

Analyse du discours

Licence 2 : Lettres modernes

Séance 3 : Ventilation et concordance

marine.delaborde@cyu.fr



CERGY PARIS
UNIVERSITÉ

Organisation des séances

3 / 6

Séances	Points abordés
Séance 1 : 18/01/23 (salle 207)	Introduction
Séance 2 : 22/02/23 (salle 121)	Le projet, iTrameur et le corpus
Séance 3 : 22/03/23 (salle 121)	Ventilation et concordance
Séance 4 : 29/03/23 (salle 121)	Spécificités
Séance 5 : 05/04/23 (salle 121)	Cooccurrents et segments répétés
Séance 6 : 12/04/23 (salle 121)	Retour sur le projet

Évaluation et projet (rappels)

- **Projet** : réalisation d'un dossier d'analyse du discours
 - **Constitution et exploitation de corpus** :
 - collecte de 10 articles de presse,
 - **formatage du corpus : le corpus fait partie du rendu (reproductibilité des expériences),**
 - application d'une méthode qualitative et d'une méthode quantitative,
 - rédaction de 10 pages hors page de couverture et mise en forme du dossier
 - travail en groupe de 2 ou 3 personnes
 - **Dossier** :
 - 10 pages dactylographiées à interligne 1,5 et numérotées, précédées d'un page couverture (sujet+auteur+table des matières+date+formation)
 - Note 1 : Ni la page couverture (qui ne sera pas numérotée), ni les annexes ou la bibliographie ne seront comptées dans le nombre de 10 pages. La bibliographie est formée de références théoriques sur le sujet, d'ouvrages qui analysent le même événement.
 - Note 2 : Indiquez le corpus, c'est-à-dire la liste des articles de presse analysés dans les annexes.
- **Évaluation** : participation en cours et investissement (20%), qualité du dossier d'analyse d'un événement médiatique (80%).

Analyse textométrique

Objectifs du projet (rappels)

- **Objectif** : apporter des éléments de réponse à une question de recherche spécifique en s'appuyant sur les résultats de mesures textométriques appliquées à un corpus de 10 articles.
 - Formuler une **problématique** en lien avec l'évènement abordé dans les 10 articles.
 - Identifier des **hypothèses** qui seront **confirmées** ou **infirmer** par les données.
 - Si les mesures permettent d'infirmer une hypothèse, c'est quand même intéressant !
 - Ne pas modifier les hypothèses formulées au début : **prise de recul** pour comprendre les résultats.
- **Annotation** : ajout d'une balise pour l'annotation du thème des paragraphes.
- **Partition** : Identifier une ou deux partitions pertinente(s) en lien avec l'hypothèse (ex : date, genre, journal, etc.).

iTrameur

Pour commencer...

→ iTrameur : <http://www.tal.univ-paris3.fr/trameur/iTrameur/>

→ Un corpus :

- Un corpus de test : corpus_chronologique_journaux.txt
 - **Attention : la structure de ce corpus n'est pas celle qui est attendue pour le projet !**
- Votre corpus si possible
 - Voir structure d'exemple si besoin : structure-corpus-date-genre-paragraphe.txt

Chargement de la base

iTrameur Analyse textométrique de données

Paramètres +

1. **Chargement** Trame Cadre SR/Patron Section Coocs Bi-Texte Dépendance Sélection Export Aide

Création d'une nouvelle base / Importation d'une base

Deux possibilités pour charger des données dans iTrameur :

1. Charger un fichier (nouvelle base) au format TXT brut, encodé en UTF-8, en ayant préalablement partitionné son contenu (cfonglet Aide).
2. Importer une base annotée déjà constituée (cfonglet Aide pour le format de cette base).

Une fois la base chargée, les données textuelles sont représentées sous la forme d'une *Carte des sections* (sections définies via le délimiteur de contexte choisi) qui apparaît au bas de cette page.

