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 **infirmées** 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

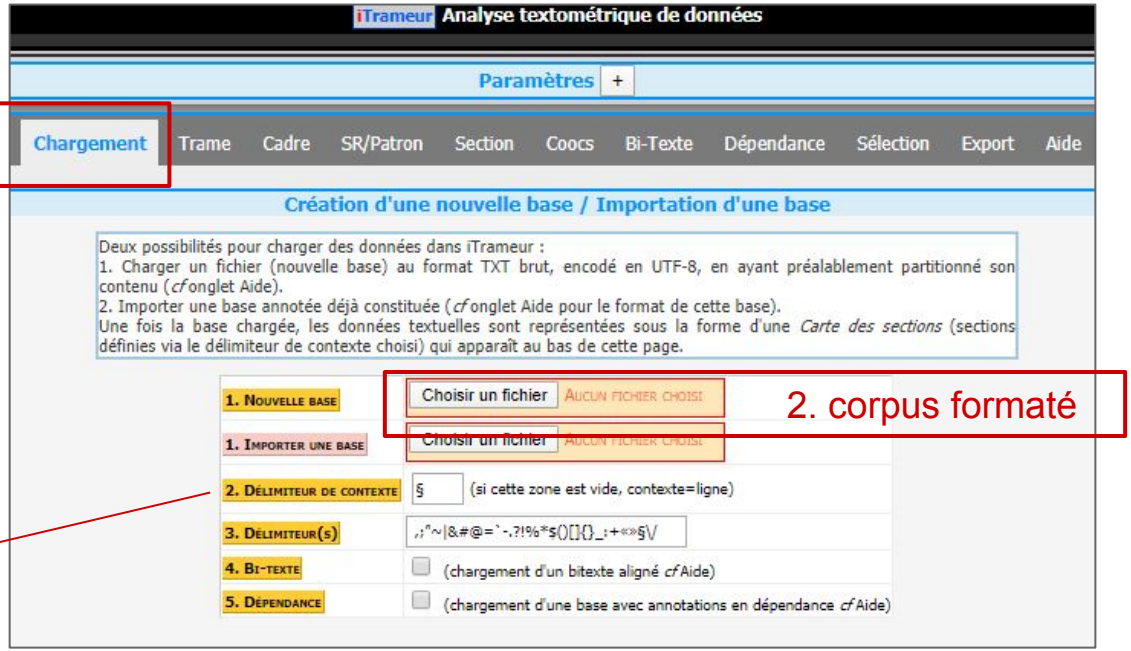
Pour commencer…

→ 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



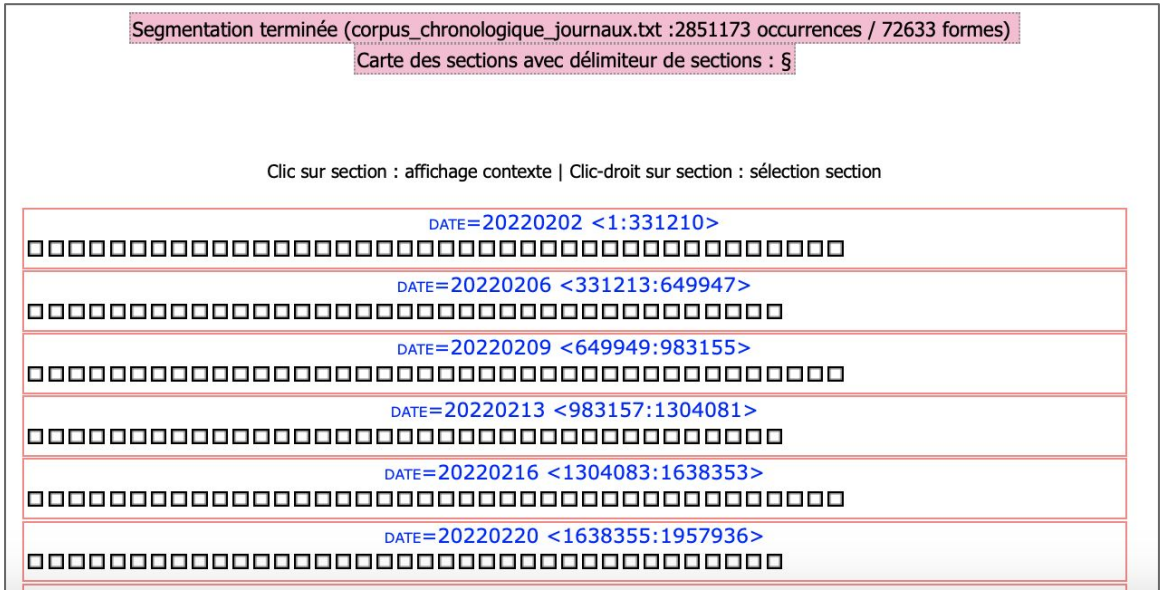
Entre 1. et 2. : vérifier le délimiteur de sections

