



Institut européen

des métiers de la **traduction** | IEMT

Université de Strasbourg

Web, corpus, traduction : exploitations

Sketch Engine

Enzo Doyen

2025 - M1

Sketch Engine

<https://auth.sketchengine.eu/>

Tableau de bord

DASHBOARD

French Web 2023 (frTenTen23)



[Get more space](#)



FRENCH WEB 2023 (FR TENTEN23)

CORPUS INFO

MANAGE CORPUS



Word Sketch

Collocations and word combinations



Thesaurus

Synonyms and similar words



Parallel Concordance

Translation search



N-grams

Multiword expressions (MWEs)



Trends

Diachronic analysis, neologisms



OneClick Dictionary

Automatic dictionary drafting



Word Sketch Difference

Compare collocations of two words



Concordance

Examples of use in context



Wordlist

Frequency list



Keywords

Terminology extraction



Text type analysis

Statistics of the whole corpus



Bilingual terms

Bilingual terminology extraction

RECENTLY USED CORPORA

NEW CORPUS

French Web 2023 (frTenTen23)	French	23,874,070,858	
French Web 2020 (frTenTen20)	French	15,115,914,647	
Europarl spoken parallel – English	English	53,837,625	
Europarl spoken parallel – French	French	59,145,988	

English Trends corpus



Explore our largest English corpus with 80+ billion words.











Automatic updates: 70 million new words every week.

OPEN CORPUS

Sélection d'un corpus

DASHBOARD

type to search  


 AMR_Jaimes_2023, English	English	613,826
 AMR_Jaimes_2023, Spanish	Spanish	761,352
 ex	French	64,594
 iemt_env	French	61,199
 inc2	French	85,706
 Inclusion	French	66,551
 RSE	English	120,744
 samaritanus_bonus_EN	English	15,841
 test_iss	English	
 test_steamworks, English	English	


902 corpora


☐ Show description


[ADVANCED SEARCH](#) [CREATE CORPUS](#)


FRENCH WEB 2023 (FRT)

 **Word Sketch**
Collocations and word combinations

 **Thesaurus**
Synonyms and similar words

 **Parallel Concordance**
Translation search

 **N-grams**
Multiword expressions (MWEs)

 **Trends**
Diachronic analysis, collocations


Sélection d'un corpus : recherche avancée


BASIC

ADVANCED

MY CORPORA

SHARED WITH ME

Corpus or language...
840 corpora 108 languages

Any language

Only with word sketches ☐

<u>Language</u>	Name ↑
English	ACL Anthology Reference Corpus (ARC)
Afrikaans	Afrikaans Web 2024 (afTenTen24)

Catégories de corpus

CORPUS CATEGORY

ALL ⓘ 840

RECENT ⓘ 0

MY CORPORA ⓘ 8

SHARED WITH ME ⓘ 8

FEATURED ⓘ 15

GENERAL PURPOSE ⓘ 359

WEB ⓘ 370

NON-WEB ⓘ 470

PARALLEL ⓘ 216

SPOKEN ⓘ 91

SPECIALIZED ⓘ 358

DIACHRONIC ⓘ 172

MULTIMEDIA ⓘ 2

LEARNER ⓘ 8

ERROR-ANNOTATED ⓘ 5



I. Analyse de corpus monolingues

Fonctionnalités des corpus **monolingues** intéressantes dans une perspective **traductionnelle**

- ♦ *Word Sketch* ;
- ♦ *Word Sketch Difference* ;
- ♦ *Thesaurus* ;
- ♦ *Concordance* ;
- ♦ *Wordlist* (surtout avec un corpus personnalisé) ;
- ♦ *Keywords* (surtout avec un corpus personnalisé).

Word Sketch

Voir les cooccurrences d'un mot dans différents contextes.

