ENTRE CANON ET ARCHIVE, UNE ÉTUDE DES DYNAMIQUES TEXTUELLES À GRANDE ÉCHELLE.

Jean Barré

École Nationale des Chartes *jean.barre@chartes.psl.eu*

15 Décembre 2021

Table des matières

Introduction

Le canon littéraire

Le canon est une règle, un modèle, une norme à imiter. Les études littéraires s'intéressent à cet ensemble qui ne représente qu'une partie infime de la production littéraire des siècles passés.

La lecture distante[Moretti, 2013]

Concept que l'on doit à Moretti. Le but est d'explorer le passé littéraire (et ce qu'en dit la théorie et l'histoire littéraire) avec avec les méthodes computationnelles. Cela nous est rendu possible par la constitution de grand corpus d'ouvrages numérisés.

Traitement Automatique des Langues

La partie technique de mon mémoire se trouve dans l'implémentation de techniques du TAL pour parcourir les textes et en extraire les informations pertinentes.

Le mémoire de M1

Enjeux et hypothèses

- Mettre au jour l'existence de dynamiques textuelles entre canon et non-canon.
- Hypothèse d'une sélection successive des textes les plus informatifs au détriment d'une majorité d'autres.
- Cette simplification nous permet de rentrer dans les textes avec des éléments quantifiables, comme par exemple la variété lexicale.
- Travail fondé sur les recherches réalisés au Stanford Literary Lab, qui montraient des différences textuelles forte entre canon et non-canon[Hewitt, 2016].

Corpus ANR Chapitres

- 2968 romans du XIX^{e} et XX^{e} siècle
- Équilibre entre les œuvres appartenant ou non au canon.

