Latent Structures of Intertextuality in French Fiction

How literary recognition and subgenres are framing textuality

Jean Barré[1,2,*]

[1]*École normale supérieure - Université PSL, 45 rue d'Ulm, Paris, 75005, France*

[2]*LaTTiCe (Langues, Textes, Traitements informatiques, Cognition), 1 rue Maurice Arnoux, Montrouge, 92049, France*

Abstract

Intertextuality is a key concept in literary theory that challenges traditional notions of text, signification or authorship. It views texts as part of a vast intertextual network that is constantly evolving and being reconfigured. This paper argues that the field of computational literary studies is the ideal place to conduct a study of intertextuality since we have now the ability to systematically compare texts with each others. Specifically, we present a work on a corpus of more than 12.000 French fictions from the 18th, 19th and early 20th century. We focus on evaluating the underlying roles of two literary notions, sub-genres and the literary canon in the framing of textuality. The article attempts to operationalize intertextuality using state-of-the-art contextual language models to encode novels and capture features that go beyond simple lexical or thematic approaches. Our findings suggest that both subgenres and canonicity play a significant role in shaping textual similarities within French fiction. These discoveries point to the importance of considering genre and canon as dynamic forces that influence the evolution and intertextual connections of literary works within specific historical contexts.

1. Introduction

How has textuality been shaped over time? Can we account for the dynamics of influence and imitation in literary history? What roles have played the underlying structures, such as the literary canon or the literary genres?

Intertextuality is the concept that centralizes these questions. It was introduced by Kristeva [1], stating that "every text is a mosaic of quotations; every text is the absorption and transformation of another text." This perspective suggests that a text is no longer finite and that its complete interpretation and understanding must involve unraveling the network of textual relationships. A few years later, Barthes [2] further develops the definition, stating that "every text is an intertext; other texts are present in it, at varying levels, in more or less recognizable forms: texts of the previous culture and those of the surrounding culture; every text is a fabric of past quotations." This metaphor of fabric, recalling that the etymology of the word "text" refers to the idea of weaving, illustrates that all literary creation is in reality a web of collective contributions. The text, conceived as a complex network of interwoven threads, represents a meeting point where various influences, voices, and cultural traditions interlace.

Intertextuality has been defined differently by successive researchers and can have multiple approaches[1]. The main definition of intertextuality lies in the detection of explicit citation. Literary studies (computational or not) have naturally taken up tracking citations from one text to another to understand phenomena of influence or rewriting in literature ([4], [5], [6]). A second approach to intertextuality, which we will call "weak" to contrast with the first, consists of a simple allusion or the-

---

[1]For an overview of the concept, see Allen [3]'s work.

matic and linguistic similarities between a given text and a set of other texts. Computational methods have enabled the development of this approach in recent years, with examples such as Ganascia [7] who conducted automatic detection of textual fragments that evoke reuse due to their similarity by n-grams. This approach allows for the identification of intertextual relationships between texts that may not be immediately apparent through manual analysis.

In this way, computational literary studies can be extremely useful for analyzing the intertextual network in large corpora of text. Researchers can now identify patterns and trends in the use of intertextuality across different genres, time periods, and cultural contexts. One example of this is the seminal paper by Manjavacas et al. [8], which proposes a thematic and lexical approach to intertextuality. By grouping authors and texts that share topical or lexical similarities, their aim was to evaluate whether it was possible to detect intertextual phenomena from a given source (i.e. the Bible) in a particular corpus. In contrast, our research views potential intertexts as endogenous to the corpus, and we will assess whether the most significant intertexts are also part of the literary canon or not.

Before expanding on our experiment, we need to define two literary notions that may impact intertextuality: the first one is the literary canon. The dynamics of prestige, defined as an author or a book being included in school curricula or reviewed in a prestigious literary journal, can influence textuality. This is because the literary canon represents what is considered to be paradigmatic. Previous computational research has uncovered disparities in textual content between canonical and non-canonical works across various corpora and cultural backgrounds ([9], [10], [11], [12], [13]). These studies have shown that formalist and transcendent definitions of the literary canon seem to be relevant across different genres and time periods. More interestingly, Underwood found an increase in the probability of a work being canonized over time, suggesting that canonical literary works constitute what Altieri [14] terms a "cultural grammar." In other words, canonical works function as foundational texts shaping the norms, values, and conventions within a specific cultural tradition.

The second is the concept of literary genre. According to Genette [15], one of the five dimensions of intertextuality (or transtextuality, as Genette refers to it) is architextuality[2], which is dedicated to literary genres. This dimension refers to the interconnectedness and interdependence of various texts within a literary genre, encompassing "all that sets the text in relationship, whether obvious or hidden, with other texts". Genette considers genre to be a crucial component of a work's overall meaning, arguing that the way in which texts reference and relate to one another can reveal important thematic and structural elements. Jauß [17] supports this argument by introducing the concept of the "horizon of expectations" of the audience, which may lead authors to adhere to certain expected norms and styles. As a result, intertextuality is stronger between texts from the same genres, and genres constitute a structuring element for intertextuality.