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

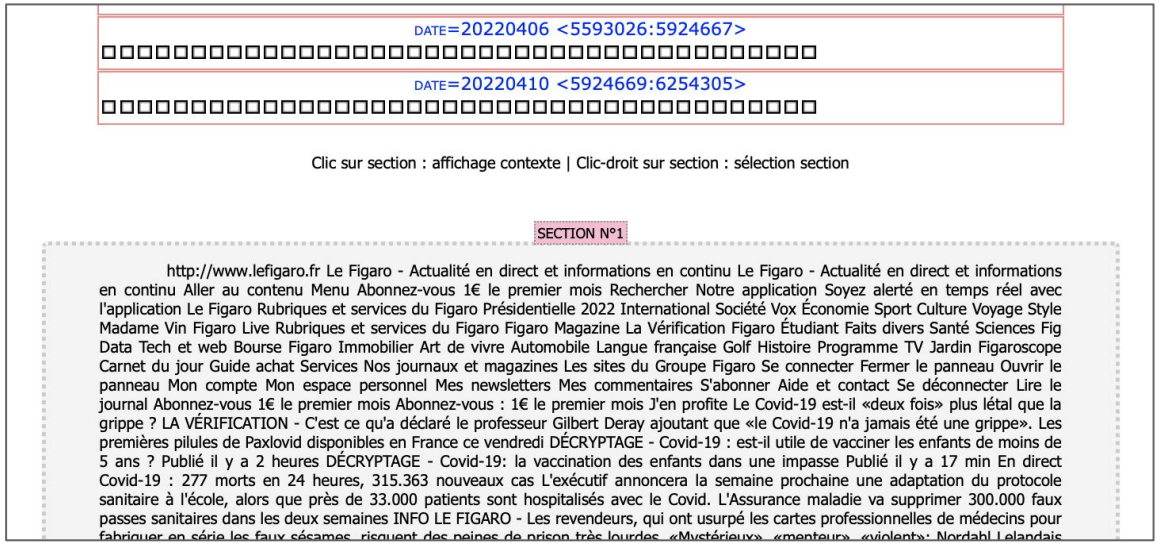


Visualisation du corpus de test

Carte des sections

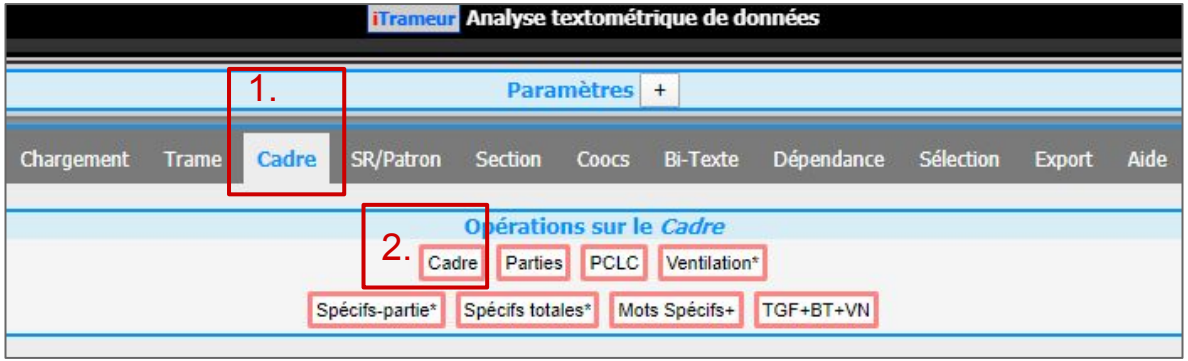
- Quand on clique sur un carré, le texte correspondant à cette section s’affiche en dessous.
- Vérifier en parcourant quelques sections que tout est normal (pas de sections vides, pas de messages d’erreurs).

