# Analyse du discours

*Licence 2 : Lettres modernes*

Séance 4 : Spécificités

marine.delaborde@cyu.fr

CERGY PARIS UNIVERSITÉ

# Organisation des séances

4 / 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

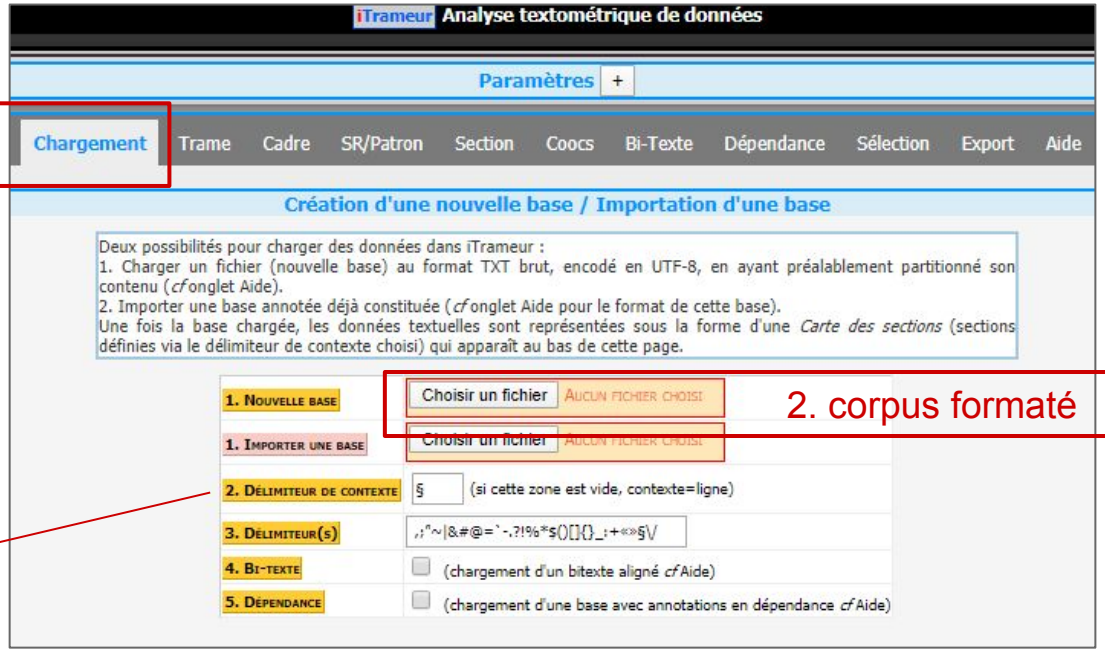
Pour commencer…

→ 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`

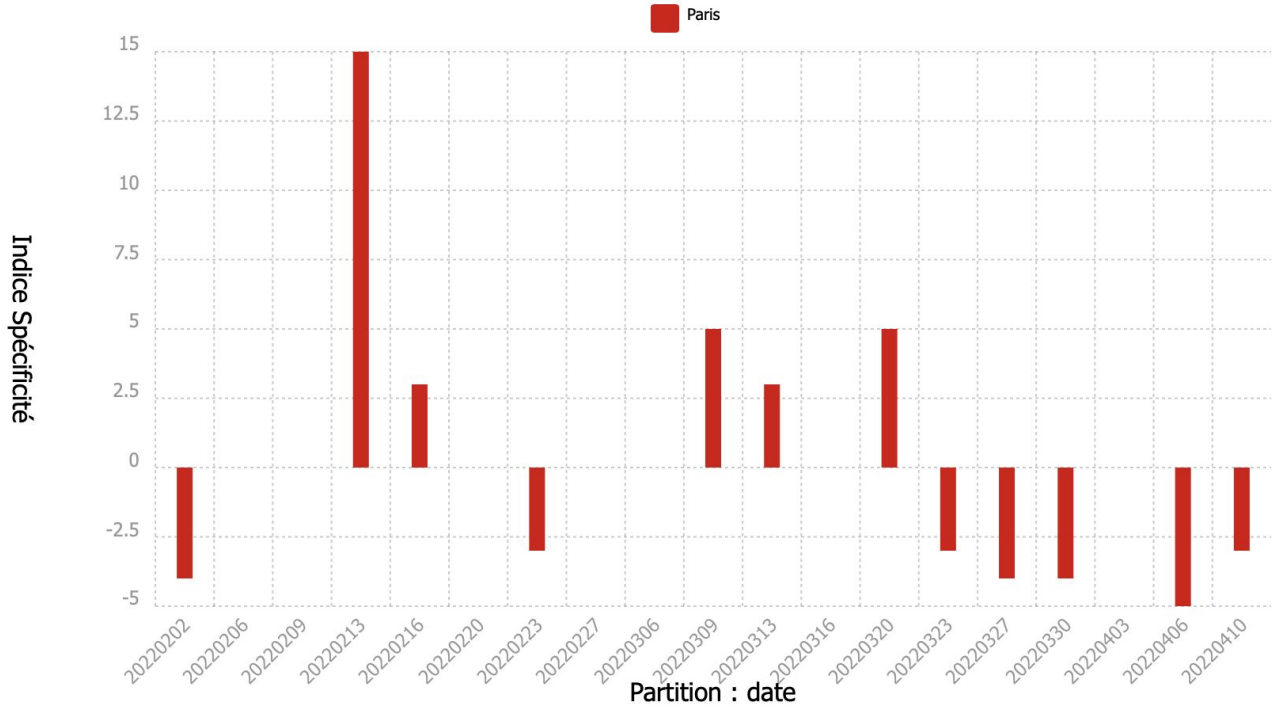
# Chargement de la base (rappel)




