article) après le verbe *faire* ? Avec un article ? Est-ce suffisant pour étudier les compléments d'objet préférés du verbe *faire*, ou bien y a-t-il du **silence** ?

Créez un corpus de 4 œuvres de Zola : La Curée, Germinal, L'Œuvre et L'Argent. Créez une liste de mots fléchis liés à la notion de dette : *dette*, *créance*, etc. Dans lequel de ces 4 romans ces mots sont-ils les plus fréquents ?

Quels sont les compléments d'objet direct préférés du verbe *faire* ? Citez un biais qui peut provoquer du **silence**.