These two concepts can be seen as structuring elements of intertextuality. Their impact on writing varies depending on their nature: unconsciously through the authority or influence of certain works, or consciously through literary subgenres. Our goal is to develop a methodology capable of automatically detecting the works and authors that have influenced the evolution of the intertextual network, and to determine which of our two concepts has the greatest impact. To achieve this, we rely on a massive corpus of over 12,000 French-language fiction.

As a proxy for prestigious literature, we will rely on novels that have been republished over time. We were assuming that books that have been republished multiple times have either sold well or are considered important enough to be reread or included in school curricula. The goal is to evaluate the role of canonical works in shaping the literary tradition. Our hypothesis is that the influence of canonical works on the intertextual network is more enduring than the pace of change in the archive, which is heavily represented by texts from popular literary sub-genres. In other words, we expect that canonical works will continue to shape the literary tradition over a longer period of time, while the popularity and influence of the others may be more short-lived.

Outline of the paper

---

[2]For an attempt to operationalize the concept at the passage level using a computational approach, see Barré [16]

The structure of the paper is as follows: We start with a detailed description of the method we used to model intertextuality (section 2), including the corpus description (subsection 2.1), the metadata construction (subsection 2.2) and the operationalization pipeline (subsection 2.3). Then, we present the results in section 3, including an evaluation of the method (subsection 3.1), and an analysis of the main results (subsection 3.2), the individual (subsection 3.3) and collective ones (subsection 3.4). The article concludes with a discussion and the perspectives opened up by this research (section 4).

2. Modeling Intertextuality

2.1. Corpus

Our corpus relies on a subset of the collection "Fictions littéraires de Gallica" [18] which represents 19,240 literary fictions drawn from Gallica, the extensive digitization initiative of the National Library of France. It is made up of works initially categorized as prose literary fiction spanning the period of 1600-1950. The collection's massive scale can be attributed to the efficiency of the French legal deposit system[3]. This legal requirement mandates that publishers submit copies of all published works to the National Library. As a result, it is estimated that approximately 40% [18] of all novels published in France during the 19th century are available in Gallica, based on the BNF catalog. Thus, it offers a representative sample of 19th and early 20th-century literature, encompassing various works such as lesser-known books and subgenres. Figure 1 shows the time distribution of the corpus, pointing the large peak of production in the late nineteenth century, with almost 3.000 novels in the 1890s.
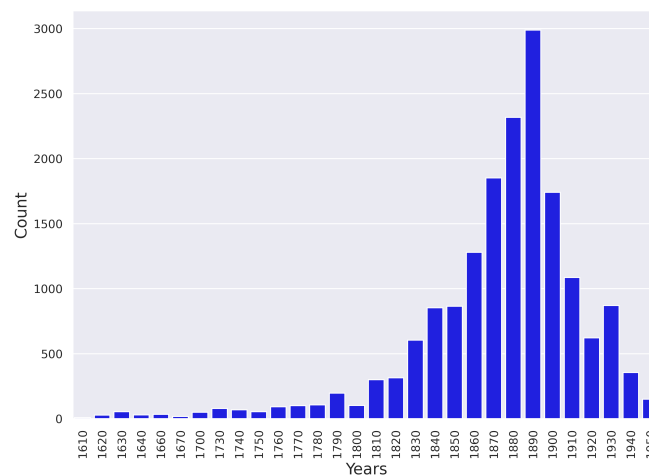

Figure 1: Number of novels over time

In France, texts generally enter the public domain 70 years after the death of their authors. As the crucial date of 1950 approaches, the corpus of novels significantly decreases in size. This is because works published after 1950 are still under copyright, and their full text is not available for use without permission from the rights holders. This limitation affects the overall size and diversity of the corpus, particularly for more recent literature.

The raw corpus contained several issues, such as the optical character recognition (OCR) quality, the presence of complete works from an author, or multiple publications of the same work. To address these issues, we removed all versions of complete works, as our focus was on individual texts. For novels with multiple editions, we selected the first publication to have the closest date associated with a text and its writing date (e.g., 6 editions of Hugo's *Toilers of the Sea* from 1866 to 1894). This allowed us to maintain a more accurate representation of the texts in relation to their original publication dates.
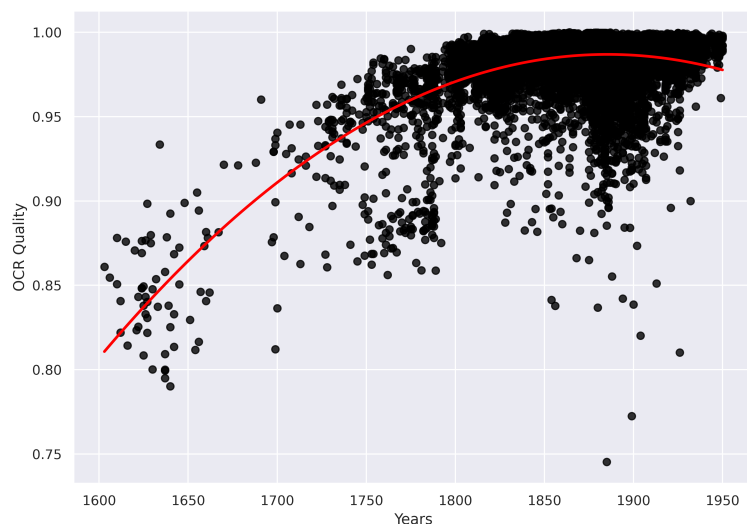
Figure 2: OCR quality over time

Figure 2 shows the temporal distribution of the works in the corpus based on a calculation of the proportion of correct words in each work. The goal is to detect works whose OCR quality is too poor for computational analysis. This evaluation of OCR quality was done by creating a dictionary of French words from a manually corrected set of literary texts. Each word in a given text was then compared to this dictionary, allowing us to compute a proxy for the word error rate metric.

