CA Journal of Cultural Analytics

Barré, Jean, et al. "Operationalizing Canonicity: A Quantitative Study of French 19th and 20th Century Literature." *Journal of Cultural Analytics*, Oct. 2023, https://doi.org/10.22148/001c.88113.

ARTICLE

Operationalizing Canonicity: A Quantitative Study of French 19th and 20th Century Literature

Jean Barré[1], Thierry Poibeau[1], Jean-Baptiste Camps[2]

[1] École Normale Supérieure - PSL, [2] École Nationale des Chartes

Keywords: Literary history, Canon studies, Computational Literary Studies, French Literature, machine learning, distant reading, natural language processing, text mining, stylometry

https://doi.org/10.22148/001c.88113

Journal of Cultural Analytics

This article delves into the literary canon, a concept shaped by social biases and influenced by successive receptions. The canonization process is a multifaceted phenomenon, emerging from the intricate interplay of sociological, economic, and political factors. Our objective is to detect the underlying textual dynamics that grant certain works exceptional longevity while jeopardizing the transmission of the majority. Drawing on various criteria, we present an operational framework for defining the French literary canon, centered on its contemporary reception and emphasizing the role of institutions, particularly schools, in its formation. Leveraging natural language processing and machine learning techniques, we unveil an intrinsic norm inherent to the literary canon. Through statistical modeling, we achieve predictive outcomes with accuracy ranging from 70% to 74%, contingent on the chosen scale of canonicity. We believe that these findings detect what Charles Altieri calls a "cultural grammar", referring to the idea that canonical works in literature serve as foundational texts that shape the norms, values, and conventions of a particular cultural tradition. We posit that this linguistic norm arises from biased latent selection mechanisms linked to the role of the educational system in the canon-formation process.

1. Introduction

In 1895, the French literary critic Lanson posed a pivotal question: "How are the choices made regarding which works and names endure in immortality?" A case in point is Stendhal, who now holds a distinguished position within the French literary canon. Stendhal, however, only rose to literary prominence long after his demise, which raises questions about the mechanisms that contributed to his canonization as an author. What factors led to the preservation of his novels within the annals of literary history, and how do they differ from the abyss of what Cohen calls the "*Great Unread*", comprising works consigned to literary oblivion?

This matter has long captivated sociocultural research. Investigations into the mechanisms governing the attribution of literary significance have notably centered on the background in which the works were conceived, as well as the sociological path taken by the authors during the canonization process. As exemplified by Bourdieu, the value assigned to an author or a novel emerges as a collective endeavor involving an array of agents and institutions within the literary realm. These encompass critics, historians, salons, political entities, the educational system, and even editorial marketing strategies, all contributing to the formulation of the work as a literary entity.

The canonization processes that underlie the compilation of texts and authors making up the canon, shape, as outlined by Pollock, a "selective tradition". This intricate trajectory is marked by a succession of biases, encompassing dimensions of gender, race, and social class. The attribution of canonicity to these texts and authors ensures their enduring presence within the literary landscape, imbuing them with a preeminent position within the standards of cultural legitimacy.

The concept of literary canon was initially introduced within the realm of literary studies to denote the collection of texts included in university syllabi and analyzed therein. As showed by scholars such as Felperin, the canon plays a vital role in the pedagogical realm of literature: "The institutional study of literature is inconceivable without a canon. Without a canon, without a corpus or syllabus of exemplary texts, there can be no interpretive community". Consequently, the canon constitutes the foundational body of texts upon which the teaching and research in literature rest. This notion is underscored by Casanova, who asserts that the canon inherently "embodies literary legitimacy itself". Essentially, the canon represents the reference set for what is officially recognized as literature, then used in the evaluation of other works.

Thus, the literary canon is a complex notion to address and the mechanisms behind this temporal filtration are numerous, whether they are linked to cultural and academic policies or to aesthetic and critical criteria. In this article, we want to see what is actually happening at the textual scale, and to map the textual differences between canonized texts and non-canonized texts.

Our study falls within the field of computational literary studies and distant reading (Moretti, "Conjectures on World Literature"). By the large scale quantitative study of literary works, it strives to go beyond the study of the few hundred works making the literary canon. Doing so, it hopes, as theorised by Underwood, to identify important structuring lines of literary history, that traditional approaches can fail to notice. In other terms, we wish to gain insight into what is happening inside the "slaughterhouse" of literature (Moretti, "The Slaughterhouse of Literature").

Our focus is directed towards a reexamination of the texts themselves and their intrinsic content. Our objective is to assess the extent of the filtration process applied to literary works. We hypothesize that there is a particular norm in the textual content of the canonized novels, and that it can be detected quantitatively. The question remains whether the textual attributes we seek to identify signify a causal phenomenon — where texts are selected due to their specific characteristics — or if they instead emerge as products of the canonization process itself, reflecting not inherent selection value but rather the biases intrinsic to the canonization trajectory.

Our research will revolve around the presentation of diverse criteria employed in detecting *canonicity* from the textual content. To comprehend the concept of the literary canon within the context of French literature, we built a contemporary reception-based literary canon, rooted in multiple factors. We then used text mining, natural language processing (NLP), and machine learning techniques to delve into the intricate layers of the literary canon.

2. Literature Review

The literary canon has been the object of many studies in computational literary studies. A first approach was to quantitatively describe the lists that constituted the various canons. Pamphlet 8, "Between Canon and Corpus: Six Perspectives on 20th-century Novels" (Algee-Hewitt and McGurl) from Stanford Literary Lab characterized the literary canon and demonstrated the inherent lack of inclusivity of these lists towards non-Western literatures. A similar approach was adopted by González et al. in their work on Hispanic Studies syllabi in US universities. They studied the diversity of the canon with entropy measures of canonical populations over time. Attempts have been made to characterize the notion of literary canon through the composition of these lists, particularly during the period of their emergence (Tolonen et al.).

Other studies have gone beyond the literary canon and have taken up Bourdieu's binary construction of the literary field, between popularity and prestige. Porter showed that these axes seemed relevant for mapping literary and cultural space, while Verboord classified authors according to their position in the literary field, using this dichotomy. He showed that Institutional Literary Prestige (drawn from academic studies, among other things) was fruitful for classification.

The second approach to understand the literary canon is to measure differences between canonical and non-canonical works in the texts themselves, using natural language processing methods. In this regard, the paper by Algee-Hewitt et al. is very instructive. Their hypothesis was that novels were selected in the canon because they were less redundant. The team measured lexical variety with entropy and found that their hypothesis was confirmed.

Underwood and Sellers devoted an article to the automatic classification of literary prestige based on textual data from poetry. They defined literary prestige as the likelihood of a text being reviewed in specialized literary journals. The main question they asked was: "Is the social boundary between elite taste and the rest of literary production associated with recognizable stylistic differences ?" With simple NLP tools (bags of words) and a predictive algorithm (logistic regression), good results were obtained, on the order of 75% accuracy for the statistical model. They showed that the literary discourse contained in the text is related to the reception of the said text, and that this relationship is statistically robust.

In the wake of these discoveries, numerous studies have addressed the question of literary prestige, focusing on the style of works consecrated by the canon, and its potential difference from other styles. This subject has been particularly addressed in the Netherlands, notably by Koolen et al., who showed that the degree of literariness perceived by humans is quantifiable and can be modeled. van Cranenburgh et al. and van Cranenburgh and Bod explored this perceived literariness using word vectors and obtained interesting results showing that the concept of literariness can be predicted to some extent based on textual features.

The paper by Brottrager, Stahl, and Arslan, proposed a formalization of literary historical reception. They analyzed and compared the relationship between the concept of canonicity based on extrinsic data (i.e. the contexts of the works) and intrinsic features (i.e. their textual content). The results showed a clear lack of correlation between the two methods. As an extension of this research, Brottrager, Stahl, Arslan, et al. evaluated how literary reception as a social process can be linked to textual qualities. They obtained a 78% accuracy in predicting if a text was reviewed in literary periodicals in the English context.

Empirical research on literary prestige is scarce in France, and few experiments have been carried out on French corpora. An exception is a study on the successive selection of works for the Prix Goncourt 2020 (Bernard). However, the results of this study did not show a clear tendency, suggesting that no textual dynamics was at stake in the selection.

The present paper is therefore part of a dynamic research context but one in which investigations on French data are lacking. Our work consists in operationalizing a wider definition of canonicity, in order to better understand this complex phenomenon. Little work has been done to evaluate quantitatively the role of the school system in the canonization process. As we will see in section 3, it is arguably a much stronger route to immortality than reviews in specialized magazines. The first step was to collect relevant metadata to build a French literary canon. In a second step, we modeled canonicity based on textual features using machine learning and natural language processing methods.

3. Determining canonical factors

One of the main tasks of this study was the construction of a literary canon. For this purpose, we enriched our corpus[1] with information about the contemporary reception of the texts and authors. Admittedly, the literary canon is neither monolithic nor temporally stable, and defining it by finite criteria is in itself reductive and neglects the complexity of the phenomenon. Nevertheless, formalizing a complex notion requires making choices to be able

1 See section 5 for the corpus description

to grasp it. One of the main restrictions we imposed was to focus on the contemporary reception of the works, in order to grasp the literary canon that has reached us today.

We sought to focus on elements that have already been discussed and analyzed by literary criticism and studies on this subject. One of the aspects we focused on is the role of institutions in the formation of literary prestige, for as Bourdieu said "It is only post mortem, and after a long process, that the school institution, [...] grants the infallible sign of consecration, namely the canonization of works as classics by including them in school curricula". The school institution constructs its own representation of literature and determines the good use of it, with chronological divisions (periodizations, literary schools, generations), categories (romanticism, naturalism, surrealism), and the development of a canon by a selection of authors. According to Guillory, the process of canon formation within the educational system can be interpreted as a matter of distributing cultural capital in schools, with these established classics being presented as exemplars that communicate a specific aesthetic standard. We focused precisely on this norm, which we aimed to quantitatively identify.

