



Conference Reader
3rd Annual Conference of
Computational Literary Studies
CCLS 2024 Vienna
June 13-14, 2024

updated version from June 18, 2024

Venue:	Haus der Musik Seilerstätte 30 1010 Vienna
Local Organizer:	Austrian Centre for Digital Humanities and Cultural Heritage at OeAW
Contact:	acdch-events@oeaw.ac.at
Hashtag:	#CCLS2024

Conference Programme

Thursday | June 13, 2024

1:00 p.m. to 1:30 p.m. | Opening

1:30 p.m. to 3:00 p.m. | Session 1 (Chair: Svenja Guhr)

- Daniel Brodén, Jonas Ingvarsson, Lina Samuelsson, Victor Wåhlstrand Skärström:
Visualization as Defamiliarization: Mixed-Methods Approaches to Historical Book Reviews
- Pascale Feldkamp, Yuri Bizzoni, Ida Marie S. Lassen, Mads Rosendahl Thomsen, Kristoffer L. Nielbo: **Measuring Literary Quality. Proxies and Perspectives**
- Marijn Koolen, Joris van Zundert, Eva Viviani, Carsten Schnober, Willem van Hage, Katja Tereshko: **From Review to Genre to Novel and Back. An Attempt To Relate Reader Impact to Phenomena of Novel Text**

3:30 p.m. to 4:30 p.m. | Session 2 (Chair: Élodie Ripoll)

- Frédérique Mélanie-Becquet, Jean Barré, Olga Seminck, Clément Plancq, Marco Naguib, Martial Pastor, Thierry Poibeau: **BookNLP-fr, the French Versant of BookNLP. A Tailored Pipeline for 19th and 20th Century French Literature**
- Matthew Wilkens, Elizabeth F. Evans, Sandeep Soni, David Bamman, Andrew Piper: **Small Worlds. Measuring the Mobility of Characters in English-Language Fiction**

5:00 p.m. to 6:00 p.m. | Keynote

- Maciej Eder: **Text Analysis Made Simple (Kind of), or Ten Years of Stylo (Abstract)**

7:00 p.m. | Conference Dinner

Friday | June 14, 2024

9:30 a.m. to 10:30 a.m. | Session 3 (Chair: Daniil Skorinkin)

- Paschalis Agapitos, Andreas van Cranenburgh: **A Stylometric Analysis of Seneca's Disputed plays. Authorship Verification of "Octavia" and "Hercules Oetaeus"**
- Botond Szemes, Mihály Nagy: **Repetition and Innovation in Dramatic Texts. An Attempt to Measure the Degree of Novelty in Character's Speech**

11:00 a.m. to 12:00 p.m. | Session 4 (Chair: Henny Sluyter-Gäthje)




- Erik Ketzan, Martin Eve: **The Anxiety of Prestige in Stephen King's Stylistics**
- Benjamin Gittel, Florian Barth, Tillmann Dönicke, Luisa Gödeke, Thorben Schomacker, Hanna Varachkina, Anna Mareike Weimer, Anke Holler, Caroline Sporleder: **Neither Telling nor Describing. Reflective Passages and Perceived Reflectiveness 1700-1945**

12:00 p.m. to 12:30 p.m. | Closing

Visualization as Defamiliarization

Mixed-Methods Approaches to Historical Book Reviews

Daniel Brodén¹ 
 Jonas Ingvarsson¹ 
 Lina Samuelsson² 
 Victor Wählstrand Skärström³ 

1. Department of Literature, History of Ideas and Religion, University of Gothenburg , Gothenburg, Sweden.
2. School of Education, Culture and Communication, Division of Language and Literature, Mälardalen University , Eskilstuna, Sweden.
3. Department of Electrical Engineering, Chalmers University of Technology , Gothenburg, Sweden.

Citation

Daniel Brodén, Jonas Ingvarsson, Lina Samuelsson, and Victor Wählstrand Skärström (2024). "Visualization as Defamiliarization. Mixed-Methods Approaches to Historical Book Reviews". In: *CCL52024 Conference Preprints* 3 (1). [10.26083/tuprints-00027397](https://doi.org/10.26083/tuprints-00027397)

Date published 2024-05-28

Date accepted 2024-04-04

Date received 2024-01-25

Keywords

book reviews, mixed methods, visualizations, close re-reading, digital humanities, defamiliarization

License

CC BY 4.0 

Reviewers

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 3rd Annual Conference of Computational Literary Studies at Vienna, Austria, in June 2024.

Abstract. This paper employs a dialectical mixed methods approach to revisit a previous study in comparative literature on discourses in literary criticism, using data visualizations to analyze the original material, 700 digitized literary book reviews from the years 1906, 1956, and 2006. The aim is to explore alternative ways of understanding the review material by comparatively examining visualizations on word and sentence levels, publication years, and genre categorizations. In the paper, we discuss significant patterns that emerge in the visualizations and how a combination of computational and interpretative analysis provide complementary perspectives on the text collection. Furthermore, drawing upon Russian formalist Viktor Shklovsky, we suggest the notion of "defamiliarization" as a conceptual framework for the process of looking at familiar research material anew through the lens of visualization, potentially uncovering previously overlooked aspects of the data. We conclude by stressing the criticality of a contextual sensibility for understanding the visualizations.

1. Background

In the study "The Order of Criticism: Swedish Book Reviews in 1906, 1956, 2006" (*Kritikens ordning: Svenska bokrecensioner 1906, 1956, 2006*) from 2013, literary scholar Lina Samuelsson analyzed what characterized literary criticism as an institution and practice, mapping dominant themes, values and discourses, at different points in time.¹ Combining a sociological and historical perspective with a Foucauldian discourse analysis, the study traced what has historically constituted a literary book review and what norms literary reviewers followed at different points in time.²

The current research project "The New Order of Criticism: A Mixed-Methods Study of 150 Years of Book Reviews in Sweden," repeats, extends and challenges the original

1. Samuelsson 2013. Since Samuelsson's study is cited repeatedly in the following, references will be made with page numbers in brackets.

2. Samuelsson examines what Foucault refers to as a "discursive practice," i.e., the "anonymous, historical rules, always determined in the time and space that have defined a given period, and for a given social, economic, geographical, or linguistic area, the conditions of operation of the enunciative function." Foucault 1972, 117. See also Samuelsson 2013, 11

study (Samuelsson being a member of the project team), drawing upon data-driven approaches to explore how “traditional” and “digital” methods can contribute to enhancing each other, both in practical and epistemological terms.³ Thus, the project ties into the ongoing critical discussion in digital humanities about the need for integrative interdisciplinary approaches and to reflect on the positivist claims made within the field (Moretti 2013; Jockers 2013). As digital historian Jo Guldi argues, without the insights of the humanities, data-driven approaches risk producing analyses that are empty or misleading. According to Guldi, data-intensive analysis lacking a historical sensibility and an awareness of the data’s original context often raises more questions than it answers (Guldi 2023, 1, 27, 83). Turning the argument around on proponents of the presumed scientificity of distant reading and macro analysis, digital literary historian Katherine Bode suggests that an exclusive focus on textual signals could be understood merely as an enactment of a de-contextualised understanding of text as data, emphasizing that aggregating text data involves a stripping of context (Bode 2018; Berry and Fagerjord 2017; Dobson 2019). Consequently, Bode argues for the importance of an interpretative and contextual understanding of both the data and the results.⁴

In this paper, we revisit the review material that the original study, ‘The Order of Criticism’, was based on from a mixed methods perspective to discuss the possibility of an analytical interplay between data visualization and close reading. Rather than engaging in the debate concerning the prerequisites of data as evidence or the need for criticality when creating data visualizations, we explore the possibility of discovering alternative ways of looking at a particular material through a dialectical mixed methods approach. Thus, in this particular context, we are less interested in evaluating the original study or interrogating the creation of the visualizations (nor the methodology of the original discourse analysis), than exploring how data-driven and interpretative methods can provide complementary analytical perspectives on a text collection, focusing on significant data patterns that emerge in visualizations and comparing them with the original analysis. Essentially, our discussion will emphasize performative and interpretative affordances of the visualizations rather than computational aspects (Bode 2020).

In total, the original study, *The Order of Criticism*, was based on 700 book reviews, which can be considered a rather substantial material for a ‘traditional’ literary history study, even though it can be considered a small dataset in a digital humanities context.⁵ However, in digital humanities, data-driven analyses of literary criticism and reception have been performed on less extensive but more curated datasets and, notably, the collection used for *The Order of Criticism* exceeds for instance the two corpora of English and German historical book reviews (605 and 547, respectively) from the long 18th and 19th century created by Brottrager et al. for automated sentiment detection (Brottrager

3. When we state that we want to “challenge” the results from the previous study, it means that we do not take for granted what results the digital analyses will generate. If the observations of the original study are confirmed by the digital methods, it is equally interesting from an epistemological perspective as if the data-driven methods lead to different conclusions or hypotheses. Regardless, it ultimately pertains to methodological discussions, and why the results turn out as they do. See Ingvarsson et al. 2022, where we also present an overview of the project’s main tasks.

4. For discussions on the epistemological consequences of digitalization for the humanities, see for example Bode 2018, 5 and 17-36; Bode 2023; Liu 2014; and Ingvarsson 2021, 1-28.

5. A note on the translation of Swedish titles: the first time the title is mentioned, an English translation is presented immediately after, in brackets. If there is an existing English title it will first be displayed in italics, still in brackets. For recurring references, and for the readability of the text, the English translation is used in italics, even though the text doesn’t exist in an English version.

et al. 2022).

To delineate our approach, we begin by situating our study within the field of mixed methods and highlighting our dialectical approach, emphasizing that while so-called quantitative and qualitative methods tend to generate different results, they can nevertheless be intermingled, making the answer to a research question more complex and flexible. We then describe the process of generating text data visualizations based on the book reviews originally investigated in *The Order of Criticism*, using TF-IDF (Term Frequency – Inverse Document Frequency) and an interface developed within our current project (<https://dh.gu.se/kno/>). Turning to the analysis, we examine data visualizations of word frequencies, publication years, and genre categorizations, respectively, in the review material from the original study, focusing on results that raise questions in relation to the prior results concerning the literary discourse in 1906, 1956 and 2006. The analysis leads up to a concluding discussion about the criticality of a contextual sensibility for understanding how we can analyze text data visualizations, but also the possibility of attributing an estranging quality to them. Drawing upon Russian formalist Viktor Shklovsky, we suggest the concept of *defamiliarization* (*priëm ostraneniya*) as a conceptual framework for understanding the process of being able to look anew at a seemingly familiar research material ("the already analyzed") through the lens of visualizations, potentially turning the analytical gaze toward overlooked aspects (Shklovsky 1990 (1929)).

2. Mixed Methods – Pragmatic and Dialectical Approaches

In digital humanities, there is a growing interest in critical reflection on "what is happening" or "what should happen" at the concrete intersections between data-driven and interpretative methods (Ahnert et al. 2023). Concerning data-intensive studies of newspaper data and literary criticism, the discussion has primarily revolved around the future potential of computational methods and productive approaches, rather than the very nature of interdisciplinary syntheses (Underwood 2018; Piper 2020). Only in recent years there has appeared a clearly articulated theoretical interest within digital humanities in developing a more organic interdisciplinarity with integrated workflows and there remains a lack of systematic reflection on the relationship between different interdisciplinary and methodological syntheses (Oberbichler et al. 2021).

However, such modes of reflection can be found within the field of mixed methods that centers on the creation and reflection of syntheses between quantitative and qualitative approaches (Johnson et al. 2007; J. W. Creswell and J. D. Creswell 2022). Much of the research practices associated with mixed methods are, of course, not necessarily "new", but the field has nevertheless come to serve as a distinct space for self-reflexive discussion. According to philosopher Yafeng Shan, the heterogeneous field of mixed methods can be discussed at various levels in scientific practice, including material selection, method selection, research purpose, and epistemology (a method's epistemological implications) (Shan 2023). Shan further identifies a number of fundamental approaches to mixed methods, including a *pragmatic* and a *dialectical* approach, which can be used to frame our study (Shan 2023, 3–4).

From a pragmatic standpoint, researchers (individually or in groups) are free to use

the method – quantitative or qualitative – that they believe best suits their task without considering one method a priori better than the other. Shan sees this as a “weaker” category insofar as the pragmatic position is open to the possibility of integrating quantitative and qualitative methods without necessitating their combination (Shan 2023, 6–8). Somewhat akin to the pragmatic stance is the dialectical one. Here, the different epistemological approaches underlying quantitative and qualitative methods are also accepted, but it is emphasized that they lead to different results. Thus, it is not just about choosing the method that “works best,” but also about accepting that different methods complement each other due to their distinct epistemological consequences. Adopting different perspectives makes the answer to a research question more complex and flexible. Therefore, Shan understands the dialectical approach as a “strong” category of mixed methods because it starts from the premise that research questions cannot be answered by only one quantitative or qualitative method, but are better understood by combining them (Shan 2023, 8).

Our investigation is based on the stronger, dialectical mixed methods approach. In digital humanities the rhetoric about computer-assisted analyses leading to more “objective” knowledge and a higher degree of “scientificity” has been prominent up until more recently, when we have partly seen a shift toward more epistemologically reflective stances. Our study is, thus, influenced by what Geoffrey Rockwell and Stéfan Sinclair call a dialogical collaboration between humanities researchers and data analysts, within which “[s]mall experiments generate hermeneutical theories as the products of interpretation: texts and tools”, and “[m]ethods, and their instantiation in tools, are discussed reflexively throughout the experiment” (Rockwell and Sinclair 2016, 8; see also Nelson 2020, 3–42). However, Shan furthermore points to an axiological dimension of mixed methods regarding questions of value or use (Shan 2023, 3 and 5). In our case, this is primarily about how traditional and digital methods can complement each other and, working together, enrich the understanding of literary criticism in Sweden. As noted above, rather than problematizing the quantitative method underlying the visualizations, we primarily seek to explore a way in which visualizations of previously researched material can make way for renewed close reading of the texts in focus. Thus, we will primarily treat the visualizations as a vehicle for defamiliarization to provide a modelled overview of a certain material, proceeding on the assumption that the encounter between a traditional analysis and data visualization may prove productive on different levels.

3. Data Visualizations

Emphasizing the rhetorical power of data visualizations, Johanna Drucker asserts that they always involve calculations that are graphically represented to communicate specific aspects of the underlying data (Drucker 2021, 86). In our case, data visualizations create a multi-dimensional “map” of various relationships between book reviews based on their linguistic characteristics at both the word and sentence levels. By studying these visualizations, we can explore the potential of a quantifying method to elucidate significant patterns in the texts in comparison with a prior study based on the same material. Consequently, we are primarily interested in patterns in the visualizations that go against our expectations based on previous results. In this, we are inspired by

Andrew Piper's and Mark Algee-Hewitt's work on the creation of topological models for visualising the lexical relationality between Goethe's *The Sorrows of Young Werther* and the author's oeuvre, bringing into view textual relationships through the form of the diagram (Piper and Algee-Hewitt 2014). Reading "words in space", rather than within sentences, as Piper and Algee-Hewitt put it, allows them to bring to light "the latency of the lexically manifest" or the potential "meaning of the distributed recurrences of language that can easily escape our critical consciousness," provoking new close readings of Goethe's texts (Piper and Algee-Hewitt 2014, 157 and passim).

In *The Order of Criticism*, 700 literary book reviews from newspapers and periodicals were examined to provide a systematic and fairly representative sample of literary criticism for the years 1906, 1956, and 2006. Each year was studied through two delimited samples that provided the study with roughly the same number of reviews from each year (198, 272 and 230 reviews from 1906, 1956 and 2006, respectively). In 1906, the samples were based on one month in spring and one month in autumn, and in 1956 and 2006, one week each in spring and autumn. While one of the aims in our current research project is to determine whether this sampling of book reviews is in fact representative (using text mining of reviews in newspaper collection of the National Library of Sweden (Kungliga Biblioteket, KB)), in the present paper we will stick with the original selection for comparative purposes.⁶

Methodologically, the study took inspiration from the so-called year study method, meaning that the reviews were analyzed from a synchronic rather than a diachronic perspective, without aligning them into a continuous historical account or "narrative", primarily comparing what could be analytically distinguished through peepholes into the past (18) (North 2001; Gumbrecht 1997). Notably, as part of the work process, the reviews were transcribed by hand, primarily from newspapers on microfilm, creating a collection, and compiled as a rudimentary database in the form of a spreadsheet containing metadata about publication year, reviewed author, reviewed work, work's publication year and language as well as reviewer and organ of publication. Information about the gender of authors and reviewers was also included when available (in some cases, the name of an author or a reviewer is lacking because they wrote anonymously or used an unfamiliar pseudonym or signature).⁷

In generating data visualizations based on the original text material, we opted for quantifying the differences between the transcribed reviews, expressed as a form of distance, leading to the placement of texts closer or farther apart. More specifically, the text in each review was lemmatized (i.e., different inflectional forms of a word have been combined) and transformed using TF-IDF, a method that emphasizes words that are unique to a specific text and downplays words that are common to all texts (e.g., "the," "it," "that," "be") (Spärck Jones 1972), while at a sentence level, we use the Sentence Transformer model trained by the National Library of Sweden (Rekathati 2021), in an

6. Although there are potentially many ways to represent our text data in visualizations, we have for comparative purposes opted for maintaining the book reviews in their entirety.

7. The category "review" refers to an assessment of a work of fiction, published either as a separate article or in a collection of several other works. When individual assessments could be distinguished in the collective review, only the part of the text that belonged to each work was related to this review's entry in the database. If this was not possible, in cases where the works were treated "integrated," the same text was repeated for each entry. In other words, a collective review in the data, as well as in the visualizations, was treated as multiple reviews where possible.

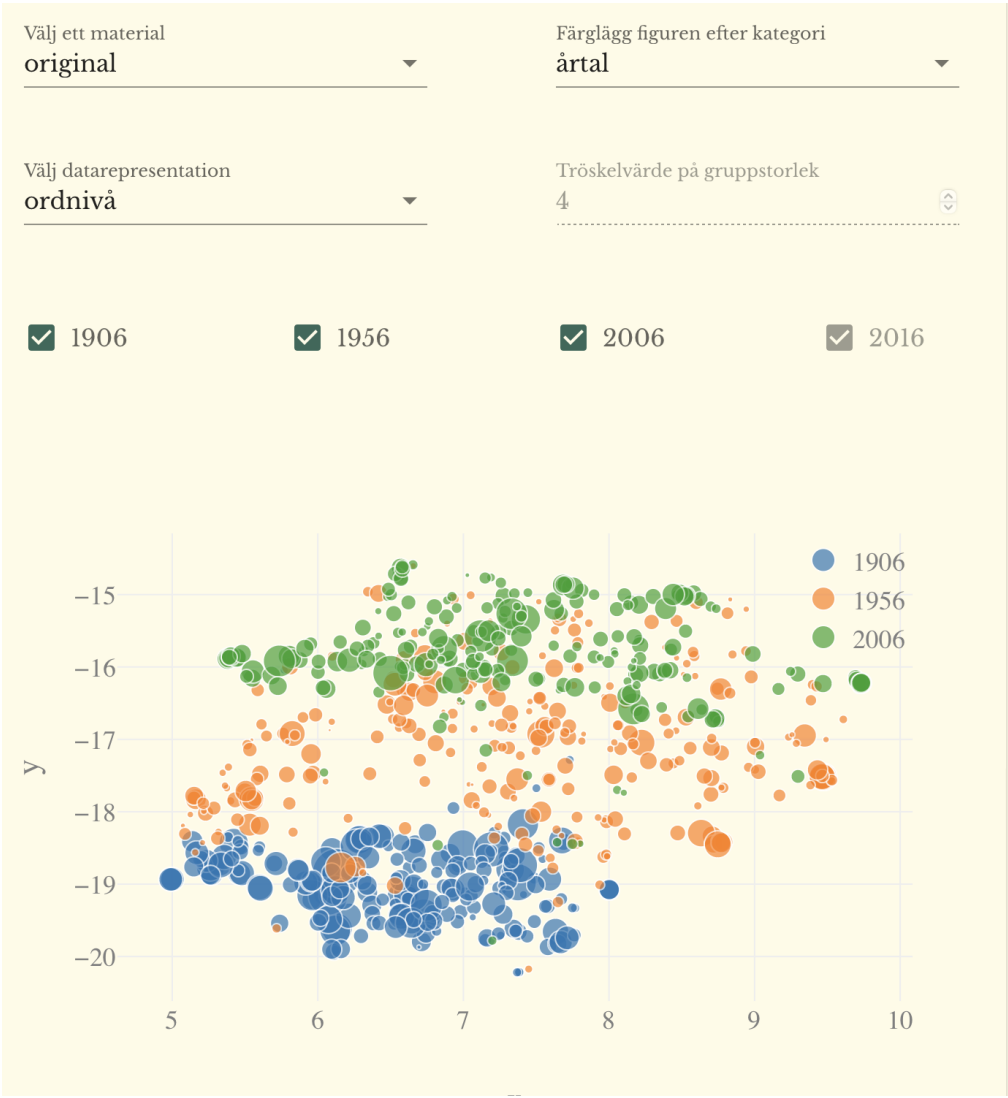


Figure 1: The "Map", showing 700 book reviews, here presented by year ("årtal") and word level ("ordnivå").

