École nationale des Chartes Université Paris, Sciences & Lettres

Jean Barré

Diplômé de licence de Lettres

Diplômé de licence d'Informatique

OPÉRATIONNALISER LA CANONICITÉ

Étude des dynamiques à grande échelle des processus de canonisation

Mémoire de deuxième année du master

« Humanités Numériques »

2022

Résumé

Ce projet interroge le canon littéraire, une notion construite avec les biais de la société et modelée par les réceptions successives. L'objectif de ce rapport est de mettre en lumière l'existence de dynamiques textuelles qui assurent une longévité exceptionnelle à certains ouvrages et menacent au contraire la transmission d'une majorité d'autres. Les méthodes quantitatives du traitement automatique des langues et de l'apprentissage machine nous permettent de mettre au jour une esthétique intrinsèque au canon littéraire. Nous proposons une modélisation statistique de la canonicité avec des résultats prédictifs allant jusqu'à 90% d'efficacité.

Mots-clés : études littéraires computationnelles ; canon littéraire ; classique ; littérature ; stylométrie ; histoire littéraire ; humanités numériques ; lecture distante ; traitement automatique de la langue ; apprentissage machine

Informations bibliographiques : Jean Barré, *Opérationnaliser la canonicité. Étude des dynamiques à grande échelle des processus de canonisation*, mémoire de master 1 « Humanités Numériques », dir. [Thierry Poibeau, Jean-Baptiste Camps], Université Paris, Sciences & Lettres, 2021.

Abstract

This project will interrogate the literary canon, a notion constructed with the biases of society and shaped by successive receptions. The aim of this report is to shed light on the existence of textual dynamics which ensure an exceptional longevity to some works and threaten the transmission of a majority of others. The quantitative methods of natural language processing and machine learning allow us to uncover an intrinsic aesthetic of the literary canon. We propose a statistical modeling of canonicity with predictive results of up to 90% accuracy.

Keywords : computational literary studies ; canon ; classic ; archive ; literature ; stylometry ; digital humanities ; distant reading ; text mining ; natural langage processing ; machine learning

Bibliographic Information : Jean Barré, *Operationalizingcanonicity. Large scale dynamics of canonization processes*, M.A. thesis « Digital Humanities », dir. [Thierry Poibeau, Jean-Baptiste Camps], Université Paris, Sciences & Lettres, 2021.

Remerciements

Je tiens à remercier chaleureusement mon directeur de mémoire Monsieur Thierry Poibeau, pour son enthousiasme pour le projet, sa grande disponibilité, et ses innombrables retours. Je tiens également à remercier mon directeur de mémoire Monsieur Jean-Baptiste Camps pour sa bienveillance, ses réflexions toujours pertinentes et le support technique qu'il m'a apporté.

J'aimerais aussi remercier Messieurs Simon Gabay et Chahan Vidal-Gorène pour leurs conseils et l'intérêt qu'ils ont montré pour ce travail.

Ce mémoire n'aurait pas été possible sans le soutien de mon amoureuse Marie Delille et mes amis proches, qui m'ont soutenu malgré mes trop longs exposés quotidiens sur le sujet.

Je tiens enfin à remercier mes parents, ma sœur et mon frère pour leurs corrections bienvenues, mais surtout pour leur amour et leur soutien indéfectible.

Table des matières

Introduction

Les plus beaux livres sont écrits dans une sorte de langue étrangère.

Marcel Proust, Contre Sainte Beuve, 1954

On peut aussi construire un modèle du *processus de canonisation qui conduit à l'institution des écrivains*, à travers une analyse des différentes formes que le panthéon littéraire a revêtues, aux *principes de classement* eux-mêmes *préconstruits.*

Pierre Bourdieu, Les règles de l'art, 1992

Stendhal est aujourd'hui considéré comme un auteur incontournable de la littérature française. Pourtant, la présence de cet auteur majeur dans la conscience collective n'a pas toujours été évidente. La renommée de son œuvre n'arrive que longtemps après sa mort. Quels processus ont joué un rôle dans la canonisation de cet auteur ? La reconnaissance des romans de Stendhal est un exemple parmi les milliers de romans du XIX^{e}et du XX^{e}siècle, alors pourquoi et comment se démarquent-ils des autres ouvrages laissés à l'abandon littéraire ?

L'ensemble des auteurs que l'on appelle *les classiques* forme le canon littéraire. Nous pouvons définir ce dernier comme le résultat d'une tradition sélective [1] dont la mémoire collective a gardé le souvenir parce qu'elle leur confère un statut prééminent.

Historiquement, les études sur les processus d'attribution de la valeur littéraire se sont intéressées aux contextes dans lesquels étaient produites les œuvres, et aux processus de canonisation des auteurs et de leurs œuvres. Dans *Les règles de l'art* [2], un ouvrage majeur concernant notamment le prestige littéraire, Pierre Bourdieu montre que les mécanismes derrière la distribution du prestige littéraire sont fondés sur des facteurs appartenant aux contextes des œuvres. Ces facteurs sont liés entre autres aux dynamiques de pouvoir au sein du champ littéraire. Les contextes de production expliqueraient deux « modes de vieillissement » des auteurs, distinguant, d'une part, les auteurs consacrés par la critique et l'enseignement et, d'autre part, les auteurs voués à une *mort* littéraire rapide.

Le prestige littéraire est donc une notion complexe à envisager et les mécanismes derrière cette filtration temporelle sont nombreux, qu'ils soient liés à des politiques culturelles ou à des critères autonomes, d'ordre esthétique et critique.

Nous voulons, dans ce mémoire, revenir aux textes et à leur contenu. Notre hypothèse est de dire qu'il y a une esthétique particulière dans le canon, et qu'on peut la détecter et la décrire. Notre travail consistera à présenter les différents éléments mis en place pour prédire la *canonicité* avec le contenu textuel des ouvrages. Nous nous fonderons sur des outils du traitement automatique des langues, de la stylométrie et des techniques d'apprentissage machine pour caractériser cette notion de canon littéraire.

Les humanités numériques, et plus précisément les études littéraires computationnelles, s'inscrivent dans une nouvelle approche de compréhension de la production culturelle des siècles passés. Un des concepts sur lequel se fonde cette nouvelle approche est le *« distant reading »* [3], théorisé par Franco Moretti dans les années 2000. Cette *« lecture distante »* a pour ambition d'explorer le passé littéraire avec des méthodes scientifiques

1. Griselda Pollock, *Differencing the canon : feminist desire and the writing of art's histories*, London ; New York, 1999 (Re visions).

2. Pierre Bourdieu, *Les règles de l'art genèse et structure du champ littéraire*, Paris, 1992.

3. Franco Moretti. « Conjectures on World Literature », *New Left Review*, 2000. On notera une synthèse des écrits autour de cette notion dans un ouvrage dédié au *« distant reading »*.Franco Moretti, *Distant reading*, Country : GB Contient des textes déjà publiés. Includes bibliographical references and index. Vendor-supplied metadata., London, 2013.

sur des corpus massifs et numérisés.

Le nouveau paradigme ouvert par le numérique en stylistique et en histoire littéraire est précisément de dépasser la seule étude du canon littéraire. Quelques centaines d'œuvres composent ce canon. Des milliers d'autres en sont exclues, et sortent de ce fait du champ des études littéraires. Un des constats posé par Ted Underwood [4] est que les études littéraires ne connaissent pas vraiment les grandes lignes structurantes de l'histoire littéraire. En effet, la discipline s'est focalisée des années durant sur une poignée d'œuvres du canon littéraire. Cette sélection est assimilable pour certains chercheurs à un *« abattoir littéraire »* [5]. Ce dernier contient selon Moretti une histoire bien différente de celle contenue dans les canons académiques, et le *« distant reading »* peut permettre de prendre en considération ces milliers de volumes oubliés.

L'utilisation de techniques du Traitement Automatique des Langues (TAL) pour l'analyse de corpus littéraires a donné lieu à de nombreuses études, que ce soit pour la modélisation du genre [6], du suspense [7], des thèmes [8], ou de la paternité des œuvres [9]. Il s'agit d'explorer des motifs et des tendances historiques sur la longue durée [10], à partir de corpus littéraires massifs et numérisés. C'est un domaine de recherche très actif aux États-unis et en Europe, mais beaucoup moins en France, bien qu'il existe également de puissants outils du TAL pour le français, ainsi que des corpus accessibles.

Cette approche ne vise à effacer ni les siècles de lecture proche ni la subjectivité du chercheur. Elle permet d'élargir la focale pour interroger nos savoirs critiques hérités mais aussi pour en produire de nouveaux.

Ce mémoire s'inscrit nécessairement dans une étude de la réception et un temps sera consacré à recueillir des méta-données pour construire un canon littéraire que l'on questionnera dans les textes. L'objectif premier est d'enrichir un corpus avec le contexte

4. Ted Underwood, *Distant horizons : digital evidence and literary change*, Chicago, 2019, chap. 1 p 1 - 33

5. F. Moretti, « The Slaughterhouse of Literature », *Modern Language Quarterly*, 61–1 (1er mars 2000), p. 207-228, DOI : 10.1215/00267929-61-1-207.

6. Ted Underwood, *The life spans of genres* in T. Underwood, *Distant horizons...*, p 34 - 67

7. Andrew Piper, Sunyam Bagga, Laura Monteiro, Andrew Yang, Marie Labrosse et Yu Lu Liu, « Detecting Narrativity Across Long Time Scales », dans *Proceedings of the Conference on Computational Humanities Research CHR2021*, dir. Maud Ehrmann, Folgert Karsdorp, Melvin Wevers, Tara Lee Andrews, Manuel Burghardt, Mike Kestemont, Enrique Manjavacas, Michael Piotrowski et Joris van Zundert, Amsterdam, the Netherlands, 2021 (CEUR Workshop Proceedings), t. 2989, p. 319-332, URL : http://ceur-ws.org/Vol-2989/#long_paper49.

8. Matthew L. Jockers et David Mimno, « Significant themes in 19th-century literature », *Poetics*, 41–6 (déc. 2013), p. 750-769, DOI : 10.1016/j.poetic.2013.08.005.

9. Florian Cafiero et Jean-Baptiste Camps, « 'Psyché' as a Rosetta Stone ? Assessing Collaborative Authorship in the French 17th Century Theatre », dans *Proceedings of the Conference on Computational Humanities Research CHR2021*, dir. Maud Ehrmann, Folgert Karsdorp, Melvin Wevers, Tara Lee Andrews, Manuel Burghardt, Mike Kestemont, Enrique Manjavacas, Michael Piotrowski et Joris van Zundert, Amsterdam, the Netherlands, 2021 (CEUR Workshop Proceedings), t. 2989, p. 377-391, URL : http://ceur-ws.org/Vol-2989/#long_paper51.

10. Fernand Braudel, *La Méditerranée et le monde méditerranéen à l'époque de Philippe II. Fernand Braudel*, Dixième édition, Malakoff, 2017.

social de production et de réception.

Dans un premier temps, nous définirons les enjeux de recherche autour de la notion de prestige littéraire. Dans un second temps, nous construirons un canon littéraire fondé sur une multiplicité de facteurs diagnostiqués par la recherche en sociologie et en histoire littéraire. Dans un troisième temps, nous modéliserons la notion de canon littéraire pour mettre au jour une esthétique propre. Enfin, un retour qualitatif sur les inférences statistiques de nos différents modèles sera nécessaire pour caractériser avec plus de finesse cette esthétique canonique.

Première partie

Préliminaires

Chapitre 1

Les modalités du canon littéraire

1.1 *Canon* et *classiques* de la littérature

Nous utiliserons dans ce mémoire la notion de *canon littéraire* dans le sens où ce dernier est composé d'ouvrages *classiques*, ou d'auteurs *classiques*.

Le mot canon vient du grec « Κανων » qui signifie « roseau » ou « tige », littéralement « droit comme une tige de roseau » [1] utilisé comme un instrument de mesure. Avec le temps le terme prend le sens de « règle » ou de « loi », et c'est avec cette signification qu'il arrive dans les langages européens modernes. C'est au quatrième siècle après JC [2], avec le canon biblique, que le mot prend le sens actuellement utilisé dans les études littéraires. L'Église instaure une liste de textes formant les Saintes Écritures. Le canon avait alors une valeur d'autorité qui permettait à l'Église de contrôler un récit commun pour la postérité.

La notion a été introduite dans les études littéraires pour désigner les syllabus des universités et les textes qui y étaient étudiés. Selon certains chercheurs, le canon est nécessaire à l'enseignement de la littérature : « L'étude institutionnelle de la littérature est inconcevable sans un canon. Sans canon, sans corpus ou syllabus de textes exemplaires, il ne peut y avoir de communauté interprétative. » [3]. Le canon est ainsi l'ensemble des textes sur lesquels se fonde la discipline et la recherche en littérature. Pour Pascale Casanova, le canon « incarne la légitimité littéraire elle même, c'est à dire ce qui est reconnu comme *la littérature*, et qui servira d'unité de mesure spécifique » [4].

W. Harris distingue différentes fonctions du canon littéraire [5] : Il permet de « fournir

1. William Marx, cours au collège de France, https://www.college-de-france.fr/site/william-marx/course-2021-04-13-14h30.htm

2. John Guillory, « Canon » in Frank Lentricchia et Thomas McLaughlin, *Critical terms for literary study*, OCLC : 813244781, Chicago [Ill., 2012, URL : http://www.credoreference.com/book/uchicagols (visité le 22/11/2021), p.233-245

3. Howard Felperin, *Beyond deconstruction : the uses and abuses of literary theory*, Oxford, 1985, les traductions des citations ont été réalisés par mes soins

4. Pascale Casanova, *La république mondiale des lettres*, Édition revue et corrigé e, Paris, 2008 (Points Série essais, 607).

5. Wendell V. Harris, « Canonicity », *PMLA/Publications of the Modern Language Association of America*, 106–1 (janv. 1991), p. 110-121, DOI : 10.2307/462827.

des modèles, de transmettre l'héritage de la pensée, par la provision de savoirs culturels nécessaires pour interpréter les textes du passé ». Le canon crée des cadres de référence communs pour faire société par identification à un ensemble de mythes et de textes sacrés. Un des problèmes soulevés par ce genre d'approche du canon est que si l'on se penche sur les listes des grands auteurs qui les composent, on trouvera très peu de femmes, et encore moins d'auteurs non blancs, ou de classe sociale défavorisée [6]. Les processus de formation du canon littéraire sont souvent exclusifs, et il faudrait déconstruire les préjugés sur lesquels se fonde le canon.

Les polémiques sur le canon littéraire ont été particulièrement virulentes dans les campus universitaires étasuniens à partir des années 1980. La synthèse d'Eric Fassin [7] est à cet égard fort instructive. Deux camps s'affrontent, avec d'un côté les défenseurs du canon (compris comme la quintessence de la tradition littéraire occidentale [8]) au nom des valeurs universelles qu'il incarne. De l'autre côté, ses détracteurs, qui relèvent les conditions sociologiques, géographiques, idéologiques et culturelles de la production des critères de définition et de sélection du canon littéraire [9].

En France, la notion de canon n'est pas si répandue et doit son existence aux polémiques venant d'outre-Atlantique. C'est le terme *classique* qui est utilisé, et qui n'est d'ailleurs pas dédié uniquement à l'espace littéraire. Le nom *les Classiques* désigne aujourd'hui les grands écrivains de la littérature française, mais cela n'a pas toujours été le cas.

Le mot *classique* vient du vocabulaire de la richesse et de la propriété. Du latin *classicus*, signifiant « citoyen de la première classe » [10]. L'adjectif *classique* prend son sens de jugement esthétique au XVII^{e}siècle, où il signifie ce qui mérite d'être copié ou de servir de modèle. A la fin du XVII^{e}siècle, la signification dérive et prend le sens de ce qui était enseigné en classe. Au cours du XVIII^{e}siècle, le terme désigne les auteurs antiques grecs ou latins.

Finalement au cours du XIX^{e}siècle, on appelle *classiques*, par opposition aux auteurs romantiques contemporains, les auteurs du temps de Louis XIV et cette époque comme le classicisme français [11]. Ainsi, dans son sens le plus restreint, le *classique* désigne les auteurs de théâtre du XVII^{e}siècle, considéré comme le grand siècle de la littérature française puisque cette période littéraire s'est le plus rapproché de la perfection antique.

6. John Guillory, *Cultural capital : the problem of literary canon formation*, [Nachdr.], Paperback ed. 1994, Chicago, 1998.

7. Éric Fassin, « La chaire et le canon. Les intellectuels, la politique et l'Université aux États-Unis », *Annales. Histoire, Sciences Sociales*, 48–2 (1993), p. 265-301, DOI : 10.3406/ahess.1993.279133.

8. Harold Bloom, *The Western canon : the books and school of the ages*, OCLC : 624578000, New York, 1994, URL : http://site.ebrary.com/id/10879075 (visité le 13/04/2022).

9. Marie-Pierre Harder, *(Dé)construire le canon Introduction*, 2013, URL : http://www.crlc.paris-sorbonne.fr/pdf_revue/revue4/1_INTRO_Harder.pdf (visité le 13/04/2022).

10. Trésor de la langue française informatisé ; https ://atilf.fr/ressources/tlfi

11. Alain Viala, « Qu'est-ce qu'un classique ? », *Littératures classiques*, 19–1 (1993), p. 11-31, DOI : 10.3406/licla.1993.1737.

L'enjeu du terme *classique* se joue au niveau de la réception. Ce n'est pas l'auteur qui choisit d'être classique, c'est l'institution scolaire ou la critique, et ses représentants, qui choisissent et émettent le jugement esthétique et politique de son appartenance au cercle des *classiques*.

Sainte-Beuve, critique et écrivain français conservateur du milieu XIX^{e}siècle nous donne un bon échantillon de la perception des classiques au XIX^{e}siècle :

> « Un vrai classique, comme j'aimerais à l'entendre définir, c'est un auteur qui a enrichi l'esprit humain, [...] qui a découvert quelque vérité morale non équivoque, [...] qui a parle à tous dans un style à lui et qui se trouve aussi celui de tout le monde, dans un style [...] nouveau et antique, aisément contemporain de tous les âges. » [12]

Plusieurs éléments sont intéressants. La notion de style - que nous discuterons plus tard dans cette partie - apparaît centrale pour définir le classique. Sainte-Beuve présente le style classique de façon ambiguë. Il est à la fois personnel et commun, « nouveau et antique ». Il y a une forme de sacralisation du classique, qui est selon Sainte-Beuve un idéal littéraire que peu d'écrivain ont atteint. Sainte-Beuve essentialise le classique en le rapprochant de la notion de vérité, et justifie ainsi la hiérarchie entre les classiques et les autres.

1.2 Le canon et l'enseignement

La supériorité esthétique mise en avant par Sainte-Beuve n'est pas suffisante pour expliquer les différents processus de canonisation des auteurs. La relativité du jugement esthétique des critiques ne permet pas d'expliquer la stabilité du canon littéraire.

C'est dans l'enseignement de la littérature que l'on peut constater les dynamiques de construction et de modification du canon littéraire français. L'institution scolaire est un des lieux majeurs où le canon s'élabore. En effet, elle produit des panthéons d'auteurs et de textes, souvent sous la forme de morceaux choisis [13] pour éduquer des générations d'écoliers. Dès la première liste nationale de 1803, le canon des auteurs français est constitué. Il est repris ensuite pour être diffusé grâce au système scolaire centralisé.

L'institution scolaire construit une représentation de la littérature qui lui est propre. Elle en détermine le bon usage, avec des découpages chronologiques (périodisations, écoles, générations), une utilisation de catégories (romantisme, naturalisme, surréalisme), et enfin l'élaboration d'un canon par une sélection d'auteurs. Ces classiques sont le fruit d'une sélection particulièrement étroite. Ils sont présentés comme des modèles, véhiculant une

12. Sainte-Beuve, *Causeries du lundi*, cité dansAntoine Compagnon, « Sainte-Beuve and the Canon », *MLN*, 110–5 (1995), Publisher : Johns Hopkins University Press, p. 1188-1199, URL : http://www.jstor.org/stable/3251396 (visité le 08/03/2022)

13. Martine Jey, *La littérature au lycée : invention d'une discipline (1880-1925)*, Paris, 1998 (Recherches textuelles, no 3).

norme esthétique particulière. C'est cette norme qui nous intéresse et que nous voulons repérer quantitativement.

Le canon littéraire construit par l'institution scolaire devient ainsi matrice de formes spécifiques en tant qu'il est le représentant de ce qui est considéré comme de la *bonne* littérature. Pourtant le canon n'est pas monolithique et s'ouvre au cours du temps. Les garants du bon ordre littéraire (l'institution scolaire et, dans une moindre mesure, la critique) nourrissent et ouvrent le canon sous la pression et l'usure du temps aux œuvres qui semblent le plus en accord avec une certaine idée de la littérature.

1.3 Un canon politique

Le canon littéraire porte une dimension politique, tant dans sa construction que dans ce qu'il représente en tant que tel. Selon Alain Viala, les auteurs retenus dans le canon « remplissent une fonction d'identification culturelle » [14], c'est à dire qu'ils représentent le socle commun de la construction culturelle d'une nation. Dans le contexte français, l'évolution du canon littéraire est le témoin de la légitimation et de l'affirmation d'une langue et d'une culture française par rapport au latin. La première liste institutionnelle d'auteurs de 1803 associe un auteur grec ou latin à un auteur français du XVII^{e}siècle. Avec la structuration et la centralisation du système scolaire sous la troisième république, le canon se cristallise et devient un objet politique [15]. La littérature dans l'enseignement, associée à un canon littéraire, se trouve investie d'un rôle d'éducation des masses et de diffusion de valeurs nationales. On peut trouver un exemple dans le canon des agrégations, qui est « ramassé autour de ce qui semble incarner les valeurs nationales » [16]. Différentes réformes de l'enseignement scolaire ont façonné le canon et la façon d'enseigner la littérature au cours du temps.

Après cette première ouverture du canon littéraire, ce dernier a longtemps été la chasse gardée des auteurs du XVII^{e}siècle, et très peu de genres littéraires étaient représentés. Seul le théâtre et la poésie étaient les genres autorisés, puisqu'ils étaient les garants d'un classicisme formel, modèle d'ordre, de clarté, le tout dans un équilibre et une harmonie parfaits.

A la fin du XIX^{e}siècle, la société évoluant vers une laïcisation des mœurs sociales, l'enseignement de la littérature fait face à des difficultés. Gustave Lanson, célèbre historien et critique littéraire, le constate :

> « C'est une absurdité de n'employer qu'une littérature monarchique et chrétienne à l'éducation d'une démocratie qui n'admet point de religion d'Etat. » [17]

14. A. Viala, « Qu'est-ce qu'un classique ? »...
15. A. Compagnon, *La Troisième République des lettres, de Flaubert à Proust*, Paris, 1983.
16. M. Jey, « Le canon aux agrégations du XIX ^{e} siècle », *Revue d'histoire littéraire de la France*, 114–1 (2014), p. 143, DOI : 10.3917/rhlf.141.0143.
17. Gustave Lanson, «L'étude des auteurs français», Revue universitaire, 1894

Pour G. Lanson, la République doit se doter de nouveaux « textes sacrés » [18], et le canon littéraire doit s'ouvrir aux écrivains et aux genres littéraires contemporains. Les missions de la littérature sont alors de moraliser les moeurs sociales et de fabriquer un consensus national. Le roman entre ainsi au panthéon de la littérature, et avec lui la plupart des auteurs romanesques considérés encore aujourd'hui comme des classiques de la littérature française.

