Support de formation : Introduction à Hyperbase

Margaux Nguyen Ngoc Minh

JEUDI 23 JANVIER 2025
18H-20H *en visio*

CONSTITUER ET TRAITER SES *CORPUS*

Séminaire doctoral CHCSC & DYPAC
Méthodologie de la thèse en SHS

Introduction à Hyperbase

Margaux NGUYEN NGOC MINH
IGE Humanités Numériques MSH PS

Présentation de l’outil HYPERBASE

Hyperbase
corpus, langage

- Accessible en ligne gratuitement : https://hyperbase.unice.fr/
- Créé en 1989 par Etienne Brunet
- Diffusé par le CNRS et l'Université Nice Côte d'azur. Conçu et développé au sein de l'UMR Bases, Corpus, Langage
- Outil pour l’analyse documentaire et statistique de textes – lexicométrie, textométrie

Présentation de l'outil HYPERBASE

- Lemmatiseurs (POS): TreeTagger et Cordial
- Algorithmes de deep learning pour la reconnaissance automatique de motifs linguistiques profonds (voir Laurent VANNI)
- Analyse arborée selon Xuan LUONG
- Comparaison avec le corpus du dictionnaire en ligne TLF, Google Books et le British National Corpus
- Traitement du français, latin, espagnol, anglais, allemand, italien, portugais, russe, grec
- Imports et exports aux formats XML et TEI

Exemples de bases et variantes

- ECRIVAINS : base qui compare tout le vocabulaire de 70 écrivains de la littérature française (55 millions de mots)
- CHRONO : base qui permet de suivre l'évolution du vocabulaire littéraire de 1600 à nos jours (117 millions de mots)
- FRANCIL : base issue d'une enquête sur le français parlé et écrit dans les pays francophones (4,5 millions de mots)
- Bases composées de monographies intégrales d'auteurs français
- Bases composées de corpus politiques / médiatiques
- Bases téléchargeables : http://ancilla.unice.fr/hyperbase/pages/bases/

- CD de littérature latine (Sylvie Mellet) : https://hyperbase.unice.fr/bases
- CD de littérature algérienne (Marie Virolle) : sur demande
- Version portugaise d'Hyperbase (Carlos Maciel, projet PORTEXT) : http://bcl.unice.fr/portext/

