



## « Humanités numériques »

Vendredi 27 octobre 2017  
Censier 221

## « Frantext »

David Moucaud  
Langue & Littérature françaises et latines

## « Humanités numériques »

- I. Intro « humanités numériques » (M)
- II. Généralités Datamasse (K)
- III. Recherche d'informations et Expressions régulières (D)
- IV. N-gramme, requêtes en google n-gramme (M)
- V. Gromoteur (K)

## **VI. Frantext**

- VII. Ressources lexicales (L)
- VIII. Encodages (D)
- IX. Formats de fichiers (M)
- X. Mots, fréquences, statistiques (K)
- XI. Spécificités, Gromoteur (K)
- XII. Examen (D)



# Frantext

*Caiss'sèkça ?!*

# **I. Un grand corpus textuel**

## « Frantext »

Qu'est-ce donc que Frantext ?

Une base de données textuelles adossée à un laboratoire CNRS

Le « fonds » des dictionnaires universitaires TLFi et DMF

Une bonne raison de se mettre (enfin) aux expressions régulières

## « Frantext »

... le TLFi ?

ou « Trésor » de la langue française,

(du latin *thesaurus* > thésauriser, Trésor Public...)

un dictionnaire « total » et totalement informatisé

... une base de données indexée & son moteur de recherche !

... hébergé par le CNRTL/Ortolang comme un service public

[www.cnrtl.fr/lexicographie/trésor](http://www.cnrtl.fr/lexicographie/trésor)

110 %



Rechercher

url explicite  
(pensez au  
« coût du  
clic » !)

CNRTL



Ortolang

Outils et Ressources pour un Traitement Optimisé de la LANGue

Centre National de Ressources Textuelles et Lexicales

■ Accueil

■ Portail lexical

■ Corpus

■ Lexiques

■ Dictionnaires

■ Métalexicographie

■ Outils

Morphologie

Lexicographie

Etymologie

Synonymie

Antonymie

Proxémie

Concordance



TLFi

Académie  
9ème éditionAcadémie  
8ème éditionAcadémie  
4ème éditionBDLP  
FrancophonieBHVF  
attestationsDMF  
(1330 - 1500)

■ Entrez une forme

trésor

options d'affichage

catégorie : toutes

Chercher

■ TRÉSOR, subst. masc.

A. -

1. Ensemble de choses de valeur (or, argent, objets précieux, piergeries, titres, etc.) acc  
soigneusement cachés. Synon. *bien(s)* (v. *bien<sup>3</sup>II*), *fortune*, *ressources*, *richesses*, *magot<sup>2</sup>* (f  
immense, *inestimable*; *trésor en pièces d'or*, *en pierres précieuses*; *les trésors de Crésus*, *en  
enfoncer*, *enterrer*, *garder un trésor*; *découvrir*, *détrerrer*, *exhumier*, *mettre à jour un trésor*; *carte*, *pla  
au trésor*; *chercheur de trésors*. Il n'y avait dans Saumur personne qui ne fût persuadé que Mons  
trésor particulier, *une cachette pleine de louis*, et ne se donnât nuitamment les ineffables jouiss  
vue d'une grande masse d'or (BALZAC, E. Grandet, 1834, p. 14). On dirait le *trésor d'un raj  
poignards à manches de nacre, vêtements couturés de saphirs, aigrettes d'émeraude sur les  
turquoises* (TAINÉ, Notes Paris, 1867, p. 332).

- [Comme terme de compar., d'évaluation, à valeur de superl.] *Tous les trésors du monde*. [La v  
d'un contentement intérieur plus précieux mille fois que tous les trésors de l'univers (J. DE MAISTRE  
t. 1, 1821, p. 218). Le pauvre diable aurait donné volontiers tous les trésors de la terre pour rem  
de la langue (THARAUD, Dingley, 1906, p. 98).

- En partic.

♦ DR. La propriété d'un trésor appartient à celui qui le trouve dans son propre fonds: si le trés  
fonds d'autrui, il appartient pour moitié à celui qui l'a découvert, et pour l'autre moitié au propriéta  
est toute chose cachée ou enfouie sur laquelle personne ne peut justifier sa propriété, et qui est

Général

Recherche

Vie privée et sécurité

Compte Firefox



## Moteurs de recherche accessibles en un clic

Selectionnez les moteurs de recherche alternatifs qui apparaissent sous la barre de recherche lorsque vous commencez à saisir un mot-clé.

### Moteur de recherche

- Qwant
- Google
- DuckDuckGo
- Portail Lexical - CNRTL
- Wikipédia (fr)

### Mot-clé

[Restaurer les moteurs de recherche par défaut](#)

[Découvrir d'autres moteurs de recherche](#)

Rechercher

Recherche Google

Recherche Portail Lexic



Paramètres de

## « Frantext »

### Qu'est-ce donc ? (site Frantext)

- ... près de **5 116 références** (décembre 2016)
- ... la seule [base] à proposer des recherches sur des textes qui vont **de 950 à nos jours**
- ... avec un fonds contemporain particulièrement riche  
(1 026 textes sont postérieurs à 1950)

Deux bases générales:

- ***Frantext intégral*** totalité des textes
- ***Frantext catégorisé*** 1 200 textes étiquetés grammaticalement

Des bases spécialisées :

- ***Frantext Moyen Français*** 219 textes (1330-1500)
- ***Frantext CTLF*** Corpus de Textes Linguistiques Fondamentaux

## « Frantext »

### Ce que n'est pas/pas encore/pas directement Frantext

Une collection d'*ebooks* récupérables (le droit d'auteur s'applique)