-



Visualisation du corpus

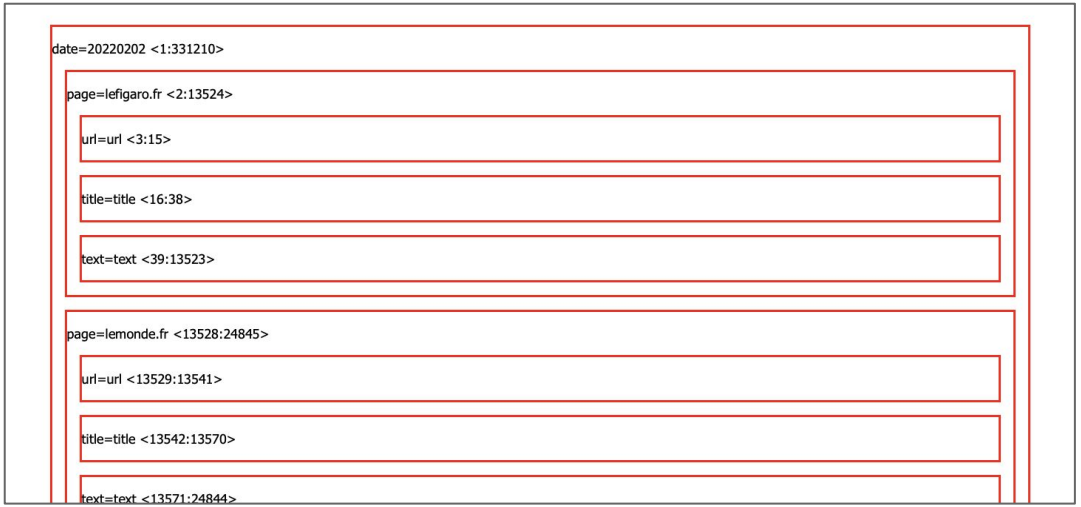
Le cadre



Visualisation du corpus

Le cadre

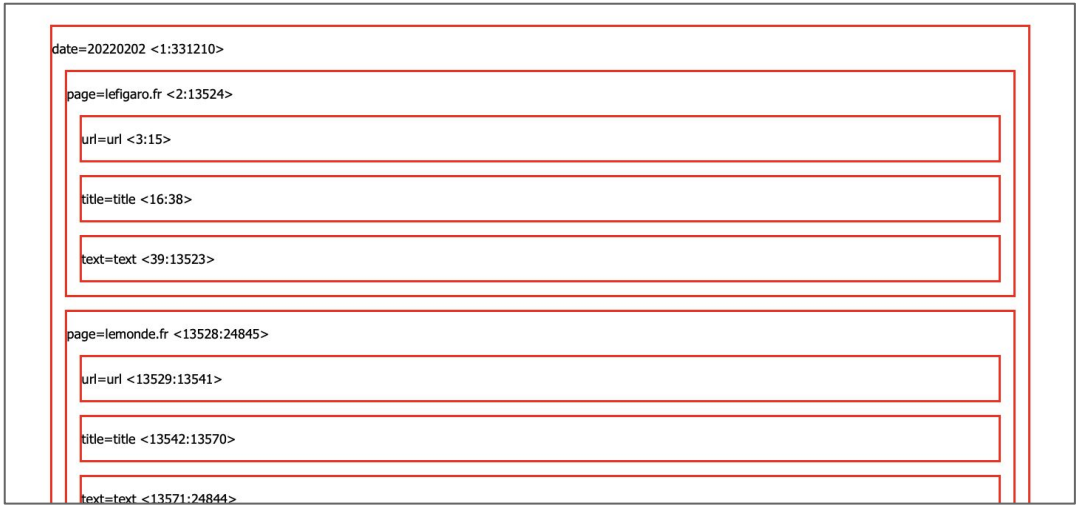
- Parties définies par les balises du corpus (date, page, url, title, text).
- Vérifier que les balises sont imbriquées correctement.
- Vérifier que les bordures des rectangles représentant les parties ne se croisent pas.