Prise en main d'Hyperbase

Hyperbase
corpus, langage

Nouvelle base

Base existante

Bibliothèque

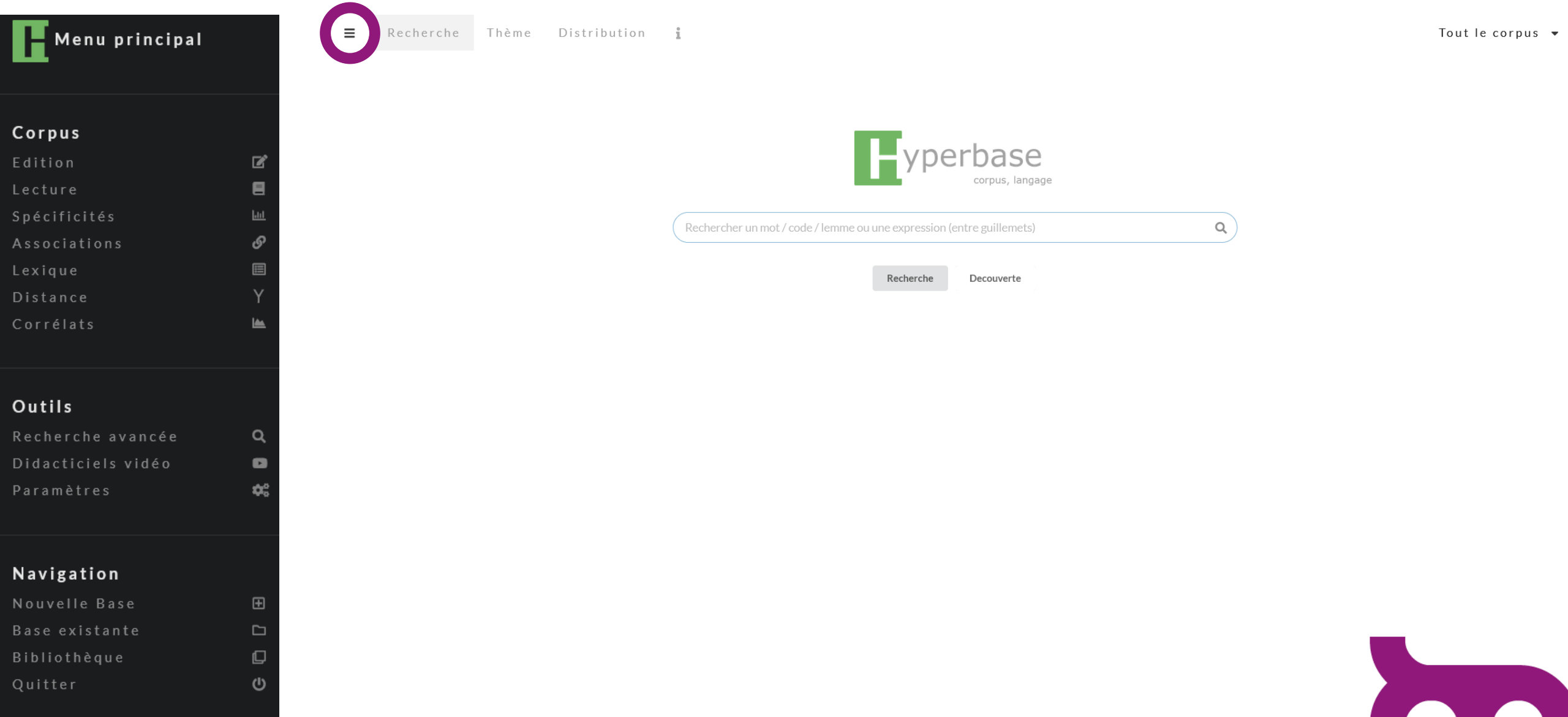
Bibliothèque des bases

Toutes les bases ▾ Toutes les langues ▾

Dernières bases consultées

elysee | fr Discours presidentiels français de 1958 à aujourd'hui	1878 vues il y a 2 jours
lasla | la Ensemble des textes latins classiques du corpus du L.A.S.L.A. (ULg) di...	808 vues il y a 9 jours
latin | la selection des textes latins de la base classique du L.A.S.L.A. (ULg)	464 vues il y a 12 jours
Campagne2022 | fr Campagne présidentielle 2022	843 vues il y a 15 jours
Campagne2017 | fr Campagnes présidentielles 2017	260 vues il y a 27 jours
mitterrand | fr Mitterrand	87 vues il y a 29 jours
plavtvs | la Comédies de Plaute traitées par le L.A.S.L.A.	21 vues il y a 29 jours
montesquieu | fr Montesquieu	2 vues il y a 30 jours
Campagne2007 | fr Campagne présidentielle 2007	37 vues il y a 30 jours

Prise en main d'Hyperbase



Requête par code grammatical

Comment formuler mes requêtes

Un mot
je

Plusieurs mots
je nous on

Un code grammatical
PRON

Un lemme
LEM:penser

Une expression
"je pense"

Remplacer un mot
"LEM:je ??? ADJ"

Remplacer plusieurs mots
"LEM:je *** ADJ"

Remplacer un caractère
alloc.tion

Remplacer plusieurs caractères
re.*

Une requête complexe
"LEM:je VERB:Pres CCONJ PRON LEM:être"

Liste des codes

Catégorie	
ADJ	Adjectif
ADP	Adposition
ADV	Adverbe
AUX	Auxiliaire
CCONJ	Conjonction De Coordination
DET	Déterminant
INTJ	Interjection

Catégorie	
NOUN	Nom
NUM	Numéral
PRON	Pronom
PROPN	Nom Propre
PUNCT	Punctuation
SCONJ	Conjonction De Subordination
SENT	Fin De Phrase