1. NOUVELLE BASE	Choisir un fichier	AUCUN FICHIER CHOISI
1. IMPORTER UNE BASE	Choisir un fichier	AUCUN FICHIER CHOISI
2. DELIMITEUR DE CONTEXTE	\$ (si cette zone est vide, contexte=ligne)	
3. DELIMITEUR(S)	,;~ &#@='~.?!*\$%{}[]_!+<>\$/\	
4. BI-TEXTE	<input type="checkbox"/> (chargement d'un bitexte aligné cfAide)	
5. DEPENDANCE	<input type="checkbox"/> (chargement d'une base avec annotations en dépendance cfAide)	

2. corpus formaté

Entre 1. et 2. :
vérifier le
délimiteur de
sections

Visualisation du corpus de test

Carte des sections

- Une ligne = un élément *date*
- Un carré = un journal = un élément *page* = une section

Pour votre projet : un carré = un paragraphe = une section

Segmentation terminée (corpus_chronologique_journaux.txt :2851173 occurrences / 72633 formes)

Carte des sections avec délimiteur de sections : §

Clic sur section : affichage contexte | Clic-droit sur section : sélection section

Visualisation du corpus de test

Carte des sections

- Quand on clique sur un carré, le texte correspondant à cette section s'affiche en dessous.
- Vérifier en parcourant quelques sections que tout est normal (pas de sections vides, pas de messages d'erreurs).

DATE=20220406 <5593026:5924667>
□ □
DATE=20220410 <5924669:6254305>
□ □
Clic sur section : affichage contexte Clic-droit sur section : sélection section
SECTION N°1

http://www.lefigaro.fr Le Figaro - Actualité en direct et informations en continu Le Figaro - Actualité en direct et informations en continu Aller au contenu Menu Abonnez-vous 1€ le premier mois Rechercher Notre application Soyez alerté en temps réel avec l'application Le Figaro Rubriques et services du Figaro Présidentielle 2022 International Société Vox Économie Sport Culture Voyage Style Madame Vin Figaro Live Rubriques et services du Figaro Figaro Magazine La Vérification Figaro Étudiant Faits divers Santé Sciences Fig Data Tech et web Bourse Figaro Immobilier Art de vivre Automobile Langue française Golf Histoire Programme TV Jardin Figaroscope Carnet du jour Guide achat Services Nos journaux et magazines Les sites du Groupe Figaro Se connecter Fermer le panneau Ouvrir le panneau Mon compte Mon espace personnel Mes newsletters Mes commentaires S'abonner Aide et contact Se déconnecter Lire le journal Abonnez-vous 1€ le premier mois Abonnez-vous : 1€ le premier mois J'en profite Le Covid-19 est-il «deux fois» plus léthal que la grippe ? LA VÉRIFICATION - C'est ce qu'a déclaré le professeur Gilbert Deray ajoutant que «le Covid-19 n'a jamais été une grippe». Les premières pilules de Paxlovid disponibles en France ce vendredi DÉCRYPTAGE - Covid-19 : est-il utile de vacciner les enfants de moins de 5 ans ? Publié il y a 2 heures DÉCRYPTAGE - Covid-19: la vaccination des enfants dans une impasse Publié il y a 17 min En direct Covid-19 : 277 morts en 24 heures, 315.363 nouveaux cas L'exécutive annoncera la semaine prochaine une adaptation du protocole sanitaire à l'école, alors que près de 33.000 patients sont hospitalisés avec le Covid. L'Assurance maladie va supprimer 300.000 faux passages sanitaires dans les deux semaines INFO LE FIGARO - Les revendeurs, qui ont usurpé les cartes professionnelles de médecins pour fabriquer en série les faux sérumes, risquent des peines de prison très lourdes. «Mystérieux», «mentaux», «violents»: Nordal Lelandais

