Introduction aux Études Littéraires Computationnelles

Jean Barré

5 octobre 2023

Doctorant École Normale Supérieure - Université PSL - LATTICE

Les Études Littéraires Computationnelles :

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

Fondements conceptuels

- **Lecture Distante** (Moretti, 2000) [2]
- Question principale - Comprenons-nous les grandes lignes de l'histoire littéraire ? (Underwood, 2019) [3]

Concept -> Formalisation -> Analyse en lecture proche et distante

Table des matières

Classifier le sous-genre du roman policier

Classifier le sous-genre du roman policier

Le roman Policier - culture populaire, de la fin 19ème à nos jours

- Quelles caractéristiques textuelles propres ?
- Le roman policier de la fin du 19ème est-il le même que celui du début 21ème siècle ?
- un modèle statistique peut-il reconnaître automatiquement le roman policier ?

Corpus

Corpus Chapitres (Leblond, 2022)[1] :

- 2961 romans
- Labels annotés : aventures, policier, autobiographie, SF, ...

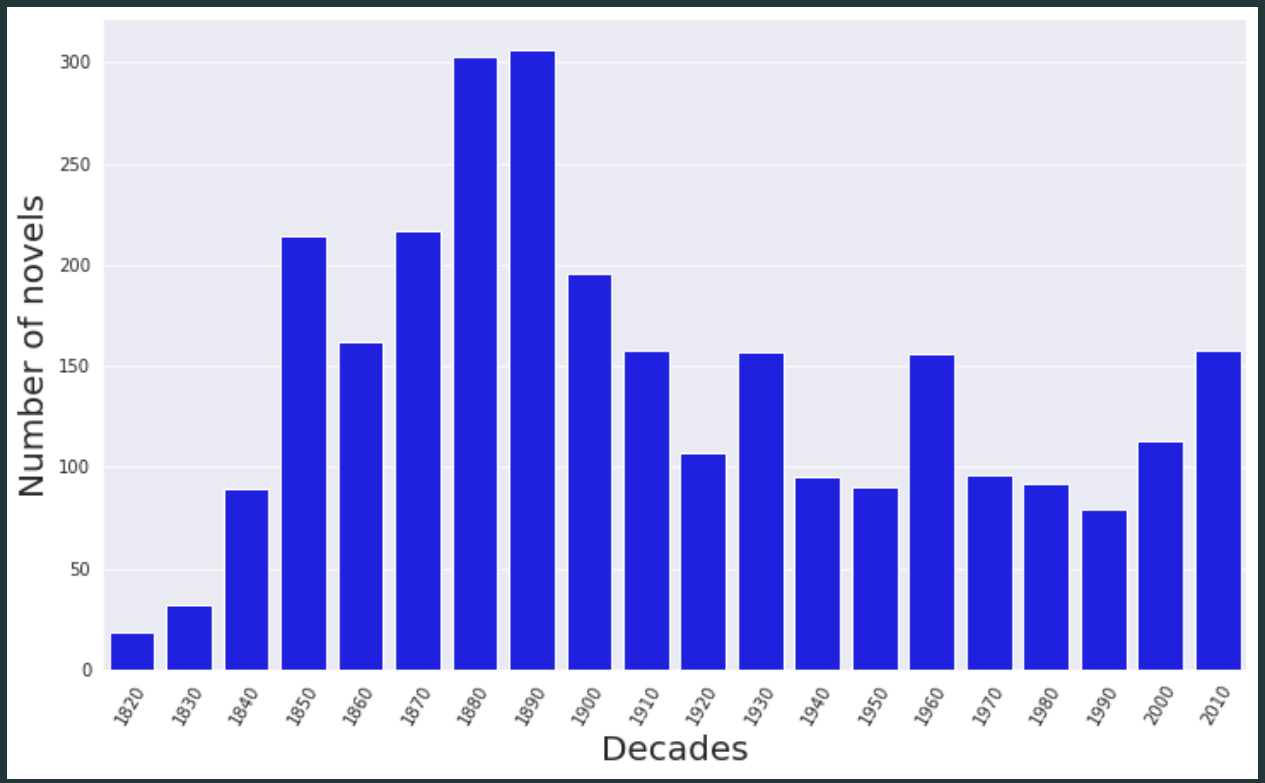

Figure 1 : Distribution du nombre de roman dans le temps

Méthode

Un ensemble de caractéristiques textuelles

- **Lexique** : 1000 mots les plus fréquents. Sac de mots.
- **Thèmes** : Cinquante thèmes principaux identifiés automatiquement. Nous récupérons leur proportion dans chaque roman.
- **Caractérisation** : Adjectifs et verbes qui caractérisent le personnage dans le récit.
- **Chronotope** - fr-BookNLP : LOC + FAC + VEH + TIME.