Lors de la poussée des nationalismes du début du XX^{e}siècle, la littérature doit affirmer l'unité voire la supériorité culturelle de la France. Ces processus ont été étudiés par Anne-Marie Thiesse dans son livre *La fabrique de l'écrivain national : entre littérature et politique*.

> « La nationalisation de l'état passe au premier chef par un intensif travail d'éducation de masses, visant à inculquer dans l'ensemble de la population le sentiment d'appartenance commune » [19].

Ainsi, le canon littéraire incarne les fondements du récit national. Il peut être vu comme la construction politique de la culture nationale. C'est d'ailleurs un paradoxe que souligne Pascale Casanova dans son livre *La république mondiale des lettres* [20]. Les cultures européennes se fondent sur des canons nationaux, tout en revendiquant les valeurs transnationales et universelles de ces derniers, qui prendraient racine dans un passé lointain des antiques grecs et latins.

1.4 Le style et les classiques

La construction du canon littéraire est éminemment politique, sa mise en place et sa diffusion relevant de politiques culturelles de mise en avant d'une littérature spécifique, censée porter les valeurs des régimes successifs de l'histoire de France. Les architectes du canon ont fondé leur sélection sur des critères linguistiques. Ils ont sélectionné ce qui, selon eux, représentait le meilleur usage de la langue française. Le classique est un modèle de style, et c'est pour cela qu'il est enseigné. Si nous pouvons questionner cette sélection en tant qu'elle lui donne un aspect arbitraire, le jugement esthétique étant par nature relatif, nous ne pouvons pas l'écarter complètement.

On peut définir le style comme « l'ensemble des traits expressifs qui dénotent l'auteur dans un écrit » [21]. En littérature, on assimile souvent le *style* à l'expression de la singularité de l'auteur, en tant qu'il s'écarte de la *norme*. Pour Georges Molinié, grand stylisticien, « Il n'y a de style que dans la mesure où des régularités dans le choix permettent de

18. A. Compagnon, *Le démon de la théorie : littérature et sens commun*, Paris, 1998 (La couleur des idées).

19. Anne-Marie Thiesse, *La fabrique de l'écrivain national : entre littérature et politique*, Paris, 2019 (Bibliothèque des histoires).

20. P. Casanova, *La république mondiale des lettres*...

21. Trésor de la Langue Française informatisé, http://stella.atilf.fr/

caractériser une écriture » [22]. Le style est un ornement formel, qui est propre à chaque auteur.

Toujours selon Georges Molinié, les qualités stylistiques des classiques pourraient expliquer leur filtration temporelle :

> « Dans la masse des romans et des pièces de théâtre, très peu, [...] sont encore lus de nos jours : l'invention et la gestion des situations dramatiques sont bien moins en cause que la qualité, la force, l'éclat d'écriture. C'est d'abord cette différence stylistique qui a séparé Pradon de Racine ; c'est la phrase de Madame de La Fayette qui a durablement séduit, parmi d'autres écritures effectives, indépendamment des atavismes ou des ruptures de la composition et de l'invraisemblable des situations. [...] Il est difficile de nier cette primauté du style. » [23]

La pérennité du canon littéraire peut s'expliquer par la « primauté du style » présente dans les *classiques* de la littérature. La langue littéraire, que l'on pourrait définir comme la langue des écrivains, se caractérise par son style, en opposition à la langue de tous les jours, qui manque de style.

Il faudrait revenir sur notre définition du style, et à la tension qui existe entre norme et style. On a vu que le style était traditionnellement vu comme l'expression d'une singularité, c'est à dire d'un idiolecte propre à un auteur. Le style peut aussi être vu comme une généralité, une langue littéraire, c'est à dire un sociolecte utilisé par les écrivains. Selon la formule de Gilles Philippe, la littérature, au sens du canon littéraire, est un « conservatoire consacrant la norme et un laboratoire célébrant la liberté du locuteur [24] ». Le style est indissociable de ces deux aspects, mais le premier a longtemps été minimisé, au profit du second, par les études stylistiques.

Le style présent dans le canon littéraire serait « une norme étalon à partir de laquelle l'on attribue ou l'on refuse une valeur esthétique à une production langagière » [25]. Il y aurait une norme stylistique haute dans le canon littéraire ; c'est cette norme que nous voulons détecter quantitativement. Les méthodes quantitatives semblent les mieux adaptées pour envisager l'ensemble des caractéristiques qui forment l'esthétique de la langue littéraire, ou plus simplement, l'esthétique canonique.

Le style est une notion que les études quantitatives ont abordé. On peut notamment présenter l'article de Herrmann et al, « Revisiting style, a key concept in literary studies ». Ces chercheurs ont proposé une définition du style compatible avec les méthodes quantita-

22. *Qu'est-ce que le style ? actes du colloque international*, dir. Georges Molinié et Pierre Alain Cahné, 1re éd, Paris, 1994 (Linguistique nouvelle).

23. G. Molinié, « Style et littérarité », *Littératures classiques*, 28–1 (1996), p. 69-74, DOI : 10.3406/licla.1996.2519.

24. *La langue littéraire : une histoire de la prose en France de Gustave Flaubert à Claude Simon*, dir. Gilles Philippe et Julien Piat, Paris, 2009.

25. G. Philippe, *Pourquoi le style change-t-il ?*, Bruxelles, 2021.

tives : « Le style est une propriété du texte constitué par un ensemble de caractéristiques formelles, pouvant être observées quantitativement ou qualitativement » [26]. La puissance de calcul et l'étude de grand corpus dans le cadre des études littéraires computationnelles peut aider à désingulariser la notion de style. Nous proposons, dans ce mémoire, d'étudier le style comme un phénomène collectif, qui caractériserait l'esthétique spécifique du canon littéraire.

Dans le prochain chapitre, nous verrons comment définir l'approche des études littéraires computationnelles pour appréhender des concepts littéraires tels que celui du style.

26. J. Berenike Herrmann, Karina van Dalen-Oskam et Christof Schöch, « Revisiting Style, a Key Concept in Literary Studies », *Journal of Literary Theory*, 9 (2015), p. 25-52.

Chapitre 2

Approches quantitatives

2.1 La mesure dans les études littéraires

Grâce aux progrès de l'informatique et du traitement du langage naturel, l'utilisation d'approches statistiques et mathématiques dans le domaine des études littéraires a considérablement augmenté ces dernières années.

La notion de *« distant reading »* [1], définie en introduction n'est finalement qu'une nouvelle approche de description et de production de savoirs empiriques en histoire littéraire.

Ce nouveau champ de recherches permet d'augmenter la focale en prenant en compte la production littéraire dans un ensemble le plus grand possible. Cela permet de comprendre les contrastes entre des oeuvres tirées de différentes périodes ou de différents contextes sociaux.

Franco Moretti et le laboratoire littéraire de Stanford [2] ont contribué à développer cette nouvelle approche des textes littéraires fondée sur la mesure et le calcul de fréquences d'éléments du texte. Leurs pamphlets 1, « Quantitative Formalism : an Experiment » [3], 6, « "Operationalizing" : or, the function of measurement in modern literary theory » [4] et 12, « Literature, Measured » [5] conceptualisent le rôle de la mesure pour appréhender des objets culturels. Ces pamphlets fondent une discipline, les études littéraires computationnelles.

1. Franco Moretti. « Conjectures on World Literature », *New Left Review*, 2000. On notera une synthèse des écrits autour de cette notion dans un ouvrage dédié au *« distant reading »*.F. Moretti, *Distant reading*...

2. https ://litlab.stanford.edu/

3. Sarah Allison, Mark Algee-Hewitt, Ryan R. Heuser, Matthew Jockers, F. Moretti et Michael Witmore, « Quantitative Formalism : an Experiment », Pamphlets of the Stanford Literary Lab–1 (2011), URL : https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf..

4. F. Moretti, « "Operationalizing" : or, the function of measurement in modern literary theory », Pamphlets of the Stanford Literary Lab–6 (2013), URL : https : / / litlab . stanford . edu / LiteraryLabPamphlet6.pdf..

5. Id., « Literature, Measured », Pamphlets of the Stanford Literary Lab–12 (2016), URL : https://litlab.stanford.edu/LiteraryLabPamphlet12.pdf..

Ces dernières modifient l'approche conceptuelle de l'objet d'étude des analyses littéraires traditionnelles. Un roman n'est plus vu comme un objet physique singulier, mais plutôt comme un agrégat de signes caractéristiques et quantifiables, dont l'ensemble ne forme qu'un échantillon parmi un corpus de milliers d'autres ouvrages. Mark-Algee Hewitt, un chercheur du domaine, souligne que « l'enjeu de cette transformation est donc notre compréhension du roman lui-même : ce qu'il est, les informations qu'il contient et ce qu'il peut faire en tant qu'objet littéraire et critique » [6]. Le roman est ainsi envisagé comme un ensemble de données, dont la signification mathématique peut nous apprendre des éléments critiques sur un auteur, un genre littéraire ou même un espace culturel.

Cet examen systématique des données contenues dans les textes écrits est conceptualisé par Matthew L. Jockers comme une analyse à l'échelle macroscopique [7] de la littérature. Cette approche *macro-analytique* permet de comprendre « le degré auquel la littérature et les auteurs individuels qui la fabriquent répondent ou réagissent aux tendances littéraires et culturelles ». C'est une nouvelle manière de décrire la position de l'écrivain dans le champ littéraire.

Avec ces nouvelles méthodes, le chiffre et la mesure prennent ainsi une influence particulièrement importante. Ted Underwood explicite leur rôle : « Les chiffres ne sont pas plus objectifs que les mots. Ils ne sont que des signes qui permettent aux observateurs humains de se débattre avec des questions de degré. Nous avons besoin de chiffres pour comprendre les longues chronologies littéraires » [8]. Le chiffre permet de manipuler de grands ensembles de données et de chercher du sens face à la quantité énorme de textes écrits. Le livre d'Andrew Piper - « Énumérations [9] » - pose la question des implications de la conception de la littérature comme *quantité*. La multiplicité des possibilités des analyses et la quantité massive de données imposent au chercheur une rigueur scientifique.

La méthode à l'oeuvre ne consiste pas à lancer des calculs exploratoires sur de grands ensembles de données, trouver des motifs de répétitions et d'en tirer les conclusions qui en découlent. Au contraire, elle se fonde sur des hypothèses et des savoirs historiques pour les questionner dans les textes. Les nouvelles méthodes que nous avons décrites fonctionnent en tandem avec la théorie littéraire et l'histoire littéraire. Mettre de côté les immenses savoirs hérités de l'analyse littéraire traditionnelle serait contre-productif et réduirait les capacités d'interprétation des études computationnelles.

6. M. Algee-Hewitt, Erik Fredner et Hannah Walser, « The Novel as Data », dans *The Cambridge Companion to the Novel*, dir. EricEditor Bulson, 2018 (Cambridge Companions to Literature), p. 189-216, DOI : 10.1017/9781316659694.013

7. M. L. Jockers, *Macroanalysis : Digital Methods and Literary History*, 1re éd., 2013, DOI : 10.5406/illinois/9780252037528.001.0001.

8. T. Underwood, « Why Literary Time is Measured in Minutes », *ELH*, 85–2 (2018), p. 341-365, DOI : 10.1353/elh.2018.0013.

9. A. Piper, *Enumerations : data and literary study*, Chicago ; London, 2018.

2.2 La modélisation en études littéraires

Dans la dernière décennie, le champ des humanités numériques a pu profiter des grands progrès des méthodes quantitatives. En effet, s'éloignant de la mesure de variables ou de fréquences d'occurrences dans les textes, les recherches actuelles utilisent désormais des modèles statistiques pour appréhender des concepts littéraires. Les recherches sont passées de méthodes statistiques descriptives à des méthodes prédictives de modélisation.

Dans ce mémoire, nous mettons en oeuvre ce genre de techniques et nous voulons ici préciser les modalités de leur emploi. L'introduction de l'apprentissage machine dans les processus de réflexion des sciences humaines va au delà des mesures en lecture distante. Selon Richard J. So, « La lecture à distance est innovante parce qu'elle a introduit non seulement des "données" ou des "algorithmes" dans les études littéraires, mais aussi, et surtout, la modélisation quantitative comme forme de raisonnement et d'analyse » [10]. La modélisation statistique permet un changement de perspective car elle introduit dans les humanités numériques la notion d'incertitude et de probabilité de la mesure. Elle ne présente pas des résultats tranchés mais plutôt un compte rendu de ce que le modèle a cherché à mesurer et les limites de sa capacité à produire le résultat.

Pour comprendre plus précisément le changement de paradigme auquel nous avons à faire, nous pouvons encore une fois revenir à Franco Moretti et son article « Style, Inc : Réflexions sur 7 000 titres » [11]. Le chercheur y atteste la baisse de la longueur des titres de romans dans le temps. Il fait des mesures simples, avec des moyennes par année de la longueur des titres, et les projette sur un graphe. Ce sont des analyses statistiques descriptives, qui ne permettent pas de conclure définitivement sur la portée de la tendance ou la significativité statistique des résultats. Andrew Piper porte l'idée que la modélisation littéraire introduit « une manière relativiste de penser la traversée des échelles d'analyse critique » [12], c'est à dire que les modèles statistiques permettent une approche plus nuancée sur les données et les résultats quantifiés. Dans le contexte des études littéraires computationnelles, l'utilisation récente de l'apprentissage automatique dans des tâches complexes de classification sur des textes littéraires a été l'objet de nombreuses recherches, telles que l'étude des spécificités intrinsèques d'un texte de fiction en comparaison à d'autres productions textuelles (philosophie, essais, ...). Le travail d'Andrew Piper [13] montre que les indices textuels et leur modélisation permettent de reconnaître et d'attester un texte comme étant fictionnel ou non-fictionnel avec plus de 94% d'efficacité. Si la tâche est en elle-même peu complexe, et que l'on pouvait s'attendre à obtenir de tels

10. Richard Jean So, « "All Models Are Wrong" », *PMLA/Publications of the Modern Language Association of America*, 132–3 (mai 2017), p. 668-673, DOI : 10.1632/pmla.2017.132.3.668.

11. F. Moretti, « Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850) », *Critical Inquiry*, 36–1 (2009), p. 134-158, DOI : 10.1086/606125.

12. A. Piper, « Think Small : On Literary Modeling », *PMLA/Publications of the Modern Language Association of America*, 132–3 (mai 2017), p. 651-658, DOI : 10.1632/pmla.2017.132.3.651.

13. Id., « Fictionality », *Journal of Cultural Analytics* (, 20 déc. 2016), DOI : 10.22148/16.011.

résultats, Andrew Piper explique que :

> « l'intérêt du point de vue quantitatif est qu'il nous permet de mieux comprendre la manière dont un type d'écriture particulier signale aux lecteurs une orientation particulière [...]. Cela n'exclut pas les innombrables façons dont les lecteurs peuvent trouver leur propre version de l'importance du roman. Mais cela nous permet de mieux comprendre le roman en tant que catégorie sociale ».

Ainsi, l'algorithme d'apprentissage machine prend les caractéristiques associées à chaque texte et calcule le degré auquel elles les distinguent ce texte comme appartenant à la catégorie qui leur est attribuée. Pour autant, Ted Underwood précise que « l'ordinateur ne sait rien de l'histoire littéraire : il ne modélise que les éléments que nous lui fournissons. Le modèle ne peut mesurer aucune dimension universelle du langage ; il indique simplement si un texte donné ressemble plus aux exemples du groupe A ou du groupe B » [14].

Des travaux similaires se sont focalisés sur la notion de genre littéraire [15]. Il faut bien comprendre que le concept de genre n'est pas arrêté, et que la frontière est fine entre, par exemple, la *science fiction* et la *fantasy*. L'apprentissage machine ne cherche pas à décrire une vérité qui serait absolue, mais cherche plutôt à détecter les contours de concepts que l'humanité a forgé au cours des écritures et lectures des siècles passés.

On pourrait citer bien d'autres travaux, modélisant différents concepts littéraires avec différentes approches computationnelles mais nous terminons cette section avec un élément, souligné par le chercheur Richard J. So [16], qui nous semble de première importance. Il s'agit de l'interprétabilité des modèles statistiques. En effet, c'est un des enjeux majeurs de la recherche en humanités numériques si l'on veut démocratiser ces approches et réconcilier les méthodes quantitatives avec les études littéraires traditionnelles et plus largement les sciences humaines et sociales. La plupart des modèles statistiques mettent en place des coefficients pour réaliser leurs inférences. Récupérer ces coefficients peut approfondir notre compréhension des prises de décisions de ces algorithmes [17] et ainsi amener de nouvelles formes de raisonnement pour traiter de questions propres aux sciences humaines et sociales. Nous proposons dans ce mémoire une synthèse entre une approche

14. T. Underwood, « Machine Learning and Human Perspective », *PMLA/Publications of the Modern Language Association of America*, 135–1 (janv. 2020), p. 92-109, DOI : 10.1632/pmla.2020.135.1.92.

15. T. Underwood, Michael L. Black, Loretta Auvil et Boris Capitanu, « Mapping mutable genres in structurally complex volumes », dans *2013 IEEE International Conference on Big Data*, Silicon Valley, CA, USA, 2013, p. 95-103, DOI : 10.1109/BigData.2013.6691676.

16. Hoyt Long et R. J. So, « Literary Pattern Recognition : Modernism between Close Reading and Machine Learning », *Critical Inquiry*, 42–2 (janv. 2016), p. 235-267, DOI : 10.1086/684353.

17. Marco Tulio Ribeiro, Sameer Singh et Carlos Guestrin, « "Why Should I Trust You ?" : Explaining the Predictions of Any Classifier », dans *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA, 2016, p. 1135-1144, DOI : 10.1145/2939672.2939778.

historique de l'analyse des textes littéraires avec une approche computationnelle. Cette dernière, décrite notamment par le Text Lab de Chicago [18], permet une approche plus empirique par l'utilisation des grands nombres pour interagir avec les textes littéraires, mais un retour à une granularité fine est essentiel pour comprendre les inférences réalisées par nos modèles statistiques.

2.3 État de l'art : Le canon littéraire au révélateur des méthodes quantitatives

Le canon littéraire est une notion très discutée dans les études littéraires computationnelles. On peut, une fois encore, citer les travaux du laboratoire littéraire de l'université de Stanford, dont certains pamphlets abordent cette notion.

Une première approche a été de décrire quantitativement les listes qui constituaient les différents canons. Le pamphlet 8, « Between Canon and Corpus : Six Perspectives on 20th- Century Novels » [19] caractérise le canon littéraire et montre le caractère exclusif de tels ensembles. Une approche similaire a été reproduite récemment, avec un travail sur les syllabus des études hispaniques dans les universités étasuniennes [20]. Les chercheurs ont étudié la diversité du canon avec des mesures d'entropie des populations canoniques dans le temps. D'autres études essaient de caractériser la notion de canon littéraire par la composition de ces ensembles, notamment dans leur construction naissante [21].

D'autres études sont allées au delà du canon littéraire et ont repris la construction binaire du champ littéraire de Pierre Bourdieu [22], entre popularité et prestige. Marc Verboord [23] a classifié les auteurs selon leur position dans le champ littéraire, en utilisant les spécificités propres au prestige et à la popularité. Une étude du laboratoire littéraire de Stanford, le pamphlet 17 [24], a montré que ces axes semblaient pertinents pour cartogra-

18. H. Long et R. J. So, « Literary Pattern Recognition… ».

19. M. Algee-Hewitt et M. McGurl, « Between Canon and Corpus : Six Perspectives on 20th- Century Novels », Pamphlets of the Stanford Literary Lab–8 (2015), URL : https://litlab.stanford.edu/LiteraryLabPamphlet8.pdf. Published inM. Algee-Hewitt, *Canon/Archive : studies in quantitative formalism from the Stanford Literary Lab*, dir. Franco Moretti, New York, 2017

20. José Eduardo González, Elliott Jacobson, Laura García García et Leonardo Brandolini Kujman, « Measuring Canonicity : Graduate Reading Lists in Departments of Hispanic Studies », *Journal of Cultural Analytics* (, 19 mars 2021), DOI : 10.22148/001c.21599.

21. Mikko Tolonen, Mark J. Hill, Ali Zeeshan Ijaz, Ville Vaara et Leo Lahti, « Examining the Early Modern Canon : The English Short Title Catalogue and Large-Scale Patterns of Cultural Production », dans *Data Visualization in Enlightenment Literature and Culture*, dir. Ileana Baird, Cham, 2021, p. 63-119, DOI : 10.1007/978-3-030-54913-8_3.

22. P. Bourdieu, *Les règles de l'art…*, p 176

23. Marc Verboord, « Classification of authors by literary prestige », *Poetics*, 31–3 (juin 2003), p. 259-281, DOI : 10.1016/S0304-422X(03)00037-8.

24. J.D. Porter « Popularity/Prestige », Pamphlets of the Stanford Literary Lab 17 (2018), url : https://litlab.stanford.edu/LiteraryLabPamphlet17.pdf. Published in M. Algee-Hewitt, *Canon/Archive…*

phier l'espace littéraire et culturel.

La seconde approche très populaire pour appréhender le canon littéraire est de mesurer dans les textes mêmes, à l'aide des méthodes du traitement automatique des langues, des différences entre ouvrages canoniques et non-canoniques.

Le pamphlet 11 [25], *Canon/Archive. Large-scale Dynamics in the Literay Field*, du laboratoire littéraire de Stanford est très instructif à cet égard. L'article montre que la variété lexicale et la quantité d'information sont plus importants dans les oeuvres retenues par le canon littéraire que dans les autres ouvrages.

Dans une synthèse sur l'utilisation de la théorie de l'information pour mesurer des éléments propres aux sciences humaines [26], Dallas Liddle explique que la redondance perçue par le lecteur humain n'est pas la même que celle calculée statistiquement, et qu'il ne faudrait pas porter de conclusions hâtives. Pourtant, il montre qu'il y a bien une pression informative dans la sélection de textes littéraires.

Ted Underwood consacre un article [27] à la classification automatique du prestige littéraire fondée sur des données textuelles. Il travaille sur de la poésie et définit le prestige littéraire comme étant la probabilité d'un texte à être examiné dans des revues littéraires spécialisées. La question principale qu'il se posait était la suivante : « Est-ce que la frontière sociale entre le goût d'une élite et le reste de la production littéraire est associée à des différences stylistiques reconnaissables ? » Avec des outils simples du TAL (des sacs de mots) et un algorithme prédictif, (une régression logistique), Ted Underwood obtient de bons résultats, de l'ordre de 75% d'efficacité pour son modèle statistique. Il montre que le discours littéraire contenu dans le texte est en lien avec la réception du-dit texte, et que ce lien est statistiquement solide.

Dans le sillage de ces découvertes, de nombreuses recherches ont investi la question du prestige littéraire. Leur prisme d'approche se focalise sur le style des ouvrages consacrés, et sa démarcation potentielle par rapport aux autres. Ce sujet est très abordé aux Pays-Bas, on peut citer notamment le travail de Karina van Koolen, qui montre que le degré de littérarité perçue par les humains est quantifiable et modélisable [28]. Deux articles d'Andreas van Cranenburgh et al [29] [30], explorent cette littérarité perçue à l'aide de vec-

25. Mark Algee-Hewitt et al. « Canon/Archive. Large-scale Dynamics in the Literay Field », Pamphlets of the Stanford Literary Lab 11 (2016), url : https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf. Published in *Ibid.*

26. Dallas Liddle, « Could Fiction Have an Information History ? Statistical Probability and the Rise of the Novel », *Journal of Cultural Analytics* (, 2019), DOI : 10.22148/16.033.

27. T. Underwood et Jordan Sellers, « The "Longue Durée" of Literary Prestige », *Modern Language Quarterly*, 77–3 (sept. 2016), p. 321-344, DOI : 10.1215/00267929-3570634.

