Beyond Canonicity

Jean Barré[1,2,*], Thierry Poibeau[1,2] and Thomas Conrad[1]

[1]*École nationale supérieure - Université PSL, 45 rue d'Ulm, Paris, 75005, France*
[2]*Laboratoire Lattice (Langues, Textes, Traitements informatiques, Cognition), 1 rue Maurice Arnoux, Montrouge, 92049, France*

Abstract

This study offers a fresh perspective on the Canon/Archive problem in literature through computational analysis. Following Tynianov's understanding of literature, we adopt a dynamic approach to literature by proposing a model of literary variability using the Kullback-Leibler divergence. We retrieve key authors and works that are shaping the broad outlines of literary change. Our aim is to evaluate the importance of canonical authors on literary variability. We opt for a cohort-driven setup to analyze the variability brought by a given text, focusing on specific textual aspects such as topics, lexicon, characterization, and chronotope. The findings reveal that canonical authors tend to contribute slightly more to literary change than those from the archive.

Keywords

Literary history, Computational literary studies, Literary variability, Canon, Cohorts-driven model,

1. Introduction

The Canon/Archive problem is a well-known issue in the field of the Computational Literary Studies (CLS). It has been and continues to be a fundamental aspect of the CLS field, as computational methods allows researchers to expand their investigations beyond the limited study of the Canon and its restricted number of texts. With the ability to process vast amounts of digitized texts in a matter of hours, researchers can now engage in distant reading, as proposed by Moretti [1], and conduct experiments on the textual content of literary works. This approach enables scholars to zoom in and out from the literary past, leading to a better understanding of general trends describing literary evolution.

This introduction of new perspectives and alternative modes of inquiry raises a fundamental question: "Do we understand the broad outlines of literary history ?". Underwood [2, 3] eloquently poses this question, contemplating whether the preserved texts thus far adequately represent the entire spectrum of literary production, or if the literary discipline has been constrained by narrow perspectives throughout its existence.

This line of investigation is not entirely new, as Iouri Tynianov expressed similar concerns in 1927 when he stated that "The theory of values in literature leads us to the perilous study

CHR 2023: Computational Humanities Research Conference, December 6 – 8, 2023, Paris, France
*Corresponding author.
jean.barre@ens.psl.eu (J. Barré); thierry.poibeau@ens.psl.eu (T. Poibeau); thomas.conrad@ens.psl.eu (T. Conrad)
https://crazyjeannot.github.io/ (J. Barré)
0000-0002-1579-0610 (J. Barré); 0000-0003-3669-4051 (T. Poibeau); 0000-0003-3669-4051 (T. Conrad)

of principal but isolated phenomena, and reduces literary history to the *dominant's history*" [4]. However, Tynianov offered a way out by suggesting that the value of a given literary phenomenon should be understood in terms of its "*significance and evolutionary qualities*" [4]. According to Tynianov, literary recognition is a dynamic process, and analyzing it requires studying *literary variability*. This refers to the diversity and range of formal elements present in literary works. It encompasses the different ways authors employ language, style, themes, narrative structures, characterization, settings, and other literary elements to create unique and distinct pieces of literature. This perspective sees literary history not as a linear chronology but considers literature within a dynamic and indivisible process that is constantly evolving. Every written text available in a library has the potential to influence the process of writing. Accounting for this perpetual movement necessarily requires an understanding of how each new text seeks to formally distinguish itself from its predecessors while still being shaped, for example, by a specific, conscious or unconscious, generic intertextuality.

Are canonical novels playing a driving role in literary variability? Canonical novels, being recognized for their importance and influence in the literary tradition, can serve as reference points for writers and readers [5]. Their impact on literary practices is manifest in various ways: inspiring new writing styles, introducing innovative themes, or encouraging formal experiments. Writers can be influenced by these canonical novels, either seeking to differentiate themselves or align with their influence [6]. From this perspective, the traditional concept of canonicity is seen as biased and limited in capturing the evolution of formal practices. The complex nature of the canonization process, influenced by external factors such as the education system and editorial policies [7], hinders its ability to incorporate *avant-garde* literary changes.

This study aims at evaluating to what extent canonical works are reliable witnesses in terms of literary variability. Previous research uncovered disparities in the textual content between what is considered canonical and non-canonical across various corpora and cultural backgrounds ([8], [9], [10], [11]). This previous work showed that canonical sets share to some extent (at least for specific timespans) an intrinsic norm. Is this specific norm capable of capturing literary variability? Or is it missing something?

In this paper, we present an operational model of the notion of formal variability in 19th and 20th century novels. Our approach is grounded in the canonical sets established in prior research on contemporary reception in France ([8]). Our objective is to investigate whether the canonized sample from contemporary reception accurately reflects the overall extent of change within literature as a whole. For this purpose, we try to identify the key works and key authors that drive literary variability. By analyzing formal aspects and considering literature as a dynamic system, we seek to gain insights into the flow of literary variability and challenge assumptions about the representativeness of canonical novels.