The work by Jey and Perret on the role of the school institution in the constitution of such canonical sets has shown that secondary and higher education have an enormous impact on the formation of the canon (in the making and especially the preservation of this canon) of authors and texts. It thus appeared relevant to approach the literary canon mainly from the perspective of the reception of works by educational institutions. While this approach is acknowledged in the humanities, it is not an exhaustive one, as other factors such as political, economic or sociological criteria also come into play in the constitution of the canon.

We therefore established the following non-exhaustive set of criteria to characterize a literary canon that we then investigated quantitatively.

3.1. The school canon

As we consider the public school system as the place where the literary canon is disseminated and conserved, it seemed important to take into account what is expected of pupils when they leave compulsory schooling, that is to say what constitutes, for the authors of these lists, the minimal literary culture for the construction of citizenship. The work by Jey, gives a detailed description of the construction of a discipline, literature, around texts guaranteeing a certain language and a certain morality, which must be disseminated to educate the masses. She analyzes the process by which works are integrated into school syllabi which is in fact a process of canonization. We therefore took the programs of the secondary school examinations, i.e. the *Brevet* (equivalent to GCSE) and the *Baccalauréat* (high school diploma), from 2000 to 2018 as part of our criteria.

3.2. The academic canon

Lists established for Higher education examinations are also of interest. We retrieved lists and programs of literary and scientific preparatory classes from the *École Normale Supérieure* competitive examination. These lists are established to evaluate and select, on the basis of literary knowledge, candidates who will become future college professors. Schmitt and Viala adopted a similar approach by listing the number of times certain authors were cited in student essays. Since their data were not available, we stuck to the examination lists, from 2008 to 2019. We also retrieved the programs of the competitive examinations for the *agrégation de lettres modernes*, the highest competitive examination for the recruitment of teachers of French as it seemed significant to note which authors and texts were selected to train the national elite of teachers of French. For an overview of the *agrégation* exams, the research by Jey, and by Chervel and by Chevrel was of great help. As these programs did not include many novels, we decided to enlarge the period of reception considered, extending the metadata back to 1950.

3.3. The canon of publishers

Next, we looked at the world of publishing, which is also one of the major actors in the canonization process. The thesis by Jipa on the collection of the "Grands écrivains de France" clearly showed the importance of editorial logics in the construction of a national consensus around a pantheon of authors.

We focused on the *Pléiade* collection, which is a prestigious collection of classic works of French literature. The *Pléiade* editions are highly regarded for their scholarly annotations, introductions, and critical notes that provide valuable insights into the literary and historical context of the works. The publication in the *Pléiade* of an author's complete works is often seen as a mark of recognition and prestige for an author's contributions to literature. It is a major sign that the said author belongs to the literary canon.

For added nuance, we incorporated novels from the "Classical literature" collection by *Garnier-Flammarion*. This collection stands out for its comprehensive critical apparatus accompanying each novel, signifying the work's depth that warrants exploration—both the literary work and the contexts in which they originated. This is relevant in this context as it mirrors a prevalent pedagogical perspective on the literary canon, often cited to uphold the existing literary framework in France and beyond.

3.4. The canon of criticism

To incorporate literary criticism, we looked at literary awards. This aspect of our canon is the least resistant to time, because these awards are strongly influenced by the economic and sociological context of their era, as highlighted by English in his study on cultural value circulation. Despite this limitation, we

aimed to evaluate the impact of these awards on literary trends. Consequently, we compiled lists of French literary awards, ranging from the prestigious *Goncourt* prize to the *Femina* award.

Additionally, we incorporated contemporary research by leveraging the online literary platform "Fabula".[2] If a query concerning an author yielded at least ten results, the author was deemed canonical. This dimension was already present in the corpus metadata, and we opted to retain it.

3.5. The political canon

There are also political implications in the canon formation process. As Viala puts it: the canon "fulfills a function of cultural identification", in other words, canonized texts represent a common base for the cultural construction of a nation. In the context of early 20th-century France, the canon formation embodied, in the words of Thiesse, "the political establishment of the national narrative". With the structuring and centralization of the education system during the Third Republic in France, the canon crystallized and became a political object (Compagnon), particularly concerning novelistic production. Literature in education, coupled with a literary canon, was assigned the role of educating the masses and disseminating national values. Various political reforms of the education system have thus shaped the canon and the methods of teaching literature over time.

To capture this political dimension in our canon, we also took a list of the 150 literary texts selected in 2018 by the *Ministère de l'Éducation Nationale* (French Ministry of Education).[3] Those texts represent what is assumed to be the French literary canon from a political standpoint.

4. Our French literary canon

Constructed with several factors, our literary canon seeks to include a wide variety of actors in the literary field who define, nourish and preserve the literary canon. Our approach to the issue of whether a work belongs to the literary canon or not adopted a twofold granularity: that of the individual novel and that of the author. This allows us, on the one hand, to construct a highly restrictive canon by considering that literary immortality is attributed to a specific text rather than to an individual writer. On the other hand, the figure of the author still holds sway in literary textbooks and various cultural depictions of literature, making it impossible for us to disregard this dimension. This second scale encompasses the entire body of work by an author as canonical, resulting in a much broader canon.

2 https://www.fabula.org/

3 https://gallica.bnf.fr/blog/18012018/150-epub-gallica-selectionnes-par-le-ministere-de-leducation-nationale

Table 1. Number of novels in the corpus for each canonical factor

bac	brevet	sup	prix	gf	gouv
104	51	42	91	117	45

bac =: Baccalauréat
brevet =: Brevet
sup =: Agrégation and ENS examinations
prix =: Literary Awards
gf =: Garnier Flammarion Collection
gouv =: Ministry of Education

To ensure consistency, all the lists were meticulously curated to only include novels that were present within our corpus of texts. This process aimed to align our corpus with the canon that we had constructed. To achieve this alignment, a simple membership test was employed: if the title of a novel appeared in at least one of the established lists, that particular novel was considered as part of the canonical body of work. Similarly, for authors, if an author's name was featured on any of the established lists, all of that author's works within our corpus were deemed to be part of the canon. We generated two binary variables—one for the novel level and the other for the author level—with the options being either *canon* or *non_canon*.

Thus, the number of works in the corpus that are in our canon amounts to 306 items (10% of the corpus), while the number of works whose authors are in our canon is 1173 novels (40%). Table 2 shows the number of novels from our corpus present in each canonical list.[4]

We calculated the cosine similarity between our canonical lists to assess their level of coherence. The heatmap of the results can be seen in Figure 1. It shows that the lists are far from being identical, even if there are certain similarities in the three school-based factors (brevet, bac, sup). It is interesting to note that the *Garnier-Flammarion* list (gf) is also close to the school-based factors, presumably because this collection, which includes a critical apparatus, is designed to be used by the school system. The literary awards list (prix) is very different from the other canonical factors. While the relevance of this list in the canon is debatable, we decided to keep it because it enables us to capture a more contemporary canon than the one captured by our school canon.

See appendix A.1 and A.5 for further details on data availability and construction.

4 For a fuller description of our canonical samples, see the online supplement to this article

Figure 1. Heatmap of the cosine similarity between our canonical lists

5. Corpus

The corpus used in this study is that collected by the project "ANR Chapitres",[5] a corpus of nearly 3000 French literary texts (Leblond). The goal of the research was to evaluate the pace of change in the length of chapters over two centuries of literature. The corpus is structured in XML (eXtended Markup Language) with TEI[6] (Text Encoding Initiative) encoding, to add metadata to the texts. The corpus consists of 2,960 novels, totaling 14,982,817 sentences and 234,175,471 tokens. A significant bias inherent in this corpus lies in its compilation of digitized novels available online. This selection process inherently reflects texts that have been chosen, published, and preserved over time, which, in turn, represents only a fraction of the entire body of written production.

The period concerned extends over two centuries of novelistic production, from the beginning of 19th to the early 21th century, as can be seen in Figure 2. The temporal distribution of novels within the corpus displays a relatively balanced spread, although the latter half of the 19th century stands out, encompassing nearly 40% of the novels. Notably, the 1880s alone contribute almost 10% of the novels. This distribution poses a challenge in terms of potential biases, as there is a risk of magnifying this period's impact in statistical measures.

5 https://chapitres.hypotheses.org/

6 TEI Consortium, eds. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 1.0. TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

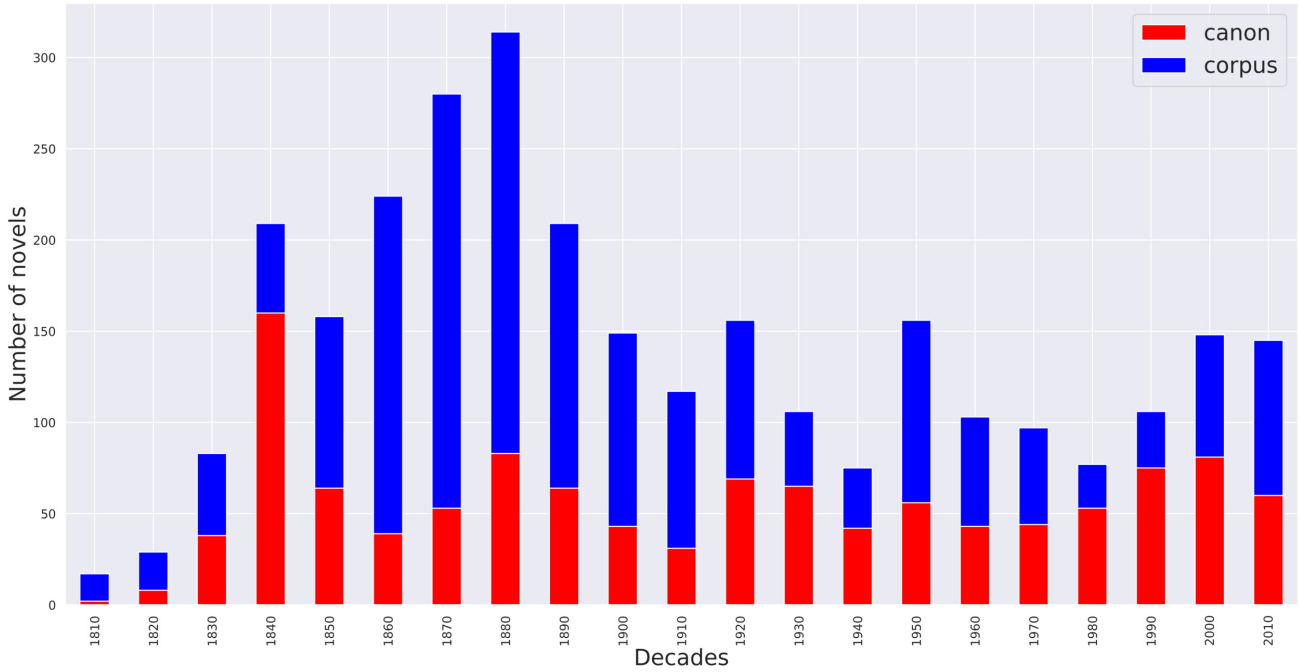

Figure 2. Distribution of the number of novels over time, broken out by canonicity tags, canon at the author scale