As we can see, the texts before the beginning of the 19th century suffer from rather poor OCR quality or orthographic standardization. On the other hand, the decrease at the end of the century can be explained by a decline in the quality of paper used for certain works. This lower paper quality can negatively impact the OCR process, resulting in a lower overall text quality for these specific works. Filtering the corpus with an estimated OCR higher than 95%, removing multiple publications and complete work from an author, we obtained a list of 12.176 novels.

2.2. Metadata construction

The notion of literary canon or the one of genre are not fixed and stable entities, but rather dynamic and contested constructions that reflects the values and ideologies of specific cultural and historic time. Therefore, it is important to approach these concepts with a critical and historical perspective, taking into account the complex and shifting dynamics of literary production and reception.

Setting aside the whole context of publication and reception of the work, which has an impact on textuality but cannot be recovered, the idea is to reintroduce historical context using proxies for literary reception and events that occurred during the life of the work. The challenge with this approach lies in the difficulty of finding large-scale viable metadata. To address this, we relied on previous research [12] which focused on canonicity in the context of contemporary reception in French literature. We then applied it to our corpus, resulting in 11,202 non canonical elements and 1,083 canonical ones.

For the genre labels, we decided to focus on a single subgenre, as we required clearly defined and coherent labels. We concentrated on "adventure novels" a dominant subgenre in the late nineteenth century. For this purpose, we relied on the work of Letourneux [19], who defines a specific period (1870-1930) for the genre, allowing him to identify its key constants: "The importance of exotic settings […] and the central role attributed to violent action, where the hero faces the risk of death or at least physical peril." We then generated synthetic metadata using a binary SVM classifier [20], trained on Letourneux's labels (with only 102 adventure labels retrieved), and as a result, we identified 2,114 "adventure novel" labels in the corpus under study.

2.3. Operationalizing textuality

This article employs intertextuality as its theoretical foundation, drawing upon the work of structuralists who initially developed the concept. Our aim is to merge their theoretical perspectives with empirical experiments on an extensive collection of literary texts. It can be argued that computational literary studies inherently operationalize intertextuality through quantitative text comparisons. These comparisons may encompass various textual aspects, such as lexical, semantic, or thematic dimensions. However, a challenge arises from the absence of a clear definition of the specific textual components that intertextual theorists refer to.

Computational literary scholars have been striving to develop methods for comparing different texts. To achieve this, they have utilized various techniques to extract information from texts, such as topic-based (LDA, neural models), lexicon-based (BoW), or more semantic comparisons centered on entities or characters [21]. These methods were selected due to their high interpretability, as basic machine learning techniques could determine topic or lexical importance for specific purposes. Nevertheless, certain textual dimensions, like word order or plot progression, have often been disregarded to obtain more interpretable features. Preventing this, the use of text embeddings has developed in the past few years, with methods such as Paragraph Vectors [22] applied to prestige inquiries [23] or genre clustering ones [24]. However, the "black box" nature of these methods has deterred some researchers. At the level of passages, embedding models have shown their strength in retrieving complex textual elements. The primary reason for using these methods is the uncertainty surrounding the specific information we want to retrieve from the texts. Some literary analysis tasks may involve different textual elements, formal or semantic dimensions, and the embedding model is supposedly encoding these informations in its latent space.

In this study, we employ an encoder model based on a transformer language models, specifically the M3-Embedding dense model [25], an open-source model developed by the Beijing Academy of Artificial Intelligence. At the time of the experiment, this model led the multilingual MTEB evaluation benchmark[4]. We also chose this model because we could fine-tune it on French literary language, a crucial factor for our research since many encoder models may lack sufficient French language data in their training corpora, and particularly French literary fiction. This model features an 8192 tokens window, which is significantly larger than that of a more traditional BERT (512), allowing us to process a wide range of text lengths. This extended window size enables the model to capture more contextual information, making it more suitable for analyzing longer parts of literary works.

Consequently, we fine-tuned our model on a passage similarity task. We constructed the training corpora by selecting 400,000 random passages from our corpus. The "query" resulted in a paragraph of ten sentences. For the "positive" relation with the query, we used the ten subsequent sentences. As the "negative" relation, we chose ten random sentences from the entire corpus. The underlying assumption is that authors maintain a consistent language in their novels, especially when dealing with consecutive passages. We also replaced all proper names with a specific token "[PROPN]" to prevent the encoder model to cluster passages only based on character names. Proper names of characters have been extracted using BookNLP-fr [27], a state-of-the-art NER pipeline for literary entities. The model will iteratively enhance its performance on the author attribution task by leveraging various cues, including formal elements, thematic content, and stylistic features. As it undergoes training, it is expected to develop a deeper understanding of these characteristics, allowing it to more accurately identify the unique voice and style of different texts and authors.