28. Corina Koolen, K. van Dalen-Oskam, Andreas van Cranenburgh et Erica Nagelhout, « Literary quality in the eye of the Dutch reader : The National Reader Survey », *Poetics*, 79 (avr. 2020), p. 101439, DOI : 10.1016/j.poetic.2020.101439.

29. A. van Cranenburgh, K. van Dalen-Oskam et J. van Zundert, « Vector space explorations of literary language », *Language Resources and Evaluation*, 53–4 (déc. 2019), p. 625-650, DOI : 10.1007/s10579-018-09442-4.

30. A. van Cranenburgh et Rens Bod, « A Data-Oriented Model of Literary Language », dans *Procee-*

teurs de mots et parviennent à des résultats intéressants. Le prestige littéraire est ainsi associé à un style et à une esthétique textuelle particuliers. L'article analyse les mots, les thèmes et les vecteurs de documents qui sont associés à cette dernière. Il est néanmoins difficile de caractériser cette esthétique, et ce mémoire essaiera de poursuivre ce travail de clarification, avec une double approche entre *distant* et *close reading*.

Un des articles fondamentaux pour notre état de l'art est le tout récent travail de Judith Brottrager et al [31], qui analyse et compare la relation entre le concept de canon fondé sur les contextes des oeuvres et les éléments textuels intrinsèques à ces dernières. Les résultats témoignent d'une absence de corrélation évidente entre les deux méthodes, avec cependant des éléments intéressants à plus petite échelle.

Ces recherches empiriques sur le prestige littéraire sont peu présentes en France, et peu d'expérimentations ont eu lieu sur des corpus d'ouvrages francophones. Il faudrait citer une étude sur la sélection successive d'ouvrages au prix Goncourt 2020 [32], qui n'obtient pas de résultats convaincants sur un corpus limité.

Ainsi, ce mémoire s'inscrit dans un contexte de recherches dynamiques dans le monde anglo-saxon. Peu de travaux ont été réalisés sur des données francophones, nous ouvrons donc un champ de recherche qui peut être foisonnant. Un premier temps sera donc consacré à recueillir des méta-données pertinentes pour construire un canon littéraire français. Dans un deuxième temps, nous modéliserons les spécificités textuelles du canon littéraire à l'aide des méthodes de l'apprentissage machine et du traitement automatique des langues. Nous essaierons différentes approches pour améliorer le niveau de l'état de l'art qui se trouve aux alentours de 75% avec les travaux de Ted Underwood déjà cités [33].

dings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers, Valencia, Spain, 2017, p. 1228-1238, DOI : 10.18653/v1/E17-1115.

31. Judith Brottrager, Annina Stahl et Arda Arslan, « Predicting Canonization : Comparing Canonization Scores Based on Text-Extrinsic and -Intrinsic Features », dans *Proceedings of the Conference on Computational Humanities Research CHR2021*, dir. Maud Ehrmann, Folgert Karsdorp, Melvin Wevers, Tara Lee Andrews, Manuel Burghardt, Mike Kestemont, Enrique Manjavacas, Michael Piotrowski et Joris van Zundert, Amsterdam, the Netherlands, 2021 (CEUR Workshop Proceedings), t. 2989, p. 195-205, URL : http://ceur-ws.org/Vol-2989/#short_paper21.

32. Michel Bernard, « Goncourt 2020 : mais qu'a-t-il de plus que les autres ? », *Humanités numériques*–4 (1er déc. 2021), Number : 4 Publisher : Humanistica, DOI : 10.4000/revuehn.2297.

33. T. Underwood et J. Sellers, « The "Longue Durée" of Literary Prestige »...

Deuxième partie

Matériel et méthode

Chapitre 3

Les défis du corpus

3.1 Les implications du corpus en méthodes quantitatives

La constitution d'un corpus est fondamentale dans les études littéraires computationnelles. C'est à la fois sa grande force, car son caractère massif justifie l'utilisation de méthode quantitatives et sa plus grande faiblesse, puisque c'est le moment critique où les biais rentrent en compte et peuvent fausser les calculs et les résultats de l'étude.

Dans un premier temps, il est important de préciser que le corpus ne vise pas à l'exhaustivité des textes écrits et publiés. Cela est simplement impossible, et l'idée est de prendre la fenêtre la plus large possible, pour représenter la majorité des courants et sous-genres littéraires. Nous visons un échantillon représentatif de ce qu'a pu être la production littéraire à une certaine époque. Bien que cette représentativité soit un idéal hors de portée, il faut tendre vers cet objectif.

Comme nous l'avons dit, un des risques majeurs réside dans la constitution de biais qui fausseront l'ambition initiale de production de savoirs empiriques en histoire de la littérature. En effet, les biais sont des distorsions, des déformations systématiques d'un échantillon statistique, en l'occurrence un corpus de textes. Cela peut se caractériser par exemple par des biais temporels, qui focaliseront l'étude sur un grand nombre de textes d'une période donnée en minimisant l'importance d'autres périodes littéraires.

De plus, si l'on veut étudier les pratiques littéraires génériques d'une période donnée, on ne peut pas se restreindre à un choix d'un petit nombre d'auteurs canoniques. Ce mémoire veut mettre au jour la présence de normes dans les pratiques stylistiques du canon littéraire. Si notre corpus contient une présence trop importante d'un sous-genre littéraire, par exemple le roman policier - réputé non-canonique - alors il est probable que l'on détecte non pas une norme générique, mais un ensemble de traits stylistiques propres à ce sous-genre précis.

Il nous faut donc un corpus équilibré, autant dans le temps, que dans les sous-genres

littéraires représentés. Enfin, il nous faut décrire rigoureusement le corpus sur lequel les analyses sont réalisées pour vérifier l'absence de biais, ou pour le moins en avoir conscience lors de l'analyse et de l'interprétation des résultats.

3.2 Le corpus d'étude

Une des chances de cette présente étude est qu'elle arrive après une première salve de fouille de texte à grande échelle, dont une des missions principales était justement la constitution de grands ensembles de textes.

Le projet « ANR Chapitres » [1] a recueilli un corpus massif de près de 3000 textes littéraires. Ce corpus est structuré en XML (eXtended Markup Language) avec un encodage TEI (Text Encoding Initiative), ce qui permet d'ajouter des méta-données aux textes, tels que le titre, l'auteur, l'éditeur, la date de parution ou encore le sous-genre romanesque. Nous mettons en annexe 7.5.3 la présentation de cette ANR. La période concernée s'étend sur deux siècles de production romanesque, du XIX^{e}au XX^{e}siècle, comme on peut le voir en figure 3.1.

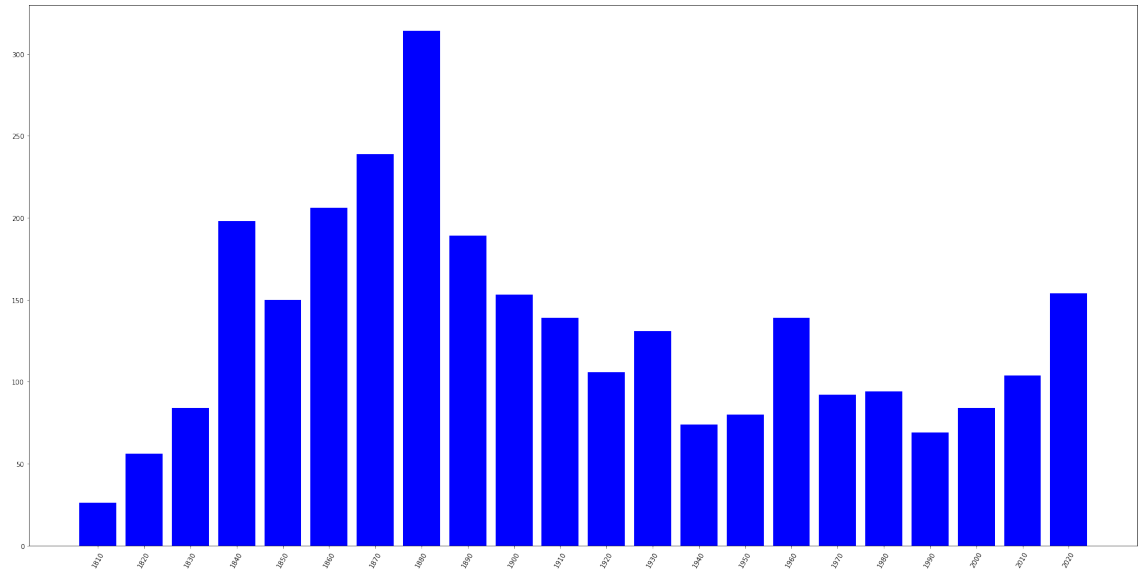

FIGURE 3.1 – Répartition du corpus dans le temps

La répartition dans le temps des romans de notre corpus est plutôt bonne, avec cependant la deuxième parie du XIX^{e}siècle qui comporte près de 40% des romans. La décennie 1880 représente à elle seule près de 10% des romans. Il faudra tenir compte de ces biais potentiels dans nos analyses, le risque étant de sur-représenter cette période dans les mesures statistiques.

Le tableau 3.1 décrit le corpus avec des statistiques simples.

1. https ://chapitres.hypotheses.org/

TABLE 3.1 – Statistiques du corpus

Romans	2953
Phrases	14.982.817
Mots	234.175.471
Moyenne de mots par roman	79301

L'un des biais principaux de ce corpus est qu'il rassemble les œuvres romanesques numérisées et disponibles sur internet. Cela implique nécessairement que ces textes ont été sélectionnés, publiés et conservés dans le temps, ce qui représente une partie infime de la production d'écrits.

Pourtant, nous pensons que les ouvrages non-canoniques du corpus représentent un bon échantillon de ce qu'à pu être la production littéraire populaire, de par leur nombre - plus des 9/10 (cela dépend du critère que nous choisissons, nous verrons cela au prochain chapitre), mais aussi par la diversité des sous-genres représentés. La figure 3.2 montre les différents sous-genres du corpus, qui vont des romans policiers aux romans d'aventures, en passant par la science fiction. Il est important de noter que 2/3 des romans ont une étiquette de genre littéraire, ce qui représente selon nous un échantillon représentatif. Nous commenterons la répartition du canon dans le corpus en section 4.2.

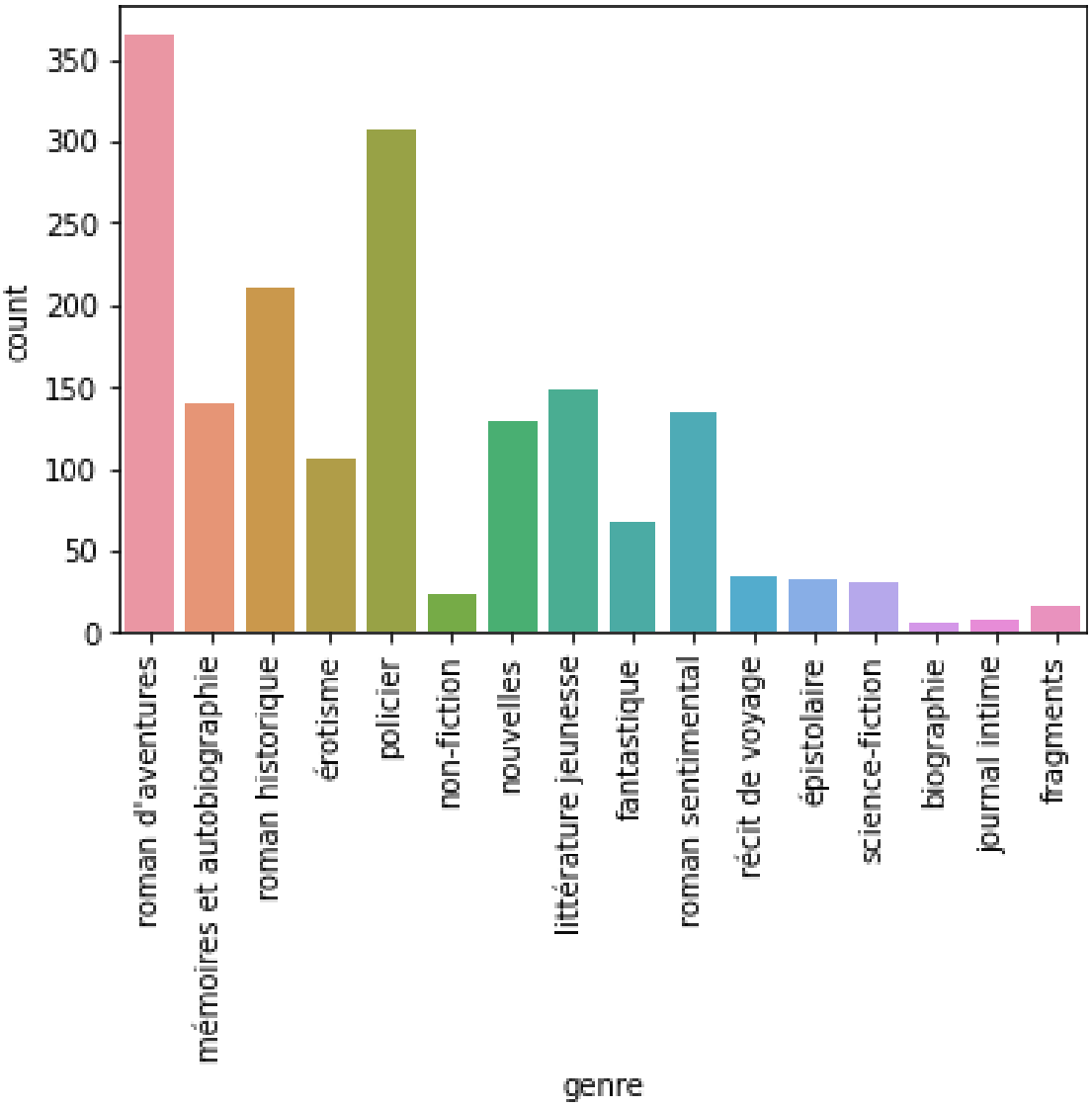

FIGURE 3.2 – Sous-genres littéraires du corpus

Chapitre 4

Caractériser un canon littéraire

4.1 Méta-données : Enrichir le corpus

Une des tâches principales du mémoire a été la construction d'un canon littéraire. Pour cela, nous entreprenons d'enrichir notre corpus avec les éléments de la réception actuelle des textes et des auteurs. Le canon littéraire n'est ni monolithique ni temporellement stable, le définir par des critères finis est en soi réducteur et néglige la complexité du phénomène. Mais nous allons essayer de nous focaliser sur des éléments déjà discutés et analysés par la critique littéraire et les études réalisées sur ce sujet. Un des aspects sur lequel nous allons nous concentrer est le rôle des institutions dans la formation du prestige littéraire. En premier lieu, tout le travail réalisé par Martine Jey et Laetitia Perret [1] sur le rôle de l'institution scolaire dans la constitution de tels ensembles est prépondérant. Toutes deux ont montré que l'enseignement secondaire et supérieur avaient un impact énorme sur les processus de canonisation (dans la fabrication et surtout la conservation de ce canon) des auteurs et des textes.

Travailler sur le canon littéraire à partir de la réception réalisée par les institutions scolaires nous paraît pertinent parce que cela s'inscrit dans une approche déjà éprouvée, bien qu'elle soit non exhaustive. En effet, d'autres facteurs entrent en jeu dans la constitution du canon, que ce soient des critères politiques, économiques ou sociologiques.

Il est important de noter que nous allons analyser la réception du canon littéraire par la réception actuelle, ce qui constitue aussi un élément discutable, mais la disponibilité des sources des méta-données (grâce à leur format numérique), nous a permis de recueillir énormément de données dans un temps réduit.

Ainsi, nous avons établi un ensemble de critères non exhaustifs pour caractériser le canon littéraire que nous allons investiguer quantitativement.

1. *L'idée de littérature dans l'enseignement*, dir. M. Jey et Laetitia Perret, Paris, 2019 (Rencontres, Série Littérature des XXe et XXIe siècles, 380. 36).

4.1.1 Le canon scolaire

En premier lieu, nous nous sommes intéressés aux listes et au programmes des examens du secondaire, c'est à dire le brevet et le baccalauréat. Si l'on considère l'école républicaine comme le lieu de diffusion et de conservation du canon littéraire, alors il nous semble important de prendre en compte ce que l'on attend d'un élève lorsqu'il sort de l'enseignement obligatoire, c'est à dire ce qui constitue, pour les auteurs de ces listes, la culture littéraire minimale à la construction citoyenne. Le travail de Martine Jey, « La littérature au lycée : invention d'une discipline (1880-1925) » [2] décrit avec précision la construction d'une discipline, la littérature autour de textes garants d'une certaine langue et d'une certaine morale, qu'il faut diffuser pour éduquer les masses. Elle y analyse le processus d'intégration des œuvres dans le corpus scolaire qui est de fait un processus de canonisation. Au sein de l'enseignement scolaire et supérieur, les études littéraires produisent des formes distinctes de connaissances linguistiques [3]. Il faudra donc récupérer des méta-données sur ces institutions qui construisent le canon littéraire.

4.1.2 Le canon de l'enseignement du supérieur

Dans un deuxième temps, nous nous sommes intéressés aux concours de l'enseignement du supérieur, qui incarnent le même rôle que précédemment mais avec un degré de prestige plus important. Nous avons récupéré les listes et les programmes de concours du supérieur (classes préparatoires littéraires et scientifiques du concours de l'ENS). Il s'agit d'évaluer et de sélectionner sur des connaissances littéraires, des candidats qui deviendront de futurs professeurs de collège. Michel P. Schmitt [4] a réalisé une démarche similaire en tenant des listes du nombre d'occurrences de citation d'auteurs dans les dissertations au CAPES interne. Sans passer par l'intérieur des copies, nous nous intéressons plutôt aux programmes des concours cités.

4.1.3 Le canon des concours de l'agrégation

Nous avons également récupéré les programmes des concours des agrégations de lettres modernes, plus haut concours de recrutement dans la fonction publique pour des professeurs. Il nous semblait significatif de constater quels étaient les auteurs et les textes sur lesquels on formait l'élite des professeurs de lettres de la République. Pour un panorama des épreuves de l'agrégation, les recherches de Martine Jey [5], d'André Chervel [6] et

2. M. Jey, *La littérature au lycée…*

3. John Guillory, « Canon » in F. Lentricchia et T. McLaughlin, *Critical terms for literary study…*, p.43

4. Michel P. Schmitt et A. Viala, « Les cotes aux concours », *Littératures classiques*, 19–1 (1993), p. 281-291, DOI : 10.3406/licla.1993.1753.

5. M. Jey, « Le canon aux agrégations du XIX e siècle »…

6. André Chervel, *Histoire de l'agrégation : contribution à l'histoire de la culture scolaire*, Paris, 1993 (Collection "Le Sens de l'histoire").

de Yves Chevrel [7] ont été d'une grande aide. Comme ces programmes ne comportaient pas beaucoup de romans, nous avons décidé d'agrandir la période de réception considérée. Pour cette méta-donnée précisément, nous remontons jusqu'en 1950.

4.1.4 Le canon des éditeurs

Ensuite, nous nous sommes intéressés au monde de l'édition, qui est aussi un des acteurs majeurs dans le processus de canonisation des oeuvres. La thèse de Dragos Jipa [8] a bien montré, avec la collection des « Grands écrivains de France », l'importance des logiques éditoriales dans la construction d'un consensus national autour d'un panthéon d'auteurs.

Nous nous sommes intéressés à la collection de la Pléiade qui est l'exemple typique du rôle des éditeurs dans la canonisation d'auteurs. La publication des écrits d'un auteur en oeuvres complètes vient souvent parachever une carrière littéraire, souvent après la mort de l'auteur. Elle est un signe majeur dans la reconnaissance du-dit auteur en tant qu'appartenant au canon littéraire. C'est aussi une des portes privilégiées pour les éditeurs car elle permet de sacraliser de nouvelles têtes et d'en tirer profit par effet de levier. La collection de la Pléiade comporte finalement peu d'auteurs de romans, et consacre un auteur dans sa totalité. Malgré tout, nous récupérons tous les auteurs publiés dans la Pléiade.

Pour plus de finesse, nous prenons également l'édition Garnier-Flammarion et plus précisément la collection « littérature et civilisation ». Un des traits majeurs de cette collection est qu'elle présente les romans avec un appareil critique qui accompagne la lecture de l'ouvrage. Cet appareil critique n'est pas anodin et montre que l'ouvrage a une portée qui requiert explication. Le canon littéraire est rempli d'oeuvres qui valent la peine que l'on se penche sur elles et leur contextes. Cela nous intéresse parce que cette vision pédagogique du canon littéraire est très répandue et est même un des arguments au statu-quo littéraire en France et ailleurs [9].

4.1.5 Le canon de la critique

Pour prendre un témoin de la critique littéraire, nous nous sommes intéressés aux prix littéraires. Ces derniers constituent la partie de notre canon la moins résistante au temps, car ils sont embourbés dans des dimensions économiques et politiques très fortes comme l'a résumé James F. English dans son livre sur la circulation de la valeur cultu-

7. Yves Chevrel, « Les Lettres modernes et la formation des professeurs de français : » *L'information littéraire*, Vol. 55–3 (1er sept. 2003), p. 3-10, DOI : 10.3917/inli.553.0003.

8. Dragoş Jipa, *La canonisation littéraire et l'avènement de la culture de masse : la collection Les grands écrivains français (1887-1913)*, ISBN : 9783631672419 Series Number : Volume 302 Series : Publications universitaires européennes, thèse de doct., Frankfurt, Peter Lang Academic research, 2016.

9. J. Guillory, *Cultural capital*...

relle [10]. Malgré ces remarques, nous voulions mesurer son effet dans les pratiques littéraires. Ainsi nous avons récupéré les listes des prix littéraires les plus importants, du prix Goncourt au prix Femina.

Enfin nous prenons en compte la recherche contemporaine, avec une canonicité à l'échelle de l'auteur et la revue littéraire en ligne « Fabula [11] ». Cet aspect était déjà présent dans les méta-données du corpus Chapitres (voir en annexe 7.5.3), et nous avons décidé de le conserver.

4.2 Notre canon

Ainsi construit avec plusieurs facteurs, notre canon littéraire se trouve moins discutable puisqu'il cherche à multiplier les entrées dans les différents acteurs du champ littéraire qui définissent, nourrissent et conservent le canon littéraire.

Une grande partie de ces données, notamment celles du brevet et du baccalauréat ont été récupérées grâce à l'immense travail du collectif *Le deuxième texte* [12] qui a mis en ligne ses données [13] en open source. D'autres ont été récupérées automatiquement à l'aide de scripts python sur les pages web de la collection Garnier-Flammarion et de la Pléiade, mais aussi à la main pour les auteurs présents dans le Lagarde et Michard (Bordas) du XIX^{e}siècle et du XX^{e}siècle.

Nous avons décidé d'établir une double approche quand à l'appartenance ou non au canon littéraire. La canonicité revient d'une part à la granularité du roman, et d'autre part à celle de l'auteur.

Nous construisons ainsi deux canons littéraires, à l'échelle des auteurs et des textes. Nous mettons en annexes les listes détaillées de ces deux ensembles, avec en annexe 7.5.3 la liste des auteurs canoniques, et en annexe 7.5.3 la liste des romans canoniques, triée par année de première parution. L'échelle de l'auteur agrandit énormément le cercle, puisque toutes les oeuvres d'un auteur canonique sont considérées comme canoniques. Si l'approche la plus intéressante nous paraissait la première à l'échelle de l'œuvre, nous avons conservé les deux qui méritaient tout autant des tests approfondis.