WORD SKETCH

French Web 2023 (frTenTen23)



BASIC

ADVANCED

AS A LIST

ABOUT

Search ?

bibliothèque

GO

bibliothèque as noun 1,143,626×

...

verbs with "bibliothèque" as object	verbs with "bibliothèque" as subject	modifiers of "bibliothèque"
fréquenter ... fréquenter la bibliothèque	brûler ... c' est une bibliothèque qui brûle	municipal ... la bibliothèque municipale • concentrated in: reference ? • concentrated in: arts ? • concentrated in: reference/encyclopedia ?
abriter ... abrite la bibliothèque • concentrated in: reference ? • concentrated in: reference/encyclopedia ?	renfermer ... bibliothèque renferme	universitaire ... la bibliothèque universitaire • concentrated in: politics & government ? • concentrated in: education ? • concentrated in: legal ?
enrichir ... enrichir la bibliothèque	déborder ... bibliothèque déborde	itunes ... la bibliothèque iTunes • concentrated in: technology & IT ?
léguer ... lègue sa bibliothèque • concentrated in: reference ? • concentrated in: reference/encyclopedia ?	conserver ... La bibliothèque conserve • concentrated in: reference ? • concentrated in: reference/encyclopedia ?	numérique ... la bibliothèque numérique • concentrated in: education ?
garnir ... bibliothèque bien garnie	regorger ... bibliothèque regorge de	départemental ... la bibliothèque départementale • concentrated in: legal ?
enceindre ... enceintes bibliothèque • concentrated in: discussion ?	enrichir ... la bibliothèque s' enrichit	
spécialiser ...	rouvrir ... bibliothèque rouvre	
	contenir ...	

Word Sketch : options avancées

BASIC

ADVANCED

AS A LIST

ABOUT

Search ?

bibliothèque

Part of speech ?

auto

noun

verb

adjective

adverb

Subcorpus ?

none (the whole corpus) ▾ 🔒 +

Minimum frequency ?

auto

Minimum score ?

0

☐ Translate ?

Text types ? ^

Genre ▾

Topic ▾

Top-level domain (e.g. com) ▾

Website (e.g. cnn.com) ▾

Web domain (e.g. news.blogs.cnn.com) ▾

Heading ▾

expand all

collapse all

GO

Word Sketch Difference

Comparer les cooccurrences de deux mots.

WORD SKETCH DIFFERENCE

French Web 2023 (frTenTen23)



BASIC

ADVANCED

ABOUT

First lemma ?

bibliothèque

Second lemma ?

librairie

GO

bibliothèque
1,143,626×

librairie
413,720×

↔

⚙️

🔍

✕

"bibliothèque/librairie" and/or ...

archive	7,820	64	...
Archives	1,563	10	...
étagère	1,787	21	...
musée	5,628	320	...
librairie	4,374	582	...
médiathèque	4,324	716	...
bibliothèque	3,210	4,374	...
imprimerie	120	1,008	...
disquaire	27	278	...
kiosque	29	978	...
EUR	0	351	...
papeterie	0	588	...

▼

↔

⚙️

🔍

✕

verbs with "bibliothèque/ librairie" as object

enceindre	616	0	...
enrichir	1,290	37	...
léguer	431	9	...
garnir	657	23	...
abriter	1,549	154	...
fréquenter	1,966	522	...
installer	2,011	1,307	...
écumer	151	96	...
compiler	168	139	...
spécialiser	2,801	4,322	...
dévaliser	72	101	...
achalander	42	69	...

▼

Thesaurus

Liste de synonymes ou de noms au sémantisme proche.

THESAURUS

French Web 2023 (frTenTen23)



BASIC

ADVANCED

ABOUT

Search ?

bibliothèque

GO

	<u>Lempos</u>	<u>Frequency</u> ²
1	bureau	2,828,659 ...
2	musée	1,315,201 ...
3	archive	976,469 ...
4	collection	2,436,530 ...
5	établissement	2,981,550 ...
6	école	5,127,929 ...
7	salle	4,783,319 ...
8	bâtiment	2,383,369 ...
9	librairie	413,720 ...
10	salon	2,037,948 ...

	<u>Lempos</u>	<u>Frequency</u> ²	
1	bureau	2,828,659	Word Sketch Difference
2	musée	1,315,201	Thesaurus
3	archive	976,469	Concordance
4	collection	2,436,530	Concordance with macro
5	établissement	2,981,550	Word Sketch
6	école	5,127,929	
7	salle	4,783,319	...
8	bâtiment	2,383,369	...
9	librairie	413,720	...
10	salon	2,037,948	...

Concordance

Permet des recherches complexes à l'aide des requêtes CQL (voir cours dédié).

Wordlist : exemple sur un corpus personnalisé sur l'écologie

Liste de mots avec leur fréquence.