Un répertoire *complet* de la littérature française

Un outil de *textométrie*

Un flim sur le *cyclisme*

## « Frantext »

***Frantext, le TLF, le DMF sont des bases de données « XML »***

XML est langage informatique assez sommaire, qui ne sert pas à programmer, mais plutôt

- à **contenir** (structure de données : des "balises" encapsulent la donnée selon un standard)
- à **décrire** (annotation des données : "métadonnées" qui enrichissent la donnée)
- et à **afficher** des **données** (il est "orienté [vers les] données" = "data-driven")

## « Frantext »

***Frantext, le TLF, le DMF sont des bases de données « XML »***

### **Contenir, décrire et enrichir des données**

|                     |  |
|---------------------|--|
| donnée              | Le petit chat est mort.  |
| donnée annotée      | <citation>Le petit chat est mort.</citation>   |
| donnée enrichie (1) | <citation auteur="Molière" année="1694">Le petit chat est mort</citation>  |
| donnée enrichie (2) | <cit aut="Molière" année="1694" src="l'école des femmes"><br><w lemme="le" type="déterminant">Le</w><br><w lemme="petit" type="adjectif">petit</w><br><w lemme="chat" type="substantif">chat</w><br><w lemme="être" type="verbe">est</w><br><w lemme="mourir" type="verbe">mort</w><br><w type="ponctuation">.</w><br></cit> |

## « Frantext »

***Frantext, le TLF, le DMF sont des bases de données « XML »***

Ce **standard** est, pour ainsi dire, à la base du Web sous sa forme « hypertextuelle » : le langage html en est une version particulière.

Comme standard de description et de structuration de données complexes, il est utilisé :

- pour l'archivage (on peut « récupérer » sous ce format tous les SMS d'un téléphone ou les messages présents sur Gmail ; les administrations en utilisent une version spécifique, le XML-EAD)
- pour et l'affichage de textes à variantes (manuscrits médiévaux : XML-TEI)
- pour l'annotation linguistique (comme dans Frantext ou dans de nombreux outils textométriques)
- sous la forme du xml-html pour l'essentiel de la partie visible d'internet...

## « Frantext »

« - Où ?

- *Y clique à gauche, y clique tout droit... »*

- La plateforme Ortolang (qui donne accès au TLF) permet d'interroger la base à partir d'un moteur de recherche extrêmement simplifié.



■ Entrez une forme

trésor

Chercher

■ TRÉSOR

Concordances de "trésor" - Résultats de 0 à 30 sur 889



lageait encor. Une femme modeste est un rare  
 mbre qui règle les plateaux. 5e LIVRE (i) Le  
 nous comptions bien continuer de grossir ce  
 isez... \*Cyrano " ma chère, ma chérie, " mon  
 me de l'orthographe : si l'on écrit rapsode,  
 s : le latin a toujours été la réserve et le  
 uré ; on l'entend dans mol, dans seuil, dans  
 e sais qu'il y a dans la chapelle un immense  
 affaires, tuer \*Bougrelas et nous emparer du  
 ale de \*Varsovie \*Mère \*Ubu. -où donc est ce

[trésor](#) ; elle obéit toujours et jamais n'importe  
[trésor](#) du verger et le jardin en fête, les fleurs d'  
[trésor](#) . CHAPITRE LXXIII Aux premiers jours d'oc  
[trésor](#) ... " \*Roxane d'une voix... \*Cyrano " mon am  
[trésor](#) , trône, il n'y a aucun motif raisonnable d'é  
[trésor](#) où il a puisé les ressources qu'il n'osait p  
[trésor](#) , dans impair, dans nef, dans jamais, dans dé  
[trésor](#) , nous le distribuerons. \*Père \*Ubu. -misérab  
[trésor](#) . ACTE IV, SCÈNE I la crypte des anciens roi  
[trésor](#) ? Aucune dalle ne sonne creux. J'ai pourtant

## « Frantext »

« - Où ?

- *Y clique à gauche, y clique tout droit... »*

- La plateforme Ortolang (qui donne accès au TLF) permet d'interroger la base à partir d'un moteur de recherche extrêmement simplifié.
- <http://www.frantext.fr>  
*ou mieux encore :*
- <http://www.frantext.fr.ezproxy.univ-paris3.fr/>

# « Frantext »



**atiff**  
ANALYSE ET TRAITEMENT  
INFORMATIQUE  
DE LA LANGUE FRANÇAISE

**Accueil**

**Corpus de travail**

**Recherche dans les textes**

**Calculs de fréquence**

**Étude de voisinage**

**Listes de mots**

**Grammaires**

**Administration**

<http://www.frantext.fr>

 

## Base textuelle FRANTEEXT



### Bienvenue

Accueil Documentation Dysfonctionnement Contact

Bienvenue Quoi de neuf ? Nous citer

**FRANTEEXT Intégral**

**5 118 références, 297 586 780 mots, du Xe au XXIe siècle.**

**Définition du corpus de travail** : sélectionner les textes selon mes critères.

**Mes corpus**

**Sélectionner tous les textes**

**Nouveaux corpus**

Agrégation 2018 Textes au programme de l'Agrégation 2017  
12 textes, 824 931 mots.

## « Frantext »

### Pour quoi faire ?

- Étudier un auteur/une œuvre/une période ou un mot/une tournure  
(le premier pas vers la textométrie)
- Parcourir la data-masse « littéraire » francophone

## « Frantext »

### Comment ?

- Recherche sur un « mot », une expression, un « motif »

Frantext est une ressource lexicale, appuyée sur le plus grand dictionnaire du français et la plus grande collection de textes.

