Méthodologie

Master : Sciences du langage

marine.delaborde@cyu.fr

Séance 3 : Interroger des corpus

Organisation du cours

Liens utiles

- **Fiche d'informations** (à remplir une seule fois) : https://forms.gle/tSBzQ6gvZeBVjsH37

- Site du cours : https://marine-delaborde.github.io/methodologie

Les séances

1/4 séances

Séances	Points abordés
Séance 1 : 16/11/2022 13h30 - 15h30	Recherche documentaire
Séance 2 : 13/12/2022 14h30 - 16h30	Traitement de textes
Séance 3 : 18/01/2023 13h30 - 15h30	Interroger des corpus
Séance 4 : 25/01/2023 13h30 - 15h30	Examen

Pour aujourd'hui

Exercice

1. Récupérer le template du mémoire (sciences de l'éducation) sur le site de l'université
2. Ajouter les informations connues liées au mémoire (informations personnelles, sujet, etc.)
3. Supprimer les informations non nécessaires
4. Ajouter la bibliographie de la séance 1 (à l'aide du gestionnaire dans le traitement de texte ou de zotero)

Résultat à m'envoyer par mail avant aujourd'hui (18/01/2023).

FRANTEXT

Un corpus et ses outils

- **ATILF - CNRS** : http://www.atilf.fr/spip.php?rubrique60
- Projet parallèle à celui du dictionnaire “**Trésor de la langue française**”
  - Accessible en ligne : www.cnrtl.fr Portail lexical > Lexicographie
- **Corpus** :
  - En ligne depuis **1998** - mis à jour en **septembre 2022** (nouvelles interface / fonctionnalités depuis 2018)
  - 5 571 références → 265 millions de mots
  - Textes du IX^{ème} au XXI^{ème} siècle
- **Outil** : https://frantext.bibdocs.u-cergy.fr/ (accès via CY)
  - Recherches dans les textes du corpus
    - Recherches simples (forme)
    - Recherches complexes (lemmes, catégories grammaticales, listes…)
- **Documentation** : https://wiki.frantext.fr/bin/view/Main/Manuel%20d%27utilisation/Corpus/

Application

Exercices sur Frantext intégral

1. Métadonnées :
   - 1.1. Quelle période est la plus représentée ?
   - 1.2. Quel domaine est le moins représenté en ancien français ?
   - 1.3. Quel est le genre principal des textes de linguistique ? Quel problème peut-on relever à propos de cette recherche ?

2. Créer des sous-corpus :
   - 2.1. Créer deux sous corpus de textes, l'un constitué de romans écrits entre 1900 et 1999 et l'autre constitué de poésies écrites entre 1900 et 1999.
   - 2.2. Dans chaque corpus, combien de fois apparaissent les mots *rêve* ou *rêves* ?
   - 2.3. Est-ce suffisant pour savoir si les rêves relèvent plus d'un genre littéraire que d'un autre ? Quelles remarques peut-on faire ?

Recherche avancée

- **CQL** (Corpus Query Language) : langage d'expression de requêtes
  - Développé par : **Corpora and Lexicons Group** (University of Stuttgart)
  - Utilisé dans différents outils d'exploration de corpus :
    - Frantext
    - Sketch Engine
    - TXM
    - SARA (BNC)

Recherche avancée

- **CQL** (Corpus Query Language) : langage d'expression de requêtes
  - **Expression CQL** = chaîne de caractères exprimant un **motif linguistique** en fonction de :
    - **Formes graphiques**
      - Mot réservé : **word** (mot)
      - Ex : `[word="bonheur"]` → bonheur
    - **Formes lemmatisées**
      - Mot réservé : **lemma** (lemme)
      - Ex : `[lemma="aimer"]` → aime, aimer, aimera...
    - **Catégories grammaticale**
      - Mot réservé : **pos** (part of speech)
      - Ex : `[pos="VINF"]` → manger, courir, voir...