Visualisation du corpus

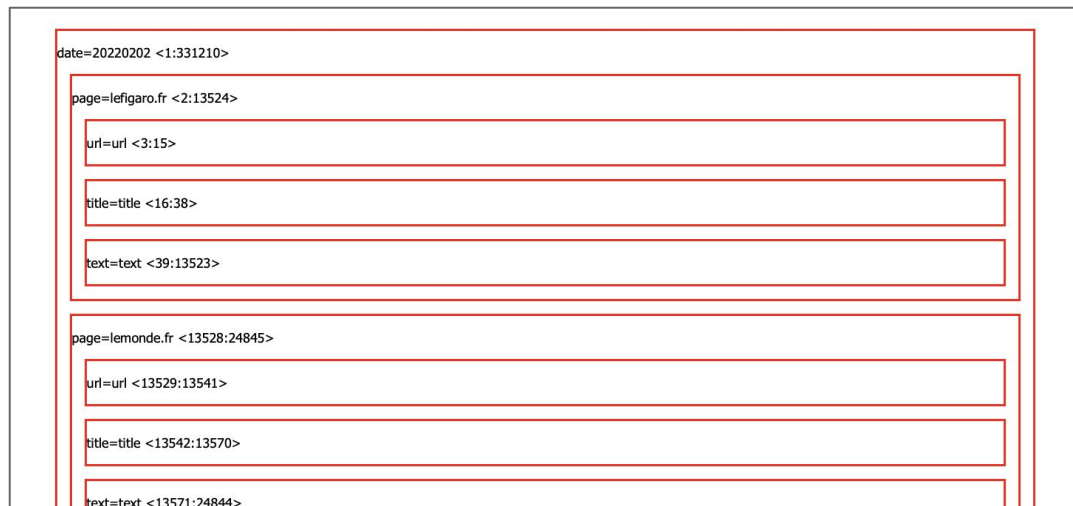
Le cadre



Visualisation du corpus

Le cadre

- Parties définies par les balises du corpus (date, page, url, title, text).
- Vérifier que les balises sont imbriquées correctement.
- Vérifier que les bordures des rectangles représentant les parties ne se croisent pas.



Visualisation du corpus

Le cadre

- Si l'affichage est différent : sûrement un problème de balises
- Dans l'onglet « Paramètres », regarder ce qui s'affiche (le nom des balises) dans le menu « Partition »

The image shows a corpus visualization interface with two document entries. Each entry is enclosed in a red rectangular frame. The first entry has the following fields: `date=20220202 <1:331210>`, `page=lefigaro.fr <2:13524>`, `url=url <3:15>`, `title=title <16:38>`, and `text=text <39:13523>`. The second entry has the following fields: `page=lemonde.fr <13528:24845>`, `url=url <13529:13541>`, `title=title <13542:13570>`, and `text=text <13571:24844>`.

The image shows the 'Paramètres' (Parameters) tab in the corpus visualization interface. The interface is divided into several sections with various parameters and their values. The 'PARTITION' section is highlighted with a red box. The 'PARTIES' section shows the value '20220202'. The 'Lg CONTEXTE' section shows the value '10'. The 'Fq MAX' section shows the value '5'. The 'Co-Freq' section shows the value '2'. The 'SR FQMIN' section shows the value '10'. The 'RELATION' section shows the value 'REL'. The 'Pôle SOURCE' section shows the value 'entrez la forme pôle'. The 'PARTIES' section shows the value '20220202'. The 'SEUIL' section shows the value '5'. The 'SR LGMAX' section shows the value '12'. The 'ANNOTATION SORTIE' section shows the value '1:Forme'. The 'ANNOTATION*' section shows the value '1:Forme'. The 'GRAPHE H' section shows the value '400'. The 'GRAPHE L' section shows the value '800'. The 'NB SÉLECTION SECTION' section shows the value '1'. The 'ANNOTATION RELATION' section shows the value '4'.