Il n'embarque cependant pas, pour l'instant, d'outils textométriques avancés permettant d'y faire des recherches « non pilotées », à l'aveugle (du type : « qu'est-ce qui est saillant ? ») = Pour schématiser, on n'y fait pas de « nuage de mots »

## « Frantext »

### Quelle différence avec *Google Books* et son Ngram-viewer ?

De l'extérieur, **le principe** semble être le même : une masse de textes, brassée par un moteur de recherche.

Mais **la précision** est accrue :

- Il s'agit d'un corpus **stabilisé**, vérifié et sans parasites :
  - La « qualité » de la numérisation, surtout pour les textes avant 1850, est très nettement supérieure
    - (la qualité des éditions également)
- Des **métadonnées** linguistiques enrichissent les textes
  - chaîne de caractères ~~~~> marques de paragraphes, découpage en phrases et surtout **lemmatisation** des mots
    - (... *puisque c'est lié au Dictionnaire !*)

## « Frantext »

### Une base limitée mais considérable

Comparez ces quelques chiffres :

**TLF**

**270 000 lemmes (entrées)**

Le Grand Robert

100 000

**Frantext**

**297 586 780 mots**

**(dont 13 millions d'ancien/moyen fr.)**

British National Corpus

100 000 000 mots

(le corpus de référence pour l'anglais contemporain : une base synchronique)

## **II. Votre nouvel ami :**

**un *bac à sable*  
littéraire & linguistique**

# Un bac à sable littéraire & linguistique

En cliquant sur le mot dans le concordancier minimaliste du CNRTL, on obtient un contexte plus large et la source de l'occurrence.

The screenshot shows a web browser window displaying a CNRTL search result. The URL in the address bar is [www.cnrtl.fr/utilites/SHOR?name=K276.xml&offset=56750&id=1&len=6](http://www.cnrtl.fr/utilites/SHOR?name=K276.xml&offset=56750&id=1&len=6). The page content includes a header with a red square icon and the text "Bibliographie". Below it, a purple circle highlights a list of bibliographic details: Titre Iphigénie, Auteur Jean MORÉAS, Année 1904, and Edition Paris : Mercure de France, 1921. To the right, a purple box contains the text "EXEMPLE : CNRTL (concordancier simplifié)". At the bottom, another purple circle highlights a section titled "Concordance" containing a French text excerpt. The word "trésor" is highlighted in red in the text. At the very bottom, there is a blue "Fermer" button.

① www.cnrtl.fr/utilites/SHOR?name=K276.xml&offset=56750&id=1&len=6 130 % ABP

**Bibliographie**

Titre Iphigénie  
Auteur Jean MORÉAS  
Année 1904  
Edition Paris : Mercure de France, 1921.

**EXEMPLE : CNRTL (concordancier simplifié)**

**Concordance**

te gagner mon amour. T'ai-je depuis ce temps donné sujet de plainte ? Content dans ta maison et la quittant sans crainte, près de moi, ton ennui se soulageait encor. Une femme modeste est un rare **trésor** ; elle obéit toujours et jamais n'importe ; mais la méchante femme est chose plus commune. Ah ! Pour tout cet excès de zèle et de douceur, pour tous mes tendres soins, tu me perces le coeur, et je

[Fermer](#)

## Un bac à sable littéraire & linguistique

Mais c'est très limité.

Les informations ne sont pas immédiatement visibles, la mise en forme (ici, des vers) est supprimée...

Retenez qu'il existe

- d'excellents outils pour construire un concordancier à partir de fichiers textuels.
- de précieux concordanciers appuyés sur une base textuelle comme Frantext pour d'autres langues :

Exemples d'une base « littéraire » très riche pour le latin :

<http://latin.packhum.org/> (le moteur est très limité)

et de la base « non littéraire » (et synchronique) de référence pour l'anglais :

<http://www.natcorp.ox.ac.uk/> (interrogeable en CQL)

aeger

282 instances

[Liv.AUC.35.14.1.1](#)[Quint.DeclMaior.5.t.1](#)[Serv.A.4.35.5](#)[Serv.A.1.208.5](#)[Ov.Rem.228](#)[Serv.A.9.811.2](#)[Prob.frg.30.1](#)[SenRhet.ConExc.3.9.pr.1](#)[Serv.A.9.811.4](#)[Cels.Med.3.12.2.2](#)[SenRhet.Con.2.4.4.17](#)[Curt.Alex.7.7.5.3](#)[Quint.DeclMaior.5.18.12](#)[SerSamm.Med.22.405](#)[Cels.Med.2.16.1.2](#)[Diff.525.29](#)

## EXAMPLE (latin.packhum.org)

[Concordance](#)

**aeger** Pergami substitit; Uillius cum Pisidiae bello occupatum

**Aeger** redemptus Liberi parentes in egestate aut alant aut

stultitiam iudicet in iuventa aetate nuptiis abstinere. **'aeger'** enim animo dicitur, 'aegrotus' corpore.

post ait 'premit altum corde dolorem'. aeger **'aeger'** est et tristis et male valens, aegrotus autem sive dolenda feres. Saepe bibi sucos, quamvis invitus, amaros **Aeger**, et oranti mensa negata mihi. Ut corpus redimas, ferru quavit aeger anhelitus Probus ait "commodius hic est **'aeger'**, quam in quinto 'vastos quavit aeger anhelitus artus': nihil astra praeter Videl et undas. commodius hic est **'aeger'** quam in quinto 'uastos quavit aeger anhelitus artus':

Crux Servi Venenum Domino Negantis **Aeger** dominus petit a servo, ut sibi venenum daret; non dedi

sta mutandi". quidam 'acer' legunt: et volunt in quinto **'aeger'** aptius dictum de sene, hic de iuvene 'acer' melius

aliud fieri potest, quam ut primis diebus bene abstineatur **aeger**, deinde sub decessu febris eius, quae grauissima est,

et, quo magis concupiscam, habui. Misit ad me adfectus, **aeger**. Non ibo? Mihi crede, aliter tu audis de coherede.