WORDLIST

iemt_env 🔍 ⓘ

word (8,159 items | 73,742 total frequency)

	Word	Frequency ? ↓
1	,	3,779 ***
2	de	3,372 ***
3	.	2,032 ***
4	la	1,923 ***
5	les	1,838 ***
6	et	1,627 ***
7	l'	1,571 ***
8	des	1,544 ***
9	d'	1,449 ***
10	en	1,354 ***

	Word	Frequency ? ↓
11	à	1,197 ***
12	le	1,033 ***
13)	786 ***
14	(784 ***
15	énergie	669 ***
16	pour	637 ***
17	dans	583 ***
18	un	578 ***
19	une	566 ***
20	est	538 ***

Wordlist : exemple sur un corpus personnalisé sur l'écologie

Résultats initiaux pas très utiles...

Wordlist : exemple sur un corpus personnalisé sur l'écologie

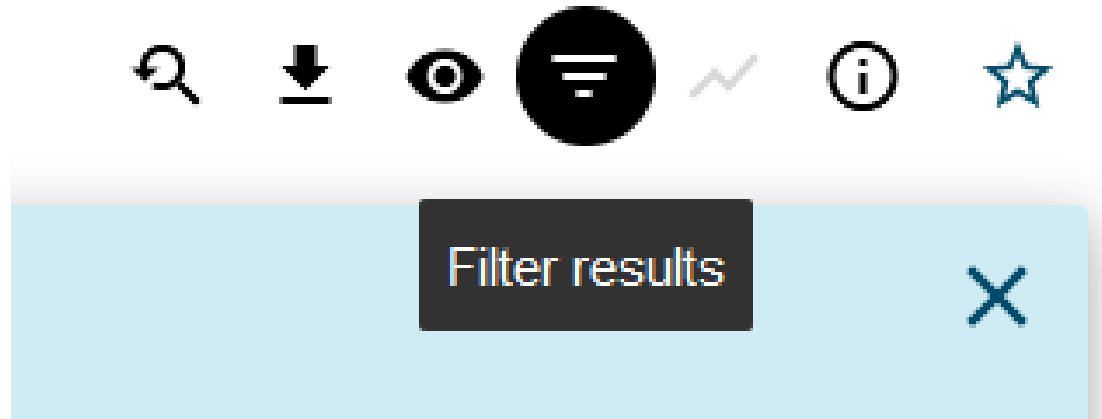
Résultats initiaux pas très utiles...

Mais on peut utiliser les **options en haut à droite** !



Wordlist : exemple sur un corpus personnalisé sur l'écologie

Filtrage des résultats



FILTER RESULTS		
Filter		
•→ ▼ éco		
	Word	Frequency ? ↓
1	écosystèmes	228 ...
2	écosystème	210 ...
3	écologique	71 ...
4	écologiques	27 ...
5	économie	20 ...
6	économique	19 ...
7	écologie	18 ...
8	écosystémiques	15 ...
9	économiques	13 ...
10	économies	9 ...

Keywords : exemple sur un corpus personnalisé sur l'écologie

Extraction terminologique effectuée sur la base d'un **corpus de référence**.

Notion de « corpus de référence »

Définition

Corpus de référence : corpus utilisé comme point de référence par rapport au corpus étudié, afin d'effectuer des comparaisons.

Un **corpus de référence** est généralement un corpus de **langue standard**, et est ainsi utilisé pour voir comment certains textes **s'éloignent du langage courant** (textes spécialisés, notamment).

SINGLE-WORDS ✓

MULTI-WORD TERMS ✓



reference corpus: French Web 2023 (frTenTen23) (items: 5,144)

	<u>Lemma</u>	
1	gw	...
2	trophique	...
3	renouvelable	...
4	trophiques	...
5	biomasse	...
6	écosystème	...
7	enr	...
8	hydroélectricité	...
9	twh	...
10	renewable	...

Keywords : exemple sur un corpus personnalisé sur l'écologie (multimots)

<u>Term</u>		
1	énergie renouvelable	...
2	mémoire écologique	...
3	niveau trophique	...
4	niveau trophiques	...
5	combustible fossile	...

<u>Term</u>		
14	consommation finale	...
15	producteur primaire	...
16	énergie solaire	...
17	éolien en mer	...
18	transfert d' énergie	...



II. Analyse de corpus bilingues

Fonctionnalités des corpus **bilingues**

- ♦ *Parallel Concordance,*
- ♦ *Bilingual terms.*

Parallel Concordance

Recherche de traductions en contexte, dans les deux sens (langue source, langue cible).