Paramètre	Valeur
PARTITION	date
PARTIES	20220202
Lg CONTEXTE	10
Fq MAX	5
Co-Freq	2
SR FQMIN	10
RELATION	REL
Pôle SOURCE	entrez la forme pôle
PARTIES	20220202
SEUIL	5
SR LGMAX	12
ANNOTATION SORTIE	1:Forme
ANNOTATION*	1:Forme
GRAPHE H	400
GRAPHE L	800
NB SÉLECTION SECTION	1
ANNOTATION RELATION	4

Paramètres

Cliquer sur le + pour dérouler le menu

iTrameur Analyse textométrique de données

Paramètres -

PARTITION	date ▼	PARTIES	20200226 ▼	LG CONTEXTE	10	Fq Max	5
GRAPHE H	400	SEUIL	5	Co-FREQ	2	INDSPMIN	5
GRAPHE L	800	SR LGMAX	12	SR FQMIN	10	NB SÉLECTION SECTION	1
ANNOTATION*	1:Forme ▼	ANNOTATION SORTIE	1:Forme ▼	RELATION	REL	ANNOTATION RELATION	4

PÔLE SOURCE entrez la forme pôle

- **Partition** : critère utilisé pour diviser un corpus
- **Parties** : morceaux de corpus obtenus en appliquant une partition

Exemples :

- la partition *date* permet d'avoir une partie pour chaque *aspiration* (1 aspiration = 1 date à laquelle on a aspiré les Unes de 70 journaux)
- la partition *page* permet d'avoir une partie pour chaque *journal*, quelle que soit la date

Paramètres

→ Plus d'informations dans l'onglet "Aide"

DÉLIMITEUR DE CONTEXTE : Cette zone de saisie doit contenir le caractère utilisé pour la construction de la *Carte des Sections* (et aussi pour déterminer les contextes utilisés pour le calcul des cooccurents via la *Carte des Sections*).

DÉLIMITEUR(S) : Cette zone de saisie contient la liste des caractères délimiteurs utilisés pour segmenter le texte en formes graphiques.

BI-TEXTE : Cette case à cocher permet de charger un bi-texte aligné (et les fonctionnalités associées).

DÉPENDANCE : Cette case à cocher permet de charger les fonctionnalités associées aux traitements sur une base contenant des annotations en dépendance.

PARTITION : Liste permettant de sélectionner une partition.

PARTIE : Liste permettant de sélectionner une partie de la partition choisie.

ANNOTATION : Cette liste, mise à jour à l'issue du chargement d'une base annotée, permet de sélectionner une annotation pour réaliser le calcul visé.

ANNOTATION SORTIE : Cette liste, mise à jour à l'issue du chargement d'une base annotée, permet de sélectionner l'annotation à utiliser pour afficher les zones textuelles (concordance, section, contexte cooccurentielle) quelle que soit l'annotation utilisée. Par exemple : concordance du pôle NOM (annotation n°3) et affichage en sortie via les formes graphiques (annotation n°1) des contextes visés. Par défaut l'annotation en sortie a la même valeur que l'annotation sélectionnée pour les calculs.

SEUIL : Par défaut, l'indice de spécificité est calculé avec un seuil de probabilité fixé à 5 %.

CO-FREQ : Par défaut, le calcul de cooccurrence est calculé en ne retenant que les candidats cooccurents dont la co-fréquence est supérieure à la valeur donnée.

INDSPMIN : Par défaut, le calcul de cooccurrence est calculé en ne retenant que les candidats cooccurents dont l'indice de spécificité est supérieur à la valeur donnée.

FQ MAX : Par défaut, le calcul des spécificités totales est calculé en ne retenant que les formes dont la fréquence est supérieure à la valeur donnée. Idem pour le calcul du "Réseau de cooccurents".

LG CONTEXTE : Longueur du contexte pour l'affichage d'une concordance.

GRAPHE H : Par défaut, les graphiques construits ont une hauteur correspondant à la valeur donnée.

GRAPHE L : Par défaut, les graphiques construits ont une largeur correspondant à la valeur donnée.

PÔLE : Zone de saisie utilisée pour définir le pôle visé (remplissage par auto-complétion).