2. Methods

2.1. Corpus

This study is based on the corpus collected in the framework of the "ANR Chapitres"[1], a massive corpus of nearly 3000 French novels [12]. The goal of this project was to evaluate the pace of change in the length of chapters over two centuries. The corpus is structured in XML-TEI[2] (Text Encoding Initiative) encoding, to add metadata to the texts. The period concerned extends over two centuries of novel production, from the 19th to the 20th century, as can be seen in Figure 1.

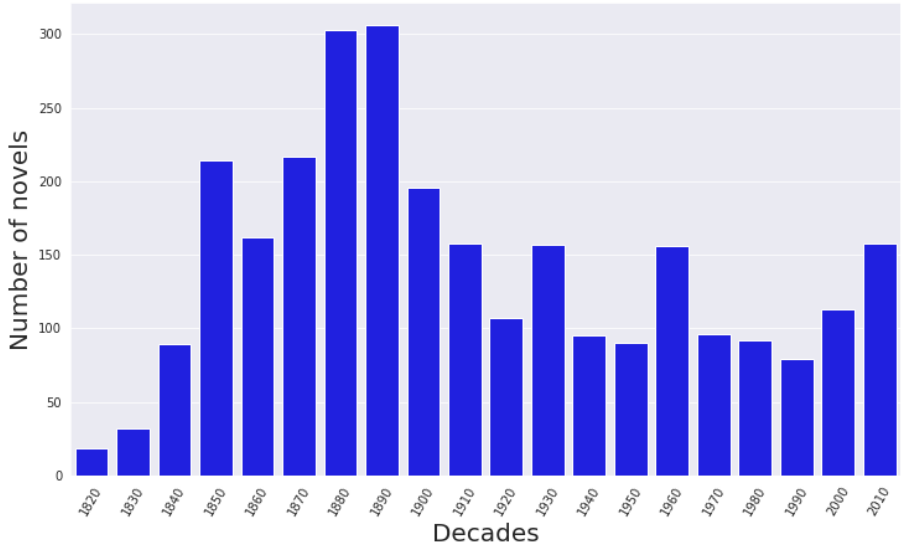

Figure 1: Distribution of the number of novels over time

Each text in the corpus is enriched with metadata, including subgenre tags, authors' dates (birth and death). The latter is really relevant for our work as we focus on cohorts effect on the pace of literary change.

2.2. Textual features

The concept of literary variability encompasses a broad spectrum of possibilities, and it can manifest itself in various ways within a text. By examining specific elements like themes, characterization, vocabulary, and chronotope, we aim to understand how novels have evolved across different time periods. However, some notions (like 'the plot') are hard to formalize and are thus not included in this study.

We first implemented topic modeling methods to extract topics from the texts. The Python library Bertopics [13] was used in a guided setting. This refers to a set of techniques that

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

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

influence the topic modeling process by providing predefined seed topics for the model to converge towards. These techniques enable users to specify a predetermined number of topic representations that are guaranteed to appear in the resulting documents. We constructed a list of 50 topics we found relevant for our study and retrieved their proportion within each novel. [3]

We also implemented a Bag-of-ngrams approach to retrieve the lexicon dimension of literary change. To do so, we rely on the 1000 most frequent lemmas and 1000 most frequent bigrams of lemmas. This may echo the paper by van Cranenburgh and Koolen [14], which showed that only using unigrams and bigrams was sufficient to classify literary texts regarding literary quality aspects. We did not remove stopwords, since they may reflect an unconscious and automatic structural way of writing, rather than less frequent words related to the content and themes of the text. Our hypothesis is that the structural way of writing novel change over time and cohorts. Bag-of-words techniques work well for various experiments in the CLS field (stylometry and author attribution for example [15]). But they are quite controversial from a literary point of view, since they exclude lots of information, including word order and syntax. They are also limited in the sense that they do not take into account the semantic drift of words throughout time. The word "wild" does not refer to the same meaning used in an adventure novel from the late 19th century or in a climate fiction from the late 20th century. Considering this statement, we assumed that bag-of-features are still getting some dimensions of literary change, particularly on the very frequent structural elements.

One of the aims of this study was to capture *chronotope* information, which is a term coined by the Russian literary scholar Bakhtin [16]. In substance, the concept of chronotope explores how the relationship between time and space influences the portrayal of characters, the development of plotlines, and the themes conveyed within a literary text. In the Natural Language Processing (NLP) context, Lasse Kohlmeyer et al. [17] demonstrated the limitations of traditional document embeddings (optimized for shorter texts) in capturing complex facets in novels (such as time, place, atmosphere, style, and plot). To address this problem, they propose to use multiple embeddings reflecting different facets, splitting the text semantically rather than sequentially. Inspired by those findings, we adapted their methodology. By using an NLP pipeline specifically tuned for novels, (fr-BookNLP, part of the multilingual BookNLP project [18, 19]), we extracted literary entities representing the chronotope, specifically focusing on FAC, TIME, LOC, and VEH. The presence of chronotope elements in a novel is highly influenced by its subgenre categorization. We believe that this type of information is crucial for our task as it has the potential to capture significant aspects of literary variability. To obtain vector representations of the chronotope elements in novels, we trained a Paragraph Vectors model [20] (Doc2Vec) using a subset of our novel dataset. This model allowed us to generate vector embeddings for our four different facets (each with 300 dimensions) that capture the chronotope information of the texts.

