http://pro.aiakide.net, achille.falaise@univ-paris-diderot.fr, loic.liegeois@univ-paris-diderot.fr

Fréquences absolues / relatives

Il est très important de distinguer :

- La **fréquence absolue** : c'est le résultat brut du comptage.
- **La fréquence relative** : c'est la fréquence absolue divisée par la taille de l'entité. C'est toujours un nombre inférieur à 1 (rappel : 1%=0,01 ; 1‰=0,001 ; 1‱=0,0001 ; 1ppm=0,000001 ; etc.).

Quand les entités sont de taille différente, il ne faut **jamais** comparer les fréquences absolues entre elles, car cela n'a aucun sens ; on peut uniquement comparer les fréquences relatives.

Internet pour les SdL : le *Web* comme corpus

A. Savoir lire une URL

L'adresse d'une page Web (URL) se compose de 4 parties : protocole**://**domaine/chemin**?**paramètres

1. le protocole (pour le Web : http ou https),
2. le **nom de domaine** (qui se lit de droite à gauche, et « commence » donc par un code tel quel .fr, .ca, .com, .org, etc.),
3. un chemin (facultatif),
4. divers paramètres (facultatif).

> Décomposez ces URLs en entourant les 3 ou 4 parties dont elles sont constituées :
>
> http://www.lemonde.fr/voyage/video/2010/10/21/plongee-dans-le-metro-moscovite_1424766_3546.html
>
> http://www.tlfq.ulaval.ca/AXL/francophonie/HIST_FR_s1_Expansion-romaine.htm
>
> https://www.google.com/search?hl=fr&q=corpus

B. Archiver une page Web et remonter le temps : https://web.archive.org

> Cherchez la version la plus ancienne du site www.univ-paris-diderot.fr
>
> Cherchez la page qui décrit le projet de transfert de l'Université Paris Diderot sur le site Paris Rive Gauche.

C. Faire de la linguistique avec Google Search

Avant tout, le Web est-il à proprement parler un *corpus* ?

En français, dit-on « hôtel à Paris », ou bien « hôtel en Paris » ? Recherchez sur *Google Search*. Quelle conclusion en tirez-vous ?

Dans la plupart des moteurs de recherche, on peut rechercher une expression exacte en l'encadrant par des guillemets. Le caractère * peut remplacer n'importe quel mot. Le mode avancé permet de restreindre à une langue ou bien a un site Web précis.

Indice, testez ces pages : http://www.plot-generator.org.uk/story/
http://www.megabambou.com/pmg/index.html

Réessayez en vous limitant aux sites lemonde.fr, ou tripadvisor.fr, pour les expressions « à/en États-Unis », puis « à/en Isère ».

Concrètement, quelle utilisation voyez-vous pour Google Search ? Rechercher des exemples sur le Web ? Sur un site précis ? Récupérer un nombre de résultats ? Déterminer si une expression est attestée ?

Lexicographie avec Google Trends

Google Trends est un outil permettant de visualiser les requêtes effectuées par les utilisateurs de Google Search. Il est intéressant en lexicographie, mais attention, ce n'est pas un outil fait pour cela ! Il ne faut pas perdre de vue qu'il se base sur les *recherches* des utilisateurs, pas sur le contenu des sites Web.

Allez sur le site de Google Trends: http://www.google.fr/trends/explore

- Dans quelles régions françaises utilise-t-on le terme *chocolatine* ? Comparez avec *pain au chocolat*.

- Pour *pain au chocolat*, que se passe-t-il en octobre 2012 et 2016 ? Que pouvez-vous en déduire ?

Concrètement, Google Trends est-il utilisable pour rechercher des régionalismes ? Une évolution linguistique ?

Linguistique diachronique avec Google Ngram Viewer

Google Ngram Viewer est un outil permettant d'effectuer des recherches diachroniques dans Google Books.

Recherchez « dans » en français entre 1600 et 2000. Quelle échelle est-elle utilisée en abscisse ? En ordonnée ? Comparez avec « en ».

