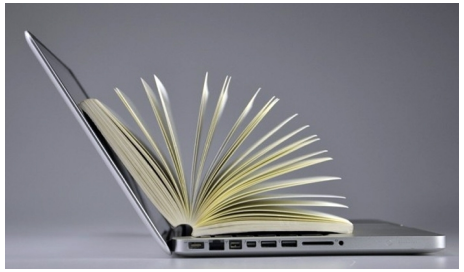
Sorbonne Nouvelle
université des cultures



Traitement et exploitation d'un objet d'étude

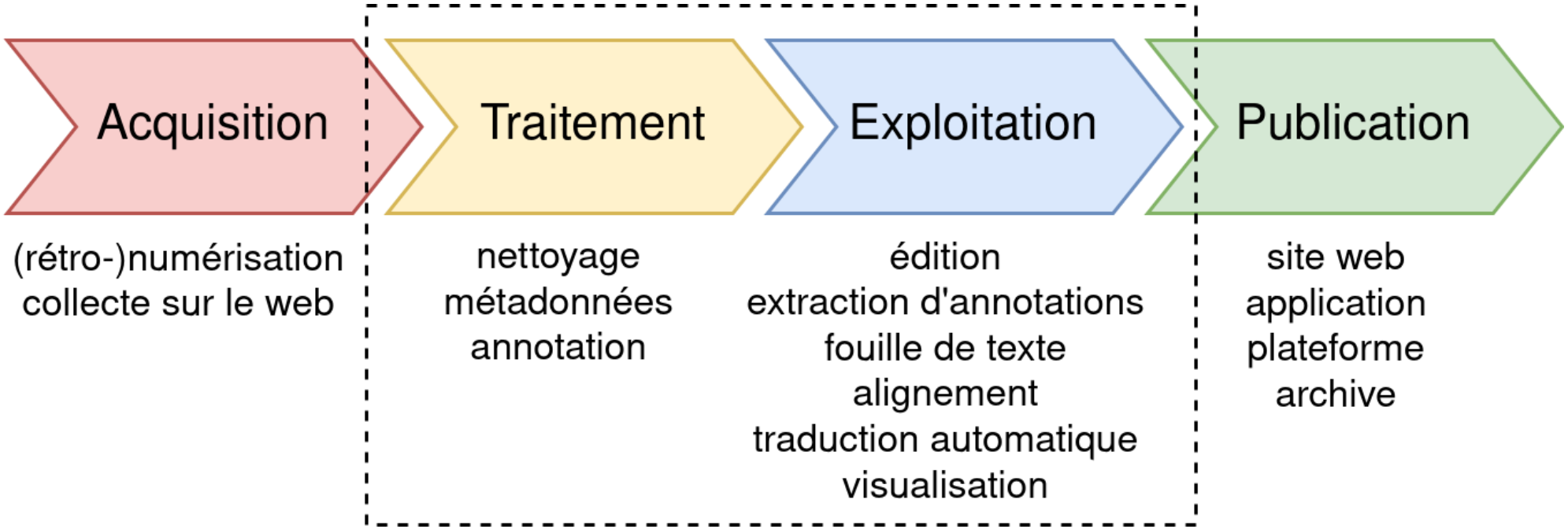
Ljudmila PETKOVIC

Introduction aux humanités numériques (L1HN001)
Mineure « Humanités numériques », licence Lettres
Paris, le 12 octobre 2023, année 2023-2024

Chaîne de traitement

Angl. *workflow*, *pipeline*

- série de processus et d'outils interconnectés et souvent automatisés conçus pour faciliter la collecte, le traitement, l'analyse et la visualisation de données numériques



Comment faire en sorte qu'une machine comprenne du texte et comment en extraire de la connaissance ?

Ordinateur *vs.* humain

- un ordinateur ne comprend pas le français comme un humain pourrait le faire
- les mots n'ont pas de sens pour une machine ⇨ moyens spécifiques pour les exploiter et les manipuler numériquement
    - le TAL adresse cette problématique de compréhension du texte

Mettre les textes en données

- constituer son corpus et récupérer des textes au format numérique — et non pas numérisé : pas de page scannée à la manière d'une photographie — dans un fichier texte ou sur une page web

Qu’est-ce que l’analyse de données textuelles ?

Mettre en données un texte, c’est en extraire des données organisées :

fréquences	nombre de mots dans une phrase, dans le corpus
concordances	occurrences d’un mot dans son contexte
index	liste hiérarchisée des mots les plus employés
étiquetage morphosyntaxique	assigner chaque mot d'un texte à sa catégorie grammaticale
relations	compte d’interactions entre tel et tel élément (p. ex. personnage)
collocations (co-occurrences)	mots associés fréquemment

- comparer des auteurs, différents chapitres d’une même œuvre, différentes œuvres d’un même auteur...