After the fine-tuning, the goal was to infer a vector representation for each novel in our corpus. Then, a challenge arises, as our encoder can only process a context window of 8096 tokens, while a typical novel contains around 100,000 tokens. We implemented the following approach to handle this crucial step. For each novel in our corpus, we first randomly draw 100 passages of it, then we run our fine tuned encoder on each passage. Finally we take the mean embedding of all passages from a novel

---

to represent it as a unique embedding. Thus we obtained 12.176 embeddings, one for each novel. As a distance metric, we opted for the cosine similarity, since it is widely used in the NLP and CLS fields. Previous work managed to show that the cosine similarity between pairs of word embeddings had a robust correlation with human similarity judgments [28]. However, when applying cosine distance to embeddings drawn from contextualized language models like ours, previous research found that a small number of "rogue dimensions" had a disproportionate impact on the cosine similarity calculation [29], [30]. To prevent this issue we ran a Standard Scaler normalization before running the cosine distance between every pair of novels from our corpus.

3. Results

3.1. Sanity check

To validate our approach at each computational step, we implemented a sanity check using a subset of one thousand novels of our corpus. For each novel, we generated five distinct random representations by selecting random passages, computing their vector representations, and averaging the resulting vectors. This produced five mean embeddings for each novel, resulting in a total of 5000 embeddings. The goal of the sanity check is to ensure that, for any given embedding, the four most similar embeddings (based on cosine similarity) correspond to the other versions of the same novel. If the method fails to identify the four closest versions correctly, we apply a penalty to the accuracy score. This test verifies whether our process of randomizing passages, vectorizing them, and averaging the vectors maintains or not the integrity of each novel's representation.
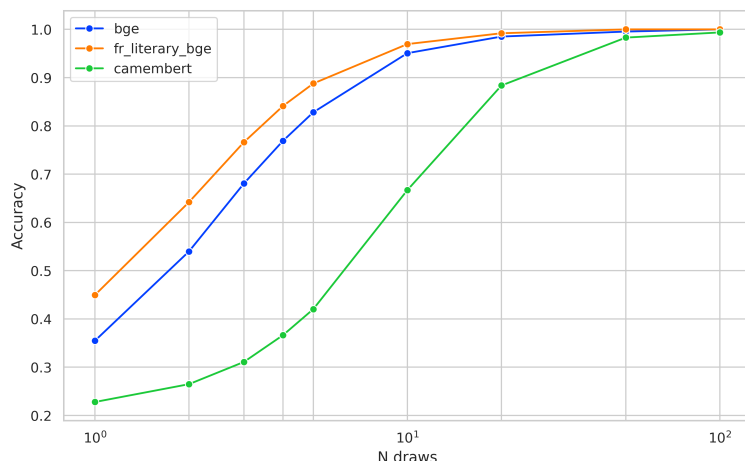
**Figure 3:** Sanity check accuracies between models and number of draws in novels

Figure 3 compares for each number of draws the performance of the base model BGE-M3 and the fine-tuned version. It shows that the fine-tuned model consistently outperforms the base model, and that the number of draws is halved to complete the sanity check (50 vs 100). This improvement highlights the effectiveness of fine-tuning, enabling an optimization of resource consumption at inference which is crucial when planning to launch the model on 12,000 novels.

Overall, these results suggest that our computational approach is effective at capturing the complexity of what constitutes a novel, while also being robust to the randomization of selected passages of the works. The evaluation we implemented turned out to be relatively easy for the model, as a 100% accuracy was achieved quickly. However, this allowed us to determine the number of random draws needed to capture the complexity and uniqueness of a given novel.

3.2. Similarity over time

After retrieving the embeddings for each of our approximately 12,000 novels, a similarity matrix is constructed, with as many rows as columns, where we measure the cosine similarity of each text with all the others. On each row, works by the same author are set to NaN to prevent the authorial idiolect from skewing the temporal analyses. Next, a similarity matrix is computed for each text with every year in the corpus. To ensure statistical relevance, we apply downsampling to each year, meaning that every year must have a minimum of 25 novels and a maximum of 50 novels. Each text is then aligned with its publication year, centering the analysis at 0 for its date of publication, resulting in the graph presented in Figure 4. This downsampling process is repeated ten times, and the standard error is displayed in the graph. We chose 30 years before and after because previous research showed it was a good window for measuring literary change[5]. A larger window would have also resulted in a significant loss of data in the analysis.
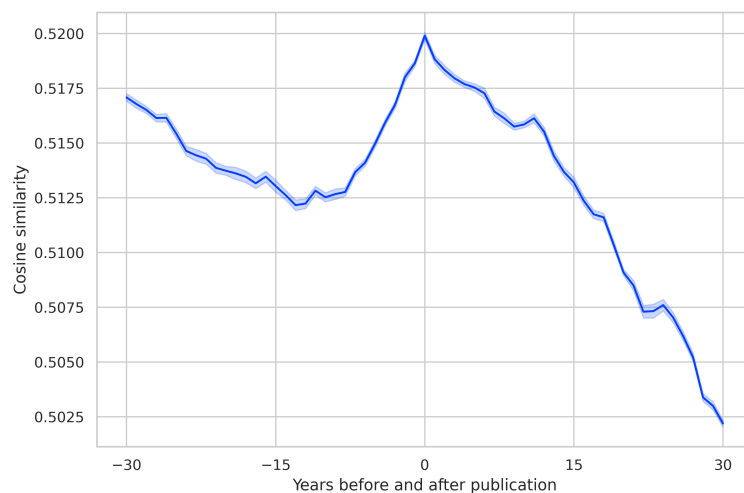
Figure 4: Similarity between texts considering dates before and after publication

The resulting graph measures the similarity between each text and its context of production. A striking pattern emerges: a clear peak in similarity is observed during the year of the novel's publication. As the distance from the publication year increases, the textual similarity diminishes. This supports clearly previous research [33], where researchers provided quantitative evidence for the concept of a literary "style of the time", highlighting a strong trend toward more contemporary stylistic influences.