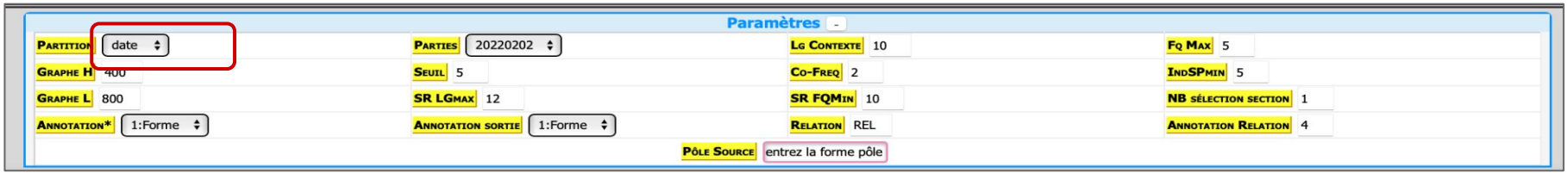


Visualisation du corpus

Le cadre

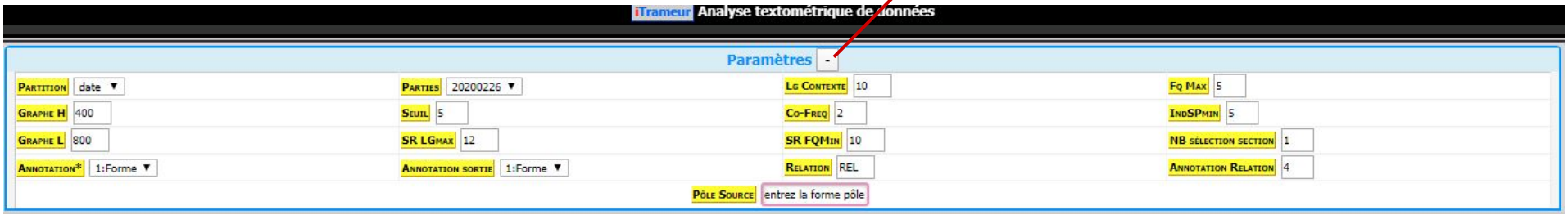
- Si l'affichage est différent : sûrement un problème de balises
- Dans l'onglet « Paramètres », regarder ce qui s'affiche (le nom des balises) dans le menu « Partition »





Paramètres

Cliquer sur le + pour dérouler le menu



- **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 cooccurrents 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 le 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 cooccurrents 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 cooccurrents 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 cooccurrents".

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).

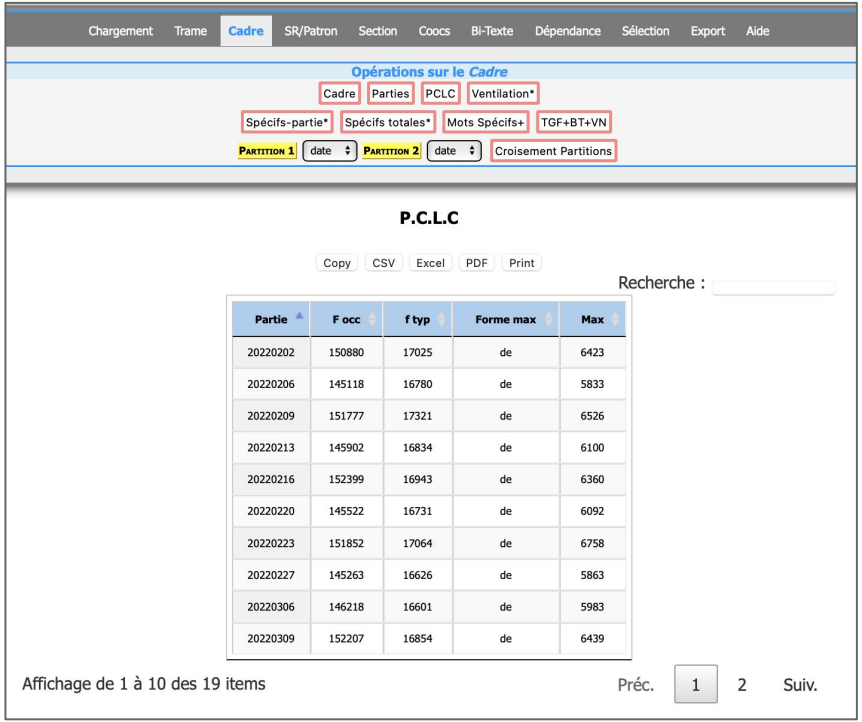
NB sélection section : Zone de saisie permettant de définir le nombre de sections à sélectionner simultanément (*via* le clic-droit).