Pour faire correspondre notre corpus au canon construit par nos soins, il a été décidé de réaliser un simple test d'appartenance, si le titre d'un roman appartient à au moins une liste, le roman en question est considéré comme canonique. Même chose pour l'auteur, s'il est présent dans une des listes établies, tous les textes de l'auteur sont considérés comme canoniques. Ainsi, le nombre des ouvrages du corpus qui sont dans notre canon s'élève à 264 éléments (moins de 1% du corpus), tandis que le nombre des œuvres dont les auteurs

10. James F English, *Economy of Prestige : Prizes, Awards, and the Circulation of Cultural Value*. OCLC : 1058248419, Cambridge, 2009.
11. https ://www.fabula.org/
12. https ://george2etexte.wordpress.com/
13. https ://www.data.gouv.fr/fr/organizations/le-deuxieme-texte/

sont dans notre canon est de 1156 romans, soit 39%.

Nous avons pensé mettre en place un indice de canonicité, pour mieux représenter la réalité du monde littéraire, qui n'est pas d'une binarité sans conditions, entre canon et non canon. Un auteur est plus ou moins canonique, et Honoré de Balzac l'est plus qu'Anatole France par exemple. Mais nous manquions de temps et de données pour présenter une telle approche, et après des essais peu fructueux, nous conservons notre test d'appartenance binaire.

Pour bien vérifier la viabilité de ces approches, nous avons réalisé des statistiques descriptives pour voir comment notre canon se comportait dans le corpus.

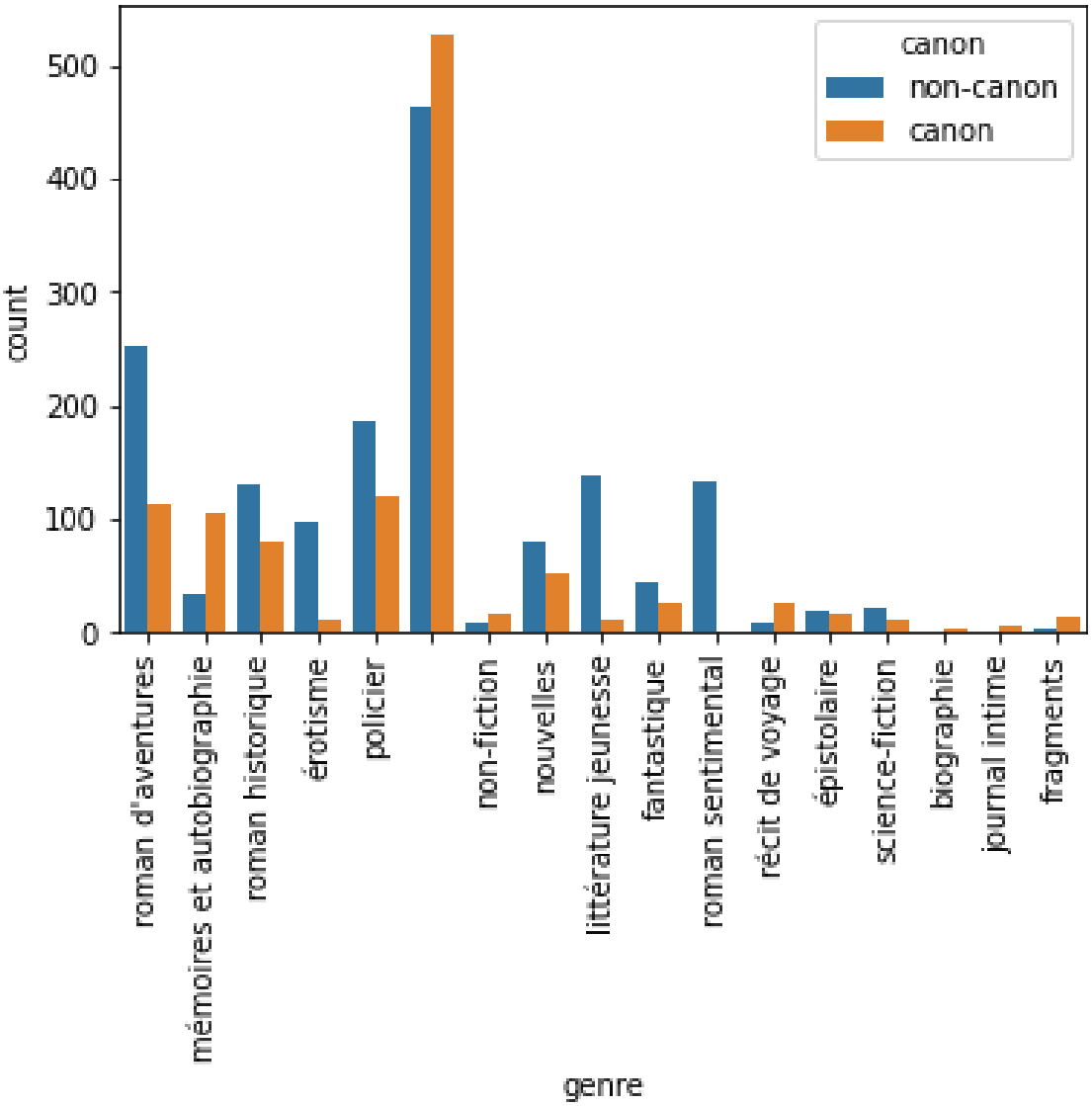

FIGURE 4.1 – Répartition du canon dans les sous-genres du corpus

On constate dans la figure 4.1 une bonne répartition des oeuvres du canon des auteurs dans les sous-genres littéraires du corpus. Il faut souligner que les ouvrages canoniques sont de fait moins représentés que ceux canoniques, puisque le corpus en comporte

seulement 39%. La seule anomalie à constater est au sein des ouvrages non-étiquetés par un genre, où le canon est plus représenté que les non-canons. Mais la différence d'une cinquantaine d'ouvrages n'est pas énorme, et l'on ne peut pas conclure dans ce cas que le canon des romans n'est qu'un agrégat de sous-genres littéraires spécifiques.

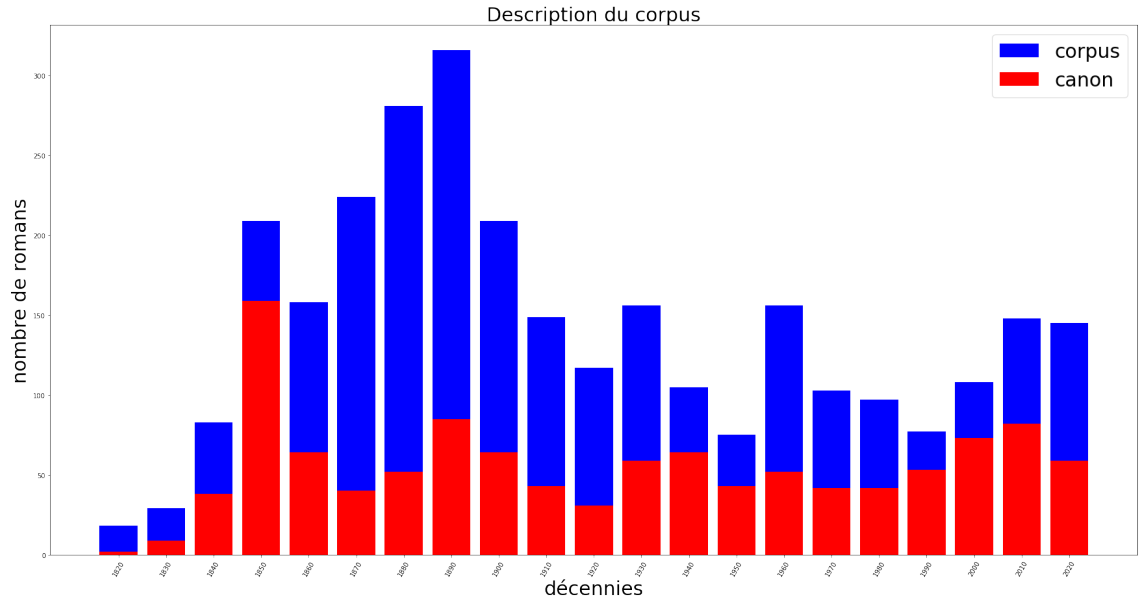

FIGURE 4.2 – Répartition du canon dans le temps

La figure 4.2 représente la présence de notre canon au fil de l'histoire littéraire. Il est assez bien réparti et suis l'évolution de la population du corpus. On remarque que la décennie de 1840 à 1850 comporte plus de canon que de non-canon, mais c'est la seule dans ce cas.

Ainsi, le genre littéraire ou la date de parution ne sont pas des facteurs définitifs pour expliquer la présence de romans ou d'auteurs dans le canon. Notre approche, qui est d'ouvrir les romans et de s'intéresser au contenu textuel de manière quantitative est justifié, dans le sens où le texte est un facteur qui pourrait contribuer à apporter des éléments de réponses.

Chapitre 5

Méthodes computationnelles

Les méthodes du traitement automatique des langues sont d'une grande assistance. En plus de nous aider au pré-traitement et au nettoyage des textes, elles nous permettent de récupérer les caractéristiques textuelles sur lesquelles notre classification automatique se fonde. Nous présentons, dans les sections qui suivent, la méthode que nous mettons en place.

5.1 Les données textuelles

Nous voulons mettre au jour une esthétique canonique dans nos textes. On s'intéresse donc à des éléments formels, qui seraient des indices linguistiques issus de la filtration des textes jusqu'à la réception actuelle. Le phénomène que nous voulons modéliser est complexe, et pas évident puisque c'est un phénomène à réception, qui viendrait consacrer un texte ou un auteur après l'écriture de ses textes. Face à la complexité du phénomène que nous envisageons, nous avons voulu simplifier les caractéristiques textuelles retenues. En effet, le meilleur moyen de prendre en compte le contenu textuel de nos romans est de récupérer des informations sur les mots de ces derniers. Cette méthode a déjà été utilisée par des chercheurs déjà cités, notamment Ted Underwood dans son article « La longue durée du prestige littéraire » [1]. Nous aurions pu prendre des éléments plus complexes, relatifs au style ou à la littérarité des ouvrages comme la longueur des phrases, l'avancée narrative ou les thèmes du récit. Mais la relation entre complexité du style, littérarité et prestige littéraire n'est pas si évidente que cela, et nous décidons d'une première approche « simple », que l'on pourrait complexifier par la suite, si ces premières recherches se trouvaient fructueuses.

Ainsi nous décidons de transformer nos ouvrages en *sac-de-mots*. Les textes sont ainsi décomposés en des listes de mots qui indexent leur fréquence relative (le nombre de fois où l'unité apparaît, divisé par la longueur totale du texte). Chaque unité est

1. T. Underwood et J. Sellers, « The "Longue Durée" of Literary Prestige »...

traitée comme une caractéristique des textes dans lesquels elle apparaît - une sorte de trait d'identification - et le texte devient un vecteur de ces traits.

Nous décidons de limiter notre sac de mots au 1000 uni-grammes les plus fréquents récupérés dans un échantillon de 200 textes tirés au hasard dans le corpus. Deux raisons à cela, une première pratique, puisque la prise en compte de tous les mots du corpus aurait nécessairement amené à des coûts computationnels très importants, car les matrices résultantes auraient été très éparses, avec beaucoup de fréquences d'apparitions nulles ou proches de zéro. La seconde raison réside dans la nature des mots que nous récupérons. Comme ce sont les mots les plus fréquents du corpus, la plupart sont des déterminants, prépositions et autres *mots-outils*. Ces derniers relèvent plus d'une écriture inconsciente et automatique des auteurs qu'à des mots moins fréquents relatifs au contenu et aux thèmes du textes. Il ne nous semble pas que les mots de nature thématique jouent un rôle dans la spécificité des textes à être canoniques ou non. Cela nous permet également de ne pas prendre en compte une grande majorité des noms communs ou des noms propres, qui ne nous intéressent pas à cette échelle.

Ces *mots-outils* sont au coeur de la stylométrie, notamment dans les attributions d'auteur [2], et dans l'étude des idiolectes [3], c'est à dire de la signature textuelle d'un écrivain. Ces méthodes ont produit de très bons résultats, allant de Hildegarde de Bingen [4], à Shakespeare [5] ou Molière [6] et Racine [7]. Si le problème auquel nous faisons face n'est pas de même nature, nous pensons que ces techniques sont pertinentes pour traiter notre problème. En effet, s'il y a bien une manière particulière de faire littérature selon les institutions qui forment le canon littéraire, alors nous devrions retrouver les marqueurs inconscients de cette sélection.

Pour quantifier les fréquences d'apparition de mots, on réduit les unités lexicales sujettes à flexion (les verbes, les substantifs, les adjectifs) à leur unité lexicale commune. On appelle ce processus *lemmatisation*. Pour ce traitement, nous utilisons la librairie Spacy. Elle nous permet de tokeniser, lemmatiser et de nettoyer les romans en contrôlant l'étiquetage morphosyntaxique des tokens. La figure 5.1 représente la chaîne de traitement de la récupération des données textuelles.

2. J. Burrows, « 'Delta' : a Measure of Stylistic Difference and a Guide to Likely Authorship », *Literary and Linguistic Computing*, 17–3 (1er sept. 2002), p. 267-287, DOI : 10.1093/llc/17.3.267.

3. Olga Seminck et Thierry Poibeau, *The Evolution of the Idiolect over the Lifetime : A Quantitative and Qualitative Study on French 19th Century Literature*, 2022.

4. M. Kestemont, « Function Words in Authorship Attribution. From Black Magic to Theory ? », dans *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, Gothenburg, Sweden, 2014, p. 59-66, DOI : 10.3115/v1/W14-0908.

5. Petr Plecháč, « Relative contributions of Shakespeare and Fletcher in *Henry VIII* : An analysis based on most frequent words and most frequent rhythmic patterns », *Digital Scholarship in the Humanities*, 36–2 (29 sept. 2021), p. 430-438, DOI : 10.1093/llc/fqaa032.

6. F. Cafiero et J.B. Camps, « Why Molière most likely did write his plays », *Science Advances*, 5–11 (nov. 2019), eaax5489, DOI : 10.1126/sciadv.aax5489.

7. Id., « 'Psyché' as a Rosetta Stone ?... ».

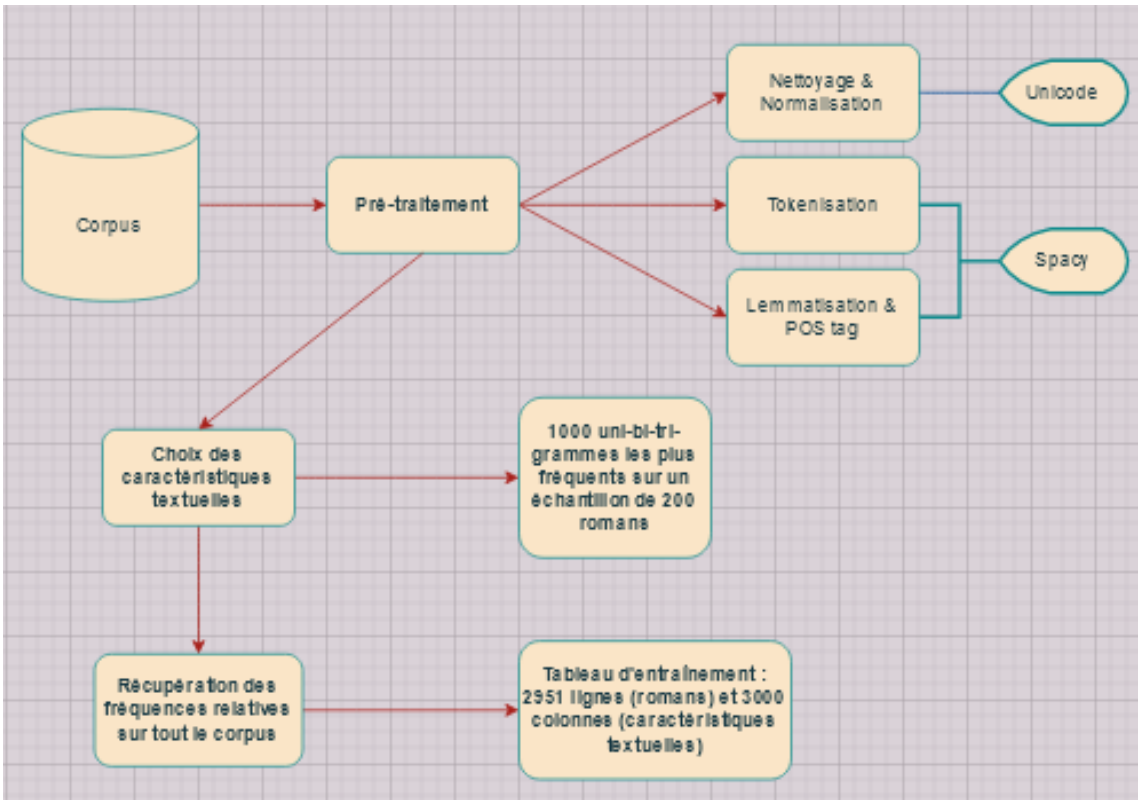

FIGURE 5.1 – Récupération des données textuelles : Flux de travail

Des recherches récentes en classification de textes littéraires ont montré que la prise en compte des contextes d'un mot pouvait avoir des impacts important sur les résultats d'une classification [8]. Ainsi, nous décidons de prendre en compte, en plus des uni-grammes, des bi-grammes et des tri-grammes de lemmes. Ces derniers sont agrégés par la concaténation de deux ou trois lemmes. Nous procédons de la même manière que pour les uni-grammes, en prenant les 1000 éléments les plus fréquents pour chaque type.

Finalement, nous obtenons un grand tableau de données, avec en colonnes les trois mille caractéristiques textuelles que l'on vient de définir, et en lignes chaque roman de notre corpus. Le tableau est rempli avec les fréquences relatives pour chaque caractéristique.

5.2 Outils de programmation

Nous fondons notre travail sur le langage de programmation Python et les librairies construites au-dessus. Pour l'analyse des données et leur manipulation nous utilisons Pandas [9] et Numpy [10]. Pour le traitement du texte à proprement dit, nous employons la

8. A. van Cranenburgh et C. Koolen, « Identifying Literary Texts with Bigrams », dans *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, Denver, Colorado, USA, 2015, p. 58-67, DOI : 10.3115/v1/W15-0707.
9. https ://pandas.pydata.org/
10. https ://numpy.org/

librairie Spacy [11]. Cette dernière est très performante pour une analyse sur de grandes quantités de données et couvre tous les traitements d'une chaîne de TAL classique. Cette librairie est un peu moins performante pour le français des textes littéraires, parce qu'elle a été entraînée sur des contenus de presse récente. D'autres librairies proposent des modèles un peu plus performants, comme la librairie Stanza [12] ou Pie-Extended [13]. Pour autant, après des tests de ces différentes librairies, Spacy se montrait la plus efficace en terme de temps d'exécution, avec un nombre d'erreurs plus important que les deux autres mais relativement réduit dans l'ensemble. Avec nos moyens informatiques limités, nous avons privilégié cette librairie pour parser nos 3000 romans. Cette dernière a aussi l'avantage d'être très bien documentée et comporte plusieurs modèles de langages pour le français. Au vu des performances des différents modèles, nous prenons la décision d'utiliser le modèle fr_core_news_lg qui a un très bon rapport d'utilisation de ressources temporelles et matérielles entre exécution et performance. Nous utilisons aussi Scikitlearn [14], qui donne des outils efficaces pour l'analyse prédictive de données. Scikitlearn est assez simple à utiliser et implémente des algorithmes d'apprentissage machine au niveau de l'état de l'art.

5.3 Modélisation statistique

Nous voulons observer si des différences statistiques majeures existent entre nos deux sous-corpus, celui canonique et celui non-canonique. Nous faisons appel au champ de recherches de l'apprentissage machine. L'objectif de la modélisation statistique est de tirer des généralisations sur de grands jeux de données. Un exemple explicite est la régression linéaire, qui cherche à établir une relation entre des variables diverses.

L'apprentissage machine fait référence à toute une série d'algorithmes statistiques qui traitent chaque texte comme un amalgame de certaines caractéristiques quantifiables. Cette approche part du principe que l'on peut quantifier leur répartition dans les textes de manière à identifier les différences entre ces derniers, afin de classifier ou prédire la catégorie à laquelle un texte est susceptible d'appartenir.

La classification automatique de textes est un problème très étudié en statistiques. Plusieurs estimateurs sont au niveau de l'état de l'art : des modèles linéaires, comme le naïve bayes ou la classification ridge. Une famille de modèle retient particulièrement notre attention parce qu'elle obtient de bons résultats pour la classification de textes littéraires : les Machines à Vecteur de Support (SVM) [15]. Les SVM ont pour but de

11. https ://spacy.io/
12. https ://github.com/stanfordnlp/stanza
13. https ://github.com/hipster-philology/nlp-pie-taggers
14. https ://scikit-learn.org/stable/index.html
15. B. Yu, « An evaluation of text classification methods for literary study », *Literary and Linguistic Computing*, 23–3 (5 sept. 2008), p. 327-343, DOI : 10.1093/llc/fqn015.

trouver les plans qui séparent les points de données avec les marges maximales entre les frontières de décision. Ils traitent les caractéristiques comme des coordonnées dans un espace cartésien à haute dimension et tentent de tracer une ligne qui divise au mieux les caractéristiques uniques d'une classe [16]. Pour l'estimateur SVM, un texte n'est qu'une combinaison de caractéristiques qui tendent à apparaître plus souvent dans une classe de textes que dans une autre. Le modèle assigne des coefficients à chaque mot pour estimer la probabilité que le roman soit canonique.

Les SVM ont l'avantage de réduire le risque de sur-apprentissage. Ce dernier se caractérise lorsqu'un modèle statistique se spécialise trop sur ses données d'entraînement, ce qui réduit sa performance de généralisation. Nous utilisons dans ce mémoire la famille de SVM développé par l'équipe scikit-learn [17] depuis 2011.

Nous avons affaire à une approche très connue en apprentissage machine : l'apprentissage supervisé. Ce dernier est assez simple à comprendre, puisqu'il associe un ensemble de données avec une certaine classe labellisée. Un modèle statistique est évalué sur sa capacité à faires des inférences entre les particularités des données et une certaine classe. Les classes de nos romans correspondent aux méta-données récupérées, c'est à dire si le roman en question appartient ou non au canon littéraire.

5.3.1 Implémentation de l'apprentissage machine

Nous mettons en place les bases de l'apprentissage machine. Le jeu de données est séparée en deux échantillons.

— Un premier sur lequel nous entraînons le modèle statistique, c'est à dire que nous lui donnons le label associé pour chaque roman.

— Un autre sur lequel nous évaluons ses performances de prédictions sur des données qu'il n'a jamais vu.

Nous mesurons ainsi la capacité du modèle à généraliser.

Nous voudrions que la taille de l'échantillon du corpus d'entraînement soit la plus large possible, pour donner toutes ses chances au modèle. Pour autant, il est important de garder une taille conséquente pour l'échantillon de test afin de mesurer à quel point le modèle est capable de réaliser de bonnes prédictions sur un grand nombre de données. Nous fixons la taille de l'échantillon test à 20% du total.