Surprisingly, before the similarity peak at the time of publication, there is a noticeable downward trend in the cosine similarity between -30 and -15 years before publication. The high similarity 30 years before publication suggests that earlier works had a strong influence on the author's development. This influence decreases as the publication date approaches, reflecting the gradual divergence from older literary models. The trend aligns with previous research on the cohort effect, where an author's writing is shaped by books from their formative years. However, this goes beyond the analysis of this paper, and we have no evidence to support this hypothesis here. Underwood et al. [31] provided strong evidence of cohorts driving literary change, though their work did not specifically explore their effect on textual similarity.
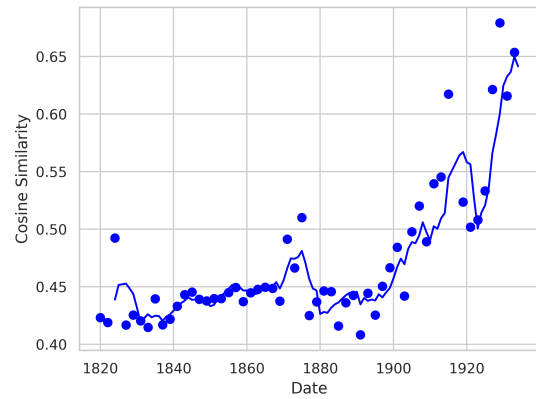
3.3. Individual similarity over time

Then, we wanted to grasp the individual trends of specific novels to investigate how they move in the textual similarity network. For these experiments, we took back the downsampled similarity matrix
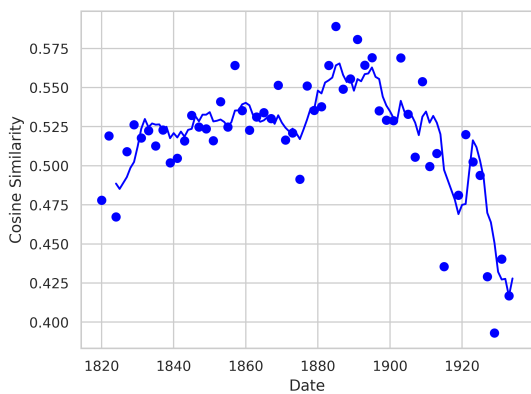
---

[5]See [31] and Moretti [32]'s work on genres, their analysis focused on the cycle of change, considering a timeframe of 25 to 30 years.
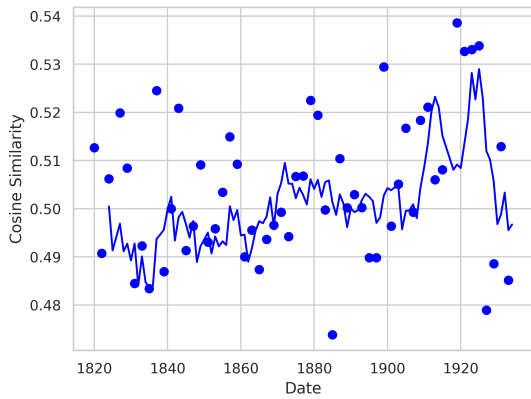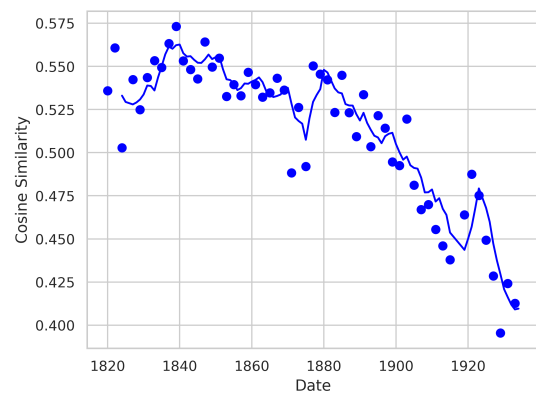
(a) Cosine similarity over time for Proust's *Le temps retrouvé*