We also recognized the significance of characterization in our task, as we believed that changes in literature could influence the portrayal of characters to readers. We thus focused on identifying key verbs that drive the actions of the main characters and the adjectives used to describe them. In line with Woloch [21]'s concept of the character space as "the encounter between an individual human personality and a determined space and position within the

[3] For the topic modeling process, see appendix A.1

narrative as a whole", we used coreference resolution techniques, specifically those offered by fr-BookNLP[4], to automatically detect and analyze the distribution of character mentions throughout the narrative [22]. We used the Spacy parser to extract the verbs and adjectives associated with each character mention. By analyzing the syntactic structure of the text, we identified the verbs that represented the actions of the characters and the adjectives that characterized them. For each novel, we selected the top five main characters and generated two vector embeddings of 300 dimensions: one representing the adjectives associated with the characters and the other representing the verbs. These embeddings capture the semantic information related to the characters' traits and actions, providing a compact representation of their characteristics within the narrative.

By incorporating these various aspects, each novel can be represented as a multidimensional vector with 3850 dimensions, 50 for the topics, 2000 for the bag-of-ngrams, 1200 for chronotope elements and 600 for the characterization. Therefore, our vector representation provides a comprehensive formalization of the novel, enabling further analysis and comparisons.

2.3. Measuring Literary Variability

Basing our work on the aspects presented above (topics, bag-of-ngrams, chronotope, and characterization), we had to find a way to grasp literary variability. We decided to use a commonly used measure, the Kullback-Leibler divergence (KLD). It is a type of statistical metric that enables us to quantify the dissimilarity between two probability distributions: the target distribution P and a reference distribution Q. Within this framework, we assess the variability of a text by measuring the surprise or deviation of that text from a set of other texts. Specifically, the KLD from Q to P is defined as follows:

$$KL(P||Q) = \sum_{x} P(x) \log(\frac{P(x)}{Q(x)})$$

where P represents our formal features, normalized as a probability distribution, and Q stands for the average of all the texts we wish to compare P with. This measure, derived from information theory, finds application in various fields, including assessing sample diversity in ecology and examining elements of linguistic evolution [23]. Barron et al. [24] applied KLD to a corpus of debates in the french revolution's first parliament, assessing both the novelty of a particular speech compared to prior speeches and its transience compared to future ones.

In the realm of literature, Algee-Hewitt et al. [25] proposed this method to determine the informational content of texts by evaluating the predictability of word-to-word transitions, taking into account the range of possible transitions. Liddle [26] also discussed that mathematical information theory may be relevant to literary analysis by showing statistically significant correlations between national histories of the novel and information-theoretical pressures.

Previous research showed that literary change throughout an author's life was powerful enough to predict the publication date of a given text [27]. However, other studies demonstrated that literary change brought about by an author throughout their life remains limited compared to the cohort effect [28]. In other words, literary change appears to be driven by cohort renewal,

[4]A discussion on the evaluation of fr-BookNLP and its limitations is provided in the appendix A.2.

which is indeed relevant since events that shape an author's life, likely to have an impact on their writing style, also influence all authors within the same generation.

Measuring the variability of a text in relation to a set of other texts immediately places us in a dual configuration: we can measure the variability of a given text with the works that precede it and with those that follow it. Studying the circulation, selection, and propagation of literary patterns in a group of texts, we can understand the dynamics of literary change and the extent to which an author's language patterns influence and are adopted by others. From a literary perspective, measuring change between two successive texts may not make much sense, as numerous factors can come into play, such as affiliation with a particular subgenre, a specific period, a literary school, or even the author themselves. Therefore, a specific framework is necessary to conduct our experiments, which revolves around the notion of generation.

Béhard defines a generation as a concept that aims to "understand the succession of aesthetic productions based on a community of upbringing, interests, and ideas specific to the same age group of writers, following a periodicity of approximately 30 years linked to historical and political cycles"[29]. This generation-based approach allows us to examine the changes and innovations introduced by authors within their respective cohorts, while also considering the broader historical and literary context in which these works emerge. It provides a more nuanced understanding of how literary variability is shaped and influenced by various factors, contributing to a richer analysis of the dynamic nature of literature.

To further support Béhar's argument, Moretti [30] also evaluated the regularity of the replacement of literary subgenres that he examines. He suggested that "a sort of generational mechanism seems to be the best way to account for the regularity of the cycle of novelistic production". Moretti's analysis focused on the cycle of change, considering a timeframe of 25 to 30 years. These studies are complemented by the research of Underwood et al. [28], who showed the significance of cohorts on literary change. Their findings indicated that cohorts have such a substantial impact that they account for more than half of the amount of change in literature.

