LETTRES SORBONNE UNIVERSITÉ

2025

Le concordancier Antconc (V4)

Crédits : Y. Dupont, J. Bezançon et G. Lejeune 2025

Objectifs

- Exploiter des Corpus
- Faire des manipulations simples sur un concordancier
- Vérifier des emplois
- Comparer des corpus

Exercice 1 : Installation de Antconc

Téléchargez le concordancier ANTCONC à partir de `https://www.laurenceanthony.net/software/antconc/`. Choisissez la verson adaptée à votre système d'exploitation.

Exercice 2 : Démarrer avec Antconc

Allez regarder sur `https://www.gutenberg.org/` où il y a un grand nombre de livres librement disponibles. Recherchez le livre de Jules Verne 20.000 lieues sous les mers, enregistrez la version au format `txt` (utf-8) dans un dossier corpus.

1. Ouvrez Antconc
2. A partir du Menu FILE, chargez le fichier précédemment téléchargé (sélectionnez QUICK CORPUS)

Nous allons tout d'abord utiliser l'onglet KWIC d'Antconc. Faites quelques recherches avec des mots de votre choix (barre de recherche en bas de la fenêtre). Ensuite, cherchez les mots suivants :

- Nautilus
- capitaine
- Anderson
- Nemo

Vous pouvez voir que les résultats suivent un certain ordre, trouvez lequel en parcourant les résultats de la fenêtre centrale. Cette fenêtre montre les usages d'un mot en contexte. On parle de KWIC (Key Word In Context) que l'on oppose à KWOC (Key Word Out of Context), allez sur Wikipédia pour avoir une description de ces deux concepts. Testez les options du premier menu déroulants de `Sort Options` pour trier selon les mots du contexte gauche ou droit ou une combinaison des deux.

Observez plus attentivement les contextes (gauche et droit) d'apparition de Nemo.

Si nous regardons plus en détails autour de la barre de recherche, nous pouvons voir différentes options. Observez le comportement de l'option CASE avec les mots "Nemo" et "Capitaine".

Exercice 3 : Quelques statistiques

Allez sur l'onglet WORD pour obtenir quelques statistiques sur notre document. Cliquez sur `start` pour faire apparaître la liste de tous les mots du texte par ordre décroissant d'effectif (*frequency*).

Utilisez maintenant la fonction INVERT ORDER pour observer les hapaxs de ce texte (mots qui n'ont qu'une occurrence). Décochez la case pour retrouver l'ordre précédent.

Exercice 4 : Expressions régulières

Naviguez dans le menu SETTINGS, puis GLOBAL SETTINGS et enfin dans la catégorie SEARCHES (fenêtre de gauche). Observez la signification des différents caractères qui y sont présentés. À l'aide de ces caractères, nous pouvons effectuer des recherches portant non pas sur une forme (un mot, un groupe de mots) mais un ensemble de variantes d'un mot ou d'un groupe de mots en jouant sur leurs caractéristiques communes.

À l'aide de ces caractères, recherchez les les mots qui ...

1. ... finissent par "able" ?
2. ... finissent par "ible" ?
3. ... commencent par "anti". Parmi eux, quels sont les mots qui utilisent le préfixe "anti-" dans le sens "contre/à l'opposé" ?
4. ... commencent par "pré" et qui finissent par "ment" ?

Ce que vous utilisez ici se nomme des expressions régulières, allez voir sur WIKIPEDIA la section "Opérateurs de base" de l'article "Expressions régulières" pour vous informer sur ce sujet.

Une expression régulière suit des "règles", on part donc d'un certain nombre d'observations pour essayer d'expliquer à la machine ce que l'on recherche. En effet, rechercher chaque cas à la main peut s'avérer très fastidieux.

On va chercher à récupérer **toutes les dates** mentionnées dans les textes. Toujours avec le même corpus de Jules Verne, on va regarder comment sont formées les noms de dates que l'on est amené à trouver (pensez à garder cochée la case regexp dans ANTCONC, disponible en bas de chaque onglet).

Pour démarrer, on va supposer que toutes les dates comportent une année avec un "1", on cherche donc le caractère `1`. **Si vous n'avez aucun résultat c'est que le corpus est mal tokénisé, (cf. instructions sur la Token Definition au début du TD)**. On a évidemment trop de résultats, on affine donc par étapes :

- En se disant que la plupart des dates sont au 19ème siècle, on cherche donc: `18`
- On a des résultats mais on a encore du **bruit**[1] : 18 août, page 184 ...
- On peut raffiner en disant que l'on recherche des suites de 4 chiffres
- A partir de la page Wikipedia (ou du CM) trouvez comment écrire une expression régulières qui recherche 4 chiffres consécutifs : une liste des chiffres possibles (avec les crochets) et en limitant le nombre d'occurrences à 4 (avec les accolades)

[1] Au sens de bruit documentaire, c'est à dire des résultats "en trop"

- Cherchez maintenant toutes les années précédées du mot “janvier”
- En utilisant les parenthèses et la barre verticale cherchez toutes les années précédeés de janvier ou février.
- génralisez à tous les mois
- Essayez maintenant de détecter toutes les dates complètes de la forme suivante : nombre + mois + année quand elle est présente, par exemple :
  - 14 janvier 1862 ; 16 janvier ; 30 janvier 1805 ...