Fr-BookNLP

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

Classification automatique du roman policier

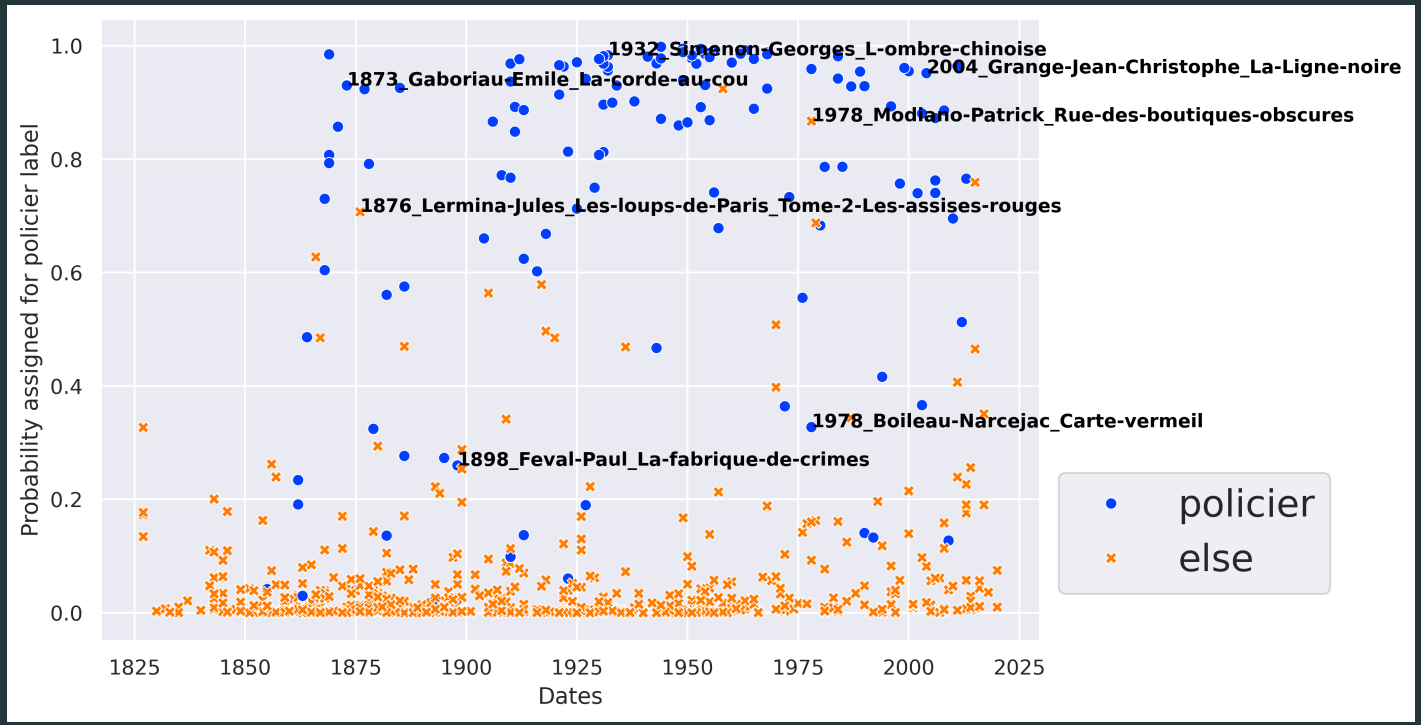

Figure 2 : Probabilité prédite d'appartenir au roman Policier

Quelles caractéristiques discriminantes ?

