



Introduction aux Études Littéraires Computationnelles

Jean Barré

2 septembre 2025

Doctorant École Normale Supérieure - Université PSL - LaTTiCe

Table des matières

1. Introduction

2. Le TAL : Toute une histoire

L'Index Thomisticus : des débuts fondateurs

Le TAL aujourd'hui

3. La chaîne de traitement du TAL

4. Le TAL dans le contexte des Humanités Numériques - Exemple de la stylométrie

Mesurer le style ?

La stylométrie en action : Deux exemples d'applications

5. Quelques exemples de travaux au Lattice

Introduction

Introduction

Les Études Littéraires Computationnelles :

- Carrefour de plusieurs disciplines (Histoire/Théorie Littéraire, Stylo-métrie, TAL, Apprentissage Machine)
- Des avancées techniques et pratiques récentes
- Tradition française de la Textométrie

Deux concepts clés :

- **Opérationnalisation**

Définir et formaliser des concepts de la théorie littéraire

- **Lecture distante**

Évolution dans la longue durée

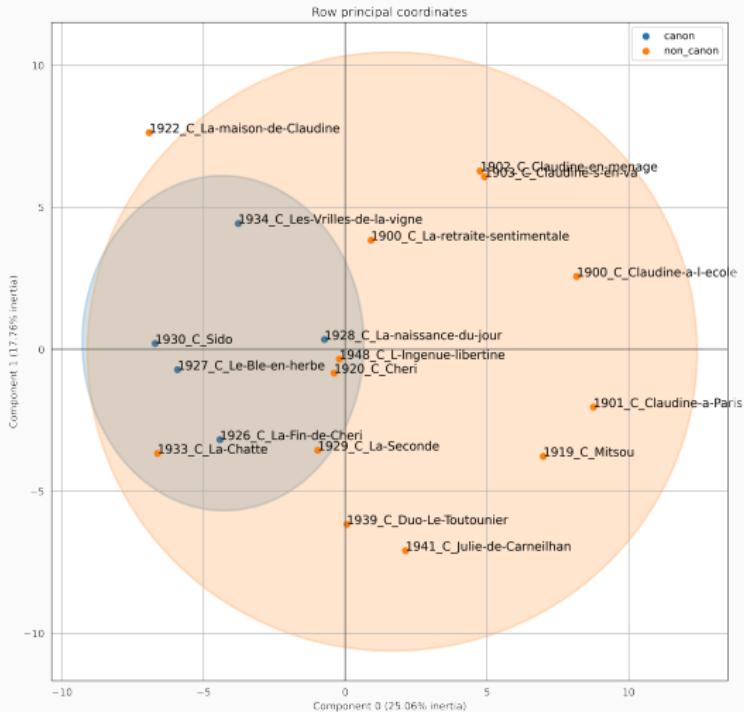


Figure 1 : Ouvrages Canoniques chez Colette (Barré et al, 2023)

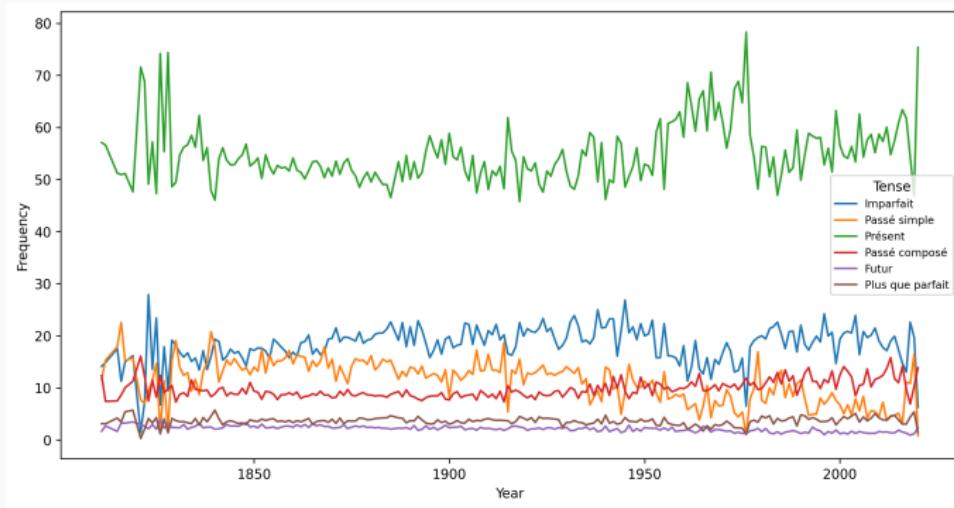


Figure 2 : Évolution des temps verbaux dans le corpus Chapitres (Barré et al, 2025)

