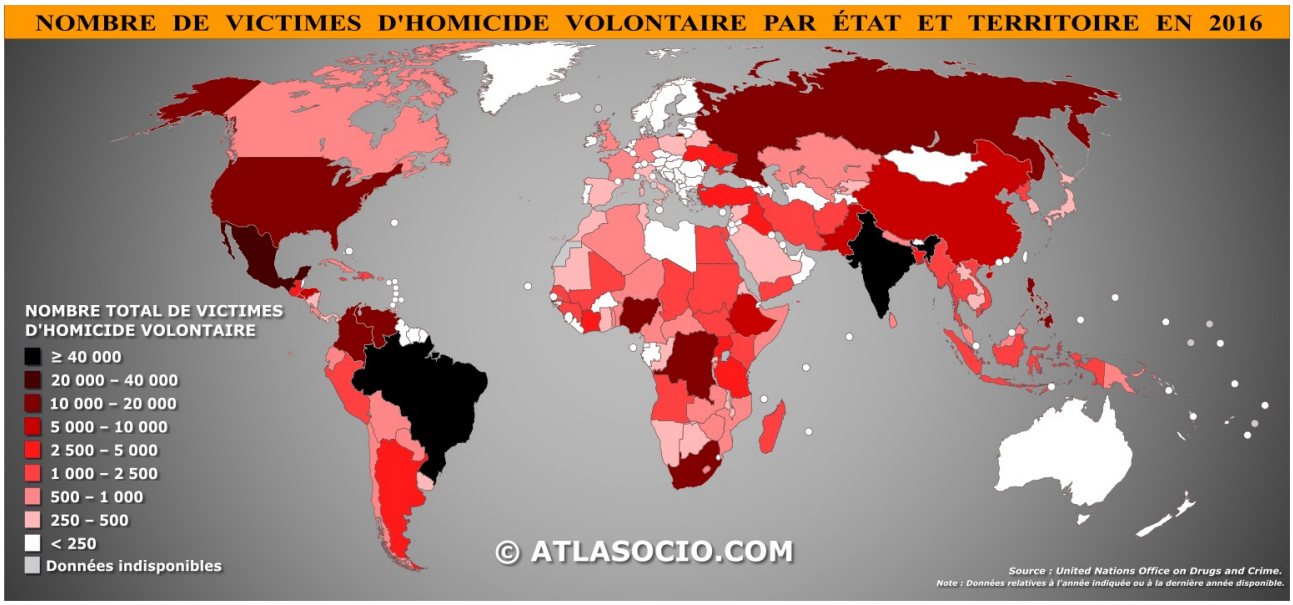
Révision de maths, encore quelques cartes…

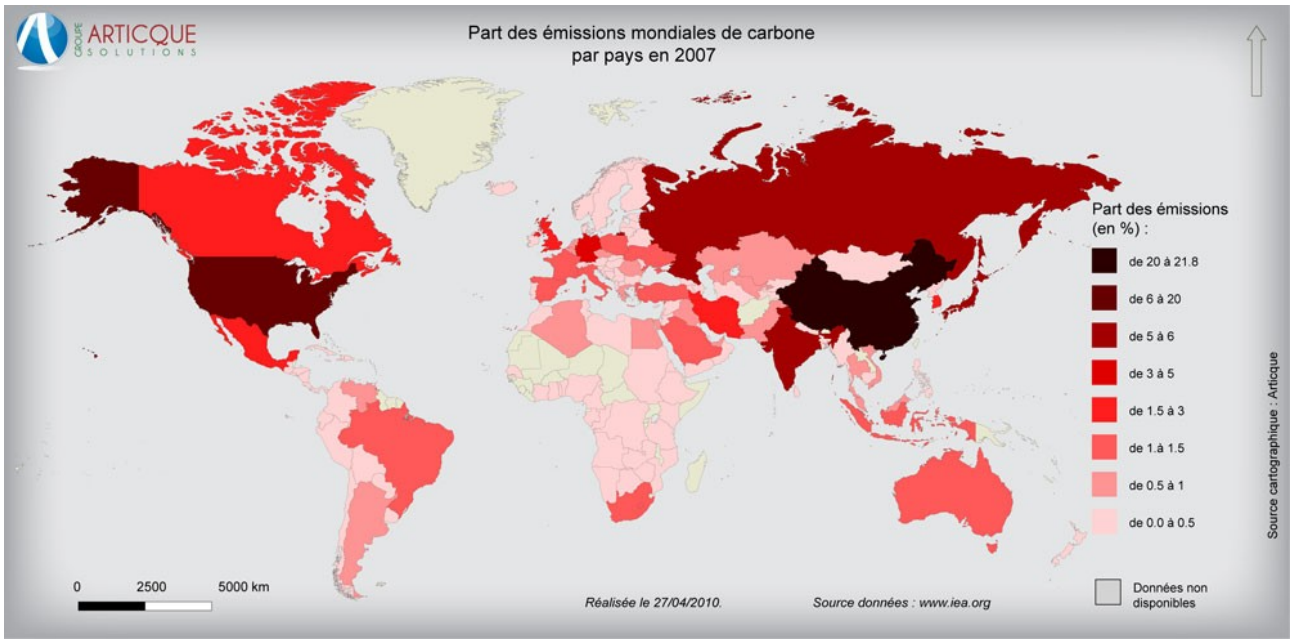
Le Portugal est-il plus sûr que l'Espagne ? L'Espagne est-elle plus sûre que la France ?