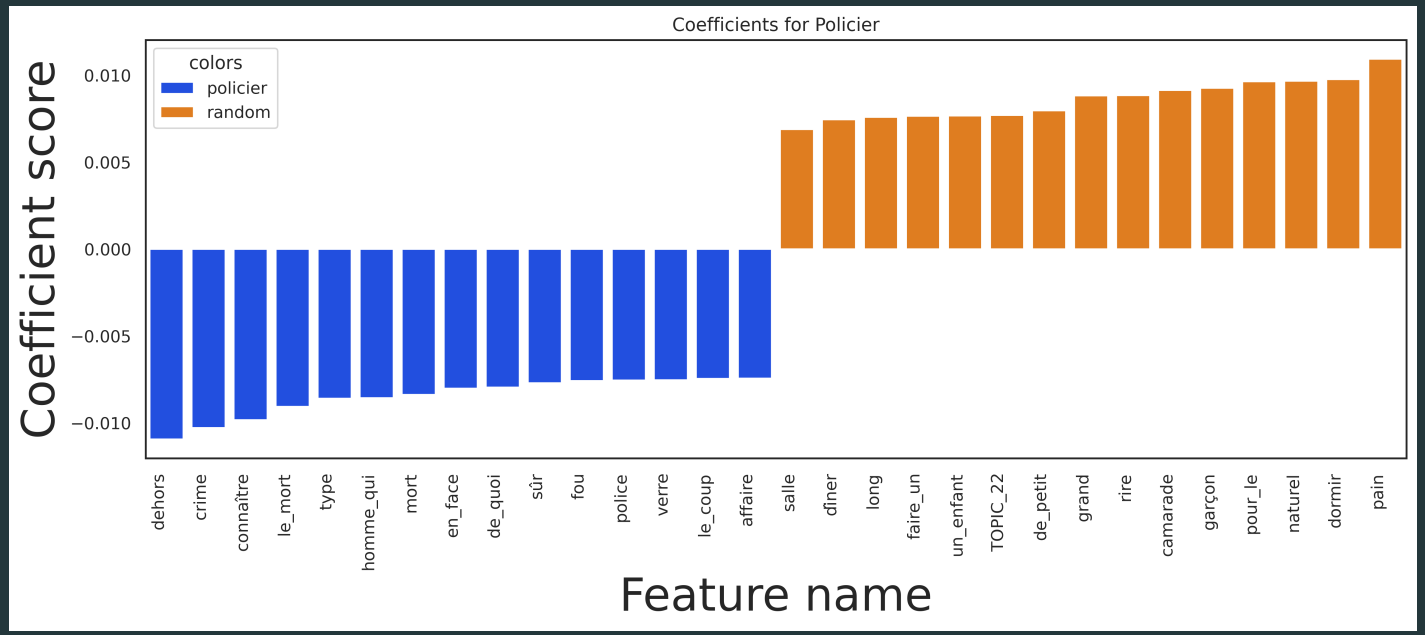

Figure 3 : Coefficients discriminants pour le modèle

Bilan

- Le roman policier a des caractéristiques textuelles qui lui son propres
- Stabilité dans la manière d'écrire le roman policier : Sous-genre très codifié
- La modélisation permet une abstraction et une simplification d'un problème très long à faire à la main
- Il nous reste l'interprétation !

Une histoire computationnelle du genre dans la fiction

Principale question de recherche

Quels sont les enjeux dans la représentation du genre dans la fiction sur ces deux derniers siècles de production littéraire ?

- Évaluez les signes genrés que les écrivains utilisent pour décrire des personnages.
- Les hommes fictifs sont-ils très différents des femmes fictives ?
- Dans quelle mesure les signes publics du genre influencent la caractérisation en général ?

cf (Underwood, 2018) [4]

Étapes de travail :

Prédiction du genre à partir des mots qui caractérisent les personnages

- Récupération de données - Spacy, Fr-BookNLP
- Manipulation de données - Numpy, Pandas
- Annotation de données - Humain
- Apprentissage machine - SKLearn
- Visualisation - Matplotlib & Seaborn

Résultats

27 528 personnages annotés automatiquement.
17 604 (64%) sont des hommes et **9 924** (36%) sont des femmes.
Sur-représentation des hommes dans la fiction

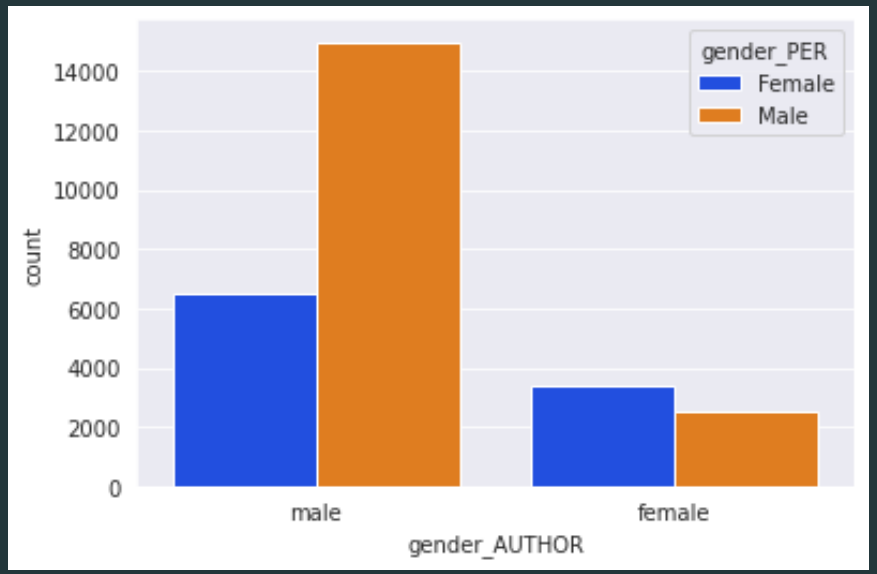

Figure 4 : Proportion de la caractérisation par des auteurs et des autrices

Autrices - 57% Personnages féminins, 43% Personnages masculins
Auteurs - 30% Personnages féminins, 70% Personnages masculins