Hafvets stjärna Vilhelm Ekelund					
H. J. Kritiker					
lyrik Genre					
1906 Årtal					
ARB Forum					
UTFORSKA RECENSION					
INNEHÅLL GRANNAR					
Sök på grannstitlar					
ID	Titel	Årtal	Författare	Kritiker	Forum
1	Hafvets stjärna	1906	Vilhelm Ekelund	H. J.	ARB
55	Källorna	1906	Sven Lidman	Oscar Levertin	SVD
36	Källorna	1906	Sven Lidman	Bo Bergman	DN
52	Göran Delling	1906	Göran Forsslund	Olof Rosén	SVD
84	Sånger och syner	1906	Ellen Lundberg-Nyblom	John Atterbom	GHT
95	Hilligenlei	1906	Gustav Frenssen	okänt	STD
37	Göran Delling	1906	Karl-Erik Forsslund	Georg Nordensvan	DN
75	Hafvets stjärna	1906	Vilhelm Ekelund	Nils Peter Svensson	GHT
35	Hafvets stjärna	1906	Vilhelm Ekelund	Bo Bergman	DN
22	Hafvets stjärna	1906	Vilhelm Ekelund	Carl David af Wirsén	VL

Figure 2: Some of the neighbors ("grannar") to the review by the signature "H.J." of Vilhelm Ekelund's poetry collection *Hafvets stjärna* ("The Star of the Sea").

approach similar to e.g. Van Cranenburgh et al. 2019. In these representations some texts appear more similar than others – for simplicity, we refer to them as neighbors ("grannar") – based on vocabulary or sentence structure. The similarity between the texts was then visualized as distances in the form of a "map" (<https://dh.gu.se/kno/>), where reviews appear as a cloud of dots, each dot corresponding to a review whose metadata (publication year, reviewed author, etcetera) is displayed when the user activates the dot with a click in the interface, the size of the dots in the visualization being determined by the length of the review texts (Figure 1). The positioning, or embedding, of the reviews is calculated at the word level from the TF-IDF representation and at sentence level using the Sentence Transformer representation using UMAP (Uniform Manifold Approximation and Projection) as an approximation of the aforementioned distance between the review texts (akin to for example multidimensional scaling, MDS), being solely based on linguistic factors and independent from the metadata in the spreadsheet (McInnes et al. 2020; Borg and Groenen 2005).

Recensioner av litteratur

I denna visualisering presenteras samlingen av recensionstexter som ett *moln* av punkter, där varje punkt motsvarar en recensionstext med tillhörande verk, kritiker, författare med mera. Samma verk kan ha recenserats flera gånger av olika kritiker, och motsvaras då av flera punkter i visualiseringen. Positioneringen, eller *inbäddningen*, hos recensionerna är beräknad endast utifrån recensionens text, och är därför endast språkligt betingad - och inte avhängig metadata som årtal, genre eller liknande. Positioneringen är vidare projicerad från en högre dimensionalitet än ett plan, vilket betyder att avståndet mellan recensionerna inte är exakt bevarad. En passande liknelse är en karta skapad från den fysiska jorden, som ju på grund av sin klotform inte bevaras exakt på en platt karta.

Välj ett material

original

Färglägg figuren efter kategori

medietyp

Välj datarepresentation

meningsnivå

Tröskelvärde på gruppstorlek

4

☒ 1906

☒ 1956

☒ 2006

☒ 2016

Figure 3: The interface for choosing parameters in the visualization, in this example based on media type ("medietyp" – newspaper or journal), and sentence level ("meningsnivå").

In these visualizations, the embedding is projected onto a two-dimensional plane, which means that the distance between reviews is not reproduced exactly. Rather, this relationship is multidimensional and complex (comparable to a map of the Earth, a body that, due to its spherical shape, cannot be accurately represented on a flat map) or, as Drucker would put it, "any point or mark used as a specific node in a humanistic graph is assumed to have many dimensions to it – each of which complicates its identity by suggesting the embeddedness of its existence in a system of co-dependent relations" (Drucker 2011, §20). The true embedding distance is displayed in the "neighbors" column ("Grannar" in Figure 2), which may be used to confirm which reviews are actually close to each other locally. While it is indeed possible to globally quantify inter- and intra-group dispersion as in Van Cranenburgh et al. 2019, we judge that a local neighborhood of reviews remains more interpretable for a reader. In our interface, the visualizations display how the reviews position themselves in relation to each other based on factors such as year of publication, genre categorization, critic, publishing organ, and author of reviewed work (Figure 3). Unlike other explorative methods, such as topic modelling, this study is mainly interested in the characterization of reviews per the existing metadata.

On a more abstract level, our approach to vizualisation ties into the discussion of "performative materiality" to counteract an overestimation of the truth-value of data representations. Since data involves simplifications of the phenomena they describe, Katherine Bode stresses that in data-rich literary research we should consider the fact that the qualities of computational analysis are performative rather than representative. Bode describes this performative dimension in data representations as "sites – or apparatuses – for engaging with literary texts as emergent events, always arising from and altering

JCLS 3 (1), 2024, 10.26083/tuprints-00027397

8

conference version

how the literary past is (re)configured” (Bode 2020). A way to affirm this performative dimension on a technical level is, as advocated by Bode, to incorporate a self-reflective function into an interface. However, our approach to the visualizations rather raises another performative issue: a certain *defamiliarizing* quality.

In a discussion of Roberto Busa’s pioneering work in computer-driven text processing through the Index Thomisticus that began in 1946, Stephen Ramsay writes that the indexing of words in Thomas Aquinas’s collected works in the form of punch cards gave rise to a particular effect, “not the immediate apprehension of knowledge, but instead what the Russian Formalists called – the estrangement and defamiliarization of textuality. One might suppose that being able to see texts in such strange and unfamiliar ways would give such procedures an important place in the critical revolution the Russian Formalists ignited” (Ramsay 2011, 3). The concept of defamiliarization has been associated with various meanings in literary theory, but one can say that the concept is generally associated with aesthetic effects that create a distance between a work and its observer to provoke reflection. Notably, defamiliarization has traditionally been linked to modernist thought, which is characterized by the idea that consciously complex formal language somehow paves the way for a deeper understanding of reality. While our study obviously does not concern art in this sense or the imperative to stimulate a deeper reflection on the world, it is nevertheless crucial that data visualizations may not only provide an abstracted and modelled overview of a certain material, but also create a distance between us, as observers, and the material, thereby making it possible to speak of a defamiliarizing quality.

4. Comparative Re-reading

Turning to our analysis, we have chosen to focus on three factors – word and sentence levels, year of publication, and genre categorization – to show how data visualizations can inspire re-readings and provide complementary perspectives on a familiar material.

4.1 Word and Sentence Levels

In *The Order of Criticism*, Samuelsson writes: “As a genre, reviews have not undergone major changes over the past hundred years. In 1906, as well as in 1956 and 2006, descriptions, interpretations, and evaluations of one or more works constitute the core of criticism. Different functions may be more or less dominant, criteria and rhetoric may vary, but the genre of the review remains stable” (155).⁸ Other literary scholars of Swedish book reviews have made similar observations. For instance, Tomas Forser calls reviews “a genre of great durability,” and Per Rydén describes it as “a traditional, almost static genre” (Forser 2002, 155; Rydén 1987, 33). However, although the genre as a whole exhibits striking similarities over time, it is clear that over a century, the content has changed, to the extent that a data-driven analysis distinguishes a clear difference between reviews from different time periods.

If we return to Figure 1, we can see that reviews tend to group together based on differ-

8. “Som genre har recensionen inte genomgått några större förändringar under de senaste hundra åren. Såväl år 1906 som 1956 och 2006 är det beskrivningar, tolkningar och värderingar av ett eller flera verk som utgör kritikens kärna. Olika funktioner kan vara mer eller mindre dominerande, kriterier och retorik varierar, men recensionsgenren är stabil” (155).

ences and similarities at the word level, predominantly according to year of publication. Furthermore, there is a clear distance between them. The differences between 1906 (blue) and 2006 (green) are more significant than those between 1956 (orange) and 1906 or 2006, indicating some form of chronological change.⁹ In short, the visualization shows that reviews from, for example, 1906 in terms of word choice are as similar to each other as they are different from texts from 1956 and 2006. For the middle year 1956, reviews are slightly more dispersed in the visualization, with some ending up with reviews from 2006 and others from 1906. A few reviews from 2006 are placed among reviews from 1906: Jim Kelly's detective novel *Måntunneln* (*Moon Tunnel*) and the children's books *Skämmarkriget* (*The Shaming War*) by Lene Kaaberbøl, *Min syster flygande Flavia* (*My Sister the Flying Flavia*) by Helena Öberg, and *När Johan vaknar upp en morgon är han stark* (*When Johan Wakes Up One Morning He is Strong*) by Petter Lidbeck and Lisen Adbåge, which we will return to below.

Notably, one should pay attention to which words determine a text's placement in a particular year cluster. While it is not possible to draw any conclusions about this solely based on the most represented words in an individual text (since positioning is determined by a complex system of relative occurrences among the reviews), it is relevant to take into account which words are over- or underrepresented for each individual year in groupings. Over- and underrepresentation are calculated here using Dunning's log-likelihood method, a familiar algorithm in corpus and discourse analysis, which quantifies how unexpected a word is in a text given the words in all other texts within a certain group, such as years (Dunning 1993). One possible explanation for reviews grouping so clearly by year may, of course, be language changes over time. For instance, words that are particularly characteristic of specific years, according to data analysis, include "skald" (poet) and "författarinna" (female author), as well as the word form "äro" (are) for 1906. However, such words seem outdated in 2006 when terms like "fiktiv" (fictional), "identitet" (identity), and "relation" (relationship) are prominent.¹⁰

One way to get closer to the factors that determine the placement of reviews in the visualization is to compare the words that vary most in frequency between the years, i.e., those that are over- or underrepresented for a specific year.¹¹ Other words that are particularly characteristic of appearing in a 1906 review include "han" (he), "hon" (she), "djup" (depth), "akt" (act), "förf" (auth, abbreviation for author), and "öfrig" (other). The latter ("öfrig") can be related to spelling reform, while "akt" is probably connected to more plays being reviewed in 1906 than in the other years. The use of "förf" (auth) likely results from it being a common abbreviation for "författare" (author) at

9. As mentioned above, the original study refrained from diachronic perspectives and adhered to the logic imposed by the single-year perspective to see each individual year as a (media) archaeological object in its own right, rather than as a passing point in historiographical progress.

10. In Sweden, the spelling reform that was implemented in 1906, although it gained broader acceptance a few years later, may have some influence.

11. In this particular context, we do not consider words that – in comparison to the others – are notably infrequent in a specific year. However, it can be noted here that "talang" (talent), "dylik" (similar), "själ" (soul), "natur" (nature), and "god" (good) for 2006; "andlig" (spiritual), "sorg" (grief), "dotter" (daughter), "son" (son), "språk" (language), "röst" (voice), "liv" (life), and "vi" (we) for 1956; and "centrum" (center), "självt biografisk" (autobiographical), "debut" (debut), "mamma" (mom), "identitet" (identity), "barn" (child), "klass" (class), "miljö" (setting), and "språk" (language) for 1906 appear in these reviews. These words indicate how language usage has changed but also reflect the order of critical discourse that the study describes (certain things are obvious to talk about at a certain time, while others are uninteresting or peripheral).

that time. Furthermore, the more frequent use of "hon" (she) and "han" (he) in 1906 than in later years could be explained by how reviews at the time dedicated significant space to content summaries, often focused on describing and explaining characters and their actions.

Equivalent typical words for reviews from 1956, for example, are "roman" (novel), "social" (social), "urval" (selection), "miljö" (setting), "analys" (analysis), "avsnitt" (section), "fin" (fine), "politisk" (political), "höst" (autumn), "spela" (play), "uppleva" (experience), "människa" (human), "diktare" (poet), and "beroende" (dependence). The presence of some of these words can probably be explained by the topics and themes of the literary works that were most frequently reviewed, as well as the fact that the term "diktare" replaced "skald" (skald). The interest in formal features and close reading that has been associated with New Criticism during this period can be noted in the use of terms such as "analysis" and "section" (76–77). The high-frequency words also testify to a certain societal engagement in the criticism, as evidenced by the presence of words like "political," "environment," and "social." This is also noted in *The Order of Criticism*, where it is related to the reflections of the time, in the aftermath of World War II, on "humanity," "mankind," and the human psyche, something that can also be seen in the recurring use of the term "human" (84, 88).

For 2006, on the other hand, the most distinctive words are "jag" (I), "skriva" (write), "text" (text), "språk" (language), "roman" (novel), "bli" (become), "berättelse" (story), "läsa" (read), "mamma" (mom), "pappa" (dad), "barn" (child), "far" (father), "handla" (act), and, as mentioned above, "relation" (relationship), "identitet" (identity), and "fiktiv" (fictional). Here, we observe several words that can be related to the fact that the discussed works – and perhaps in some cases reflections on the critics' own lives – revolve around relationships and family dynamics ("mom," "dad," "child," "father," "relationship"). Other words are indicative of how literature is discussed and described ("write," "language," "novel," "story," "fictional," "act"). The distinguishing words confirm the prior observations in *The Order of Criticism* about a more present and subjective critical subject, as well as a significant interest in identity issues (125–127; 134–136; 145–148).¹²

A visualization at the sentence level (Figure 4) provides a much more heterogeneous result, which can support the above argument that the *form* of criticism has not changed significantly, while the visualization at the word level in Figure 1 indicates that the *content* expressed or valued has changed over time.¹³ In this way, one can say that the data-driven analysis actually seems to confirm the earlier assumptions of literary critics that literary criticism as a whole is a relatively stable – or, if you will, conservative – genre of text.

12. A quick look at the overrepresented words for each year reveals that the evaluative words that we might normally attribute great importance to within literary criticism, at least quantitatively, do not play a significant role in the material. For 1906, the word "djup" (depth) remains, in 1956, "fin" (fine), while in 2006, we do not find any such words at all (perhaps a sign of the times). However, a word's frequency says nothing about how significant it is in context. In this regard, both the original study and the data visualization could benefit from being supplemented with some sort of sentiment analysis, in order to organize and study evaluative words and attitudes in their immediate context.

13. The visualization of the distances between review texts at the sentence level does not consider the text as a collection of individual words, but as a collection of sentences, preserving structures and formulations. Formally, a SentenceTransformer is used to produce equivalent embeddings as on the word level. See Rekathati 2021.

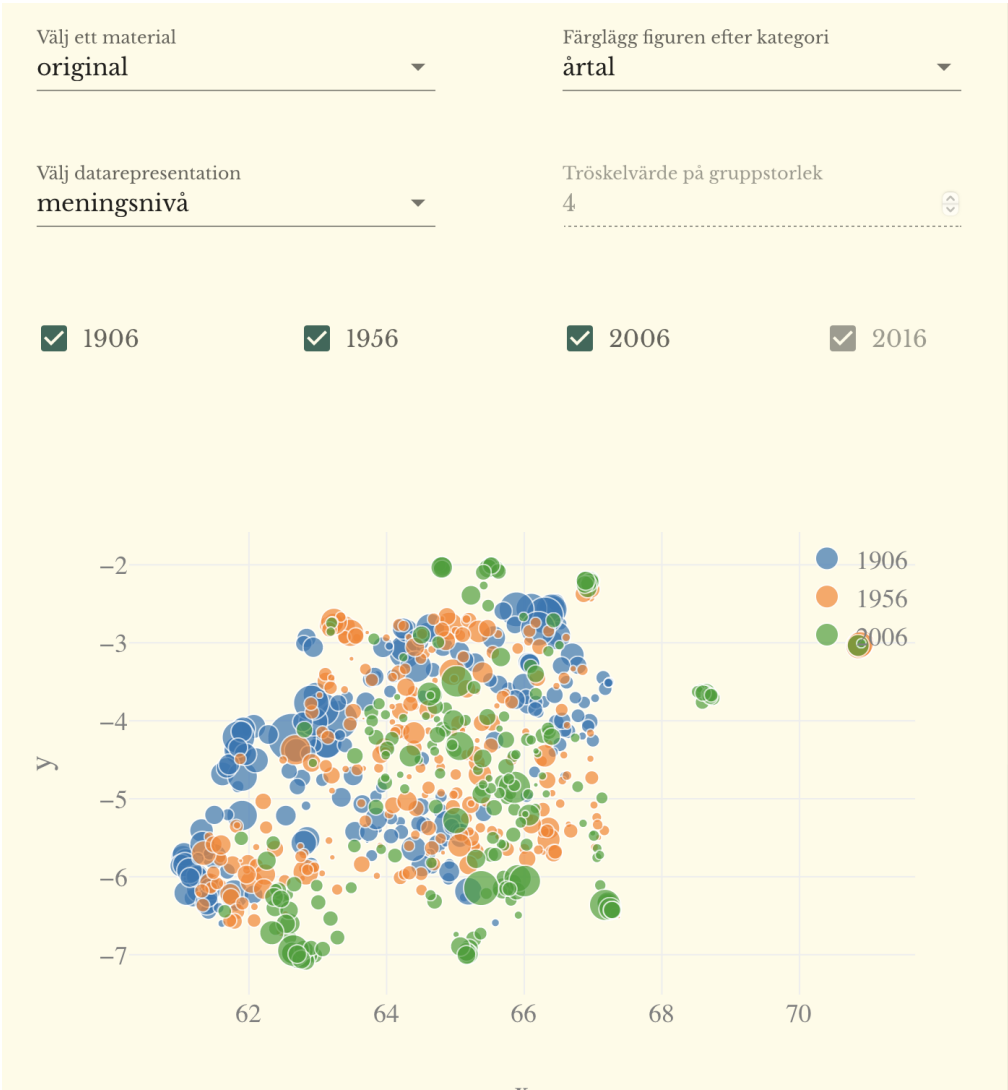


Figure 4: Visualization of the material by year ("årtal"), based on the sentence level ("meningsnivå").

INNEHÅLL <u>GRANNAR</u>					
Sök på granntitlar Q					
ID	Titel	Årtal	Författare	Kritiker	Forum
623	Måntunnel	2006	Jim Kelly	Jan Broberg	SDS
589	Hjältar	2006	Claudia Marcks	Terje Holtet Larsen	O&B
59	Dissonantser	1906	Christian Krogh	August Brunius	SVD
531	Lunar Park	2006	Bret Easton Ellis	Johan Dahlbäck	GP
79	Mary	1906	Björnstjerne Björnson	Karl Warburg	GHT
393	Långt bort härifrån	1956	Tora Dahl	Stig Carlson	Persp
13	Saul	1906	Nils Gustaf	J. A. Runström	AB
169	Doktor Glas	1906	Hjalmar Söderberg	Fredrik Böök	O&B
522	The lay of the land	2006	Richard Ford	Andres Lokko	Expr
160	Sonen	1906	Hugo Öberg	Fredrik Böök	O&B
Träffar per sida 10 1-10 av 10 < >					

Figure 5: The neighbors to Jan Broberg’s 2006 review of Jim Kelly’s *Moon Tunnel*, four of them being from 1906.

4.2 Publication Year

324

As a distinct example of the defamiliarizing qualities of the visualization, we can compare the reviews that end up far from others within the same group (i.e., outlier dots) to study common distinguishing features. For example, *Moon Tunnel* by Jim Kelly, reviewed in *Sydsvenska Dagbladet* in 2006, can be seen on the map surrounded by reviews from 1906. Looking at the neighbors, they are indeed reviews from different years, but a significant number of them are from 1906 (Figure 5). Since this text, unlike most of the others from 2006, has neighbors from 1906, there is a reason to consider why this is the case.

The review of *Moon Tunnel* is part of a collective review where Kelly's work is discussed in pair with Peter Robinson's *En bit av mitt hjärta* (*Piece of My Heart*), but the text is clearly divided in the sense that the first half deals with Robinson's work, and the second with Kelly's. The visualization is based on the database, which treats these texts as two separate segments (as mentioned above). The review of Robinson's work, unlike the review of Kelly's, is located near the cluster of 2006 reviews but is also surrounded by reviews from 1956. It's worth noting that these reviews, even though they appear in the same article, were separated in the original study for analytical purposes and are thus treated as separate texts in the database. This makes the collective review particularly interesting for our purposes, as the same text gives rise to two different placements in the visualization. Do they differ significantly?

Let's start with the review that landed in the center of the 1906 review cluster, *Moon Tunnel* by Jim Kelly. The words that the computational analysis has identified as significant, aside from those related to the plot, include words like "obestridd" (undisputed), "lättköpta" (easily bought), "återigen" (again), "elegi" (elegy), "udda" (odd), "mästerskap" (mastery), "lansera" (launch), "lovande" (promising). In this context, *significant* means the weighting an individual word has on the placement of the work in the visualization. Words like "promising," as well as others listed further down like "nå" (achieve), "steg" (step), and "författare" (author), are terms that could be related to the typical characteristics of literary criticism around 1906 and a tendency to assess how well the author has developed artistically, and to determine if an author is worthy of their title (as true authors).¹⁴ Clear evaluative words like "undisputed" and "mastery" could be linked to this discourse, which becomes evident upon closer examination of the text.