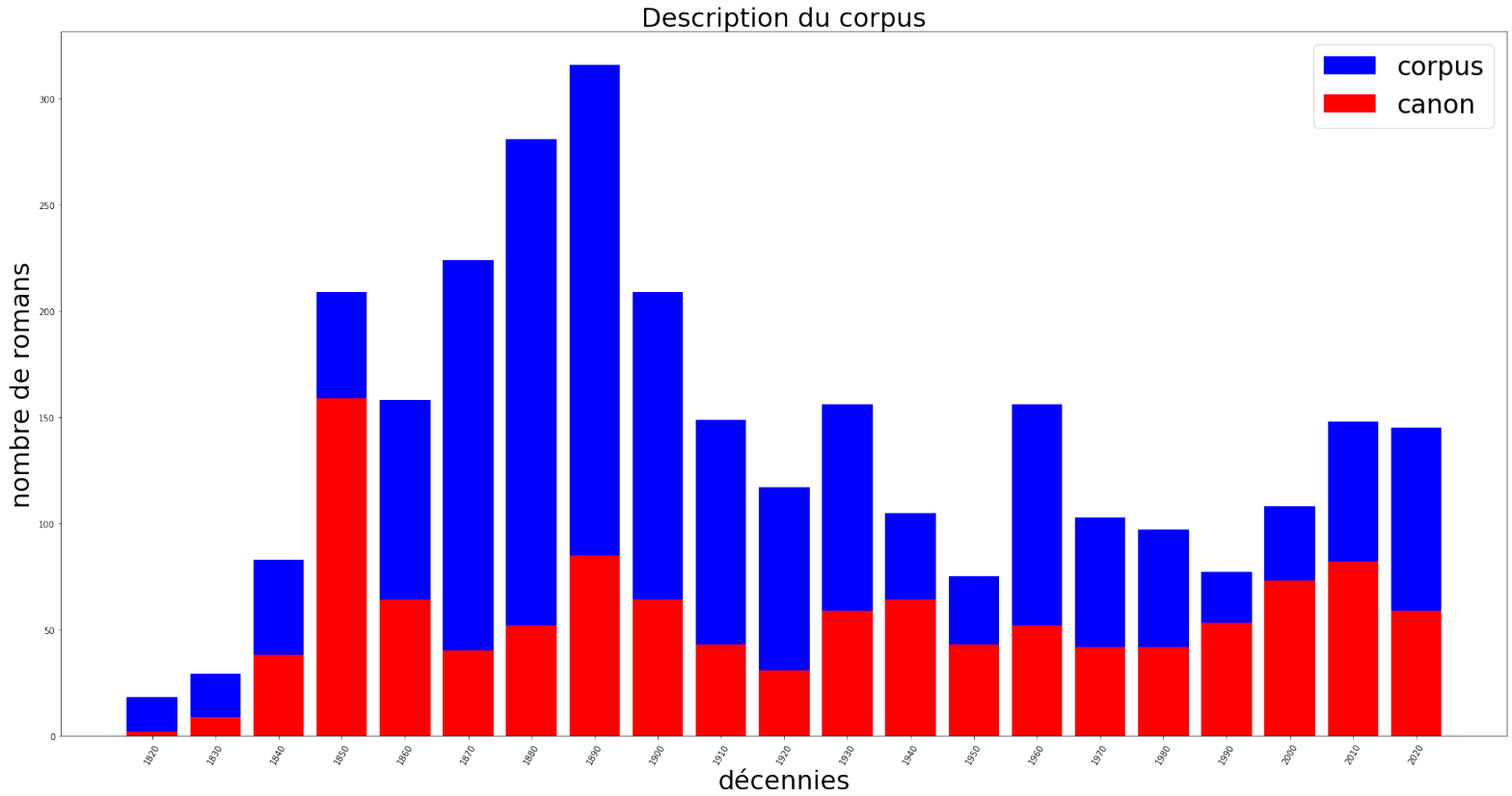

Figure – Répartition du corpus dans le temps

Chaîne de traitement

Trois composantes majeures :

1. Spacy