Catégorie	
SYM	Symbole
VERB	Verbe
X	Autre

Paramétrage

Paramètres

Unité textuelle

Forme

Unité statistique

Spécificité

Filtres

Catégories grammaticales

Choix des contextes

- (•) Nombre de mots 10 Bidirectionnel
- ( ) Paragraphe
- ( ) Phrase

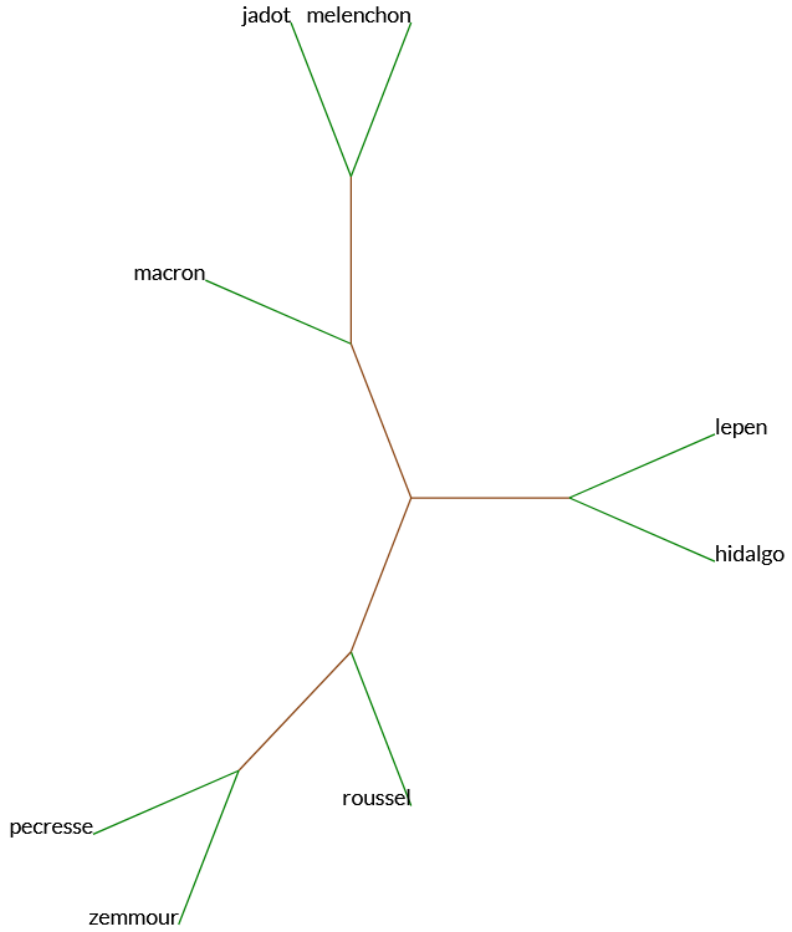
Taille des résultats

100

Annuler Valider

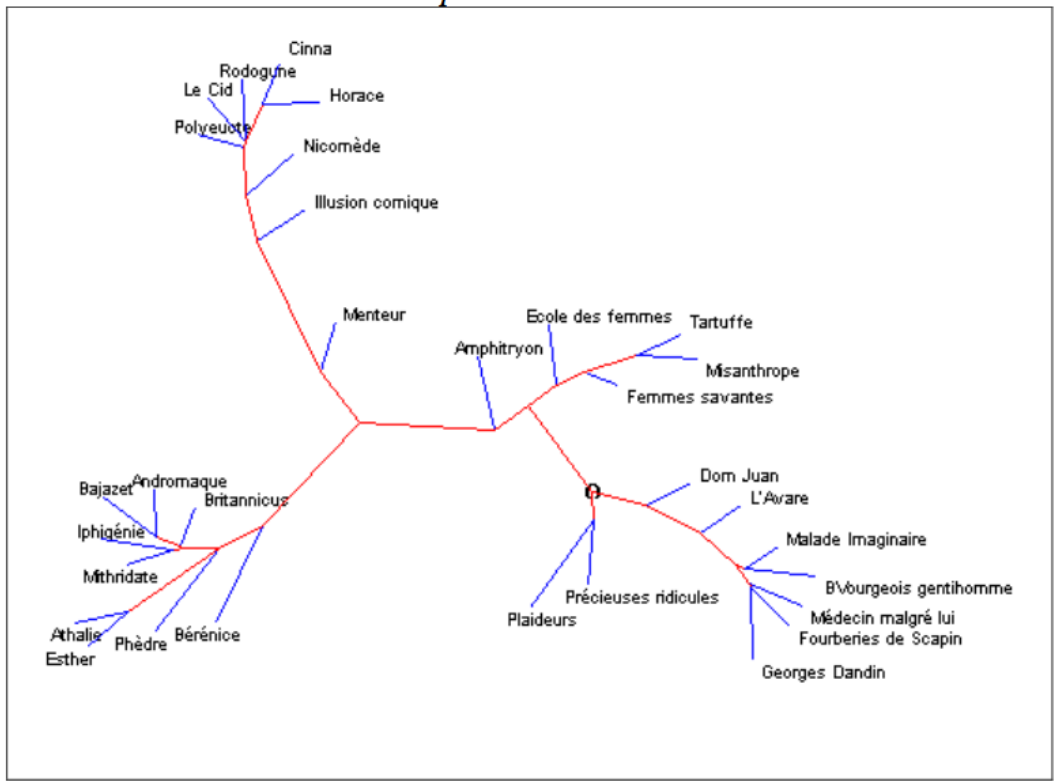
Distance

Distance intertextuelle



Base « Campagne2022 » en ligne sur Hyperbase.