The distribution of canonical novels at the author scale appears to be consistently spread across the entire corpus, minimizing the risk of any temporal bias impacting our experiment. There are approximately 50 canonical novels per decade, with the exception of the 1840s where there is a notable increase to over 150 canonical novels. This anomaly can be attributed, in part, to Balzac, as the editions of his 85 novels present in the corpus are predominantly from this particular decade.

We believe that the non-canonical works in the corpus are a good sample of what the archive may have been, not only by their number – they account for nearly 90% of the novels at the novel scale, and 65% at the author scale – but also by the diversity of the sub-genres represented.

The *Chapitres* corpus provides additional information about each text, with approximately two-thirds of them accompanied by details about their sub-genre. Figure 3 illustrates the distribution of the literary canon across various sub-genres within the corpus, encompassing genres from detective novels to travelogues. Notably, there is no overrepresentation of the canon within any specific sub-genre. However, an intriguing observation is the partial or complete absence of canonical works within the sub-genres of sentimental novel and children's literature. This observation appears to align with the notion that these two sub-genres lack the literary recognition associated with, for instance, adventure novels. While the validity of the sub-genre labels can be debated, our focus here lies in the balanced distribution of canonical works among these diverse categories.

See appendix 10 for details about the distribution of the canon at the novel scale.

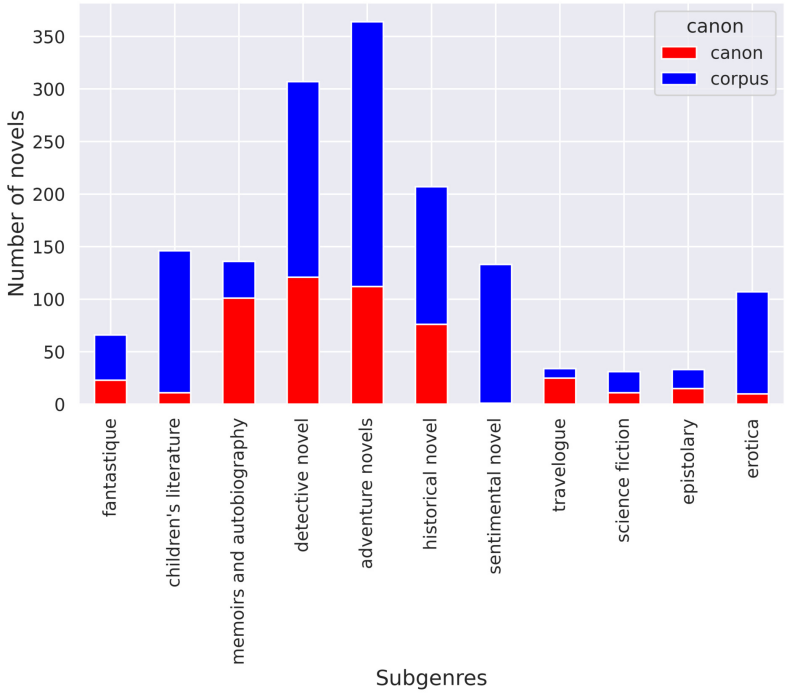

Figure 3. Literary sub-genres in the corpus, broken out by canonicity tags, canon at the author scale

6. Methods

With our text collection and our operable definition of the canon in hand, we started the quantitative analysis. This was based on text features and a classifier, trained to predict canonicity.

6.1. Textual features

In view of the complexity of the phenomenon studied, we wanted to simplify the textual features retained to train the classifier. The classification was therefore based on a bag-of-words model with relative frequencies. Lemmas were used to build n-grams and sequences of both lemmas and POS-tags. We chose two configurations of these patterns, one with the lemmas of content words and POS-tags of the function words, and the other one vice versa as we wished to test how relevant function words were to characterize canonical information. Each type of feature was limited to a bag-of-words of the 1000 most frequent n-grams retrieved from a sample of 200 texts randomly drawn from the corpus.

Our hypothesis was that function words should be very helpful, because they are more related to an unconscious and automatic structural writing (Pennebaker) than less frequent words related to the contents and the themes of the text. van Cranenburgh et al. showed that thematic information does not play a huge role in the literariness of texts, and we extrapolated these results to our case study (the specificity of a text to be canonical or not). This also allowed us to ignore most of the common nouns or proper nouns, which are not relevant to this study. Function words are at the heart of stylometry, notably in authorship attribution (Mosteller and Wallace), and in the study of idiolectal

evolution (Seminck et al.), i.e. the textual signature of a writer. These methods have produced very good results on several authors, from Hildegarde de Bingen (Kestemont), to Shakespeare (Plecháč) or Molière (Cafiero and Camps) and Racine (Gabay). Although the nature of the challenge we encountered may differ, we considered that these techniques were applicable to our inquiry. This is because if there exists a distinct manner in which novels are crafted based on the institutions that shape the literary canon, then using stop words as features may reveal the subconscious indicators of this selection process.

6.2. Prediction

We based our work on the canonical labels defined for each text in the corpus. These were then used as ground truth for our binary classification. Two distinct experiments were conducted for the two canonical scales retained.

The automatic classification of texts is a well studied problem in statistics. One family of models, Support Vector Machines (SVM), is of particular interest here because it obtains good results (Yu) when classifying literary texts, and has the advantage of reducing the risk of over-fitting. In this paper, we used the family of SVMs developed by the Scikit-learn team since 2011 (Pedregosa et al.), and more specifically the SVC estimator.

We ran our model in a basic 5-fold cross-validation set up. The dataset is split into 5 consecutive folds and each fold was used once as a validation while the 4 remaining folds formed the training set. Given the nature of the features used in authorship attribution, we wanted to avoid over-fitting on an author's writing style. To do so, we implemented Scikit-learn's Group strategy. All works by the same author (group) were placed in the same fold; thus, each group will appeared exactly once in the test set across all folds. In this manner the model cannot cheat and recognize the same idiolectal information in both the training and the test sets. See appendix A.3 for the detailed prediction setup, in particular for how we handled the baselines.

Data imbalance was especially challenging at the novel scale, given that our canonical sample represents only 10% of the dataset. Since SVM models are quite sensitive to such imbalanced classes, we re-balanced the classes before implementing the classification by taking the 306 canonical novels and randomly adding 306 non-canonical novels (50% canon, 50% non_canon). We implemented this random selection a hundred times and for each resulting sample the model was run in a 5-fold cross-validation setting. The following results are aggregated from this process.

Table 2. Results of the evaluation of the model, novel scale

	precision	recall	f1-score	support	accuracy
canon	0.728	0.668	0.697	306	
non_canon	0.691	0.748	0.719	306	
full dataset				612	0.708

7. Results

7.1. Results at the novel scale

The model achieved 70.8% accuracy at the novel scale which is better than the baseline, which scored at 51% accuracy. This shows that the SVM is able to separate the two classes based on latent textual reasons. For each class, the metrics are coherent (5% gap between precision and recall). The F1-score for non-canonical works is a little better than for canonical ones.

Surprisingly, the model achieved its best performance using only uni-grams and bi-grams of lemmas as features. This observation resonates with the findings of van Cranenburgh and Koolen, whose research demonstrated the effectiveness of bi-grams in classifying literary texts. Given these outcomes, our strategy based mostly on stopwords distribution and structural information from texts appears to be notably effective. What these findings seem to indicate is that the detected canonical norm operates beneath consciousness.

In Figure 4, we projected the predicted probability of each novel to belong to the literary canon. All of these probabilities are drawn from the 5 test samples of the 5-fold cross-validation, from which we evaluated the generalization performance of the model. The blue circles represent the novels actually classified in our metadata as canonical and the orange crosses represent the non-canonical ones. As can be seen, the SVM has trouble discriminating the two classes, and there are noisy errors throughout the whole period.

The timeframe during which the model demonstrates effective performance is the span from 1850 to 1900, during which the two categories are clearly differentiated. It is worth noting that this might stem from a corpus bias, as the period from 1850 to 1900 is relatively over-represented in the corpus, as depicted in Figure 2. The model has access to a larger volume of training data from this particular timeframe, leading to a specialization in this era. Nonetheless, this over-fitting does not appear to hinder the model's performance, as it continues to perform well.

The red non-linear regression is fitted on the predicted probability for each text to be canonical. This prediction is retrieved from the test set, meaning that the model has seen neither the novel in question nor the writer's other works. There is a huge increase in this probability over time, from 0.2 to 0.6 while it should be around 0.5 since our dataset is balanced. This result is discussed in the next section.