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



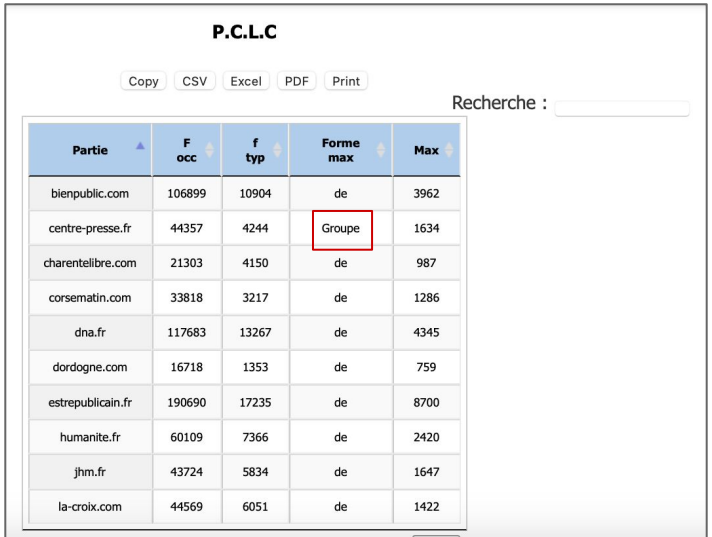
Calcul qui permet d'obtenir pour chaque partie

(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)



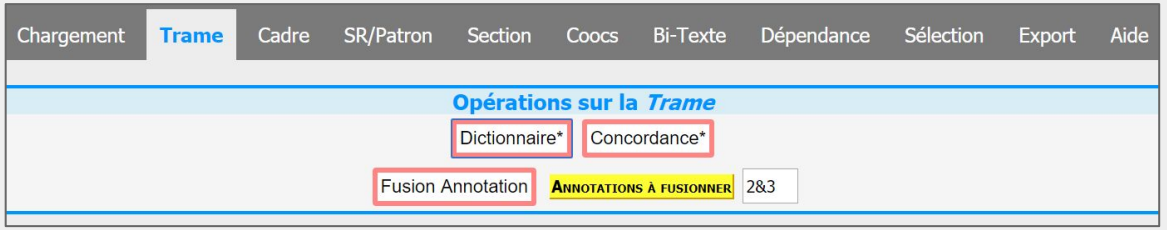
P.C.L.C

Copy CSV Excel PDF Print

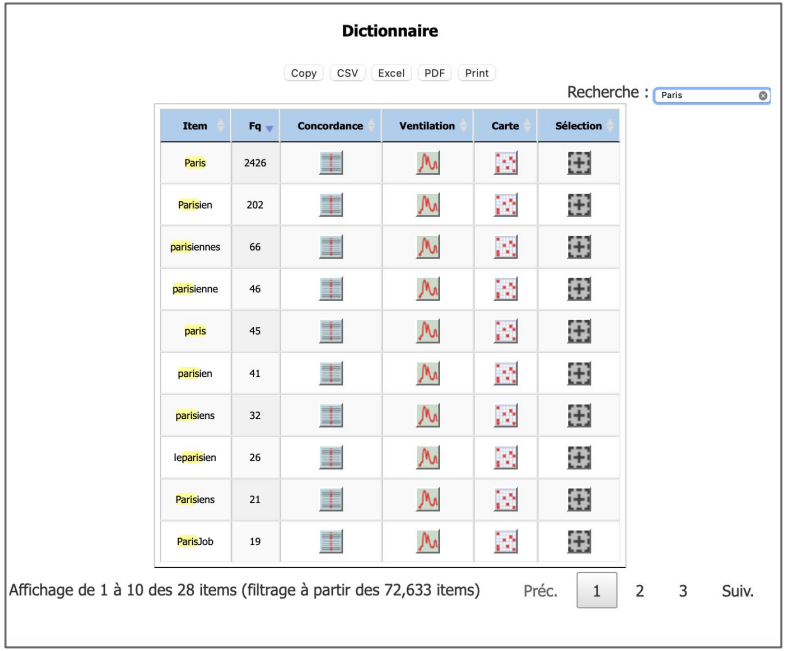
Recherche :