(b) Cosine similarity over time for Leroux's *Le mystère de la chambre jaune*

(c) Cosine similarity over time for Hugo's *Les Misérables*

(d) Cosine similarity over time for Reybaud's *Mézélie*

Figure 5: Different patterns of individual textual similarity evolution over time

computed for each text with every year in the corpus. Figure 5 represents four different patterns of textual similarity over time. Figure 5a shows the cosine similarity plot for Proust's *Le temps retrouvé*, the final volume of his monumental *À la recherche du temps perdu*. The graph reveals a significant peak around the time of its post-mortem publication in 1927. The detected pattern is not the author's signature, as texts by the same author are excluded from the analysis. Instead, it reflects a kind of "average language" or a broader intertextual network at the linguistic level. This peak aligns deeply with the modernist focus on subjectivity and inner consciousness. Like many modernist writers of the time, Proust was influenced by emerging psychological theories, especially those of Sigmund Freud, which emphasized the exploration of the unconscious mind. In *Le temps retrouvé*, Proust explores the protagonist's memories, perceptions, and internal reflections, creating a narrative that is more concerned with the flow of time as experienced subjectively than with external events or linear plot progression. Proust's plot exhibits the highest level of noise among the four. This can be explained by insights from Compagnon [34]'s work which provides a framework for understanding how Proust's *À la recherche du temps perdu* reflects a dialogue between past literary forms and the new modernist literary movement that was gaining prominence in the early 20th century. This concept of Proust being "between two centuries" helps explain why Proust's work resonates with both 19th-century novelists and modernist writers, which could explain the disparities in similarity across time.

Figure 5b shows the similarity over time of Leroux's *Le mystère de la chambre jaune*, published in 1907. A huge increase in cosine similarity appears from the beginning of the twentieth century to the end of the observed time period. This reflects perfectly the novel's status as a "proto-detective"

story [35] that laid the groundwork for the *whodunit* detective genre. This peak can be attributed to the rise of detective fiction following Leroux's or Doyle's innovations, with authors building on their narrative techniques and conventions. As detective stories proliferated in the early 20th century, the post-publication similarity of *Le mystère de la chambre jaune* became increasingly apparent.

In Figure 5c, Victor Hugo's *Les Misérables* offers a more intriguing pattern. If a slight peak around its publication date in 1862 appears, it is much less pronounced than the decades that follow between 1880 and 1900. This could be explained by the growing canonical status of both Hugo and *Les Misérables* influencing subsequent generations of writers who either drew inspiration from its themes or echoed its narrative structures. The vivid depictions of social injustice, poverty, and the struggles of the lower classes in *Les Misérables* resonate with key themes of Naturalism, despite the fact that they will take this social critique further, focusing on deterministic views of human behavior influenced by environment and heredity. The drop at the end of the graph is likely due to linguistic and thematic changes that are too significant to maintain high textual similarity.

The last individual trend in Figure 5d is *Mézélie*, published in 1839 by Madame Charles Reybaud. The graph exhibits a gradual decline in cosine similarity over time. Despite her notable success and popularity during her time, Reybaud's works, including *Mézélie*, have largely faded into obscurity and is now part of the so called "archive".

3.4. Collective Structuring Frames of Textual Similarity

Our final experiments use the same type of analysis as in subsection 3.2, but we aimed to explore the extent to which collective factors, such as genres or canonicity, could influence the different trends in textual observed earlier. To that end we relied on metadata described in subsection 2.2. To compare our samples (canon vs archive or adventure vs general) with equal temporal cardinality, we implemented stratified sampling. We randomly selected the appropriate number of elements from the general and non-canon samples to match those in the adventure and canonical sets, while also maintaining the temporal distribution per decades of these specific groups.
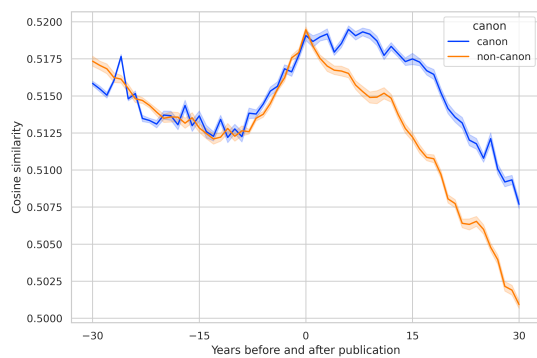
(a) Similarity between texts considering dates before and after publication, Canon in blue, Archive in orange
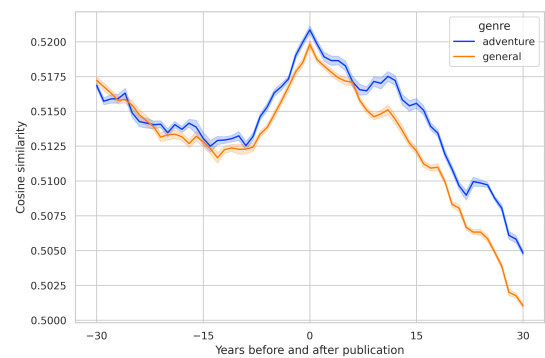
(b) Similarity between texts considering dates before and after publication, Adventure in blue, in orange

Figure 6: Different patterns of collective textual similarity evolution over time

Figure 6a shows that while both canonical and archival novels follow a similar similarity trend before publication, the curves diverge notably after the publication date. The pattern is particularly striking, suggesting that canonical works tend to exhibit higher similarity scores with later publications compared to archival works. This trend persists over time, indicating that canonical works exert a more enduring influence on the intertextual network, whereas non-canonical works have a more limited and shorter-lived impact.