Analyse arborée de la distance lexicale (calculée sur N)
Représentation radiale



E. Brunet, *Manuel de référence - Hyperbase*, 2011, p. 54.

Spécificité

Spécificités de jadot

Rechercher

Formes

Écart	Corpus	Texte	Mot
16.76	138	75	climat
13.47	608	131	;
13.13	49	37	écologistes
11.55	47	32	écologiste
11.34	124	49	écologie
11.21	106	45	énergies
10.4	11144	942	les
10.29	2205	258	sur
9.87	228	59	évidemment
8.8	14	14	pulsions
8.39	164	43	climatique
8.06	12	12	lobby
7.95	142	38	Poutine
7.81	99	31	énergétique
7.58	57	23	Vladimir
7.39	40	19	forts
7.21	81	26	logements
7.12	116	31	Françaises
7.01	180	39	démocratie
6.89	5542	464	nous
6.88	483	71	projet

Codes

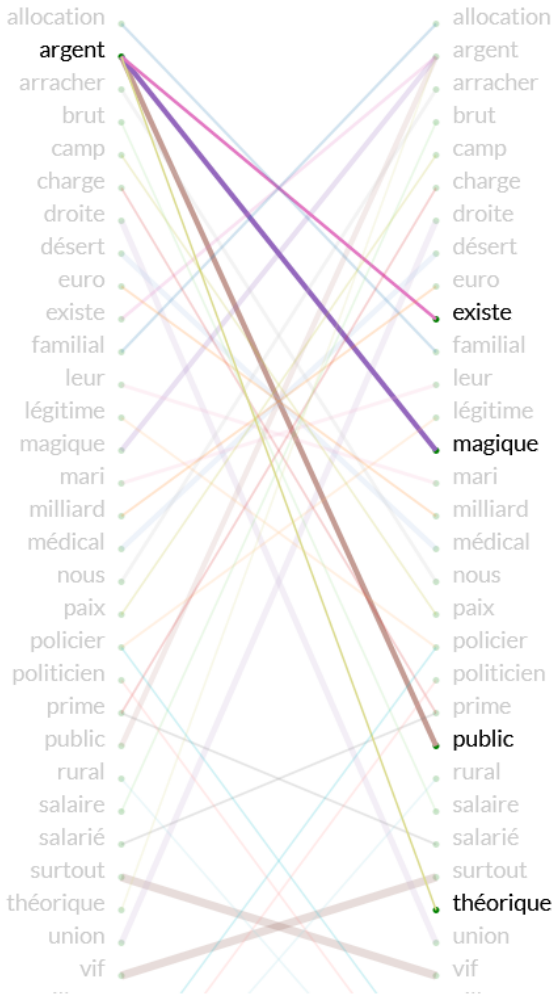
Écart	Corpus	Texte	Mot
12.93	36524	2785	SENT
12.43	57645	4173	Def
10.93	111221	7548	Plur
10.38	73281	5085	Art
9.87	109598	7375	NOUN
6.77	94394	6211	Fem
6.25	79470	5238	DET
5.74	34952	2381	ADJ
4.11	5420	403	Card
3.05	4660	332	NUM
2.6	3842	270	Fut
2.44	81141	5087	ADP
2.39	115886	7224	Masc
1.28	123	10	X
1.08	18273	1134	CCONJ
0.7	1306	80	Sub
-1.01	49	2	PART
-1.47	1215	65	Ord
-1.64	9131	528	2pers
-1.69	23606	1392	AUX
-2.1	104	2	INTJ

Lemmes

Écart	Corpus	Texte	Mot
17.45	92	65	écologiste
16.76	138	75	climat
13.47	608	131	;
11.34	124	49	écologie
10.42	48252	3458	le
10.24	2270	263	sur
9.85	540	97	public
9.8	275	65	énergie
9.42	245	59	évidemment
9.4	26	20	lobby
9.06	185	49	climatique
8.2	135	38	Poutine
8.2	7038	597	nous
8.19	16	14	pulsion
8.15	110	34	énergétique
7.65	56	23	Vladimir
7.25	591	84	projet
7.04	195	41	logement
6.97	56	21	paysan
6.85	47	19	élevage
6.77	197	40	démocratie