Le TAL : Toute une histoire

Comprendre la pensée par l'usage du langage

- **Roberto Busa** (1913-2011), jésuite italien, spécialiste de *Thomas d'Aquin*.
- Intuition : on accède à la pensée d'un·e auteur·ice en maîtrisant **son usage du langage**.
- Objectif : construire un **concordancier exhaustif** des œuvres de Thomas (toutes les occurrences, avec contexte).
- Finalité : permettre des **recherches systématiques** (mots, lemmes, tournures, cooccurrences).
- <https://www.corpusthomisticum.org/it/index.age>

De l'idée à l'industrialisation

- Constat immédiat : **trop vaste** pour un traitement manuel.
- Quête d'une "machinerie" (« any gadget that might help », Busa, 1980) ⇒ **collaboration avec IBM**.
- **179 textes** transcrits pour machines (cartes perforées); > **30 personnes** mobilisées.
- **10 632 980 mots indexés**; ~1500 km de câbles; ~10 000 h de calcul; ~1 000 000 h de travail humain.
- Chaîne de traitement : **encoder** → **normaliser/lemmatiser** → **indexer** → **interroger**.

Le chantier matériel



Figure 3 : *

Cartes perforées et tri mécanique

Pourquoi c'est fondateur (TAL & Humanités)

Apports durables

- **Échelle** : embrasser une **masse de données** et poser des questions **impossibles autrement**.
- **Analyse** : du *close reading* à la *distant reading* (patterns globaux, fréquences, cooccurrences).
- **Méthodes** : pipeline *text mining* avant l'heure (nettoyage, lemmatisation, indexation, requêtes).
- **Héritage** : acte précurseur reliant **TAL** et **humanités numériques**.

À retenir

Un outil pour la pensée : mieux comprendre Thomas d'Aquin en objectivant son **usage du langage**.

Les tâches du TAL

- Récupération d'informations linguistiques (syntaxe, schéma de dépendances)
- Traduction automatique
- Classification automatique de texte (spam/non spam)
- Résumé de texte - Récupération de thèmes spécifiques
- Questions / Réponses - Chatbots
- Génération de texte

Le TAL aujourd'hui

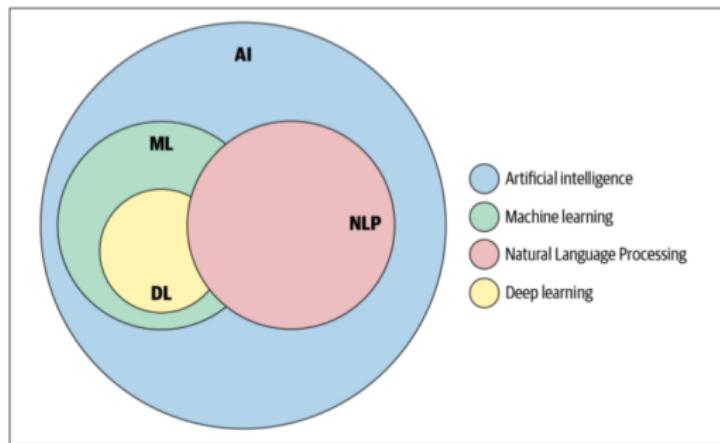


Figure 1-8. How NLP, ML, and DL are related

Figure 5 : Définition du champ. Source : (Sowmya, 2020)

La chaîne de traitement du TAL

La chaîne de traitement classique du TAL

1/3

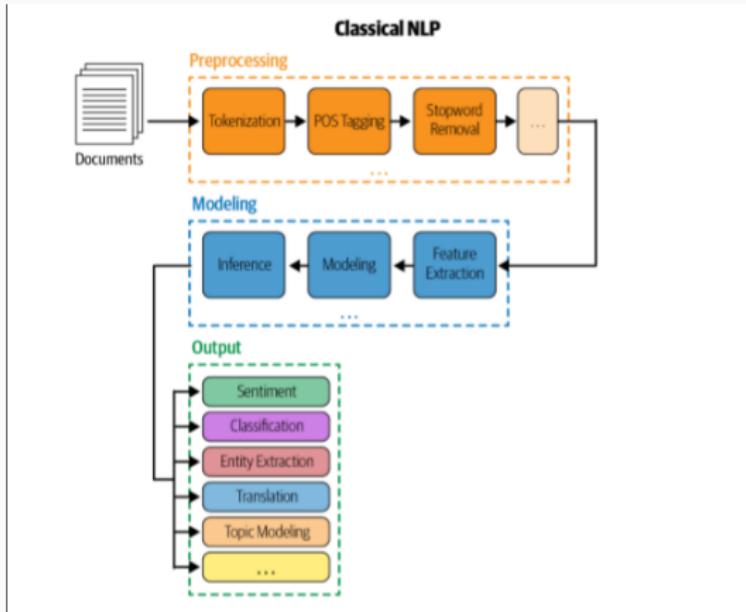


Figure 6 : Source : (Sowmya, 2020)

