

Analyse du discours

Licence 2 : Lettres modernes

Séance 5 : Cooccurents et segments répétés

marine.delaborde@cyu.fr



CERGY PARIS
UNIVERSITÉ

Organisation des séances

5 / 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 : ??/??/23 (salle 121)	Ventilation et concordance
Séance 4 : ??/??/23 (salle 121)	Spécificités
Séance 5 : ??/??/23 (salle 121)	Cooccurents et segments répétés
Séance 6 : ??/??/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 la thématique abordée dans les 10 articles.
 - Identifier des hypothèses qui seront confirmées ou infirmées par les données.
 - Si les mesures permettent d'infirmar une hypothèse, c'est quand même intéressant !
 - Ne pas modifier les hypothèses formulées au début : prise de recul pour comprendre les résultats.

iTrameur

Pour commencer...

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

→ Un corpus :

- Un corpus de test : corpus_chronologique_journaux.txt
- Votre corpus si possible

Chargement de la base (rappel)

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 (*cf*onglet Aide).
2. Importer une base annotée déjà constituée (*cf*onglet 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 <small>AUCUN FICHIER CHOISI</small>
1. IMPORTER UNE BASE	Choisir un fichier <small>AUCUN FICHIER CHOISI</small>
2. DELIMITEUR DE CONTEXTE	\$ (si cette zone est vide, contexte=ligne)
3. DELIMITEUR(s)	.,!~ &#@='-.?!%*\$(){}_!+«»\$V
4. BI-TEXTE	<input type="checkbox"/> (chargement d'un bitexte aligné <i>cf</i> Aide)
5. DEPENDANCE	<input type="checkbox"/> (chargement d'une base avec annotations en dépendance <i>cf</i> Aide)

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

2. corpus formaté



Les spécificités

Rappels

- 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

Rappels

- 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.

Segments répétés

= suite de formes dont la fréquence est égale ou supérieure à 2 dans le corpus

Paramètres -

PARTITION date ▼	PARTIES 20200128 ▼	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

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

Opérations sur les Segments Répétés

SR

Segments répétés* Carte Sections(SR)* Ventilation(SR)* Concordance(SR)*

1. Paramétrage par défaut

2. Bouton "Segments répétés"































Segments répétés

- On obtient tous les **segments répétés** du corpus triés par **fréquence absolue**, avec l'information de **longueur du segment**.
- On peut explorer les résultats et lancer des **opérations** sur les items (concordance, ventilation, carte des sections).

Segments Répétés

Copy CSV Excel PDF Print

Recherche :

SR	Fq	Lg	Concordance	Ventilation	Carte
de la	16907	2			
de l	12289	2			
Hier à	8059	2			
à la	7335	2			
à l	6065	2			
Toute l	5456	2			
en Ukraine	5102	2			
Faits divers	4231	2			
dans le	3937	2			
Présidentielle 2022	3883	2			

Affichage de 1 à 10 des 152,146 items

Préc. 1 2 3 4 5 ... 15215 Suiv.

Segments répétés

- Possibilité de **rechercher le segment** que l'on veut et de projeter les résultats sur la **carte des sections**, une **ventilation** ou un **concordancier**.

The screenshot shows the 'Opérations sur les Segments Répétés' window. At the top is a menu bar with 'Chargement', 'Trame', 'Cadre', 'SR/Patron' (highlighted), 'Section', 'Coocs', 'Bi-Texte', 'Dépendance', 'Sélection', 'Export', and 'Aide'. Below the menu bar is a toolbar with a trash icon, a search input field with a yellow 'SR' label, and four buttons: 'Segments répétés*', 'Carte Sections(SR)*', 'Ventilation(SR)*', and 'Concordance(SR)*'. Two red boxes with arrows provide annotations: Box 1 points to the search field with the text '1. Champ de recherche des segments répétés'. Box 2 points to the three operation buttons with the text '2. Opérations possibles pour projeter les résultats'.

Opérations sur les *Segments Répétés*

SR

Segments répétés* Carte Sections(SR)* Ventilation(SR)* Concordance(SR)*

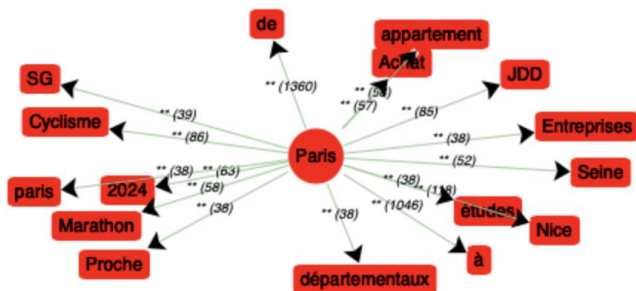
1. Champ de recherche des segments répétés

2. Opérations possibles pour projeter les résultats

Cooccurrents

Présence simultanée dans un fragment de texte des occurrences de deux formes données.