Nous implémentons grâce à Scikitlearn un pipeline avec un pré-traitement des données, un StandardScaler et un estimateur, le SVM. Ce pré-traitement permet de norma-

16. Pour de plus amples informations sur les SVM, voir l'article fondateur de Vladimir Vapnik dans Corinna Cortes et Vladimir Vapnik, « Support-vector networks », *Machine Learning*, 20–3 (sept. 1995), p. 273-297, DOI : 10.1007/BF00994018

17. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, « Scikit-learn : Machine Learning in Python », *Journal of Machine Learning Research*, 12 (2011), p. 2825-2830.

liser nos données. La normalisation d'un ensemble de données est une exigence commune à de nombreux estimateurs d'apprentissage automatique : ils peuvent mal se comporter si la distribution statistique ne ressemble pas à des données distribuées sous forme d'une loi normale. Nous avons implémenté le SVM du logiciel SuperStyl [18], de Jean-Baptiste Camps, puis nous avons codé un algorithme mieux adapté à notre approche, avec des estimateurs plus performants et une méthode spécifique pour une classification binaire. [19]

Nous contrôlons la robustesse de nos résultats par une validation croisée en cinq parties. Cette validation est une procédure qui permet d'éviter le sur-apprentissage et de garantir la fiabilité de nos modèles. Elle sépare l'ensemble du jeu de données en cinq parties, et réalise cinq entraînements distincts en prenant à chaque fois quatre parties pour l'entraînement et une pour s'évaluer. La moyenne de l'efficacité du modèle sur ces cinq entraînements est ainsi un score beaucoup plus robuste et fiable.

5.3.2 Métriques d'évaluation du modèle

Nous évaluons notre modèle grâce à des métriques d'évaluation de la performance : l'efficacité, la précision, le rappel et un f1-score.

Pour comprendre ces métriques, il faut se familiariser avec les notions de vrai-positif (TP), vrai-négatif (TN), faux-positif (FP) et faux-négatif (FN). Dans notre cas, un TP est un roman considéré comme canonique par nos méta-données et prédit comme tel. Un TN est un roman non-canonique prédit comme non-canonique. Un FP est un roman non-canonique prédit comme canonique. Un FN est un roman canonique prédit comme non canonique. Une prédiction sans erreur équivaut donc à minimiser les mauvaises attributions, c'est à dire minimiser FP et FN.

L'efficacité est la métrique la plus simple à comprendre, puisque c'est le pourcentage d'éléments prédits correctement par le modèle.

$$Efficacite = \frac{TP + TN}{TP + TN + FP + FP}. \tag{5.1}$$

Comme on peut le voir sur la figure 5.2, la précision est la fraction d'éléments pertinents parmi les éléments extraits. Autrement dit, la précision est définie par :

$$Precision = \frac{TP}{TP + FP}. \tag{5.2}$$

Le rappel est la fraction d'éléments pertinents qui ont été extraits. Autrement dit, le rappel est définie par :

$$Rappel = \frac{TP}{TP + FN}. \tag{5.3}$$

18. J.B. Camps, *SUPERvised STYLometry (SuperStyl)*, version ... 2021, DOI : ....
19. Voir en annexe 7.5.3 le lien du github avec le code du mémoire

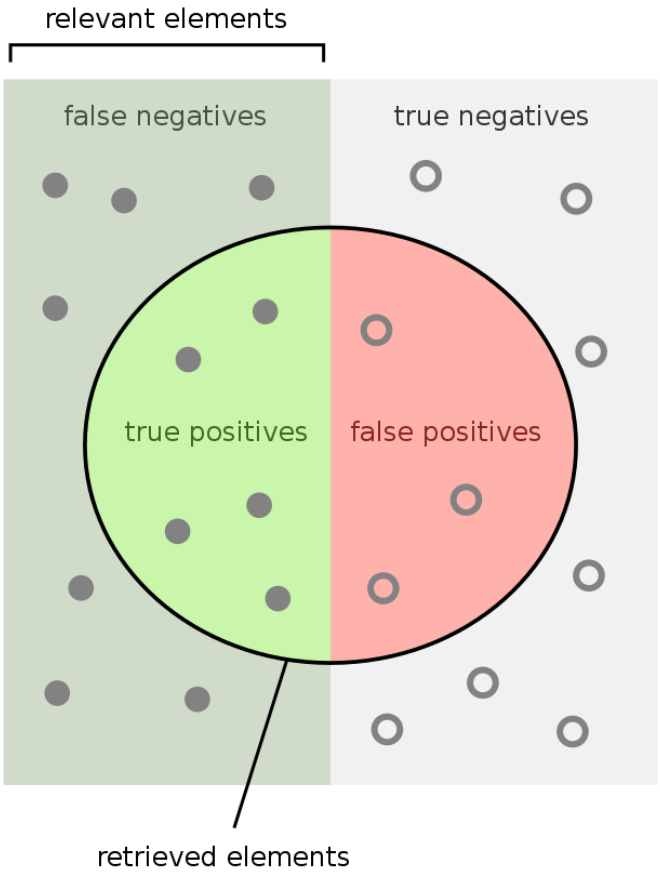

FIGURE 5.2 – Précision et rappel 1/2 [20]

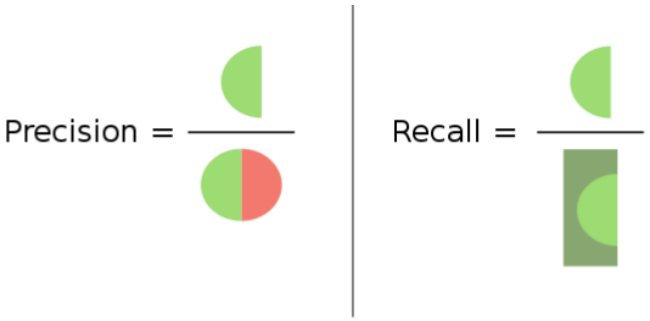

FIGURE 5.3 – Précision et rappel 2/2 [21]

Le score F1 peut être interprété comme une moyenne harmonique de la précision et du rappel, où un score F1 atteint sa meilleure valeur à 1 et son pire score à 0. La formule pour le score F1 est la suivante : F1 = 2 * (précision * rappel) / (précision + rappel)

Il faut être attentif à ces trois métriques pour s'assurer de la viabilité du modèle et

20. *Source : https : //en.wikipedia.org/wiki/Precision_and_recall*
21. *Source : https : //en.wikipedia.org/wiki/Precision_and_recall*

de sa performance. En effet, même lorsque l'efficacité du modèle est très bonne, les autres mesures peuvent, par exemple dans le cadre d'un jeu de données peu équilibré, montrer les limites du modèle en donnant des scores très bas.

Ainsi, nos expérimentations se fondent sur trois piliers principaux. Un corpus d'étude, un canon construit par nos soins et les méthodes quantitatives. A partir de cela, nous proposons de modéliser la notion de canon littéraire pour poser un diagnostique quantitatif de la filtration du contenu littéraire au cours du temps.

Troisième partie

Résultats et discussions

Chapitre 6

Une esthétique canonique multi-échelle

Dans ce chapitre, nous présentons les résultats de nos différentes approches. Les hypothèses du mémoire semblent se vérifier statistiquement, et nous évaluons leur solidité statistique avec différentes métriques et différentes approches.

6.1 A l'échelle du roman

Le modèle parvient à 75.3% d'efficacité en validation croisée à l'échelle du roman. Dans notre cas de classification binaire, cela veut dire que la prédiction de la canonicité se révèle être bonne pour trois textes sur quatre. Il est admis que les résultats d'un modèle statistique sont fiables lorsque ses performances atteignent entre 75% et 80% d'efficacité, si les autres métriques du modèle convergent vers une stabilité autour de ces valeurs.

	precision	recall	f1-score	support	accuracy
canon	0.746	0.598	0.662	52	
non-canon	0.754	0.858	0.806	76	
full dataset				128	**0.753**
macro-average	0.752	0.728	0.732	128	
weighted average	0.751	0.751	0.746	128	

TABLE 6.1 – Résultats de l'évaluation du modèle en validation croisée

Le modèle prédit mieux les ouvrages non-canoniques que les autres, mais le score de précision est plutôt bon (0.746). C'est le rappel (recall) qui pose des problèmes à notre modèle, c'est à dire que lorsqu'il prédit le label canonique il se trompe rarement, mais il n'arrive pas à le prédire dans un grand nombre de cas. Nous voyons là des signes du déséquilibre encore présent dans nos échantillons d'entraînement.

Pour renforcer les résultats, nous entreprenons une démarche aléatoire pour nous assurer que le modèle statistique détecte bien des différences textuelles associées à nos

méta-données plutôt que d'arriver artificiellement à séparer les deux classes. Cette approche aléatoire se réalise assez facilement, en désignant une classe canon ou non canon aléatoirement pour tous nos romans. Les résultats oscillent entre 45% et 55% d'efficacité, ce qui montre bien que le SVM n'arrive pas à séparer artificiellement nos deux classes sans raisons textuelles latentes. Ainsi, il se passe véritablement quelque chose dans les ouvrages canoniques. Nous désignons par dialecte canonique ce résultat, dont nous discuterons dans la suite du mémoire.

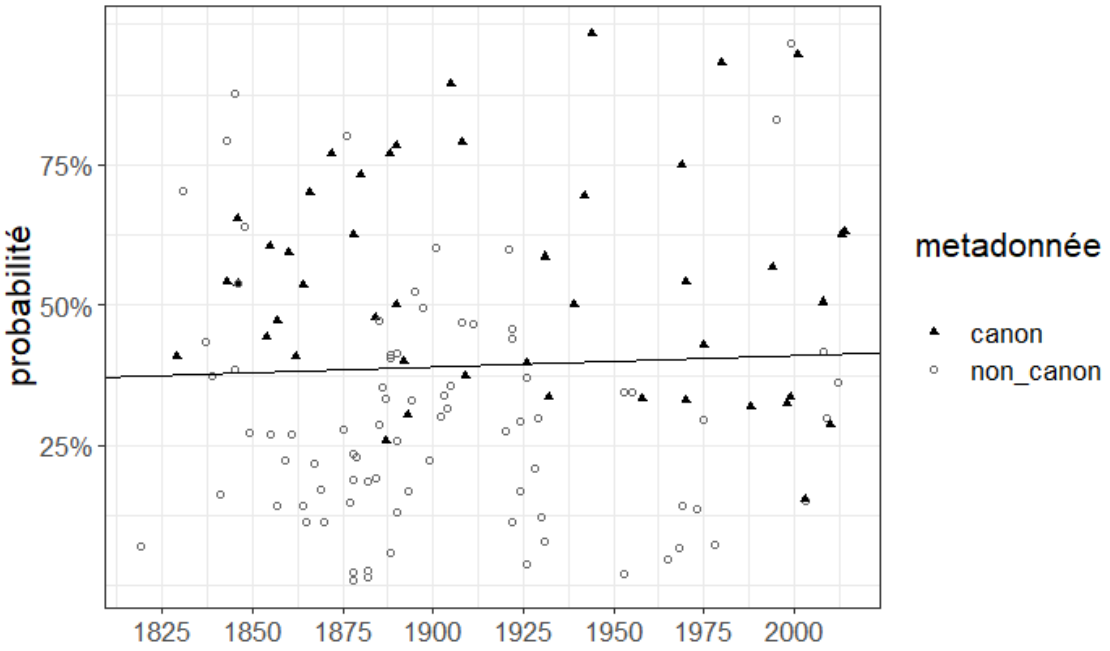

FIGURE 6.1 – Probabilité prédite d'appartenir au canon littéraire, canon des romans

Dans la figure 6.1 [1] nous projetons la probabilité prédite de chaque roman à appartenir au canon littéraire. Tous ces romans proviennent de l'échantillon de test, sur lequel on évalue la performance de généralisation du modèle. Les triangles représentent les romans classés comme canoniques et les cercles ceux non-canoniques. Le SVM arrive assez bien à détecter les deux classes, peu de romans canoniques sont classés non-canoniques, à l'exception de la fin du XX^{e}siècle, où le modèle fait quelques erreurs. La droite correspond à une régression linéaire de l'ensemble des probabilités canoniques assignées à chaque roman par notre modèle. Nous constatons ainsi une légère hausse de cette probabilité au cours du temps et nous commenterons ce résultat qui se renforce à l'échelle des auteurs.

1. Nous reproduisons ici la visualisation de Ted Undewood dans son livre T. Underwood, *Distant horizons*..., page 80, voir Id., *Tedunderwood/Horizon : Data And Code To Support Distant Horizons*, 24 mars 2018, DOI : 10.5281/ZENODO.1206317 pour le code

6.2 A l'échelle de l'auteur

Le modèle atteint 90.4% d'efficacité en validation croisée à l'échelle de l'auteur. Les résultats sont ainsi nettement meilleurs qu'à l'échelle des romans.

	precision	recall	f1-score	support	accuracy
canon	0.866	0.888	0.878	231	
non-canon	0.928	0.911	0.921	361	
full dataset				592	**0.904**
macro-average	0.898	0.902	0.898	592	
weighted average	0.904	0.904	0.904	592	

TABLE 6.2 – Résultats de l'évaluation du modèle en validation croisée

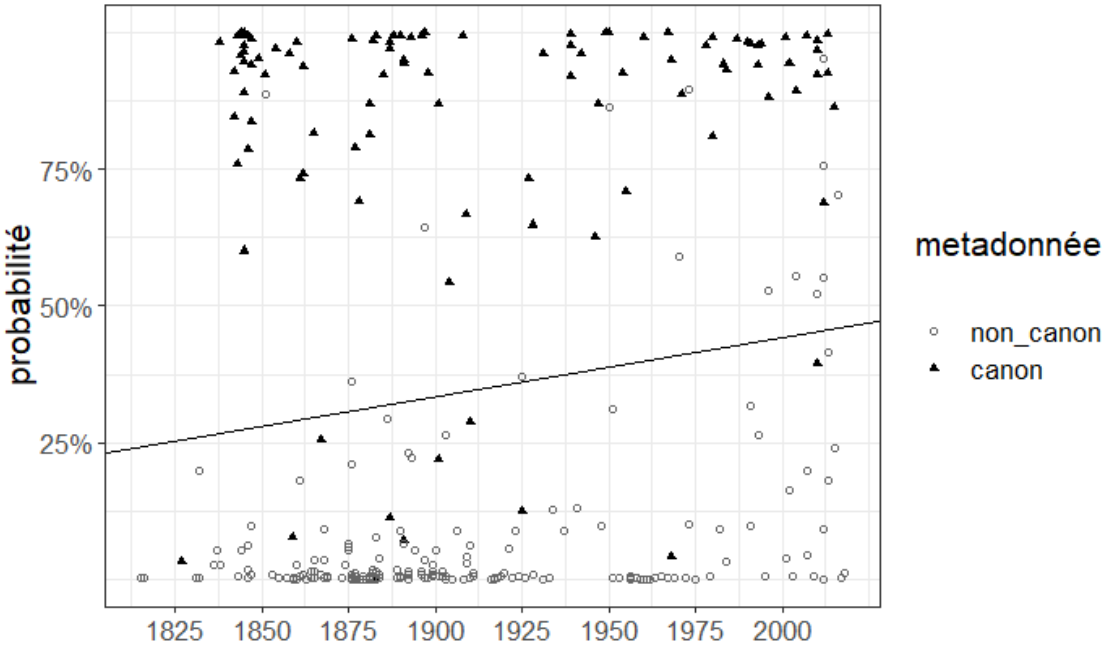

FIGURE 6.2 – Probabilité prédite d'appartenir au canon littéraire, canon des auteurs

Dans la figure 6.2 nous projetons de la même manière la probabilité prédite de chaque roman à appartenir au canon littéraire, avec les méta-données canoniques à l'échelle de l'auteur. On remarque que le SVM est bien meilleur à cette échelle, c'est à dire bien plus confiant dans ses prédictions qu'à l'échelle des romans. Les deux groupes sont très distincts sur la figure 6.2, et les erreurs sont moins nombreuses.

Au regard de ces résultats, nous nous sommes demandés si le modèle ne détectait pas des indices linguistiques d'un groupe d'auteurs.

Nous voulions savoir si le modèle ne reconnaissait pas l'idiolecte des auteurs canoniques sur lequel il s'était entraîné présents dans les œuvres de l'échantillon test. Si cela se vérifiait, notre modèle serait capable de réaliser des attributions d'auteurs et non de détecter un dialecte canonique.

Ainsi nous avons entrepris de créer des jeux de données ne comportant qu'un roman par auteur pour exclure la probabilité que le modèle ne reconnaisse les ouvrages canoniques par la signature textuelle de leur auteur. Les résultats de cette démarche gardent une très bonne stabilité, c'est à dire que le modèle détecte bien quelque chose sans relation explicite avec les auteurs.

Une des raisons qui pourraient expliquer la différence dans les performances entre les deux échelles de canonicité serait la taille de l'échantillon sur lequel le modèle est entraîné. En effet un défaut du SVM est de favoriser la classe la plus importante du jeu de données. Notre canon littéraire, à l'échelle du roman, n'est constitué que de 264 éléments, et même avec des procédures d'apprentissage non-équilibrées, il faut faire attention à la répartition des classes dans le jeu de donnée de l'échantillon d'entraînement. Ainsi, nous avons mis en place un ratio de répartition d'au moins 35% de romans canoniques et 65% de non canoniques au maximum. Le problème est que lorsque l'on travaille avec l'indice de canonicité à l'échelle des romans, la taille de l'échantillon de travail est moins important que celui à l'échelle des auteurs, de l'ordre de 800 éléments pour l'échelle de canonicité des romans et tout le corpus (3000 romans) à l'échelle des auteurs. De nombreux auteurs considérés comme canoniques ont beaucoup écrit d'ouvrages, cela explique la grande différence entre nos deux approches. Avec ce ratio et des procédures implémentées grâce à la librairie python imblearn [2], le modèle fait des progrès de l'ordre de 5% d'efficacité.

La régression linéaire projetée sur le graphe témoigne d'une tendance globale détectée par notre modèle. La probabilité d'appartenir au canon littéraire augmente avec le temps. Techniquement, cette hausse est une erreur. Les romans ne sont pas plus susceptibles d'appartenir au canon littéraire parce qu'ils sont publiés plus tard. Mais cela veut dire que le modèle échoue à produire des critères valides pour deux siècles de production littéraire. Les livres publiés plus tard comportent plus de signes linguistiques associés au canon littéraire. On remarque qu'une majorité des erreurs du modèle se trouvent entre les années 1975 jusqu'à nos jours. Des romans non canoniques (dont l'auteur n'est pas considéré comme canonique), obtiennent un bon score de canonicité. Le modèle perd en assurance et de nombreux romans se trouvent dans un entre-deux. On pourrait expliquer cela par une usure de la norme canonique, qui n'est plus aussi facile à discerner depuis les années 1975, malgré une distinction très forte durant 150 ans.

En plus de ces hypothèses, on pourrait invoquer les limites de l'expérience pour expliquer cette tendance. En effet, on remarque que beaucoup de données tests se situent dans la deuxième partie du XIX^{e}siècle, où par ailleurs le modèle est très performant.

2. https ://imbalanced-learn.org/stable/index.html

C'est peut-être un biais du corpus que nous avons là, puisque cette période est proportionnellement très représentée dans notre corpus, comme on peut le constater en figure 3.1. Le modèle a plus de données d'entraînement sur cette période, donc il se spécialise dessus. Pour autant, ce sur-apprentissage ne semble pas rédhibitoire puisque les données sont très bonnes sur les autres périodes aussi. Cela témoigne plutôt de la stabilité de la norme esthétique, puisque notre modèle repose en partie sur un léger sur-apprentissage des données plus nombreuses de la deuxième partie du XIX^{e}siècle, mais est capable de généraliser sur la longue durée.

Chapitre 7

Discussions

7.1 Étude des cas limites

L'étude des erreurs les plus importantes de nos modèles peut dire beaucoup de la manière dont ils se comportent.

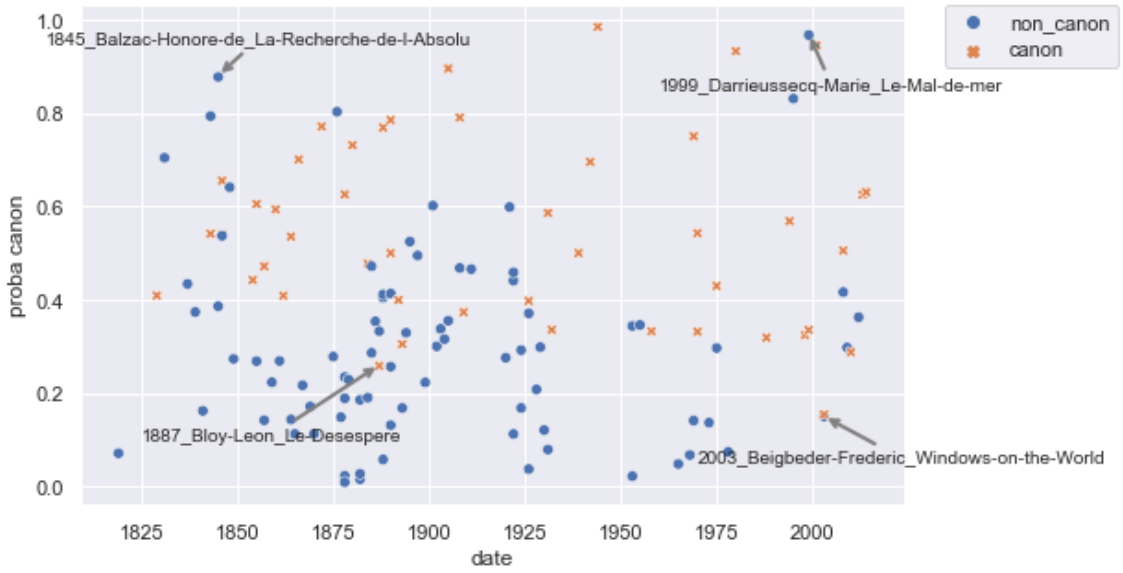

FIGURE 7.1 – Cas limites du modèle entraîné à l'échelle des romans

La figure 7.1 met en avant quatre erreurs du modèle. Nous rappelons qu'à cette échelle le modèle performe à 75% d'efficacité, donc les erreurs sont assez nombreuses. La canonicité prédite du roman de Balzac *La recherche de l'absolu* est très importante, avec près de 90% de probabilité. Pourtant, ce roman ne fait pas partie de notre canon littéraire. Ce roman de Balzac est classé dans les Études philosophiques de La Comédie humaine. Pour notre modèle, ce roman remplit tous les critères linguistiques relatifs au canon littéraire. Il en va de même pour le roman de Marie Darrieussecq, *Le mal de mer*, qui pourtant est publié 150 années plus tard. Ce dernier n'est pas particulièrement resté dans les mémoires, du moins il n'est pas rentré dans un processus de canonisation. Pourtant,

notre modèle le détecte comme tel. Cela peut re-légitimer cette œuvre, qui, au moins dans son contenu linguistique, fait aussi bien que les grands classiques de la littérature française. On constate d'autres erreurs du modèle dans le sens inverse. Des textes appartenant à notre canon littéraire ne sont pas bien notés par le modèle statistique. Par exemple, *Windows on the world*, de Frédéric Beigbeder, fait parti de notre canon puisqu'il a gagné le prix Interallié en 2003. Ce roman relate les dernières minutes des victimes des attaques du World Trade Center le 11 septembre 2001. On constate ici les limites de nos méta-données, puisque le roman en question a pour vocation de raconter une histoire qui puisse toucher beaucoup de gens, en les amenant sur un sujet qui a bouleversé la société, plus que celle de rentrer dans le canon littéraire. *Le désespéré*, est le premier roman de Léon Bloy, publié en 1887. Ce roman est canonique parce qu'il a été republié dans la collection *Littérature Classique* de Garnier Flammarion en 2010. Il n'est pas aisé de comprendre pourquoi ce roman ne correspond pas à l'esthétique canonique perçue par notre modèle. Pour autant, la singularité de cette œuvre a été soulignée par Pierre Glaudes dans l'appareil critique de la réédition : « *Le désespéré* est surtout un aérolithe littéraire, écrit dans une langue barbelée de mots rares, étrangement mystique, une œuvre d'une surprenante modernité ». Il semble que notre modèle donne raison à Pierre Glaudes, le roman sort des conventions et de l'esthétique du canon littéraire.