Associations

Associations de zemmour



Lemmes

Écart	Corpus	Texte	Association
52.55	413	62	vif surtout
37.47	349	43	droite union
35.29	934	66	argent public
32.92	141	34	médical désert
32.21	928	33	argent magique
23.63	1952	30	leur mari
20.91	7062	39	nous arracher
19.80	984	24	argent existe
17.35	170	19	familial allocation
15.56	214	18	salarié prime
15.27	861	52	euro milliard
15.20	143	16	policier légitime
15.08	993	19	ville village
13.79	200	16	charge prime
13.69	293	15	salaire brut
13.26	5822	53	vous politicien
12.47	921	13	argent théorique
12.42	308	15	paix camp
12.22	302	17	policier violence

1 →

Lexique

Lexique issu du récit d'I. Jablonka, *Laëtitia ou La Fin des hommes*, Seuil, 2016.

Rechercher

A	6
a	1321
aa	6
abaisse	1
abandon	3
abandonne	2
abandonner	1
abandonné	5
abandonnée	2
Abandonnées	1
abandonnées	1
abandonnés	6
abandonnique	2
abandons	1
abattue	2
abattues	1
aberration	1
abêti	1
abîme	1
abîment	1
abîmé	1
abject	1
aboie	1
abolition	1
abominable	3
abomination	2
abondamment	1
abord	14
aborde	2
aborder	3
abordé	1
abords	6
aboutir	1
aboutissement	1

Rechercher

&	1
+	1
02	1
027	1
1,50	1
1,60	1
1,64	1
1,85	1
105	1
1156	1
11h	1
122	1
129	1
12h12	1
139	1
140	1
1481	1
1485	1
1614	1
170	1
175	1
1760	1
1789	1
1791	1
1793	1
1798	1
1802	1
1810	1
1824	1
1830	1
1836	1
1865	1
1870	1
1875	1

Rechercher

,	9839
.	5439
de	5219
la	3028
à	2221
le	2043
et	2006
les	1802
l'	1801
un	1550
est	1508
des	1452
d'	1374
en	1364
a	1321
une	1164
«	983
»	983
du	893
il	880
Laëtitia	873
dans	848
elle	832
que	779
pas	741
au	726
qu'	716
qui	651
pour	639
son	592
sur	592
se	569
sa	506

Corrélats

A partir du récit d'I. Jablonka, *Laëtitia ou La Fin des hommes*, Seuil, 2016.