gesturus, cum in conspectu eius obequitaret hostis, adhuc **aeger** ex vulnere, praecipue voce deficiens, quam et modicus

gratias quin immo fortunae, gratias ago, quod adhuc **aeger** sentit, intellegit. alioquin cadaver acceperam et pretia

ex hederae tornantur pocula lignis: hinc trahet adsuetos **aeger** quoscunque liquores. aut uiridis coctorum holerum

uero duo genera sunt, alterum ubi nihil adsumit **aeger**, alterum ubi non nisi quod oportet. Initia morborum p

iacitur, ut levetur onere navis. Aegrum et aegrotum. **aeger** animo, aegrotus corpore. Altitudinem et alimentum et

# Un bac à sable littéraire & linguistique

La base Frantext permet bien plus grâce à :

➤ une catégorisation des textes (auteurs, dates, genres)

➤ la lemmatisation des mots indexés

(et donc la flexion des mots recherchés)

rappel : le **lemme** est la forme standardisée correspondant à une entrée de dictionnaire : /aimer/ pour *aimerai, aimassiez, aimant, aimée...*

➤ de puissants opérateurs de recherche : jokers, troncatures, (pseudo-) « expressions rationnelles »

[Frantext 2, en ligne dans quelques semaines, permettra aussi les requêtes CQL – voir cours « *Recherche d'information* »]

# Un bac à sable littéraire & linguistique

Annexe : à quoi ressemble un texte « lemmatisé » ?

The screenshot shows a web-based lemmatization tool interface. At the top, there are tabs for 'FRANTEXT intégral' and 'Le lemmatiseur du pauvre'. Below the tabs, a back arrow, a link to 'obvil.lip6.fr/alix/lem.jsp', and a '+' button are visible. The main content area displays the poem 'La Beauté' followed by its lemmatized version and a detailed morphological analysis table.

La Beauté

Je suis belle, ô mortels! comme un rêve de pierre,  
Et mon sein, où chacun s'est meurtri tour à tour,  
Est fait pour inspirer au poète un amour  
Éternel et muet ainsi que la matière.

Je trône dans l'azur comme un sphinx incompris;  
Tlunis un cœur de poix à la blancheur des cyprès.

**Envoyer**

| Graphie | Forme   | Catégorie  | Lemme  | Index |
|---------|---------|------------|--------|-------|
| La      | la      | DETart     | le     | 0–2   |
| Beauté  | beauté  | SUB        | beauté | 3–9   |
| Je      | je      | PROpers    | je     | 11–13 |
| suis    | suis    | VERBaux    | être   | 14–18 |
| belle   | belle   | ADJ        | beau   | 19–24 |
| ,       | ,       | PUNcl      |        | 24–25 |
| ô       | ô       | EXCL       | ô      | 26–27 |
| mortels | mortels | SUB        | mortel | 28–35 |
| !       | !       | PUNsent    |        | 35–36 |
| comme   | comme   | CONJsubord | comme  | 37–42 |
| un      | un      | DETart     | un     | 43–45 |
| rêve    | rêve    | SUB        | rêve   | 46–50 |
| de      | de      | PREP       | de     | 51–53 |
| pierre  | pierre  | SUB        | pierre | 54–60 |
| ,       | ,       | PUNcl      |        | 60–61 |
| Et      | et      | CONJcoord  | et     | 62–64 |

fichier tabulaire  
(*ligne à ligne, ~ tableau*)

*associant à chaque occurrence (forme) ses caractéristiques linguistiques*

## **III. Exploration**

### **(back to the *regex*)**

# Exploration

## Protocole Frantext

- 1° délimiter un corpus / une tranche [borner sa recherche]
- 2° rechercher {l'expression X} [formuler sa recherche]
- 3° parcourir les résultats
  - ou
  - 2° analyser des fréquences absolues et relatives

# Exploration

*différents menus  
pour des accès  
légèrement  
différents*



The screenshot shows the ATILF logo at the top left, followed by the text "ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE". Below this is a vertical menu list enclosed in a white border:

- Accueil**
- Corpus de travail**
- Recherche dans les textes**
- Calculs de fréquence**
- Étude de voisinage**
- Listes de mots**
- Grammaires**
- Administration**

Below the menu is the URL <http://www.frantext.fr>. At the bottom are logos for CNRS and Université de Lorraine.

# Exploration : définir un corpus

## Corpus de travail

[Formulaire](#)[Multicritères](#)[Auteurs](#)[Date](#)[Genre littéraire](#)[Corpus](#)[Mes corpus](#)

### ■ Recherche dans un élément bibliographique

[dans l'auteur](#)[dans le titre](#)[dans le genre littéraire](#)[dans la date](#)[cote Frantext](#)[\\*](#)

### ■ Corpus de travail

**Nombre de textes :** 5118

**Nombre de mots :** 297 586 780

**Sélectionner tous les textes**

**Visualiser**

**Stats**

**Vider**

### Options

- sensible à la casse
- sensible aux diacritiques
- sous-chaîne
- bibliographie détaillée

**Astuce : les expressions de date**

# Exploration : définir un corpus

## ■ Expressions de date valides

|             |  |
|-------------|--|
| 1789        | une date exacte                                    |
| 1700-1800   | tous les textes entre 1700 et 1800 bornes incluses |
| 1700-       | tous les textes à partir de 1700, borne incluse    |
| $\geq 1700$ | 2 syntaxes possibles                               |
| <1900       | tous les textes jusqu'à 1900, borne exclue         |
| -1900       | tous les textes jusqu'à 1900, borne incluse        |
| $\leq 1900$ | 2 syntaxes possibles                               |
| >1900       | tous les textes après 1900, borne exclue           |