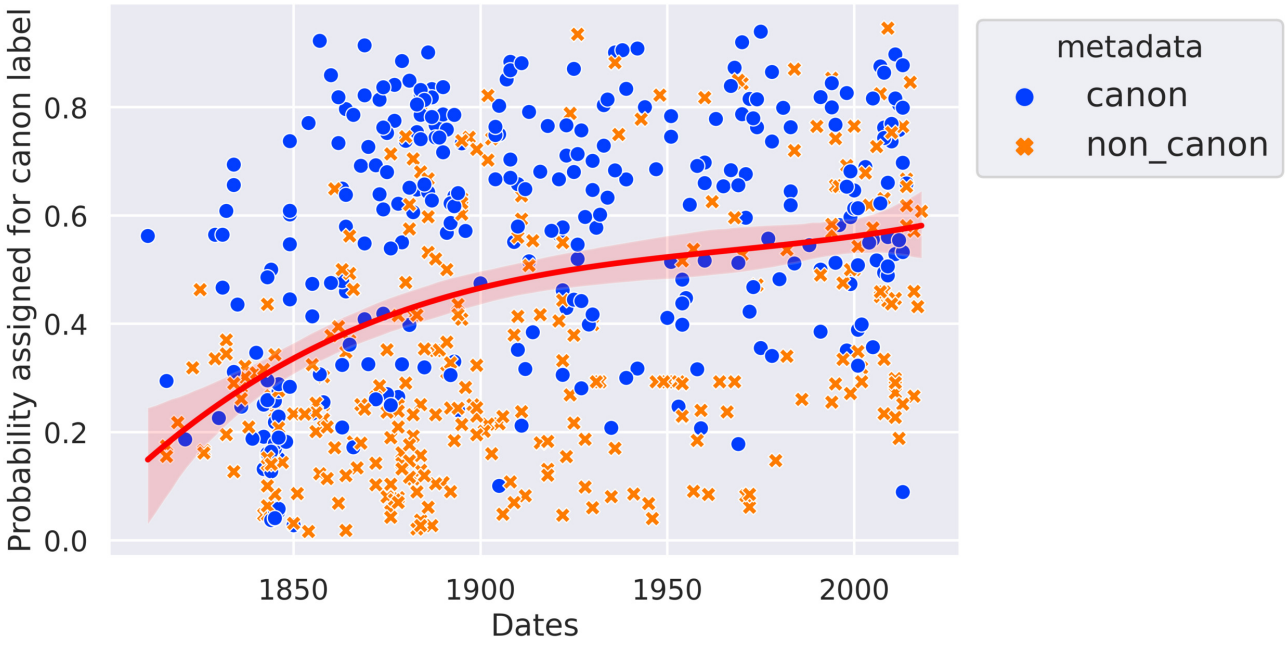

Figure 4. Predicted probability to be canonical, novel scale

To gain deeper insights into the implications of these findings, we focused on two authors and works that are clearly distinguished by the model's assessments. Our model gave Gustave Flaubert's novel *L'Éducation sentimentale* (*Sentimental Education*) an extremely high canonical score (0.914). Published in 1869, the novel offers a profound exploration of the lives of the depicted characters against the backdrop of the political and social upheavals of mid-19th-century France. The exceptional canonicity score of the novel aligns well with its revered status within French literature. Flaubert's skillful interweaving of personal desires, historical context, and enduring themes has firmly secured the novel's place in the literary canon, and the model's recognition of this exemplifies its aptitude in discerning and evaluating the intricate facets that define canonical literature.

In contrast, the novel *Borgia* published in 1906 by Michel Zévaco gets a very low canonical score (0.04). The novel is a historical adventure fiction novel that clearly falls outside the bounds of canonical literature. The novel's focus on political intrigue, scandal, and sensational storytelling aligns with the model's identification of works that deviate from canonical norms.

Table 3. Results of the evaluation of the model, author scale

	precision	recall	f1-score	support	balanced accuracy
canon	0.721	0.645	0.681	1173	
non_canon	0.782	0.836	0.808	1787	
full dataset				2960	**0.741**

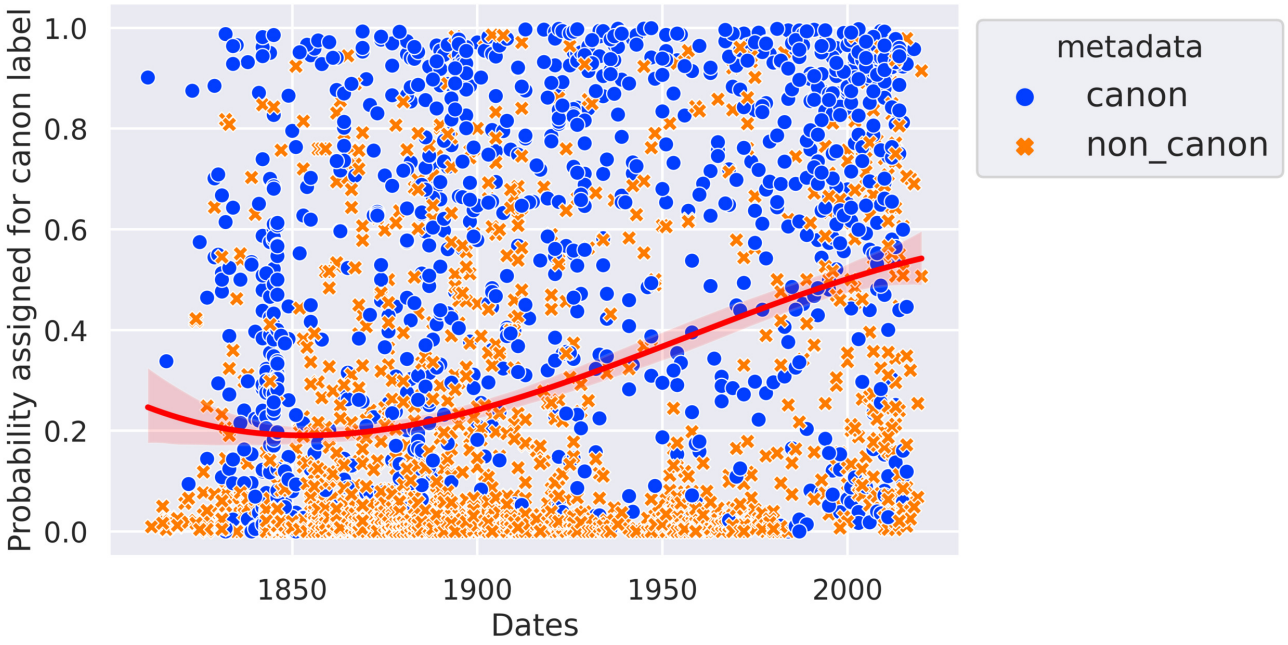

Figure 5. Predicted probability to be canonical, author scale

7.2. Results at the author scale

The model reaches 74.1% balanced accuracy at the author level. The results are better than the performances at the novel scale, but only marginally so. This result is interesting because it might indicate that canonicity can be defined in a very restrictive manner at the novel scale. Although the prominence of individual authors is significant in literary history, it is also quite intuitive that the process of canonization operates within an author's body of work, celebrating a limited selection of novels. We will elaborate on this argument in section 8.

In Figure 5 we projected the predicted probability of each novel to belong to the literary canon, with the canonical metadata at the author scale. The SVM performs better at this scale, i.e. it is a little more confident in its predictions than at the novel scale.

The red non-linear regression projected onto the graph shows an overall trend detected by our model. The probability of belonging to the literary canon increases over time, from 0.2 to 0.5 while it should revolve around 0.4. Technically, this increase is an error. Novels are not more likely to belong to the literary canon because they were published later. It is hard to say whether it is a data related issue or an actual trend in literary history. There is an increase in

the canonical percentage in the last decades of our corpus (from the 1980s), as we can see in figure 2. But it does not explain everything, since the same trend is found at the novel scale, without the increase in our corpus (see appendix A.2 and figure 10). Similar findings are discussed by Underwood, and the assumptions drawn were *i*) that the model failed to produce valid criteria for two centuries of literary production and, *ii*) that books published later have more linguistic signs associated with the standards that govern reception. Our results endorse this prior research and support these hypotheses. This trend is not solely linked to the distribution of the canon over time, but rather seems to be connected to a form of convergence of the overall novelistic production towards the canonical norm. We will attempt to provide further analysis in the discussion in section 9.

In the period just before 1850, we observed a significant increase in misclassifications, particularly regarding canonical novels receiving unexpectedly low canonical scores. Upon closer examination, we noticed a noteworthy pattern where certain prolific writers, including Eugene Sue, Alexandre Dumas (the elder), and George Sand —well-known figures in French literature— were inadequately predicted by our model. While they are acknowledged figures, they may not be as firmly canonized as some others, considering the popular and serialized nature of Sue and Dumas's works, which deviate from the traditional elitist canon. Additionally, gender bias might be influencing the model's assessment, given the relatively fewer women represented in the canon. It is important to note that the model's scores are based on patterns identified within the corpus, and its performance might be influenced by the availability and distribution of data. The model's inability to establish a valid norm for 200 years of history may also reveal a shift within this norm, which could potentially be explained by several factors. One possible explanation is the evolution of language and writing styles over time. As societal norms and linguistic conventions change across centuries, the model may struggle to capture a consistent norm that spans such a wide timeframe.

It is also noticeable that a large amount of the model's errors for non-canonical works are found between the 1980s to the present. The model loses confidence and many novels fall in between. This could be due to an attrition of the canonical standard, which has become challenging to discern since the 1980s.

See appendix A.4 for information about our additional results, in particular on the idiolectal bias.

7.3. Discriminant features analysis

One of the fundamental benefits of machine learning for the field of literary studies lies in the ability to delve into the inferences made by models, shedding light on the intricate mechanisms that drive their predictions. We retrieved in figure 6 the 40 most discriminating features for the model. The coefficients derived from the predictive model offer intriguing insights into the factors that contribute to the classification of works as either canonical or non-canonical.