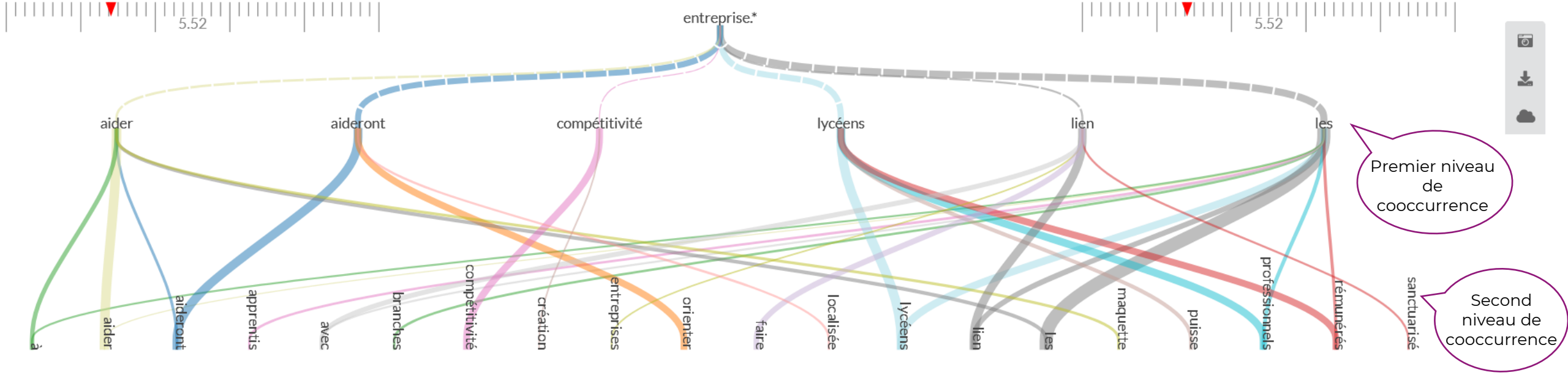
Thème – arborescence

Hyperbase

entreprise.*

Recherche | Thème | Distribution

macron



Indice de spécificité	Probabilité	Fréquence corpus	Fréquence partie	Taille corpus	Taille partie	Mot
7.7	3.39e-14	11144	73	640393	1520	les
7.15	1.56e-12	7	5	640393	1520	lycéens
7.07	2.65e-12	2	4	640393	1520	aideront
5.8	6.33e-9	102	7	640393	1520	aider
5.71	1.07e-8	63	6	640393	1520	lien
5.52	3.1e-8	14	4	640393	1520	compétitivité

Thème – nuage de mots

Recherche | Thème | Distribution | macron

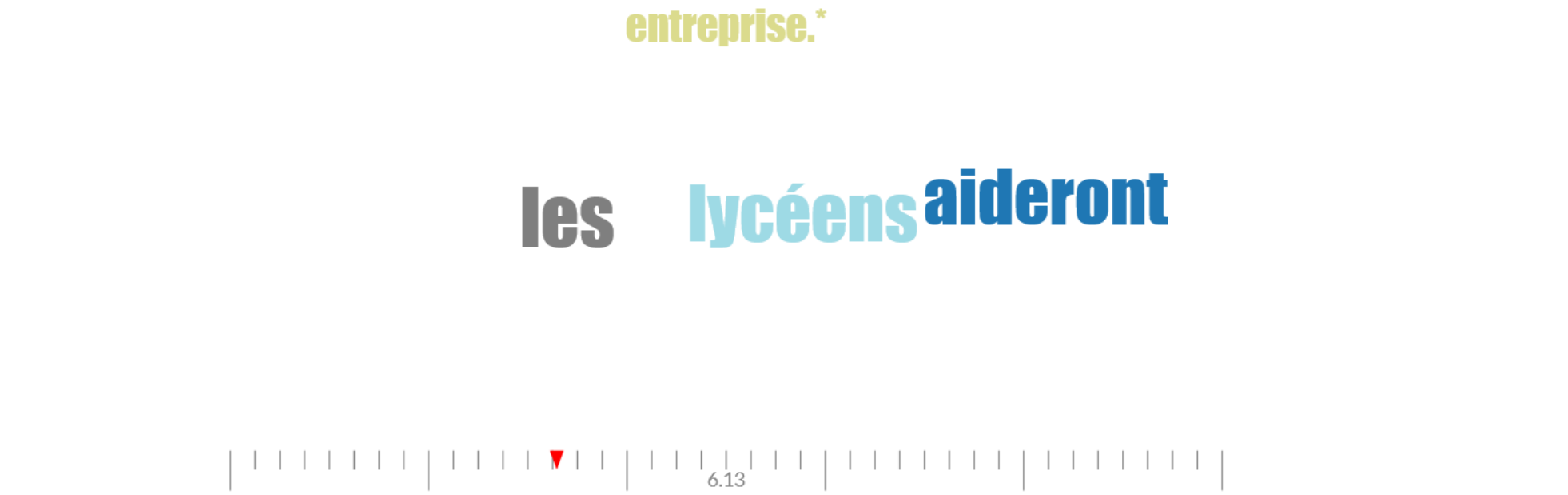
Premier niveau de cooccurrence

5.10

cité	Probabilité	Fréquence corpus	Fréquence partie	Taille corpus	Taille partie	
	3.39e-14	11144	73	640393	1520	
	1.56e-12	7	5	640393	1520	
	2.65e-12	2	4	640393	1520	
	6.33e-9	102	7	640393	1520	

Thème – nuage de mots

Recherche | Thème | Distribution | macron



icité	Probabilité	Fréquence corpus	Fréquence partie	Taille corpus	Taille partie
	3.39e-14	11144	73	640393	1520
	1.56e-12	7	5	640393	1520
	2.65e-12	2	4	640393	1520
	6.33e-9	102	7	640393	1520

Thème – nuage de mots

"féministe" dans la campagne présidentielle 2022
Triomphe lexical ou une lutte sémantique ?