Recherche avancée

- Liste des catégories grammaticales utilisées (POS) :

CODE	Notion grammaticale
ADJ	adjectif
ADV	adverbe
CC	conjonction de coordination
CS	conjonction de subordination
CLO	clitique objet
CLS	clitique sujet
DET	déterminant
ET	mot étranger
I	interjection
NC	nom commun
NP	nom propre
P+D	préposition + déterminant
PONCT	ponctuation
PRO	pronom
PROREL	pronom relatif
PROWH	pronom interrogatif
P	préposition
V	verbe conjugué
VINF	verbe à l'infinitif
VPP	verbe participe passé
VPR	verbe participe présent
X	mot non traité

Recherche avancée

- **Combiner des requêtes** (opérateurs booléens) :
  - **&** → ET
  - **|** → OU
  - **!** → NON
- **Exemples** :

```
1. [word= "grand" & pos= "NC"]
2. [lemma= "grand" & pos= "ADJ"]
3. [word= "grand" | word= "petit"][word= "sandwich"]
4. [word= "grand|petit"][word= "sandwich"]
5. [lemma="grand" & !(pos="NC")]
6. [lemma="grand" & pos!="NC"]
7. [lemma= "petit"][lemma= "loup"]
```

Recherche avancée

- **Variantes** :

  Exemple : `[word="État"]` → État

  - **%c** : insensible à la casse
    - `[word="État"%c]` → État , ÉTAT , état
  - **%d** : insensible aux diacritiques
    - `[word="État"%d]` → Etat, État
  - **%cd** : insensible à la casse et aux diacritiques :
    - `[word="État"%cd]` → etat, état, Etat, État, ETAT, ÉTAT
  - **%l** : insensible aux expressions régulières (correspondance littérale)
    - `[word="("%l]` → (

- **Listes** : `[word="chien"] &liste("couleurs")`

Recherche avancée

- **Remarques sur les unités lexicales :**
  - "aujourd'hui", "parce que" → 1 unité lexicale
    - `[word="aujourd'hui"]`
  - "l'amitié" → 2 unités lexicales
    - `[word="l'amitié"]`
    - `[word="l'"][word="amitié"]`
  - `[word=" bonheur "]` (espaces entre les guillemets)
  - `[ word = "bonheur" ]` (espaces entre les crochets)
  - `[]` → **Joker** : n'importe quelle unité lexicale.

Astuce de visualisation

- **Tri des résultats par fréquence :**

> Fréquence > Niveau 1 > Position > Pivot > Calculer

Recherche avancée

- **Expressions régulières** (bonus) **:**
  - `[word="libertés?"]` → liberté ou libertés
  - `[word="âgé?e?s?"]` → âg, âgé, âgée, âgées, âges, âge, âgs, âgés
  - `[word="nation.*"]` → nation, nations, national, nationalisme...
  - `[word=".+able"]` → table, semblable...
  - `[word="..."]` = ".{3}" → que, est, les, sur...
  - `[word="\."]` → .
  - `[word="[tsf]able"]` → table, sable, fable
  - `[word="guerre|paix"]` → guerre, paix
  - `[word="(re|ap|sur)prendre"]` → reprendre, apprendre, surprendre
  - `[word="\d"] [word="janvier"]` → 1 janvier, 2 janvier...

Application

Exercices sur Frantext intégral

3. Recherches assistées :
   - 3.1. Combien y a-t-il de *pantalon* de couleur (le mot *pantalon* au singulier suivi d'une couleur) dans le corpus moderne ?
   - 3.2. Combien y a-t-il d'occurrences de « un visage familier » dans le corpus contemporain ?
     - 3.2.1. Faire la même recherche avec des lemmes, quels autres résultats obtient-on ?
     - 3.2.2. Ajouter un joker optionnel (pouvant aller jusqu'à 10 mots) entre le nom et l'adjectif. Quels résultats obtient-on ?

4. Faire des expressions CQL :
   - 4.1. Combien y a-t-il d'occurrences du nom propre *Florence* dans le corpus moderne ?
   - 4.2. Combien y a-t-il d'occurrences du verbe croire (et ses variantes flexionnelles) dans le corpus contemporain ?
   - 4.3. Le nom commun *chouette* (et ses variantes flexionnelles) est-il plus fréquent que l'adjectif *chouette* (et ses variantes flexionnelles) dans Frantext intégral ?
   - 4.4. Dans le corpus contemporain, quels sont les adjectifs qui suivent le nom commun *ordinateur* ?