L’attention / temps d’écran des personnages

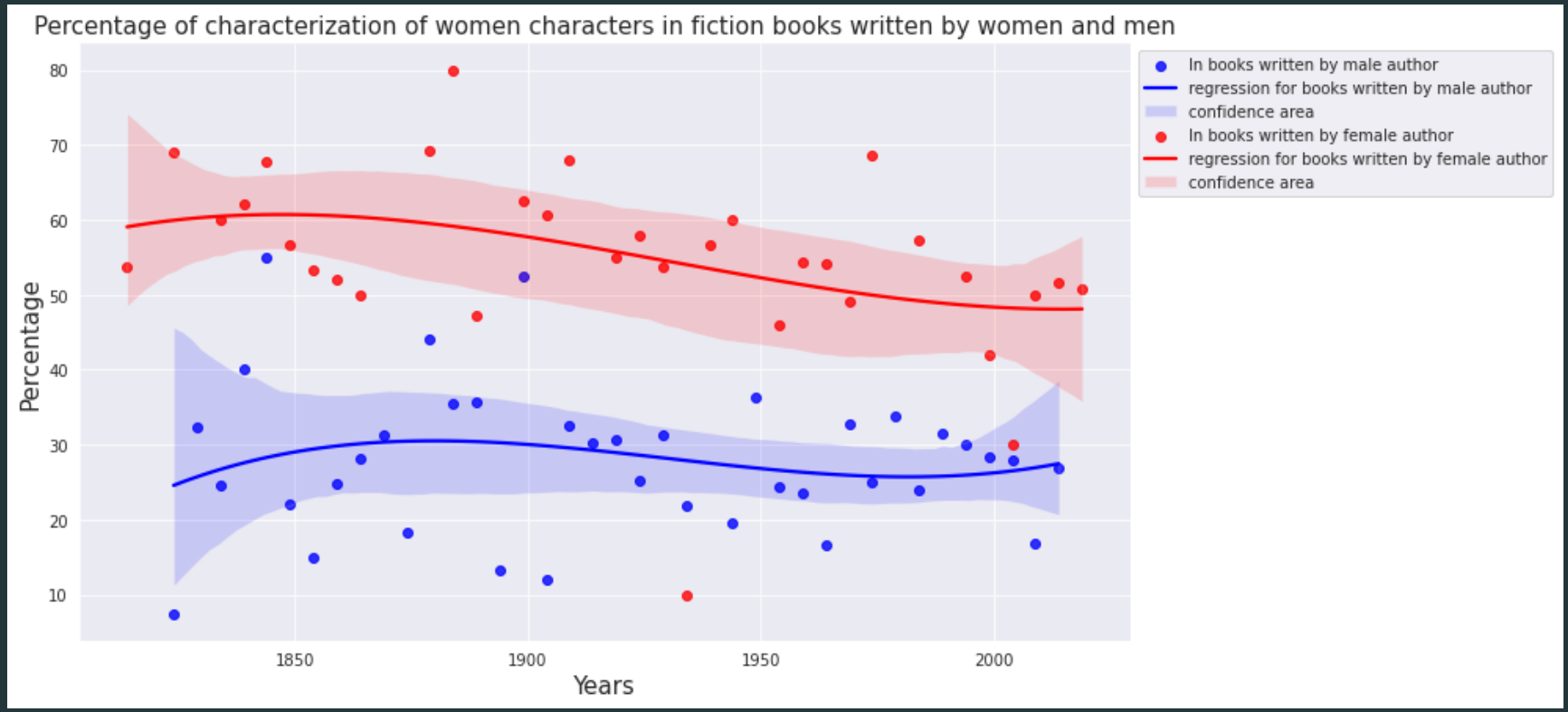

Figure 5 : Proportion de la caractérisation des femmes par des auteurs et des autrices, moyenne tous les cinq ans