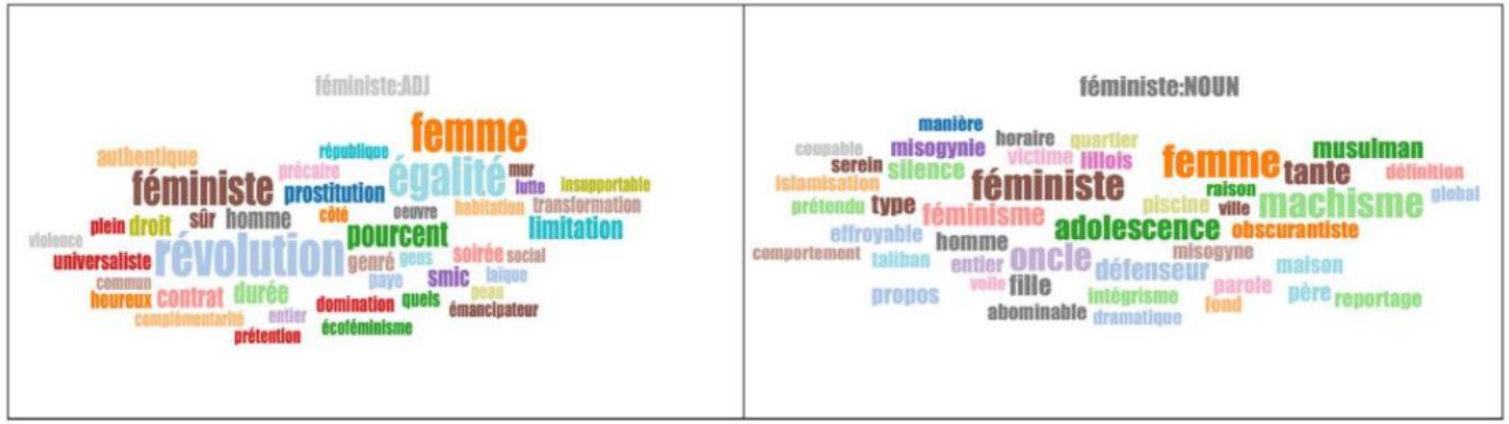


Figure 2. Lemmes co-occurrents de l'adjectif « féministe » (corpus gauche) et du nom « féministe » (corpus extrême droite)

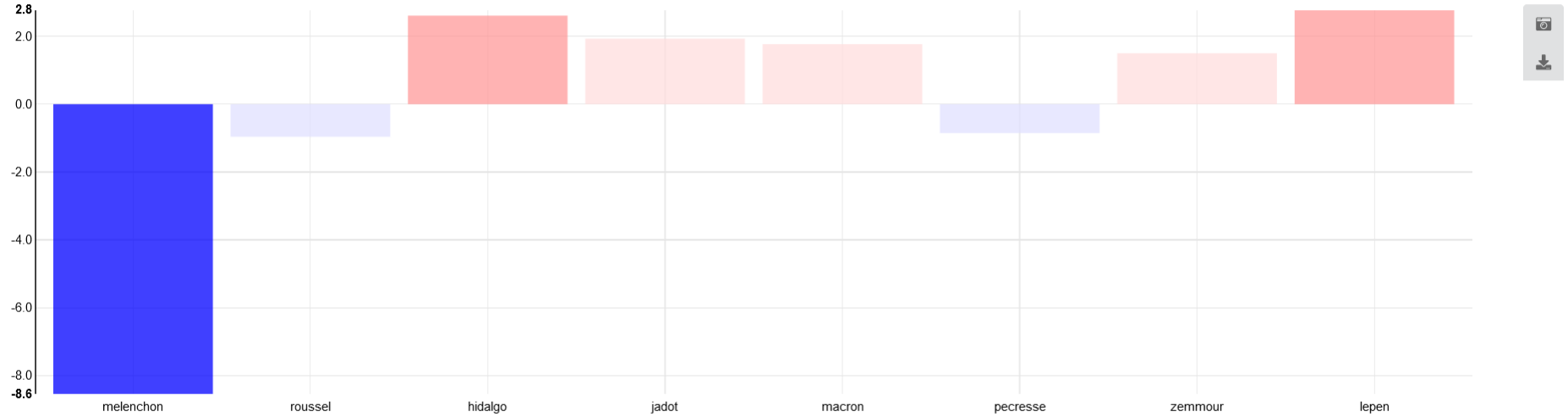
Magali Guaresi. Les mots "féminisme" et "féministe" dans la campagne présidentielle : un triomphe lexical ou une lutte sémantique ?. Corpus, 2023, 24, ⟨10.4000/corpus.7714⟩. ⟨hal-03979102⟩