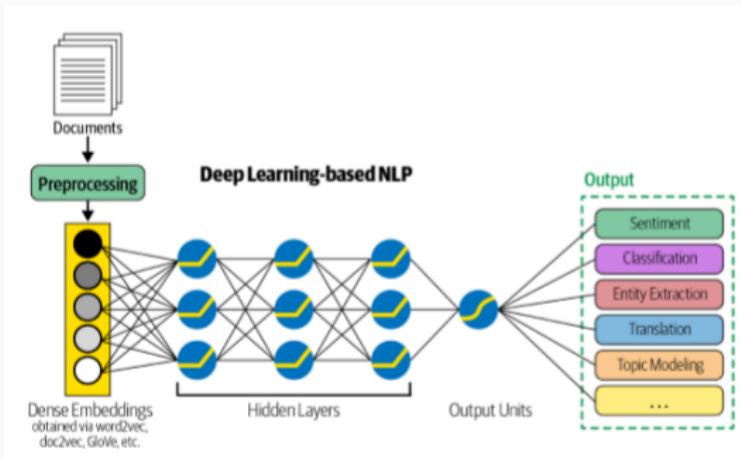


Figure 7 : Source : (Sowmya, 2020)

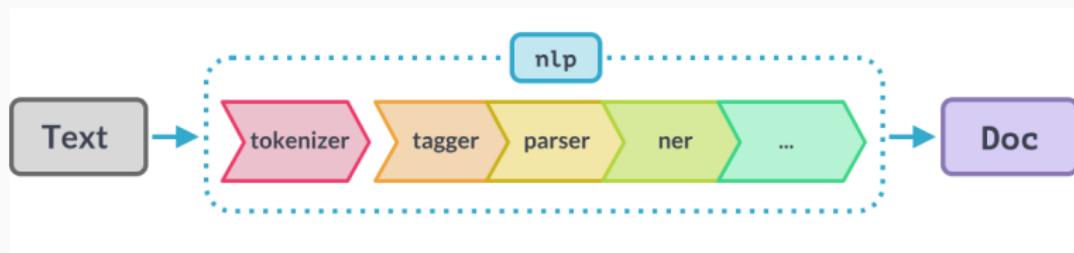


Figure 8 : Source : <https://spacy.io/usage/processing-pipelines/>

Le TAL dans le contexte des Humanités Numériques - Exemple de la stylométrie

La stylométrie

Analyser des données textuelles

- dater, localiser des textes ;
- regrouper des textes selon des caractères stylistiques ;
- attribuer une pièce disputée entre plusieurs auteurs ;
- détecter des collaborations.

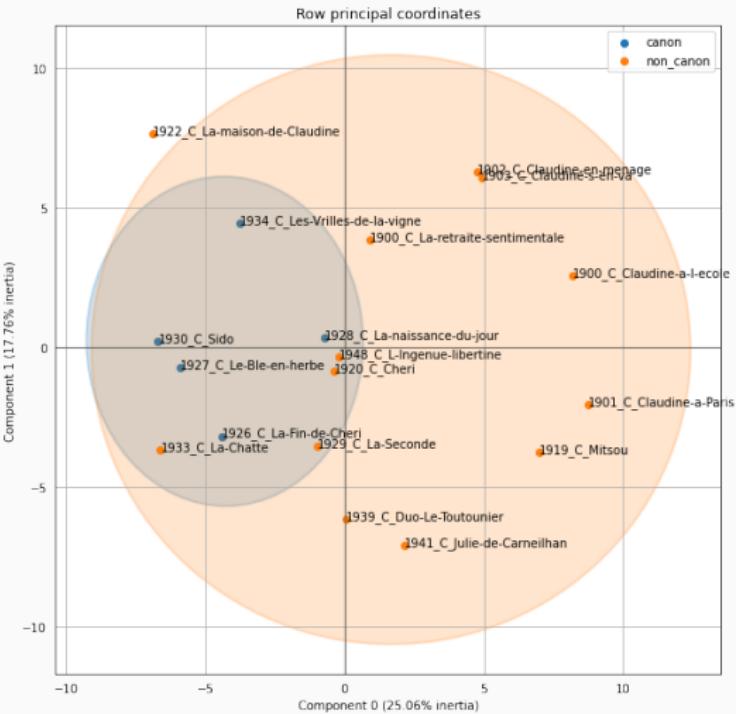


Figure 9 : (Barré, 2023)