2. Stylométrie roulante
3. Entropie et Type-token ratio

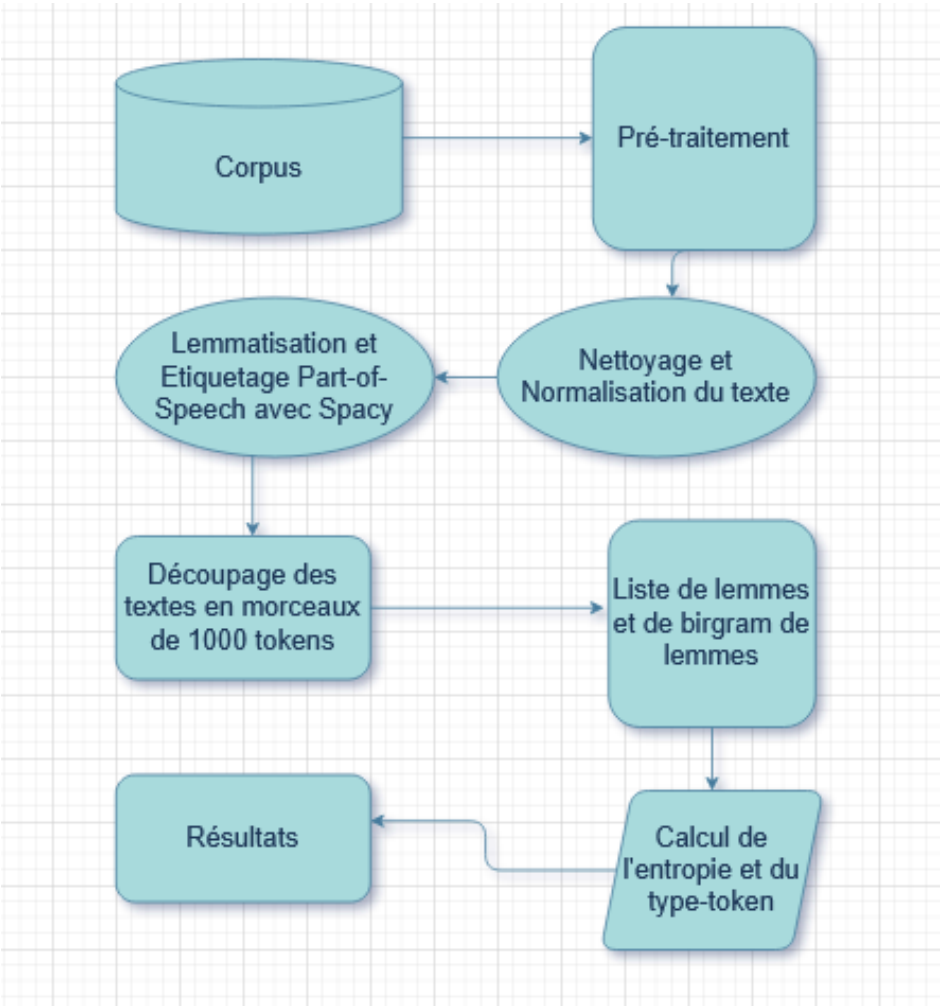

Figure – Chaîne de traitement

Les mesures

Theorem (Entropy)