NB SÉLECTION SECTION : Zone de saisie permettant de définir le nombre de sections à sélectionner simultanément (*via* le clic-droit).

RELATION : Zone de saisie permettant de saisir le nom d'une relation.

ANNOTATION RELATION : Zone de saisie permettant le numéro d'annotation portant le nom de la relation visée.

Statistiques sur le corpus : PCLC

= Principales Caractéristiques
Lexicographiques du Corpus

Chargement Trame **Cadre** SR/Patron Section Coocs Bi-Texte Dépendance Sélection Export Aide

Opérations sur le Cadre

Cadre Parties PCLC Ventilation*

Spécifs-partie* Spécifs totales* Mots Spécifs+ TGF+BT+VN

PARTITION 1 date PARTITION 2 date Croisement Partitions

P.C.L.C

Copy CSV Excel PDF Print

Recherche :

Partie	F occ	f typ	Forme max	Max
20220202	150880	17025	de	6423
20220206	145118	16780	de	5833
20220209	151777	17321	de	6526
20220213	145902	16834	de	6100
20220216	152399	16943	de	6360
20220220	145522	16731	de	6092
20220223	151852	17064	de	6758
20220227	145263	16626	de	5863
20220306	146218	16601	de	5983
20220309	152207	16854	de	6439

Affichage de 1 à 10 des 19 items

Préc. 1 2 Suiv.

Calcul qui permet d'obtenir pour chaque partie

(selon la partition choisie) :

- le nombre de formes totales
- le nombre de formes différentes
- la forme la plus fréquente
- le nombre d'occurrences de cette forme

Statistiques sur le corpus : PCLC

- Regarder les PCLC pour la partition page permet de vérifier qu'il n'y a pas de résultats étranges et donc des erreurs (→ correction/nettoyage du corpus)
- Dans la plupart des cas, la forme la plus fréquente est un « mot-outil » → “de”, “à”, “le”, etc. (mots les plus fréquents en français, même s'ils ne sont pas réellement des vecteurs de sens)

P.C.L.C

Copy CSV Excel PDF Print

Recherche :

Partie ▲	F occ	f typ	Forme max	Max
bienpublic.com	106899	10904	de	3962
centre-presse.fr	44357	4244	Groupe	1634
charentelibre.com	21303	4150	de	987
corsematin.com	33818	3217	de	1286
dna.fr	117683	13267	de	4345
dordogne.com	16718	1353	de	759
estrepublicain.fr	190690	17235	de	8700
humanite.fr	60109	7366	de	2420
jhm.fr	43724	5834	de	1647
la-croix.com	44569	6051	de	1422

Dictionnaire

- Permet de **vérifier l'existence d'un mot** dans le corpus.
- Affiche les items du corpus selon un **classement du plus au moins fréquent**.
- Il est possible de sélectionner ou projeter chaque item sur une **concordance**, une **ventilation** ou une **carte des sections**.

Dictionnaire

Copy CSV Excel PDF Print

Recherche : Paris

Item	Fq	Concordance	Ventilation	Carte	Sélection
Paris	2426				
Parisien	202				
parisiennes	66				
parisienne	46				
paris	45				
parisien	41				
parisiens	32				
leparisien	26				
Parisiens	21				
ParisJob	19				

Affichage de 1 à 10 des 28 items (filtrage à partir des 72,633 items) Préc. 1 2 3 Suiv.

Concordance

Les contextes d'apparition du « mot » recherché : le retour au texte

Afficher items

Recherche :