Premières pistes d'analyses

- « **liberté** de presse », « **liberté** de pensée »...

Terme	Collocation	Total (contexte)
liberté*	presse	57
liberté*	liberté	45
liberté*	pensée	15
liberté*	tribune	14
liberté*	théâtre	11
liberté*	peut	10
liberté*	dire	10

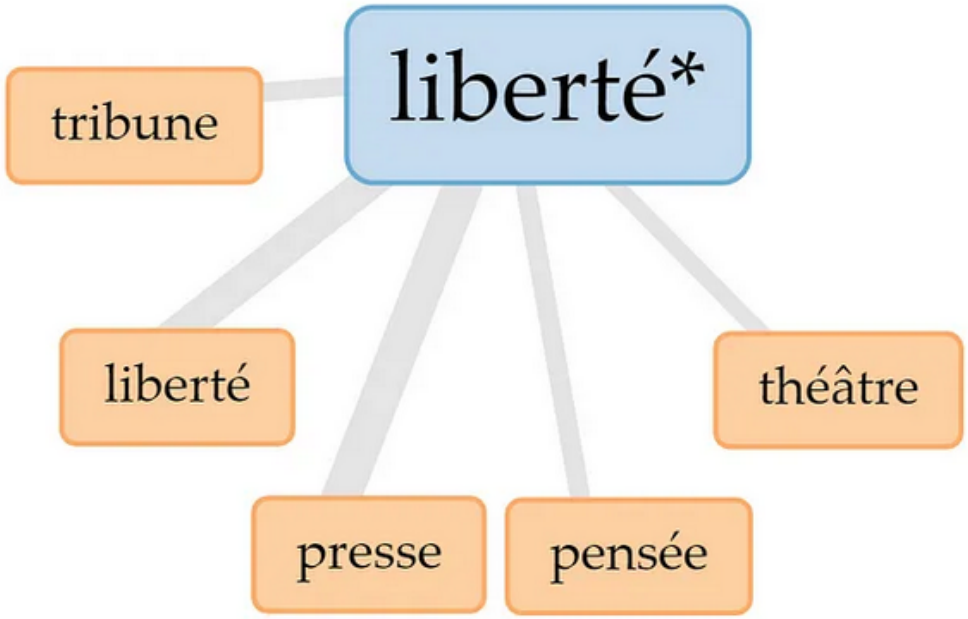


Tableau et représentation des collocations du mot « liberté » dans *Actes et paroles* de Victor Hugo.