L'Ukraine est-elle moins sûre que la France ?



Ce n'est pas parce qu'il y a des % qu'on peut faire n'importe quoi...

Comment est calculé le % utilisé sur la carte ci-dessous ? Pourquoi l'Allemagne a-t-elle le plus haut % d'Europe ? Peut-on dire qu'on allemand émet 10 fois plus de carbone qu'un suisse ?



Anatext, http://phraseotext.univ-grenoble-alpes.fr/anaText

Anatext travaille sur du texte brut. On lui donne du texte simplement par copier/coller.

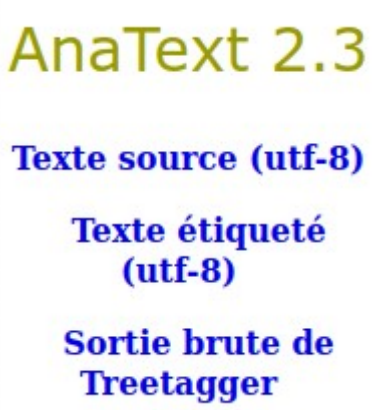
On va travailler sur la pièce de théâtre *Le Mariage inattendu de Chérubin* (Olympe de Gouges, 1786). Comme beaucoup de textes classiques, elles est disponible librement sur Wikisource : https://fr.wikisource.org/wiki/Le_Mariage_inattendu_de_Ch%C3%A9rubin

En quelle langue ce texte est-il écrit ?

À votre avis, quels caractères risquent de poser problème si on utilise la langue « Français » dans Anatext ?

Utilisez un éditeur de texte pour chercher/remplacer les caractères qui vont poser problème (on parle de « prétraiter » ou « nettoyer » le texte).

Copiez/collez le texte nettoyé dans Anatext, et cliquez sur « Analyser le texte ».



← Commencez par consulter le « Texte source (utf-8) ».

Si il y a un problème de codage des caractères (mojibake), vérifiez quel codage de caractères utilise votre navigateur (sur Firefox : Ctrl+i). Corrigez-le (sur Firefox : menu Affichage).

← Consultez ensuite la « Sortie brute de Treetagger ». TreeTagger Même chose pour le codage des caractères. **À quoi correspond chaque ligne et chaque colonne ?**

TreeTagger fait plusieurs choses :

- **tokenisation** : https://en.wikipedia.org/wiki/Lexical_analysis#Tokenization
- **étiquetage morpho-syntaxique** : https://fr.wikipedia.org/wiki/%C3%89tiquetage_morpho-syntaxique
- **lemmatisation** : https://fr.wikipedia.org/wiki/Lemmatisation

Pour cela, il s'appuie sur des **modèles de langue** (*parameter file* dans le jargon de TreeTagger). Combien y en a-t-il pour le français sur le site officiel de TreeTagger ? Comment sont-ils obtenus ?

Lemmes spécifiques

Lemmes spécifiques

Copy CSV Excel PDF Print

Show 10 entries Search:

Rang	Lemme	Fréquence	CorpusRef (par million)	LogLike (spécificité)
1	suif	36	1.165	490.426
2	comte	52	45.66	337.414
3	prussien	22	3.855	214.720
4	manufacturier	7	0.035	131.433
5	boule	23	45.94	112.029
6	officier	25	83.905	96.883

^--- À quoi correspond la colonne « Fréquence » ?

À quoi correspond la colonne « CorpusRef » ? (v--- indice en vert sur la page d'accueil d'AnaText).

Il n'y a pas de colonne pour la fréquence relative des mots dans le texte analysé (*Le Mariage inattendu de Chérubin*). Calculez celle du mot le plus fréquent !

Calcul de spécificité

En linguistique de corpus, le calcul de spécificité sert à comparer deux corpus. En général, on compare (A) un corpus qu'on veut étudier, avec (B) un corpus « de contraste » ou « de référence ». On essaye de trouver les mots qui sont « spécifiques » du corpus (A), c'est à dire des mots dont la fréquence relative est significativement plus élevée dans (A) que dans (B). Ce degré de spécificité est un nombre qu'on appelle « indice de spécificité » ou « score de spécificité ».