Building upon their conclusions, we compute KLD comparing successive cohorts in a timespan of 30 years. For instance, if analyzing a novel published in 1970 by an author born in 1930, we compare it with all the books written by authors born between 1870 to 1900 to evaluate the extent of change. This approach enables us to consider cohorts as a rolling phenomenon, since defining arbitrary cohorts could not be representative of the cohort succession phenomenon. This methodology allows us to view literature as a continually evolving synchronic system, framed by cohorts.

3. Results

3.1. Literary dynamics : between novelty and influence

Figure 2 represents the amount of variability for each text : The x-axis represents the entropy of each text, relative to the cohort preceding the text's author, indicating the level of surprise or formal novelty in the text compared to its preceding cohort. A text with high "surprise" score on this axis would be considered formally innovative. The y-axis represents the entropy of each text, relative to the cohort following the text's author, indicating the level of surprise or

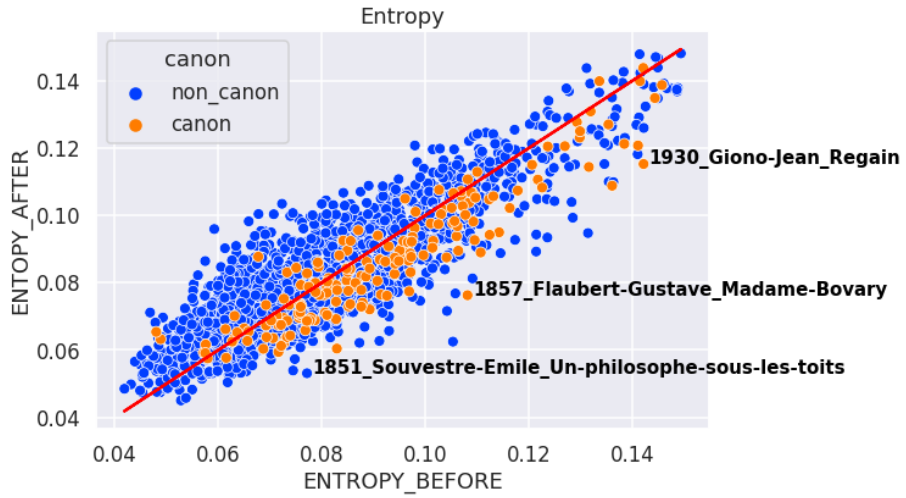

Figure 2: Amount of variability comparing cohort before and cohort after

influence the text has on the next cohort. A text with high "surprise" on this axis would indicate that it has little influence on what follows. A text that is both highly influential and innovative would receive a high value on the x-axis and a low value on the y-axis.

The graph, being complex, required some elements to facilitate interpretation: three novels are depicted for a better understanding of how the graph works. Gustave Flaubert's *Madame Bovary* stands out with high novelty and high influence scores, thanks to its groundbreaking narrative style, character development, and enduring impact on literature. Jean Giono's *Regain* receives a high novelty score for its exploration of resilience and human connection to nature, but its low influence suggests that it didn't gain immediate widespread recognition. Émile Souvestre's *Un Philosophe sous les Toits* addresses social issues and garnered high influence despite not being part of the French literary canon, making it historically significant. Canonical texts are highlighted in orange. It is notable that these texts are distinctively positioned below the x=y line. This suggests that canonical works exhibit greater variability compared to the preceding cohort but slightly less variability compared to the following generation. This representation of the canon indicates a higher level of innovation and influence.

3.2. Signal of literary variability

These initial findings were highly intriguing, prompting us to pursue a complementary approach to gain a deeper understanding. We focused on the x-axis, which we deemed more comprehensible from a literary standpoint. The change introduced by a text (or an author) in relation to the previous generation is intuitively grasped as each text finds a way to differentiate itself from the broader literary production of a given period. By associating the entropy value obtained for each text with the author's birth date, we were able to represent in figure 3 the signal of change over time.

Through visual representation, we showcased the patterns and fluctuations observed in the analysis of KL divergence or entropy. This approach allowed us to capture the dynamic nature of literary change and observe how it manifests itself over different historical periods. The resulting graph provides a visual narrative of the evolving literary landscape, shedding light on the distinctive voices, innovative works, and influential shifts in the realm of literature. It offers a compelling visual representation of the signal of literary change, enabling a more nuanced understanding of the complex processes at play in cultural and artistic evolution.