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

# Ventilation : spécificités (rappel)

## Partition en dates



# Les spécificités


- Métrique permettant de donner un aperçu global du vocabulaire surreprésenté ou sous-représenté dans un ensemble d'items.
    - Valeur positive = **surreprésentation** = statistiquement anormal que cet item soit aussi fréquent dans cet ensemble.
    - Valeur négative = **sous-représentation** = statistiquement étonnant que cet item soit aussi peu fréquent dans cet ensemble.

# Les spécificités

## Le vocabulaire spécifique à une partie

- Un vocabulaire **spécifique à une date** en particulier tous journaux confondus

  = sûrement lié à **l'actualité**

- Un vocabulaire **spécifique à un journal** en particulier toutes dates confondues

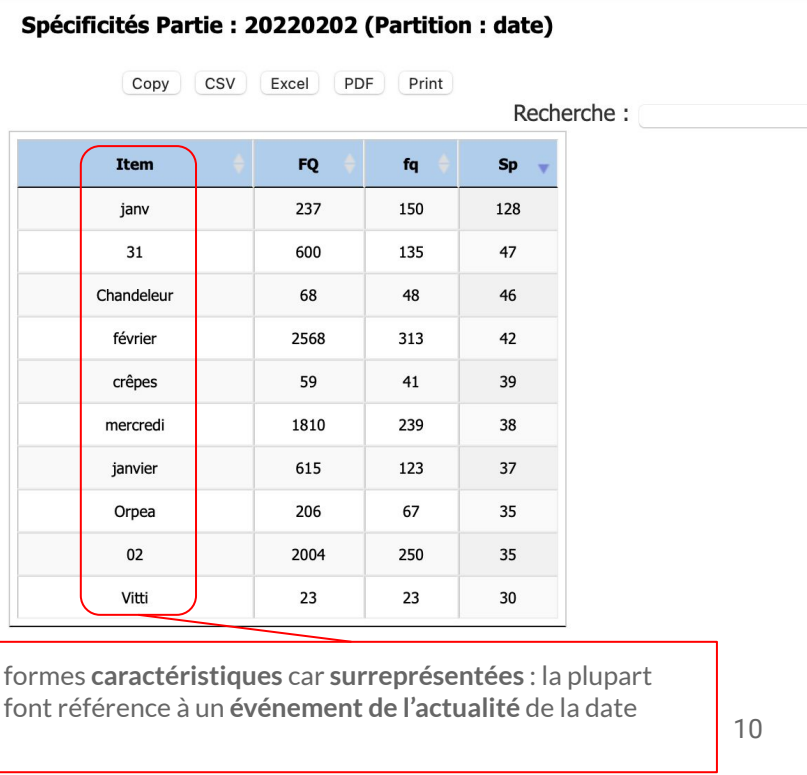
  = sûrement lié à sa **ligne directrice**

- Cette fonctionnalité permet aussi de faire un **tri sur le contenu aspiré** : représentation des éléments indésirables et des caractères mal encodés.

# Les spécificités

## Le vocabulaire spécifique à une partie

- **Spécificités par partie**
  - Calcule le **vocabulaire spécifique de la partie visée** (sélectionnée dans les paramètres) par rapport aux autres parties (attention, bien choisir la partition : choisir une partition en pages pour pouvoir sélectionner un journal en tant que partie).
  - Tableau : informations de fréquence totale et de spécificité des items présents dans la partie.
  - Si un item est **surreprésenté**, la forme est dite « **caractéristique** » de la partie car sa fréquence est anormalement élevée par rapport au reste du corpus.
  - On peut **rechercher** des items dans le tableau et **trier** en fonction des fréquences/de l'ordre alphabétique des items.



