Frantext intégral

Fréquences

Fréquence relative = Fréquence absolue / taille du corpus

Familiarisation avec le corpus

Allez dans Documentation > Didacticiel > Frantext 1, Présentation

Quel critère de sélection : **genre textuel, période** ?

But de recherche ?

Quelle taille ? Nombre de textes, nombre mots ?

Premier contact avec l'outil

Accédez au corpus intégral. Comme beaucoup d'outils, l'interface de Frantext s'organise en plusieurs étapes : (1) **sélection du corpus de travail**, (2) **recherche dans les textes**, (3) **calcul de fréquence**. Contrairement à d'autres, Frantext utilise aussi beaucoup des **listes de mots**.

Sélectionnez tout le corpus, puis effectuez des recherches dans les textes pour déterminer précisément ce qu'est un « mot » pour Frantext.

Les signes de ponctuation sont-ils des mots ?

La recherche est-elle sensible à la **casse** ?

« aujourd'hui », « parce que », « pomme de terre »… combien de mots ?

Les majuscules comportent-elles des **diacritiques** ?

Le corpus utilise-t-il les **ligatures** ?

Loi de Zipf

La loi de Zipf est une observation empirique concernant la fréquence des mots dans un texte.

> Dans les années 30, un scientifique de l'université de Harvard, G. K. Zipf, a montré qu'en classant les mots d'un texte par fréquence décroissante, alors, on observe que la fréquence d'utilisation d'un mot est inversement proportionnel à son rang. La loi de Zipf stipule que la fréquence du second mot le plus fréquent est la moitié de celle du premier, la fréquence du troisième mot le plus fréquent, son tiers, etc. Cette loi peut s'exprimer de la manière suivante : Fréquence d'un mot de rang N = (Fréquence du mot de rang 1) / N
>
> D'après : Emmanuel Giguet, *La loi de Zipf*, https://giguete.users.greyc.fr/java/zipf.html, consulté le 2017-03-06.

D'après la **fréquence des mots du corpus de travail** (menu Calculs de fréquences), l'œuvre de Zola vérifie-t-elle la loi de Zipf ?

Mots les plus fréquents

Recherchez maintenant les mots les plus fréquents chez Balzac, et comparez avec Zola. Que pouvez-vous (très grossièrement !) en déduire sur leurs différences de style ? Comparez maintenant ces fréquences avec les mots les plus fréquents du roman *la Disparition* de Perec.

Expressions régulières

N’importe quel mot : **^.*$**

N’importe quel mot qui contient *vert* : vert

Le mot vert exactement (il commence *et* finit par *vert*) : ^vert**$**

N’importe quel mot se terminant par -*âtre* : âtre**$**

N’importe quel mot commençant par *anti-* : ^anti

Le mot *vert*, éventuellement suivi d’un *s* : ^verts**?**$

Le mot *vert*, éventuellement fléchi : ^vert**(**s|es**)?$** ou bien ^verte**?**s**?$** ou bien ^(vert|verte|vertes)$

… et aussi *verdeur(s)* : ^ver**(**t|deur**)**e**?**s**?$** ou bien ^(vert|verte|vertes|verdeur|verdeurs|verdeures)$

Les parenthèses et les barres (...|...) fonctionnent en mode « expression régulière », mais aussi dans les autres modes.

Les mots en -isme

Créez la liste des mots en -*isme(s)*.

Quels sont les mots en -*isme* les plus fréquents entre 1701 et 2000 ?

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

Quel sont les mots en -*isme* les plus fréquents pour chaque siècle ? (faites une recherche par siècle)

Mots fléchis

Recherchez **les flexions du verbe** *faire*, puis *taire*. Quelle est la limite de ce mode de recherche ? Comment paraît fonctionner la recherche de mots fléchis ?

Listes de mots

Pour en avoir le cœur net, créez la liste des flexions de *faire* et *taire*, et analysez cette liste. Décelez-vous un risque d’ambiguïté ?

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* ?