$$H\prime = -\sum_{i=1}^{R} pilnpi \tag{1}$$

Theorem (Type-token ratio)

$$TTR = 100 * nbtypes/nbtokens \tag{2}$$

Visualisation des résultats - 1/2

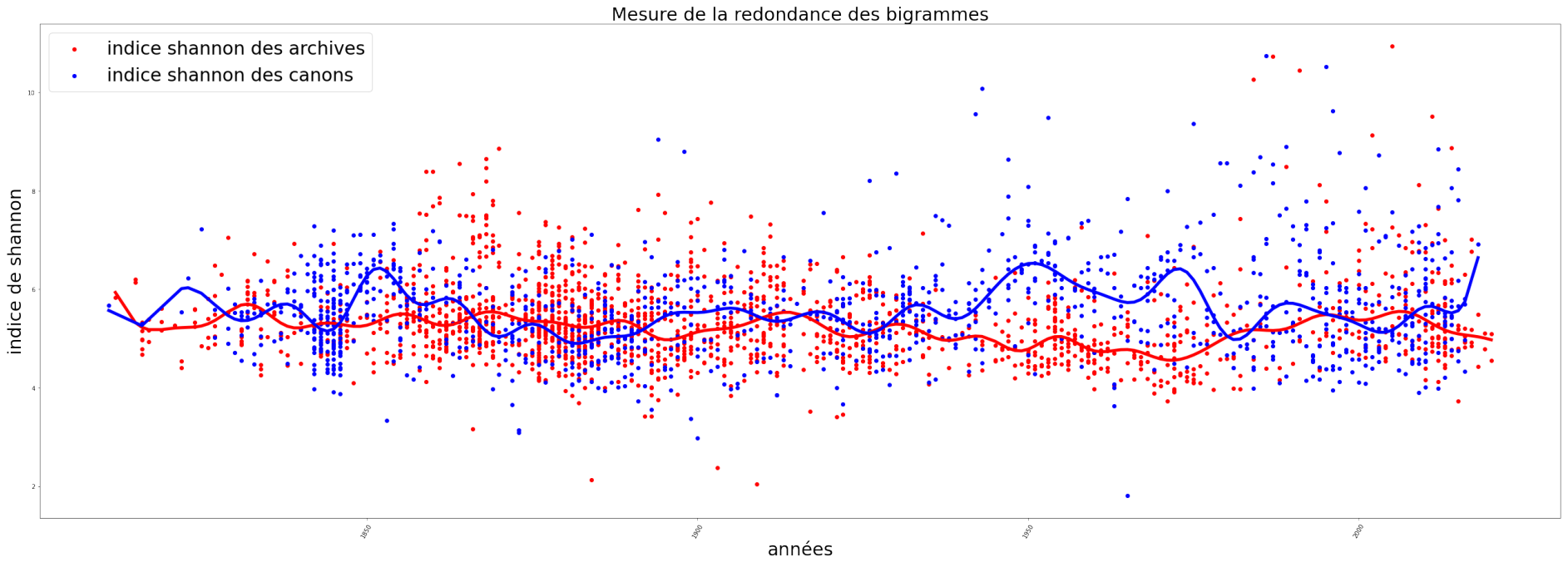

Figure – Mesure de la redondance des bigrammes

Visualisation des résultats - 2/2

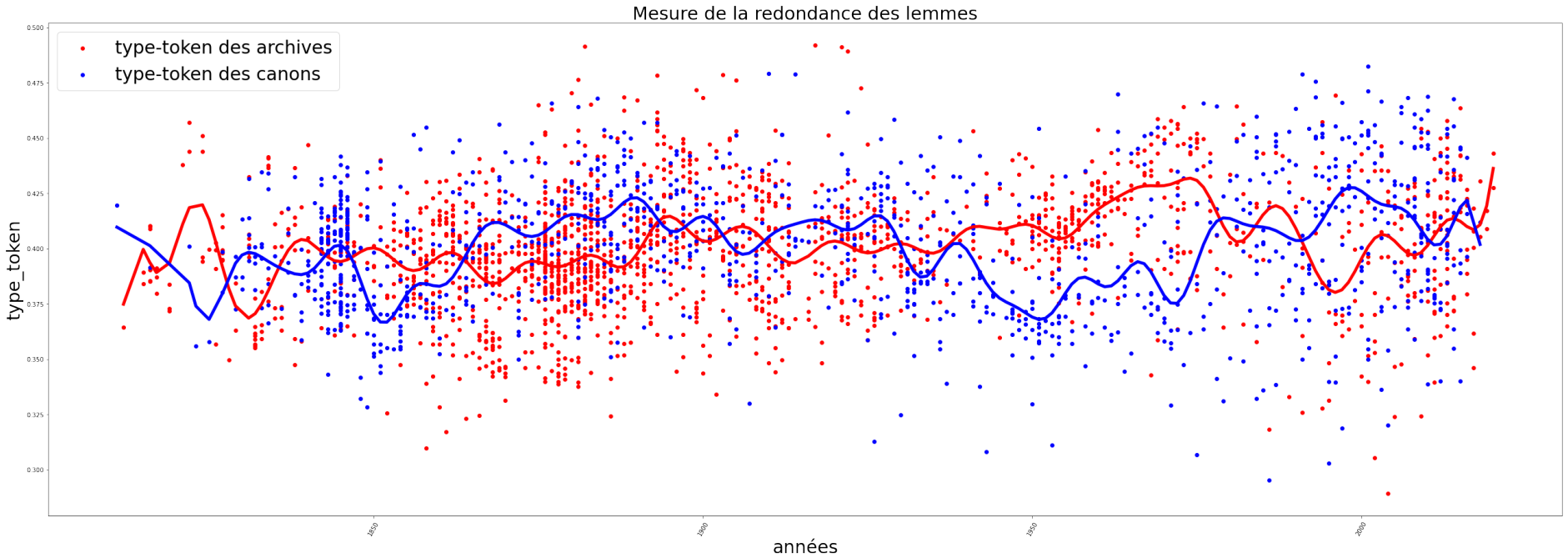

Figure – Mesure de la redondance des lemmes

Analyse des résultats

- Nous n'obtenons pas les mêmes résultats que dans l'étude du Literary Lab.
- L'hypothèse n'est pas validée empiriquement.
- On constate une démarcation de nos deux groupes de textes sur une période de 50 ans.
- Les textes canoniques connaissent une hausse de redondance sur cette période. Comment interpréter cela ?

Conjectures

Les mesures détecteraient le Nouveau Roman :