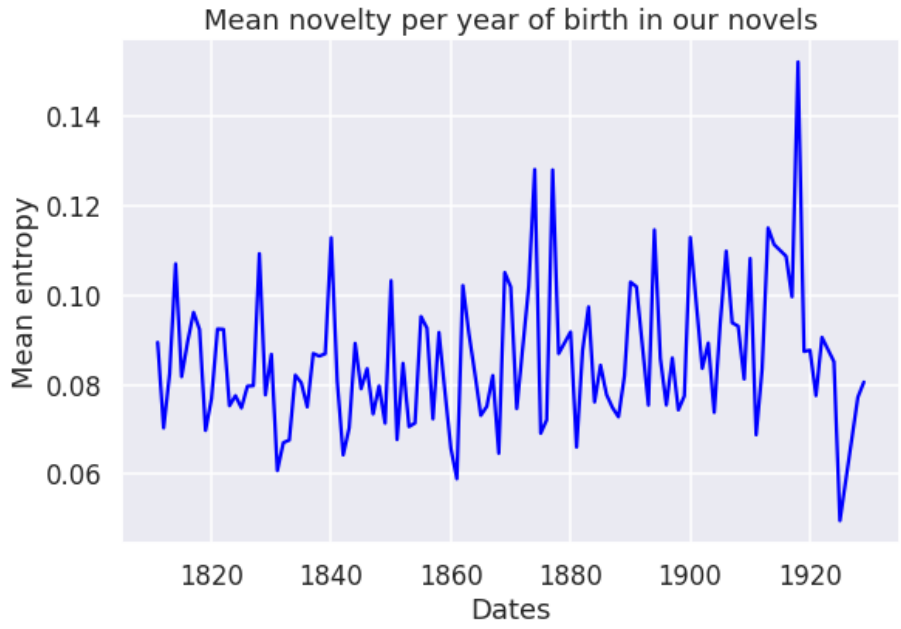

Figure 3: Signal of change comparing previous cohort

Each peak corresponds to a significant variability introduced by a specific author (or a group of authors sharing the same birth date). This approach allows us to identify the names of key works and key authors that drive literary variability. One shall notice that the last peaks should be ignored due to lack of authors born around 1920 and later in our corpus.

Jules Verne, born in 1828, is the author who mainly explains the second peak in variability. His works, particularly *Vingt-mille lieues sous les mers*, published in 1870 (with 0.149 KLD), and *L'Île Mystérieuse*, published in 1875 (with 0.232 KLD) exemplify his innovative approach to literature. In the first one, Verne introduced readers to Captain Nemo's underwater vessel, the Nautilus, which travels beneath the seas and explores uncharted depths. This visionary depiction of a futuristic submarine, powered by electricity and equipped with advanced technology, set the stage for the emergence of the science fiction genre. In *L'Île Mystérieuse*, Verne combined elements of adventure and survival on a remote island with the exploration of technology and engineering. The novel tells the story of a group of castaways who use their knowledge and resourcefulness to survive and thrive on the island.

Verne's innovative storytelling can be seen as a response to the social fascination with progress

and exploration. During the late 19th century, the world was witnessing rapid advancements in science and technology, driven by the Industrial Revolution and scientific discoveries. This era of progress and innovation deeply influenced the literary landscape, as writers such as Verne sought to capture the spirit of exploration and curiosity prevalent in society. This way, Verne laid the foundation for a mixture between adventure and science fiction subgenres.

The peak from 1877 is led by Raymond Roussel, a lesser-known but highly innovative writer, particularly with his works *Impressions d'Afrique*, published in 1910 (with 0.145 KLD) and *Locus Solus*, published in 1914 (with 0.21 KLD). The first one is a novel that defies traditional narrative conventions and follows a dreamlike, non-linear structure. The story revolves around a group of travelers who embark on a journey through Africa, encountering strange and surrealistic occurrences along the way. The second narrative is also characterized by its intricacy and complexity, as it contains multiple layers of storytelling that takes you on a tour of the estate of a scientist named Martial Canterel, where he showcases a series of bizarre and macabre inventions. Roussel's experimental and imaginative storytelling style set him apart as a pioneer in avant-garde literature, and his works can be seen as early surrealism.

René Crevel and Nathalie Sarraute lead the peak in 1900. René Crevel's work *Le roman cassé*, published in 1935 (with 0.17 KLD) and Nathalie Sarraute's *Tropismes*, published in 1939 (with 0.159 KLD) both exemplify their innovative approaches to literature, as they challenged conventional narrative structures and scrutinized the inner workings of human consciousness.

The first novel breaks away from traditional linear storytelling and embraces a fragmented and non-linear narrative style. Crevel's exploration of the subconscious mind and his use of stream-of-consciousness writing make the novel a precursor to the surrealist and modernist movements. The novel centers around the mental states of its characters, delving into their thoughts, dreams, and desires. The title *Le roman cassé* itself, which translates to *The Broken Novel*, is indicative of Crevel's intention to dismantle traditional narrative conventions and explore new modes of expression. His work can be seen as an early example of the deconstruction of the novel form, where the focus shifts from external events and plot-driven storytelling to an exploration of the characters' inner lives and psychological states.