Figure 6b shows a less striking pattern but remains interesting and statistically solid. Both curves follow a similar trend before publication, yet they begin to diverge approximately 10 years before and

up to 5 years after the publication date. This pattern suggests that genres have a precise contextual existence within a specific historical moment. Following the decline in similarity, we observe that adventure novels maintain a stronger similarity with later texts (from +10 to +30 years). This is harder to interpret, but one hypothesis could be that many works within the genre were canonized retrospectively, which, as seen in Figure 6a results in a higher similarity with subsequent works.

4. Discussion

In this article, we demonstrated that both subgenres and canonicity function as collective structuring frames for textuality. While our operationalization using language model encoders and cosine similarity is undoubtedly limited in its ability to fully capture the complexity of a novel, our approach nonetheless uncovered novel patterns of similarity within French fiction.

Firstly, we showed that canonical works tend to be more deeply integrated into the intertextual network after their publication. This investigation sheds light on the collective representations that shape the cultural frameworks within which we operate. One way to understand this phenomenon is through the concept of "cultural grammar," as proposed by Altieri [14]. According to this view, canonical literary works serve as foundational texts that help to shape the norms, values, and conventions within a particular cultural tradition.

Secondly, we highlighted that texts within the same genre share distinct textual similarities during the relatively brief period when the genre is being defined. This could be attributed to the commercial nature of genre, which contributes to increasing the similarity of certain texts within a specific historical moment.

Ultimately, both canonical and subgenres works establish a set of shared references and expectations that guide the production and reception of subsequent texts, and contribute to the formation of a shared cultural imagination.

Narrowing the scope of study appears essential for future research in two ways: reducing the number of works analyzed and focusing on passages rather than entire novels. This will help to reestablish a more concrete foundation based on textual evidence. It would be interesting for example, to examine these dynamics at the scale of a specific sub-genre, taking the example of detective novels in order to track the dynamics of intertextuality within a highly codified and well-defined sub-genre. This will also provide an opportunity to test the distant reading observations on the subgenre's canonical novels, to determine whether the phenomenon persists at a smaller scale.

Acknowledgments

References

[1] J. Kristeva, Sèméiotikè: recherches pour une sémanalyse, Points, Éditions Point, Paris, 2017.

[2] R. Barthes, Texte (théorie du), Encyclopædia Universalis (1974). URL: https://www.universalis-edu.com/encyclopedie/theorie-du-texte/.

[3] G. Allen, Intertextuality, The new critical idiom, third edition ed., Routledge, Taylor & Francis Group, 2022.

[4] T. Allen, C. Cooney, S. Douard, R. Horton, R. Morrisse, M. Olsen, G. Roe, R. Voyer, Plundering Philosophers: Identifying Sources of the Encyclopédie, Journal of the Association for History and Computing (2010). URL: http://hdl.handle.net/2027/spo.3310410.0013.107.

[5] M. Büchler, G. Crane, M. Moritz, A. Babeu, Increasing Recall for Text Re-use in Historical Documents to Support Research in the Humanities, Springer Berlin Heidelberg, 2012, p. 95−100. doi:10.1007/978-3-642-33290-6_11.

[6] J.-G. Ganascia, P. Glaudes, A. Del Lungo, Automatic detection of reuses and citations in literary texts, 2014. doi:10.48550/ARXIV.1404.2997.

[7] J.-G. Ganascia, Détection automatique de phénomènes intertextuels, Genesis (2020) 63−77. doi:10.4000/genesis.5671.

[8] E. Manjavacas, F. Karsdorp, M. Kestemont, A Statistical Foray into Contextual Aspects of Intertextuality, volume 2723, CEUR Workshop Proceedings, 2020, pp. 77−96.

[9] 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.

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

[11] J. Brottrager, A. Stahl, A. Arslan, U. Brandes, T. Weitin, Modeling and predicting literary reception. a data-rich approach to literary historical reception, Journal of Computational Literary Studies 1 (2022). doi:10.48694/jcls.95, publisher: Universitäts- und Landesbibliothek Darmstadt.

[12] J. Barré, J.-B. Camps, T. Poibeau, Operationalizing canonicity: A quantitative study of french 19th and 20th century literature, Journal of Cultural Analytics 8 (2023). doi:10.22148/001c.88113.

[13] J. Barré, T. Poibeau, Beyond canonicity: Modeling canon/archive literary change in french fiction, in: CEUR Workshop Proceedings CHR2023, 2023, pp. 814−830.

[14] C. Altieri, An idea and ideal of a literary canon, Critical Inquiry 10 (1983) 37−60. doi:10.1086/448236.

[15] G. Genette, Introduction à l'architexte, in: G. Genette, T. Todorov (Eds.), Théorie des genres, number 181 in Points, Éd. du Seuil, 1986, pp. 110−148.

[16] J. Barré, Détection automatique de l'architextualité dans le roman d'aventures, in: Humanistica 2024, Stylométrie, Association francophone des humanités numériques, Meknès, Morocco, 2024. URL: https://hal.science/hal-04559749.

[17] H. R. Jauß, Toward an aesthetic of reception, number 2 in Theory and history of literature, nachdr. ed., Univ. of Minnesota Press, 2010.

[18] P.-C. Langlais, Fictions littéraires de Gallica / Literary fictions of Gallica, 2021. doi:10.5281/zenodo.4751204.