The presence not only of individual words, but how evaluative words function in the review of *Moon Tunnel* that resemble the order of criticism in 1906, becomes apparent when one considers the review as a text rather than as text data. The review begins

14. "A work can receive praise while its author is told that he or she is not a poet or bard. When Oskar Hoffmann's children's book *Bland Marsmänniskor* (Among Martians) is reviewed, the critic points out that it is "a work by a faiseur, not a poet." Axel Klinckowström's verse epic *Örnsjö-tjuren* (The Örnsjö Bull) is even called a debut work, despite the reviewer knowing that the author has previously published both poetry collections and prose works. He explains: "I deliberately write debut, for in the not so few poems he previously published with Old Norse subjects, the poetic berserker rage struggled too hard with literary amateurism for the result to be the intended."

(Ett verk kan få lovord samtidigt som dess författare får veta att han eller hon inte är någon diktare eller skald. När Oskar Hoffmanns barnbok *Bland Marsmänniskor* recenserar påpekar kritikern att den är 'ett verk af en faiseur, icke af en skald'. Verseposet *Örnsjö-tjuren* av Axel Klinckowström kallas till och med för ett debutantverk – trots att anmälaren vet att författaren utgivit både diktsamlingar och prosaverk tidigare. Han förklarar: 'Jag skrifver med flit debuterat, ty i de ej så få poem han förut utgifvit med fornnordiska ämnen brottades det poetiska bärsärkaraseriet allt för hårdt med den litterära diletantismen för att resultatet skulle blifva det afsedda')" (41).

with: "Jim Kelly does not reach the now undisputed mastery of Robinson, but his latest detective novel, *Moon Tunnel*, is still a step forward for this promising English author."¹⁵ Here, one can observe stylistic features that are described in *The Order of Criticism* as characteristic of 1906. The critic's evaluation is evident – Kelly is considered "inferior" to Robinson, who is described as a "master." Similarly, the development of the author's work is assessed, and the reviewer believes that the novel is "a step forward for this promising English author." This can be compared to reviews from 1906 where a critic might praise aspects such as "an unusually straightforward developmental trajectory," while another critic laments a poetry collection that is "all too similar to its older siblings" (33).¹⁶

Looking at the reviews of *The Shaming War* and *My Sister the Flying Flavia*, which also have neighbors from a century ago, both stand out for consisting of plot summaries, concluding with a clear assessment from the critic. "With *My Sister the Flying Flavia*, copywriter Helena Öberg has created a sympathetic and easily readable story for those between seven and nine," writes *Sydsvenska Dagbladet*, and the critic from *Upsala Nya Tidning* concludes the review of Lene Kaaberbøl's *The Shaming War* with the judgment that: "The Shaming series is not a complicated fantasy work, rather a fairly simply told saga, with not too large a cast of characters or an advanced structure. But due to some truly scary scenes, it is still not suitable reading for very young fantasy fans."¹⁷ Helena Öberg's *When Johan Wakes Up One Morning he is Strong* is also reviewed in *Upsala Nya Tidning*, alongside another illustrated chapter book. This text is also relatively short and primarily focused on the plot.

The reason these children's book reviews are close to the 1906 cluster likely lies in the significant use of words describing the content of the literary works, which is typical also of early 20th-century criticism, along with words declaring clear concluding judgment.¹⁸ Furthermore, the critics do not refer to themselves in the above-mentioned reviews of Öberg's, Kaaberbøl's, and Kelly's books: there are no "I," "my," "mine," or other references to the critic as a person. This distinguishes these reviews from the descriptions of literary criticism in 2006 encountered in *The Order of Criticism*, which highlights the presence of the critical subject, while the absence of reference to the writing subject is typical of critics from a hundred years earlier.

But, returning to the crime fiction review discussed above: how do the texts about Robinson's and Kelly's detective novels differ from each other – after all, the books are reviewed in the same review but end up in different places in the visualization (Broberg 2006)? Why does the text about Robinson's end up among reviews from 1956 but much

15. "Till Robinsons numera obestridda mästerskap når Jim Kelly inte upp, men dennes senaste deckare, *Måntunneln*, är ändå ett steg framåt för den här lovande engelske författaren" (Broberg 2006).

16. "En ovanligt rakt uppstigande utvecklingslinje" and "blott allt för lik sina äldre syskon" (33).

17. "Med *Min syster flygande Flavia* har copywritern Helena Öberg skapat en sympatisk och lättläst berättelse för den som är mellan sju och nio." Frieberg 2006; and "Skämmerskeserien är inte något komplicerat fantasyverk, snarare en hyggligt enkelt berättad saga, utan alltför stort persongalleri eller avancerad struktur. Men på grund av en hel del riktigt otäcka scener är det ändå inte läsning för alltför unga fantasyfans" (Tammerman 2006).

18. Another possibility is that the words related to the plot of the novels are also common in literary works from 1906. However, in these reviews from 2006, we find words such as "strid" (battle), "mörk" (dark), "oförrätt" (injustice), "ärkefiende" (archenemy), and "rättmätig" (rightful) (in the context of *The Shaming War*); "förälder" (parent), "bo" (home), "skola" (school), "tärtljus" (cake candles), "pilla" (fiddle), "utblåsa" (blow out), "fosterhem" (foster home), and "rosenbusk" (rosebush) (in the context of *My Sister the Flying Flavia*); and "morgon" (morning), "pyjamasskjorta" (pyjama shirt), "hulkenstil" (Hulk style), "plågoande" (tormentor), and "moppe" (moped) – which does not support such an interpretation.

closer to other 2006 reviews than the later part of the text discussing Kelly? 394

Of the words listed as significant for the placement of the Robinson review (among 395 those not related to the plot), we can note terms such as "förtjänst" (merit), "höstbok" 396 (autumn book), "engelsk" (english), "deckararena" (approx. detective genre), "roman" 397 (novel), "konststycket" (the feat), "komplexitet" (complexity), "mysterium" (mystery), 398 "täthet" (density), "eminent" (eminent), "levandegöra" (bring to life), "förbrylla" (baf- 399 fle), "personteckning" (characterization), "invända" (object), "nyanserad" (nuanced), 400 "parentes" (parenthesis), "händelseförlopp" (sequence of events), "invändning" (ob- 401 jection), "ovänta[d]" (unexpected), and "bidra" (contribute). One can also note more 402 words related to the critic and their task, such as "recensera" (review), "recension" 403 (critique), "läsare" (reader). Furthermore, several evaluative expressions are present, 404 such as "ny" (new), "bra" (good), "favorit" (favorite), "positiv" (positive), which 405 align more with the literary critical discourse of 1956 and 2006 than 1906 (134–135). 406 Looking at the actual review, it also starts with a clear focus on the critic himself: "That 407 Peter Robinson belongs to *my* favorites in the detective genre today, has surely become 408 apparent from *my* reviews over the years," [our emphasis]. Following this, which is 409 quite typical for the reviews of 2006, is a reservation that simultaneously emphasizes 410 the qualities of the work: "It could possibly be argued that the author does not play 411 entirely fair with the reader in a certain respect, but it is still an objection that carries 412 little weight considering all the other merits of the novel." The critic talks about the 413 novel as dense and complex, the characterization nuanced, and the setting vivid. 414

Primarily, the Robinson review focuses on evaluation, and it's a positive one. Despite 415 recurring phrases related to the plot of the novel, there isn't a direct description of the 416 plot, but rather, they serve as summaries: it is in the vividly depicted English landscape 417 where "the events unfold," and it is the "portrayal of the youth culture that plays a 418 significant role in the plot" that makes the novel complex. We don't get to know much 419 more about what is being depicted. This brevity in plot summaries is more characteristic 420 of 1956 and 2006, than of 1906 reviews, where we have seen that the course of events 421 can be described in some detail. However, the Robinson review ends in the spirit of 422 1906 critics with an assessment of the author's progression: "Yes, Robinson has certainly 423 developed since entering the detective genre." 424

Thus, there are clear differences in language use at the word level between reviews from 425 1906, 1956, and 2006, but somewhat less at the sentence level, which in this case could be 426 interpreted as the rhetoric and typical genre features of the criticism. Some discursive 427 features noted to apply to the different years are supported by the data-driven analysis, 428 but there is also room to discover other patterns, such as how different literary categories 429 are reviewed. This will be the focus of the next observation about the defamiliarizing 430 quality of our visualizations. 431

4.3 Genre Categorization 432

During the writing process of *The Order of Criticism* the data were compiled regarding the 433 genres in which reviewed works were categorized according to the National Library of 434 Sweden's catalog Libris: prose, poetry, drama, children's literature, and "other" (which 435 includes, among other things, audiobooks and comic books). It goes without saying that 436 literary genres are far more complex and ambiguous than what these categories reflect. 437

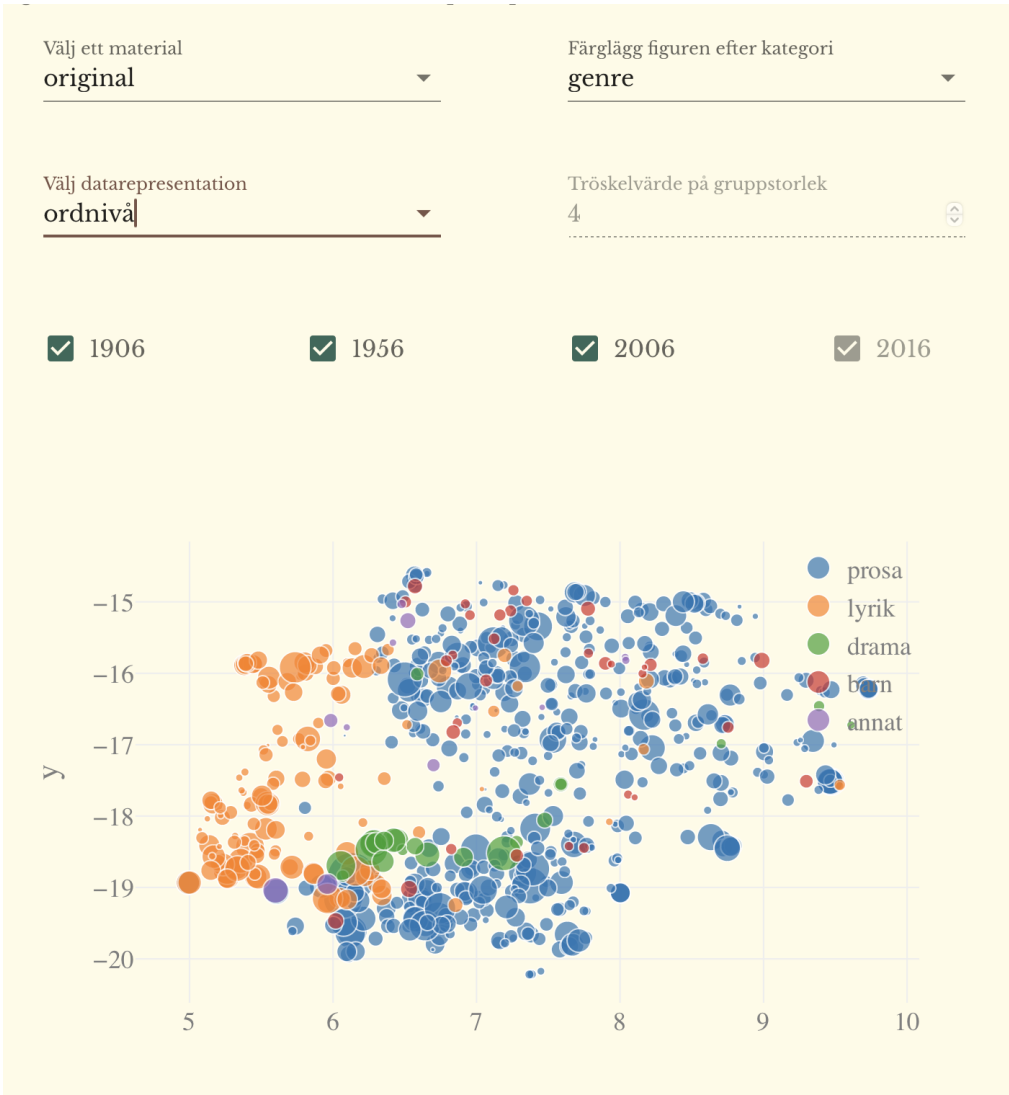


Figure 6: Visualization on the word level, based on the reviewed work’s genre. “Prosa” = Prose; “Lyrik” = Poetry; “Drama” = Play; “Barn” = Children’s literature; “Annat” = other.

Institutionalized classifications are just one part of the networks of cultural meaning-making and historical processes that contribute to our understanding of which genres a particular book can be understood in relation to. Genres consist of a constantly changing, multifaceted, and contradictory palette of aesthetic traditions and labels, where libraries are one actor, and the audience, the book industry, reviewers, and researchers are others. Nevertheless, the Libris catalog can be used to create a rudimentary perspective on the relationships between different literary works and their reception, as computerized analysis can easily track differences and similarities at the text level based on attributed genres.

To avoid delving into a complex genre theoretical discussion, for the sake of simplicity, we choose to refer to these variables as "genre categorizations." Even though the Libris catalog might be considered an authority in this context, there are plenty of indications that library classifications can be discussed. For example, "children's literature," rather than being a more distinct genre, should be seen as a collective term for literature written by adults for a child audience, which can encompass both prose and poetry as well as plays for children. Nevertheless, in critical practice, there is a tendency for different reviewers to be assigned works from different genres: one critic reviews prose, another reviews drama, a third reviews poetry, and someone else writes about children's literature.¹⁹

In Figure 6, where the visualization is color-coded at the word level based on assigned genres in Libris, we can see that the reviews, as in the case of publication years, clearly group by category. The same holds true at the sentence level, as shown in Figure 7.²⁰ At the word level, almost all poetry (orange) is concentrated on the left. Likewise, drama (green) forms a distinct cluster. Similarly, prose (blue), which constitutes the largest category, is cohesive. The most dispersed category is children's literature (red), both at the word and sentence levels, which can likely be explained by the fact that children's literature, as mentioned earlier, encompasses a range of forms of expression. It may also be due to significant variations within children's literature criticism. An indication of this is that the "other" category, which includes, among other things, comic books and essays, can also be described as heterogeneous and scattered in the visualization.

As in the case of publication years, it is reasonable to make some observations about noteworthy placements here. In Figure 6, we can note that a limited number of poetry reviews ended up among prose reviews, but there are no prose works in the poetry section on the left. In this sense, one can speak of a significant consistency within poetry criticism. Some of the prose reviews that are placed near the poetry reviews (and have several poetry neighbors) are reviews of Vendela Fredricsson's *Landar* (Landing) from 2006. In this context, it is relevant to mention that Landing is a prose-lyric short novel that made *Expressen's* critic wonder "if the alleged debut novelist [...] actually wants to write semi-surrealistic poetry."²¹ The colleague in *Helsingborgs Dagblad* noted that "[a]t

19. It would be an interesting study in its own regard to explore the discrepancy between the critical practice and the literary analysis regarding genre categorizations.

20. The following analysis will be based on the placement in the graph of the reviews at the word level, but we can thus conclude that unlike how the reviews grouped themselves in relation to years, there does not seem to be any significant difference regarding genres in the works being reviewed whether the visualization is done at the sentence or word level.

21. "[...] om det egentligen är semisurrealistisk poesi som den påstådda romandebutanten [...] vill skriva" (Lekander 2006).

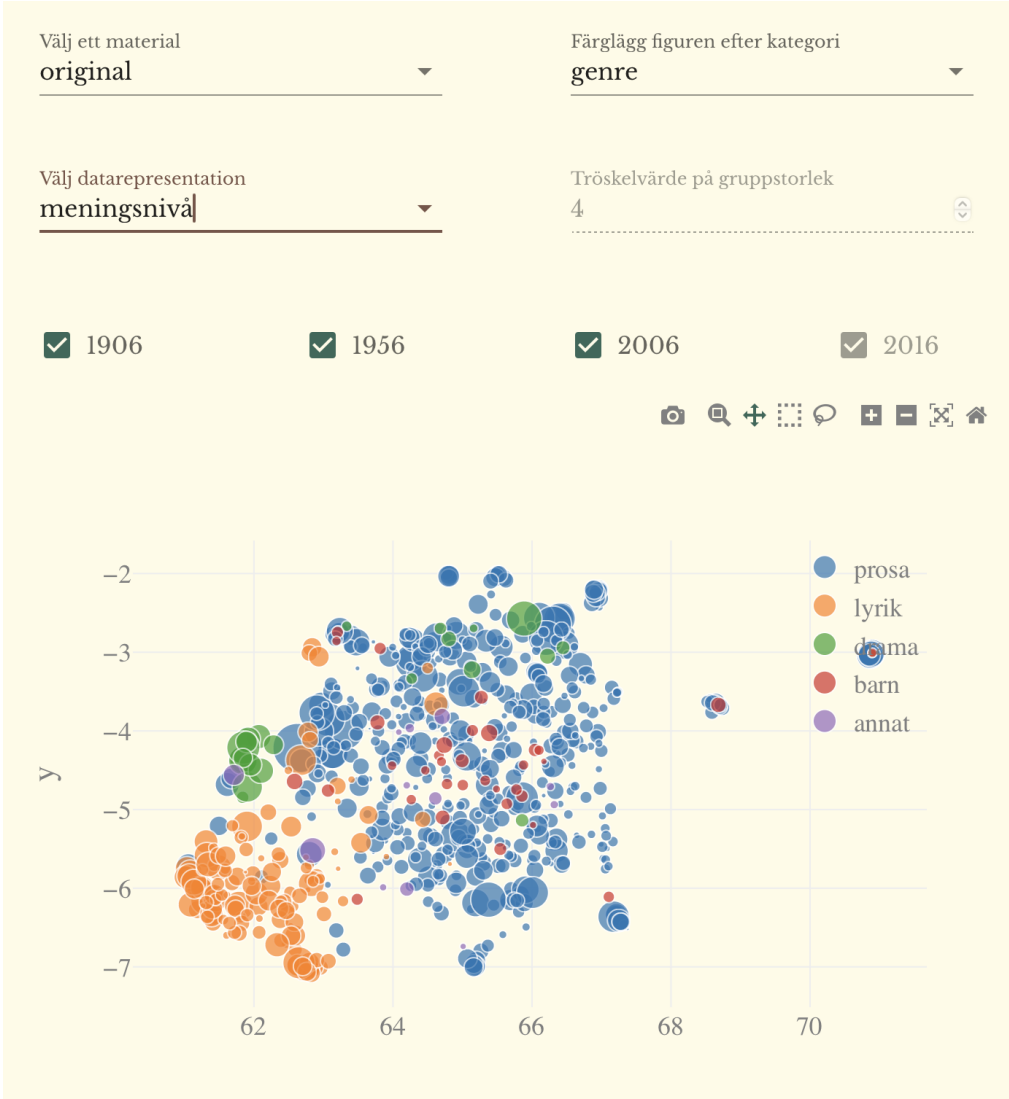


Figure 7: Visualization on the sentence level, based on the reviewed work's genre.

times, *Landing* feels more like poetry than a novel” (Lingebrandt 2006). *Landing* was also reviewed by *Göteborgs-Posten*, but its critic, unlike the others, did not focus on the work’s lyrical aspect but rather discussed its plot (a love triangle) in some detail. This review is also placed far from the other reviews of the same book.

The “drama cluster” in Figure 6 includes a limited number of works that were reviewed in several newspapers, mainly in 1906. However, we find some drama reviews placed further away together with prose, including Cecilia Nelson’s *Öknen* (The Desert), reviewed in *Norrländska Socialdemokraten* in 2006, as well as a collective review in the magazine *Perspektiv* in 1956 of four comedy plays. It should be mentioned in this context that only a few plays were reviewed during the examined periods of 1956 and 2006. The fact that these are placed far from the others indicates possible historical changes and differences in both the drama category and the criticism of drama. In the review of *The Desert* there is actually no discussion about the genre itself – that is, the play – except that it mentions that it is Nelson’s “debut play.” Among the words that have influenced the review’s placement in the visualization are those related to the work’s plot, including “kamel” (camel) and “möte” (meeting), and adjectives like “politisk” (political) and “verklig” (real).

Another indication that the reviewed works have more influence on the groupings than the reviewer or the category is that the reviews from 1956 of Erland Josephson’s drama *Sällskapslek* (Party Games), Jean Anouilh’s *Ornifle eller Luftgästen* (Ornifle: A Play), Hans Hergin’s *O, sköna Tasmanien* (O, Beautiful Tasmania), and Bo Widerberg’s *Skiljas* (Divorce). These four plays are included in the same collective review, but are not placed next to each other. Although works in the same category often become neighbors in the visualization, this is not surprising in itself. The content of a work is reflected in the text that deals with it, often through quotes and plot summaries. However, it is still worth noting that even though the visualization does not take metadata into account, it creates a striking pattern.

Let’s take an example from 1906: Anders Österling’s play *Nattens röster* (Voices of the Night). When reading the reviews, it becomes clear that they are remarkably similar to each other. This is evident not least through the words that are most significant for the placement of the reviews in the visualization. Several of the recurring words are related to the play’s form and content, such as “akt” (act), “musik” (music), and “mor” (mother).²² Other recurring words are related to the genre itself, such as “dramatisk” (dramatic), “drama” (drama), “vers” (verse), and “lyrisk” (lyrical).

When it comes to the prose category, reviews of the same book also group together. In Figure 8, we have sorted out the works that were reviewed at least five times in 1906 and marked them in different colors. Here, it is evident that even though some reviews of the same work are so close that they overlap, while others have a wider spread, reviews of the same title are usually neighbors. Essentially, the same holds true for 1956 and 2006. In short, reviews tend to group with their peers in terms of both categories, publication years, and titles.