A l'échelle des auteurs, le modèle est bien plus performant, mais quelques erreurs restent à noter comme on peut le voir en figure 7.2.

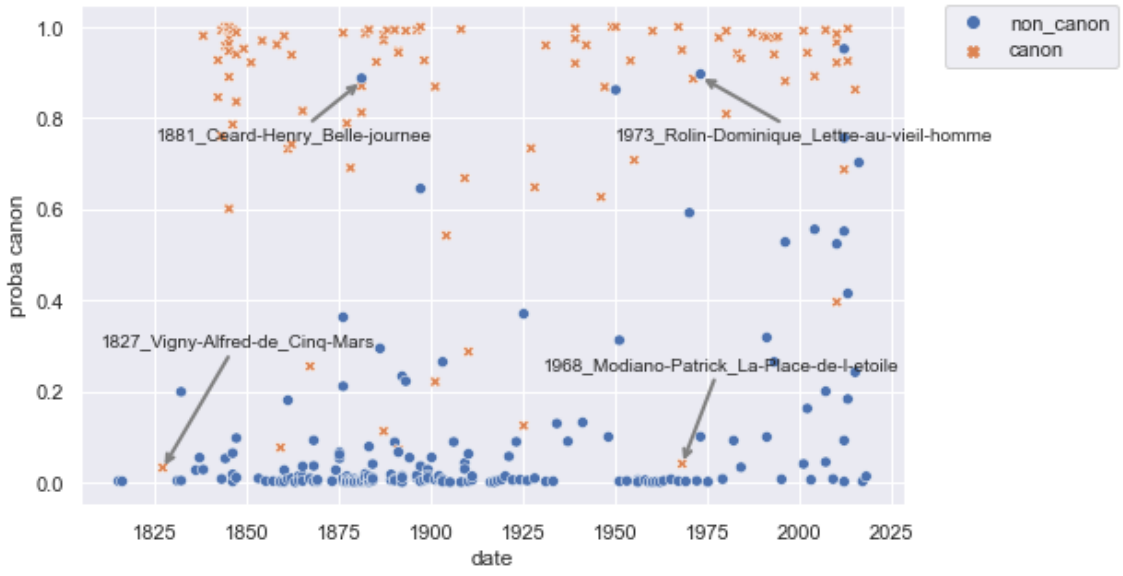

FIGURE 7.2 – Cas limites du modèle entraîné à l'échelle des auteurs

On a par exemple la *Belle journée* d'Henry Céard qui est très bien noté par le modèle. Paru en 1881, ce roman est une ré-interprétation de *Madame Bovary*, en forme d'hommage à Flaubert, mort un an auparavant. Henry Céard est un disciple de Flaubert et de Zola, son style se rapproche de ces derniers, même si l'histoire littéraire n'a pas retenu

cet auteur. On peut donc expliquer cette erreur par la démarche même du roman qui est une imitation du style et du thème du roman de Flaubert. De la même manière, le livre de Dominique Rolin *Lettre au vieil homme*, publié en 1973 est perçu comme canonique par le modèle, alors que l'écrivaine n'est pas retenue dans notre canon littéraire.

Au contraire, le roman *Cinq mars*, d'Alfred de Vigny, est dans notre canon mais n'est pas admis comme tel par le modèle. Ce roman est considéré comme le premier roman historique, et l'on peut émettre l'hypothèse que c'est en cela qu'il se démarque de la manière commune de faire des romans pour notre modèle statistique. Enfin, le roman de Patrick Modiano *La place de l'étoile*, publié en en 1968 aux éditions Gallimard est lui aussi très mal noté par le modèle. Le registre parodique du roman semble expliquer le fait qu'il n'ait pas sa place dans le canon littéraire perçu par le modèle.

7.2 Caractéristiques discriminantes du modèle statistique

Il faudrait maintenant comprendre comment nos modèles statistiques sont capables de prédire la canonicité. Un des intérêts de l'apprentissage machine est d'ailleurs la possibilité de plonger dans les inférences réalisées par le modèle. En effet, on peut récupérer les coefficients que le modèle assigne à chaque caractéristique pour séparer nos deux groupes. Dans la figure 7.3, nous projetons les 40 caractéristiques les plus discriminantes pour le modèle. En bleu, nous avons les 20 caractéristiques qui donnent le plus de poids à l'esthétique canonique, et en rouge celle non-canonique.

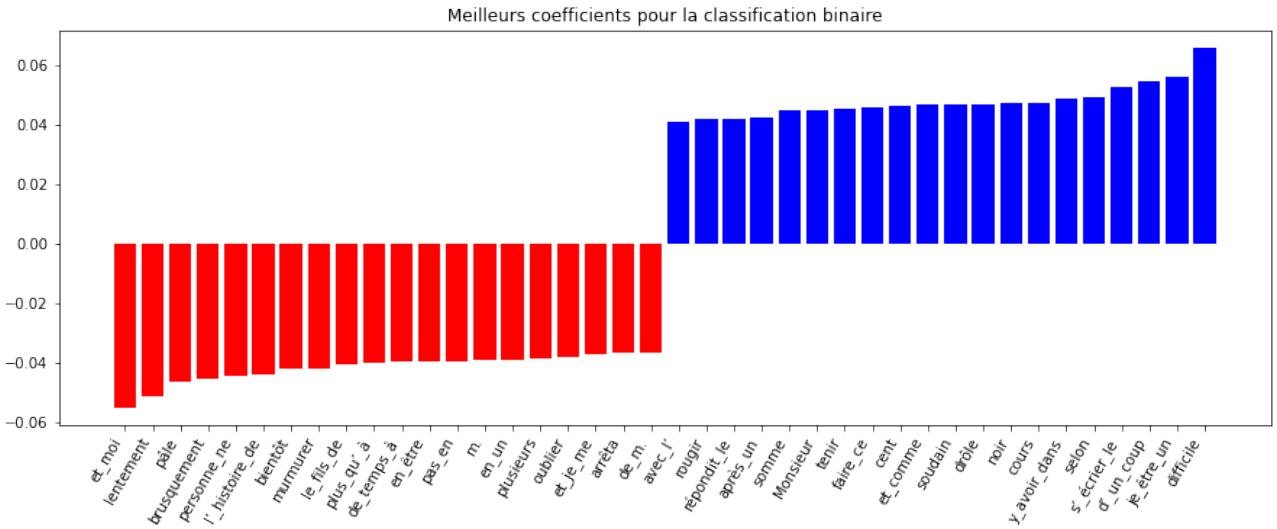

FIGURE 7.3 – coefficients discriminants pour le modèle, canon des auteurs

Il est difficile d'appréhender ce qui se joue dans ces coefficients. Les mots *difficile*,

drôle ou *rougir* sont considérés plus canoniques que les mots *pâle*, *oublier* ou *murmurer*. Le problème interprétatif auquel nous faisons face n'est pas dû à l'apprentissage machine mais plutôt à la complexité du phénomène que nous modélisons, la réception littéraire. Il est compliqué d'attribuer la canonicité à un groupe de mots précis, et comme nous avons fondé notre travail sur une approche linguistique avec des mots outils, nous avons en plus à interpréter des mots très courants.

Pour comprendre ce que le modèle détecte, nous représentons dans le texte les coefficients de ce dernier. Les 500 caractéristiques les plus canoniques sont coloriés en bleu, les 500 les plus non-canoniques en rouge et le reste n'est pas représenté. Nous projetons en figure 7.2 ces coefficients sur un extrait de *Sido*, roman écrit par Colette et publié en 1929.

– et pourquoi cesserais -je d' être de mon village ? il n' y faut pas compter . te voilà bien fière , mon pauvre min
et - chéri , parce que tu habites paris depuis ton mariage . je ne peux pas m' empêcher de rire en constatant combie
n tous les parisiens sont fiers d' habiter paris , les vrais parce qu' ils assimilent cela à un titre nobiliaire , l
es faux parce qu' ils s' imaginent avoir monté en grade . à ce compte - là , je pourrais me vanter que ma mère est n
ée boulevard bonne-nouvelle ! toi , te voilà comme le pou sur ses pieds de derrière parce que tu as épousé un parisi
en . et quand je dis un parisien ... les vrais parisiens d' origine ont moins de caractère dans la physionomie . on
dirait que paris les efface ! elle s' interrompait , levait le rideau de tulle qui voilait la fenêtre : – ah ! v
oici mlle thévenin qui promène en triomphe , dans toutes les rues , sa cousine de paris . elle n' a pas besoin de le
dire , que cette dame quériot vient de paris : beaucoup de seins , les pieds petits , et des chevilles trop fragiles
pour le poids du corps ; deux ou trois chaînes de cou , les cheveux très bien coiffés ... il ne m' en faut pas tant
pour savoir que cette dame quériot est caissière dans un grand café. une caissière parisienne ne pare que sa tête et
son buste , le reste ne voit guère le jour . en outre , elle ne marche pas assez et engraisse de l' estomac . tu ver
ras beaucoup , à paris , ce modèle de femme - tronc . ainsi parlait ma mère , quand j' étais moi - même , autrefoi
s , une très jeune femme . mais elle avait commencé , bien avant mon mariage , de donner le pas à la province sur pa
ris . mon enfance avait retenu des sentences , excommunicatoires le plus souvent , qu' elle lançait avec une force
d' accent singulière . où prenait -elle leur autorité , leur suc , elle qui ne quittait pas , trois fois l' an , son
département ? d' où lui venait le don de définir , de pénétrer , et cette forme décrétale de l' observation ? ne
l' eussé-je pas tenu d' elle , qu' elle m' eût donné , je crois , l' amour de la province , si par province on n' en
tend pas seulement un lieu , une région éloignés de la capitale , mais un esprit de caste , une pureté obligatoire d
es mœurs , l' orgueil d' habiter une demeure ancienne , honorée , close de partout , mais que l' on peut ouvrir à to
ut moment sur ses greniers aérés , son fenil empli , ses maîtres façonnés à l' usage et à la dignité de leur maison
. en vraie provinciale , ma charmante mère , « sido » , tenait souvent ses yeux de l' âme fixés sur paris . théâtr
es de paris , modes , fêtes de paris , ne lui étaient ni indifférents , ni étrangers . tout au plus les aimait -elle
d' une passion un peu agressive , rehaussée de coquetteries , bouderies , approches stratégiques et danses de guerre
. le peu qu' elle goûtait de paris , tous les deux ans environ , l' approvisionnait pour le reste du temps . elle re
venait chez nous lourde de chocolat en barre , de denrées exotiques et d' étoffes en coupons , mais surtout de progr
ammes de spectacles et d' essence à la violette , et elle commençait de nous peindre paris dont tous les attraits ét
aient à sa mesure , puisqu' elle ne dédaignait rien . en une semaine elle avait visité la momie exhumée , le musée
agrandi , le nouveau magasin , entendu le ténor et la conférence sur la musique birmane . elle rapportait un manteau
modeste , des bas d' usage , des gants très chers . surtout elle nous rapportait son regard gris voltigeant , son
teint vermeil que la fatigue rougissait , elle revenait ailes battantes , inquiète de tout ce qui , privé d' elle ,
perdait la chaleur et le goût de vivre .

FIGURE 7.4 – Extrait de Sido, écrit par Colette, canonicité coloriée

Dans cet extrait, nous voyons bien que le modèle ne détecte pas une manière d'être canonique. Colette écrit avec des mots considérés comme non-canonique par le modèle, comme les pronoms possessifs *mon* ou *ma*. Le crible du modèle statistique révèle un ensemble de traits linguistiques spécifique mais pas exclusif à ces œuvres canoniques. Le verbe *rougir* ou l'adjectif qualificatif *charmante* en fin d'extrait ne sont pas uniquement dédiés aux auteurs canoniques, mais le choix de ces mots est d'avantage réalisé par ces derniers.

7.3 Des motifs stylistiques

Pour palier ce problème interprétatif, nous avons revu notre démarche, avec l'objectif de détecter plus d'éléments relatifs au style. Nous avons implémenté pour cela une approche similaire à la technique des motifs stylistiques. Selon, Dominique Legallois, Thierry Charnois et Meri Larjavaara,

> « la caractérisation du style d'un auteur, peut bénéficier d'une nouvelle d'une nouvelle méthode en linguistique de corpus, la découverte de modèles séquentiels ou de "motifs", c'est-à-dire des chaînes contiguës de formes de mots/lemmas/POS tag. L'analyse des motifs peut être considérée comme complémentaire aux approches discrètes » [1].

Dans cette section, les motifs ne vont pas nous servir à caractériser un style d'auteur, mais plutôt à spécifier l'esthétique canonique que nous détectons. Nous simplifions l'approche pour construire nos motifs. On récupère de la même manière des uni-grammes, bi-grammes et tri-grammes, avec cette fois ci l'étiquetage morpho-syntaxique des mots-outils et le lemme des autres mots. Nous mettons le tableau des scores en annexe 7.5.3. Les résultats sont sensiblement similaires à notre première approche, ce qui s'explique par la proximité des caractéristiques textuelles utilisées. Ces résultats semblent toutefois renforcer notre première intuition : il y a bien une esthétique particulière sélectionnée dans le canon littéraire construit par nos soins.

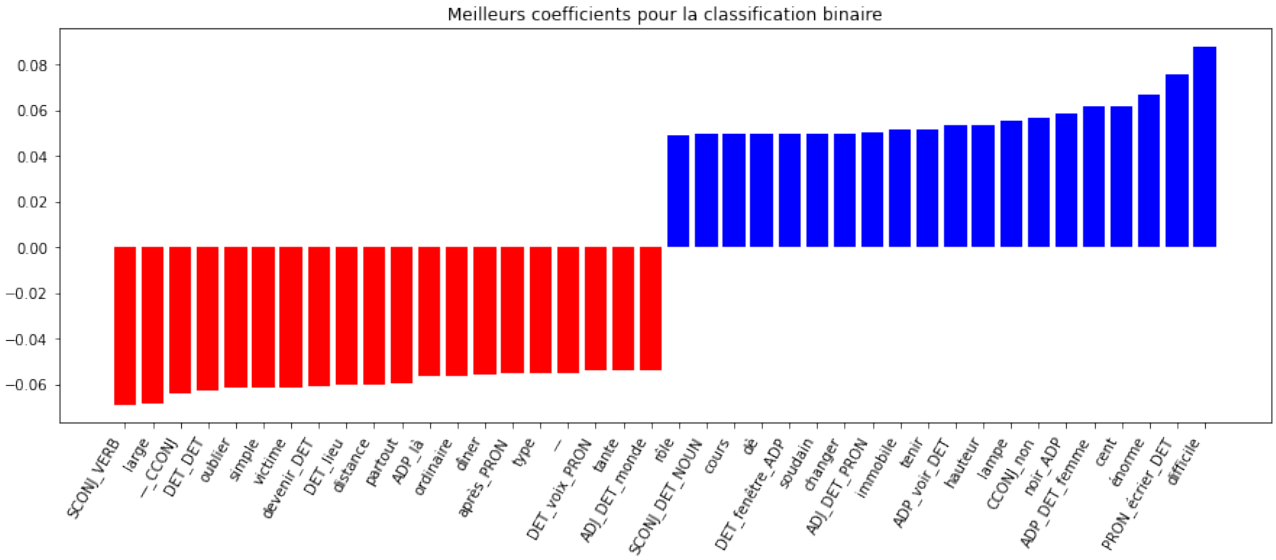

FIGURE 7.5 – coefficients discriminant pour le modèle des motifs, canon des auteurs

1. Dominique Legallois, Thierry Charnois et Meri Larjavaara, « The balance between quantitative and qualitative literary stylistics : how the method of "motifs" can help », dans *The Grammar of Genres and Styles*, dir. Dominique Legallois, Thierry Charnois et Meri Larjavaara, 2018, p. 164-193, DOI : 10.1515/9783110595864-008.

Dans la figure 7.5 nous projetons les coefficients discriminants. On voit bien qu'en plus de mots simples, le modèle assigne de bons scores à des manières de construire une phrase. Le bi-gramme SCONJ_VERB est la concaténation de deux mots consécutifs, en l'occurrence une conjonction de coordination avec un verbe, par exemple « Il n'est pas heureux, *parce qu'habiter* en province n'est pas dans son intérêt ». Cette forme se rapproche plus selon le modèle d'une écriture non-canonique. Au contraire, le tri-gramme PRON_écrier_DET, qui donne par exemple« s'écria le personnage », avec la forme pronominale du verbe écrier, est plus proche d'une écriture canonique. C'est une manière d'introduire du discours direct dans le récit, l'on ne peut pas conclure d'une utilisation plus prononcé du discours direct dans les romans canoniques avec cet unique indice.

7.4 Test sur des données hors domaine d'étude

Pour comprendre un peu mieux ce que le modèle détecte concrètement, nous lui avons donné des textes hors de son domaine d'étude pour voir quel score leur était attribué. En effet le modèle peut donner un score à tout texte écrit, même si ce n'est pas un texte littéraire. Il nous fallait des textes assez conséquents en terme de longueur, pour éviter d'avoir une matrice d'entraînement trop vide. Nous lui avons présenté deux différents textes :

— La page wikipédia du théorème de Bayes. [2]

— Un extrait du Code Civil, le chapitre 2 du titre 2 du livre II. [3]

Ce sont deux textes très formels, le premier présente une vulgarisation scientifique d'un théorème mathématique, avec une histoire et une contextualisation du théorème. L'autre comporte un chapitre du Code Civil, le tout représentant 8 articles avec de nombreux alinéas. En terme de longueur, les deux textes sont assimilables à des nouvelles suffisamment longues pour être modélisées correctement. Voici les résultats :

	proba canon	proba non-canon	prediction
Article du Théorème de Bayes	0.375779	0.624221	non_canon
Code Civil	0.454353	0.545647	non_canon

TABLE 7.1 – Résultats des expérimentations hors domaine, canon des œuvres.

Le modèle reposant sur le canon des œuvres réagit assez bien à ce test, dans le sens où il n'a l'air de reconnaître ni les traits canoniques ni ceux non-canoniques. Si les deux sont classés « non-canon », la probabilité d'appartenance montre que le modèle n'est

2. https://fr.wikipedia.org/wiki/Théorème_de_Bayes
3. https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006070721

pas vraiment sûr de lui. Le fait que ces deux énoncés très formels soient classés comme non-canon pourrait soutenir l'idée d'un style plus développé dans le canon littéraire.

	proba canon	proba non-canon	prediction
Article du Théorème de Bayes	0.484885	0.515115	non_canon
Code Civil	0.580976	0.419024	canon

TABLE 7.2 – Résultats des expérimentations hors domaine, canon des auteurs

Le modèle reposant sur le canon des auteurs est lui aussi dubitatif quand à ces textes et ne sait pas comment les classer. A cette échelle le modèle se montrait très sûr de lui, et très peu de scores étaient entre 0.8 et 0.2. Cela montre bien que ces deux textes ne ressemblent en rien aux traits canoniques ou non-canoniques.

7.5 Sélectivité canonique dans la production d'un auteur

Nos modèles détectent une filtration linguistique d'une certaine manière d'écrire des romans. Cette sélection ne se fait pas seulement au niveaux des auteurs, que l'on considère comme classique ou non, mais aussi au sein de la production littéraire d'un même auteur. Nous présentons ici les résultats des expériences menées à partir des romans de Colette, Georges Perec et Guy de Maupassant.

Pour mieux comprendre ce qui se joue à cette échelle, nous avons mis en place d'autres expériences. Avec des réductions de dimension, et plus précisément des analyses en composantes principales (ACP), nous projetons tous les ouvrages d'un même auteur sur un seul plan pour pouvoir comparer les ouvrages entre eux.

L'ACP est une méthode bien connue de réduction de dimension qui va permettre de visualiser des données complexes pour y discerner des regroupements, des typologies. L'ACP entraîne une perte d'information car les dimensions sont réduites, mais les informations conservées sont sensées être les plus significatives.

7.5.1 Colette

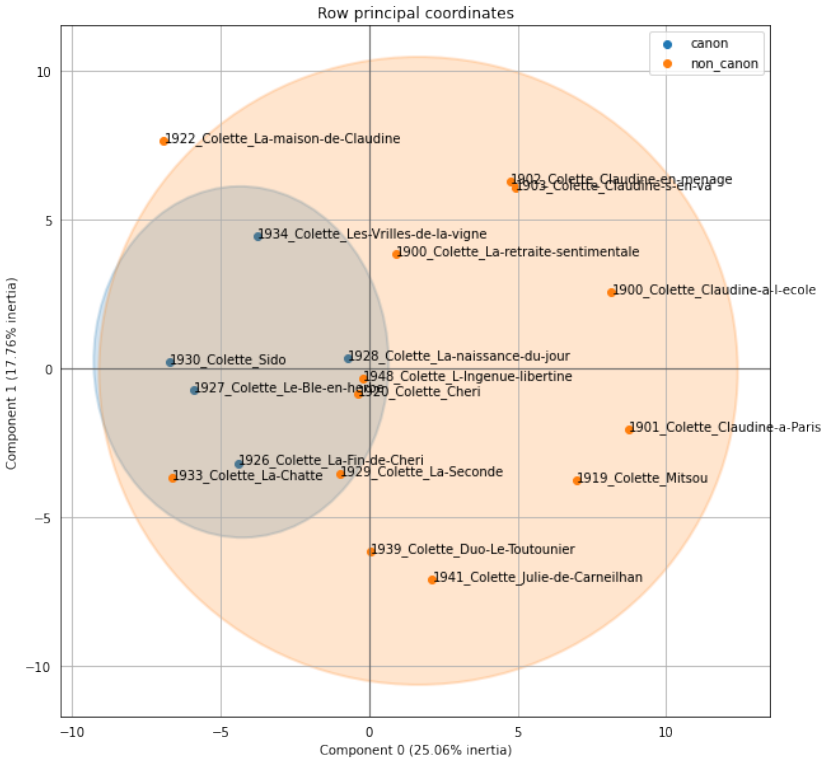

FIGURE 7.6 – Sélectivité canonique chez Colette

Nous avons réalisé des réductions de dimension en figure 7.6 sur les écrits de Colette,

une écrivaine célèbre du début du XX^{e}siècle.

Deux éléments sont mis en avant dans ce graphique, d'une part en orange les romans de Colette non-canoniques, et d'autre part en bleu les ouvrages considérés comme canoniques. On remarque que ces derniers forment un groupe assez distinct au sein de la production littéraire de Colette. L'ACP appose naturellement les romans canoniques dans la même partie du graphique, c'est à dire qu'il y a des similarités conséquentes entre ces romans. Les cinq romans canoniques sont tous écrits entre 1926 et 1934, ce qui pourrait expliquer la présence de ce groupe en tant qu'il correspond à un moment littéraire précis dans la vie de l'auteur. Loin de ce groupe se trouve la série des *Claudine*, qui ont été des ouvrages très populaires mais pas intégré dans le canon. Ces ouvrages ont fait la popularité de l'auteur au début de sa carrière, mais ne correspondaient pas aux critères de sélection du canon.

7.5.2 Georges Perec

Georges Perec est un écrivain qui a publié la majorité de ses œuvres dans la deuxième partie du XX^{e}siècle. Notre canon a retenu deux de ses ouvrages, *Les choses*, publié en 1965 et *La vie mode d'emploi*, publié en 1978.