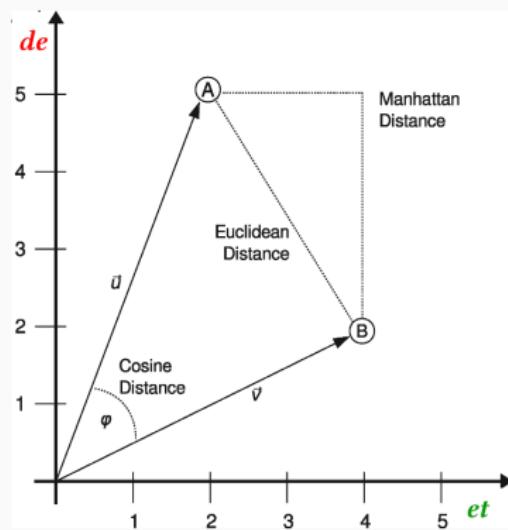
Mesure de distances entre des textes

Texte A

et de joie, il dit à son voisin de classe de gauche qu'il senthousiasmait de cette conférence et de cette journée.

Texte B

et de chaque côté de lui, il vit chiens et chats et poules et canards.



Mesurer le style ?

Wincenty Lutosawski - Principes de stylométrie (1890)

Objectif : Classer chronologiquement les œuvres de Platon

Postulat :

Chaque individu emploie une langue démontrant des propriétés particulières et mesurables, appelées stylome ou idiolecte.

Idiolecte :

Ensemble de traits linguistiques caractéristiques d'un individu

On peut détecter cela quantitativement - Importance des mots outils (mots les plus fréquents)

Propriétés inconscientes du style

Chaîne de traitement du TAL

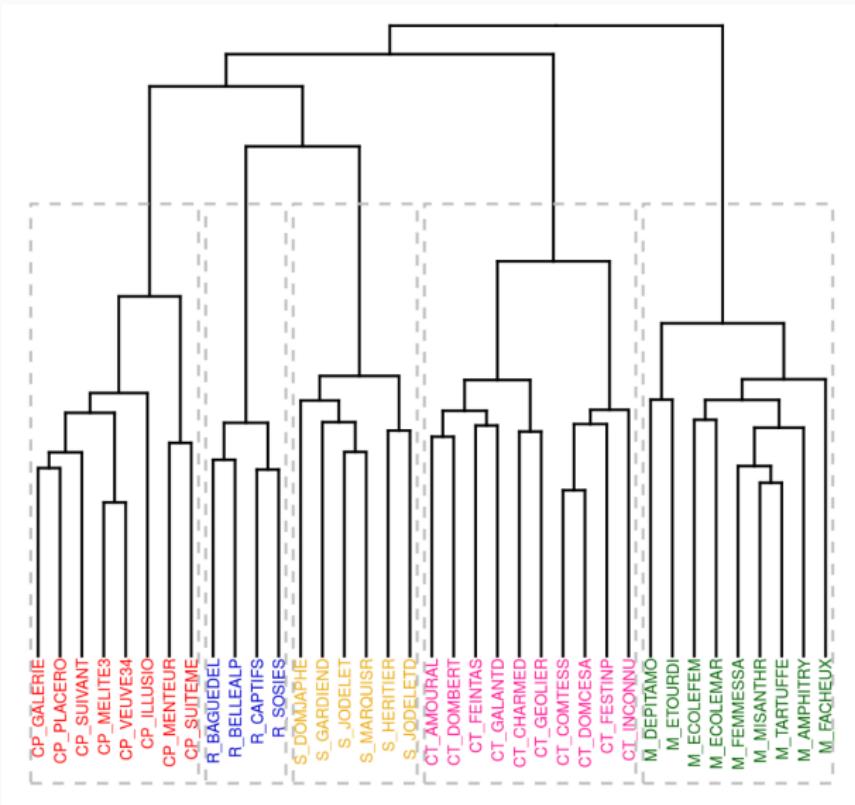
- Optical Character Recognition / Handwritten Text Recognition
- Tokenisation
- Lemmatisation (lemme = partie canonique - entrée de dictionnaire)
- Étiquettagé morpho-syntaxique (partie du discours - nom, verbe, pronom, substantif, ...)
- Sac de mots / Sac de séquence de mots - de POS

Contexte & stylométrie

- Remise en cause de la paternité de certaines œuvres de Molière.
- P. Corneille étant le vrai auteur ?
- Analyse du lexique, des lemmes, des rimes, de la morphosyntaxe sur un corpus de comédies.