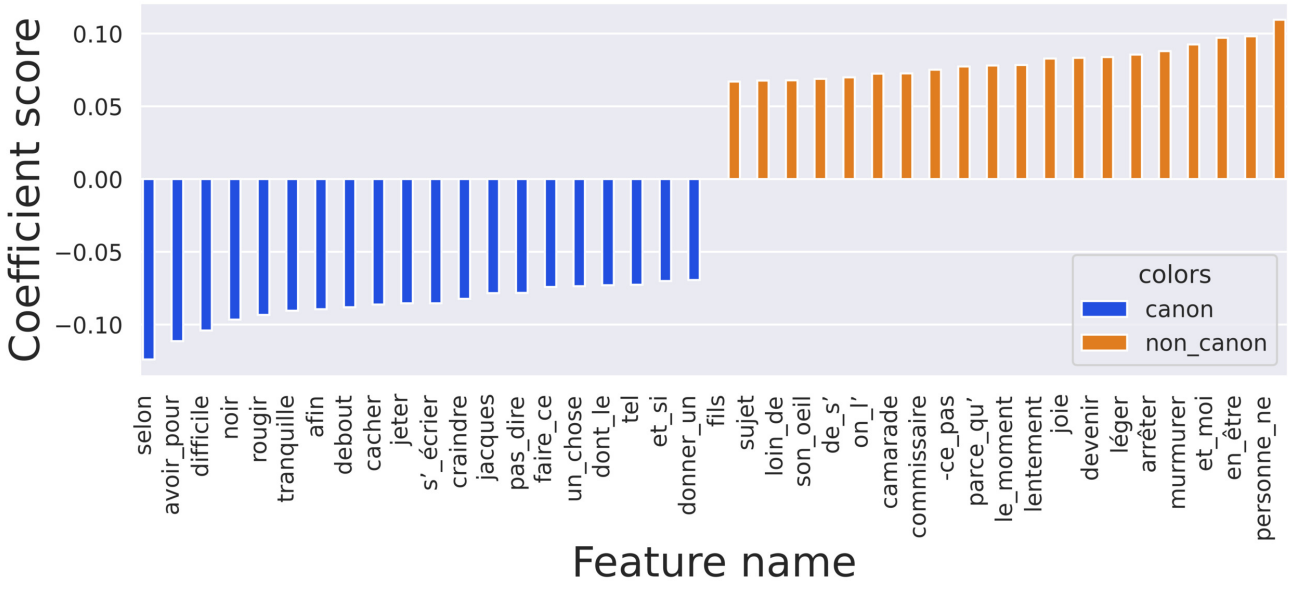

Figure 6. Top 40 discriminant coefficients for the model, canon at the author scale

Examining the elements associated with non-canonical labeling, a distinct pattern emerges. Certain ngrams such as “personne ne” (nobody), “en être” (to be in it), and “et moi” (and me) appear frequently. These phrases, while seemingly innocuous, often characterize colloquial language or informal dialogue. Their prevalence may reflect a tendency towards more mundane or everyday narratives. Similarly, words such as “murmurer” (to murmur), “arrêter” (to stop), and “devenir” (to become) hint at simpler action-driven narratives, often prevalent in genres like adventure or detective fiction. Notably, specific subgenre affiliations can also be deduced from the coefficients. The presence of words such as “commissaire” (detective) in non-canonical labeling might be indicative of works associated with crime or detective intrigue, genres that may be deemed less canonical due to their distinct narrative priorities. In contrast, canonical labeling features words such as “jacques” (a proper name) or “fils” (son) which may allude to more character-driven narratives.

On the other hand, examining the elements contributing to canonical labeling, a different linguistic and thematic spectrum comes into focus. Phrases such as “donner un” (to give a), “et si” (and if), “tel” (such) or “avoir pour” (have for) project a level of linguistic sophistication. These constructions often involve greater syntactical complexity (with more auxiliaries), potentially indicating a propensity for intricate, nuanced narrative structures. Similarly, terms such as “dont le” (of which the), and “faire ce” (to do this) suggest attention to detail and precision in language use.

However, it is important to proceed with caution in generalizing these patterns. The literary landscape of the 19th and 20th centuries was incredibly diverse, encompassing a myriad of styles, themes, and subgenres. While these coefficients offer intriguing insights, the complexity of literature often defies simplistic categorizations.

8. Canonicity at the novel scale

We showed in 7.1 the model's ability to detect the canonical norm with nearly the same performance at both the author and novel scales. This suggests that certain works within an author's oeuvre might align more closely with the established norms and criteria of the literary canon, while others might deviate or be less congruent with those norms. It demonstrates that the process of canon formation is not solely constrained at the level of individual authors, but it also operates within the body of work produced by a single author. This phenomenon could be attributed to various factors, such as shifts in an author's creative intent, experimentation with different narrative techniques, or a response to evolving literary trends.

Focusing on specific authors, we conducted a more targeted experiment, to gain a deeper understanding of what was at play at the author scale. We computed Principal Component Analysis (PCA) thanks to the Python library Prince (Halford). See appendix A.4 for further details on the method. We present in this section the visualization of experiments conducted on the novels of Colette, Victor Hugo and Guy de Maupassant.

8.1. Colette

Figure 7 shows the PCA of the writings of Colette, a famous early 20th century writer. Two elements are highlighted in this graph, on the one hand in orange the non-canonical novels of Colette, and on the other hand in blue the works considered as canonical. The latter form a rather distinct group within Colette's literary production. The PCA positions the canonical novels within a shared region of the graph, indicating a noticeable level of similarity among these works. It is worth noting that all five canonical novels were composed between 1926 and 1934. This temporal alignment might offer an explanation for their clustering, as it corresponds to a distinct literary phase in the author's career. Far from this group is the series of *Claudine*, that were very popular novels which she published under her husband's name. These novels made the popularity of the author at the beginning of her career, but did not correspond to the selection criteria of the canon. It was only later that Colette's identity as a writer was firmly established (Ladimer), and that her works gained prestige. The novel *Sido* is a fictionalized memoir that delves into Colette's relationship with her mother. Its placement within the canonical norm, and its departure from the Claudine series, reflects Colette's transformation as a writer. The novel presents a more reflective and introspective side of Colette's writing, as

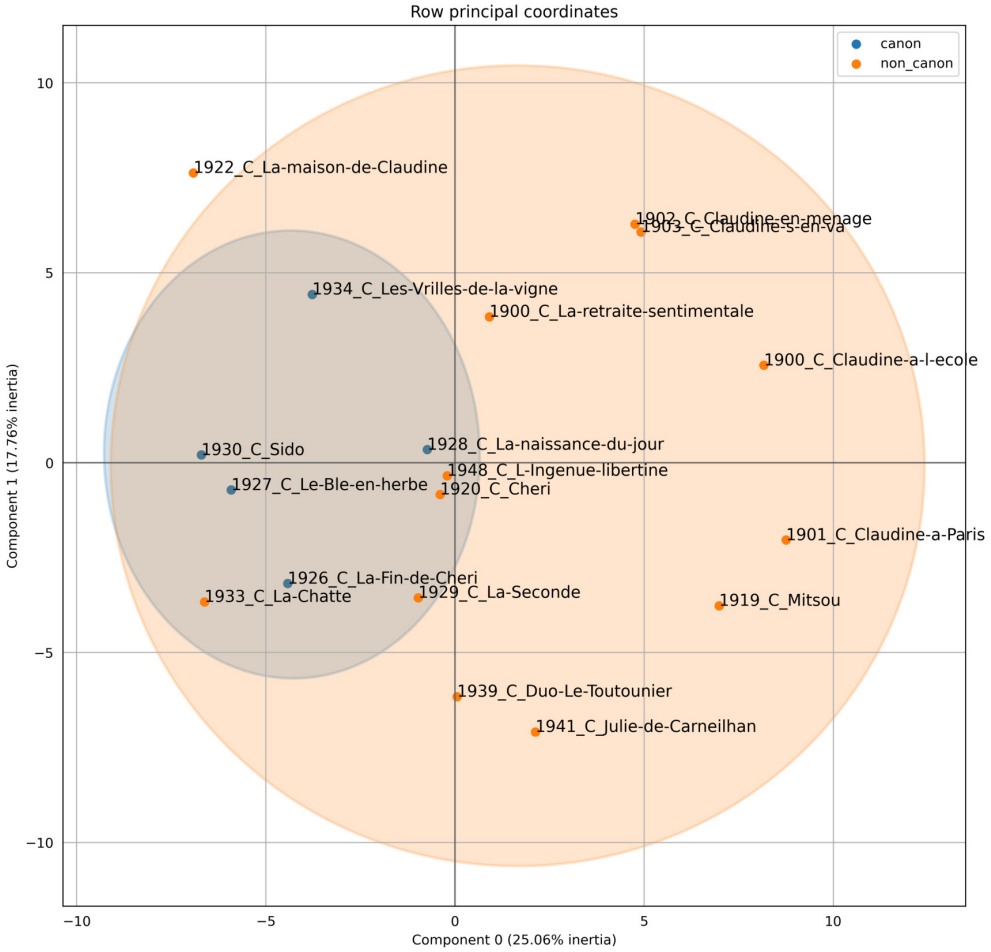

Figure 7. Canonical selectivity in Colette, canonical works in blue and noncanonical ones in orange

she contemplates her personal history, Note, however, that a late novel such as *Julie de Carneilhan*, published in 1941, is far from our canonical specificity, so the PCA does not only detect some chronolectal aspects of Colette's work.

8.2. Victor Hugo

Victor Hugo is one of the most famous and canonical French authors. Not all his writings are equally canonized, however, and some of his novels tend to be forgotten. This is the case for *Han d'Islande*, an early novel by the young Victor Hugo, and for *Le Rhin*, a travel guide with stories about the Rhine river, published in 1842. The three volumes of *Le Rhin* present in our corpus are unsurprisingly very close. Once again, the PCA detects the signature of the author's chronolect, roughly describing two writing periods of Hugo, the first around the 1830s and the second during the author's later period (see figure 8). The two *non canonized* novels are at the margin of the idiolectal signature of Victor Hugo, and stand out from the canonical selection.