22. As can be seen in the list of significant words, “mala” and “ering” are also recurring, which are actually the names of the protagonists Mala and Ering. This, in turn, reminds us that digital analysis normally excludes proper names, but in this case, they are not perceived as such because they look like ordinary words. The title of the work and other metadata are also filtered out, and therefore, words like “natt” (night) or “röst” (voice) are not included.

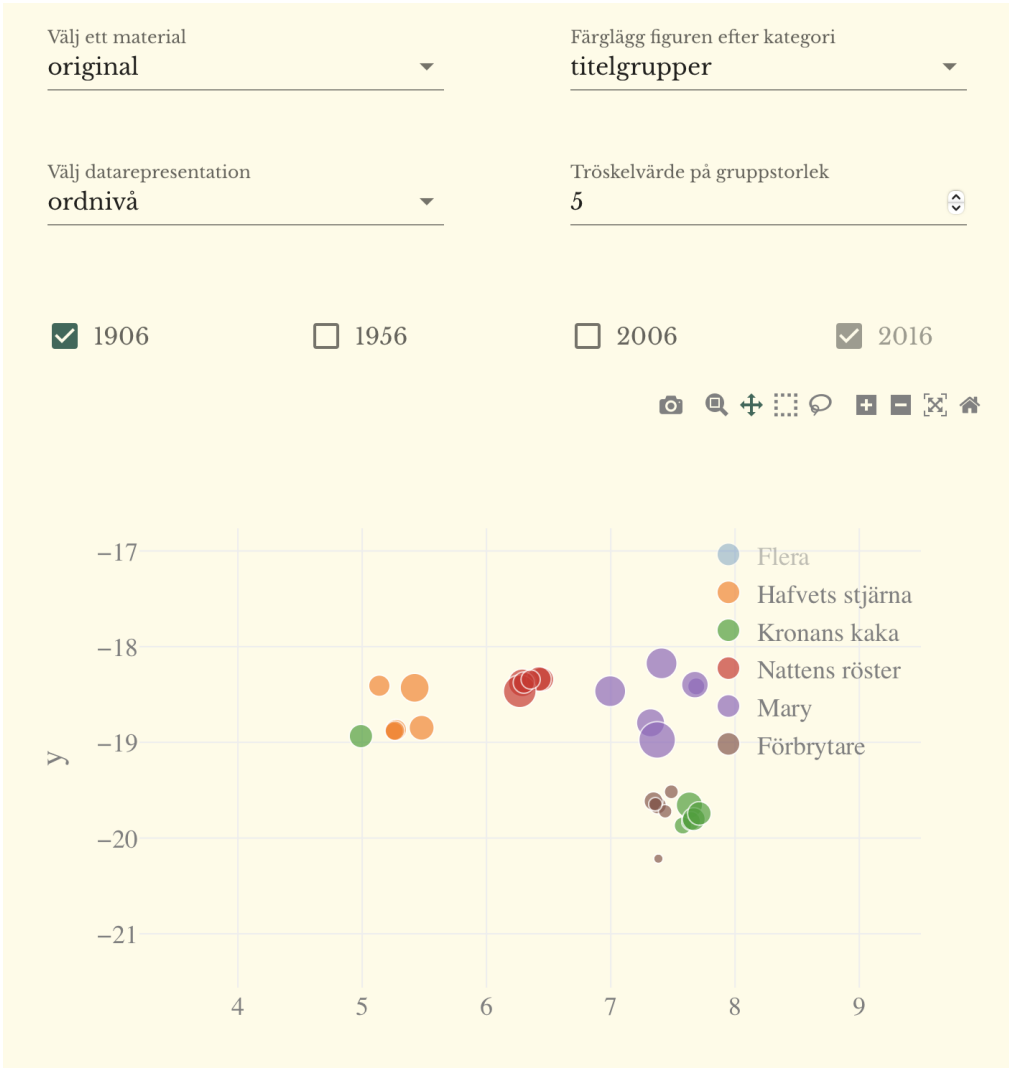


Figure 8: Visualization on word level of reviews where the same title has been reviewed more than five times 1906. Different colors mean different literary works.

5. Conclusion – Contextualization and Defamiliarization 518

Initially, we described our use of a mixed methods approach to the study of literary criticism in terms of what Shan refers to as a dialectical position, which means that the investigation does not prioritize a quantitative method over a qualitative method, and vice versa. Rather, we recognize that different approaches generate different results that taken together, nevertheless, can enrich the understanding of what has characterized the norms of literary criticism at different points in time, as analyzed in a previous study (Shan 2023, 8). According to Shan, mixed methods can be applied at different levels in scientific practice, including method selection, and epistemology, which has bearing on our analysis of data patterns emerging in visualizations of a corpus of book reviews previously examined in a study in comparative literature. *Methodologically*, we have combined a quantification of differences and similarities between book review text with close re-reading, taking the historical context of the texts into account. *Epistemologically*, following Piper and Algee-Hewitt, we have explored how dialectically combining traditional and digital analysis may contribute to new knowledge about a particular research material (Piper and Algee-Hewitt 2014).

Therefore, there is a point in discussing the results on a both concrete and abstract level. Concretely, our visualizations of overrepresented and underrepresented words in literary criticism from different periods confirm assumptions made in the original study, for example that reviews in 1906 devoted more space to plot summaries and evaluation of authorship, while reviews in 1956 reflected a different societal engagement, and those in 2006 tended to emphasize the “I” of the critic. However, by visualizing linguistic characteristics in relation to publication year, we not only found that reviews grouped themselves into clusters roughly in line with our expectations, but also that reviews sharing strong thematic similarities challenged chronological expectations, and grouped together regardless of significant historical distances. An example being a review from 2006 of a detective novel that contained a rhetoric very similar to how reviews in 1906 tended to evaluate authors based on their perceived artistic development towards “mastery.” Our visualizations of genre categorizations also called for closer examination. The fact that a review of a prose-lyrical short novel ended up near the cluster of poetry reviews, rather than prose reviews, was likely due to how the reviewers tended to emphasize the book’s fusion of prose and poetry. At the same time, a single review of the novel in question that did not touch upon this aspect, ended up far from the others. Thus, here the visualization directed our attention to the extent to which reviews foreground genre characteristics, a critical aspect not discussed in *The Order of Criticism*. Notably, these results point to the importance of contextual approach when analyzing our text data visualizations. Without knowledge about the historical contexts of literary criticism, it would be hard to make such observations about the clustering and breaks in the expected pattern.

Furthermore, our analysis highlights the usefulness of the concept of defamiliarization in our analytical context. Here, we can specifically turn to Victor Shklovsky’s conceptualization of how defamiliarization slows down or de-automates perception, allowing familiar assumptions to be renegotiated. Analyzing Shklovsky’s notion of defamiliarization and the perceptual processes that a work sets in motion, literary scholar Beata Agrell makes an important distinction (Agrell 1997b, 26–58, 1997a, 87–89). Agrell argues

that, according to Shklovsky's theory, the work in question "is thus not autonomous but 563
directed towards a certain type of observation, which it simultaneously *invokes* through 564
 its built-in devices" (Agrell 1997b, 28).²³ Hence, in a transferred sense, one may say 565
 that our data visualizations de-automatize the perception of the text material and also 566
 defamiliarize the original conclusions in *The Order of Criticism*. The fact that our results 567
 confirm many conclusions in the prior study can in this context be viewed as a strength, 568
 as it indicates that the visualizations can indeed capture significant patterns in the 569
 material. Perceiving something in a radically different way does not necessarily mean 570
 seeing radically different things. Rather, a key point in thinking about visualizations in 571
 terms of Shklovsky's concept of defamiliarization is that they offer a "double vision" or 572
 a shift between different positions from which to study the texts. Arguably, one may 573
 talk about a potential to evoke shifts in perspective and to direct analytical attention 574
 to overlooked aspects of a specific material. Thus, rather than ultimately leading to a 575
 "better" path to truth, visualizations could potentially generate new research questions 576
 about familiar materials. Which seems significant enough. 577

6. Acknowledgements 578

This paper has been written with the financial support of Riksbankens Jubileumsfond, 579
 funding the project *The New Order of Criticism: 150 years of Book Reviews in Sweden* 580
 (2020–2024). Special thanks to Aram Karimi, GRIDH (Gothenburg Reserach Infrastruc- 581
 ture in Digital Humanities) at the University of Gothenburg, for helping out with the 582
 LaTe implementation. 583

7. Author Contributions 584

Daniel Brodén: Supervision, writing and editing 585

Jonas Ingvarsson: Supervision, writing and editing 586

Lina Samuelsson: Writing and editing, author of original study 587

Victor Wählstrand Skärström: Data visualization, quantitative analysis and writing 588

References 589

- Agrell, Beata (1997a). "Brukslitteratur, skönlitteratur och medkänslans estetik – betrak- 590
 telse över läsarter." In: *Den litterära textens förändringar*, 87–99. 591
- (1997b). "Konsten som grepp : formalistiska strategier och emblematiske tanke- 592
 former". In: *Tidskrift för litteraturvetenskap* 26.1, 26–58. 593
- Ahnert, Ruth, Emma Griffin, Mia Ridge, and Georgia Tolfo (2023). *Collaborative Historical* 594
Research in the Age of Big Data: Lessons from an Interdisciplinary Project. Elements in 595
 Historical Theory and Practice. Cambridge University Press. ISBN: 1009175556. 596
- Berry, David and Anders Fagerjord (2017). *Digital Humanities: Knowledge and Critique in* 597
a Digital Age. Polity. ISBN: 0745697666. 598



23. "Konstverket är således inte autonomt, utan *inriktat* på en viss typ av betraktande, som det samtidigt, via 599
 sina inbyggda grepp, *frammanar*" (Agrell 1997b, 28).

- Bode, Katherine (2018). *A World of Fiction: Digital Collections and the Future of Literary History*. Digital Humanities. The University of Michigan Press. ISBN: 0472130854. 599 600
- (2020). “Data beyond representation: From computational modelling to performative materiality.” In: MLA Convention 2020. <https://katherinebode.wordpress.com/home/mla-convention-2020/>. 601 602 603
- (2023). “What’s the Matter with Computational Literary Studies?.” In: *Critical Inquiry* 49.4, 507–529. [10.1086/724943](https://doi.org/10.1086/724943). 604 605
- Borg, Ingwer and Patrick J. F. Groenen (2005). *Modern Multidimensional Scaling. Theory and Applications*. Springer Series in Statistics. Springer. ISBN: 9780387289816. 606 607
- Broberg, Jan (2006). “Elegant och elegiskt”. In: *Sydsvenska Dagbladet* 02/10/2006. 608
- Brottrager, Judith, Annina Stahl, Arda Arslan, Ulrik Brandes, and Thomas Weitin (2022). “Modeling and Predicting Literary Reception. A Data-Rich Approach to Literary Historical Reception”. In: *Journal of Computational Literary Studies* 1.1, 1–27. <https://doi.org/10.48694/jcls.95>. 609 610 611 612
- Creswell, John W and J David Creswell (2022). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches (6th ed.)*. SAGE. ISBN: 9781071817940. 613 614
- Dobson, James E. (2019). *Critical Digital Humanities: The Search for a Methodology*. Topics in the Digital Humanities. University of Illinois Press. ISBN: 0252084047. 615 616
- Drucker, Johanna (2011). “Humanities Approaches to Graphical Display”. In: *Digital Humanities Quarterly* 5.1. <https://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html>. 617 618 619
- (2021). *The Digital Humanities Coursebook*. Routledge. ISBN: 9781003106531. 620
- Dunning, Ted (1993). “Accurate Methods for the Statistics of Surprise and Coincidence”. In: *Computational Linguistics* 19.1, 61–74. <https://dl.acm.org/doi/10.5555/972450.972454#sec-terms>. 621 622 623
- Forser, Tomas (2002). *Kritik av kritiken: 1900-talets svenska litteraturkritik*. Anthropos. ISBN: 9185722235. 624 625
- Foucault, Michel (1972). *Archaeology of Knowledge: and the Discourse on Language (translated by A. M. Sheridan Smith)*. Tavistock Publications Ltd. ISBN: 9780422736503. 626 627
- Guldi, Jo (2023). *The Dangerous Art of Text Mining: A Methodology for Digital History*. Cambridge University Press. ISBN: 9781009262989. 628 629
- Gumbrecht, Hans-Ulrich (1997). *In 1926: Living at the Edge of Time*. Harvard University Press. ISBN: 0674000560. 630 631
- Ingvarsson, Jonas (2021). *Towards a Digital Epistemology: Aesthetics and Modes of Thought in Early Modernity and the Present Age, 2nd Ed.* Palgrave Pivot. Palgrave. ISBN: 9783030787233 632 633
- Ingvarsson, Jonas, Daniel Brodén, Lina Samuelsson, Victor Wåhlstrand Skärström, and Niklas Zechner (2022). “The New Order of Criticism. Explorations of Book Reviews Between the Interpretative and Algorithmic”. In: <https://ceur-ws.org/Vol-3232/paper20.pdf>. 634 635 636 637
- Jockers, Matthew Lee (2013). *Macroanalysis: Digital methods and literary history: Topics in the digital humanities*. Topics in the digital humanities. University of Illinois Press. ISBN: 9780252079078. 638 639 640
- Johnson, R Burke, Anthony J. Onwuegbuzie, and Lisa A Turner (2007). “Toward a Definition of Mixed Methods Research”. In: *Journal of Mixed Methods Research* 1.2, 112–133. <http://dx.doi.org/10.1177/1558689806298224>. 641 642 643
- Lekander, Nina (2006). “En fiffigare finhet”. In: *Expressen* 10/04/2006. 644
- Lingebrandt, Ann (2006). “Gnistrande”. In: *Helsingborgs Dagblad* 10/04/2006. 645

- Liu, Alan (2014). "Theses on the Epistemology of the Digital: Advice for the Cambridge Centre for Digital Knowledge". In: *Alan Liu (blog)*. <https://liu.english.ucsb.edu/theses-on-the-epistemology-of-the-digital-page/>.
- McInnes, Leland, John Healy, and James Melville (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv: 1802.03426 [stat.ML].
- Moretti, Franco (2013). *Distant Reading*. Verso. ISBN: 9781781680841.
- Nelson, Laura K. (2020). "Computational Grounded Theory: A Methodological Framework". In: *Sociological Methods Research* 49.1, 3–42. <https://doi.org/10.1177/0049124117729703>.
- North, Michael (2001). "Virtual Histories: The Year as Literary Period". In: *Modern Language Quarterly* 62.4, 407–424.
- Oberbichler, Sarah, Emanuela Boros, Antoine Doucet, Jani Marjanen, Eva Pfanzelter, Juha Rautiainen, Hannu Toivonen, and Mikko Tolonen (2021). "Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians". In: *Journal for the Association for Information Science and Technology* 2.73, 225–239. <https://doi.org/10.1002/asi.24565>.
- Piper, Andrew (2020). *Can We Be Wrong? The Problem of Textual Evidence in a Time of Data*. Elements in Digital Literary Studies. Cambridge University Press. ISBN: 9781108922036.
- Piper, Andrew and Mark Algee-Hewitt (2014). "The Werther Effect I: Goethe, Objecthood, and the Handling of Knowledge". In: *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*. Vol. 146. Studies in German Literature Linguistics and Culture, 155–184. <https://doi.org/10.1515/9781571138903-008>.
- Ramsay, Stephen (2011). *Reading Machines: Toward an Algorithmic Criticism*. Topics in the Digital Humanities. University of Illinois Press. ISBN: 0252078209.
- Rekathati, Faton (2021). "The KBLab Blog: Introducing a Swedish Sentence Transformer". In: *The KBLab Blog*. [10.31235/osf.io/w48rf](https://doi.org/10.31235/osf.io/w48rf).
- Rockwell, Geoffrey and Stéfan Sinclair (2016). *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. MIT Press. ISBN: 9780262332064.
- Rydén, Per (1987). *Domedagar: svensk litteraturkritik efter 1880*. Litteraturvetenskapliga institutionen, Lund. ISBN: 9185152161.
- Samuelsson, Lina (2013). *Kritikens ordning: svenska bokrecensioner 1906, 1956, 2006 (diss.) bild, text form*. ISBN: 9789198044713.
- Shan, Yafeng (2023). "Philosophical Foundations of Mixed Methods Research". In: *Philosophy Compass* 17.1, 1–12. [10.1111/phc3.12804](https://doi.org/10.1111/phc3.12804).
- Shklovsky, Viktor (1990 (1929)). "Art as Device". In: *Theory of Prose, translated by Benjamin Sher*, 1–14.
- Spärck Jones, Karen (1972). "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". In: *Journal of Documentation* 28.1, 11–21. <https://doi.org/10.1108/eb026526>.
- Tammerman, Ann-Mari (2006). "Fantasier lev ut". In: *Upsala Nya Tidning* 03/11/2006.
- Underwood, Ted (2018). *Distant Horizons: Digital Evidence and Literary Change*. Digital Humanities. The University of Chicago Press. ISBN: 9780226612836.
- Van Cranenburgh, Andreas, Karina van Dalen-Oskam, and Joris van Zundert (2019). "Vector space explorations of literary language". In: *Language Resources and Evaluation* 53.4, 625–650.

Measuring Literary Quality Proxies and Perspectives

Pascale Feldkamp¹ 
Yuri Bizzoni¹ 
Ida Marie S. Lassen¹ 
Mads Rosendahl Thomsen² 
Kristoffer L. Nielbo¹ 

1. Center for Humanities Computing, Aarhus University , Aarhus, Denmark.
2. Comparative Literature – School of Communication and Culture, Aarhus University , Aarhus, Denmark.

Citation

Pascale Feldkamp, Mads Rosendahl Thomsen, Kristoffer L. Nielbo, and Yuri Bizzoni (2024). “Measuring Literary Quality. Proxies and Perspectives”. In: *CCLS2024 Conference Preprints* 3 (1). [10.26083/tuprints-00027391](https://doi.org/10.26083/tuprints-00027391)

Date published 2024-05-28

Date accepted 2024-04-04

Date received 2024-01-09

Keywords

literary quality, literary success, canonicity, literary culture, computational literary studies, 19th-20th century literature

License

CC BY 4.0 

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 3rd Annual Conference of Computational Literary Studies at Vienna, Austria, in June 2024.

Abstract. Computational studies of literature have adopted approaches from statistics and social sciences to perform large scale studies of fiction, and recent work has sought to approximate the success of literary texts using some proxy for literary quality, such as collections of human judgments, sales-numbers or lists indicating canonicity. However, most quantitative studies of literary quality use one such measure as a golden standard of literary judgement without fully reflecting on what it represents. Conclusions drawn from these studies are nonetheless bound to mirror a particular conception of literary quality associated with the chosen metric. To address this issue, we provide a discussion of the interrelation of various “proxies of literary quality” within a corpus of novels published in the US in the late 19th and 20th century, performing correlations and comparisons across 14 different proxies. We start with a heuristic distinction between expert-based literary judgments, such as those represented by college syllabi and literary anthologies, and crowd-based judgments, such as GoodReads’ ratings, and explore the differences between these and other proxies that fall in-between, such as library holding numbers, prestigious literary prizes, and classics book series. Our findings suggest that works favored in expert-based judgments tend to score lower on GoodReads, while those long-listed for awards tend to score higher and enjoy greater circulation in libraries. Generally, two main kinds of “quality perception” emerge as we map the literary judgment landscape: one associated with canonical literature, and one with more popular literature, which may indicate that judgements of canonicity or literariness are not equal to popularity among readers. Additionally, our study suggests that prestige in genre-literature, as represented by main genre-fiction awards such as the Hugo or World Fantasy Award, constitute distinct proxies on their own, though more closely aligned to popular than canonical proxies.

1. Introduction

The concept of quality in literature is a fascinating riddle: it would seem that the idiosyncratic nature of reading precludes any objective standard for what constitutes a “good” book – and yet certain texts seem to have an enduring appeal: they interest

readers across time and national borders and are consecrated in the institutional canons of different cultures. This paradox lies at the heart of discussions about what literary quality is, as well as of attempts to define, measure or predict it.¹

The challenge of defining literary quality is complicated by the diversity of preferences of individual readers and reader-types (Riddell and Dalen-Oskam 2018), and even the tendency of readers to change their opinion about a text (Harrison and Nuttall 2018; Kuijpers and Hakemulder 2018). Moreover, the question of what constitutes literary quality and where it resides (in style, plot, emotional engagement, themes, etc.) quickly becomes a complicated matter of its own, one that schools of literary criticism have grappled with in many different ways (Bjerck Hagen et al. 2018).

While the evaluation of texts and the question of quality has naturally been prominent in literary criticism, its significance has often been eclipsed within scholarly discourse by various disciplinary shifts. Ethical and postcolonial shifts calling attention to canon representativity (Peer 2008), methodological transformations of the 20th century moving the focus from evaluation towards interpretation (Bjerck Hagen et al. 2018), and the expansion of the conceptual boundaries of literature to encompass texts ideologically opposed to aestheticism or “pleasing” the reader (Wellek 1972), are examples that have played a role in making terms like “literary quality”, or “classics” unpopular – said to belong to the “precritical era of criticism itself” (Guillory 1995). However, to attribute the longevity or popularity of certain books to purely contextual factors and reject the notion of literary quality altogether would seem to be at odds with both the resilience of canons and consensuses among readers at the large scale, which appear far from volatile (Archer and Jockers 2017; Bizzoni et al. 2021; Maharjan et al. 2017, 2018; Wang et al. 2019).² Moreover, literary cultures have consistently established and upheld proxies of literary excellence in practice, such as literary awards, classics book series, or prescriptions in creative writing courses. Thus, a disparity appears to have arisen between a scholarly “denial of quality” (Wellek 1972) and the multitude of evaluative criteria actualized within literary culture.