Molière a-t-il vraiment écrit ses pièces ?

2/2



Psyché, Texte issue d'une collaboration :

- Molière a dressé le plan de la pièce. Il a rédigé le début de chaque acte
- Corneille s'est lui occupé de finir les actes
- Quinault - rédige les paroles du choeur (début et/ou fin des actes)
- Peut-on détecter quantitativement le signal de cette collaboration ?

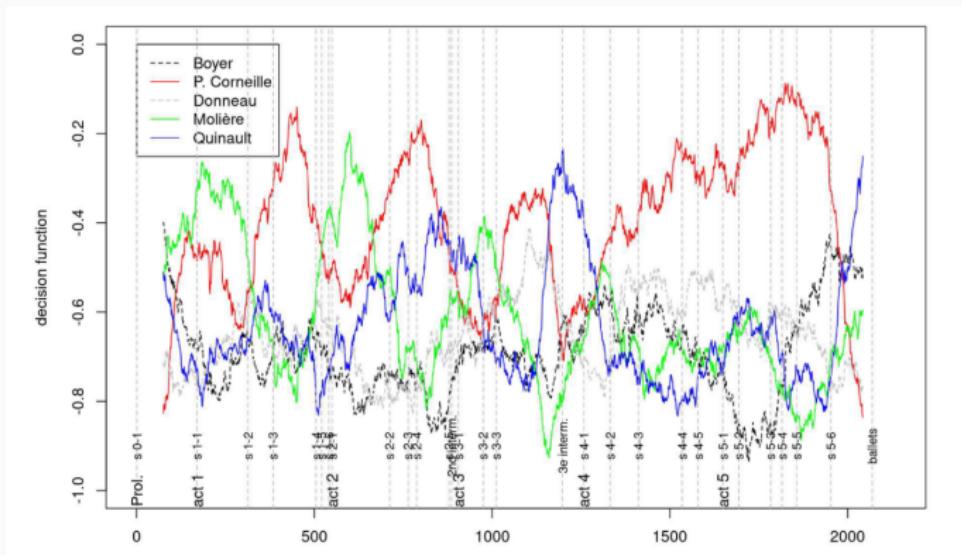


Figure 11 : Source : (Cafiero, 2021)