PARALLEL CONCORDANCE

Europarl spoken parallel – English

BASIC

ADVANCED

ABOUT

Search ?

celebration

in

English

Translated as (optional) ?

in

French

SEARCH

simple **celebration** • 121

7.13 per million tokens • 0.00071%



align ▾



Europarl spoken parallel – French

① #1043295

<s> For that reason, yes, it is a cause for **celebration** ! </s>

① #1043703

<s> We will be celebrating in December, and I hope that what will come afterwards will not be a hangover but the positive memory of the **celebration** itself and the reasons for it. </s>

① #1046292

<s> It is only in this way, paying attention to those who remain outside, that the **celebrations** for opening up the borders can be complete for us. </s>

<s> Alors oui, nous pouvons **fêter** cet instant! </s>

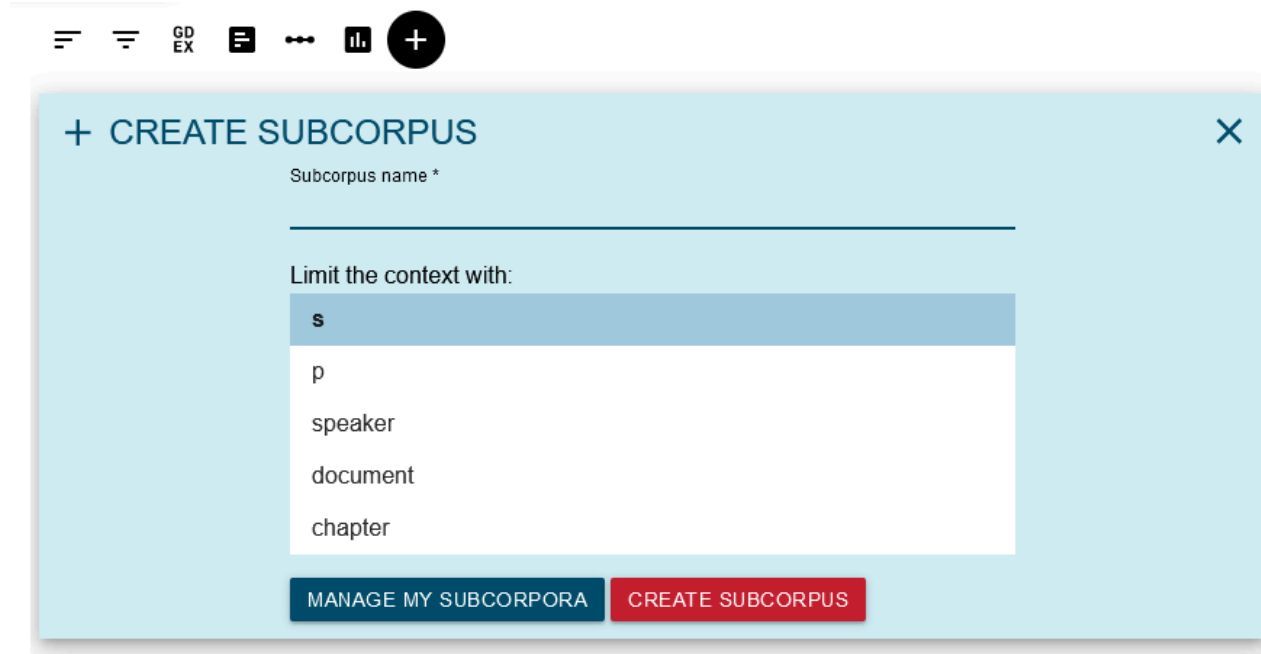
<s> L' heure sera à la fête en décembre et j' espère que nous n' en garderons pas une gueule de bois, mais le souvenir positif de cette **célébration** et des raisons qui y ont conduit. </s>

<s> Ce n' est qu' à cette condition, que si nous prêtons attention à ceux qui restent à l' extérieur, que la **fête** marquant l' ouverture des frontières sera totale pour nous. </s>

Possibilité de télécharger les résultats



Parallel Concordance : sous-corpus via résultats



The screenshot shows a software interface with a top navigation bar containing icons for menu, list, GDEX, document, link, and a plus sign. A modal dialog box titled '+ CREATE SUBCORPUS' is open, featuring a close button (X) in the top right corner. Inside the dialog, there is a text input field labeled 'Subcorpus name *'. Below this, a section titled 'Limit the context with:' contains a list of options: 's' (selected), 'p', 'speaker', 'document', and 'chapter'. At the bottom of the dialog are two buttons: 'MANAGE MY SUBCORPORA' and 'CREATE SUBCORPUS'.