**Spécificités Partie : 20220202 (Partition : date)**

Copy CSV Excel PDF Print

Recherche :

| Item | FQ | fq | Sp |
|---|---|---|---|
| janv | 237 | 150 | 128 |
| 31 | 600 | 135 | 47 |
| Chandeleur | 68 | 48 | 46 |
| février | 2568 | 313 | 42 |
| crêpes | 59 | 41 | 39 |
| mercredi | 1810 | 239 | 38 |
| janvier | 615 | 123 | 37 |
| Orpea | 206 | 67 | 35 |
| 02 | 2004 | 250 | 35 |
| Vitti | 23 | 23 | 30 |

formes **caractéristiques** car **surreprésentées** : la plupart font référence à un **événement de l'actualité** de la date

# Les spécificités

## Le vocabulaire spécifique à une partie

- **Spécificités totales**
  - Calcule les **spécificités pour chaque partie en fonction de la partition sélectionnée** dans les paramètres
  - Tableau : affichage pour chaque item de sa fréquence totale dans le corpus, de sa fréquence et de son indice de spécificité dans chaque partie
  - On peut **rechercher** des items dans le tableau et **trier** en fonction des fréquences/spécificités/ordre alphabétique des items

**Tableau Général des Items** (FQ > 5 | annotation:1)
**Partition : date**

Copy CSV Excel PDF Print

Recherche :

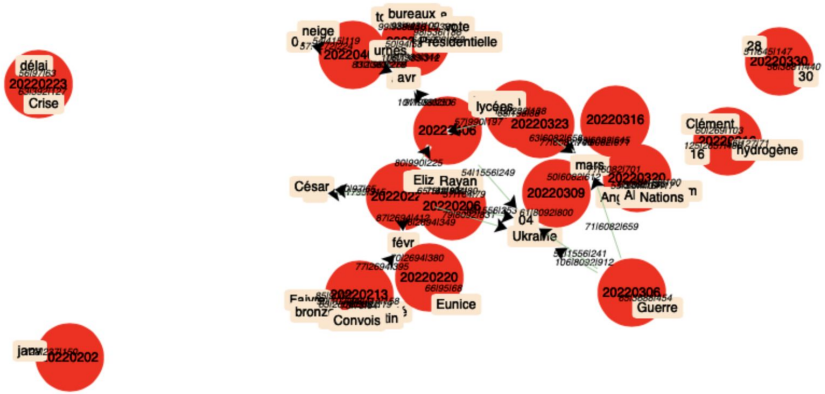
| Item | FQ | 20220202 / fq | 20220202 / sp | 20220206 / fq | 20220206 / sp | 20220209 / fq | 20220209 / sp | 20220213 / fq |
|---|---|---|---|---|---|---|---|---|
| Ukraine | 8092 | 41 | -133 | 38 | -129 | 37 | -138 | 133 |
| mars | 6082 | 27 | -104 | 27 | -99 | 53 | -81 | 26 |
| Guerre | 3688 | 2 | -84 | 5 | -75 | 3 | -82 | 4 |
| Hier | 8112 | 209 | -34 | 546 | 11 | 161 | -53 | 517 |
| 03 | 1938 | 16 | -28 | 88 | 0 | 13 | -30 | 35 |
| samedi | 2668 | 40 | -25 | 192 | 7 | 49 | -21 | 308 |
| avr | 989 | 1 | -23 | 0 | -23 | 0 | -24 | 1 |
| guerre | 1975 | 29 | -19 | 19 | -24 | 18 | -27 | 18 |
| Russie | 1486 | 19 | -17 | 26 | -12 | 11 | -23 | 58 |
| avril | 1383 | 16 | -17 | 8 | -22 | 26 | -11 | 11 |

Affichage de 1 à 10 des 23,499 items  Préc. 1 2 3 4 5 … 2350 Suiv.

# Les spécificités

## Le vocabulaire spécifique à une partie

- **Vocabulaire spécifique positif par partie**
  - Trouve les **mots de spécificité positive d'une partition**
  - Résultats présentés sous forme de tableau et de graphique : pour chaque partie on associe ses mots spécifiques
  - Il faut parfois **augmenter le seuil** de spécificité positive retenu (*IndSPmin - ic = 50*) pour **limiter le nombre d'unités à afficher**




**Les mots spécifiques de la partition : date**

Copy CSV Excel PDF Print

Recherche :

| Item | Partie | FQ | fq | Sp |
|---|---|---|---|---|
| janv | 20220202 | 237 | 150 | 128 |
| reine | 20220206 | 144 | 81 | 65 |
| 04 | 20220206 | 1556 | 253 | 59 |
| Elizabeth | 20220206 | 119 | 81 | 75 |
| II | 20220206 | 164 | 79 | 57 |
| févr | 20220206 | 2694 | 349 | 56 |
| Rayan | 20220206 | 104 | 80 | 82 |
| 11 | 20220213 | 2609 | 358 | 63 |
| liberté | 20220213 | 414 | 158 | 93 |
| Valentin | 20220213 | 312 | 119 | 70 |

Affichage de 1 à 10 des 61 items Préc. 1 2 3 4 5 6 7 Suiv.