- **Examen**

Examen final : **50%**

- **Projet**

Projet à rendre : **30%**

- **Travaux pratiques notés**

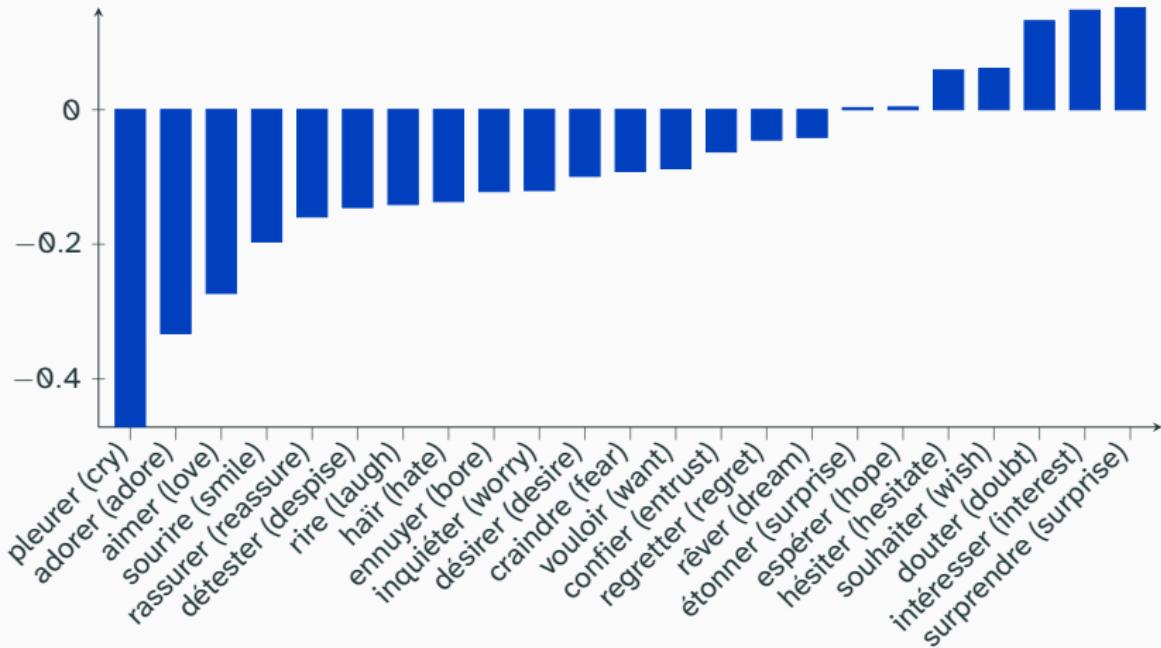
TP noté : **20%**

Quelques exemples de travaux au Lattice

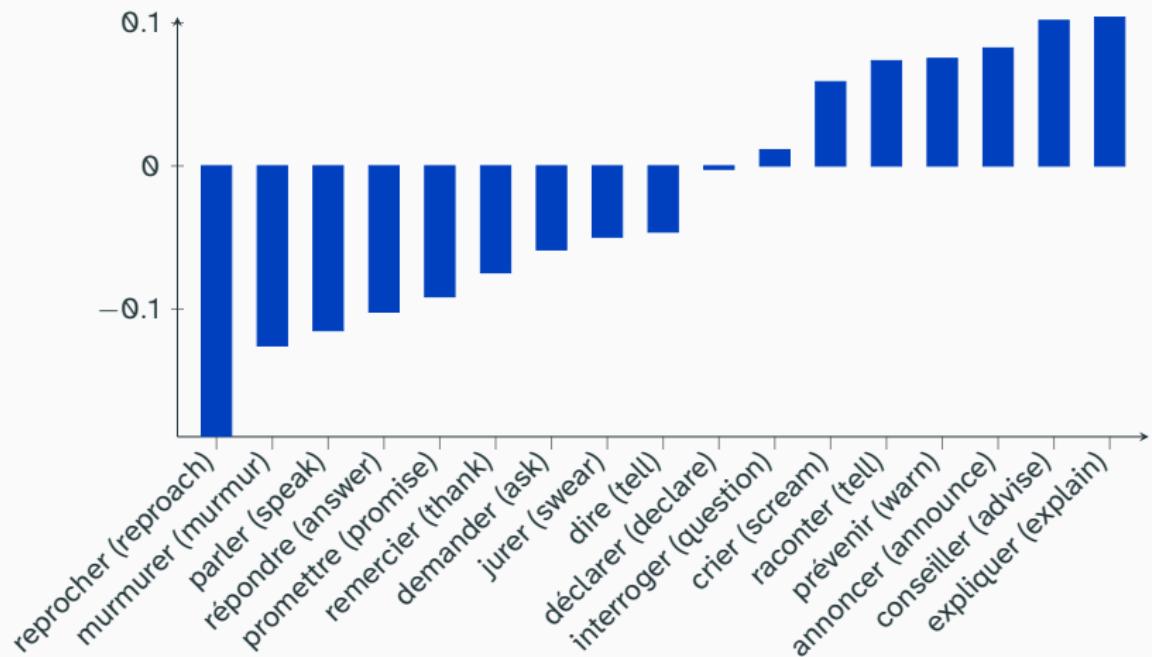
Algorithme de TAL à l'échelle des romans

- NER / Reconnaissance d'entités nommées (PER, FAC, TIME, ORG, LOC, ...)
- Regroupe les dénominations des personnages :
Arsène Lupin, Monsieur Lupin, Lupin, le gentleman cambrioleur
-> ARSENE_LUPIN
- Résolution de la co-référence : pronoms, noms, .. : Arsène, son ami, ce monsieur

Stéréotypes de Genre **CHR2023** : Verbes d'Émotions Associés avec les Personnages Féminins et Masculins