Quelques mots genrés dans la fiction 1/6

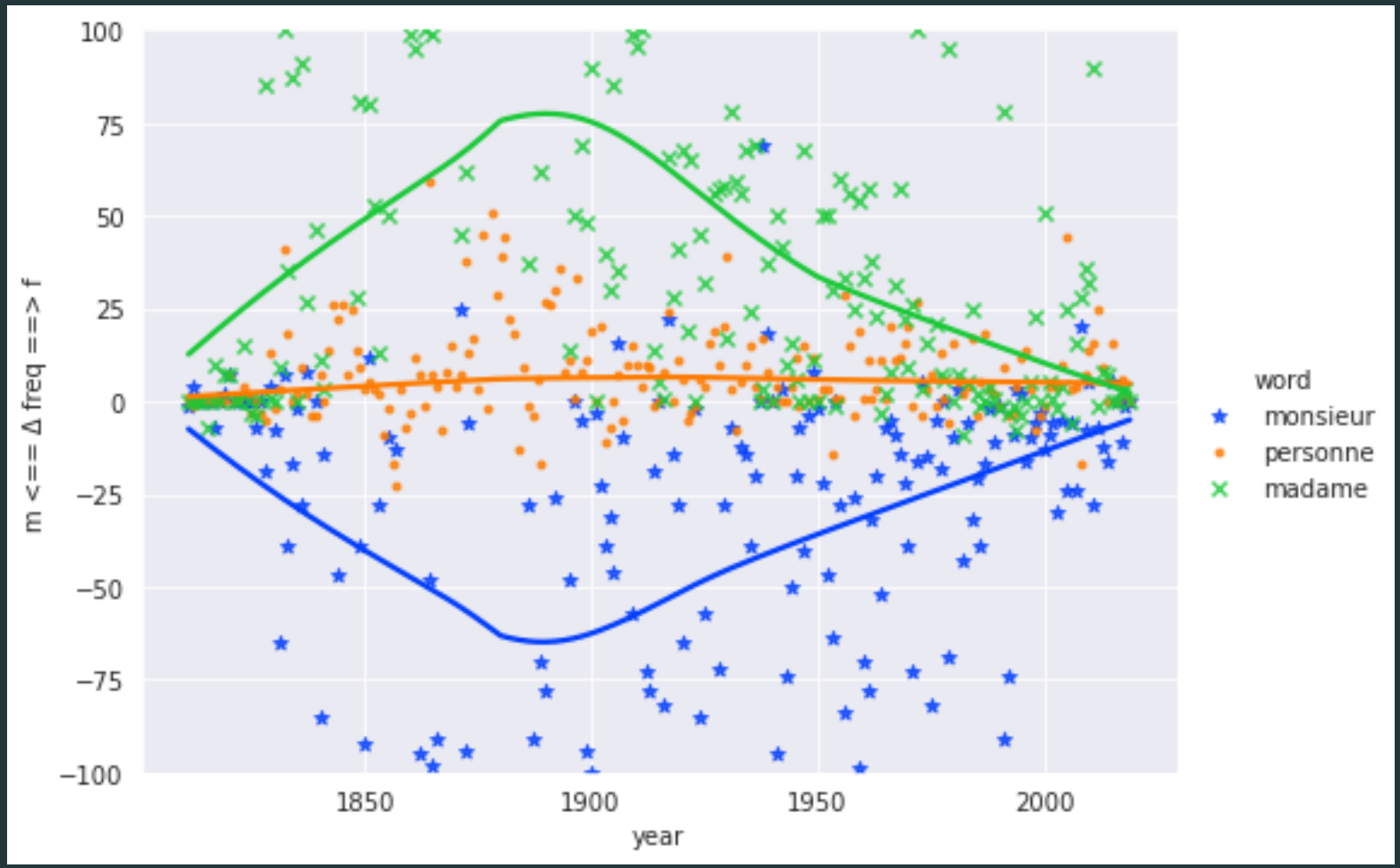

Figure 6 : Comment les hommes et les femmes sont caractérisés par des mots évidents : homme et femme