+ CREATE SUBCORPUS

















Subcorpus name *

Limit the context with:

- s
- p
- speaker
- document
- chapter

MANAGE MY SUBCORPORA CREATE SUBCORPUS

Parallel Concordance : ajout de plusieurs langues

BASIC		ADVANCED		ABOUT	
Search 		Translated as (optional) ?		  	
<input type="text" value="celebration"/>		<input type="text"/>		<input type="text"/>	
in		in		in	
<input type="text" value="English"/>		<input type="text" value="French"/>		<input type="text"/>	
					
Translated as (optional) ?		Translated as (optional) ?		Translated as (optional) ?	
 		 		 	
<input type="text"/>		<input type="text"/>		<input type="text"/>	
in		in		in	
<input type="text" value="Polish"/>		<input type="text" value="Spanish"/>		<input type="text"/>	
					
<div>SEARCH</div>					

<s> Tak, to jest powód do świętowania! </s>	<s> Por esa razón, sí, ¡ es motivo de celebración! </s>
<s> W grudniu będziemy świętować i mam nadzieję, że to, co nastąpi potem, to nie będzie kac, lecz pozytywne wspomnienie samego świętowania i jego powodów. </s>	<s> Tendremos una celebración en diciembre, y confío en que lo que venga después no sea una resaca, sino un recuerdo positivo de la celebración en sí, y de las razones que la motivaron. </s>
<s> Świętowanie otwarcia granic będzie dla nas pełne tylko wtedy, gdy będziemy pamiętali o tych, którzy pozostają na zewnątrz. </s>	<s> Sólo de este modo, prestando atención a los que permanecen en el exterior, las celebraciones por la apertura de las fronteras pueden ser plenas para nosotros. </s>
<s> Wczoraj świat obchodził międzynarodowy Dzień Praw Człowieka, który ma szczególne znaczenie w tym roku, ponieważ rozpoczyna obchody upamiętniające 60. rocznicę przyjęcia Powszechnej Deklaracji Praw Człowieka. </s>	<s> Ayer se celebró en todo el mundo el Día internacional de los derechos humanos, que es especialmente importante este año porque marca el inicio de las celebraciones para conmemorar el 60º aniversario de la aprobación de la Declaración Universal de los Derechos Humanos. </s>
<s> Sądzę, że Parlament Europejski dobrze zaczął obchody 60. </s><s> Rocznicę ogłoszenia Powszechnej Deklaracji Praw Człowieka przyznając nagrodę im. Sacharowa za wolność przekonań imponującemu Salihowi Mahmoudowi Osmanowi. </s>	<s> en nombre del Grupo ALDE. - Señor Presidente, creo que el Parlamento Europeo ha iniciado con buen pie la celebración del 60º aniversario de la Declaración Universal de los Derechos Humanos con la concesión del Premio Sajarov a la Libertad de Conciencia al admirado señor Salih Mahmoud Osman. </s>
<s> w imieniu grupy ALDE. - Panie przewodniczący! </s><s> Pan poseł Schulz ma całkowitą rację, mówiąc, że nadszedł tydzień świętowania, stwierdzając uroczystie, że Karta chroniąca naszych obywateli przed przemocą w związku z ogromną władzą, jaką dysponuje Unia, oraz podpisanie traktatu lizbońskiego umacnia naszą zdolność działania i w niezwykle wysokim stopniu poprawia jakość naszej demokracji. </s>	<s> en nombre del Grupo ALDE. - Señor Presidente, el señor Schulz tiene toda la razón al decir que ésta es una semana de celebración, en la que se proclama solemnemente la Carta que protege a nuestros ciudadanos contra los abusos del enorme poder del que está investida la Unión, y se firma el Tratado de Lisboa que refuerza nuestra capacidad de actuar y que mejora decididamente la calidad de nuestra democracia. </s>

Bilingual terms

Extraction terminologique bilingue.

Ne fonctionne que sur les **corpus personnalisés** : requiert donc de disposer de **ses propres documents**.

Fonctionne autant avec les **documents alignés** que les **documents non alignés**.