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

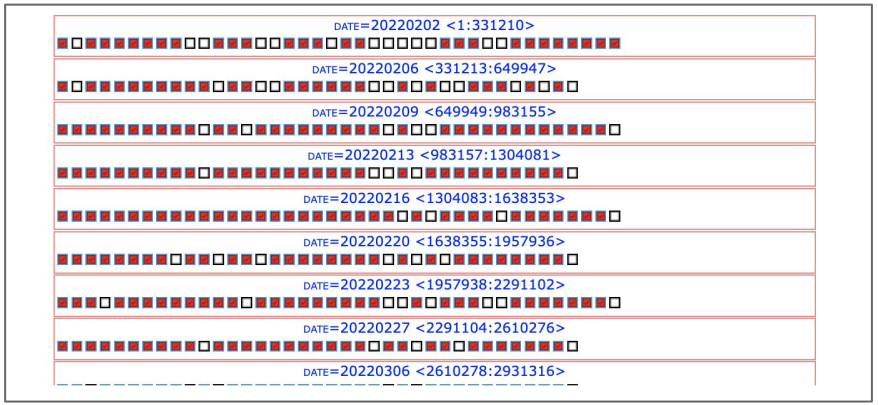
Dictionnaire



- 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.**



Carte de sections à partir du dictionnaire



- Carrés en rouge = pages qui contiennent la forme « Paris » dans chaque aspiration.
- Si on clique sur un carré : on obtient le texte de la section correspondante avec la forme recherchée surlignée en rouge

SECTION N°40

http://limportant.fr L'Important : l'info par les réseaux sociaux L'Important : l'info par les réseaux sociaux #L'info par les réseaux sociaux Connexion Inscription #L'info par les réseaux sociaux Inscription Connexion Politique Eco Monde MO Afrique Société Santé Tech Culture Planète Régions Politique Eco Monde MO Afrique Société Santé Tech Culture Planète Régions Tribunes Reportages Sur l'écran d'accueil ABONNEZ-VOUS À LA NEWSLETTER! fr.style.yah fr.style.yah 1 femme sur 3 humiliée par son partenaire à cause de ses règles @chloethibaud fr.style.yah fr.style.yah 1 femme sur 3 humiliée par son partenaire à cause de ses règles @chloethibaud bfmtv.com bfmtv.com Grand **Paris**: les nouvelles restrictions pour les véhicules polluants repoussées à 2023 @BFMTV lavoixdunord lavoixdunord Grippe aviaire : près de 3 millions de volailles abattues depuis fin novembre @MarieMarysienka wan-ifra.org wan-ifra.org 41% des femmes travaillant dans les médias victimes d'harcèlement sexuel sur leur lieu de travail @metamedia pinknews.co. pinknews.co. Iran :