Exemple issu de la présentation d'Hyperbase par L. Vanni : https://hyperbase.unice.fr/doc/presentation/index.html

Distribution

Hyperbase "notre pays"

Recherche | Thème | Distribution | Tout le corpus



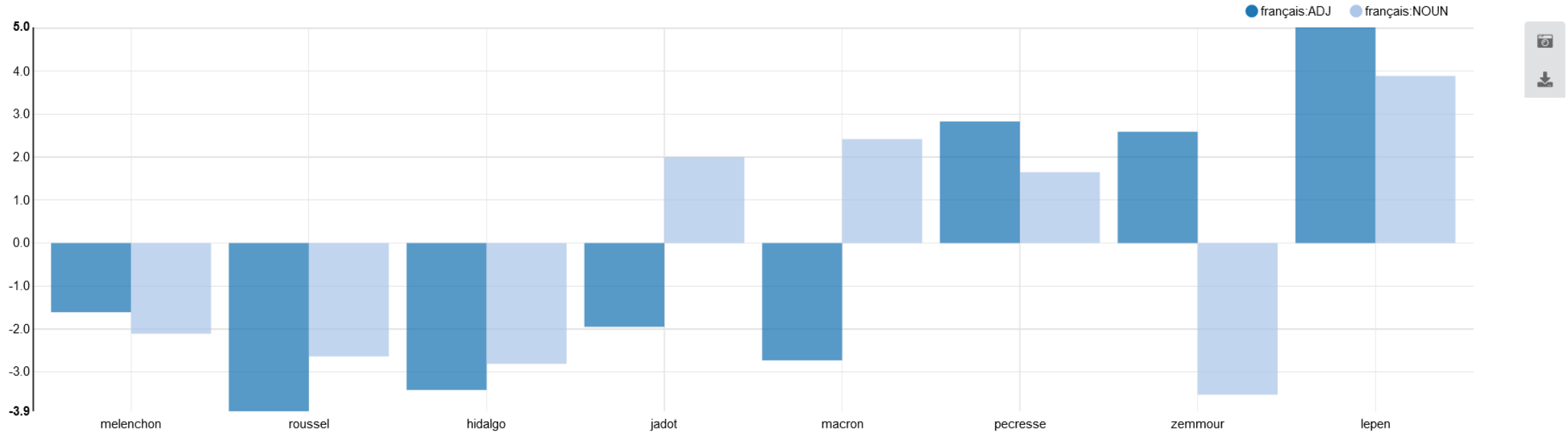
Mots	Total	melenchon	roussel	hidalgo	jadot	macron	pecresse	zemmour	lepen
notre pays	392	-8.55	-0.96	2.61	1.93	1.77	-0.85	1.5	2.78

Distribution

Hyperbase

français:ADJ français:NOUN

Recherche | Thème | Distribution | Tout le corpus



Mots	Total	melenchon	roussel	hidalgo	jadot	macron	pecresse	zemmour	lepen
français:ADJ	309	-1.61	-3.92	-3.42	-1.95	-2.73	2.83	2.59	5.03
français:NOUN	91	-2.11	-2.64	-2.81	2	2.42	1.65	-3.53	3.89

Merci !

Mail : margaux.nguyen_ngoc_minh@ens-paris-saclay.fr

formation@msh-paris-saclay.fr

Page de la Plateforme Humanités Numériques PS : https://msh-paris-saclay.fr/plateforme-humanites-numeriques-paris-saclay/

Sitographie - Bibliographie :

- E. Brunet, *Manuel de référence - Hyperbase*, 2011. [Disponible en ligne sur le site de l'UNICE, URL : http://ancilla.unice.fr/hyperbase/manuel.pdf]

- M. Guaresi, « Les mots "féminisme" et "féministe" dans la campagne présidentielle : un triomphe lexical ou une lutte sémantique ? », *Corpus*, 2023. [Disponible en ligne sur HAL, URL : https://hal.science/hal-03979102/]

- L. Vanni, Support de présentation d'Hyperbase. [Disponible en ligne sur le site de l'UNICE, URL : https://hyperbase.unice.fr/doc/presentation/index.html]

- L. Vanni, Tutoriels vidéos Hyperbase, 2024. [Disponibles en ligne sur Youtube, URL : https://www.youtube.com/@Hyperbase-Logometrie/videos]