With recent computational inquiry into literary studies, and sizeable attempts at quantifying “quality”, this disparity is even more apparent. The stricter conditions of quantitative analysis – operationalizing traditional disciplinary concepts – bring the complexity of the idea of “quality” in literature to the fore. Computational studies of literary preferences have found that reader appreciation or success can to some extent be predicted by stylistic features (Cranenburgh and Bod 2017; Dalen-Oskam 2023; Maharjan et al. 2017), as well as narrative features such as plot (Jockers 2015), emotional valence and flow

1. In this article, we will use the term “literary quality” in a general sense – as “quality in literature” – independently from kinds of texts (e.g. high-brow/low-brow) and evaluative groups (e.g. universities, online communities). That is, we do not intend to imply perceived *literariness*, but rather we aim to denote some form of appreciation of a literary work. In other words, our focus is not on whether a text appears to be high-brow, have sophisticated references to other works of literature and so forth, but rather on whether a text is considered outstanding by different types of readership.

2. A very Marxist reader, Leon Trotsky, observed how the historical and aesthetic dimensions of art are utterly independent: “If I say that the importance of the Divine Comedy lies in the fact that it gives me an understanding of the state of mind of certain classes in a certain epoch, this means that I transform it into a mere historical document, for, as a work of art, the Divine Comedy must speak in some way to my feelings and moods... Dante was, of course, the product of a certain social milieu. But Dante was a genius. He raised the experience of his epoch to a tremendous artistic height. And if we, while today approaching other works of medieval literature merely as objects of study, approach the Divine Comedy as a source of artistic perception, this happens not because Dante was a Florentine petty bourgeois of the 13th century but, to a considerable extent, in spite of that circumstance” (Trotsky 1974)

(Maharjan et al. 2018; Reagan et al. 2016; Veleski 2020), or the predictability of novels' sentiment-arcs (Bizzoni et al. 2022a,b, 2021) – not to mention text-extrinsic features such as genre, promotion, author visibility and gender (C. W. Koolen 2018; Lassen et al. 2022; Wang et al. 2019). While such studies point to the existence of certain consensuses, it should be noted that these studies define the concept of success or quality very differently. The first and possibly most complex task of quantitative studies of literary quality is that of defining a “proxy” of quality itself: from where should we take the judgments we intend to explain?

In computational literary studies, a “proxy” serves as a formal method for approximating abstract constructs or concepts through operationalization. Proxies bridge qualitative interpretation with quantitative methodologies: they translate constructs or concepts, like “quality in literature”, into measurable variables. A “quality proxy” thus means a specific operationalization of appreciation among many. For example, we might differentiate between literary “fame” and “popularity”, since fame, such as the fame of James Joyce’s *Ulysses* does not necessarily mean that it is widely read. These different forms of quality may be measured in dissimilar ways – i.e., through different “proxies” – for example by looking at how often a book is subject of literary scholarship, vs. how many copies it sells, or how often it is rated on GoodReads.³

A large number of quantitative and computational works have used votes of popularity to approximate judgments of literary quality. GoodReads is a widely used resources (Jannatus Saba et al. 2021; Maharjan et al. 2017; Porter 2018), also since it provides a single scale of scores averaged on large numbers of individual readers. The “GoodReads approach” can be seen as an example of “counting votes”, where the majority decides: the number of votes or a higher average score defines quality. On the polar opposite, a number of studies have used individual canon-lists of works selected by individual or cohorts of established literary scholars to approximate what are “quality works” of literature (Mohseni et al. 2022). Canon-lists or anthologies represent the idiosyncratic perspective of the few. Naturally this approach has advantages and disadvantages: “canon-makers” with or without institutional backing presumably have a vast knowledge of literature, but the criteria of selection are not always explicit and may or may not represent a particular taste or kind of reader. These limitations are, however, are homologous to those of the “GoodReads approach” where criteria and type of reader is likewise unknown (is it a particular type of reader who rates books online?). Studies have also modelled literary quality by whether or not a book has won a literary award (Febres and Jaffe 2017), which is akin to the “canon perspective”, but may differ in terms of the institutional affiliation of actors. Another method is to seek judgements of quality in the reading population (C. Koolen et al. 2020). Yet efforts of gauging readers’ conceptions of quality with sophisticated questionnaires is naturally limited by the difficulty and costs of conducting extensive surveys. Either of these approaches nevertheless runs the risk of modelling but one kind of “literary quality”, prompting reflections on how they are related. While some studies have tried to map the relations and overlaps between kinds of quality proxies (Manshel et al. 2019; Porter 2018), usually experiments are conducted on a limited scale, either in terms of corpus, or in terms of

3. At present, *Ulysses* has 124,536 ratings on GoodReads and a relatively low average rating of 3.75, compared to works such as Suzanne Collins’ *The Hunger Games* and J.K. Rowling’s *Harry Potter and the Sorcerer’s Stone*, with above 8 million ratings and average ratings above 4.3.

the number and types of quality proxies considered. 83

The question remains of how different proxies relate to an overall concept of literary quality: do different proxies offer windows or perspectives into a more or less universal perception of quality, or do such proxies represent vastly different forms of appreciation? 84
85
86
Do, for instance, GoodReads scores mirror, on a larger scale, the selection of experts, 87
such as for literary anthologies, or do they diverge to such an extent that we may assume 88
that what is judged to be “quality” in each proxy is based on different criteria? 89

To address the question of differences between quality proxies, we collected 14 different 90
possible proxies for literary quality, ranging from popular online platforms to university 91
syllabi and prestigious awards, and used them to annotate a corpus of over 9,000 novels 92
(note that we do not analyze the texts themselves in this article).⁴ Our central question 93
was whether and to what extent these metrics measure the same thing: if the “quality” 94
measured by GoodReads data differs from that represented by the number of library 95
holdings, the two metrics will have nothing in common; if instead there is a significant 96
overlap - that is, books popular on GoodReads are also acquired by many libraries - they 97
will correlate. To the best of our knowledge, this is the first study that tries to compare 98
several judgements of literary quality on a large collection of modern titles, trying to 99
understand, by a rigorous approach, the relation between them. 100

2. Related Works 101

Studies have found that there seems to be a consensus among readers about what works 102
are “classics”. Walsh and Antoniak (2021) tested the relation between GoodReads’ 103
Classics, a user-compiled list, and titles included in college English syllabi (as collected 104
by the OpenSyllabus project), showing that there is a significant overlap between what 105
is perceived as classics on GoodReads and what appears on college syllabi (Walsh 106
and Antoniak 2021). Thus, users seem to be replicating a particular perception of the 107
“canonicity” of titles. 108

Similarly, Koolen et al. (2020) surveyed a large number of Dutch readers, asking for both 109
judgments of how “enjoyable” and how “literary” a novel is, and have shown that there 110
is a more substantial consensus among readers about “literariness” than “enjoyability”- 111
ratings, which appear less predictable than those of literariness (C. Koolen et al. 2020). 112

Another study by Porter et al. (2018) sought to model differences in popularity and 113
prestige in their corpus, using, on the one hand, GoodReads’ average ratings and, on 114
the other hand, the Modern Language Association’s database of literary scholarship, 115
counting the number of mentions of an author as the primary subject of a scholarly work. 116
They show that there is a clear difference in the equilibrium between popularity and 117
prestige across genres. Books from genres like sci-fi are rated very often on GoodReads 118
but are sparsely represented in scholarly work, while poetry exhibits an opposite ten- 119
dency. Based on Pierre Bourdieu’s conceptualization of the literary field, they define two 120
axes of literary “success”, prestige and popularity as online popularity (on GoodReads) 121
and prestige among literary scholars (represented in the MLA database), so that their 122

4. See section 4 for a discussion of this corpus, which, it should be noted, is heavily skewed toward American and Anglophone authors.

"map" risks to look overly neat. Literary scholars, for example, may not be the primary nor most important actors in processes of literary prestige, and Manshel et al. (2019) have shown how literary prizes – appointed by committees who may be either authors themselves, scholars, or lay-readers – appear to have an important role in positively influencing both prestige and popularity.⁵

While only a few studies have tried to measure differences and convergences of literary quality judgments quantitatively, the question of how literary cultures evaluate texts has been central to sociological approaches to literature. Especially the attempts of Pierre Bourdieu to "map" the literary field is central in this context and has given rise to a string of seminal works on power dynamics in literary cultures (Bennett 1990; Casanova 2007; Guillory 1995; Moretti 2007). Bourdieu's map of the French "literary field" (1) focuses on literary genres and their interrelation in terms of prestige (and not actors in literary quality judgments *per se*). However, Bourdieu makes an important distinction between types of audiences and considers "consecration by artists, by institutions of the dominant classes, and by popular success" as distinct axes, that are more or less mutually exclusive.⁶

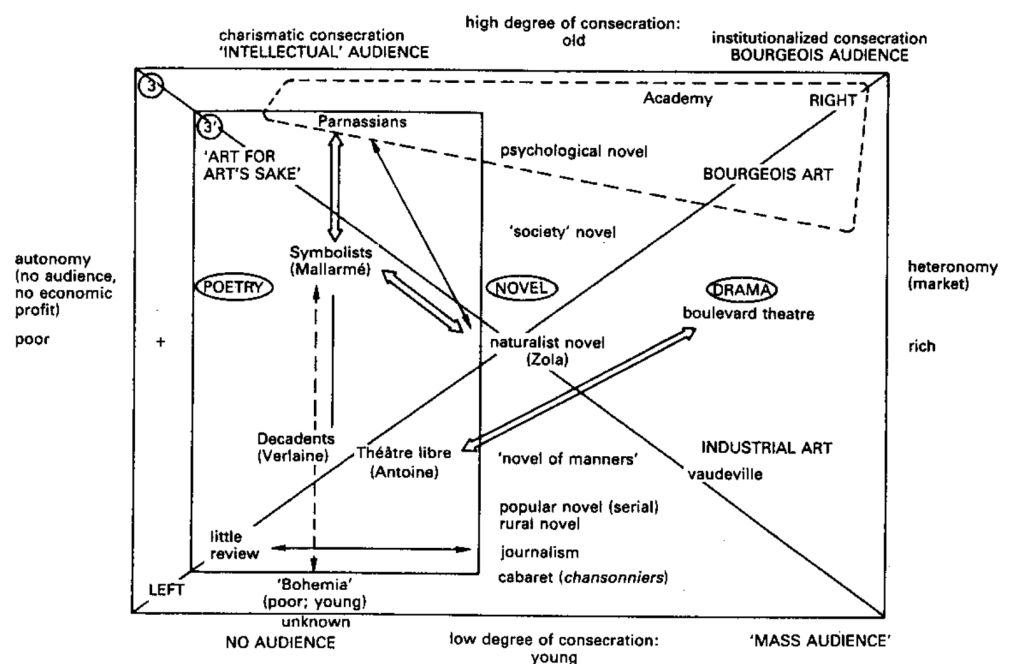


Figure 1: Bourdieu's French literary field of the late 19th century, with audience or popularity on the x-axis and consecration or prestige on the y-axis.

While the relation between these actors is only sketched out (and it is the present study's aim to inspect these more closely), Bourdieu's map can serve as a heuristic conceptualization of types of actors in literary quality judgments. Here, the idea of expert-based and crowd-based literary judgments is apparent at either pole, represented

5. Using the same definitions of popularity and prestige as Porter et al. (2018), it seems that whether or not books had received a prize significantly raised the probability of both being popular and prestigious (Manshel et al. 2019).

6. Bourdieu writes: "there are few fields [beyond the literary] in which the antagonism between the occupants of the polar positions is more total" (Bourdieu 1993, p. 46).

on one side by intellectual and bourgeois audiences, recognized intellectuals such as “Parnassians” and institutions such as *l’Académie Française*; and on the other hand by amateur and mass audience such as the artistic underdogs “bohemia” and popular media. As Porter et al. (2018) have shown, “on a broad level, real-world data about popularity and prestige appear to confirm Bourdieu’s intuitions” (Porter 2018). In their visualization the genres “Mystery & Thriller” and “Science Fiction & Fantasy” appear where Bourdieu places the “Popular novel” (at low consecration and high economic profit), while poetry is in the upper left area of the map, representing high prestige and low popularity. However, the focus of Porter et al. is on the right-hand part of Bourdieu’s map, with prestige defined as institutional or academic consecration: the place for literary works in academia. For a more comprehensive “map” based on real world data, various actors, including literary prizes and publishers, should be considered. It is to this end that the present paper uses a sizeable corpus to examine the interrelation judgments of a type of “success” in the literary field, including various actors under the general categories of expert-based and crowd-based literary success based off Bourdieu’s “map”. We discuss the selection of various proxies and what they represent, before moving on to looking at their distribution and interrelation in the Chicago corpus.

3. Selecting Types of Literary Judgments

By considering various proxies of literary quality, our aim was to examine the interrelation of conceptually different types. We considered three distinct approaches to literary quality:

1. Approaches that seek to approximate literary canonicity or quality in an institutional sense, looking at which works or authors are included in school or university syllabi, literary anthologies, or that win literary awards.
2. Approaches that seek to approximate reader-popularity, basing proxies of literary quality on larger populations, where the selection process appears more “democratic”, seeking the quality perception of “layman readers”, by collecting user-generated data such as ratings from sites like GoodReads, Amazon, or Audible.
3. In-between approaches that seek to measure the market success or market resilience of works, looking at, for example, sales figures.

3.1 Expert-based Quality Proxies

Expert-based proxies of literary quality may to an extent be synonymous with canonicity, that is, consecration and institutionalization. Often, quantitative studies of reader appreciation define canonicity or prestige through canon lists compiled by, i.a., individual magazines (Vulture 2018, as in Porter 2018), editors (Karlyn and Keymer n.d., as in Algee-Hewitt et al. 2018), or literary scholars (Bloom 1995, as in Mohseni et al. 2022). However, such lists resemble personal canons that may not have a wide reach, e.g., it is unclear how widely accepted Harold Bloom’s chosen canon is among scholars. In this study, we have preferred canonicity proxies that do not depend on the selection of

very few. To examine expert-based proxies of literary quality and estimate the amount of “canonic” literature in our dataset, we marked all titles by authors that appear in selected institutional or user-compiled proxies that indicate literary prestige: a literary anthology, the most assigned titles in English Literature course syllabi, literary awards, and a publisher’s classics series.

3.1.1 Anthologies

Students of English or of Literature will often be acquainted with anthologies that are compiled in part for educational use, facilitating easy access to some key works. In this context, the Norton Anthology in particular is a leading literary anthology (Pope 2019), with diachronic series of English and American literature that are widely used in education (Shesgreen 2009). For the present study, we marked all titles in our corpus written by authors mentioned in these two series, where the anthology of English Literature is the most widespread (Ragen 1992).

3.1.2 Syllabi

While titles assigned on Literature or English syllabi surely vary across colleges and regions, it is possible to find trends and most assigned titles via large collections of data, such as by the OpenSyllabus project, which has collected 18.7 million college syllabi in an attempt to map the college curriculum.⁷ From this data, we took all titles in our corpus by authors who appear as authors of one of the top 1000 titles assigned in English Literature college syllabi.

3.1.3 Awards and Long Lists

We collected long-listed titles (winners and finalists) for both prestigious general literature awards: The Nobel Prize in Literature, the Pulitzer Prize, the National Book Award (NBA); as well as various genre-based awards (for the full list, see Table 1). The choice of long-lists allowed us to have a more titles annotated, but also an annotation possibly less susceptible to the extrinsic factors that can influence the choice of a winner among a small selection of candidates in the moment (politics, topic, prominence of the author, and so forth).

Manshel et al. (2019) have shown that winning an award does contribute to long-term prestige – but also popularity – of titles in academia as well as on GoodReads. Interestingly, Kovács and Sharkey (2019), found that while awards may initially make a title more popular and gather more ratings on GoodReads, this may also affect a drop in average rating as the reception of a book becomes polarized. As such, the choices of award-committées do seem to be in touch with the general public, but also diverge from consensus among readers at the very large scale Kovács and Sharkey 2014. We keep genre-awards and more general literary awards separate in our analysis, as we expect titles to be received differently across genres. As our corpus catalogues mainly American and British authors, the focus of our selection was the topmost known committee-based awards in anglophone literary culture.

7. See: <https://www.opensyllabus.org>.

3.1.4 Classics Series 223

Various large publishing houses, like Vintage or Penguin⁸, maintain a classics series. As Penguin is arguably one of the biggest publishers of anglophone literature (Alter et al. 2022), we marked all titles or authors in our corpus that appear in their classics series. We looked at both the specific titles (title-based) with matches in our data, and at all titles by authors featured in the series (author-based), keeping these separate in our analysis.

3.2 Crowd-based Quality Proxies 230

Where proxies of quality are clearly vote-based and the result of equal weight for each individual in a large population, we call them “crowd-based”, remembering, however, that these votes are cast within a system and social structures (e.g., on the social platform GoodReads), which are not non-hierarchical as the term “crowd-based” generally implies, nor isolated from tendencies of expert-based proxies. For example, the canonicity perception of GoodReads’ users may have more to do with expert-based proxies of literary quality than we think (Walsh and Antoniak 2021). Among crowd-based measures, we have opted for GoodReads and Audible average rating (number of “stars” given to a title) and rating count (number of votes). We also used two GoodReads user-compiled lists: the “GoodReads classics” and the “Best books of the 20th century” which may represent canonic literature but at a larger scale than expert-based canonicity lists.

3.2.1 GoodReads 243

GoodReads is a social network or “social catalogue site” with links to other social networks (Facebook, Twitter, Instagram, and LinkedIn), designed for readers to discover, review, and share their thoughts. Otis Chandler, GoodReads’ co-founder, states on the homepage that the idea was to make a social forum akin to looking at the bookshelf at a friend’s house: “When I want to know what books to read, I’d rather turn to a friend than any random person or bestseller list.” With its 90 million users, GoodReads arguably offers an insight into reading culture “in the wild” (Nakamura 2013), as it catalogues books from a wide spectrum of genres and derives book-ratings from a heterogeneous pool of readers in terms of background, gender, age, native language and reading preferences (Kousha et al. 2017). GoodReads’ average ratings represent the average user rating of titles. Rating ranges from 0 stars (indicating low appreciation) to 5 stars (indicating high appreciation). The average score provides a general indication of the book’s reception, but is problematic as it conflates types of literary appreciation, i.e., satisfaction, enjoyment, and evaluation, to one scale. While it is important to note that these GoodReads’ ratings and number of raters (rating count) do not present an absolute measure of literary quality or even popularity (GoodReads did start with predominantly American users), they do offer a valuable perspective on a work’s overall popularity among a diverse population of readers. Beyond ratings, GoodReads also compiles vote-based lists and “shelves”, arranged according to the titles most often either assigned to a particular list or tagged to a particular shelf. These are, for example, GoodReads’ Classics, Best Books of the 20th Century, The Worst Books of All Time, etc.

8. See: <https://www.penguin.com/penguin-classics-overview/>.

For the present study, we used the top 100 of a popular list, the Best Books of the 20th Century⁹, and a shelf, the GoodReads' Classics¹⁰, where titles were *listed* by users 600 to 10,000 times, and *shelved* 15,588 to 64,903 times, respectively.

3.2.2 Audible

We use the average rating and number of ratings of title on Audible, the Amazon audiobook service. Like GoodReads, the site uses a five-star scale for user ratings, however, the amount of users and the rating counts are significantly lower for Audible compared to GoodReads: while Dan Brown's *The Da Vinci Code* has 2,259,837 ratings on GoodReads, it has 3,225 ratings on Audible at the moment of writing, and the average Audible rating is inflated in comparison to the GoodReads' average rating for our corpus, which may be an effect of a smaller number of users.

3.3 In-between Quality Proxies

The number of copies sold is often adopted as a reliable standard to estimate the success novels, for example to gauge a set of signals that land a book on the bestseller list Archer and Jockers 2017. It is interesting because a proxy like sales figures seems to stand in-between the crowd- and expert-based proxies, including a degree of resilience or canonicity of titles (as classics will continue to sell) as well as popular demand. The NPD BookScan¹¹, for example, is a popular resource in this regard (as used in Wang et al. 2019), which provides data for the publishing industry both regarding genre, prices, and weekly sales figures for all books published in the US since 2003. It is clear that such data is market- and location-specific, and is only an option for studies of more contemporary works. As with any other approximation of literary quality, but perhaps especially pertaining to sales figures, the issue is both that data pertains to more recent publications, is not readily available, and that contextual factors may influence the data. For book-sales, Wang et al. (2019) have shown that marketing, the particular publishing house, and visibility of the author plays a central role for sales numbers. Instead of sales-figures, we may use proxies that also include an aspect of resilience and popular success. Thus, we have used the number of libraries holding a given title on Worldcat and the number of translations of a work into other languages, as well as the author's presence on Wikipedia and a bestseller list. The number of library holdings as a proxy is conceptually intermediate between a completely free, crowd-based vote count and an expert-driven single choice, as the list of books held by libraries depends on both popular demand (of library-card holders) and expert choices (librarians). Similarly, the translational success of a work shows a degree of market success (if translation is seen as a token of publishers seeking to expand sales of bestselling books outside the national market) and canonicity or resilience (if translation is seen as a token of a work's cultural longevity or durable popularity). Similarly, Wikipedia rank and bestseller lists appear conceptually to include a degree of resilience and popular success.