Comparez l’évolution de l’emploi des prépositions « à » et « en » avec Martinique, Guadeloupe et Haïti.

Corpus parallèles avec Linguee

Linguee est un moteur de recherche pour des textes multilingues, notamment des sites et des documents d’organisations internationales. Il permet de rechercher une expression (ce qu’un dictionnaire ne peut pas faire) et d’afficher une série d’exemples de traduction.

Comment traduiriez-vous « conseil de classe » en anglais ? Quelle est le degré de qualité des traductions sur lesquelles Linguee se base ? (inutile de connaître l’anglais, il y a d’autres indices dans la page…)

Essayez avec « par conséquent », « en avoir le cœur net ». Quelle traduction retenir ? Comparez « sans doute » et « sans aucun doute ». Comment traduiriez-vous « *raining cats and dogs* » (facile) et « *off the charts* » (plus difficile) en français ?

Les outils à retenir : Google Search (mode avancé), Google Scholar, Web Archive, Google Trends, Google Ngram Viewer, Linguee.

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

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

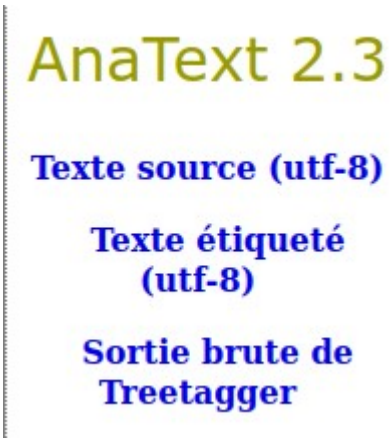
On va travailler sur la pièce de théâtre *Le Mariage inattendu de Chérubin* (Olympe de Gouges, 1786). Comme beaucoup de textes classiques, elles est disponible librement sur Wikisource : https://fr.wikisource.org/wiki/Le_Mariage_inattendu_de_Ch%C3%A9rubin

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 ?

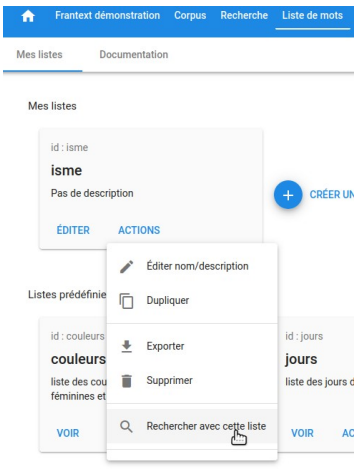
Frantext, https://www.frantext.fr/

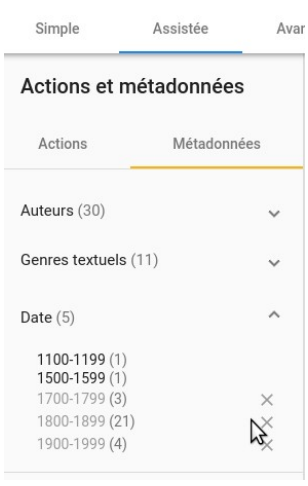
Familiarisation avec le corpus

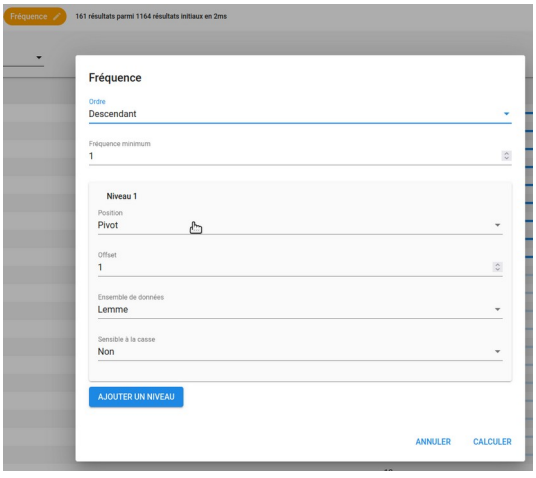
Quelle taille ? Nombre de textes, nombre mots ? Quelles autres informations sont disponibles dans l'onglet Statistiques ?