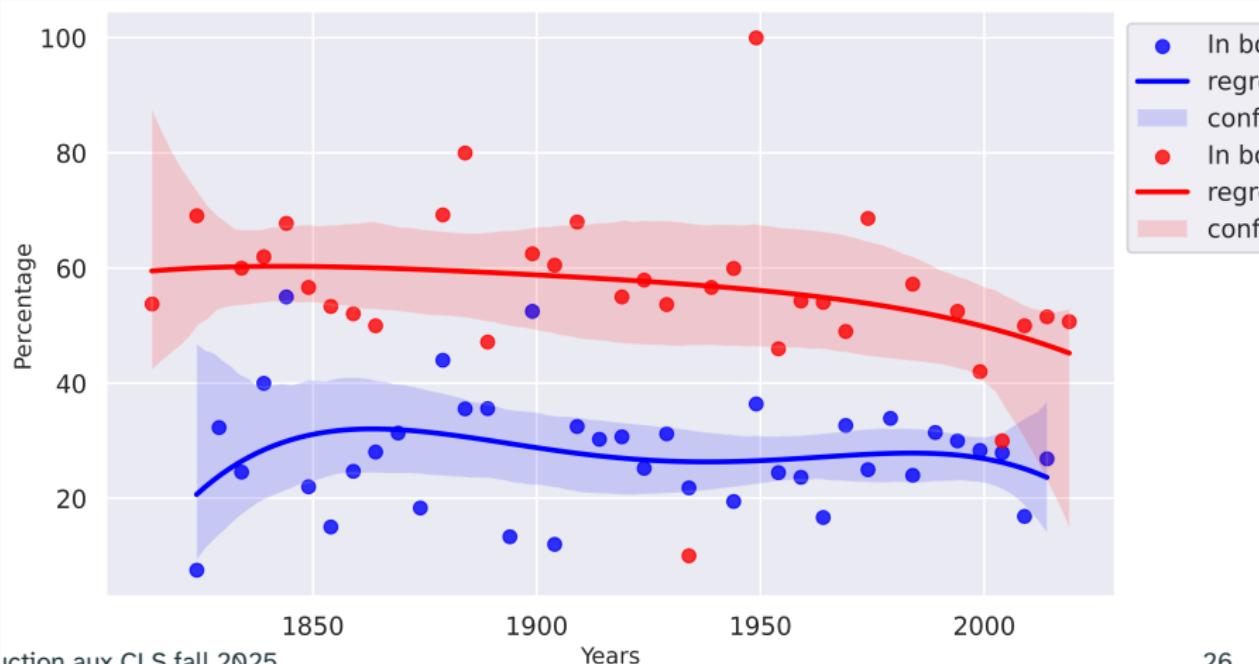


Stéréotypes de Genre **CHR2023** : Speech Verbs Associated with Male and Female Characters



Temps d'écran des personnages Hommes et Femmes dans le temps

En fonction du genre de l'auteurice (Dhai Intensive Week 2022)



Réseaux de Personnages

(Chen et al DH 2024)

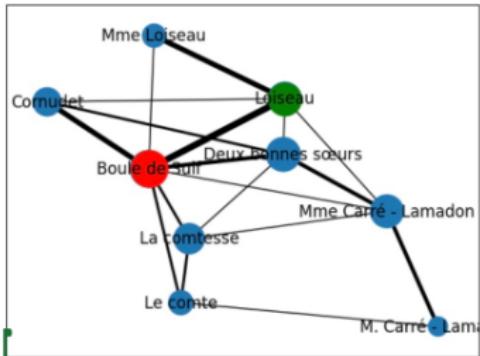


Figure 1

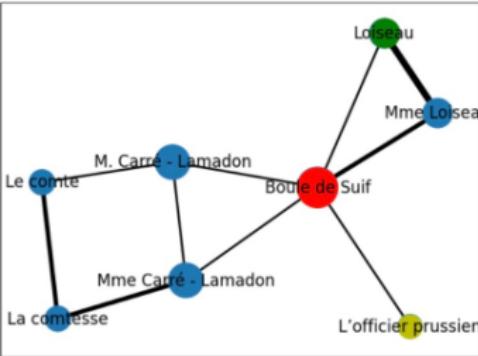


Figure 2

UMAP Projection of Characters - TF-IDF
(Min Mention Count 500)