N°	Partie	Contexte Gauche	Pôle	Contexte Droit
9	20220202	reste de son hôtel particulier de la rue de Grenelle à	Paris	et du château du Jonchet, dans l'Eure-et-Loir. Une
32	20220202	Estaing apposée à l'Assemblée nationale 13:00 Obsèques vendredi à	Paris	du couturier Thierry Mugler Voir toutes les dépêches Monde Analyse JO
33	20220202	Joe Biden Vladimir Poutine Union européenne Proche-Orient Eco Entreprises Médias	Paris	Culture Séries Cinéma Musique Livres Gastronomie Sport Football Tennis Cyclisme Rugby
34	20220202	Joe Biden Vladimir Poutine Union européenne Proche-Orient Eco Entreprises Médias	Paris	Culture Séries Cinéma Musique Livres Gastronomie Sport Football Tennis Cyclisme Rugby
38	20220202	Miyawaki", du nom d'un botaniste japonais, sortent de terre à	Paris	et dans sa région. Ces... ABONNÉS Arnaud Ngatcha, adjoint à la
39	20220202	ABONNÉS Arnaud Ngatcha, adjoint à la mairie chargé de l'international : "	Paris	a un rôle à jouer en Europe" La Ville veut se
58	20220202	toulon Sports Rugby "Une très belle performance, une joie de battre	Paris	..." Christophe Galtier réagit après la victoire de l'OGC Nice en
59	20220202	Jeux Olympiques Pékin 2022 - JO d'hiver Nos 24 paris pour	Paris	Sports départementaux Isère Haute-Savoie Savoie Drôme Ardèche Vaucluse Hautes-Alpes
60	20220202	Jeux Olympiques Pékin 2022 - JO d'hiver Nos 24 paris pour	Paris	Sports départementaux Isère Haute-Savoie Savoie Drôme Ardèche Vaucluse Hautes-Alpes
62	20220202	pied Route du Louvre Calendrier des courses Cyclisme Tour de France	Paris	-Roubaix Quatre jours de Dunkerque Natation Rugby Tennis Volley-ball Enduropale

Affichage de 1 à 10 des 302 items (filtrage à partir des 2,426 items)

Préc.

1

2

3

4

5

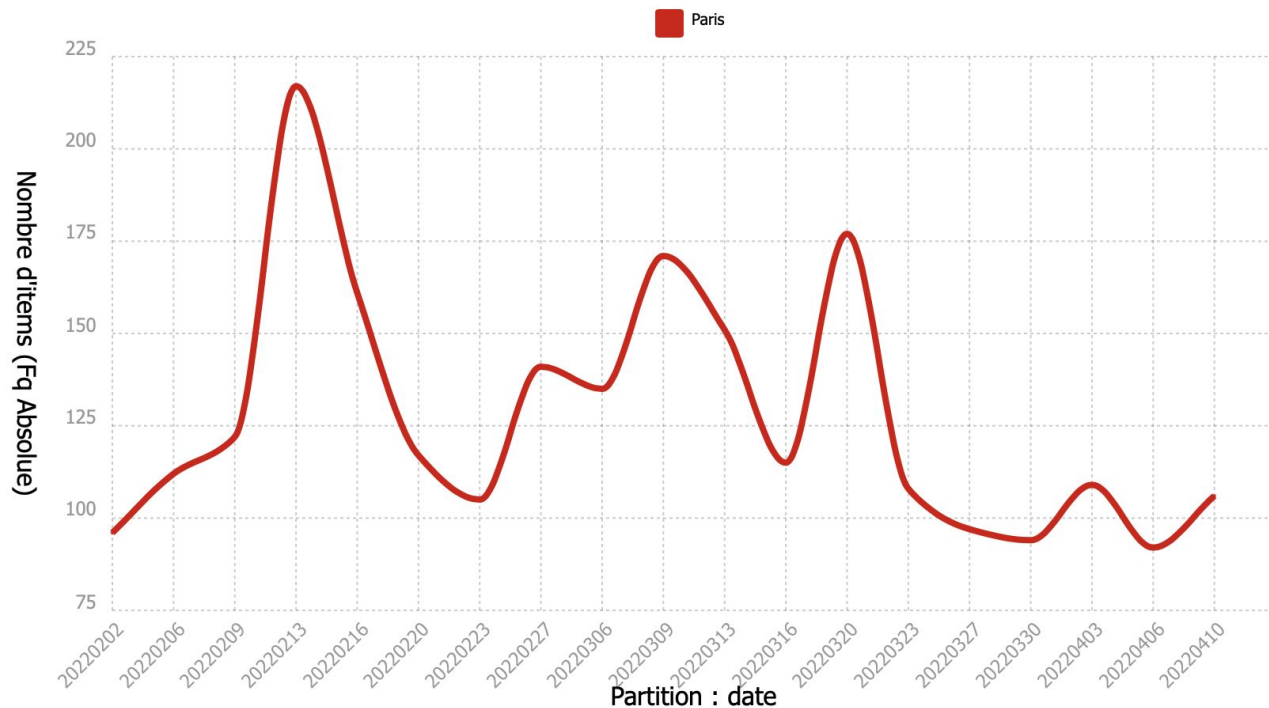
...