Outils de mise en données des textes

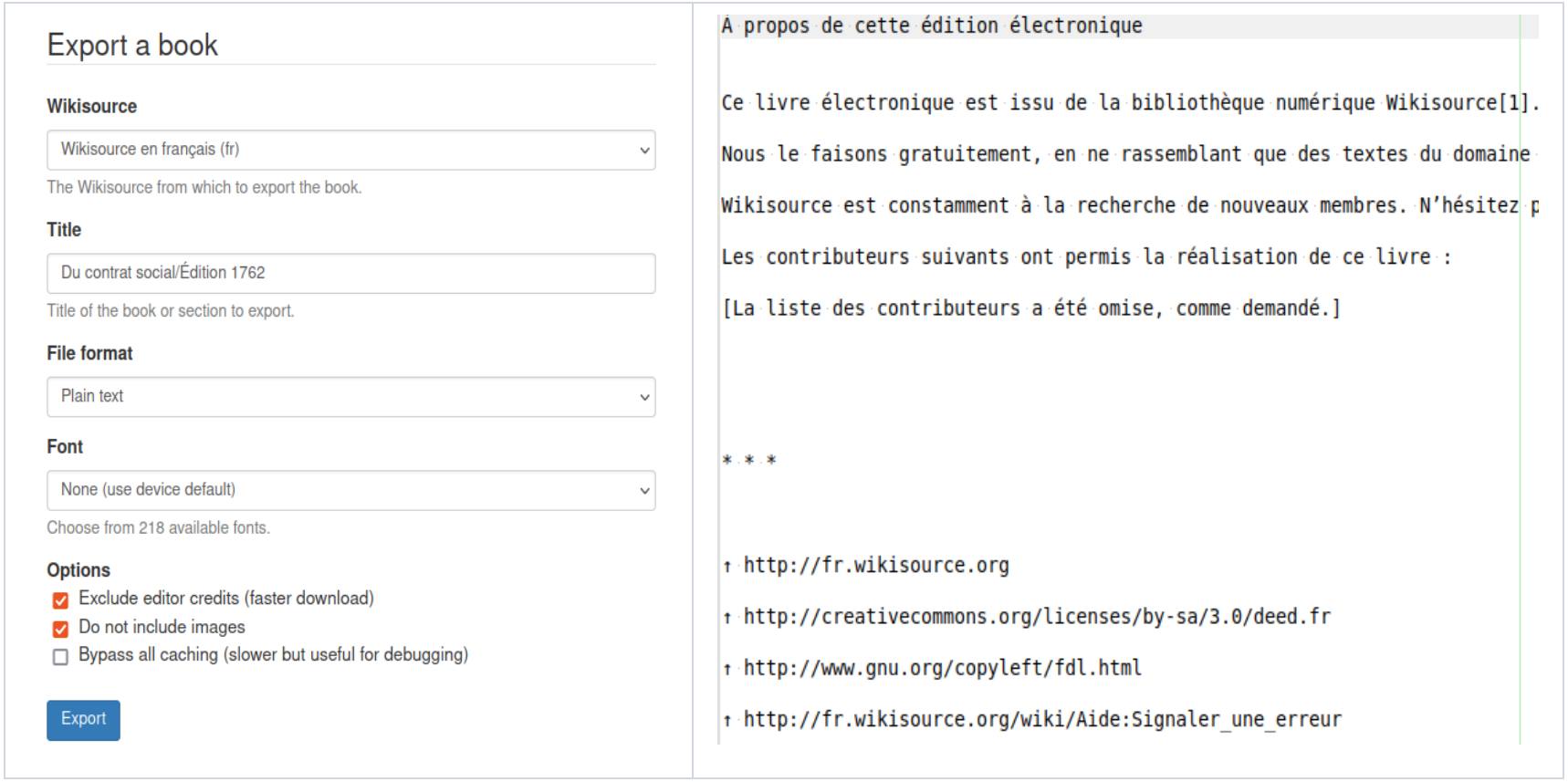
Pour transformer un texte en données organisées nous pouvons utiliser des logiciels, des services dédiés, ou entrer nous-même des données textuelles dans un tableur.

- l'exploitation des textes nécessite qu'ils soient accessibles dans un format numérique exploitable par les services ou logiciels
- privilégier l'exploitation des textes accessibles dans le domaine public ou disponibles sous licence ouverte de type *creative commons*

Avant d'être utilisables numériquement, les textes doivent subir un pré-traitement (nettoyage).

Gestion du paratexte

- crédits, mentions légales... ⇨ besoin d'un nettoyage plus ciblé (programmation)