Nathalie Sarraute's *Tropismes* is a collection of interconnected short prose pieces that explore the subtle and fleeting movements of the characters' inner thoughts and feelings. Sarraute's writing style is characterized by its precision and attention to the nuances of human behavior. She coined the term "tropismes" to describe these brief and involuntary movements of the characters' consciousness. Her innovative use of language and her focus on the psychological subtleties of her characters set her apart as a pioneer of the *Nouveau Roman* movement.

Thus, our novelty signal highlights works and authors who have significantly distanced themselves from the dominant formal rules of the previous generation. We identified authors who have contributed to the creation of new sub-genres or avant-garde writers with varying degrees of recognition, like Raymond Roussel, author from the *Archive*. These peaks might represent pivotal moments in literary history where new ideas, styles, or narrative techniques emerge, leading to a distinct shift in the literary landscape. Nevertheless, it is worth noting that any work that differs formally from the majority of other novels stands out. For instance, children's novels like *Le petit prince* (Antoine de Saint-Exupéry) also prominently emerge in the scores.

3.3. Canonical novels, drivers of literary variability?

When examining the list of authors who stand out in their contribution to literary change, many of them are well-known and directly associated with the literary canon. To assess the amount of change among canonical works compared to non-canonical works from the archive, we project in figure 4 two distinct curves onto the graph based on their canonicity labels, at the author level (considering all works by an author as canonical). By considering the canonicity distinction, we gain a deeper understanding of how these different subsets of texts contribute to the overall landscape of literary variability.

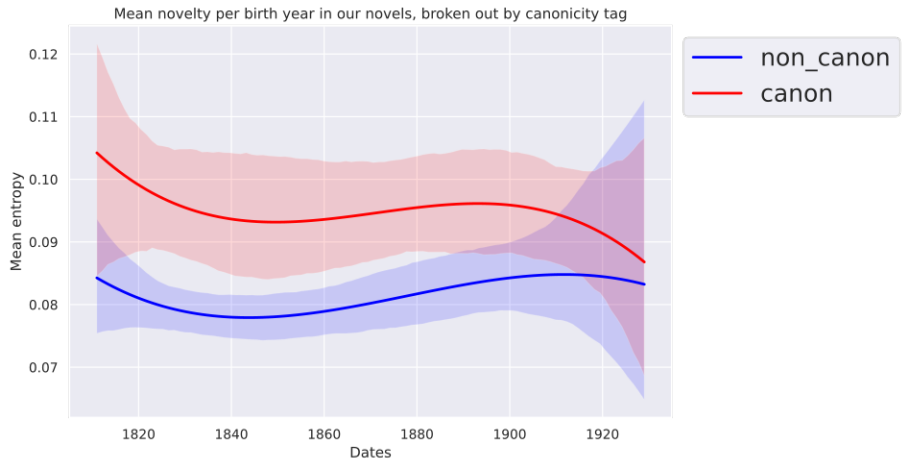

Figure 4: Mean Novelty per birth year, broken out by canonicity tag

Thus, we observe two distinct curves on the graph: the red curve representing canonical authors and the blue one representing non-canonical authors. The margin of error for canonical authors is larger due to their smaller number compared to the archival authors. The gap between both sets remains relatively stable over a century of authors' birth dates. The clear conclusion from the graph is that canonical authors tend to introduce more variability in their novels compared to non-canonical authors.

The smaller difference observed towards the end of the period suggests a few possibilities. It implies that the overall corpus becomes more limited towards the end, with a smaller pool of texts available for analysis. Furthermore, it indicates that the criterion of canonicity might be less relevant for the last generations of authors.

4. Limitations

Our approach is subject to the inherent accuracy limitations of many NLP algorithms used, including fr-BookNLP, Spacy, and Bertopics, all of which are prone to making errors.

The choice of a 30-year time frame for cohort succession in our experiments is somewhat subjective. Although it is reasonable to assume that significant changes occur within this window compared to shorter intervals like 5 or 10 years, the selection remains debatable. Furthermore,

our comparisons are limited to successive cohorts, neglecting the potential influence of earlier literary works. Authors are likely to have been influenced by canonical texts published decades or even centuries before their own works, which warrants further consideration.

We faced the challenge of conducting close readings. While we have identified distinctive authors and texts, we have not provided textual evidence of the observed changes. Given the large-scale nature of our study, it is understandable, but future research should strive to incorporate detailed textual analysis to support our findings. By examining specific passages and linguistic features, a more comprehensive understanding of the observed literary variability can be achieved.

5. Conclusion

In conclusion, this study has provided valuable insights into the dynamics of literary variability and the role of canonical works in the French literary landscape. Through our operationalization of formal variability and our cohort-driven model, we succeeded to identify the names of key works and key authors that drive literary variability. We also examined the extent to which the canon accurately represents the overall amount of change in literature. Our findings suggest that canonical authors contribute to greater variability compared to non-canonical authors. This might imply that canonical works have a significant impact on the introduction of diverse elements in literature. It is worth noting that the fading gap in variability towards the end of the corpus raises questions about the evolving nature of the canon itself.