31

Suiv.

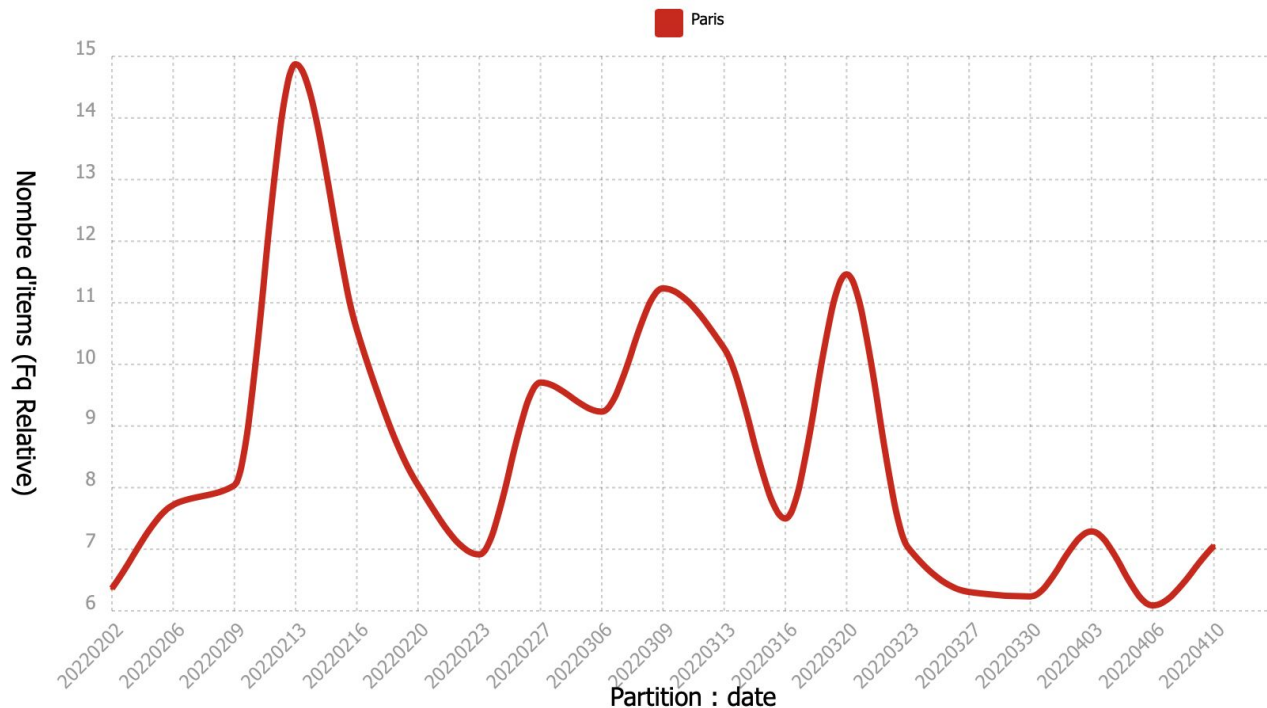
Ventilation : Fq Absolue

Le nombre d'occurrences des « mots » dans les parties



Ventilation : Fq Relative

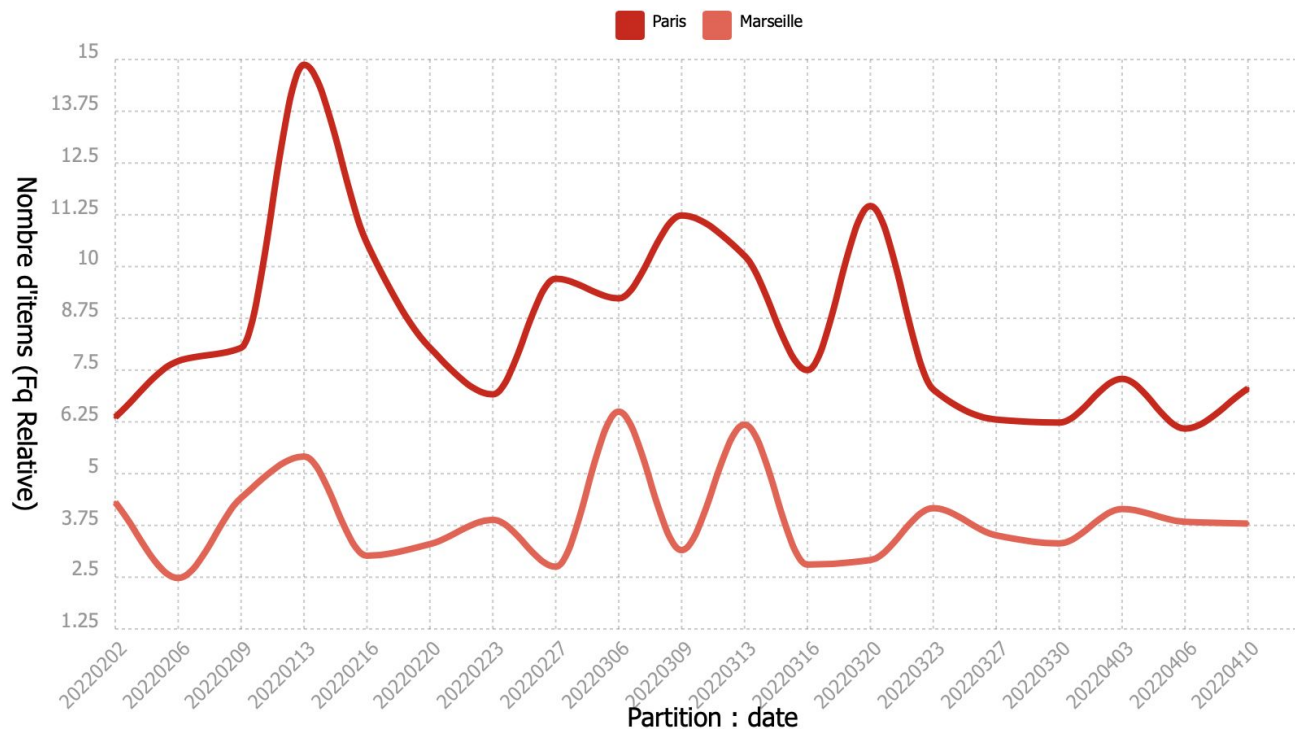
Le nombre d'occurrences des « mots » dans les parties mis en rapport avec la taille des parties



Ventilation

Comparaison de fréquences

Paramètres			
PARTITION	date	PARTIES	20220202
GRAPHIE H	400	SEUIL	5
GRAPHIE L	800	SR LGMAX	12
ANNOTATION*	1-Forme	ANNOTATION SORTIE	1-Forme
		Pôle Source	Paris Marseille
		LG CONTEXTE	10
		Co-Freq	2
		SR FQMIN	10
		RELATION	REL
		Fq Max	5
		InfoSPHIN	5
		NB SÉLECTION SECTION	1
		ANNOTATION RELATION	4



Ventilation : spécificités (prochain cours)

Partition en dates