Concordance

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

Afficher 10 items

Recherche : JO

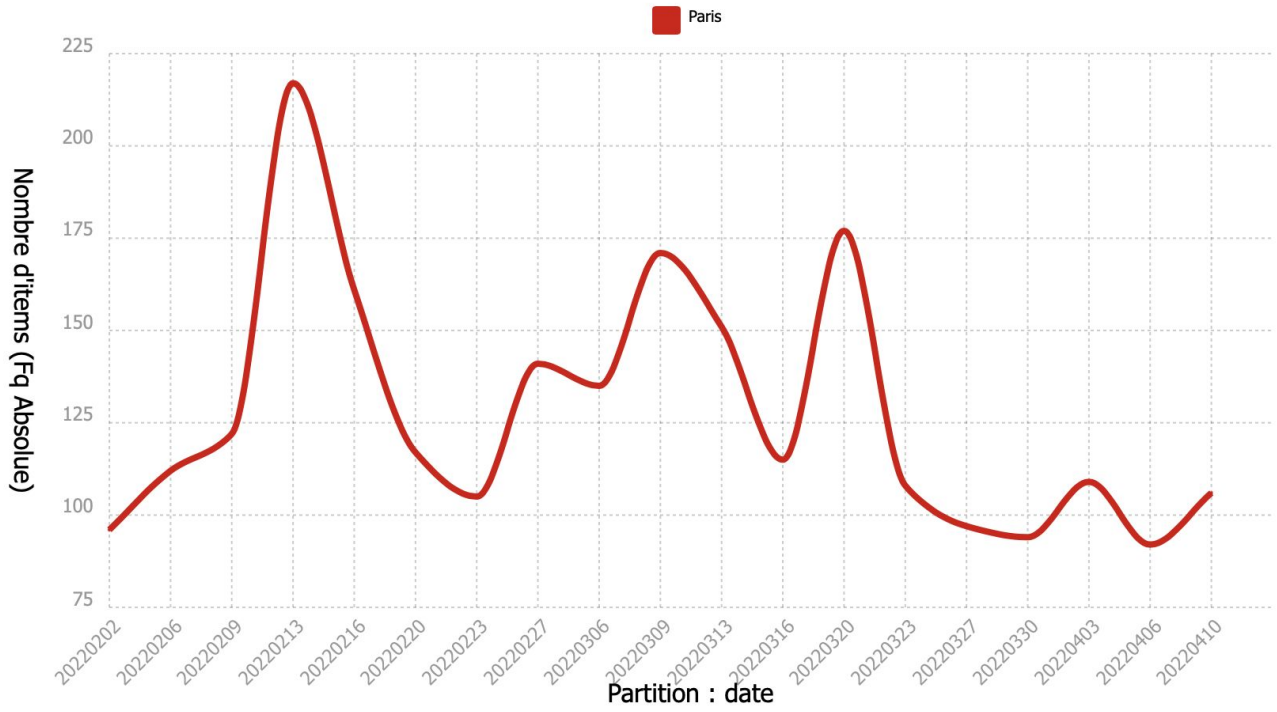
N°	Partie	Contexte Gauche	Pôle	Contexte Droit
9	20220202	reste de son hôtel particulier de la rue de Grenelle à	Paris	et du château du Jonchet, dans l'Eure-et-Loir. Une
32	20220202	Estaing apposée à l'Assemblée nationale 13:00 Obsèques vendredi à	Paris	du couturier Thierry Mugler Voir toutes les dépêches 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.