Analyzing formal aspects such as topics, styles, chronotopes, and characterization in a large corpus of novels, the study aims to uncover patterns of literary change and explore the relationships between texts, authors and cohorts. By organizing texts into generations, we established a temporal and contextual framework that allows us to capture and analyze the evolving literary dynamics over time. This approach acknowledges that texts produced within the same generation share certain characteristics and influences, providing a meaningful basis for measuring and understanding literary variability.

Moving forward, future research could focus on investigating the role of specific subgenres in driving literary change. By examining whether the emergence or growth of certain subgenres corresponds to peaks of change, we can gain a deeper understanding of how different literary trends influence the overall dynamics of literature.

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).

Lastly, the authors also wish to thank the anonymous reviewers whose comments have helped us to substantially improve this paper.

References

[1] F. Moretti, Conjectures on world literature, New Left Review (2000).

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

[3] T. Underwood, We don't already understand the broad outlines of literary history., 2013. URL: https://tedunderwood.com/2013/02/08/we-dont-already-know-the-broad-outlines-of-literary-history/.

[4] Y. Tynianov, On Literary Evolution (1927), Academic Studies Press, Boston, USA, 2019, pp. 267–282. URL: https://doi.org/10.1515/9781644690635-015. doi:10.1515/9781644690635-015.

[5] G. Pollock, Differencing the canon: feminist desire and the writing of art's histories, Revisions, Routledge, 1999.

[6] A. Mukherjee, Canonicity, 2017. URL: https://oxfordbibliographies.com/view/document/obo-9780190221911/obo-9780190221911-0054.xml. doi:10.1093/obo/9780190221911-0054, institution: Oxford University Press, Pages: 9780190221911-0054.

[7] P. Bourdieu, Les règles de l'art, Éditions du Seuil, 1992.

[8] J. Barré, J.-B. Camps, T. Poibeau, Operationalizing canonicity, ??? (2023-06).

[9] M. Algee-Hewitt, M. McGurl, Between canon and corpus: Six perspectives on 20th- century novels, Pamphlets of the Stanford Literary Lab (2015). URL: https://litlab.stanford.edu/LiteraryLabPamphlet8.pdf.

[10] T. Underwood, J. Sellers, The "longue durée" of literary prestige, Modern Language Quarterly 77 (2016-09) 321–344. URL: https://read.dukeupress.edu/modern-language-quarterly/article/77/3/321-344/47316. doi:10.1215/00267929-3570634.

[11] J. Brottrager, A. Stahl, A. Arslan, U. Brandes, T. Weitin, Modeling and predicting literary reception, Journal of Computational Literary Studies (2022). URL: https://jcls.io/article/id/95/. doi:10.48694/JCLS.95, publisher: Universitäts- und Landesbibliothek Darmstadt.

[12] ANRChapitres, Anrchapitres/2000romans19e20e: Corpus chapitres, 2022. URL: https://doi.org/10.5281/zenodo.7446728. doi:10.5281/zenodo.7446728.

[13] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022).

[14] A. van Cranenburgh, C. Koolen, Identifying literary texts with bigrams, in: Proceedings of the Fourth Workshop on Computational Linguistics for Literature, Association for Computational Linguistics, 2015, pp. 58–67. URL: http://aclweb.org/anthology/W15-0707. doi:10.3115/v1/W15-0707.

[15] M. Kestemont, Function words in authorship attribution. from black magic to theory?, in: Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL), Association for Computational Linguistics, 2014, pp. 59–66. URL: http://aclweb.org/anthology/W14-0908. doi:10.3115/v1/W14-0908.

[16] M. Bakhtin, The dialogic imagination: four essays, number 1 in University of Texas Press Slavic series, 18. paperback printing ed., Univ. of Texas Press, 2011.

[17] Lasse Kohlmeyer, Tim Repke, Ralf Krestel, Novel views on novels: Embedding multiple facets of long texts, 2021 Association for Computing Machinery. (2021-11-14).

[18] D. Bamman, T. Underwood, N. A. Smith, A bayesian mixed effects model of literary char-

acter, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2014, pp. 370–379. URL: https://aclanthology.org/P14-1035. doi:10.3115/v1/P14-1035.

[19] D. Bamman, BookNLP, 2021. URL: https://github.com/booknlp/booknlp.

[20] Q. V. Le, T. Mikolov, Distributed representations of sentences and documents, CoRR abs/1405.4053 (2014). URL: http://arxiv.org/abs/1405.4053. arXiv:1405.4053.

[21] A. Woloch, The One vs. the Many, Princeton University Press, 2003. URL: http://www.jstor.org/stable/j.ctt7srp4.

[22] J. Barré, P. Cabrera Ramírez, F. Mélanie, I. Galleron, Pour une détection automatique de l'espace textuel des personnages romanesques, in: Humanistica 2023, Corpus, Association francophone des humanités numériques, Genève, Switzerland, 2023. URL: https://hal.science/hal-04105537.