# Exploration : définir un corpus

## Corpus de travail

Formulaire Multicritères Auteurs Date Genre littéraire Corpus Mes corpus

### ■ 15 textes répondent au critère de recherche

Tri par cote Tri par date inverse Tri par auteur Tri par titre

Sélectionner tous Ne sélectionner aucun Inverser la sélection

Ajouter les textes sélectionnés au corpus de travail (attendre la fin de l'affichage des données)

|   |        |  |      |                                     |
|---|--------|--|------|-------------------------------------|
| 1 | ► S004 | LA FONTAINE (Jean de) ♂<br><i>Fables : Livres 1 à 6</i><br>poésie                      | 1668 | <input checked="" type="checkbox"/> |
| 2 | ► Q949 | LA FONTAINE (Jean de) ♂<br><i>Adonis</i><br>poésie                                     | 1671 | <input checked="" type="checkbox"/> |
| 3 | ► Q963 | LA FONTAINE (Jean de) ♂<br><i>Le Songe de Vaux</i><br>poésie                           | 1671 | <input checked="" type="checkbox"/> |
| 4 | ► Q964 | LA FONTAINE (Jean de) ♂<br><i>Recueil de poésies chrétiennes et diverses</i><br>poésie | 1671 | <input checked="" type="checkbox"/> |
| 5 | ► Q966 | LA FONTAINE (Jean de) ♂<br><i>Fables nouvelles et autres poésies</i>                   | 1671 | <input checked="" type="checkbox"/> |

# Exploration : définir un corpus

## Corpus de travail

Formulaire

Multicritères

Auteurs

Date

Genre littéraire

Corpus

Mes corpus

### ■ Recherche dans un élément bibliographique

dans l'auteur

dans le titre

dans le genre littéraire

dans la date

cote Frantext

\*

### ■ Corpus de travail

Nombre de textes :

15

Nombre de mots :

161 026

Sélectionner tous les textes

Visualiser

Stats

Vider

Options

- sensible à la casse
- sensible aux diacritiques
- sous-chaîne
- bibliographie détaillée

Astuce : les expressions de date

■ Rechercher dans les textes

Passage à l'étape 2

# Exploration : formuler une recherche (mot)

Accueil

Corpus de travail

Recherche dans les textes

Recherche par mots ou séquence

Recherche par lemmes

Recherche de cooccurrences

Recherche des mots d'une liste

Recherche dans les mots du corpus

Historique des recherches

Que signifie recherche dans les textes ?

Les expressions de séquence

Calculs de fréquence

Étude de voisinage

Listes de mots

## Recherche par mots et séquence

Mots ou séquence Lemmes Cooccurrences Mots d'une liste Mots du corpus Historique

**Mot ou séquence**

grenouilles

*texte exact*  
 *flexion d'un verbe*  
 *flexion d'un substantif ou adjectif*  
 *expression de séquence*  
 *expression régulière*  
 *flexion et variantes médiévales*  
 *flexion et variantes XVI<sup>e</sup>-XVII<sup>e</sup>*  
 *flexion moderne*

Effacer le formulaire Lancer la recherche

# Exploration *(oups : problème d'affichage)*

## Résultats 1 à 11 / 11

Concordancier avancé

Début

<<

>>

- ▶ [1] **S004** pour eux, soit pour leurs **LES** affaires. E XIV, LE LIÈVRE ET
- ▶ [2] **S004** fut un signal Pour s'enfuir **alla** devers sa tanière. Il s'en
- ▶ [3] **S004** sa tanière. Il s'en alla passer : sur le bord d'un étang
- ▶ [4] **S004** agisse en Loup; C'est le plus **LES** certain de beaucoup. LE IV,
- ▶ [5] **S004** beaucoup. LE IV, LES >  
GRENOUILLES QUI DEMANDENT UN ROI
- ▶ [6] **S004** des Dieux leur envoie une **tue** Grue, Qui les croque, qui les
- ▶ [7] **S004** lui romprons encor la tête. XII, LE SOLEIL ET LES **GRENOUILLES**
- ▶ [8] **S004** race est détruite. Bientôt on la **du** verra réduite à l'eau

GRENOUILLES> FABLE XIV LE LIÈVRE ET LES

Zoom

passer sur le bord d'un étang : Grenouilles

Zoom

Grenouilles aussitôt de sauter dans les ondes

Zoom

GRENOUILLES QUI DEMANDENT UN ROI> FABLE IV

Zoom

FABLE IV LES GRENOUILLES QUI DEMANDENT UN ROI

Zoom

, Qui les gobe à son plaisir,  
Et Grenouilles

Zoom

> FABLE XII LE SOLEIL ET LES GRENOUILLES

Zoom

Styx. Pour un pauvre Animal, Grenouilles à mon

Zoom

# Exploration : formuler une recherche (lexie + regex)

## Recherche dans les mots du corpus

Mots ou séquence Lemmes Cooccurrences Mots d'une liste **Mots du corpus** Historique