1. La situation temporelle correspond
2. Registre de langue familier
3. Oralité du discours

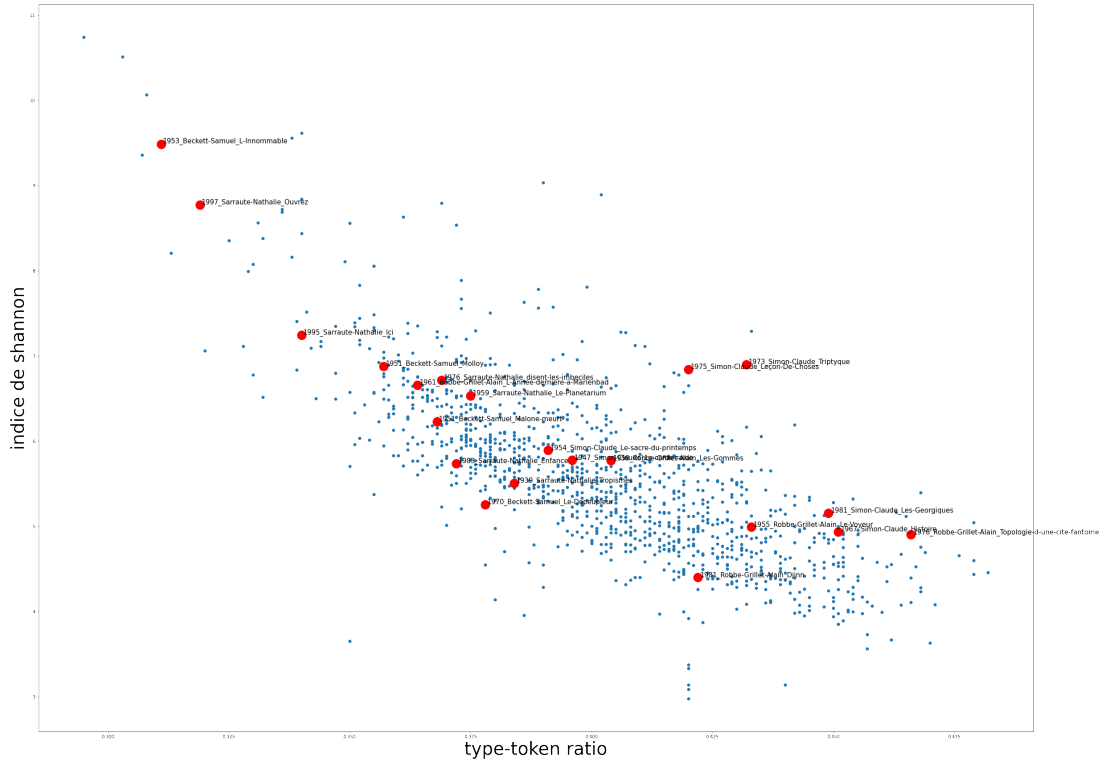

Figure – Redondance dans les romans canoniques

Les recherches en cours

Constats et Limites du mémoire de M1

- Le corpus Chapitres désigne un texte comme étant canonique à partir du nombre de résultats d'une requête sur fabula.org.
- La canonicité est définie de manière binaire à l'échelle des auteurs et non des ouvrages.
- Une hypothèse (le canon littéraire est peu redondant - plus informatif) un peu légère pour saisir des liens potentiels entre une réalité textuelle et les contextes de la production littéraire.

Définir un Canon Littéraire

- Le prestige littéraire comme probabilité qu'un auteur soit discuté, étudié et enseigné dans le champ littéraire[Bourdieu, 1992] contemporain.
- Pour identifier cela je détermine un indice de canonicité entre 0 et 1, qui décrit à quel point un texte est canonique.

Plusieurs paramètres à cet indice :

1. Publication en oeuvre complète aux éditions de la Pléiade.
2. Publication avec appareil critique aux éditions Gallimard-Flammarion
3. Récupération des listes des textes au programme de l'agrégation depuis 1969
4. Nombre de citation dans le Lagarde et Michard

Formaliser la Littérarité

Concept de littérarité

Ce qui constitue le discours littéraire.[Molinié, 1993] L'idée est de détecter et de décrire les traits formels d'une esthétique littéraire.

Hypothèse

La littérarité d'un texte est un des critères de sélection des chercheurs en littérature pour déterminer si un texte vaut la peine d'être étudier. Cela expliquerait la persistance temporelle des textes.

Littérarité formelle :

1. Longueur des phrases.
2. Complexité syntaxique des phrases.
3. Distribution lexicale des textes.
4. Détection des stylèmes, fondés sur les mots outils

Modéliser et Prédire la Canonisation

- Recherche de corrélations entre littérarité et prestige littéraire.
- Entraîner un modèle statistique (SVM) sur les caractéristiques textuelles de la littérarité.
- Comparer les résultats des deux approches, entre les contextes et les caractéristiques textuelles des textes.

Conclusion

- Je cherche si les textes étudiés par les chercheurs successifs ont une composante textuelle qui les démarquent au sein de la production littéraire.
- Je cherche si le phénomène social de la reconnaissance littéraire est en relation avec un type particulier d'écriture.

Pistes de réflexion :

- Enrichir mon corpus de texte avec le contexte contemporain de réception.
- Questionner les standards du jugement esthétique et leur évolution au cours du temps.
- Entraîner plusieurs modèles sur plusieurs périodes de réception.
- Revenir à une lecture proche : Les morceaux de texte que le modèle considère comme le plus littéraire.

Bibliographie indicative


Franco Moretti (2013)
Distant reading


Ted Underwood (2019)
Distant horizons : digital evidence and literary change


Gilles Philippe, Julien Piat (2009)
La langue littéraire : une histoire de la prose en France de Gustave Flaubert à Claude Simon


Georges Molinié, Alain Viala (1993)
Approches de la réception


M. Algee-Hewitt, S. Allison, M. Gemma, R. Heuser, F. Moretti and H. Walser. (2016)
Canon/Archive, Large scale dynamics in the literary field


John Guillory (1998)
Cultural capital : the problem of literary canon formation


Pierre Bourdieu (1992)
Les règles de l'art : Genèse et structure du champ littéraire