Rappel : alignement

EN	FR
<p><s>Steam Deck OLED has 30-50% more battery life.</s> <s>We fit a bigger battery into the case, and the OLED display draws less power.</s></p> <p><s>Combined with the updated, more efficient AMD APU, you have way more time to play your favorites.</s></p>	<p><s>L'autonomie de Steam Deck OLED est 30 à 50 % supérieure à celle du modèle LCD</s> <s>, ce grâce à une plus grande batterie et à l'écran OLED, qui est moins énergivore.</s></p> <p><s>Ajoutez à cela un nouvel APU d'AMD plus efficace, et vous [...].</s></p>

Bilingual terms

Extraction terminologique bilingue.

Ne fonctionne que sur les **corpus personnalisés** : requiert donc de disposer de **ses propres documents**.

Fonctionne autant avec les **documents alignés** que les **documents non alignés**.

Comme pour l'extraction terminologique monolingue, repose sur l'utilisation d'un **corpus de référence**.

Bilingual terms : exemple sur corpus de l'OMS

(Merci à Nelson Jaimes-Quintero)

SINGLE -WORDS [EN]MULTI -WORDS [EN]BILINGUAL TERMS SINGLE -WORDS [ES]MULTI-WORDS [ES]

PARALLEL CONCORDANCESELECT ALLEL Deselect ALLShow statisticsDOWNLOAD...

Source term	Target term
1. antimicrobial	<div><div><input checked="" type="radio"/> uso de antimicrobianos ✖</div><div><input type="radio"/> salud humana</div><div><input type="radio"/> uso responsable</div><div><input type="radio"/> RAM</div><div><input type="radio"/> sector de la salud humana</div></div>
2. antimicrobial resistance	<div><div><input checked="" type="radio"/> FAO ✖</div><div><input type="radio"/> Alimentarius</div><div><input type="radio"/> Codex</div><div><input type="radio"/> OMS</div><div><input type="radio"/> resistencia</div></div>
3. amr	<div><div><input checked="" type="radio"/> RAM ✖</div><div><input type="radio"/> sola salud</div><div><input type="radio"/> PNUMA</div><div><input type="radio"/> plan de acción</div><div><input type="radio"/> FAO</div></div>

Bilingual terms : exemple sur corpus de l'OMS

SINGLE [EN]

MULTI [EN]

BITERMS

SINGLE [ES]

MULTI [ES]

☒ Show statistics

DOWNLOAD...

Keyword ?	Frequency ?	Rel. frequency ?	Ref. frequency ?	Rel. ref. frequency ?	Score ?
1. antimicrobiano ↗	5,399	5,891.847	36,258	1.851	2,067.27
2. antibiótico ↗	3,680	4,015.928	187,753	9.583	379.58
3. aeruginosa ↗	434	473.618	5,371	0.274	372.50
4. resistance ↗	322	351.394	1,041	0.053	334.62
5. aureus ↗	483	527.091	12,727	0.650	320.14
6. BOV ↗	270	294.647	146	0.007	293.46
7. vancomicina ↗	327	356.850	4,400	0.225	292.23
8. pneumoniae ↗	371	404.867	8,128	0.415	286.86



III. Création de corpus personnalisés

Étapes pour créer un corpus

RECENTLY USED CORPORA			NEW CORPUS
French Web 2023 (frTenTen23)	French	23,874,070,858	
Europarl spoken parallel – English	English	53,837,625	
French Web 2020 (frTenTen20)	French	15,115,914,647	
Europarl spoken parallel – French	French	59,145,988	

Indiquez le nom, le type de corpus et la langue.

Build your own private corpus from texts on the web or from your own documents.

Name

required

Corpus type

☒ Single language corpus

☐ Multilingual corpus

Language

English

Description

Storage used: 294,881 of 1,000,000 words (29%)

Étapes pour créer un corpus

Recherche de textes sur le Web.