9. See: https://www.GoodReads.com/list/show/6.Best_Books_of_the_20th_Century.

10. See: <https://www.GoodReads.com/shelf/show/classics>.

11. See: <https://www.npd.com/industry-expertise/books/>.

3.3.1 Library Holdings 303

For each title, the Chicago Corpus provides the number of US libraries holding a copy of 304
it. The idea is that libraries' choices could help indicate a canon that is not arbitrary (as 305
libraries supposedly respond to institutional demands like school reading requirements) 306
but also remains essentially crowd-based (as libraries also respond to other demands, 307
including from leisure-readers). Libraries are institutions managed by experts, but 308
adding together the choices of thousands of different libraries allows the selection to 309
partly overcome the risks involved in electing one single, if well-informed, authority. 310

3.3.2 Translations 311

The *Index Translationum* database¹² collects all translations published in ca. 150 UNESCO 312
member states, compiled from their local bibliographical institutions or national libraries. 313
It catalogues more than 2 million works across disciplines. Note that the database was 314
created in 1979 and stopped compiling in 2009. Thus, we are not looking at the most 315
translated works through time, where the "classics" may be more frequent, but at a 316
particular period, and the results should be interpreted with that in mind. 317

3.3.3 Wikipedia Author-page-rank 318

Using wikipedia page-views, that is, the number of times visits to an author's page on 319
Wikipedia is also sometimes used as a proxy for popularity or resilience. Hube et al. 320
(2017) have used Wikipedia metrics to measure in the centrality of authors in digital 321
space (Hube et al. 2017), with a variation of page-rank, the original google algorithm. 322
It is an efficient way to navigate graphs: hubs or author-pages on Wikipedia that have 323
the highest number of other pages referencing them have a higher rank, which means 324
a higher rank for more referenced authors. The Wikipedia page rank thus measures 325
a type of "canonicity" of authors, but also their presence in the popular and cultural 326
sphere, if we consider that Wikipedia-pages are created both by experts and lay-readers. 327
For the present study, we used Wikipedia author-page (WAP) rank, where it should be 328
noted that ranks refer to authors, so that books by the same author will have the same 329
rank, independent from differences between individual titles. 330

3.3.4 Bestseller Lists 331

To gauge the commercial success of titles, we also marked titles in our corpus that were 332
also extant in the Publisher's Weekly American 20th century bestseller list.¹³ Publishers 333
Weekly is a trade news magazine which is published once a week (from 1872) and 334
targeted at agents within the field: publishers, literary agents, booksellers, and librarians. 335
While sales numbers are considered, the full set of selection criteria for the list are 336
unknown. 337

12. See: <https://www.unesco.org/xtrans/bsform.aspx>.

13. Extracted from the database by John Unsworth at the University of Illinois: <https://web.archive.org/web/20111014055658/http://www3.isrl.illinois.edu/~unsworth/courses/bestsellers/picked.books.cgi>.

4. Dataset: the Chicago Corpus

338

In order to quantify the possible convergence of these proxies, we need a dataset of chosen titles. A large dataset of titles would allow us to see whether different ways of scoring or judging literary works tend to have something in common (e.g. valuing similar texts) or not. Ideally, for a first experiment, we would also require a selection of texts that are not too widespread in time, written/read in the same language, and in the same narrative form (e.g. all prose novels).

We base our study on the *Chicago corpus*,¹⁴ a corpus of over 9,000 manually compiled novels that were either written or translated into English and published in the US between 1880 to 2000. The corpus was compiled based on the number of libraries holding a copy of the novel, with a preference for novels with more holdings. Beyond responding to the constraints detailed above, the Chicago corpus allows us to access, the number of libraries holding each title in the US. Moreover, the Chicago corpus has been curated and used by teams of literary scholars, and offers access to the full text of all its titles, which makes a study of correlations between quality judgments and textual features possible in the future.

Because of its unique method of compilation, the Chicago corpus is a rare dataset in terms of its diversity: it spans works from genre-fiction and popular fiction (i.a., Isaac Asimov, Agatha Christie, George R. R. Martin), to seminal works from the entire period, central modernist and postmodernist texts (e.g. James Joyce's *Ulysses* and Don DeLillo's *White Noise*), as well as winners of the Nobel Prize (i.a., Ernest Hemingway, William Faulkner, Toni Morrison), and other prestigious literary awards (i.a., Cormac McCarthy). As such, it represents a sizeable subsection of both prestigious or "canonic" works, as well as popular and genre-fiction classics.

It should be noted that the Chicago corpus contains only works either written in or translated into English, and therefore exhibits an over-representation of Anglophone authors.

We previously discussed the essential characteristics of these proxies of literary quality, as well as the kind of outlook on literary judgments that they seem to model or approximate. Some are on the free and vote-counting end of the spectrum, putting equal weight to the rating of each user. Resources like the Norton collection, as well as prestigious literary awards, arguably fall on the expert-based side of the spectrum, as they are managed by small groups of authoritative readers, usually professional literary critics.

By collecting and annotating proxies of quality for titles in the Chicago corpus, we collected a wide variety of "quality judgments" for each title, some continuous (as GoodReads' average ratings) or progressive (as the number of library holdings), some discrete, as any list that either includes or excludes titles. This, as we will see, constitutes a fundamental divide between our measures, and in some sense mirrors two different ways of assessing literary quality. The resources that in one way or another score each book – number of ratings, number of library acquisition, average rating – represent quality on a continuum, while the resources that select books – anthologies, syllabi and

14. For more on the corpus, see the resource at: https://github.com/centre-for-humanities-computing/chicago_corpus.

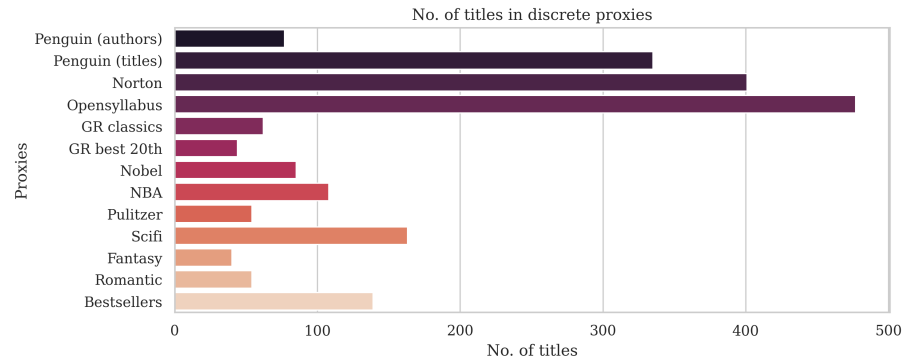


Figure 2: Sizes of discrete proxies in our corpus.

	Titles
National book award	108
Pulitzer prize	53
Nobel prize*	85
Scifi awards	163
Hugo award	
Nebula award	
Philip K. Dick award	
J.W. Campbell award	
Prometheus award	
Locus sci-fi award	
Fantasy awards	40
World fantasy award	
Locus fantasy award	
British fantasy award	
Mythopoeic award	
Romantic awards*	54
Rita awards*	
RNA awards*	
<hr/>	
Norton anthology*	401
OpenSyllabus*	477
Penguin classics series (titles)	77
Penguin classics series*	335
GoodReads' classics*	62
GoodReads' best books of the 20th century*	44
<hr/>	
20th century bestsellers (Publisher's Weekly)	139
Wikipedia AP rank*	3558
Translations	5082
GR avg. rating	8989
GR rating count	8989

Table 1: Number of titles in the corpus per quality proxy. Proxies followed by * are author-based: For these, we included all titles extant in the corpus by the author mentioned, either due to the scarcity of awards in the genre or the nature of the award/list, e.g., the Nobel prize given to authors rather than to individual titles. All other proxies are title-based.

awards – are discrete, representing quality as a threshold.

379

In the following sections, we examine the relation between these proxies, assessing the correlation between them, how they are situated in a network, and their intersections.

380

381

5. Results

5.1 Correlation

Having annotated the titles in our corpus for these proxies, we looked at the correlations between them to see how and whether they interplay. As some values are discrete and others are not, the correlation matrix is often a measure of overlap: if the correlation coefficient at the intersection of *Penguin classics* and *Norton* is a high number, the two proxies have large overlaps. Computing a Spearman or Pearson correlation between two discrete lists means checking whether and to what extent the two lists include the same items. Finally, correlations between discrete and continuous values tell us whether there is a sizable change in values when switching from one category to another – for example, whether there is a sizable change in scores between books that were long-listed for a given award and books that were not.¹⁵

conference version

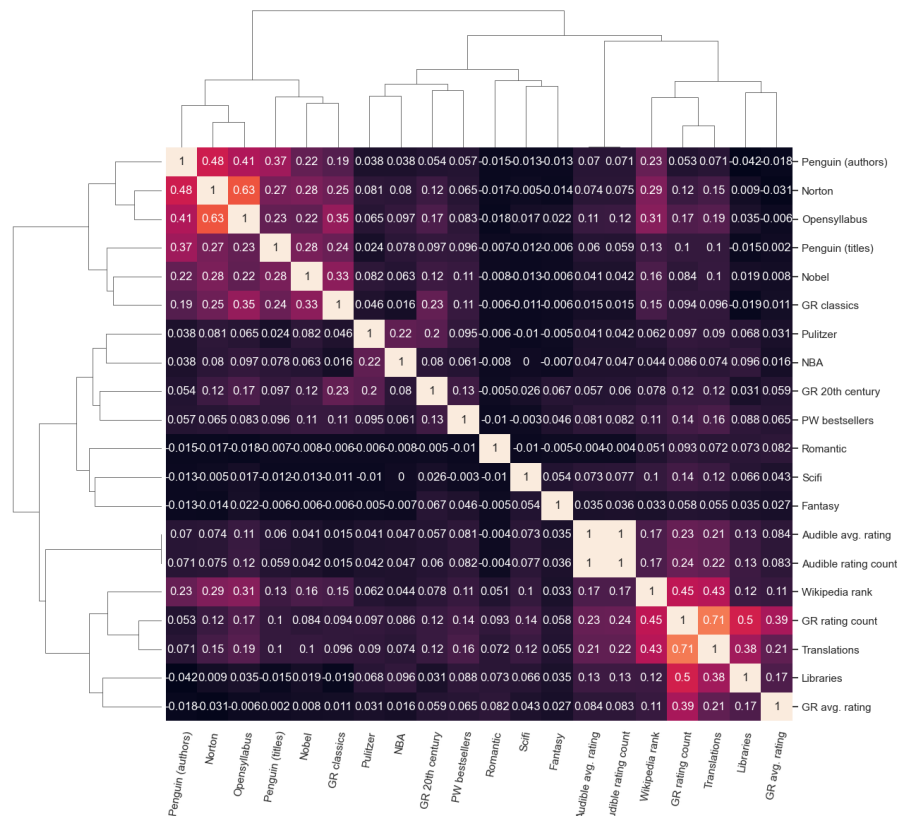


Figure 3: Correlations between discrete and continuous measures of literary quality (Spearman correlation). The matrix shows hierarchical clustering by Ward’s method.

Looking at the correlation matrix resulting from our dataset we find intriguing correlations between proxies of appreciation. Firstly, we find that there seem to be two “islands” with stronger internal correlations: one spans, roughly, GoodReads and Audible number of ratings and average ratings along with the Library holdings; the other is more or less connecting what we could call “canon lists” – GoodReads’ best books of the 20th century, GoodReads’ Classics, the Nobel, Opensyllabus, the Norton anthology, and the Penguin

15. It is crucial to remember that a correlation between a discrete and a continuous variable is not equivalent to a t-test of significance, as we will discuss later; that is, random samples from the same population could show a valid correlation, and vice versa: samples from two populations could show no correlation at all.

Classics Series, and (somewhat surprisingly) the bestsellers. Weak correlations happen out of these two areas - Wikipedia's rank correlates with Sci-fi awards, but not with the more mediatized Pulitzer prize, the award which, together with the Nobel, correlates with GoodReads' best books of the 20th century. However, these do not correlate with each other. Furthermore, the number of ratings of GoodReads and Audible shows correlations with Opensyllabus, the Norton anthology, and the Penguin Classics series.

Secondly, if we disregard the Nobel prize, which correlates with "canon" proxies such as Opensyllabus, the awards do not overlap much with one another, and do not display strong correlations with other categories. Beyond the mentioned correlations of the Pulitzer and Nobel with the GoodReads' list of best books of the 20th century, awards – and especially genre-awards – do not appear to correlate with other proxies. This lack of correlation is relevant, especially as it means that long-listed works of genre-literature appear to have no strong presence in resources like the Norton anthology or in the GoodReads' Classics list, indicating the strong presence of general fiction in these resources. However, it is still possible that the awards elicit a particular range of ratings in terms of GoodReads' ratings or libraries holdings without eliciting a detectable correlation. Also, not surprisingly, genre-fiction awards do not overlap with more literary awards (such as the Pulitzer, National Book Award, and the Nobel). At the same time, the Pulitzer and National Book Award do converge. The awards of Romantic fiction and Fantasy are the most removed, showing little convergence other proxies.

In sum, we could hypothesize that we are seeing the difference between two types of quality modeling, one that corresponds to crowd-based measures (GoodReads, Audible) and one that relates to more expert-based measures (Opensyllabus, Norton). The first category includes only measures based on counting votes - the number of people who rated a book and the average values of all users' ratings. Instead, The second category appears to be lists defined by small groups of experts that exclude or include titles, even if that group, as in the case of the GoodReads' Classics, may be lay readers.

It is notable that what we have called the "in-between" measure of library holdings correlates more strongly with crowd-based proxies (GoodReads, Audible). The correlations range from slight to robust with GoodReads' and Audible's rating count and GoodReads' average ratings. That is, books that many people rate or listen to on those platforms also tend held by many libraries. In this sense, the group consisting of "canon" lists appear like a product of the idiosyncrasies of small expert groups, to be overcome when many annotators are actually in the picture.

However, note that the second "island" of correlations does include GoodReads' classics list and, to an extent, the GoodReads' best books of the 20th century, two lists constituted through the votes of thousands or tens of thousands of individual users. Also, if the second group's selections were completely idiosyncratic and independent from each other, they would not correlate with each other, yet show evident convergence. Finally, the "expert-based" status of *Opensyllabus* can be questioned, given that it is the collection of several independent college choices, and is, in that sense, closer to the library holdings.

Thus, no clear distinction between these two clusters can be based on the method of selection (expert-based versus crowd-based), but may be based, rather, in the form of perceived canonicity or literariness that tells the second group from the first. In other

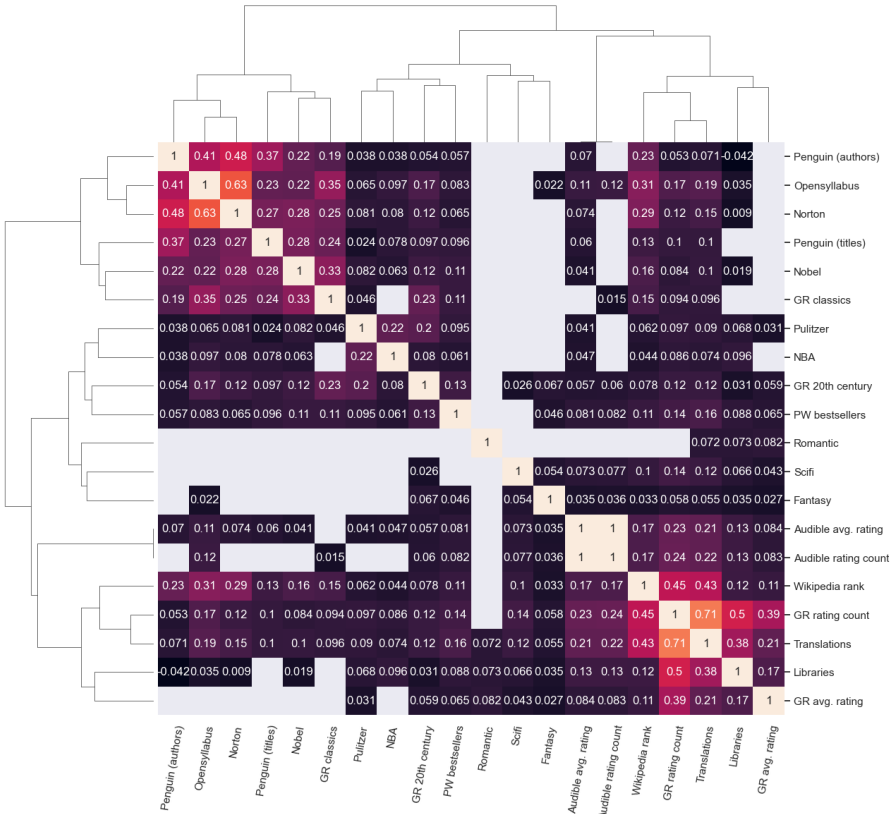


Figure 4: Again, correlations between discrete and continuous measures of literary quality (Spearman correlation), this time with non-significant correlations masked (p-value < 0.05)

words, what we are seeing might be two different “faces” of the concept of literary quality that may be perceived by the same reader. An observation supporting that there should be two main “perceptions” of quality is that the users of GoodReads seem not to give the highest ratings to the titles of the Norton anthology. Still, when GoodReads users constitute lists of “classics” and “20th century best”, they converge with the anthology on similar ground.

5.2 Network

As we have seen, continuous proxies of literary quality, such as GoodReads’ ratings and library holdings seem to correlate. However, a visualization of their convergence shows that the correlation may not be strictly linear (Fig. 5).

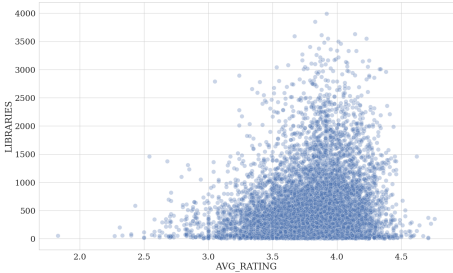


Figure 5: Scatterplot of library holdings vs. avg. rating of all titles with a threshold of 5 ratings.

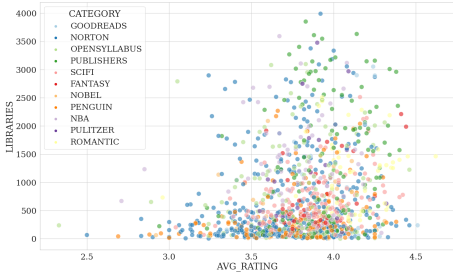


Figure 6: Scatterplot of library holdings vs. avg. rating of titles contained in one of the quality proxies.

Indeed, the interrelation between different proxies may be difficult to gauge when 454
 looking at correlation coefficients and visualizations. Proxy interrelations are better 455
 visualized in the literary quality standard landscape when visualized as a network, 456
 where each node represents one proxy and each edge the correlation (i.e., for discrete 457
 lists, the overlap) between proxies. 458

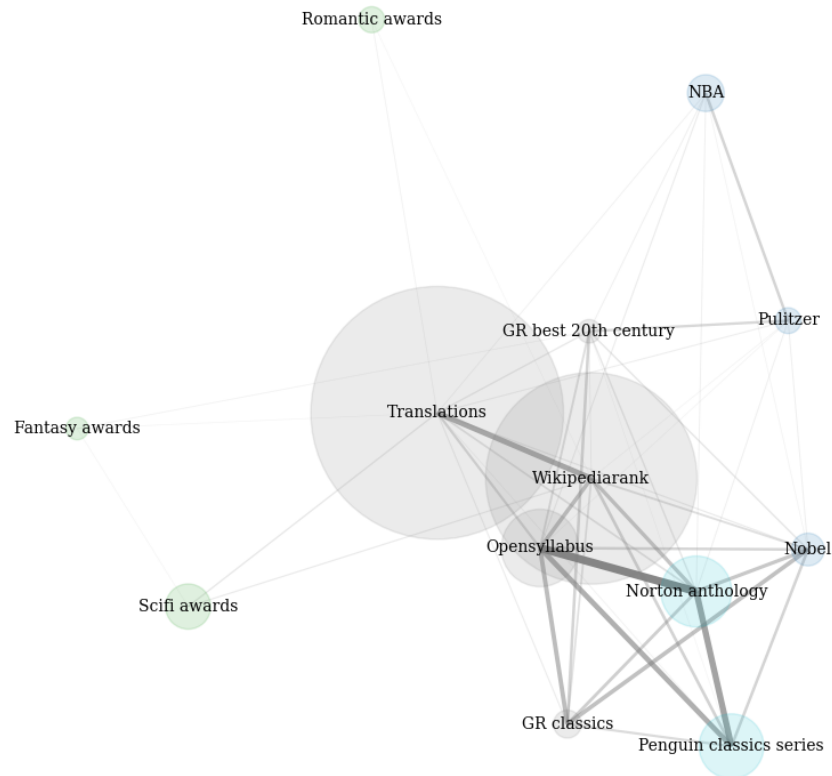


Figure 7: Network of literary quality proxies with edge-width and opacity based on the correlation coefficient between proxies (Spearman correlation), excepting the corpus-wide categories of GoodReads' ratings. We apply a coefficient threshold of 0.05 for edges being visualized. Positions are likewise determined by correlation between proxies, using the Fruchterman-Reingold force-directed algorithm for positioning. The sizes of the nodes are determined by the number of titles in each proxy. Colors are used to indicate similar types of awards: literary awards, genre-fiction awards, book-series/anthology.