**Formulaire**

(gren|citr)ouill.\*

texte exact  
 texte en début  
 texte à l'intérieur  
 texte strictement à l'intérieur  
 texte en fin  
 *expression régulière*

sensible à la casse  
 sensible aux diacritiques

**Avancé**

citrouille 2  
citrouilles 1  
grenouille 10  
grenouilles 11

Rechercher dans les textes

Créer la liste

Ajouter à la liste - ▾

# Exploration : formuler une recherche (co-occurrence)

## Recherche de cooccurrences de mots

Mots ou séquence Lemmes Cooccurrences Mots d'une liste Mots du corpus Historique

Formulaire

un accès plus complexe (« intersection »)

|                          |             |           |
|--------------------------|-------------|-----------|
| Séquence 1 : grenouilles | texte exact | ▼         |
| Séquence 2 : sauter      | -           | ▼ Voulu ▼ |
| Séquence 3 :             | -           | ▼ Voulu ▼ |

Contexte de la co-occurrence :

dans une même phrase  
 pas nécessairement dans la même phrase

Positions relatives des séquences 1 et 2 : indifférente ▼ distance maximale : 20 mots

Positions relatives des séquences 1 et 3 : indifférente ▼ distance maximale : 20 mots

Positions relatives des séquences 2 et 3 : indifférente ▼ distance maximale : 20 mots

Effacer le formulaire Lancer la recherche

# Exploration : formuler une recherche (co-occurrence)

Recherche : grenouilles + sauter...

## Résultats 1 à 1 / 1

Début << >>

► [1] S004

alla passer sur

encore ce problème d'affichage...

Zoom

On constate quelques problèmes d'affichage en ces temps de migration vers « Frantext 2 » (novembre 2017), sans conséquences sur la pertinence des résultats :

<< 100 300 500 700 1000 1500 \* >>

## Résultat 1

► [1] S004 - LA FONTAINE Jean de, *Fables* : Livres 1 à 6, 1668, p. 89

### LIVRE DEUXIÈME, FABLE XIY, LE LIÈVRE ET LES GRENOUILLES

, tout lui donnait la fièvre.  
Le mélancolique Animal,  
En rêvant à cette matière,  
Entend un léger bruit : ce lui fut un signal  
Pour s'enfuir devers sa tanière.

Il s'en alla passer sur le bord d'un étang :  
Grenouilles aussitôt de sauter dans les ondes;  
Grenouilles de rentrer en leurs grottes profondes.  
Oh ! dit-il, j'en fais faire autant

mais d'affichage seulement !

## Exploration : types de recherche

Le moteur de Frantext permet donc d'interroger la base :

|  |             |
|--|-------------|
| ➤ en cherchant des mots ou une séquence                      | x ?         |
| ➤ en cherchant des lemmes                                    |             |
| ➤ en cherchant la co-ocurrence de certains mots              | (x + y) ?   |
| ➤ en lui soumettant une liste de mots                        |             |
| ➤ par la liste des mots effectivement présents dans le texte | x in CORPUS |

Seul le dernier mode tend vers l'exploration « inductive », non pilotée.

# Exploration : types de recherche

d'ici, on peut créer un index complet

## Recherche dans les mots du corpus

Mots ou séquence

Lemmes

Cooccurrences

Mots d'une liste

Mots du corpus

Historique

### Formulaire

La regex attrape tout

\*

Appliquer

Effacer

- texte exact
- texte en début
- texte à l'intérieur
- texte strictement à l'intérieur
- texte en fin
- expression régulière

- sensible à la casse
- sensible aux diacritiques

Retour aux occurrences

Avancé

|            |   |
|------------|---|
| accru      | 4 |
| accrus     | 2 |
| accrut     | 1 |
| accueil    | 3 |
| accumulait | 1 |
| accumuler  | 1 |
| accusa     | 1 |
| accusait   | 3 |
| accusant   | 1 |
| accusateur | 1 |

Rechercher dans les textes

Créer la liste

Ajouter à la liste

Export d'un fichier exploitable  
qui est l'index du texte

12 228 formes fréquence totale 161 009

Télécharger la liste résultat

## Exploration : types de recherche

L'export des résultats fait de Frantext une « ressource » exploitable comme une étape dans un projet de recherche.

Sa fonction première est cependant documentaire (chercher une expression donnée -> ex. suivant : *il faut mourir*) et quantitative (repérer, mesurer la distribution de ces objets)

## Exploration : types de recherche

Or la recherche lexicale, même dans une base structurée comme Frantext, reste soumise à des ambiguïtés que la machine ne perçoit pas nécessairement.

Cherchons par exemple le verbe défectif « falloir », dont les occurrences que l'on pourrait trouver auront les formes : *falloir, fallu, faut, faudra, faudrait, fallait, fallut, faille, fallût*

On repère immédiatement le problème de l'ambiguïté du signe, « en surface » (la « forme », vs le lemme): qu'il faille vs une faille

Et un texte lemmatisé l'est rarement *assez bien* pour exclure toute erreur.

# Exploration : types de recherche

## Résultats 1 à 50 / 50000

Début << Km >>

### ► [1] R671 - MAROT Jean, *Le Voyage de Gênes*, 1507, p. 96

- du roy tousjours. De nuyt et jour ce peuple et villenaile Si tressouvent leur livroyent la bataile Que des François les gens diminuoyent ; Mais pour ung d'eulx, est à croire sans faille, Qu'ilz tuoyent tant de ceste coquinnaille Que cimetieres et maisons en puoyent. Ce neantmoins tousjours en recouvroyent Qui aux François estoit inrecouvrable Jusqu'aulx

Zoom

Haut

### ► [2] R970 - LA CHESNAYE Nicolas de, *La Condamnation de Banquet*, 1508, p. 73

une table ronde ou carree et se la saison est qu'on ne puisse finir de prunes, fault prendre prunes seiches ou en faire de cire qui auront forme et couleur de \*Damas. GOURMANDISE Il faut remplir noz estomachz Soit de trippes ou de jambon. FRIANDISE Fy ! fy ! C'est pour \*Jehan ou \*Thomas ! Il me suffist de pou et bon. JE-BOY-à-VOUS Voicy belle provision : Pastez,