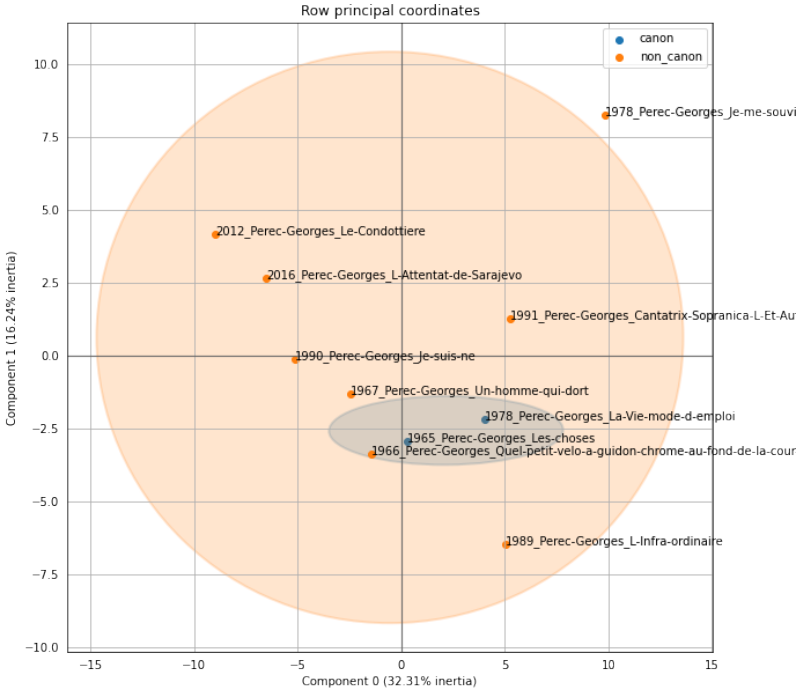

FIGURE 7.7 – Sélectivité canonique chez Georges Perec

L'ACP en figure 7.7 montre bien que ces deux ouvrages sont très similaires dans leur

contenu, malgré le fait qu'ils aient été publiés à 13 années d'intervalle. Les romans acceptés dans le canon par la réception correspondent donc bien à une esthétique particulière, au sein même de la production d'un écrivain.

On remarquera que le roman *Je me souviens* est très loin des autres romans, à cause de son caractère stylistiquement subversif. Le roman est un enchaînement de souvenirs de l'auteur, avec une anaphore « Je me souviens » sur tout le roman. C'est un roman qui a marqué la période mais il n'est pas rentré dans les critères de notre canon.

7.5.3 Guy de Maupassant

Il est important de souligner que cette expérience ne marche pas pour tous nos auteurs. Examinons l'exemple des ouvrages de Guy de Maupassant, en figure 7.8 : ce dernier est un auteur très productif, et la visualisation de la réduction de dimensionnalité ne parvient pas à séparer les canons des non-canons.

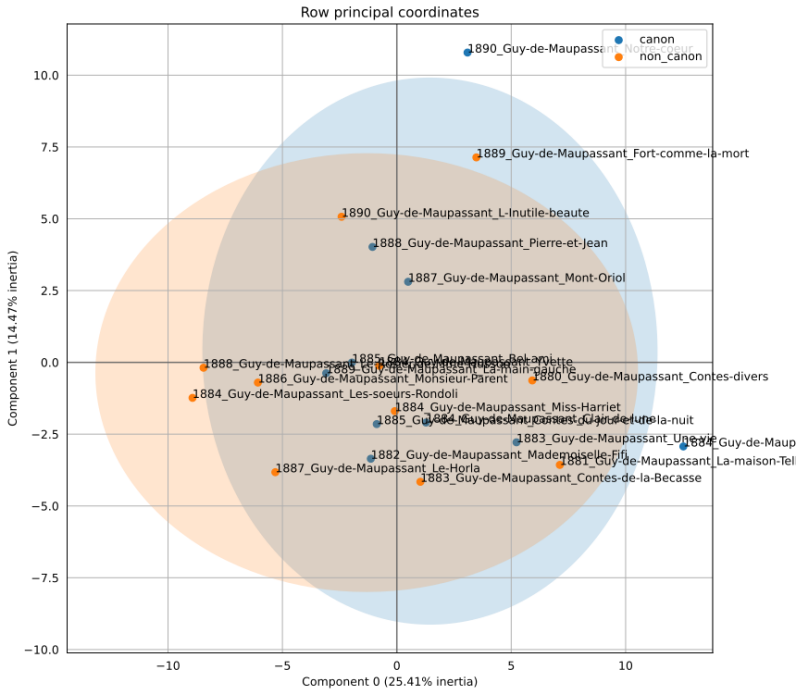

FIGURE 7.8 – Sélectivité canonique chez Maupassant

Il y a une superposition des deux ensembles des œuvres, canoniques et non-canoniques. La critique et surtout l'institution scolaire ont tellement sacralisé le style de cet auteur que la distinction entre leurs œuvres de premier et de second plan ne tient plus, comme si le tamis sélectif avait fini par accepter la manière d'écrire de cet écrivain dans son ensemble, peu importe ses ouvrages.

Ainsi, l'esthétique canonique constatée à l'échelle de milliers de romans par nos modèles statistiques semble sortir renforcée de nos différents tests et entrées de compréhension des données du modèle. L'étude des cas limites et les tests hors domaine d'étude semblent montrer la solidité du modèle ainsi que sa capacité à généraliser. Cette sélectivité du canon littéraire se constate aussi à l'échelle de la production d'un auteur, et ce malgré sa qualité de classique. Nos ACP montrent la filtration d'un certain type de contenu dans la production littéraire, entre un contenu conservé dans la mémoire collective et un autre laissé à l'abandon littéraire.

Conclusion

Ce travail propose une étude en profondeur de la notion de canon littéraire. Nous avons produit un ensemble de critères, fondés sur des constats historiques, pour caractériser le canon littéraire de la réception contemporaine. L'objectif de ce rapport est d'enrichir la vision traditionnelle qui envisage le canon comme une construction arbitraire, politique, idéologique, ou relative au hasard. Nous nous sommes intéressés au contenu textuel des ouvrages pour ajouter à cette définition un aspect formel, d'ordre esthétique et interne aux œuvres qui pourrait expliquer la sélectivité du canon littéraire. À l'aide d'un grand corpus de romans, des méthodes quantitatives de l'apprentissage machine et du traitement automatique des langues, nous avons pu modéliser en lecture distante la notion de canon littéraire. Cette modélisation se fonde sur les dynamiques textuelles qui régissent les standards du jugement esthétique et la filtration institutionnelle qui en résulte.

Un des apports principaux de ce mémoire a été de montrer qu'il existait une esthétique canonique, et qu'un modèle statistique pouvait prédire la canonicité avec 75% à 90% d'efficacité. S'il a été difficile d'interpréter et de caractériser cette esthétique, ou de revenir à des enjeux littéraires plus concrets, nous avons pu établir les analyses suivantes :

— notre modèle détecte une usure de l'esthétique canonique dans les dernières décennies de notre histoire littéraire. On pourrait discuter de la pertinence de la notion de canon sur les décennies plus contemporaines, puisque la filtration temporelle par les différentes institutions n'a pas encore *tamisé* les ouvrages.

— malgré cette usure contemporaine, nous avons montré une grande stabilité du jugement esthétique sur près de deux siècles. Notre modèle a été capable de détecter des lignes structurantes bien plus stables que décrites par l'histoire littéraire. Cela pourrait s'expliquer par la seule prise en compte de la réception contemporaine. De plus amples recherches seraient nécessaires pour conclure avec certitude.

Par ailleurs, en distinguant le canon du non-canon et en décrivant ce qui rend le roman canonique singulier dans son écriture, nous avons tenté de comprendre la fonction sociale remplie par ces textes particuliers. Les éléments textuels mis à jour sont les témoins d'une manière de *faire littérature*, manière sélectionnée et acceptée par la réception. Ils composent ce que l'on peut appeler un *dialecte canonique*, qui est constitué d'une langue sélectionnée et autorisée dans l'enseignement, instituée comme modèle de style pour des générations d'étudiants.

Ce travail ouvre de nombreuses perspectives de recherches, d'une part pour pallier les limites de la présente étude, et d'autre part, pour préciser l'esthétique canonique détectée. Notre démarche voulait saisir quantitativement les variables linguistiques qui caractérisent le phénomène social du prestige. La tâche était d'autant plus complexe qu'elle consistait à prédire des évènements arrivés à réception, c'est-à-dire après écriture. C'est pour ces raisons que nous avons opté pour une approche simple en *sac de mots*, avec un canon uniforme et des méta-données réduites. L'enjeu était de vérifier cette hypothèse

de langue littéraire. Il faudrait effectuer des recherches complémentaires pour comprendre exactement ce qui se joue dans le et les canons littéraires.

Une approche future serait de récupérer des méta-données en diachronie, c'est-à-dire au fur et à mesure de la filtration de la réception. On pourrait également fragmenter notre vision du canon par acteurs du champ littéraire (éditions, manuels scolaires, prestige académique, revues littéraires, ...). Nous pourrions aussi prendre des caractéristiques textuelles plus complexes, avec des techniques algorithmiques au niveau de l'état de l'art en traitement automatique des langues, comme par exemple les vecteurs de mots ou de paragraphes, ou encore la modélisation de sujet.

Les processus de décanonisation n'ont pas été abordés dans ce mémoire. Ils mériteraient un travail entier à eux seuls. Il faudrait identifier les ouvrages célébrés mais oubliés, qui sortent du canon au fil du temps, et étudier les raisons de cette marginalisation littéraire.

Ainsi, notre recherche s'est inscrite dans une réflexion sur la littérature et son canon en tant qu'objet de définition d'une culture nationale. Nous avons pu mettre au jour grâce à notre démarche quantitative une norme esthétique véhiculée par le canon et entretenue par la succession des politiques culturelles. Notre diagnostic témoigne des conditions sociologiques, idéologiques et culturelles de la production de la valeur littéraire et de ses critères de sélection. Ce travail est un modeste exemple de ce que peuvent apporter les méthodes quantitatives à la recherche en sciences humaines et sociales. Les humanités numériques permettent d'ouvrir de nouvelles perspectives de compréhension du monde pour façonner des savoirs historiques inédits et comprendre les normes sociales et politiques sur lesquelles s'est construite notre société.

Annexes

Disponibilité des données du mémoire

Illustrations, code et données du mémoire disponibles sur Github à cette adresse : https://github.com/crazyjeannot/canonization_process.

Version LaTeXdisponible sur Github, à cette adresse https://github.com/crazyjeannot/redaction_canonization_process.

Le canon des auteurs

Auteurs du canon	Nombre de romans classiques	Nombre total de romans
Honore De Balzac	19	85
Emile Zola	17	22
Romain Rolland	10	12
Guy De Maupassant	10	21
Alexandre Dumas	8	88
Victor Hugo	8	12
François René De Chateaubriand	7	7
Colette	5	18
Jules Verne	4	28
Alphonse Daudet	4	21
George Sand	4	51
Stendhal	4	7
Albert Camus	4	4
Edmond Et Jules De Goncourt	4	7
Jules Barbey D'Aurevilly	3	8
Louis Aragon	3	7
Patrick Modiano	3	21
Gustave Flaubert	3	7
Jean Giono	3	6
Michel Tournier	3	11
Romain Gary	2	9
Maurice Leblanc	2	33
Claude Farrère	2	5
Michel Deon	2	3
Georges Perec	2	10
Jean Echenoz	2	16
Simone De Beauvoir	2	7
Georges Bernanos	2	8
Nathalie Sarraute	2	6
Joris Karl Huysmans	2	10
Antoine De Saint Exupery	2	4
André Malraux	2	3
Jules Renard	2	4
Pierre Michon	2	6
Sainte-Beuve	2	2

Delphine De Vigan	2	4
Marguerite Duras	2	8
Laurent Gaude	2	4
Marguerite Yourcenar	2	9
Maylis De Kerangal	2	5
Andre Gide	2	10
Claude Simon	2	11
Joseph Kessel	2	5
Jules Vallès	2	4
Alain Robbe-Grillet	2	6
Henri Barbusse	2	2
Théophile Gautier	2	6
Mathias Enard	1	4
Raymond Queneau	1	4
Marcel Pagnol	1	7
Frederic Beigbeder	1	5
Antoine Volodine	1	9
Emile Moselly	1	2
Mme Tarbe Des Sablons	1	1
Octave Mirbeau	1	5
Henri Murger	1	1
Patrick Deville	1	1
Alphonse De Chateaubriant	1	2
Philippe Grimbert	1	1
Jean-Christophe Rufin	1	3
Jules Verne	1	2
Paul Fils Feval	1	6
Jean Marie Gustave Le Clézio	1	4
Marie NDiaye	1	1
Félicien Marceau	1	1
Eugène Ionesco	1	1
Henry Bauchau	1	4
Henry Bordeaux	1	3
Paul Feval	1	59
Olivier Rolin	1	8
Alain Fournier	1	1
Leon Bloy	1	3
Raymond Radiguet	1	3
Pierre Combescot	1	1

Magali	1	4
Julien Gracq	1	3
Louis Ferdinand Celine	1	4
Benjamin Constant	1	1
Georges Rodenbach	1	2
Philippe Claudel	1	4
Comtesse De Segur	1	11
Michel Leiris	1	1
Marie Cardinal	1	1
Cesarie Farrenc	1	1
Emmanuel Carrere	1	6
Erik Orsenna	1	1
Albert Cohen	1	1
Marc Dugain	1	1
Christian Gailly	1	10
Jean Philippe Toussaint	1	2
Samuel Beckett	1	4
Pascal Laine	1	1
Delly	1	95
Marie Darrieussecq	1	8
Charles Victor Arlincourt	1	7
Michel Houellebecq	1	5
Louis Pergaud	1	3
Eugene Fromentin	1	1
Sylvie Germain	1	8
Alice Zeniter	1	1
Pierre Loti	1	31
Jacques Laurent	1	6
Didier Daeninckx	1	21
Paul Verlaine	1	4
Roger Vercel	1	1
Georges Duhamel	1	5
Pascal Quignard	1	12
Marie Nimier	1	1
Marcel Proust	1	7
Emile Ajar - Gary Romain	1	4
Marcel Schwob	1	2
Pierre Benoit	1	2
Maurice Genevoix	1	2

Jean Rolin	1	4
Alfred De Musset	1	2
Alfred De Vigny	1	3
Nina Bouraoui	1	7
Paul Nizan	1	3
Mathieu Riboulet	1	4
Francois Mauriac	1	1
Louise Michel	1	1
Comte De Lautreamont	1	1
Marguerite Audoux	1	6
Catherine Cusset	1	2
Annie Ernaux	1	8
Alice Ferney	1	6
Chatrian Erckmann	1	20
Paule Constant	1	1

Le canon des œuvres

Date	Auteur	Titre
1811	Chateaubriand-François-Rene-de	Oeuvres-completes
1816	Constant-Benjamin	Adolphe
1821	Arlincourt-Charles-Victor	Le-Solitaire
1829	Hugo-Victor	Le-dernier-jour-d-un-condamne
1830	Stendhal	Le-Rouge-et-le-noir
1831	Hugo-Victor	Notre-Dame-de-Paris
1831	Hugo-Victor	Notre-Dame-de-Paris
1832	Vigny-Alfred-de	Stello
1834	Sainte-Beuve	Volupte
1834	Hugo-Victor	Claude-Gueux
1834	Sainte-Beuve	Volupte
1835	Gautier-Theophile	Mademoiselle-de-Maupin
1836	Musset-Alfred-de	La-Confession-d-un-enfant-du-siecle
1839	Stendhal	La-Chartreuse-de-Parme
1840	Sand-George	Pauline
1842	Balzac-Honore-de	La-Femme-de-trente-ans
1842	Balzac-Honore-de	Beatrix
1842	Balzac-Honore-de	Le-Contrat-de-mariage
1843	Dumas-Alexandre	Le-Corricolo
1843	Balzac-Honore-de	Le-Cure-de-Tours
1843	Balzac-Honore-de	Le-Pere-Goriot
1843	Balzac-Honore-de	Illusions-perdues
1843	Balzac-Honore-de	Pierrette
1843	Balzac-Honore-de	Eugenie-Grandet
1844	Balzac-Honore-de	Le-Colonel-Chabert
1844	Balzac-Honore-de	Splendeurs-et-miseres-des-courtisanes
1844	Chateaubriand-François-Rene-de	Vie-de-Rance
1844	Balzac-Honore-de	Le-Lys-dans-la-vallee
1844	Dumas-Alexandre	Les-Trois-Mousquetaires
1845	Dumas-Alexandre	Le-Comte-de-Monte-Cristo
1845	Balzac-Honore-de	Un-debut-dans-la-vie
1845	Balzac-Honore-de	Adieu
1845	Dumas-Alexandre	Vingt-ans-apres
1845	Dumas-Alexandre	La-Reine-Margot
1846	Sand-George	La-Mare-au-Diable

1846	Balzac-Honore-de	Physiologie-du-mariage
1846	Mme-Tarbe-Des-Sablons	Isabelle
1846	Balzac-Honore-de	Louis-Lambert
1846	Balzac-Honore-de	Une-tenebreuse-affaire
1846	Balzac-Honore-de	Le-Cure-de-village
1846	Dumas-Alexandre	La-Dame-de-Monsoreau
1846	Balzac-Honore-de	Le-Medecin-de-campagne
1848	Dumas-Alexandre	La-dame-aux-camelias
1849	Chateaubriand-François-Rene-de	Memoires-d-Outre-Tombe
1849	Chateaubriand-François-Rene-de	Memoires-d-Outre-Tombe
1849	Chateaubriand-François-Rene-de	Memoires-d-Outre-Tombe
1849	Sand-George	La-petite-Fadette
1849	Chateaubriand-François-Rene-de	Memoires-d-Outre-Tombe
1849	Chateaubriand-François-Rene-de	Memoires-d-Outre-Tombe
1850	Dumas-Alexandre	Le-Vicomte-de-Bragelonne
1854	Barbey-d-Aurevilly-Jules	L-ensorcelee
1855	Balzac-Honore-de	Les-Paysans
1855	Stendhal	Chroniques-italiennes
1857	Feval-Paul	Le-Bossu
1857	Flaubert-Gustave	Madame-Bovary
1858	Segur-comtesse-de	Les-Malheurs-de-Sophie
1860	Erckmann-Chatrian	Contes-fantastiques
1860	Goncourt-Edmond-et-Jules-de	Charles-Demailly
1862	Hugo-Victor	Les-Miserables
1862	Flaubert-Gustave	Salammbô
1863	Verne-Jules	Cinq-Semaines-en-ballon
1863	Fromentin-Eugene	Dominique
1863	Farrenc-Cesarie	La-jalousie
1863	Gautier-Theophile	Le-capitaine-Fracasse
1864	Verne-Jules	Voyage-au-centre-de-la-Terre
1864	Barbey-d-Aurevilly-Jules	Le-Chevalier-des-Touches
1864	Goncourt-Edmond-et-Jules-de	Germinie-Lacerteux
1864	Goncourt-Edmond-de-Goncourt-Jules-de	Renee-Mauperin
1866	Hugo-Victor	Les-travailleurs-de-la-mer
1868	Zola-Emile	Therese-Raquin
1869	Hugo-Victor	L'homme-qui-rit
1869	Lautreamont-comte-de	Les-chants-de-Maldoror
1869	Murger-Henri	Scenes-de-la-vie-de-Boheme
1869	Flaubert-Gustave	L'education-sentimentale

1870	Zola-Emile	La-fortune-des-Rougon
1870	Verne-Jules	Vingt-mille-lieues-sous-les-mers
1872	Zola-Emile	La-curee
1873	Zola-Emile	Le-ventre-de-Paris
1873	Daudet-Alphonse	Contes-du-lundi
1874	Flaubert-Gustave	La-tentation-de-saint-Antoine
1874	Hugo-Victor	Quatrevingt-Treize
1874	Barbey-d-Aurevilly-Jules	Les-Diaboliques
1875	Daudet-Alphonse	Jack
1875	Zola-Emile	La-faute-de-l-abbe-Mouret
1876	Zola-Emile	Son-excellence-Eugene-Rougon
1876	Sand-George	Horace
1877	Goncourt-Edmond-et-Jules-de	La-fille-Elisa
1877	Flaubert-Gustave	Trois contes
1878	Verne-Jules	Un-capitaine-de-quinze-ans
1878	Zola-Emile	Madeleine-Ferat
1879	Huysmans-Joris-Karl	Les-soeurs-Vatard
1879	Daudet-Alphonse	Lettres-de-mon-moulin
1880	Zola-Emile	Nana
1881	Valles-Jules	Le-Bachelier
1881	Valles-Jules	L-Enfant
1881	Flaubert-Gustave	Bouvard-et-Pecuchet
1882	Guy-de-Maupassant	Mademoiselle-Fifi
1883	Zola-Emile	Au-bonheur-des-dames
1883	Guy-de-Maupassant	Une-vie
1884	Zola-Emile	La-joie-de-vivre
1884	Guy-de-Maupassant	Clair-de-lune
1884	Daudet-Alphonse	Sapho
1884	Huysmans-Joris-Karl	a-rebours
1884	Guy-de-Maupassant	Au-soleil
1885	Zola-Emile	Germinal
1885	Verne-Jules	Mathias-Sandorf
1885	Guy-de-Maupassant	Bel-ami
1885	Guy-de-Maupassant	Contes-du-jour-et-de-la-nuit
1886	Michel-Louise	Memoires
1887	Zola-Emile	La-terre
1887	Bloy-Leon	Le-Desespere
1887	Guy-de-Maupassant	Mont-Oriol
1887	Loti-Pierre	Madame-Chrysantheme

1888	Zola-Emile	Le-reve
1888	Guy-de-Maupassant	Pierre-et-Jean
1889	Guy-de-Maupassant	La-main-gauche
1890	Guy-de-Maupassant	Notre-coeur
1890	Zola-Emile	La-bete-humaine
1891	Zola-Emile	L-argent
1892	Zola-Emile	La-debacle
1892	Rodenbach-Georges	Bruges-la-Morte
1893	Zola-Emile	Le-docteur-Pascal
1893	Verlaine-Paul	Mes-prisons
1893	Renard-Jules	Coquecigrues
1893	Leblanc-Maurice	Une-femme
1894	Renard-Jules	Poil-de-carotte
1894	Stendhal	Lucien-Leuwen
1895	Leblanc-Maurice	Contes
1896	Schwob-Marcel	Vies-imaginaires
1900	Mirbeau-Octave	Le-journal-d-une-femme-de-chambre
1904	Rolland-Romain	Jean-Christophe
1904	Rolland-Romain	Jean-Christophe
1904	Rolland-Romain	Jean-Christophe
1905	Rolland-Romain	Jean-Christophe
1905	Farrere-Claude	Les-civilises
1907	Moselly-Emile	Terres-lorraines
1908	Rolland-Romain	Jean-Christophe
1908	Barbusse-Henri	L-enfer
1908	Rolland-Romain	Jean-Christophe
1908	Rolland-Romain	Jean-Christophe
1909	Farrere-Claude	La-Bataille
1910	Audoux-Marguerite	Marie-Claire
1910	Rolland-Romain	Jean-Christophe
1911	Rolland-Romain	Jean-Christophe
1912	Pergaud-Louis	La-Guerre-des-boutons
1912	Rolland-Romain	Jean-Christophe
1913	Bordeaux-Henry	La-Maison
1913	Alain-Fournier	Le-grand-Meaulnes
1914	Gide-Andre	Les-Caves-du-Vatican
1916	Barbusse-Henri	Le-Feu
1919	Benoit-Pierre	L-Atlantide
1922	Aragon-Louis	Les-aventures-de-Telemaque