[23] C. Bentz, D. Alikaniotis, M. Cysouw, R. Ferrer-i Cancho, The entropy of words—learnability and expressivity across more than 1000 languages, Entropy 19 (2017-06-14) 275. URL: http://www.mdpi.com/1099-4300/19/6/275. doi:10.3390/e19060275.

[24] A. T. J. Barron, J. Huang, R. L. Spang, S. DeDeo, Individuals, institutions, and innovation in the debates of the french revolution, Proceedings of the National Academy of Sciences 115 (2018) 4607–4612. URL: https://pnas.org/doi/full/10.1073/pnas.1717729115. doi:10.1073/pnas.1717729115.

[25] M. Algee-Hewitt, S. Allison, M. Gemma, R. Heuser, H. Walser, F. Moretti, Canon/archive. large-scale dynamics in the literary field, Pamphlets of the Stanford Literary Lab (2016). URL: https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf.

[26] D. Liddle, Could fiction have an information history? statistical probability and the rise of the novel, Journal of Cultural Analytics (2019). URL: https://culturalanalytics.org/article/11056. doi:10.22148/16.033.

[27] O. Seminck, P. Gambette, D. Legallois, T. Poibeau, The evolution of the idiolect over the lifetime: A quantitative and qualitative study of french 19th century literature, Journal of Cultural Analytics 7 (2022-09-01). doi:10.22148/001c.37588.

[28] T. Underwood, K. Kiley, W. Shang, S. Vaisey, Cohort succession explains most change in literary culture, Sociological Science 9 (2022-05-02) 184–205. URL: https://sociologicalscience.com/articles-v9-8-184/. doi:10.15195/v9.a8.

[29] H. Béhar, La littérature et son golem, number 58 in Vol. 1: Travaux de linguistique quantitative, H. Champion, 1996.

[30] F. Moretti, Graphs, maps, trees: abstract models for literary history, paperback edition ed., Verso, 2007.

A. Appendix

A.1. Topic modeling : Detailed approach

We provided Bertopics 50 specific topics with a list of 10 words associated with each topic. These topics served as seed topics to guide the model's convergence during the analysis. Bertopics is an algorithm with several layers, for the embedding one we employed the Camembert sentence vectorizer to create embeddings for the sentences. These embeddings capture the semantic meaning of the sentences and facilitate further analysis. Then we employed Principal Component Analysis which allowed us to transform the high-dimensional embeddings into a lower-dimensional space while preserving the essential information, making the subsequent steps more efficient. Then we ran the clustering on the sentences into distinct groups based on their semantic similarities, with the HDBSCAN algorithm. HDBSCAN is a density-based clustering method that identifies clusters of varying shapes and sizes, allowing us to group sentences that share similar topics or themes. Throughout the analysis, the model was capable of retrieving more than the initial 50 predefined topics. However, in order to maintain consistency and focus on the specific topics we had predefined, we made the decision to stick with the 50 seed topics provided by Bertopics. This allowed us to have a more targeted and interpretable analysis, focusing on the topics that were of particular interest for our research.

A.2. Fr-BookNLP Evaluation

A.2.1. NER

	precision	recall	F_1
PER	85.0	92.1	88.4
LOC	59.4	54.3	56.8
FAC	73.4	66.0	69.5
TIME	75.3	36.4	49.1
VEH	68.9	63.6	66,1

In the context of evaluating the performance of the model, having better precision than recall implies that when the model identifies literary entities, it is more likely to be accurate in its predictions. Precision measures the percentage of correctly predicted literary entities out of all the predicted entities. This is beneficial for the analysis as it ensures that the identified literary entities are more likely to be correct, even though some relevant entities may be missed (lower recall). In this context, prioritizing precision helps in reducing false positives and improving the reliability of the identified literary entities. The focus of the analysis is on identifying literary entities rather than employing Named Entity Recognition (NER) techniques. Literary entities encompass specific elements related to literature, such as the titles of books, names of authors, and other literary references, which may not always be recognized as standard named entities in typical NER tasks. By tailoring the analysis to literary entities, the results are better suited to capture and interpret relevant information specific to the literary domain. This approach

ensures more accurate and contextually appropriate results for the purpose of understanding literary variability and its driving factors.

A.2.2. Coreference resolution :

Metrics	F_1	
MUC	88,0	$Average$ 76.4
B^3	69,2	
$CEAF_e$	71.8	

The issue of duplication arises when the model detects the same character multiple times within the analyzed text. In some cases, the top five literary entities identified by the model may contain instances where two or more main characters from a text are the same character in terms of name or attributes. While this duplication might seem problematic at first glance, it is essential to understand the context and purpose of the analysis. In this particular study, the primary objective is not to identify unique and distinct characters but rather to retrieve a proxy for characterization as a whole. We aim to capture the prevalence and significance of certain characters across different texts and literary works. Therefore, the focus is more on character representation and the overall impact of these characters on the literary landscape, rather than identifying completely separate and non-repeating characters.