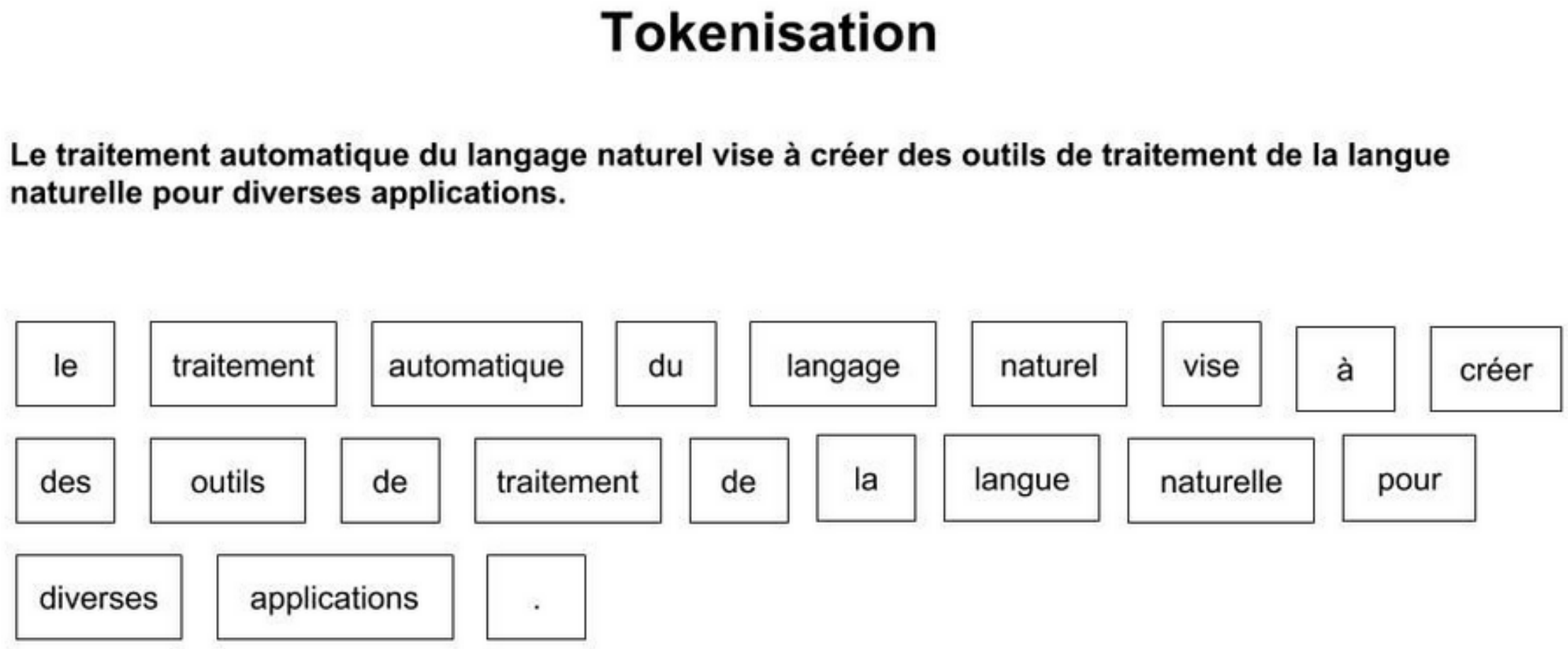
Pré-traitement

- étape inévitable, essentielle et pourtant fastidieuse pour tout projet TAL
- harmonisation : du corpus « bruité » ou « sale » au corpus « propre », p. ex.
  - tokeniser le texte
  - enlever les mots vides
  - lemmatiser le texte
  - remplacer toutes les majuscules par des minuscules (`ljudmila ≠ Ljudmila`)
  - retirer la ponctuation (`. , ? ! ; : ... ( ) [ ] « » – / { }`)
  - retirer les nombres (si ceux-ci n'apportent pas d'informations pour l'analyse)
  - supprimer les lignes vides
- soit « nettoyer » le texte ou laisser ce « bruit » s'il ne fausse pas l'analyse
- tâche non triviale, a priori effectuée de manière entièrement automatisée

Tokenisation

- décomposer une phrase, et donc un document, en *tokens*
  - token : élément correspondant à un mot ou une ponctuation
- cas particuliers :
  - mots avec un trait d'union (*peut être ≠ peut-être*)
  - dates et heures séparées par des points, des slashs, des deux points
  - apostrophes
  - caractères spéciaux : émoticônes, formules mathématiques

Tokenisation



Mots vides

- mots fonctionnels, mots outils, « antidictionnaires » (angl. *stop words*)
- mots dont la fréquence d'apparition est trop élevée et dont le contenu n'est pas pertinent pour le processus d'extraction
    - typiquement les mots grammaticaux : conjonctions, prépositions, déterminants, pronoms, adverbes (*et, dans, le, je, où*...)

> ⚠ À noter que la stylométrie s'intéresse particulièrement à l'agencement des mots outils, signature « verbale » inconsciente du discours.

Approche computationnelle et quantitative du texte, dont l'objectif est de mesurer les idiosynchrasies stylistiques, appellées « stylomes » (un ensemble de caractéristiques mesurables des produits linguistiques).

Retirer les mots vides

Lemmatisation

- chaque mot a une forme canonique (forme racine) et des formes fléchies (différentes occurrences possibles)
- objectif : associer chaque mot à une forme canonique
- analyse lexicale qui permet de regrouper les mots d'une même famille ensemble

  ⇨ regroupement par *lemme*

  - verbes ⇨ **infinitif**
  - noms, adjectifs, articles ⇨ **masculin singulier**

(le lemme correspond à l'infinitif des verbes et à la forme au masculin singulier des noms, adjectifs et articles)

Lemmatisation

les étoiles claires luisent dans la nuit noire

le étoile clair luire dans le nuit noir

Lemmatisation + élimination des mots vides

Le traitement automatique du langage naturel vise à créer des outils de traitement de la langue naturelle pour diverses applications.

traitement | automatique | langage | naturel | viser | créer | outil | traitement

langue | naturel | divers | application

Autre exemple de la lemmatisation

- texte **brut** : Le petit chat *a disparu*. C'*est* dommage, il *était* gentil le chat.
- texte **lemmatisé** : Le petit chat *avoir disparaître*. C'*être* dommage, il *être* gentil le chat.
    - nombre de **lemmes**, après avoir converti tous les mots en minuscules (**10**) : *le, petit, chat, avoir, disparaître, c', être, dommage, il, gentil*

Token vs. *type*

Le petit chat a disparu. C'est dommage, il était gentil le chat.