Il y a plusieurs manières de calculer ce score. En général, on prend la liste des fréquences relatives des mots de chacun des deux corpus (tableau ci-dessous), et on compare ces fréquences entre elles, grâce à une formule qui nous donne un score. Il existe plusieurs méthodes pour comparer ces fréquences, les plus utilisées en linguistique sont le *Loglikelihood* (souvent abrégé en *Loglike*) et les *spécificités de Lafon*. Elles sont assez complexes, mais il n'est pas indispensable de comprendre en détail comment elles sont calculées pour s'en servir.

	Fréquence relative dans le corpus A	Fréquence relative dans le corpus B
Mot 1		
Mot 2		
Mot 3		
...		
Mot *n*		

Quels sont les 5 lemmes les plus fréquents dans le texte ? (en fréquence absolue et en fréquence relative)

Quels sont les 5 lemmes les plus spécifiques du texte ? Et par rapport à quoi ?

Recherchez toutes les occurrences de l'adjectif pauvre. Combien d'occurrences ?
(si vous en trouvez 10 il y a un problème... on a dit les adjectifs)

Quel nom est le cooccurrent le plus fréquent du mot pauvre ?

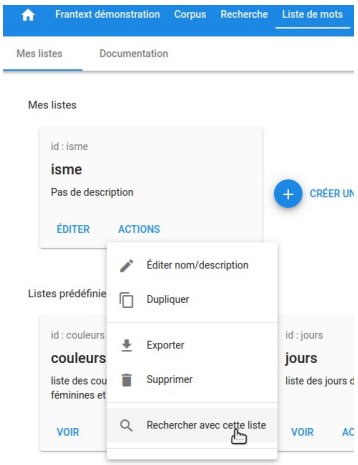
Frantext, https://www.frantext.fr/

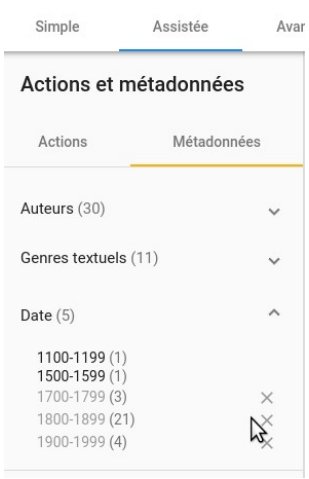
Familiarisation avec le corpus

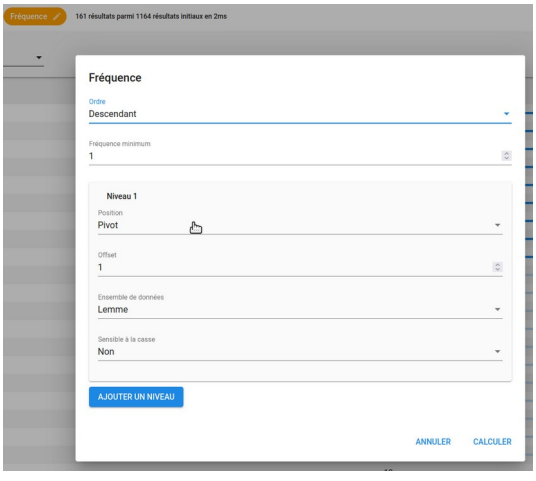
Quelle taille ? Nombre de textes, nombre mots ? Quelles autres informations sont disponibles dans l'onglet Statistiques ?

Listes de mots

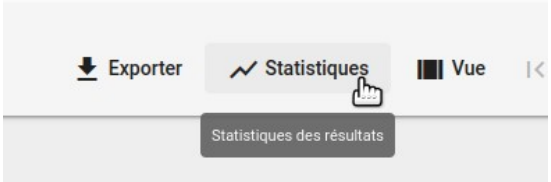
Quels sont les mots en *-isme* (expression régulière *.*isme*) les plus fréquents entre 1701 et 2000 ?







Quel siècle comporte le plus de mots en *-isme* ?



Voisinage de mots

Quels mots s'emploient le plus souvent dans la même phrase que *travail* ? Comparez avec *travaux*.

Quels mots s'emploient souvent exactement après *faire* ?

Parties du discours

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**.

Récapitulatif

	Type de texte	Structuration du corpus	Lemmes	Parties du discours	Dépendances
Anatext					
Frantext					
Google Ngrams					
Google Search					
Google Trends					
ScienQuest					
SketchEngine					
TXM					

Et encore plein d'autres !

- http://explorationdecorpus.corpusecrits.huma-num.fr/

Et plus généralement : le consortium CORLI

- https://groupes.renater.fr/sympa/info/corli
- https://groupes.renater.fr/wiki/corli/index