As was also apparent in the correlation matrix (Fig. 3), longlists of genre-fiction awards 459
 tend to be far removed from other proxies, with a slight correlation between Fantasy 460
 and Scifi-awards, which might be explained by the thematic overlap between these 461
 genres. The disconnection between more "literary proxies" like the Penguin Classics 462
 series and the Norton Anthology may also be affected by relabelling of genre-fiction 463
 in literary markets. Genre tags may act like implicit quality judgments themselves: 464
 prestigious horror is often relabelled "gothic" or "literary fiction" and doesn't even 465
 run for genre-awards (think of, i.a., Bram Stoker and Mary Shelley). Genre-labelling 466
 is a complex issue, where various cultural factors and market forces may play in. 467
 For example, works by women authors are often labeled or re-labeled into less prestigious 468
 genres, such as 'Romantic fiction' over 'literary novel' (Groos 2000). 469

In our network, books listed in the *Index Translationum* show a strong correlation with author's in our Wikipedia-page-rank data, and also have a large actual overlap: 52.7 percent of translated books are books by authors in our Wikipedia-page-rank data, and 75.3 percent of books by authors in our Wikipedia-page-rank data are also in the *Index Translationum*-list of translated works. While literary awards, National Book Award and Pulitzer do show some overlap, the cluster of most related proxies seems to be the more expert-based expert-based type of proxy: especially Opensyllabus, Norton Anthology, and the Penguin Classics series form a distinct triangle in the network. Books that are in one of these three proxies also tend to be in the other, which is particularly interesting in this case, since the underlying selection mechanisms of these the three seem distinct, split between institutional and commercial affiliations. Nevertheless, their selection still converges on some shared perception of quality of titles. Furthermore, the divergence of awards from the remaining proxies, as well as the divergence between award-types of general (National Book Award, Pulitzer) and genre-fiction is even more apparent in the network, while the Nobel prize shows stronger convergences with the mentioned triad of more canonical, expert-based proxies, indicating its difference from the other prestigious awards.

5.3 Intersection

	GR avg. rating	GR rating count	Library holdings	Translations	WAP rank
Corpus average	3.75	14246.36	535.74	6.58	0.000058
Opensyllabus	3.78	109831.81	738.05	25.22	0.000423
Penguin classics*	3.72	57105.42	463.54	16.18	0.000334
Penguin classics (titles)	3.76	194615.08	496.74	43.14	0.000418
Norton	3.74	74424.81	687.75	22.09	0.000402
GoodReads' classics	3.82	4307090.65	501.37	57.11	0.000869
GoodReads' best books of the 20th century	4.04	992225.89	998.41	98.02	0.000439
Nobel	3.81	119078.32	811.09	32.04	0.000558
NBA	3.83	62071.08	1266.10	17.28	0.000111
Pulitzer	3.91	135290.26	1498.77	33.98	0.000176
Scifi awards	3.88	73716.60	701.81	13.81	0.000135
Fantasy awards	3.92	164753.12	804.28	18.27	0.000158
Romantic awards	4.09	31595.07	1078.24	11.69	0.000037
Bestsellers	3.94	120453.92	1290.56	43.03	0.000222

Table 2: Intersectional values: mean continuous quality-measure per discontinuous proxies. Bold font indicates the highest mean within the selection of proxies. Note that the Wikipedia rank (WAP) has been multiplied by 100, because of the generally low values.

Correlations are not the only way of checking whether two categories converge: our continuous values (library holdings, GoodReads' average ratings and rating count, translation and Wikipedia page rank) may be used to distinguish between discrete proxies. For example, Pulitzer prize winners might elicit consistently higher GoodReads' ratings than the corpus average. In this example, we would propose that GoodReads' ratings exhibit a "convergence" with the Pulitzer resource. Similarly, it may be that one type of award has systematically higher ratings and more library holdings than other books, indicating an affinity to the perception of quality affecting library holdings. In other words, there may not be a correlation between but still a convergence of two categories. Examining proxy intersections in this way, we look at the distribution of continuous proxy-values of each discrete proxy, comparing this distribution to titles in our corpus that are not contained in any of our selected quality proxies.

When visualizing the distribution of titles of different categorical proxies in terms of our the continuous proxies (rating count, translations, etc.), we see that titles included

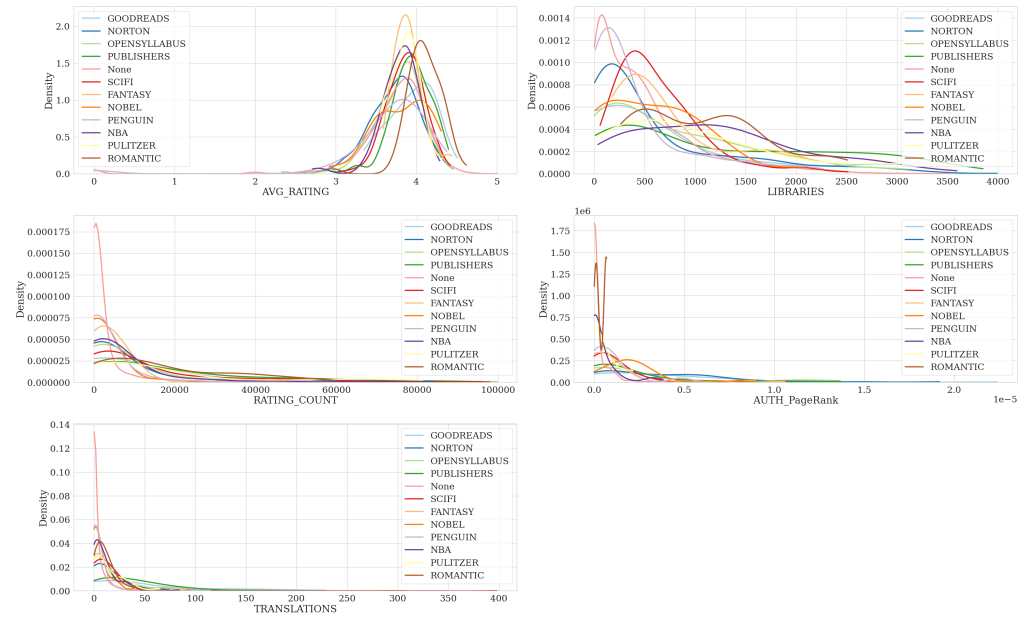


Figure 8: Kernel density estimate (KDE) plots of the distributions of measures per quality proxy. Note that rating count values above 100,000 have been filtered out for the purpose of visualization. “None” represents titles that are not in either of the proxies.

in categorical quality proxies generally have a longer tail and may have different distributions than titles not contained in any proxy of quality (“None” in Fig. 8). Looking at GoodReads’ average rating and library holdings, books included in categorical proxies seem to have smoother slopes in comparison to the rest of the corpus (“None”), whereas in terms of rating count, Wikipedia Author-page Rank and translations, we see a much higher amount of works in either proxy having very low values, with a long tail of few outliers at very high values. Measures such as rating count tend to exhibit a log-type distribution.

Moreover, different categorical proxies peak at different values within the continuous proxies. For example, the distribution of books that have won a Romantic literary award seem to peak at a higher value of GoodReads’ average rating, having also the highest mean average rating of any proxy (Tab. 2).¹⁶ Titles in GoodReads’ Classics, Nobel prize, Opensyllabus and Norton Anthology are represented more evenly across values of Wikipedia Author-page Rank, which may be expected as we also saw that these proxies seem to be closely related in our network (Fig. 7). It indicates that these base their selection on some shared perception of quality, which may also prompt their authors to have more prominent Wikipedia pages. Interestingly, the plot showing distributions over library holdings shows a somewhat opposite tendency: here, genre-fiction tends to place at higher values, so that Sci-fi, Fantasy and Romantic fiction, for example, peak at higher values, and have high mean library holdings numbers (Tab. 2). In general, the two “islands” of quality observed in our correlation matrix (Fig. 3) can be observed in the colors that peak in the different quadrants, genre fiction in some, what we could call more “higher brow” or canonical literature in others.

16. Note that the odd distribution of Romantic titles in the plots with library holdings and Wikipedia Author-page Rank rank may be an effect of the small number of titles. It may be that one author who has higher canonicity is responsible for the peak at the higher end in both plots.

Visualizing the mean values of each discrete proxy in terms of continuous proxies further aids in gauging the differences between these quality perspectives (Fig. 9-13).

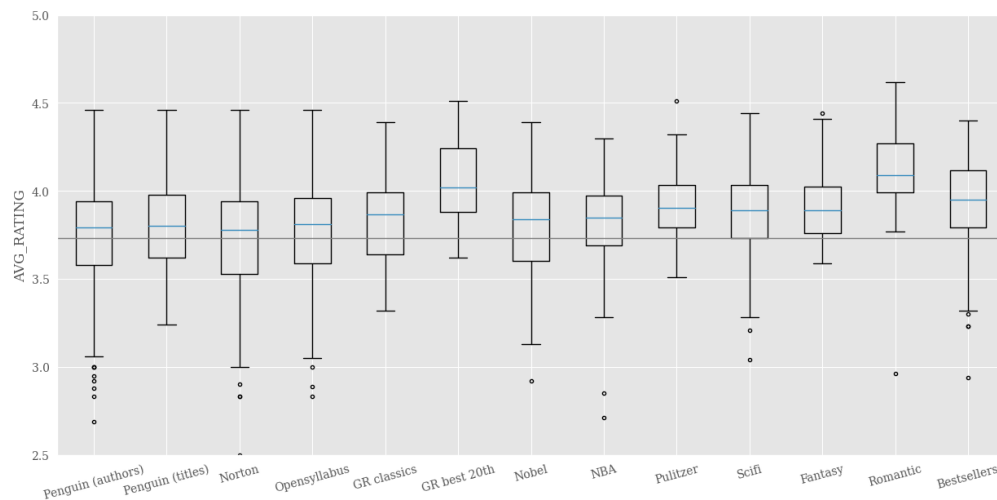


Figure 9: Boxplot of average GoodReads rating for discrete categories. The grey line indicates the corpus average rating.

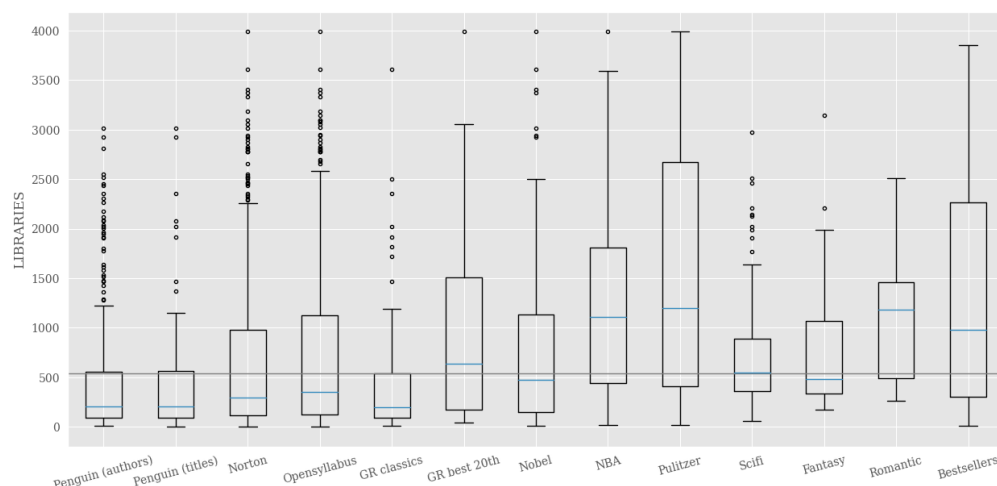


Figure 10: Boxplot of average number of library holdings for discrete categories. The grey line indicates the corpus average holdings.

GoodReads' best books of the 20th century appear to have the highest average GoodReads' ratings, closely followed by Hugo and Pulitzer titles, while the Norton and Opensyllabus titles record the lowest average ratings (Tab. 9). Overall, Opensyllabus' and Norton Anthology titles score consistently lower with respect to any other category in terms of their GoodReads' average ratings as well as their number of libraries holdings (10).

GoodReads' best books of the 20th century is the only proxy that stands out in terms of GoodReads' rating *count* (Fig. 11). Note that rating count is a problematic proxy because of its non-normal distribution, with very few titles at very high values, which is why we see a low corpus mean with many outliers for each proxy as well as long whiskers for the GoodReads' best books of the 20th century category.

Translation numbers and Wikipedia Author-page Rank are the two continuous measures that appear similar in the sense that titles longlisted for awards tend to score low in

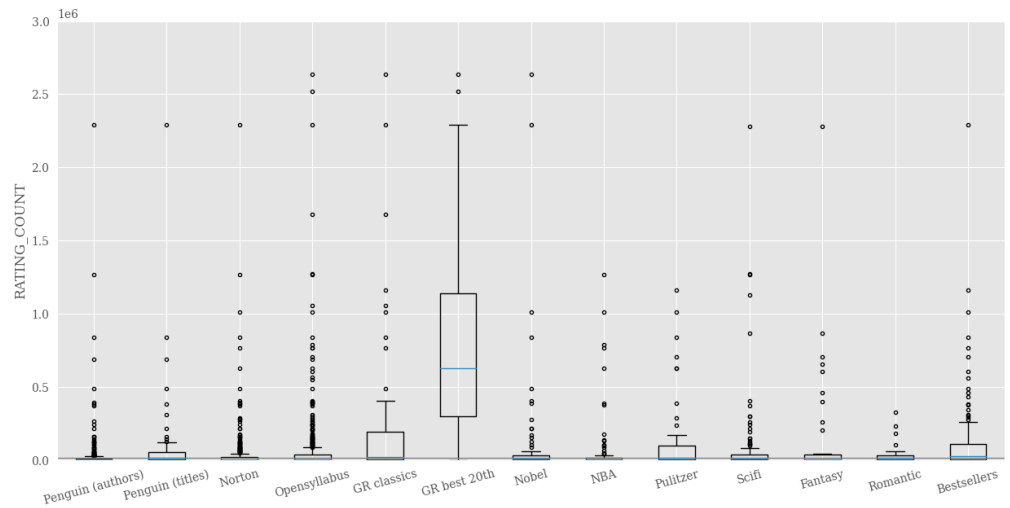


Figure 11: Boxplot of rating count of discrete categories. The grey line indicates the corpus average rank.

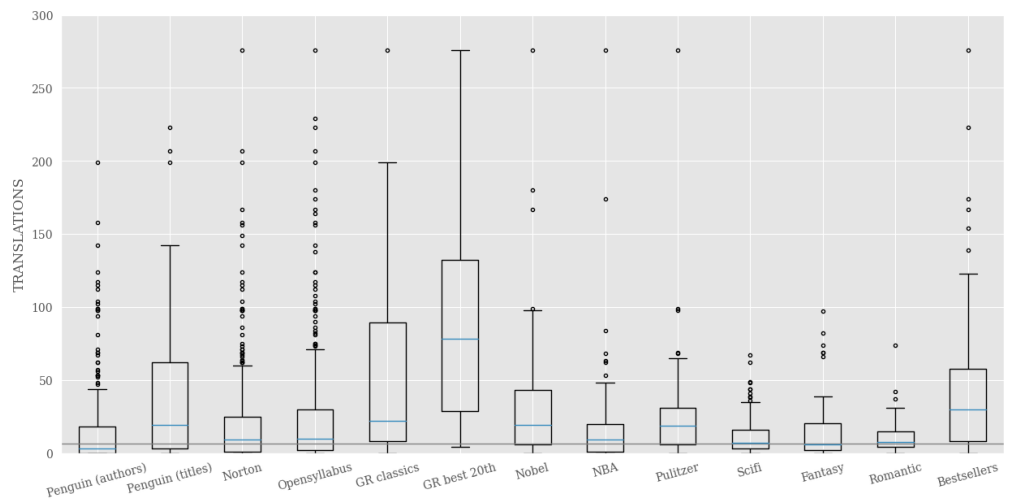


Figure 12: Boxplot of average translation numbers for discrete categories. The grey line indicates the corpus average number.

comparison to, for example, GoodReads' Classics titles. Again, there is a difference between general fiction awards (National Book Award, Pulitzer) and genre-fiction awards, where titles longlisted for genre-fiction awards tend to place lower. It is interesting that for these two plots (Fig. 12, 13), the user-generated lists GoodReads' Classics and best books of the 20th century score high, with a subtle difference between the two plots. When looking at translation numbers, we see that GoodReads' best 20th century books score higher than GoodReads' Classics, and that bestsellers are also one of the proxies with higher mean translation numbers. Conversely, when looking at the Wikipedia Author-page Rank, we see that GoodReads' Classics have a higher mean than the best 20th century books, and that the Nobel titles, as well as the more expert-based measures that showed the strongest affinities in our network (7) also have a higher mean in comparison to when looking at translation numbers. Considering each of these boxplots together, overall, we observe the following patterns:

1. Titles longlisted for awards, both general fiction and genre-awards, tend to have

higher average GoodReads' rating and library holdings.

553

2. The proxies we found to be strongly correlated in the "island" of our correlation matrix representing more "canonical" fiction (Fig. 3), Opensyllabus, Norton, and GoodReads' Classics, tend to have lower average GoodReads' ratings and library holdings.
3. There is a partial convergence between vote-based continuous scores and discrete categories. While translation numbers and Wikipedia Author-page Rank seem to ascribe higher values to more "canonical" fiction, GoodReads' users and library holdings they seem to have a higher appreciation for awards and genre-fiction, and a lower appreciation for the canon.

554

555

556

557

558

559

560

561

562

We clearly note a distinct variation among quality proxies, with an inclination of proxies of similar affiliation type – i.e., institutional, intellectual, commercial – to exhibit analogous behavior. Especially awards appear less aligned to other proxies of literary quality in terms of correlation (Fig. 3, 7). Nevertheless, titles longlisted for awards in our corpus enjoy a higher appreciation among users of GoodReads and a higher circulation in libraries. This agrees with the approach of Manshel et al. (2019), who consider awards an distinct form of quality proxy Manshel et al. 2019.

563

564

565

566

567

568

569

Looking at the different types of awards, we seem to confirm Bourdieu's intuition that the literary field is polarized: our genre-award proxies appear far removed from other proxies (including more general literary awards, see Fig.7). Yet they have higher average GoodReads' ratings and library holdings than, for example, the more institutionally oriented Norton Anthology. These characteristics would situate titles of genre-awards roughly at the place of the "popular novel" in Bourdieu's map of the literary field, which also aligns with the study of the prestige versus popularity of genre fiction by Porter 2018. In contrast, a proxy like the Norton Anthology, may be situated more toward the "intellectual" and "bourgeois" poles of Bourdieu's map, considering it is part of the inter-linked triangle of proxies observed in our network (Fig. 7), of which Opensyllabus has an institutional status. The clear divergence between proxies like the Norton Anthology

570

571

572

573

574

575

576

577

578

579

580

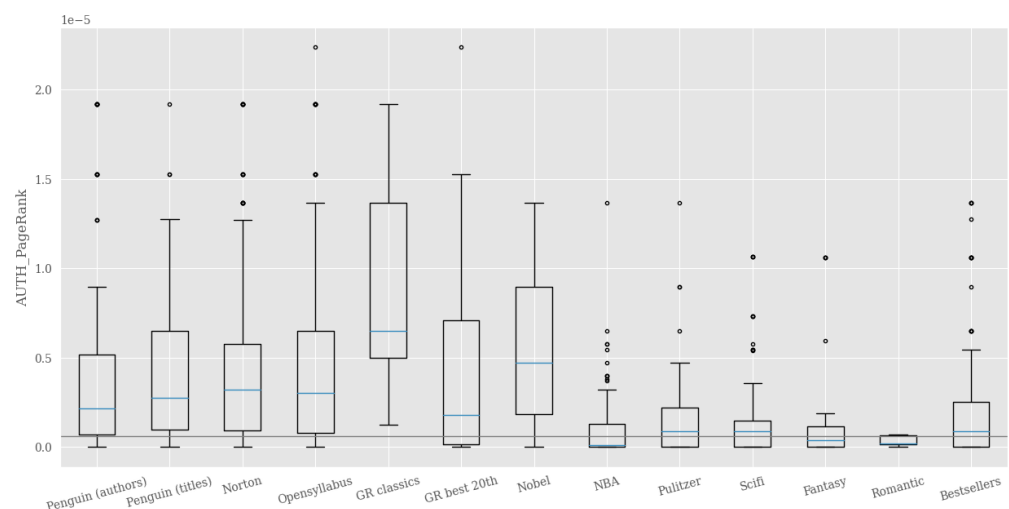


Figure 13: Boxplot of average Wikipedia AP rank for discrete categories. The grey line indicates the corpus average rank.

and genre-fiction awards may be explained by differences in style and topic of books, but studies have also suggested that different types of audiences appreciate books at different levels of readability (Bizzoni et al. 2023). Thus, the divergence may also have to do with socio-cultural factors like population literacy, where more “readable” works are preferred at the level of larger audiences, and more institutionally acclaimed works, such as those included in the Norton Anthology less so, partly because of difficulty at the sentence level.