Find texts on the web

Automatically find and download relevant texts

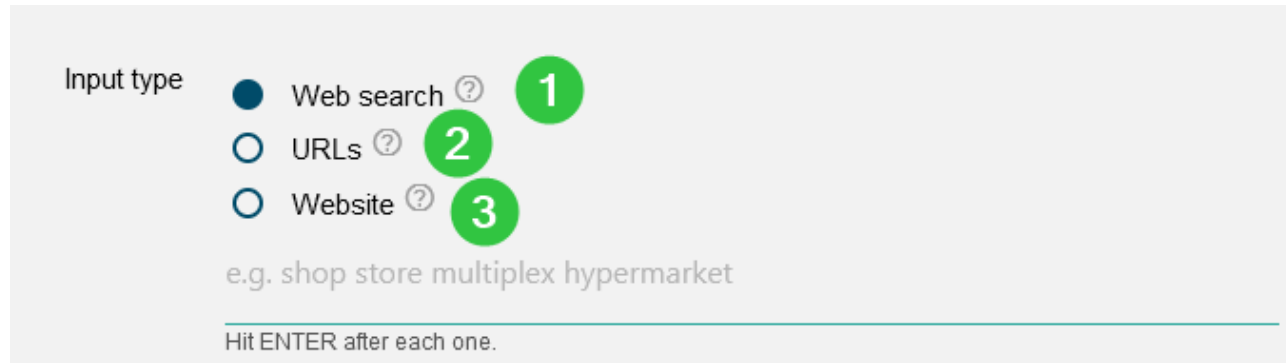


I have my own texts

Upload your own files (.txt, .pdf,...) or paste text

Étapes pour créer un corpus

1. recherche directe par mots-clés
2. liste d'adresses URL
3. toutes les pages d'un site Web.



Input type

- ☒ Web search [?] 1
- ☐ URLs [?] 2
- ☐ Website [?] 3

e.g. shop store multiplex hypermarket

Hit ENTER after each one.

Paramètres : recherche sur le Web ; liste de mots à bannir ; liste de mots à autoriser (uniquement) ; restriction de taille.

Input type

☒ Web search ?

☐ URLs ?

☐ Website ?

e.g. shop store multiplex hypermarket

Hit ENTER after each one.

Folder name ? web1

Web search settings ▼

Denylist settings ▼

Allowlist settings ▼

Size restrictions ▼

Étapes pour créer un corpus

Définition des mots-clés.

Input type

- ☒ Web search [?]
- ☐ URLs [?]
- ☐ Website [?]

llm ✕ modèles de langue ✕ nlp ✕ tal ✕

You can type additional words or phrases. Hit ENTER after each one.

Étapes pour créer un corpus

Sélection des pages desquelles extraire le contenu.

Filter

type to search

SELECT VISIBLE **DESELECT VISIBLE** **EXPAND ALL** **COLLAPSE ALL**




✓ modèles de langue • nlp • tal (23/23 selected) ^

- ✓ talnarchives.atala.org/TALN/TALN-2020/66.pdf
- ✓ aclanthology.org/2018.tal-2.1.pdf
- ✓ aclanthology.org/2021.tal-1.1.pdf
- ✓ aclanthology.org/2021.tal-2.1.pdf
- ✓ aclanthology.org/2022.jeptalnrecital-taln.26.pdf
- ✓ fr.wikipedia.org/wiki/Traitement_automatique_des_langues

Étapes pour créer un corpus

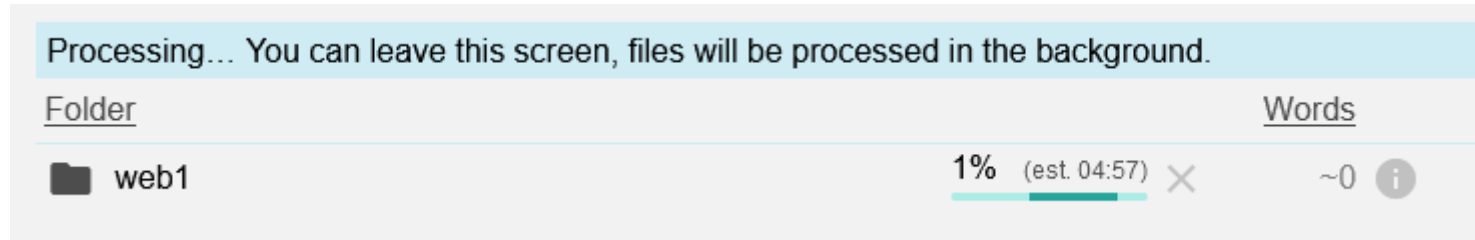
Compilation...

Processing... You can leave this screen, files will be processed in the background.

<u>Folder</u>		<u>Words</u>
 web1	1% (est. 04:57) 	~0 

Étapes pour créer un corpus

Compilation...



Et le corpus est prêt !