Le calcul des cooccurrents d'un terme se fait en mettant au jour les **termes co-fréquents surreprésentés** dans les contextes de ce terme.



Les résultats apparaissent aussi sous la forme d'un graphique :

- on peut **déplacer les flèches** pour le rendre plus lisible,
- dans lequel on peut **faire apparaître plus ou moins de formes** en **modifiant le champ « Co-Freq »** (on ne retient que les formes dont la co-fréquence est supérieure à la valeur donnée), ou en modifiant IndSPMin (l'indice de spécificité minimum).

Paramètres

PARTITION	date	PARTIES	20220202	Le Contexte	10	Fq Max	5
GRAPHES H	400	SEUIL	5	Co-Freq	2	IndSPMin	50
GRAPHES 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 Paris

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

Calcul de cooccurrents

NB TERME GAUCHE 10 NB TERME DROITE 10

Cooccurrents* Cooccurrents* sur partie sélectionnée

Réseau Cooccurrents* Réseau Cooccurrents* sur partie sélectionnée

StopList="GESTIONNAIRE DE SÉLECTION"

Paris	542	85	**
JDD	40	39	**
Seine	45	38	**
Entreprises	213	57	**
Nice	60	38	**
études	991	86	**
départementaux	415	58	**
à	304	52	**
Proche	107	38	**
Marathon	108	38	**
2024	11794	1360	**
paris	120	38	**

Cooccurents

Présence simultanée dans un fragment de texte des occurrences de deux formes données.

Accès au contexte en cliquant sur un cooccurrent.

Contextes

Afficher 10 items

Recherche :

N°	Contexte
1	#Génération 2024 #labellisation #Paris 2024 #cafés #hotellerue #
2	min Flavy Cohaut : « Paris 2024, j'y pense
3	label décerné par le Comité Paris 2024. Rugby XV -
4	préparer les Jeux Olympiques de Paris 2024 La ville d'Autun
5	défis sur la route de Paris 2024 PODCAST – Primaires :
6	défis sur la route de Paris 2024 Lire plus d'articles
7	Complexe sportif #Jeux olympiques #Paris 2024 #stade Jacquin #terrain
8	de la Flamme olympique de Paris 2024 Pour des raisons budgétaires
9	jugeant la contribution demandée par Paris 2024 trop élevée, la
10	village olympique des Jeux de Paris 2024, l'art campe

Affichage de 1 à 10 des 62 items

Préc.

1

2

3

4

5

6

7

Suiv.

Cooccurents

The screenshot shows a software interface titled "Calcul de cooccurents". It contains several input fields and checkboxes. Annotations with arrows point to specific elements, explaining their functions.

Calcul de cooccurents

NB TERME GAUCHE 10 **NB TERME DROITE** 10

Cooccurents* **Cooccurents* sur partie sélectionnée**

Réseau Cooccurents* **Réseau Cooccurents* sur partie sélectionnée**

☐ **STOPLISTE="GESTIONNAIRE DE SÉLECTION"**

Effectue le calcul des cooccurents sur l'ensemble des items de fréquence supérieure à FQ Max (cf paramètres).

Permet de faire le calcul des cooccurents sur la partie sélectionnée dans les paramètres (par exemple sur un journal en particulier).

Possibilité d'utiliser le Gestionnaire de Sélections comme une stop-liste (une liste de mots que l'on ne souhaite pas prendre en compte : par exemple les mots vides) : pour cela, il faut cocher cette case et sélectionner les items à exclure de ces calculs (par exemple à partir du dictionnaire).

Effectue le calcul des cooccurents sur l'ensemble des items de fréquence supérieure à FQ Max (cf paramètres) dans la partie sélectionnée dans les paramètres.

Gestionnaire de sélection

Possibilité de **sélectionner des formes** du dictionnaires et de les **projeter** sur la carte des sections ou sur une concordance.

Sélection "Paris" terminée...

Dictionnaire

Copy CSV Excel PDF Print

Recherche : Paris

1. Recherche d'un mot dans le dictionnaire

2. Bouton "Sélection"

3. Sélection terminée

Item	Fq	Concordance	Ventilation	Carte	Sélection
Paris	2426				

4. Projection de la sélection

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

Gestionnaire de sélection

Carte Sections (Sélection) Concordance (Sélection)

Supprimer sélections

La semaine prochaine...

Vos projets

- Corpus formaté
- Problématique
- Hypothèse