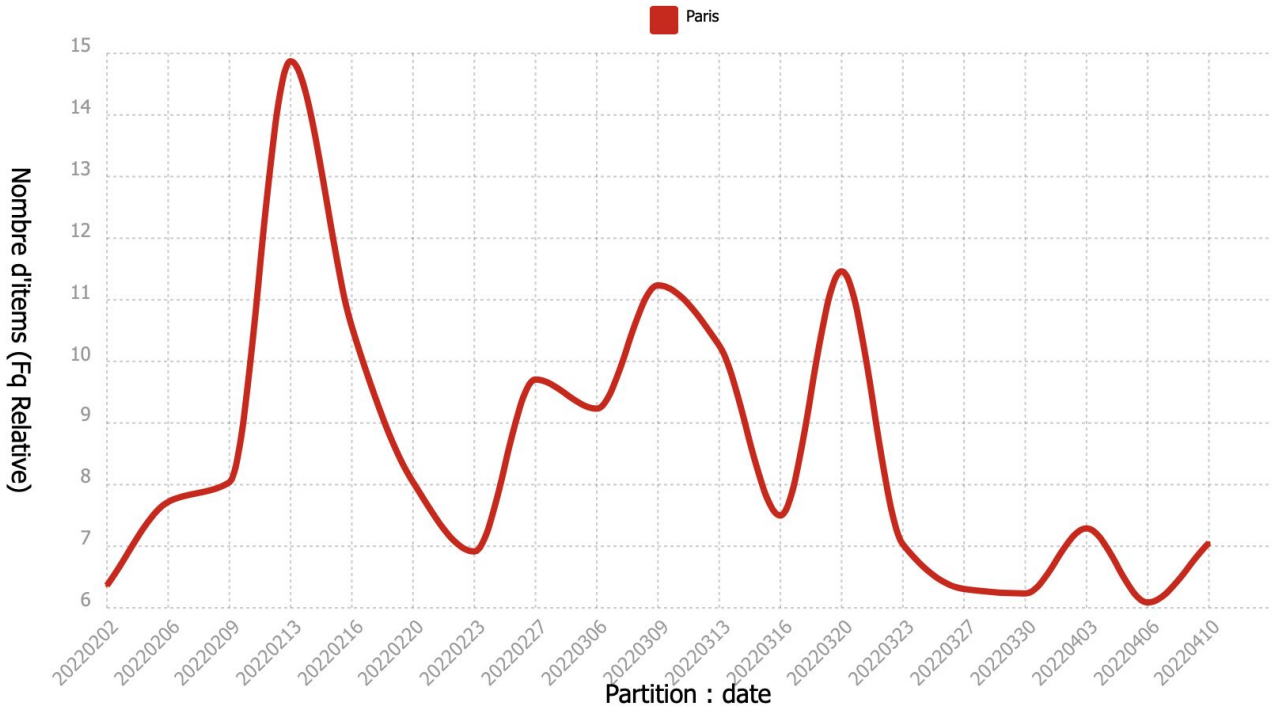
Ventilation : Fq Absolue

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



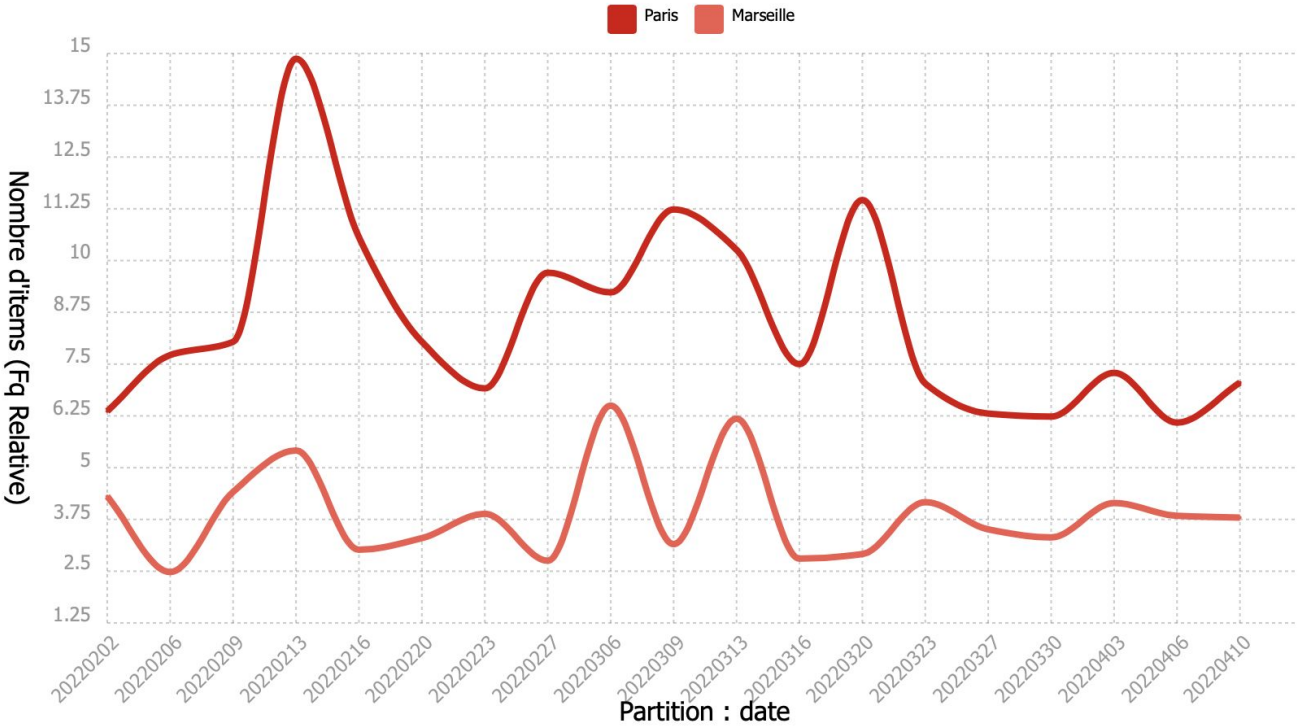
Ventilation : Fq Relative

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



Ventilation

Comparaison de fréquences



Ventilation : spécificités (prochain cours)

Partition en dates