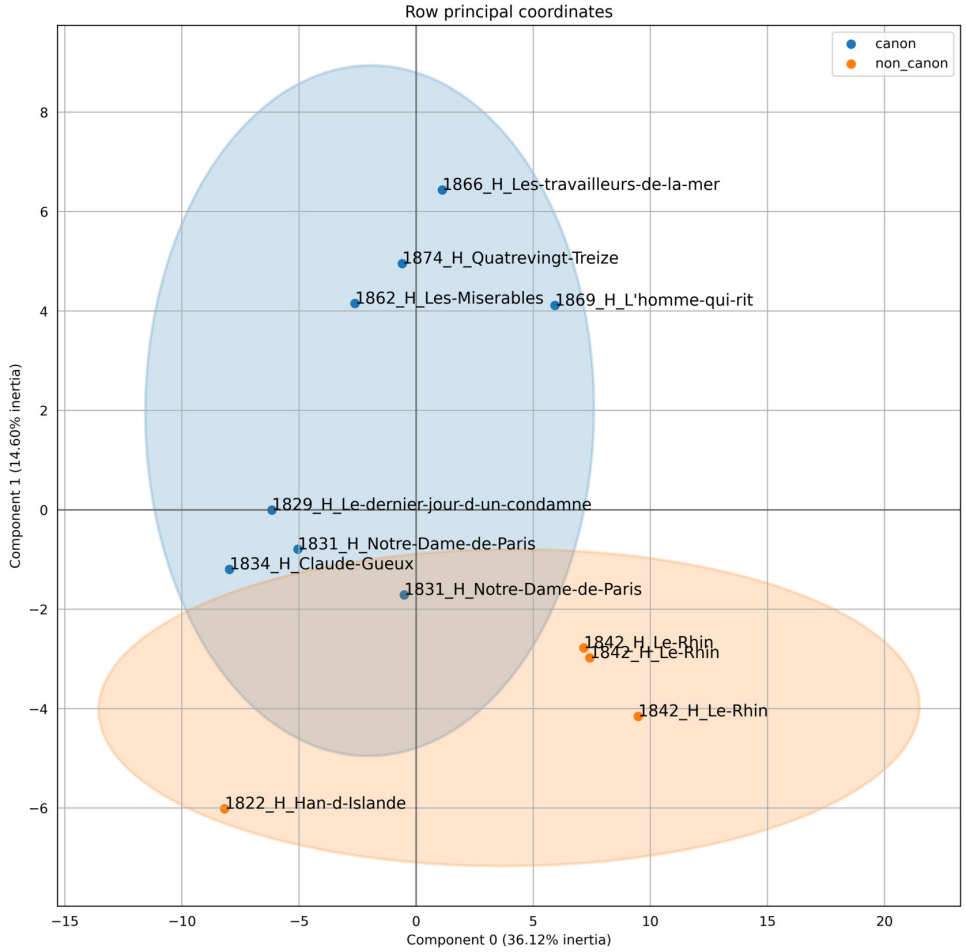

Figure 8. Canonical selectivity in Victor Hugo, canonical works in blue and non-canonical ones in orange

8.3. Guy de Maupassant

It is important to note that this experiment does not work for all our authors, as evidenced by the example of Guy de Maupassant's works, shown in Figure 9: He was a very productive author, and the PCA visualization fails to separate canonical from non-canonical texts.

The two categories of works, canonical and non-canonical, overlap. Critics and particularly the academic institution have elevated this author's novels to such a degree that the distinction between his canonized works and the others has blurred, as if the selective filter had embraced the entirety of his writing style, regardless of specific works.

Hence, the linguistic norm identified across numerous novels by our statistical model appears to gain strength from our additional experiments. This canonical norm is not solely contingent on an author's unique linguistic or temporal characteristics. The PCA experiments demonstrate the sifting of a specific type of content within an author's literary production, discerning between content enshrined in collective memory and content relegated to literary oblivion.

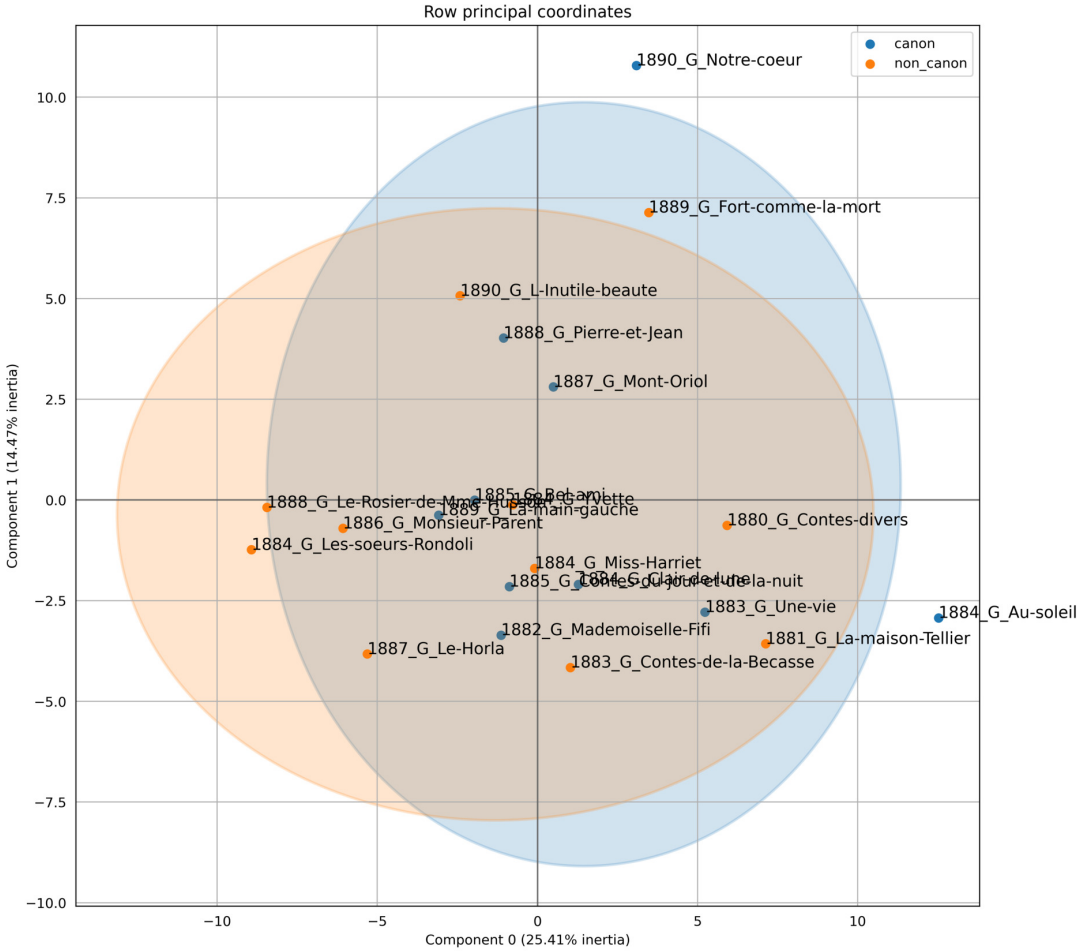

Figure 9. Canonical selectivity in Guy de Maupassant, canonical works in blue and non-canonical ones in orange

9. Discussion

The canon is a complex and multifaceted entity, simultaneously normative in the sense that it only includes a limited number of authors or novels, and dynamic, in that it reflects the constant evolution of the literary field and thus the evolution of literary reception criteria. The canon-makers (the educational institution and, to a lesser extent, the critics) nurture and expand the canon with the passage of time and include works that appear most aligned with a certain conception of literature.

The results we obtained are not particularly surprising, in the sense that the canon is inherently normative, implicitly establishing the rules of "good" literature. Altieri assigns the role of a "cultural grammar" to the canon. This concept refers to a set of linguistic and cultural norms that define the acceptable forms of expression, themes, and ideas within a given society. Just as grammar in language dictates the rules for constructing sentences, this "cultural grammar" dictates the norms for constructing literature that is deemed canonical. The identification of common linguistic features and structures within canonical works suggests that these works adhere to a specific set of rules, much like a grammatical framework, which goes beyond mere stylistic choices.

This may seem a non-intuitive way to view the canon, but it appears to be a fruitful approach to interpret our findings. We think that this norm is not prescriptive, and that it can indeed serve as a touchstone that artists and writers engage with, challenge, and respond to. This perspective emphasizes the significance of viewing the canonical tradition as a dynamic and evolving phenomenon, which continues to influence the creation and interpretation of contemporary works.

As we saw in section 3 with Bourdieu, the intricate mechanisms of canon-formation are inherently tied to the school's function in society, constructing its own representation of literature and generating, as Guillory puts it, "distinct forms of linguistic knowledge". Our approach based on an extensive analysis of the textual content of novels sought to unveil the subtle dynamics that underlie the canon, recognizing its significance as a "cultural grammar" that shapes both the creation and interpretation of literature, while acknowledging that these intricate mechanisms are inherently tied to the school's function in society.

10. Conclusion

In conclusion, this study has introduced a practical definition of the literary canon, validated through quantitative experiments. By establishing criteria rooted in historical evidence, we have delineated the contours of the literary canon within contemporary reception. Drawing on prior research, we acknowledged the educational institution as one of the most influential canon-makers. Leveraging a substantial corpus of novels and harnessing quantitative machine learning techniques and natural language processing, we conceptualized the notion of the literary canon through distant reading. A key contribution of this research has been the identification of a shared linguistic norm among canonical novels, coupled with the development of a statistical model capable of predicting the canonicity of a text with 70% to 74% accuracy.