Listes de mots

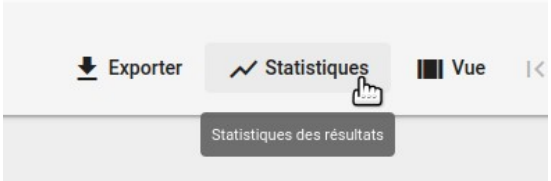
Quels sont les mots en *-isme* (expression régulière *.*isme*) les plus fréquents entre 1701 et 2000 ?







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**.

Plateforme *ScienQuest* − Corpus *Scientext - Écrits scientifiques en français*

La plateforme se situe à l'adresse : http://corpora.aiakide.net/

Familiarisation avec le corpus

À partir des informations affichées sur la page d'accueil, décrivez ce corpus.

Sélectionnez ce corpus, et allez sur l'onglet *Textes*. Comment est organisé ce corpus ?

Dans l'onglet *Recherche*, recherchez la forme *hypothèse*. Une fois la recherche terminée, cliquez sur une occurrence, puis affichez l'arbre syntaxique. Quels types d'annotations comporte ce corpus ? Quelle différence avec *Frantext intégral* et *Frantext catégorisé* ?

Lemmes

Recherchez maintenant le **lemme** *hypothèse*. Quelle différence avec une recherche sur la **forme** *hypothèse* ?

Recherchez maintenant le **lemme** *porter*. Comparez avec *porte*. Quelle différence ? Comment expliqueriez-vous les erreurs pour la recherche avec *porte* ?

Recherchez la locution adverbiale *au passage*. Que pouvez vous dire de sa répartition dans le corpus (Résultats → Statistiques), et sur son emploi dans l'écrit scientifique ?

On utilise normalement assez peu le pronom *je* dans les textes scientifiques. Grâce à une recherche sur ce pronom, et en consultant l'onglet « Résultats → Statistiques », pouvez-vous déterminer dans quels cas il est le plus utilisé ?

On conseille plutôt l'utilisation des pronoms *nous* et *on*. Dans quels genres les retrouve-t-on le plus ?

Expressions régulières

Recherchez les adverbes en *-ment* en début de phrase, sachant que [A-Z] trouve n'importe quel caractère majuscule[1]. Dans quel genre textuel trouve-t-on le plus d'adverbes en *-ment* en début de phrase ?

Recherchez les verbes conjugués au futur à la première personne (terminaison en *-rai* ou *-rons*). Y a-t-il du bruit ?

Syntaxe

Avec quels verbes emploie-t-on préférentiellement le nom *hypothèse* ? Recherchez *verbe + hypothèse*, puis *verbe + n'importe quel mot + hypothèse*.

Essayez maintenant avec une **relation syntaxique** objet direct. Dans quelle mesure les résultats diffèrent-t-ils ?

Quels verbes au futur ont-ils souvent des pronoms personnels introduisant l'auteur (*je, nous, on*) ?

« *On reconnaît un mot à ses fréquentations* » (J. R. Firth, 1957)
En linguistique, une collocation est une cooccurrence privilégiée, une association habituelle d'un mot à un autre au sein d'une phrase, un rapprochement de termes qui, sans être fixe, n'est pas pour autant fortuit, comme : « voix suave », « courir vite », « entraîner des conséquences ». *Wikipédia*

Comparez *jouer + nom* et *verbe + rôle*. L'expression *jouer un rôle* est-elle une collocation ? Idem pour *poser problème*. Comment *problème* se distingue-t-il de *question* ?

Recherche sur le passif

Vous rechercherez avec les relations syntaxiques de votre choix des structures passives où le lemme hypothèse est sujet. Ex : *l'hypothèse est formulée de la façon suivante* .

1 … mais pas les majuscules accentuées. En ne s'en occupe pas dans le cadre du TD.