- **tokens** : instances individuelles d'une unité linguistique (**13**) : *Le, petit, chat, a, disparu, c', est, dommage, il, était, gentil, le, chat*
- **types** : formes fléchies des mots, classes de tokens contenant la même séquence de caractères (**12**) : *Le, petit, chat, a, disparu, c', est, dommage, il, était, gentil, le*
    - noter deux différentes graphies de l'article défini (*Le, le*), considérées comme deux types différents, ainsi que la non-répétition du mot *chat*

Les **tokens** servent à mesurer le **volume** d'un corpus, tandis que les **types** servent à en mesurer la **richesse**, le nombre de formes **uniques**.

Ici, on ne compte **pas** les signes de ponctuations.

Mettre les textes en données (exemple)

Lettre	Lettre chiffre R	Expéditeur	Destinataire	Date
1	I	Cécile Volanges	Sophie Carnay	03/08
2	II	Marquise de Merteuil	Vicomte de Valmont	04/08
3	III	Cécile Volanges	Sophie Carnay	04/08
4	IV	Vicomte de Valmont	Marquise de Merteuil	05/08
5	V	Marquise de Merteuil	Vicomte de Valmont	07/08
6	VI	Vicomte de Valmont	Marquise de Merteuil	09/08
7	VII	Cécile Volanges	Sophie Carnay	07/08
8	VIII	La Présidente de Tourvel	Cécile Volanges	09/08
9	IX	Cécile Volanges	La Présidente de Tourvel	11/08
10	X	Marquise de Merteuil	Vicomte de Valmont	12/08
11	XI	La Présidente de Tourvel	Cécile Volanges	13/08
12	XII	Cécile Volanges	Marquise de Merteuil	13/08
13	XIII	Marquise de Merteuil	Cécile Volanges	13/08
14	XIV	Cécile Volanges	Sophie Carnay	14/08

Terme	Occurrences
c'est	1093
qu'il	1029
dit	992
là	756
d'un	524
c'était	492
jean	473
homme	458
monsieur	457
marius	443
valjean	423

	Titre	Mots	Mots/Phrase
1	Les Misérables - Tome I -…	105,796	16.6
2	Les Misérables - Tome II …	92,083	18.5
3	Les Misérables - Tome III…	81,599	16.7

Tableur des échanges épistolaires dans *Les Liaisons dangereuses* (gauche) ; index des mots et informations sur le corpus *Les Misérables* (droite).

Voyant Tools

https://voyant-tools.org/

- environnement d'analyse de texte en ligne
- plateforme d'analyse de texte assistée par ordinateur
- aucune installation de logiciel n'est nécessaire
- analyser des textes importés depuis des fichiers ou liens hypertextes
- analyser une œuvre, plusieurs pour les comparer (ou plusieurs actes/chapitres d'une même œuvre)

Voyant Tools

- tutoriel d'Aurélien Berra
- vidéo tutoriel de l'ancien labex OBVIL

Des données aux visualisations

Les données textuelles peuvent être compilées dans des tableaux et comparées, analysées, représentées graphiquement avec des outils de visualisation de données.

Objectifs :

- analyser un grand volume de texte, dépasser les capacités de lecture
- mettre en évidence des aspects du corpus, étayer une argumentation
- ouvrir de nouvelles pistes sur le corpus ou asseoir une théorie

À quoi bon ?

- le travail sur les données issues de texte, leur analyse et représentation, sert l’interprétation des œuvres
- interroger le texte, l'analyser dans un environnement technologique de l’écrit devenu numérique
- s'acculturer à ce milieu du texte, développer des habiletés propres au numérique et à la discipline pratiquée : les lettres, l’étude de textes

Références

- **Gabay, S.** (2021). « 4. Introduction à la stylométrie » [*diapositives*] https://github.com/gabays/32M7129/blob/master/Cours_04/Cours_Geneve_4.pdf.
- **Galleron, I.** (2021). « Introduction aux humanités numériques (L1HN001) [*diapositives en interne*].
- **Marques, J.** (2021). « L'analyse de données textuelles en classe au service de la compréhension et de l'interprétation des textes ». Medium https://j0annamarques.medium.com/lanalyse-de-données-textuelles-en-classe-au-service-de-la-compréhension-et-de-l-interprétation-des-5c3b96dd97c6.
- **Team Data** (2018). Introduction au NLP : le traitement de texte automatisé » https://blog.coddity.com/articles/natural-language-processing/.stav