Quelques mots genrés dans la fiction 2/6

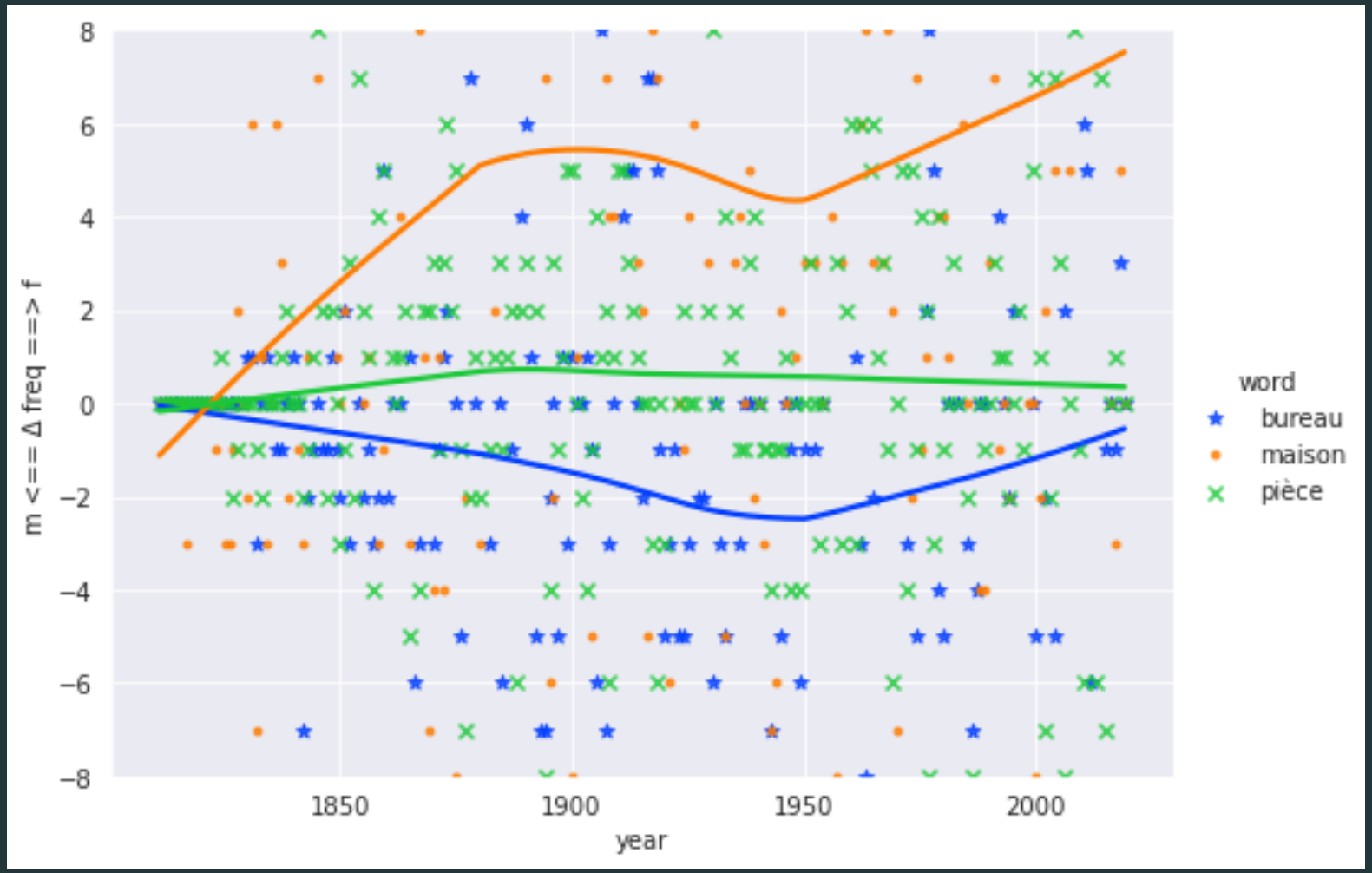

Figure 7 : Dans l'espace

Quelques mots genrés dans la fiction 3/6

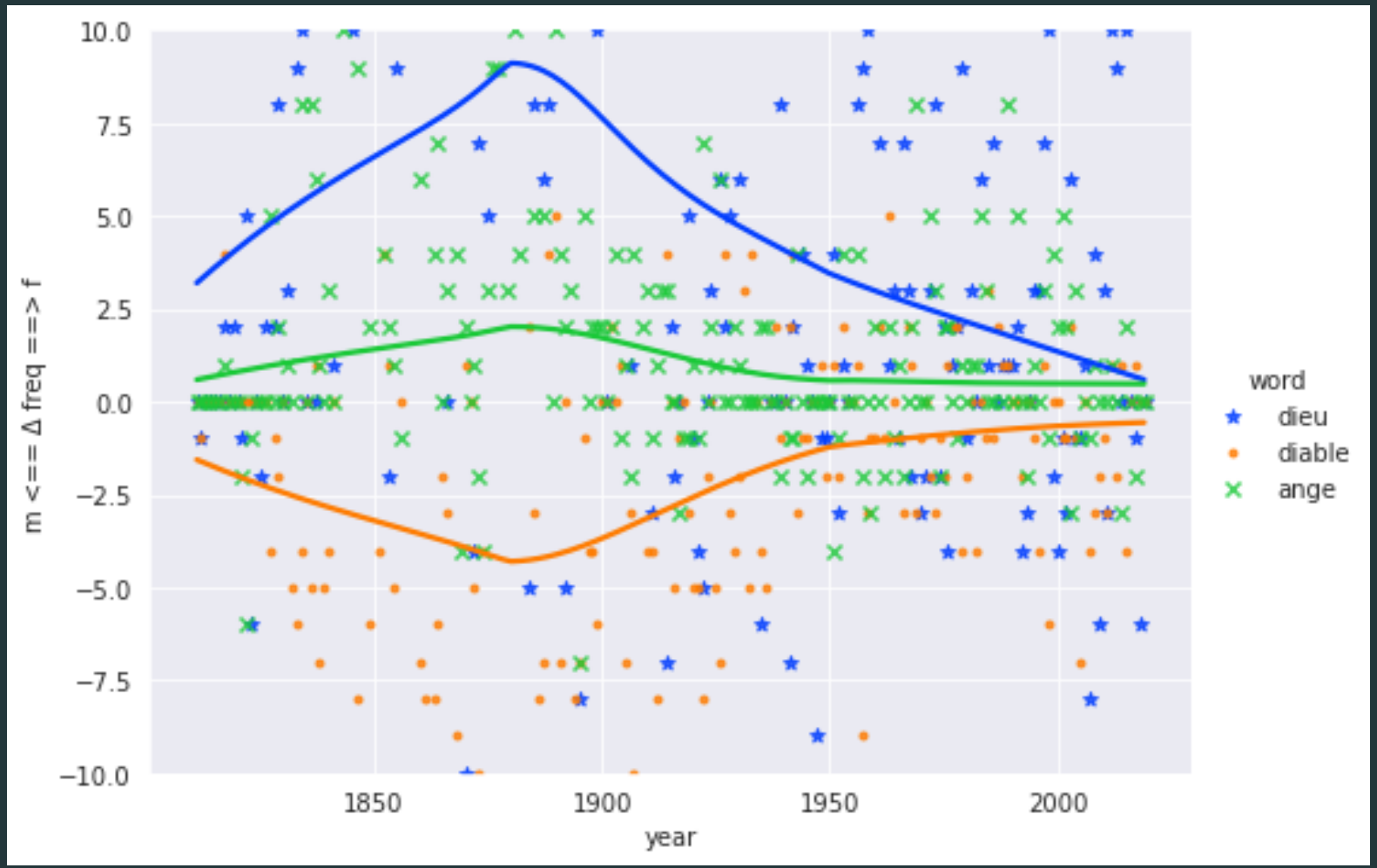

Figure 8 : Les mots religieux

Quelques mots genrés dans la fiction 4/6

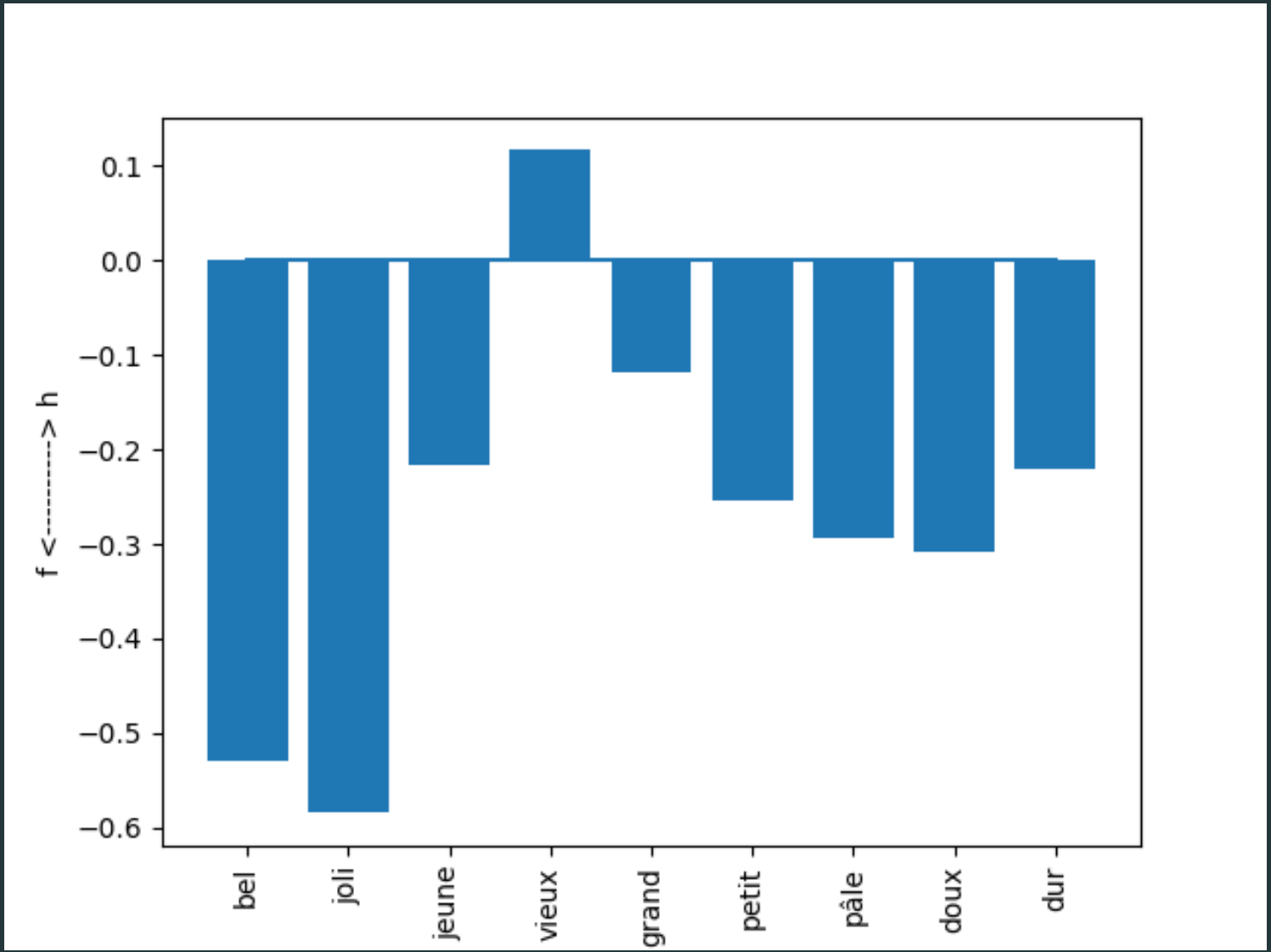

Figure 9 : Les adjectifs physiques

Quelques mots genrés dans la fiction 5/6

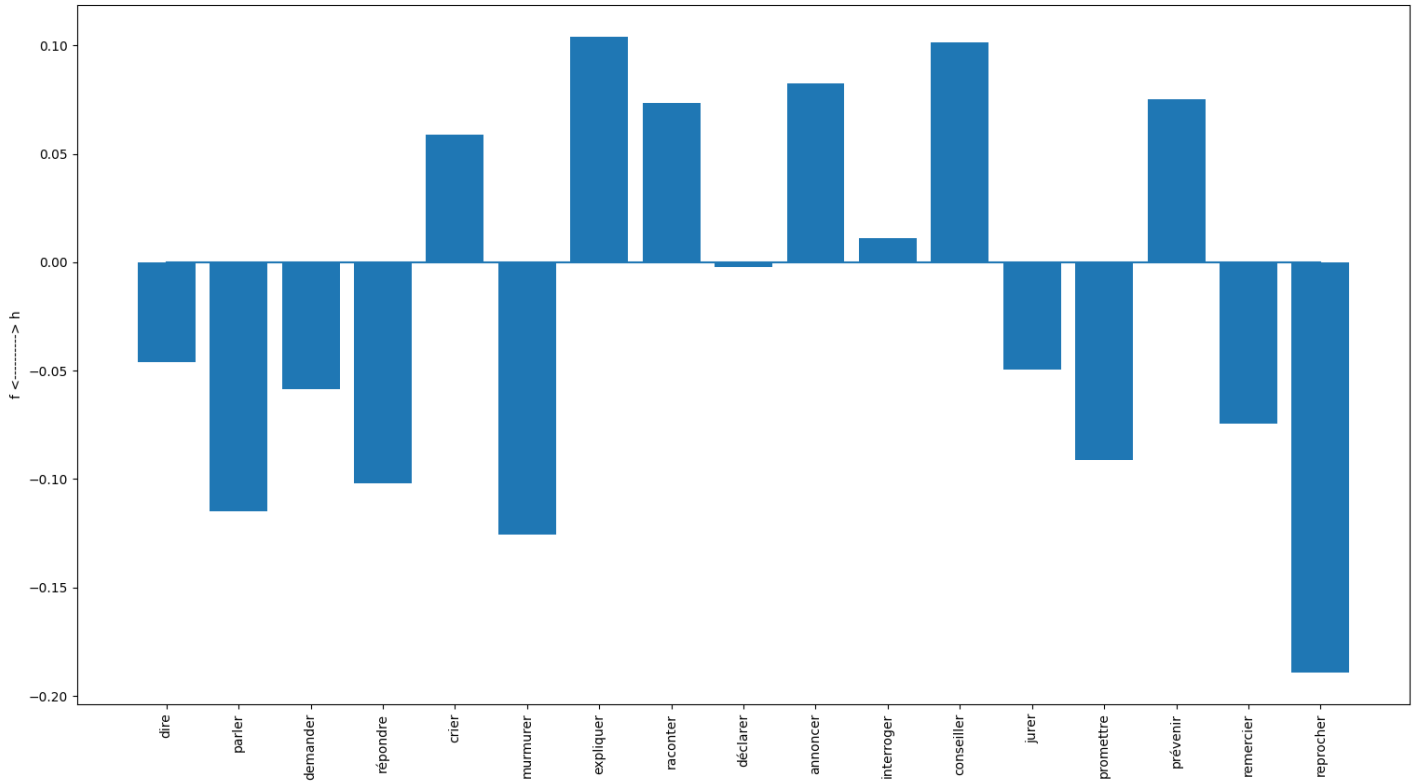

Figure 10 : Les prises de parole

Quelques mots genrés dans la fiction 6/6