Zoom

Haut

### ► [3] R970 - LA CHESNAYE Nicolas de, *La Condamnation de Banquet*, 1508, p. 137

y est semblablement, Tout entier, sans nulles myettes. Disposez si bien les apprestes ! Vueillez voz platz si bien coucher Qu'ilz treuvent leurs viandes prestes Et qu'il ne faille que trencher ! L'ESCUYER Il nous fault doncques ces platz loger. LE CUYSINIER Leurs propres lieux assignerons. Tous les platz seront serrez sur une petite table, et les nommeront

Zoom

Haut

### ► [4] R970 - LA CHESNAYE Nicolas de, *La Condamnation de Banquet*, 1508, p. 140

aprés toutes ces merelles Il fault merles et tourterelles. L'ESCUYER Et pour bailler aguiselement Belles orenges largement. LE CUYSINIER Aprés chair, selon noz usaiges, Il faut tartes à deux visages. L'ESCUYER Je vueil aussi qu'on leur propine La belle tarte jacopine. LE CUYSINIER Pour viande commune et tritte Il fault avoir la cresme fritte. L'

Zoom

Haut

## Exploration : types de recherche

L'étude de « voisinage », c'est-à-dire de la distribution des contextes, permet de chercher des expressions figées.

exercice « voisinage » : « faille »

1° qu'indique le résultat brut de cette étude ?

2° cliquez sur le 2<sup>e</sup> ou le 3<sup>e</sup> contexte privilégié et classez les résultats par ordre chronologique croissant ou décroissant

# Exploration : types de recherche

**Étude du voisinage**

**Voisinage mot**      **Voisinage liste**

## ■ Étude du voisinage d'un mot

Mot à étudier : faille

Définition du voisinage :

phrase contenant le pivot

|   |                         |
|---|-------------------------|
| 0 | nombre de phrases avant |
| 0 | nombre de phrases après |

portion de texte

|   |                      |
|---|----------------------|
| 4 | nombre de mots avant |
| 4 | nombre de mots après |

Triés des résultats :

- par ordre alphabétique des mots
- par ordre croissant des fréquences
- par ordre décroissant des fréquences

Format de sortie :

- sur l'écran
- dans un fichier à télécharger

# Exploration : types de recherche

## ■ Résultats

*Attention, le lien pour afficher le contexte est une approximation.*

|      |                 |
|------|-----------------|
| 3385 | <b>faille</b>   |
| 328  | <b>sanz</b>     |
| 67   | <b>li</b>       |
| 52   | <b>semble</b>   |
| 34   | <b>quoiqu'</b>  |
| 34   | <b>voir</b>     |
| 32   | <b>vos</b>      |
| 28   | <b>temps</b>    |
| 26   | <b>chercher</b> |
| 25   | <b>croire</b>   |
| 25   | <b>crois</b>    |
| 25   | <b>nule</b>     |
| 25   | <b>pense</b>    |
| 25   | <b>prendre</b>  |
| 24   | <b>attendre</b> |
| 23   | <b>robe</b>     |
| 23   | <b>tenir</b>    |
| 22   | <b>aller</b>    |

# Exploration : types de recherche

## Résultats 1 à 50 / 353

*la normalisation des données n'est jamais infaillible...*

Debut << km >>

- [1] 0233 - CHRISTINE DE PISAN, *LE LIVRE DE L'ADVISION CHRISTINE*, 1405, p. 49

LA FIN DE LA COMPLAINTE DE LA DAME COURONNEE [XXIX]

grace n'y remedie, que passé a long temps ne fus plus proplexe. Helas! mais comment remede du ciel  
espereroie quant aux miens si mal je le voy desservir? Et encore plus me grieve **sanz faille** le peril de  
pis ou je me voy que e mal que je seuffre, tant soie bien batue, tout ainsi que celui qui devant lui voit  
cil qui l'a navré a paour qu'il le pertue. Si te merci, ma bien

Zoom

Haut

- [2] A151 - PIZAN Christine de, *Le Livre du duc des vrais amants*, c.1405, p. 278

Le Duc des vrais amans

2365 Garir bien se penera, Et que jour assignera, Temps et heure, et en quel place Parler, ains que  
long temps passe, Porray a elle **sanz faille**, 2370 Et que la lettre me baille, Disant que se recommande  
A moy, et qu'elle me mande Que plus je ne me soussie

Zoom

Haut

- [3] A151 - PIZAN Christine de, *Le Livre du duc des vrais amants*, c.1405, p. 292

Le Duc des vrais amans

Excusacion en l'eure Trouva que la a celle heure Ere alé pour un pesant 2545 Afaire qui en present Lui  
est survenu **sanz faille**, Dont lui fault comment qu'il aille Parler au seigneur en haste, Car moult grant  
besoing le haste, 2550

Zoom

Haut

# Exploration : types de recherche

« li » est un pronom du français médiéval -> on peut s'attendre à trouver le verbe (~*lui faille*). [ordre chrono inversé ~j. 1400]

## Résultats 1 à 50 / 190

Début << Km >>

▶ [1] 5608 - \*ANONYME , Ysaÿe le Triste, 1400, p. 43

mors. » Lors vient Trons li bochus a l'ermitte en disant : « Ne lui en moustrés plus, je vous en prie. - Vous n'en avés garde, fait Sarban, car je n'en ay nulle vollenté. - Sans **faille**, fait **li** nains, ce n'est mie merveilles, car tes caulx ne font mie a atendre. » 27 - (a) En tel guise que par cest livre avés oÿ, seut ferir Ysaÿe d'espee, et depuis Trons ly nains l'