Utilisez votre expression régulière pour établir une chronologie des évènements dans chaque livre (re-séparez donc le corpus pour travailler texte par texte). Déposez la pour chaque texte sur Moodle avec un évènement par ligne (dans l'ordre de leur apprition dans le texte) . Quelque chose sous la forme :

- Date1 avec contexte de l'évènement 1 (la phrase ou le contexte d'apparition)
- Date2 avec description de l'évènement 2...
- Par exemple (dans 5 semaines en ballon):

14 janvier 1862 : Il y avait une grande affluence d'auditeurs, le 14 janvier 1862, à la séance de la Société royale géographique de Londres, Waterloo place, 3

16 janvier : Le lendemain, dans son numéro du 16 janvier, le Daily Telegraph publiait un article ainsi conçu:

30 janvier 1805 : il {Mungo-Park} repart le 30 janvier 1805 avec son beau-frère Anderson, Scott le dessinateur et une troupe d'ouvriers.

Exercice 5 : Spécificités de corpus

Nous allons ici utiliser AntConc afin de trouver des mots spécifiques à un corpus donné (qui pourraient être de bons mots-clés). Dé-sélectionnez tous les fichiers de la colonne de gauche (menu File puis Clear All Tools And Files).

- Aller dans le menu File puis Open Corpus Manager.
- Sélectionnez `Raw File(s)` et entrez le nom de votre premier corpus.
- Cliquez sur `Add File(s)` et sélectionnez "Vingt milles lieues sous les mers".
- Ensuite, donnez lui le nom "20000_lieues" (grâce au champ Corpus name) et cliquez sur `Create`. Vous avez créé un corpus contenant "Vingt milles lieues sous les mers". Ce sera notre corpus dit "Target", celui sur lequel on effectue notre analyse.
- Faites de même (`Raw File(s)` puis `Add File(s)`) pour créer un nouveau corpus qui contiendra les deux autres ouvrages ("Cinq Semaines en Ballon" et "Voyage au Centre de la Terre"), ce sera notre corpus de référence, nommez le "reference".
- En haut, pour l'option `Corpus Source`, cochez "Corpus Database". Vous devriez voir vos deux corpus créés sous User (List).

- Vous devez indiquer quel corpus sera le corpus Target et quel corpus sera la référence. Double-cliquez sur un corpus après avoir choisi un type de corpus sur la partie droite de la fenêtre.
- Une fois vos corpus assignés comme il se doit, quittez la fenêtre `Corpus Manager` (`Return To Main Window`).

À partir de cela, trouvez les mots clés de "Vingt milles lieues sous les mers" (onglet `Keyword`).

1. Quels sont les 10 mots qui sont spécifiques à "Vingt milles lieues sous les mers" (par opposition aux deux autres textes) ?
2. Que représentent les 5 mots les plus spécifiques ?

Exercice 6 : Clusters/N-grams

Allez sur l'onglet Clusters/N-grams, puis :

1. recherchez les clusters de taille 2 commençant par "capitaine". Remarquez-vous quelque chose de particulier ?
2. Quels sont les différents capitaines mentionnés dans le livre ?
3. comment pouvez-vous configurer l'outil dans le menu Tool Settings pour rendre votre recherche plus efficace ?
4. Utilisez cet outil pour vérifier que l'histoire se déroule en partie sous les eaux (pensez à configurer Cluster Size).
5. Allez sur l'onglet Collocates et trouvez les mots qui sont en collocation avec "capitaine". Y a-t-il une différence avec les clusters ?

Exercice 7 : Trouver des indices pour comprendre des textes

Dans cette partie, nous allons voir comment Antconc permet de trouver des indices dans le texte rapidement pour répondre aux deux questions :

1. le capitaine Nemo n'est pas le héros de l'histoire. Comment pouvez-vous utiliser l'onglet Plot pour vérifier cette assertion ?
2. "Vingt milles lieues sous les mers" est écrit de manière différente des autres textes, quels sont les indices qui peuvent illustrer cela ? (observez les mots outils)