Frantext intégral

Fréquences

Fréquence relative = Fréquence absolue / taille du corpus

Familiarisation avec le corpus

Allez dans Documentation > Didacticiel > Frantext 1, Présentation

Quel critère de sélection : **genre textuel, période** ?

But de recherche ?

Quelle taille ? Nombre de textes, nombre mots ?

Premier contact avec l'outil

Accédez au corpus intégral. Comme beaucoup d'outils, l'interface de Frantext s'organise en plusieurs étapes : (1) **sélection du corpus de travail**, (2) **recherche dans les textes**, (3) **calcul de fréquence**. Contrairement à d'autres, Frantext utilise aussi beaucoup des **listes de mots**.

Sélectionnez tout le corpus, puis effectuez des recherches dans les textes pour déterminer précisément ce qu'est un « mot » pour Frantext.

Les signes de ponctuation sont-ils des mots ?

La recherche est-elle sensible à la **casse** ?

« aujourd'hui », « parce que », « pomme de terre »… combien de mots ?

Les majuscules comportent-elles des **diacritiques** ?

Le corpus utilise-t-il les **ligatures** ?

Loi de Zipf

La loi de Zipf est une observation empirique concernant la fréquence des mots dans un texte.

> Dans les années 30, un scientifique de l'université de Harvard, G. K. Zipf, a montré qu'en classant les mots d'un texte par fréquence décroissante, alors, on observe que la fréquence d'utilisation d'un mot est inversement proportionnel à son rang. La loi de Zipf stipule que la fréquence du second mot le plus fréquent est la moitié de celle du premier, la fréquence du troisième mot le plus fréquent, son tiers, etc. Cette loi peut s'exprimer de la manière suivante : Fréquence d'un mot de rang N = (Fréquence du mot de rang 1) / N
>
> D'après : Emmanuel Giguet, *La loi de Zipf*, https://giguete.users.greyc.fr/java/zipf.html, consulté le 2017-03-06.

D'après la **fréquence des mots du corpus de travail** (menu Calculs de fréquences), l'œuvre de Zola vérifie-t-elle la loi de Zipf ?

Mots les plus fréquents

Recherchez maintenant les mots les plus fréquents chez Balzac, et comparez avec Zola. Que pouvez-vous (très grossièrement !) en déduire sur leurs différences de style ? Comparez maintenant ces fréquences avec les mots les plus fréquents du roman *la Disparition* de Perec.

Expressions régulières

N’importe quel mot : **^.*$**

N’importe quel mot qui contient *vert* : vert

Le mot vert exactement (il commence *et* finit par *vert*) : ^vert**$**

N’importe quel mot se terminant par -*âtre* : âtre**$**

N’importe quel mot commençant par *anti-* : ^anti

Le mot *vert*, éventuellement suivi d’un *s* : ^verts**?**$

Le mot *vert*, éventuellement fléchi : ^vert**(**s|es**)?$** ou bien ^verte**?**s**?$** ou bien ^(vert|verte|vertes)$

… et aussi *verdeur(s)* : ^ver**(**t|deur**)**e**?**s**?$** ou bien ^(vert|verte|vertes|verdeur|verdeurs|verdeures)$

Les parenthèses et les barres (...|...) fonctionnent en mode « expression régulière », mais aussi dans les autres modes.

Les mots en -isme

Créez la liste des mots en -*isme(s)*.

Quels sont les mots en -*isme* les plus fréquents entre 1701 et 2000 ?

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

Quel sont les mots en -*isme* les plus fréquents pour chaque siècle ? (faites une recherche par siècle)

Mots fléchis

Recherchez **les flexions du verbe** *faire*, puis *taire*. Quelle est la limite de ce mode de recherche ? Comment paraît fonctionner la recherche de mots fléchis ?

Listes de mots

Pour en avoir le cœur net, créez la liste des flexions de *faire* et *taire*, et analysez cette liste. Décelez-vous un risque d’ambiguïté ?

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* ?