Zoom

Haut

*Ce n'est pas le cas : « sans faille » domine*

▶ [2] 5608 - \*ANONYME , Ysaÿe le Triste, 1400, p. 52

Tenés, mon maistre vous envoie ce cheval et voeut que vous l'aiés car combien qu'ail ait abatu vo chevalier et gaignié se monture, si ne le veut il point emmener. - Sans **faille**, dist **li** portiers, se ly sires a perdu son cheval pau y aconte, car d'autres en a assés, Dieu merchi, et loch bien que tu le gardes pour toy, sy yras a cheval car plus aise yras qu'a troter

Zoom

Haut

▶ [3] 5608 - \*ANONYME , Ysaÿe le Triste, 1400, p. 81

e avoit geu. « Ce n'est pas la qu'il gist, par Dieu, fait Trons, c'est en le cambre de la. » Lors voit

Zoom

# Exploration : types de recherche

« li » est un pronom du français médiéval -> on trouve le verbe (~*lui faille*) en remontant le temps [*ordre croissant ~1150*]

Résultats 1 à 50 / 190

Début << Km >>

► [1] A022 - \*ANONYME , *Roman de Thèbes*, 1150, p. 81

-

> mes se il dui armé fussant, 8639 puis que **li** rois armez ne vet, 8640 il eüssent grant **faille** fet. 8641  
Parthonopieus par aventure

*Ici, le substantif*

Zoom

Haut

► [2] A012 - \*ANONYME , *Eneas (volume 1)*, 1155, p. 5

-

, 147 qui est deesse de bataille, 148 et pria **li** que ne li **faille**, 149 a li se tiegne au jugement, 150 et al li

Zoom

*Ici on trouve bien le verbe, mais l'occurrence est comptée deux fois...*

ut

► [3] A012 - \*ANONYME , *Eneas (volume 1)*, 1155, p. 5

- 147 qui est deesse de bataille, 148 et pria li que ne **li faille**, 149 a li se tiegne au jugement, 150 et al li  
donra

Zoom

Haut

# Exploration : périodisation

## Périodisation des résultats (étude des fréquences)

### ■ Formulaire d'étude d'un mot

---

Mot à étudier :

faille

Fréquences :

relatives  absolues

---

Tranche de corpus :  auteur par auteur

tri par ordre des fréquences  
 tri par ordre alphabétique des auteurs

---

référence par référence

tri par ordre des fréquences  
 tri par ordre alphabétique des références

---

par tranches de temps

durée d'une tranche de temps (années)

50

résultats triés par ordre des fréquences  
 résultats triés par ordre chronologique

---

Format de sortie :

sur l'écran  dans un fichier à télécharger

---

# Exploration : périodisation

## ■ Diagramme des fréquences relatives

Échelle : un astérisque représente une fréquence relative de **3** millionième(s)  
freq. abs. freq. rel.

|           |     |     |       |
|-----------|-----|-----|-------|
| 0- 49     | 267 | 111 | ***** |
| 1150-1199 | 96  | 92  | ***** |
| 1200-1249 | 69  | 91  | ***** |
| 1400-1449 | 192 | 82  | ***** |
| 1350-1399 | 135 | 43  | ***** |
| 1550-1599 | 165 | 26  | ***** |
| 1300-1349 | 21  | 21  | ***** |
| 1500-1549 | 36  | 21  | ***** |
| 1600-1649 | 230 | 21  | ***** |
| 1650-1699 | 248 | 18  | ***** |
| 1250-1299 | 4   | 16  | ***** |
| 1450-1499 | 41  | 16  | ***** |
| 2000-2049 | 187 | 12  | ****  |
| 1950-1999 | 677 | 9   | ***   |
| 1700-1749 | 97  | 7   | ***   |
| 1900-1949 | 392 | 7   | ***   |
| 1750-1799 | 122 | 5   | **    |
| 1850-1899 | 243 | 5   | **    |
| 1800-1849 | 163 | 4   | **    |

## Exploration : séquences

### Autre exemple

On veut pouvoir à l'inverse étudier une expression identifiée comme un segment récurrent dans la littérature ou la langue, et en faire la « généalogie ».

exercice « séquence » :      (faut | fault) mourir      355 *résultats*

↔ « regex » :                [Ii]l faul?t mourir      (*en panne...*)

choisissez d'abord l'affichage « quantifié » des résultats pour observer la distribution chronologique sur la totalité de la base.

alternative (séquence)      &cfalloir mourir      498 *résultats*

## Exploration : séquences

L'expression de **séquences** a sa propre « grammaire », que concurrencera bientôt, dans *Frantext 2*, l'intégration du langage **CQL** (Corpus Query Language, une manière de formuler des requêtes devenue un nouveau standard et présente dans le BNC et TXM par exemple)

& cf falloir mourir      fléchit *falloir* (et pas l'infinitif *mourir*)

il &q mourir -> il *<nimportquelmot>* mourir  
il &q(1,2) mourir -> il *<l\_ou\_2\_mots>* mourir

*Pour info ou test dans TXM, leurs équivalents en CQL :*

[lemma='falloir'] mourir

il [ ] mourir

il [ ]{1,2} mourir

# **De l'exploration au traitement des données**

## **Traitement de données :**

Exo 1 : Choisissez une période de publication plus restreinte, une expression donnée, et comparez les résultats « quantifiés » de votre recherche dans Frantext et dans le Ngram-viewer de Google Livres.

Exo 2 : Choisissez un texte exportable (il doit être libre) et entraînez-vous à faire les mêmes requêtes dans Notepad++ par expressions régulières.