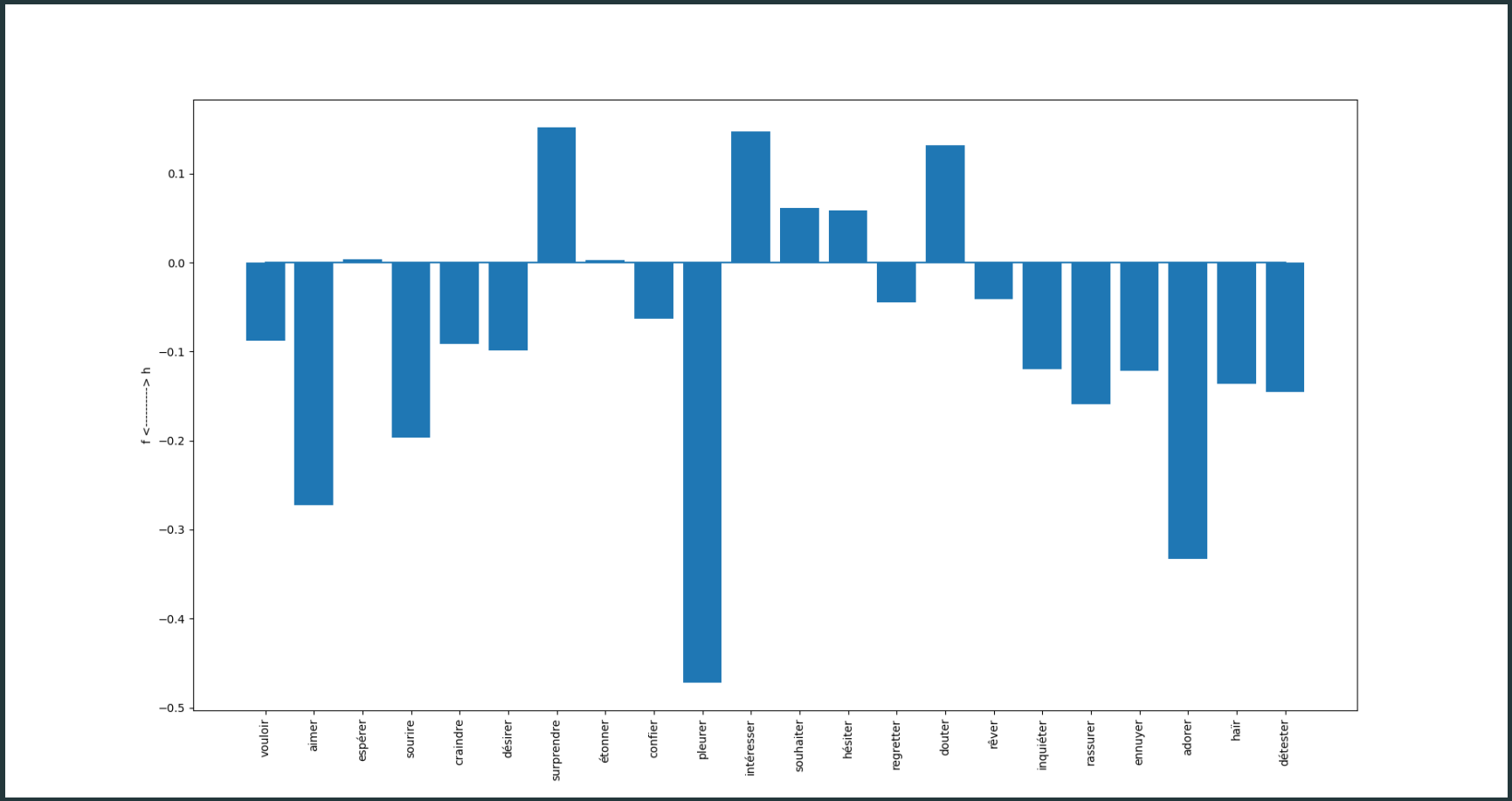

Figure 11 : Les émotions

Bilan

- Évaluation de la manière dont la caractérisation littéraire est liée à des biais de genre
- Il existe des mots individuels / champs lexicaux liés aux stéréotypes de genre
- La proportion de caractérisation des personnages féminins dépend fortement du genre de l'auteur
- Les auteurs masculins écrivent moitié moins sur les personnages féminins que les auteurs féminins

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 !

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 d'écueils 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

Bibliographie indicative i

A. Leblond.
Corpus chapitres.
Zenodo, Dec. 2022.

F. Moretti.
Conjectures on world literature.
New Left Review, 2000.

T. Underwood.
Distant horizons : digital evidence and literary change.
The University of Chicago Press, 2019.

T. Underwood, D. Bamman, and S. Lee.
The transformation of gender in english-language fiction.
Journal of Cultural Analytics, 3(2), 2018-02-13.