[19] M. Letourneux, Le roman d'aventures: 1870-1930, Presses Universitaires de Limoges et du Limousin, 2010.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825−2830.

[21] L. Kohlmeyer, T. Repke, R. Krestel, Novel views on novels: Embedding multiple facets of long texts, 2021 Association for Computing Machinery. (2021).

[22] Q. V. Le, T. Mikolov, Distributed representations of sentences and documents, 2014. URL: https://arxiv.org/abs/1405.4053. arXiv:1405.4053.

[23] A. van Cranenburgh, K. van Dalen-Oskam, J. van Zundert, Vector space explorations of literary language, Language Resources and Evaluation 53 (2019) 625−650. doi:10.1007/s10579-018-09442-4.

[24] O. Sobchuk, A. Šeļa, Computational thematics: Comparing algorithms for clustering the genres of literary fiction, 2023. arXiv:2305.11251.

[25] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. arXiv:2402.03216.

[26] N. Muennighoff, N. Tazi, L. Magne, N. Reimers, MTEB: Massive Text Embedding Benchmark, 2022.

URL: https://arxiv.org/abs/2210.07316v3.

[27] F. Mélanie-Becquet, J. Barré, O. Seminck, C. Plancq, M. Naguib, M. Pastor, T. Poibeau, BookNLP-fr, the French Versant of BookNLP. A Tailored Pipeline for 19th and 20th Century French Literature, Journal of computational literary studies (2024). URL: https://tuprints.ulb.tu-darmstadt.de/id/eprint/27396. doi:10.26083/TUPRINTS-00027396.

[28] B. Wang, A. Wang, F. Chen, Y. Wang, C.-C. J. Kuo, Evaluating word embedding models: methods and experimental results, APSIPA Transactions on Signal and Information Processing 8 (2019) –. doi:10.1017/ATSIP.2019.12.

[29] W. Timkey, M. van Schijndel, All bark and no bite: Rogue dimensions in transformer language models obscure representational quality, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 4527–4546. doi:10.18653/v1/2021.emnlp-main.372.

[30] J. Mu, P. Viswanath, All-but-the-top: Simple and effective postprocessing for word representations, 2018. URL: https://openreview.net/forum?id=HkuGJ3kCb.

[31] T. Underwood, K. Kiley, W. Shang, S. Vaisey, Cohort succession explains most change in literary culture, Sociological Science 9 (2022) 184–205. doi:10.15195/v9.a8.

[32] F. Moretti, Graphs, maps, trees: abstract models for literary history, Verso, 2007.

[33] J. M. Hughes, N. J. Foti, D. C. Krakauer, D. N. Rockmore, Quantitative patterns of stylistic influence in the evolution of literature, Proceedings of the National Academy of Sciences 109 (2012) 7682–7686. URL: http://dx.doi.org/10.1073/pnas.1115407109. doi:10.1073/pnas.1115407109.

[34] A. Compagnon, Proust entre deux siècles, Éd. du Seuil, Paris, 2013.

[35] E. d. Lavergne, La naissance du roman policier français: du Second Empire à la Première Guerre mondiale, number 7 in Études de littérature des XXe et XXIe siècles, Classiques Garnier, Paris, 2009. OCLC: ocn436637927.

[36] Jean Barré, fr_literary_bge_base, 2024. doi:10.57967/HF/3255.

A. Data, code and model availability

We have made the data and code available on GitHub[6] and released the fine-tuned model and corpus on HuggingFace.[7]

B. Finetuning supplement

Figure 7 shows a rapid decrease in loss at the beginning of the training, from 1.3 to around 0.5 during the first epoch, indicating that the model is quickly learning to distinguish positive passages (subsequent passage) from negative ones (random passage). After this initial drop, the loss decreases more slowly and stabilizes around 0.4 after 2 epochs, suggesting the model is converging. The small fluctuations are likely due to the stochastic nature of the gradient updates, as mini-batches vary in content.

---

[6]https://github.com/crazyjeannot/CHR_latent_structures
[7]https://huggingface.co/crazyjeannot/fr_literary_bge_base[36]
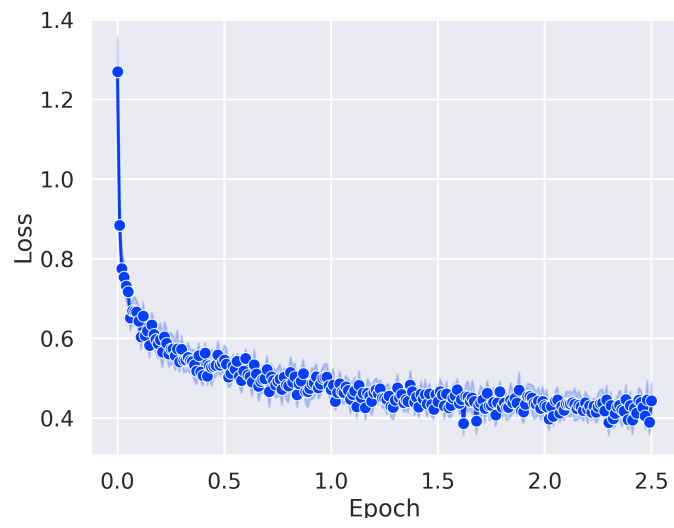
Figure 7: Loss fluctuations during finetuning