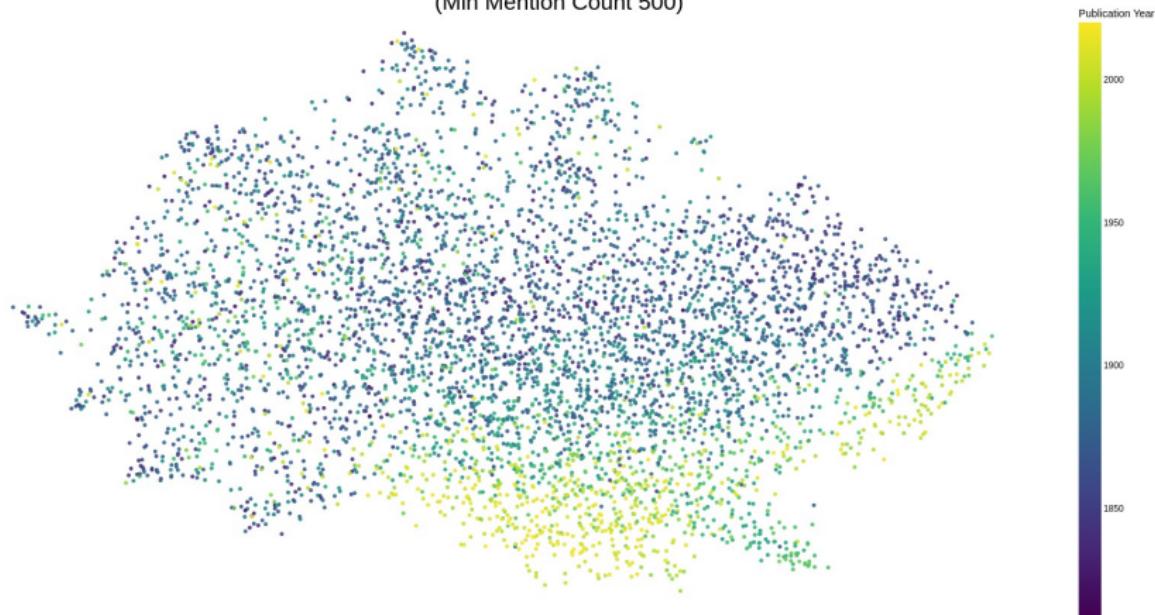


UMAP Projection of Characters by Author - TF-IDF
(Min Mention Count 500)

Main Authors
• Dumas-Alexandre
• Delly

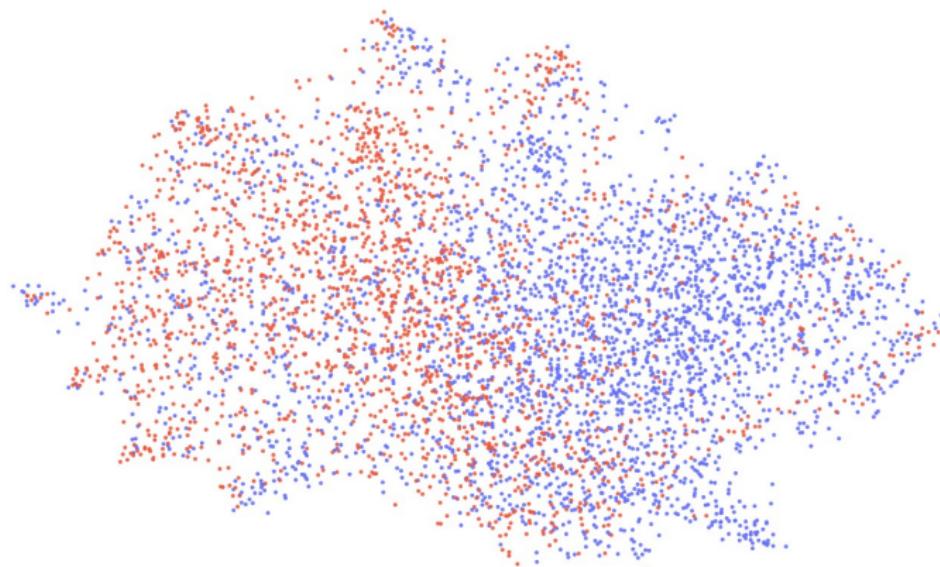


UMAP Projection of Characters by Publication Year - TF-IDF
(Min Mention Count 500)



UMAP Projection of Characters by Character Gender - TF-IDF
(Min Mention Count 500)

Character Gender
• Male
• Female



UMAP Projection of Characters by Literary Gender - TF-IDF
(Min Mention Count 500)

Literary Gender
• roman d'aventures
• policier



Enjeux théoriques des CLS

- La mesure / la modélisation dans les études littéraires
- Objet d'étude : plus le texte mais le corpus
- Passer des concepts à l'opérationnalisation
- Retour à la lecture proche : nécessaire !

3 étapes clés

- Question(s) de recherche
- Modélisation
- Analyse des résultats

Gagner du temps et déléguer des tâches à l'ordinateur

- traiter des corpus très importants ou jusqu'à un niveau très fin, sans temps supplémentaire.
- réaliser des opérations répétitives difficilement envisageables à la main, en limitant le risque d'erreur humaine;

Avoir une autre approche des mêmes données

- obtenir des réponses qu'on n'aurait pas pu obtenir par des moyens traditionnels.
- bénéficier de l'apport méthodologique d'autres champs scientifiques (biostatistiques, IA, TAL, etc.).

Ancrer son analyse dans les faits et leur mesure

- éviter un certain nombre décueils de l'analyse traditionnelle : surévaluation des phénomènes individuels, des individus aberrants, meilleure évaluation des tendances d'ensemble, etc.
- systématiser son analyse : modélisation des données, uniformité de la méthode qui leur est appliquée.

Questions ?

N'hésitez pas à m'écrire !
jean.barre@ens.psl.eu