The objective of this analysis was to augment the conventional viewpoint that often regards the canon as arbitrary, influenced by politics, ideology, or randomness. Our focus on the textual content of works aimed to imbue this definition with a formal and internal dimension, shedding light on latent selection mechanisms within the canon-formation. Indeed, these mechanisms gradually shape what is considered as prestigious literature, influenced by economic, sociological, and political dynamics. The amalgamation of these influences may steer these processes to sift through texts that adhere to specific norms established within the literary realm, thus perpetuating a replication of the literary canon over time. In essence, the canonization processes establish a framework that molds distinct literary forms. We believe that these findings might reflect what Charles Altieri calls a "cultural grammar", referring to the idea that canonical works in literature and culture serve as foundational texts that shape the norms, values, and conventions of a particular cultural and artistic tradition.

This work opens up numerous avenues for further research. Our methodology revolved around quantitatively capturing the linguistic variables that underpin the societal phenomenon of canonization. This task was particularly intricate as it entailed predicting events that transpired during the reception phase—post-writing, that is. Given these complexities, we chose to employ a straightforward *bag-of-words* approach, adopting a consistent canon and streamlined metadata. The primary objective was to test this hypothesis within the realm of French literature. Subsequent investigations are necessary to comprehensively grasp the nuances embedded in the literary canon.

A future approach would involve obtaining metadata chronologically, as reception evolves. Further possibilities encompass dissecting the canon through various literary field agents such as editions, textbooks, school, academic prestige, and literary journals. Additionally, incorporating advanced algorithmic techniques in natural language processing, such as word or paragraph vectors, topic embeddings, and transformers, could enhance the analysis of more intricate textual attributes.

Acknowledgments

Jean Barré's PhD is supported by the EUR (Ecole Universitaire de Recherche) Translitteræ (programme "Investissements d'avenir" ANR-10- IDEX-0001-02 PSL and ANR-17-EURE-0025). This research was conducted as part of a Master's thesis in digital humanities at the École nationale des Chartes. We would like to extend our gratitude to the entire teaching and administrative team, without whom this article would not have been possible.

Thierry Poibeau is supported in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-16-IDEX-0003 (I-Site Future, programme "Cité des dames, créatrices dans la cité"). This work also benefitted from the IRN (International Research Network) Cyclades (Corpora and Computational Linguistics for Digital Humanities).

Lastly, the authors also wish to thank the anonymous reviewers whose comments have helped us construct a stronger paper.

Submitted: April 03, 2023 EDT, Accepted: May 10, 2023 EDT

REFERENCES

Algee-Hewitt, Mark, et al. *Canon/Archive. Large-Scale Dynamics in the Literary Field*. no. 11, 2016, https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf.

Algee-Hewitt, Mark, and M. McGurl. *Between Canon and Corpus: Six Perspectives on 20th- Century Novels*. no. 8, 2015, https://litlab.stanford.edu/LiteraryLabPamphlet8.pdf.

Altieri, Charles. "An Idea and Ideal of a Literary Canon." *Critical Inquiry*, vol. 10, no. 1, Sept. 1983, pp. 37–60, https://doi.org/10.1086/448236.

Barré, Jean. *Replication Data for: "Operationalizing Canonicity."* Harvard Dataverse, 2023, https://doi.org/10.7910/DVN/GQQKWK.

Bernard, Michel. "Goncourt 2020 : Mais Qu'a-t-Il de plus Que Les Autres ?" *Humanités Numériques*, no. 4, Dec. 2021, https://doi.org/10.4000/revuehn.2297.

Bourdieu, Pierre. *Les règles de l'art: Genèse et structure du champ littéraire*. Éditions du Seuil, 1992.

Brottrager, Judith, Annina Stahl, Arda Arslan, et al. "Modeling and Predicting Literary Reception." *Journal of Computational Literary Studies*, vol. 1, no. 1, 2022, https://doi.org/10.48694/JCLS.95.

---. "Predicting Canonization: Comparing Canonization Scores Based on Text-Extrinsic and -Intrinsic Features." *Proceedings of the Conference on Computational Humanities Research CHR2021*, edited by Maud Ehrmann et al., vol. 2989, CEUR, 2021, pp. 195–205, http://ceur-ws.org/Vol-2989/#short_paper21.

Cafiero, Florian, and Jean-Baptiste Camps. "Why Molière Most Likely Did Write His Plays." *Science Advances*, vol. 5, no. 11, Nov. 2019, p. eaax5489, https://doi.org/10.1126/sciadv.aax5489.

Casanova, Pascale. *La république mondiale des lettres*. Éditions du Seuil, 2008.

Chervel, André. *Histoire de l'agrégation: Contribution à l'histoire de la culture scolaire*. Institut national de recherche pédagogique : Editions Kimé, 1993.

Chevrel, Yves. "Les Lettres modernes et la formation des professeurs de français." *L'information littéraire*, vol. Vol. 55, no. 3, Sept. 2003, pp. 3–10, https://doi.org/10.3917/inli.553.0003.

Cohen, Margaret. *The Sentimental Education of the Novel*. Princeton University Press, 2002.

Compagnon, Antoine. *La troisième république des lettres, de flaubert à proust*. Editions du Seuil, 1983.

English, James F. *Economy of Prestige: Prizes, Awards, and the Circulation of Cultural Value.* Harvard University Press, 2009.

Felperin, Howard. *Beyond Deconstruction: The Uses and Abuses of Literary Theory*. Clarendon press, 1985.

Gabay, Simon. "Beyond Idiolectometry? On Racine's Stylometric Signature." *Proceedings of the Conference on Computational Humanities Research CHR2021*, edited by Maud Ehrmann et al., vol. 2989, CEUR, 2021, pp. 359–76, http://ceur-ws.org/Vol-2989/#long_paper39.

González, José Eduardo, et al. "Measuring Canonicity: Graduate Reading Lists in Departments of Hispanic Studies." *Journal of Cultural Analytics*, vol. 6, no. 1, Mar. 2021, https://doi.org/10.22148/001c.21599.

Guillory, John. "Canon." *Critical Terms for Literary Study*, edited by Frank Lentricchia and Thomas McLauahlin, 2nd ed., The University of Chicago Press, 1995, pp. 233–49, https://press.uchicago.edu/ucp/books/book/chicago/C/bo3627086.html.

---. *Cultural Capital: The Problem of Literary Canon Formation*. Paperback, Univ. of Chicago Press, 1998.

Halford, Max. *Prince*. https://github.com/MaxHalford/prince.

Jey, Martine. *La littérature au lycée: Invention d'une discipline (1880-1925)*. Klincksieck, 1998.

---. "Le canon aux agrégations du XIXe siècle." *Revue d'histoire littéraire de la France*, vol. 114, no. 1, 2014, p. 143, https://doi.org/10.3917/rhlf.141.0143.

Jey, Martine, and Laetitia Perret, editors. *L'idée de littérature dans l'enseignement*. Classiques Garnier, 2019.

Jipa, Dragoș, editor. *La canonisation littéraire et l'avènement de la culture de masse: La collection les grands écrivains français (1887-1913)*. Peter Lang Verlag, 2016, https://doi.org/10.3726/978-3-653-06804-7.

Kestemont, Mike. "Function Words in Authorship Attribution. From Black Magic to Theory?" *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, 2014, pp. 59–66, https://doi.org/10.3115/v1/w14-0908.

Koolen, Corina, et al. "Literary Quality in the Eye of the Dutch Reader: The National Reader Survey." *Poetics*, vol. 79, Apr. 2020, p. 101439, https://doi.org/10.1016/j.poetic.2020.101439.

Ladimer, Bethany. *Colette, Beauvoir, and Duras: Age and Women Writers*. University Press of Florida, 1999.

Lagarde, André, et al. *XXe siècle: Les grands auteurs français: Anthologie et histoire littéraire*. Nouv. éd., Bordas, 1988.

Lagarde, André, and Laurent Michard. *XIXe siécle: Les grands auteurs français; anthologie et histoire littéraire*. Bordas, 1999.

Lanson, Gustave. *Hommes et Livres: Études Morales et Littéraires*. Hachette livre-bnf, 1895.

Leblond, Aude. *Corpus Chapitres*. Zenodo, 2022, https://doi.org/10.5281/ZENODO.7446728.

Moretti, Franco. "Conjectures on World Literature." *New Left Review*, 2000.

---. "The Slaughterhouse of Literature." *Modern Language Quarterly*, vol. 61, no. 1, Mar. 2000, pp. 207–28, https://doi.org/10.1215/00267929-61-1-207.

Mosteller, Frederick, and David L. Wallace. "Inference in an Authorship Problem." *Journal of the American Statistical Association*, vol. 58, no. 302, June 1963, p. 275, https://doi.org/10.2307/2283270.

Pedregosa, F., et al. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825–30.

Pennebaker, James W. *The Secret Life of Pronouns: What Our Words Say about Us*. 1st U.S. ed, Bloomsbury Press, 2011.

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*, vol. 36, no. 2, June 2020, pp. 430–38, https://doi.org/10.1093/llc/fqaa032.

Pollock, Griselda. *Differencing the Canon: Feminist Desire and the Writing of Art's Histories*. Routledge, 1999.

Porter, Jack D. "Popularity/Prestige." *Pamphlets of the Stanford Literary Lab*, no. 17, 2018, https://litlab.stanford.edu/LiteraryLabPamphlet17.pdf.

Schmitt, Michel P., and Alain Viala. "Les cotes aux concours." *Littératures classiques*, vol. 19, no. 1, 1993, pp. 281–91, https://doi.org/10.3406/licla.1993.1753.

Seminck, Olga, et al. "The Evolution of the Idiolect over the Lifetime: A Quantitative and Qualitative Study of French 19th Century Literature." *Journal of Cultural Analytics*, vol. 7, no. 3, Sept. 2022, https://doi.org/10.22148/001c.37588.