1922	Feval-Paul-fils	Le-Bossu
1923	Radiguet-Raymond	Le-diable-au-corps
1923	Chateaubriant-Alphonse-de	La-Briere
1925	Genevoix-Maurice	Raboliot
1925	Proust-Marcel	Albertine-disparue
1926	Bernanos-Georges	Sous-le-soleil-de-Satan
1926	Colette	La-Fin-de-Cheri
1926	Kessel-Joseph	Les-Captifs
1927	Colette	Le-Ble-en-herbe
1927	Gide-Andre	Les-Faux-monnayeurs
1928	Colette	La-naissance-du-jour
1929	Bernanos-Georges	La-joie
1930	Giono-Jean	Regain
1930	Malraux-Andre	La-voie-royale
1930	Colette	Sido
1931	Saint-Exupery-Antoine-de	Vol-de-nuit
1932	Celine-Louis-Ferdinand	Voyage-au-bout-de-la-nuit
1933	Duhamel-Georges	Le-notaire-du-Havre
1933	Malraux-Andre	La-Condition-Humaine
1934	Vercel-Roger	Capitaine-Conan
1934	Colette	Les-Vrilles-de-la-vigne
1935	Delly	Contes
1936	Aragon-Louis	Les-Beaux-Quartiers
1936	Giono-Jean	Les-vraies-richesses
1938	Nizan-Paul	La-Conspiration
1939	Saint-Exupery-Antoine-de	Terre-des-hommes
1939	Sarraute-Nathalie	Tropismes
1939	Leiris-Michel	L-Age-d-homme
1942	Camus-Albert	L-etranger
1944	Camus-Albert	Le-premier-homme
1947	Camus-Albert	La-peste
1950	Duras-Marguerite	Un-barrage-contre-le-Pacifique
1951	Gracq-Julien	Le-rivage-des-Syrtes
1951	Giono-Jean	Le-Hussard-sur-le-toit
1951	Beckett-Samuel	Molloy
1953	Robbe-Grillet-Alain	Les-Gommes
1954	Cohen-Albert	Le-livre-de-ma-mere
1954	Beauvoir-Simone-de	Les-Mandarins
1954	Beauvoir-Simone-de	Les-Mandarins

1956	Gary-Romain	Les-Racines-du-Ciel
1958	Kessel-Joseph	Le-lion
1959	Queneau-Raymond	Zazie-Dans-Le-Metro
1960	Pagnol-Marcel	Le-Temps-des-Secrets
1960	Gary-Romain	La-promesse-de-l-aube
1960	Simon-Claude	La-Route-des-Flandres
1965	Perec-Georges	Les-choses
1967	Tournier-Michel	Vendredi-ou-les-limbes-du-Pacifique
1967	Simon-Claude	Histoire
1968	Yourcenar-Marguerite	L-Oeuvre-Au-Noir
1969	Magali	La-prisonniere
1969	Mauriac-Francois	Un-adolescent-d-autrefois
1969	Marceau-Felicien	Creezy
1970	Deon-Michel	Les-poneys-sauvages
1970	Tournier-Michel	Le-Roi-des-Aulnes
1971	Laurent-Jacques	Les-betises
1971	Tournier-Michel	Vendredi-ou-la-vie-sauvage
1972	Aragon-Louis	Aurelien
1972	Modiano-Patrick	Les-boulevards-de-ceinture
1973	Deon-Michel	Un-Taxi-mauve
1973	Ionesco-Eugene	Le-solitaire
1974	Laine-Pascal	La-dentelliere
1974	Yourcenar-Marguerite	Souvenirs-pieux
1975	Cardinal-Marie	Les-mots-pour-le-dire
1975	Ajar-Emile-Gary-Romain	La-Vie-Devant-Soi
1977	Modiano-Patrick	Livret-de-famille
1978	Modiano-Patrick	Rue-des-boutiques-obscures
1978	Perec-Georges	La-Vie-mode-d-emploi
1980	Le-Clezio-Jean-Marie-Gustave	Desert
1983	Sarraute-Nathalie	Enfance
1983	Ernaux-Annie	La-place
1983	Echenoz-Jean	Cherokee
1984	Duras-Marguerite	L-Amant
1988	Orsenna-Erik	L-exposition-coloniale
1991	Quignard-Pascal	Tous-Les-Matins-Du-Monde
1991	Michon-Pierre	Rimbaud-le-fils
1991	Combescot-Pierre	Les-filles-du-Calvaire
1994	Camus-Albert	Le-premier-homme
1994	Rolin-Olivier	Port-Soudan

1995	Carrere-Emmanuel	La-classe-de-neige
1996	Rolin-Jean	L-Organisation
1998	Constant-Paule	Confidence-pour-confidence
1998	Dugain-Marc	La-chambre-des-officiers
1998	Daeninckx-Didier	Cannibale
1999	Volodine-Antoine	Des-anges-mineurs
1999	Echenoz-Jean	Je-m-en-vais
2001	Robbe-Grillet-Alain	La-reprise
2001	Gaude-Laurent	Cris
2001	Gailly-Christian	Un-soir-au-club
2001	Rufin-Jean-Christophe	Rouge-Bresil
2002	Gaude-Laurent	La-Mort-du-roi-Tsongor
2003	Beigbeder-Frederic	Windows-on-the-World
2004	Grimbert-Philippe	Un-secret
2005	Germain-Sylvie	Magnus
2005	Bouraoui-Nina	Mes-mauvaises-pensees
2006	Nimier-Marie	La-Reine-du-silence
2007	de-Vigan-Delphine	No-et-moi
2007	Claudel-Philippe	Le-rapport-de-Brodeck
2008	Enard-Mathias	Zone
2008	Bauchau-Henry	Le-Boulevard-peripherique
2008	Cusset-Catherine	Un-brillant-avenir
2009	Toussaint-Jean-Philippe	La-Verite-sur-Marie
2009	Vigan-Delphine-de	Les-heures-souterraines
2009	Michon-Pierre	Les-Onze
2009	N-Diaye-Marie	Trois-femmes-puissantes
2010	Kerangal-Maylis-de	Naissance-d-un-pont
2010	Houellebecq-Michel	La-Carte-et-le-territoire
2012	Riboulet-Mathieu	Les-oeuvres-de-misericorde
2012	Deville-Patrick	Peste-et-cholera
2013	Zeniter-Alice	Sombre-dimanche
2013	Ferney-Alice	Cherchez-La-Femme
2013	Darrieussecq-Marie	Il-faut-beaucoup-aimer-les-hommes
2014	Kerangal-Maylis-de	Reparer-les-vivants

Scores des motifs stylistiques

	precision	recall	f1-score	support	accuracy
canon	0.868	0.886	0.878	230	
non-canon	0.928	0.912	0.916	361	
full dataset				592	**0.902**
macro-average	0.896	0.898	0.896	591	
weighted average	0.902	0.898	0.902	591	

TABLE 3 – Résultats de l'évaluation du modèle en validation croisée pour les motifs

Projet ANR « Chapitres (XIXe-XXe siècles) »

Je remercie toute l'équipe de l'ANR Chapitres et en particulier Aude Leblond pour m'avoir permis de réaliser mes recherches sur leur corpus.

Équipe de l'ANR Chapitres :

— Aude Leblond
— Claire Colin
— Thomas Conrad
— Marianne Reboul
— Raphaël Baroni
— Alexandre Gefen
— Camille Koskas
— Virginie Tahar
— Michel Bernard
— Alain Schaffner
— Jérémy Naïm
— Ugo Dionne
— Dimitri Garncarzyk
— Anaïs Goudmand
— Victoire Feuillebois

Lien : https://chapitres.hypotheses.org/.

Publication principale : « Pratique et poétiques du chapitre du XIXe au XXIe siècle » [4].

4. Claire Colin, Thomas Conrad et Aude Leblond, *Pratiques et poétiques du chapitre du XIXe au XXIe siècle*, Rennes, 2017 (Interférences).

Bibliographie

Algee-Hewitt (Mark), *Canon/Archive : studies in quantitative formalism from the Stanford Literary Lab*, dir. Franco Moretti, New York, 2017.

Algee-Hewitt (Mark), Fredner (Erik) et Walser (Hannah), « The Novel as Data », dans *The Cambridge Companion to the Novel*, dir. EricEditor Bulson, 2018 (Cambridge Companions to Literature), p. 189-216, doi : 10.1017/9781316659694.013.

Algee-Hewitt (Mark) et McGurl (M.), « Between Canon and Corpus : Six Perspectives on 20th- Century Novels », Pamphlets of the Stanford Literary Lab–8 (2015), url : https://litlab.stanford.edu/LiteraryLabPamphlet8.pdf..

Allison (Sarah), Algee-Hewitt (Mark), R. Heuser (Ryan), Jockers (Matthew), Moretti (Franco) et Witmore (Michael), « Quantitative Formalism : an Experiment », Pamphlets of the Stanford Literary Lab–1 (2011), url : https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf..

Bernard (Michel), « Goncourt 2020 : mais qu'a-t-il de plus que les autres ? », *Humanités numériques*–4 (1er déc. 2021), Number : 4 Publisher : Humanistica, doi : 10.4000/revuehn.2297.

Bloom (Harold), *The Western canon : the books and school of the ages*, OCLC : 624578000, New York, 1994, url : http://site.ebrary.com/id/10879075 (visité le 13/04/2022).

Bourdieu (Pierre), *Les règles de l'art genèse et structure du champ littéraire*, Paris, 1992.

Braudel (Fernand), *La Méditerranée et le monde méditerranéen à l'époque de Philippe II. Fernand Braudel*, Dixième édition, Malakoff, 2017.

Brottrager (Judith), Stahl (Annina) et Arslan (Arda), « Predicting Canonization : Comparing Canonization Scores Based on Text-Extrinsic and -Intrinsic Features », dans *Proceedings of the Conference on Computational Humanities Research CHR2021*, dir. Maud Ehrmann, *et al.*, Amsterdam, the Netherlands, 2021 (CEUR Workshop Proceedings), t. 2989, p. 195-205, url : http://ceur-ws.org/Vol-2989/#short_paper21.

Burrows (J.), « 'Delta' : a Measure of Stylistic Difference and a Guide to Likely Authorship », *Literary and Linguistic Computing*, 17–3 (1er sept. 2002), p. 267-287, doi : 10.1093/llc/17.3.267.

Cafiero (Florian) et Camps (Jean-Baptiste), « Why Molière most likely did write his plays », *Science Advances*, 5–11 (nov. 2019), eaax5489, doi : 10.1126/sciadv.aax5489.

— « 'Psyché' as a Rosetta Stone ? Assessing Collaborative Authorship in the French 17th Century Theatre », dans *Proceedings of the Conference on Computational Humanities Research CHR2021*, dir. Maud Ehrmann, *et al.*, Amsterdam, the Netherlands, 2021 (CEUR Workshop Proceedings), t. 2989, p. 377-391, url : http://ceur-ws.org/Vol-2989/#long_paper51.

Camps (Jean-Baptiste), *SUPERvised STYLometry (SuperStyl)*, version ... 2021, doi : ....

Casanova (Pascale), *La république mondiale des lettres*, Édition revue et corrigé e, Paris, 2008 (Points Série essais, 607).

Chervel (André), *Histoire de l'agrégation : contribution à l'histoire de la culture scolaire*, Paris, 1993 (Collection "Le Sens de l'histoire").

Chevrel (Yves), « Les Lettres modernes et la formation des professeurs de français : » *L'information littéraire*, Vol. 55–3 (1er sept. 2003), p. 3-10, doi : 10.3917/inli.553.0003.

Colin (Claire), Conrad (Thomas) et Leblond (Aude), *Pratiques et poétiques du chapitre du XIXe au XXIe siècle*, Rennes, 2017 (Interférences).

Compagnon (Antoine), *La Troisième République des lettres, de Flaubert à Proust*, Paris, 1983.

— « Sainte-Beuve and the Canon », *MLN*, 110–5 (1995), Publisher : Johns Hopkins University Press, p. 1188-1199, url : http://www.jstor.org/stable/3251396 (visité le 08/03/2022).

— *Le démon de la théorie : littérature et sens commun*, Paris, 1998 (La couleur des idées).

Cortes (Corinna) et Vapnik (Vladimir), « Support-vector networks », *Machine Learning*, 20–3 (sept. 1995), p. 273-297, doi : 10.1007/BF00994018.

Cranenburgh (Andreas van) et Bod (Rens), « A Data-Oriented Model of Literary Language », dans *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, Valencia, Spain, 2017, p. 1228-1238, doi : 10.18653/v1/E17-1115.

Cranenburgh (Andreas van), Dalen-Oskam (Karina van) et Zundert (Joris van), « Vector space explorations of literary language », *Language Resources and Evaluation*, 53–4 (déc. 2019), p. 625-650, doi : 10.1007/s10579-018-09442-4.

Cranenburgh (Andreas van) et Koolen (Corina), « Identifying Literary Texts with Bigrams », dans *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, Denver, Colorado, USA, 2015, p. 58-67, doi : 10.3115/v1/W15-0707.

ENGLISH (James F), *Economy of Prestige : Prizes, Awards, and the Circulation of Cultural Value*. OCLC : 1058248419, Cambridge, 2009.

FASSIN (Éric), « La chaire et le canon. Les intellectuels, la politique et l'Université aux États-Unis », *Annales. Histoire, Sciences Sociales*, 48–2 (1993), p. 265-301, DOI : 10.3406/ahess.1993.279133.

FELPERIN (Howard), *Beyond deconstruction : the uses and abuses of literary theory*, Oxford, 1985.

GONZÁLEZ (José Eduardo), JACOBSON (Elliott), GARCÍA (Laura García) et KUJMAN (Leonardo Brandolini), « Measuring Canonicity : Graduate Reading Lists in Departments of Hispanic Studies », *Journal of Cultural Analytics* (, 19 mars 2021), DOI : 10.22148/001c.21599.

GUILLORY (John), *Cultural capital : the problem of literary canon formation*, [Nachdr.], Paperback ed. 1994, Chicago, 1998.

HARDER (Marie-Pierre), *(Dé)construire le canon Introduction*, 2013, URL : http://www.crlc.paris-sorbonne.fr/pdf_revue/revue4/1_INTRO_Harder.pdf (visité le 13/04/2022).

HARRIS (Wendell V.), « Canonicity », *PMLA/Publications of the Modern Language Association of America*, 106–1 (janv. 1991), p. 110-121, DOI : 10.2307/462827.

HERRMANN (J. Berenike), DALEN-OSKAM (Karina van) et SCHÖCH (Christof), « Revisiting Style, a Key Concept in Literary Studies », *Journal of Literary Theory*, 9 (2015), p. 25-52.

JEY (Martine), *La littérature au lycée : invention d'une discipline (1880-1925)*, Paris, 1998 (Recherches textuelles, no 3).

— « Le canon aux agrégations du XIX e siècle », *Revue d'histoire littéraire de la France*, 114–1 (2014), p. 143, DOI : 10.3917/rhlf.141.0143.

L'idée de littérature dans l'enseignement, dir. Martine Jey et Laetitia Perret, Paris, 2019 (Rencontres, Série Littérature des XXe et XXIe siècles, 380. 36).

JIPA (Dragoş), *La canonisation littéraire et l'avènement de la culture de masse : la collection Les grands écrivains français (1887-1913)*, ISBN : 9783631672419 Series Number : Volume 302 Series : Publications universitaires européennes, thèse de doct., Frankfurt, Peter Lang Academic research, 2016.

JOCKERS (Matthew L.), *Macroanalysis : Digital Methods and Literary History*, 1re éd., 2013, DOI : 10.5406/illinois/9780252037528.001.0001.

JOCKERS (Matthew L.) et MIMNO (David), « Significant themes in 19th-century literature », *Poetics*, 41–6 (déc. 2013), p. 750-769, DOI : 10.1016/j.poetic.2013.08.005.

KESTEMONT (Mike), « Function Words in Authorship Attribution. From Black Magic to Theory ? », dans *Proceedings of the 3rd Workshop on Computational Linguistics for*

Literature (CLFL), Gothenburg, Sweden, 2014, p. 59-66, DOI : 10.3115/v1/W14-0908.

KOOLEN (Corina), DALEN-OSKAM (Karina van), CRANENBURGH (Andreas van) et NAGELHOUT (Erica), « Literary quality in the eye of the Dutch reader : The National Reader Survey », *Poetics*, 79 (avr. 2020), p. 101439, DOI : 10.1016/j.poetic.2020.101439.

LEGALLOIS (Dominique), CHARNOIS (Thierry) et LARJAVAARA (Meri), « The balance between quantitative and qualitative literary stylistics : how the method of "motifs" can help », dans *The Grammar of Genres and Styles*, dir. Dominique Legallois, Thierry Charnois et Meri Larjavaara, 2018, p. 164-193, DOI : 10.1515/9783110595864-008.

LENTRICCHIA (Frank) et MCLAUGHLIN (Thomas), *Critical terms for literary study*, OCLC : 813244781, Chicago [Ill., 2012, URL : http://www.credoreference.com/book/uchicagols (visité le 22/11/2021).

LIDDLE (Dallas), « Could Fiction Have an Information History ? Statistical Probability and the Rise of the Novel », *Journal of Cultural Analytics* (, 2019), DOI : 10.22148/16.033.

LONG (Hoyt) et SO (Richard Jean), « Literary Pattern Recognition : Modernism between Close Reading and Machine Learning », *Critical Inquiry*, 42–2 (janv. 2016), p. 235-267, DOI : 10.1086/684353.

MOLINIÉ (Georges), « Style et littérarité », *Littératures classiques*, 28–1 (1996), p. 69-74, DOI : 10.3406/licla.1996.2519.

Qu'est-ce que le style ? actes du colloque international, dir. Georges Molinié et Pierre Alain Cahné, 1re éd, Paris, 1994 (Linguistique nouvelle).

MORETTI (Franco), « The Slaughterhouse of Literature », *Modern Language Quarterly*, 61–1 (1er mars 2000), p. 207-228, DOI : 10.1215/00267929-61-1-207.

— « Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850) », *Critical Inquiry*, 36–1 (2009), p. 134-158, DOI : 10.1086/606125.

— « "Operationalizing" : or, the function of measurement in modern literary theory », Pamphlets of the Stanford Literary Lab–6 (2013), URL : https://litlab.stanford.edu/LiteraryLabPamphlet6.pdf..

— *Distant reading*, Country : GB Contient des textes déjà publiés. Includes bibliographical references and index. Vendor-supplied metadata., London, 2013.

— « Literature, Measured », Pamphlets of the Stanford Literary Lab–12 (2016), URL : https://litlab.stanford.edu/LiteraryLabPamphlet12.pdf..

PEDREGOSA (F.), VAROQUAUX (G.), GRAMFORT (A.), MICHEL (V.), THIRION (B.), GRISEL (O.), BLONDEL (M.), PRETTENHOFER (P.), WEISS (R.), DUBOURG (V.), *et al.*, « Scikit-learn : Machine Learning in Python », *Journal of Machine Learning Research*, 12 (2011), p. 2825-2830.

PHILIPPE (Gilles), *Pourquoi le style change-t-il ?*, Bruxelles, 2021.

La langue littéraire : une histoire de la prose en France de Gustave Flaubert à Claude Simon, dir. Gilles Philippe et Julien Piat, Paris, 2009.

Piper (Andrew), « Fictionality », *Journal of Cultural Analytics* (, 20 déc. 2016), DOI : 10.22148/16.011.

— « Think Small : On Literary Modeling », *PMLA/Publications of the Modern Language Association of America*, 132–3 (mai 2017), p. 651-658, DOI : 10.1632/pmla.2017.132.3.651.

— *Enumerations : data and literary study*, Chicago ; London, 2018.

Piper (Andrew), Bagga (Sunyam), Monteiro (Laura), Yang (Andrew), Labrosse (Marie) et Liu (Yu Lu), « Detecting Narrativity Across Long Time Scales », dans *Proceedings of the Conference on Computational Humanities Research CHR2021*, dir. Maud Ehrmann, *et al.*, Amsterdam, the Netherlands, 2021 (CEUR Workshop Proceedings), t. 2989, p. 319-332, URL : http://ceur-ws.org/Vol-2989/#long_paper49.

Plecháč (Petr), « Relative contributions of Shakespeare and Fletcher in *Henry VIII* : An analysis based on most frequent words and most frequent rhythmic patterns », *Digital Scholarship in the Humanities*, 36–2 (29 sept. 2021), p. 430-438, DOI : 10.1093/llc/fqaa032.

Pollock (Griselda), *Differencing the canon : feminist desire and the writing of art's histories*, London ; New York, 1999 (Re visions).

Ribeiro (Marco Tulio), Singh (Sameer) et Guestrin (Carlos), « "Why Should I Trust You ?" : Explaining the Predictions of Any Classifier », dans *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA, 2016, p. 1135-1144, DOI : 10.1145/2939672.2939778.

Schmitt (Michel P.) et Viala (Alain), « Les cotes aux concours », *Littératures classiques*, 19–1 (1993), p. 281-291, DOI : 10.3406/licla.1993.1753.

Seminck (Olga) et Poibeau (Thierry), *The Evolution of the Idiolect over the Lifetime : A Quantitative and Qualitative Study on French 19th Century Literature*, 2022.

So (Richard Jean), « "All Models Are Wrong" », *PMLA/Publications of the Modern Language Association of America*, 132–3 (mai 2017), p. 668-673, DOI : 10.1632/pmla.2017.132.3.668.

Thiesse (Anne-Marie), *La fabrique de l'écrivain national : entre littérature et politique*, Paris, 2019 (Bibliothèque des histoires).

Tolonen (Mikko), Hill (Mark J.), Ijaz (Ali Zeeshan), Vaara (Ville) et Lahti (Leo), « Examining the Early Modern Canon : The English Short Title Catalogue and Large-Scale Patterns of Cultural Production », dans *Data Visualization in Enlightenment Literature and Culture*, dir. Ileana Baird, Cham, 2021, p. 63-119, DOI : 10.1007/978-3-030-54913-8_3.

UNDERWOOD (Ted), *Tedunderwood/Horizon : Data And Code To Support Distant Horizons*, 24 mars 2018, DOI : 10.5281/ZENODO.1206317.

— « Why Literary Time is Measured in Minutes », *ELH*, 85–2 (2018), p. 341-365, DOI : 10.1353/elh.2018.0013.

— *Distant horizons : digital evidence and literary change*, Chicago, 2019.

— « Machine Learning and Human Perspective », *PMLA/Publications of the Modern Language Association of America*, 135–1 (janv. 2020), p. 92-109, DOI : 10.1632/pmla.2020.135.1.92.

UNDERWOOD (Ted) et SELLERS (Jordan), « The "Longue Durée" of Literary Prestige », *Modern Language Quarterly*, 77–3 (sept. 2016), p. 321-344, DOI : 10.1215/00267929-3570634.

UNDERWOOD (Ted), BLACK (Michael L.), AUVIL (Loretta) et CAPITANU (Boris), « Mapping mutable genres in structurally complex volumes », dans *2013 IEEE International Conference on Big Data*, Silicon Valley, CA, USA, 2013, p. 95-103, DOI : 10.1109/BigData.2013.6691676.

VERBOORD (Marc), « Classification of authors by literary prestige », *Poetics*, 31–3 (juin 2003), p. 259-281, DOI : 10.1016/S0304-422X(03)00037-8.

VIALA (Alain), « Qu'est-ce qu'un classique ? », *Littératures classiques*, 19–1 (1993), p. 11-31, DOI : 10.3406/licla.1993.1737.

YU (B.), « An evaluation of text classification methods for literary study », *Literary and Linguistic Computing*, 23–3 (5 sept. 2008), p. 327-343, DOI : 10.1093/llc/fqn015.

Table des figures

Liste des tableaux