Thiesse, Anne-Marie. *La fabrique de l'écrivain national: Entre littérature et politique*. Gallimard, 2019.

Tolonen, Mikko, et al. "Examining the Early Modern Canon: The English Short Title Catalogue and Large-Scale Patterns of Cultural Production." *Data Visualization in Enlightenment Literature and Culture*, 2021, pp. 63–119, https://doi.org/10.1007/978-3-030-54913-8_3.

Underwood, Ted. *Distant Horizons: Digital Evidence and Literary Change*. The University of Chicago Press, 2019.

Underwood, Ted, and Jordan Sellers. "The *Longue Durée* of Literary Prestige." *Modern Language Quarterly*, vol. 77, no. 3, Aug. 2016, pp. 321–44, https://doi.org/10.1215/00267929-3570634.

van Cranenburgh, Andreas, et al. "Vector Space Explorations of Literary Language." *Language Resources and Evaluation*, vol. 53, no. 4, Feb. 2019, pp. 625–50, https://doi.org/10.1007/s10579-018-09442-4.

van Cranenburgh, Andreas, and Rens Bod. "A Data-Oriented Model of Literary Language." *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 1228–38, https://doi.org/10.18653/v1/e17-1115.

van Cranenburgh, Andreas, and Corina Koolen. "Identifying Literary Texts with Bigrams." *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, 2015, pp. 58–67, https://doi.org/10.3115/v1/w15-0707.

Verboord, Marc. "Classification of Authors by Literary Prestige." *Poetics*, vol. 31, no. 3–4, June 2003, pp. 259–81, https://doi.org/10.1016/s0304-422x(03)00037-8.

Viala, Alain. "Qu'est-ce qu'un classique?" *Littératures classiques*, vol. 19, no. 1, 1993, pp. 11–31, https://doi.org/10.3406/licla.1993.1737.

Yu, B. "An Evaluation of Text Classification Methods for Literary Study." *Literary and Linguistic Computing*, vol. 23, no. 3, Sept. 2008, pp. 327–43, https://doi.org/10.1093/llc/fqn015.

A. Appendix

A.1. Data Construction

A large part of our metadata, in particular those of the *brevet* and the *baccalauréat*, was recovered thanks to the immense work of the association *Le deuxième texte*,[7] which has put its data[8] online in open access. The purpose of the association is to highlight the value of women writers in the French cultural heritage. Other data were automatically retrieved using Python scripts on the web pages of the *Garnier-Flammarion* and the *Pléiade* collections, but also by hand for the authors present in the *Lagarde et Michard* compilations (Lagarde and Michard; Lagarde et al.).

A.2. Corpus distribution, canon at the novel scale

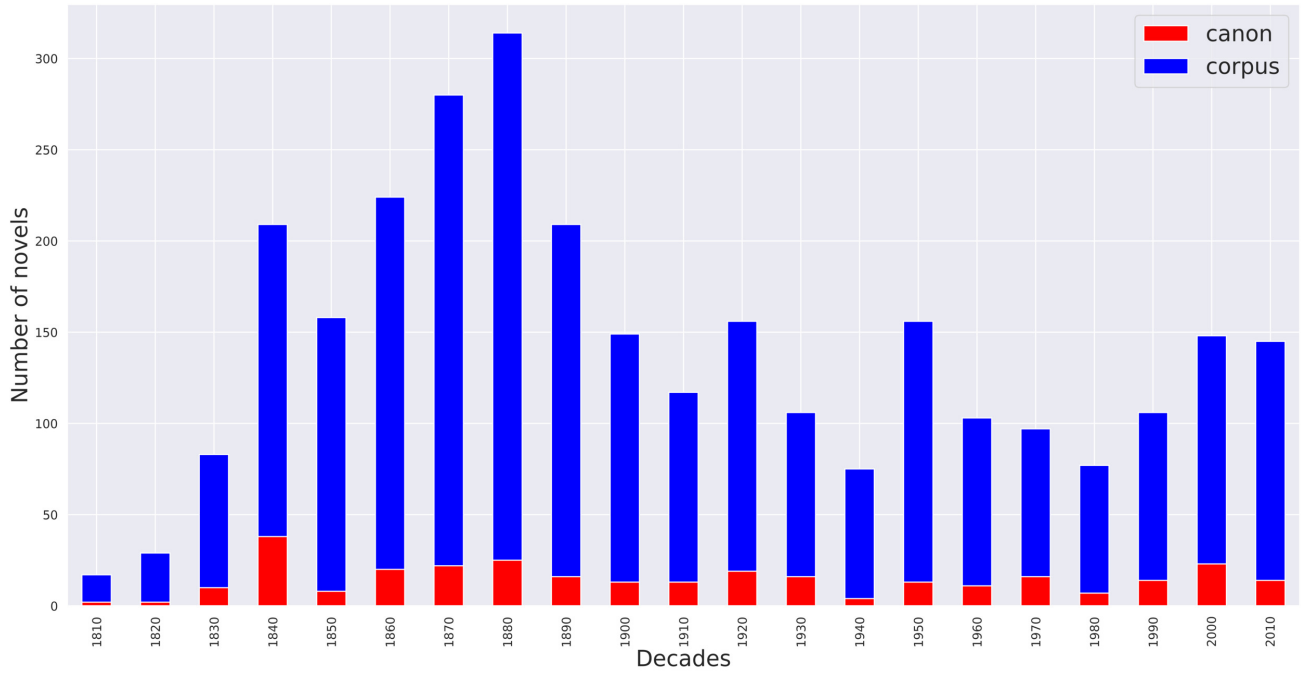

Figure 10. Distribution of the number of novels over time, broken out by canonicity tags, canon at the novel scale

A.3. Modeling Setup

We ran a grid search to find the best combination of parameters. It turned out that the best setup was the default one. The main issue we faced during the training was the imbalance between our classes as mentioned above. We therefore set the *class_weight* parameter to "balanced". This mode adjusts weights inversely proportional to class frequencies in the input data. As evaluation metrics we used balanced accuracy (average accuracy for each class), precision, recall and F1 score.

7 https://george2etexte.wordpress.com/

8 https://www.data.gouv.fr/fr/organizations/le-deuxieme-texte/

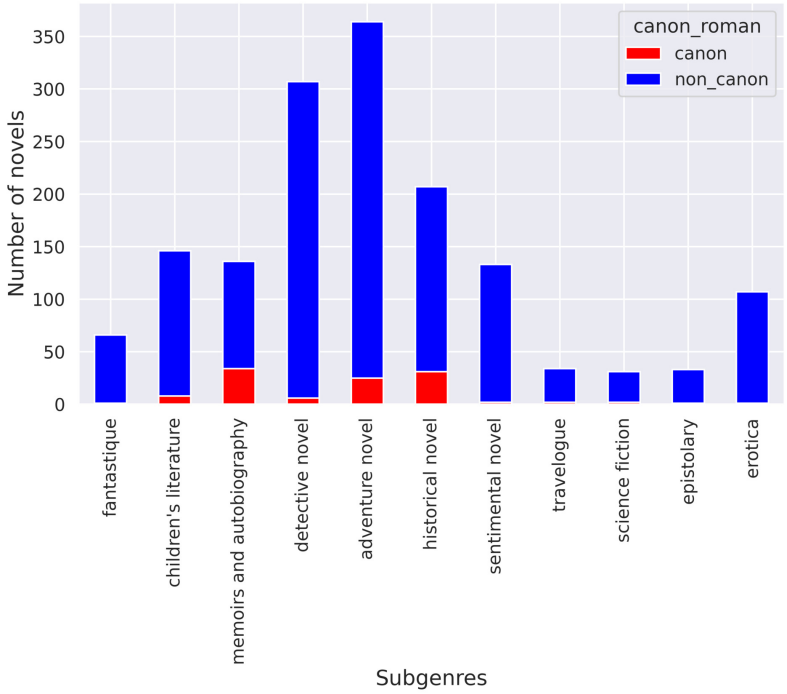

Figure 11. Literary sub-genres in the corpus, broken out by canonicity tags, canon at the novel scale

As a baseline, a random approach was adopted to ensure that the statistical model detected textual differences associated with the metadata rather than artificially managing to separate the two classes. We randomly drew our canonical or non-canonical labels for all the novels, according to their proportion in the dataset.

To handle the idiolectal bias, we implemented sklearn group strategy with three different functions: GroupKFold, StratifiedGroupKFold and LeaveOneGroupOut. Very similar results were obtained with all three. GroupKFold achieves a slightly better metric balance, so we presented its results in the paper.

A.4. Additional Experiments

We assessed the contribution of the different canonicity factors to the performance of the model. Six data-sets were created, each excluding one of the six canonical factors. The results ranged from 65% to 70% accuracy, which means that no single factor is required to carry out our classification. However, when more than one factor was removed, the score dropped significantly due to lack of data.

Furthermore, one analysis enables us to quantify the impact of idiolectal bias. Notably, when we allowed for an unconstrained distribution of an author's works between the training and test sets, the model performed significantly better. The accuracy surged from 0.78% at the novel scale to 0.91% at the author scale. The efficacy of our model in capturing idiolectal nuances can be attributed to our method's reliance on stop words. This strategy essentially acts as a "cheat code" for the model, facilitating author attribution instead of

focusing solely on canonicity. This observation carries intriguing implications, potentially underpinning the argument that the canon might be construed as an amalgamation of distinct authorial styles operating within specific subgenres.

PCA is a statistical technique for reducing the dimensionality of a dataset. This is accomplished by linearly transforming the data into a new coordinate system where most of the variation in the data can be described and represented in two dimensions to visually identify clusters of closely related data points. In our experiment, we projected all the works of the same author on a single plane to be able to compare the works, using only the 100 most frequent words.

A.5. Data Availability

The raw word relative frequencies for original texts used in this study can be downloaded on the Harvard Dataverse (Barré). All our metadata, scripts and output data are also available there.