Following Bourdieu, we might contrast actors behind the general fiction award proxies as “intellectual audiences” against those behind genre-fiction awards as a “mass audience” (Fig. 1). However, it is important to note we do not find audiences to be as polarized or distinct as Bourdieu suggested. Rather, proxies seem to transverse their actor-type affiliations. For instance, while bestsellers and Opensyllabus have dissimilar actors underlying them – institutional versus market-oriente – bestsellers had the strongest correlation with Opensyllabus as seen in Fig. 3. These findings imply the potential existence of two overarching types of “quality perception,” which overlay and interlink proxies underpinned by divergent actors or audiences. This insight emerges from the observation of two “islands” when looking at correlations (3), but also from looking into the differential favoring of each of the continuous measures contained in the first “island”. When exploring discrete proxies in terms of the continuous ones, we saw that GoodReads’ ratings and library holdings on one side, and translation numbers and Wikipedia page-rank on the other were more similar in the way they value, for example, longlisted titles for genre-awards. This suggests that actor or audience-based distinctions might not fully capture the intricate dynamics of appreciation judgments in the literary field.

When looking at proxies in terms of the distinction between expert-based or crowd-based, we do see vote-based or what we could characterize as “crowd-based” proxies cluster in terms of correlation: Audible average ratings with GoodReads’ average ratings, as wells as libraries, translation numbers and Wikipedia Author-page Rank, of which the latter may, in part, represent tastes of lay-readers (see section 3.3.3). However, continuous crowd-based proxies also differ: GoodReads’ ratings and library holdings numbers assign higher values to some proxies, like awards, which proxies like Wikipedia Author-page Rank does not. Wikipedia Author-page Rank is also the proxy which mostly strongly bridges the two “islands” in our correlation matrix, exhibiting correlations with both “islands” (Fig. 3), which may explain its different behaviour and which may more properly situate it between expert-based and crowd-based type of proxies. As such, we may use the distinction between expert-based and crowd-based proxies heuristically, though it seems that more complex judgements based on different quality “perceptions” contribute to the clusters we have observed.

6. Conclusion and Future Works

Generally, we seem to observe two types of “quality perception”, or two faces of the concept of quality, emerge through the differences and surprising convergences of the host of proxies considered in the present study.

There appears to be a perception of titles’ canonicity in expert-based proxies like Open-

syllabus that does not converge much with the popularity of a title on crowd-based resources like GoodReads. In this sense, we validated and expanded Walsh and Antoniak 2021's study, as we too observed the convergence of different canonicity proxies, including those compiled on GoodReads by large numbers of unqualified readers. This suggests the presence of two distinct modes of evaluating quality, which can mirror two macro-classes of reader types (Riddell and Dalen-Oskam 2018) or can be even accessible to individual readers as they navigate different dimensions of assessment.

This duality is reminiscent of several similar dichotomies theorized in previous works: C. Koolen et al. 2020's distinction of literariness and enjoyability, Porter 2018 and Manshel et al. 2019's distinction between prestige and popularity, and naturally of Bourdieu 1993's two axes of institutionalized vs popular art. Yet, the duality that emerges from our data is nuanced and does not represent a polar opposition, but rather fuzzy islands between different proxies. Bestseller lists agree with canonical groups and with GoodReads' metrics, and the distinctness of titles included in longlists for genre awards might even indicate a possible third – or many – different perceptions of quality, which may be connected to various extra- and intra-textual features.

This is not surprising: indeed, as we mentioned in the beginning, every literary judgment is unique insofar as it is based on idiosyncratic or internalized interpretations of the text, various expectations suggested by the genre of a title, its publication date, textual features, the cover, etc. For example, one type of book may be more demanding to read and likely set the expectation bar of readers higher, genre-codes influence readers quality judgements or attract types of readers, and so on. The consensus among readers found in recent computational studies, which suggest that textual features inform quality judgements (i.e., Bizzoni et al. 2021; Dalen-Oskam 2023; Maharjan et al. 2017; Wang et al. 2019) should therefore be interpreted with an eye to the type of proxy used in the particular study.

More complicated is the possible influence of social structures and power dynamics (Bennett 1990; Casanova 2007; Guillory 1995; Moretti 2007) on quality judgments: it is possible that we see the effect of crowd-based types of proxies being more diverse in terms of gender, reviewer background, etc. so that they appear to form a different "perception" of quality. This would not explain, however, why what we would understand as a crowd-based type of proxy, the bestseller list, seems to correlate with expert-based proxies. Examining the characteristics of titles at the textual level in conjunction with considerations of various quality proxies – but also considering likely biases influencing literary judgements – would help shed further light on the complex issue of measuring literary qualities. Nevertheless, what we have called two main "perceptions of quality" in this study cannot be completely idiosyncratic since two main groups of proxies do correlate and seem to converge on similar grounds, despite differences in their nature.

Various limitations inhere to the selection of quality proxies and to the quality proxies themselves, and it should be noted that various other proxies could be collected, among others, sales figures. Moreover, different literary cultures may vary in their ways of assessing quality, while this study is clearly situated in an Anglophone and American context. In terms of challenges in assessing the quality proxies themselves, for example, it is possible that GoodReads represents a contemporary audience so that canonical literature, assessed over decades or centuries, does not precisely align with their tastes.

In future studies, we suggest a closer inspection of possible biases, such as the publication 669
 dates of titles, as well as gender or race biases influencing literary judgements. We also 670
 suggest a stronger focus on the interplay between textual features and different types of 671
 quality proxies. For example, assessing the importance of readability for different types 672
 of proxies, which is an often underrated metric that may, among other things, likely 673
 account for the demise of certain avant-garde works over time, as well as the difference 674
 in preference between types of audiences. 675

7. Data Availability 676

Data can be found here: <https://github.com/centre-for-humanities-computing/c> 677
[hicago_corpus](https://github.com/centre-for-humanities-computing/c). 678

8. Author Contributions 679

Pascale Feldkamp: Analysis, writing, review & editing 680

Yuri Bizzoni: Analysis, writing, review & editing 681

Ida Marie S. Lassen: Methodology, project administration 682

Mads Rosendahl Thomsen: Methodology, review & editing, project administration 683

Kristoffer L. Nielbo: Methodology, project administration 684

References 685

- Algee-Hewitt, Mark, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, 686
 and Hannah Walser (2018). *Canon/archive : large-scale dynamics in the literary field*. 687
 Vol. Stanford Literary Lab. Pamphlets of the Stanford Literary Lab 11, 14. <https://l> 688
[itlab.stanford.edu/LiteraryLabPamphlet11.pdf](https://l). 689
- Alter, Alexandra, Elizabeth A. Harris, and David McCabe (July 2022). "Will the Biggest 690
 Publisher in the United States Get Even Bigger?" In: *The New York Times*. <https://ww> 691
[w.nytimes.com/2022/07/31/books/penguin-random-house-simon-schuster-anti](https://ww) 692
[trust-trial.html](https://ww). 693
- Archer, Jodie and Matthew L Jockers (2017). *The bestseller code*. London: Penguin books. 694
- Bennett, Tony (1990). *Popular Fiction: Technology, Ideology, Production, Reading*. Routledge. 695
- Bizzoni, Yuri, Pascale Moreira, Nicole Dwenger, Ida Lassen, Mads Thomsen, and Kristof- 696
 fer Nielbo (2023). "Good Reads and Easy Novels: Readability and Literary Quality 697
 in a Corpus of US-published Fiction". In: *Proceedings of the 24th Nordic Conference on* 698
Computational Linguistics (NoDaLiDa). Tórshavn, Faroe Islands: University of Tartu 699
 Library, 42–51. <https://aclanthology.org/2023.nodalida-1.5>. 700
- Bizzoni, Yuri, Telma Peura, Kristoffer Nielbo, and Mads Thomsen (2022a). "Fractal Sen- 701
 timents and Fairy Tales-Fractal scaling of narrative arcs as predictor of the perceived 702
 quality of Andersen's fairy tales". In: *Journal of Data Mining & Digital Humanities*. 703
[10.46298/jdmdh.9154](https://doi.org/10.46298/jdmdh.9154). 704






- Bizzoni, Yuri, Telma Peura, Kristoffer Nielbo, and Mads Thomsen (2022b). “Fractality of sentiment arcs for literary quality assessment: The case of Nobel laureates”. In: *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*. Taipei, Taiwan: Association for Computational Linguistics, 31–41. <https://aclanthology.org/2022.nlp4dh-1.5>.
- Bizzoni, Yuri, Telma Peura, Mads Rosendahl Thomsen, and Kristoffer Nielbo (2021). “Sentiment Dynamics of Success: Fractal Scaling of Story Arcs Predicts Reader Preferences”. In: *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*. Silchar, India: NLP Association of India (NLP AI), 1–6. <https://aclanthology.org/2021.nlp4dh-1.1>.
- Bjerck Hagen, Eric, Christine Hamm, Frode Helmich Pedersen, Jørgen Magnus Sejersted, and Eirik Vassenden (2018). “Literary Quality: Historical Perspectives”. In: *Contested Qualities*. Ed. by Knut Ove Eliassen, Jan Hovden, and Øyvind Prytz. Fagbokforlaget, 47–74.
- Bloom, Harold (1995). *The Western Canon: The Books and School of the Ages*. First Riverhead Edition. New York, NY: Riverhead Books.
- Bourdieu, Pierre (1993). *The field of cultural production: essays on art and literature*. Ed. by Randal Johnson. Columbia University Press.
- Casanova, Pascale (2007). *The World Republic of Letters*. Harvard University Press.
- Cranenburgh, Andreas van and Rens Bod (Apr. 2017). “A Data-Oriented Model of Literary Language”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, 1228–1238. <https://aclanthology.org/E17-1115>.
- Dalen-Oskam, Karina van (June 2023). *The Riddle of Literary Quality*. ISBN: 978-90-485-5814-8. <https://www.aup.nl/en/book/9789048558148/the-riddle-of-literary-quality> (visited on 04/12/2024).
- Febres, Gerardo and Klaus Jaffe (2017). “Quantifying literature quality using complexity criteria”. In: *Journal of Quantitative Linguistics* 24.1, 16–53. [10.1080/09296174.2016.1169847](https://doi.org/10.1080/09296174.2016.1169847).
- Groos, Marije (2000). “Wie schrijft die blijft? Schrijfsters in de literaire kritiek van nu”. In: *Tijdschrift voor Genderstudies* 3.3.
- Guillory, John (1995). *Cultural Capital: The Problem of Literary Canon Formation*. University of Chicago Press.
- Harrison, Chloe and Louise Nuttall (2018). “Re-reading in stylistics”. In: *Language and Literature* 27.3. SAGE Publications Ltd, 176–195. [10.1177/0963947018792719](https://doi.org/10.1177/0963947018792719).
- Hube, Christoph, Frank Fischer, Robert Jäschke, Gerhard Lauer, and Mads Rosendahl Thomsen (2017). *World Literature According to Wikipedia: Introduction to a DBpedia-Based Framework*. <http://arxiv.org/abs/1701.00991>.
- Jannatus Saba, Syeda, Biddut Sarker Bijoy, Henry Gorelick, Sabir Ismail, Md Saiful Islam, and Mohammad Ruhul Amin (2021). “A Study on Using Semantic Word Associations to Predict the Success of a Novel”. In: *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, 38–51. <https://aclanthology.org/2021.starsem-1.4>.
- Jockers, Matthew L (2015). *Syuzhet: Extract sentiment and plot arcs from text*.
- Karlynn, Danny and Tom Keymer (n.d.). *Chadwyck-Healey Literature Collection*. http://collections.chadwyck.com/marketing/products/about_ilc.jsp?collection=ncl.


- Koolen, Corina, Karina van Dalen-Oskam, Andreas van Cranenburgh, and Erica Nagel-
hout (2020). "Literary Quality in the Eye of the Dutch Reader: The National Reader
Survey". In: *Poetics* 79, 1–13. [10.1016/j.poetic.2020.101439](https://doi.org/10.1016/j.poetic.2020.101439).
- Koolen, Cornelia Wilhelmina (2018). *Reading beyond the female: the relationship between
perception of author gender and literary quality*. ILLC dissertation series DS-2018-03.
Amsterdam: Institute for Logic, Language and Computation, Universiteit van Ams-
terdam. ISBN: 978-94-028-0951-0.
- Kousha, Kayvan, Mike Thelwall, and Mahshid Abdoli (2017). "GoodReads reviews to
assess the wider impacts of books". In: *Journal of the Association for Information Science
and Technology* 68.8, 2004–2016. ISSN: 2330-1643. [10.1002/asi.23805](https://doi.org/10.1002/asi.23805).
- Kovács, Balázs and Amanda J Sharkey (2014). "The Paradox of Publicity". In: *Adminis-
trative Science Quarterly* 1, 1–33. [10.1177/0001839214523602](https://doi.org/10.1177/0001839214523602).
- Kuijpers, Moniek M. and Frank Hakemulder (2018). "Understanding and Appreciating
Literary Texts Through Rereading". In: *Discourse Processes* 55.7, 619–641. [10.1080/0163853X.2017.1390352](https://doi.org/10.1080/0163853X.2017.1390352).
- Lassen, Ida Marie Schytt, Yuri Bizzoni, Telma Peura, Mads Rosendahl Thomsen, and
Kristoffer Laigaard Nielbo (2022). "Reviewer Preferences and Gender Disparities in
Aesthetic Judgments". In: *CEUR Workshop Proceedings*. Antwerp, Belgium, 280–290.
https://ceur-ws.org/Vol-3290/short_paper1885.pdf.
- Maharjan, Suraj, John Arevalo, Manuel Montes, Fabio A. González, and Tamar Solorio
(2017). "A Multi-task Approach to Predict Likability of Books". In: *Proceedings of
the 15th Conference of the European Chapter of the Association for Computational Linguis-
tics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguis-
tics, 1217–1227. <https://aclanthology.org/E17-1114>.
- Maharjan, Suraj, Sudipta Kar, Manuel Montes, Fabio A. González, and Tamar Solorio
(2018). "Letting Emotions Flow: Success Prediction by Modeling the Flow of Emo-
tions in Books". In: *Proceedings of the 2018 Conference of the North American Chapter of the
Association for Computational Linguistics: Human Language Technologies: Volume 2, Short
Papers*. New Orleans, Louisiana: Association for Computational Linguistics, 259–265.
<https://aclanthology.org/N18-2042>.
- Manshel, Alexander, Laura B McGrath, and J.D. Porter (2019). *Who Cares about Literary
Prizes?* <https://www.publicbooks.org/who-cares-about-literary-prizes/>.
- Mohseni, Mahdi, Christoph Redies, and Volker Gast (2022). "Approximate Entropy in
Canonical and Non-Canonical Fiction". In: *Entropy* 24.2, 278. [10.3390/e24020278](https://doi.org/10.3390/e24020278).
- Moretti, Franco (2007). *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso.
- Nakamura, Lisa (2013). "'Words with Friends': Socially Networked Reading on GoodReads".
In: *PMLA* 128.1, 238–243. [10.1632/pmla.2013.128.1.238](https://doi.org/10.1632/pmla.2013.128.1.238).
- Peer, Willie van (2008). "Ideology or aesthetic quality?" In: *The quality of literature: linguis-
tic studies in literary evaluation*. Ed. by Willie van Peer. John Benjamins Publishing, 17–
29.
- Pope, Colin (2019). *We Need to Talk bout the Canon: Demographics in 'The Norton Anthology'*.
[https://themillions.com/2019/04/we-need-to-talk-about-canons-picturing-
writerly-demographics-in-the-norton-anthology-of-american-literature.ht
ml](https://themillions.com/2019/04/we-need-to-talk-about-canons-picturing-writerly-demographics-in-the-norton-anthology-of-american-literature.html).
- Porter, J.D. (2018). *Popularity/Prestige: A New Canon*. Vol. 17. Pamphlets of the Stanford
Literary Lab. Stanford Literary Lab. [https://litlab.stanford.edu/LiteraryLab
Pamphlet11.pdf](https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf).

- Ragen, Brian Abel (1992). "An Uncanonical Classic: The Politics of the "Norton Anthology""". In: *Christianity and Literature* 41.4, 471–479. <https://www.jstor.org/stable/44312103>. 799 800 801
- Reagan, Andrew J, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds (2016). "The emotional arcs of stories are dominated by six basic shapes". In: *EPJ Data Science* 5.1, 1–12. 802 803 804
- Riddell, Allen and Karina van Dalen-Oskam (2018). "Readers and their roles: Evidence from readers of contemporary fiction in the Netherlands". In: *PLOS ONE* 13.7. Ed. by K. Brad Wray, e0201157. [10.1371/journal.pone.0201157](https://doi.org/10.1371/journal.pone.0201157). 805 806 807
- Shesgreen, Sean (2009). "Canonizing the Canonizer: A Short History of The Norton Anthology of English Literature". In: *Critical Inquiry* 35.2, 293–318. <https://www.jstor.org/stable/10.1086/596644>. 808 809 810
- Trotsky, Leon (1974). *Class and Art: Problems of Culture Under the Dictatorship of the Proletariat*. New Park. 811 812
- Veleski, Stefan (2020). "Weak Negative Correlation between the Present Day Popularity and the Mean Emotional Valence of Late Victorian Novels". In: *Proceedings of the Workshop on Computational Humanities Research*. Amsterdam, the Netherlands: CEUR Workshop Proceedings, 32–43. <http://ceur-ws.org/Vol-2723/long44.pdf>. 813 814 815 816
- Vulture, editors (2018). *A Premature Attempt at the 21st Century Literary Canon*. <https://www.vulture.com/article/best-books-21st-century-so-far.html>. 817 818
- Walsh, Melanie and Maria Antoniak (2021). "The Goodreads "Classics": A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism". In: *Post45: Peer Reviewed*. <https://post45.org/2021/04/the-goodreads-classics-a-computational-study-of-readers-amazon-and-crowdsourced-amateur-criticism/>. 819 820 821 822
- Wang, Xindi, Burcu Yucesoy, Onur Varol, Tina Eliassi-Rad, and Albert-László Barabási (2019). "Success in Books: Predicting Book Sales Before Publication". In: *EPJ Data Science* 8.1, 31. [10.1140/epjds/s13688-019-0208-6](https://doi.org/10.1140/epjds/s13688-019-0208-6). 823 824 825
- Wellek, René (1972). "The Attack on Literature". In: *The American Scholar* 42.1. Publisher: The Phi Beta Kappa Society, 27–42. <https://www.jstor.org/stable/41207073>. 826 827

From Review to Genre to Novel and Back

An Attempt To Relate Reader Impact to Phenomena of Novel Text

Marijn Koolen¹ 
 Joris J. Van Zundert² 
 Eva Viviani³ 
 Carsten Schnober³ 
 Willem Van Hage³ 
 Katja Tereshko²

1. DHLab, Humanities Cluster, Amsterdam, The Netherlands.
2. Computational Literary Research, Huygens Institute , Amsterdam, The Netherlands.
3. eScience Center , Amsterdam, The Netherlands.

Citation

Marijn Koolen, Joris van Zundert, Eva Viviani, Carsten Schnober, Willem van Hage, and Katja Tereshko (2024). "From Review to Genre to Novel and Back. An Attempt To Relate Reader Impact to Phenomena of Novel Text". In: *CCLS2024 Conference Preprints* 3 (1). [10.26083/tuprints-00027398](https://doi.org/10.26083/tuprints-00027398)

Date published 2024-05-28

Date accepted 2024-04-04

Date received 2024-01-25

Keywords

reader impact, literary novels, genre, topic modeling

License

CC BY 4.0 

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 3rd Annual Conference of Computational Literary Studies at Vienna, Austria, in June 2024.

Abstract. We are interested in the textual features that correlate with reported impact by readers of novels. We operationalize impact measurement through a rule-based reading impact model and apply it to 634,614 reader reviews mined from seven review platforms. We compute co-occurrence of impact-related terms and their keyness for genres represented in the corpus. The corpus consists of the full text of 18,885 books from which we derived topic models. The topics we find correlate strongly with genre, and we get strong indicators for what key impact terms are connected to which genre. These key impact terms gives us a first evidence-based insight into genre-related readers' motivations.

1. Introduction

Already Aristotle noted the reciprocal relations between an author, the text the author creates, and the response from an audience to the text. This fundamental model of rhetorical poetics has remained relevant throughout the ages (cf. e.g. Abrams 1971; Warnock 1978). The dynamics of the relations between author, text, and reader have been heavily theorized and fiercely debated (cf. e.g. Hickman 2012; Wimsatt 1954). But if there is no lack of theory, it appears to be much harder to gain empirical insights into these relations, though not for lack of trying by practitioners in such fields as empirical and computational literary studies (e.g. Fialho 2019; Loi et al. 2023; Miall and Kuiken 1994). One effect of the immense success of the World Wide Web and softwarization and digitization of societies and their cultures (Berry 2014; Manovich 2013) is the availability of large collections of online book reviews and digital full texts from novels published as ePubs. This allows us to apply NLP techniques and corpus statistics to get empirical data on the relations between text and reader that until now could only be theorized or anecdotally evidenced. At the same time, we should acknowledge that it is no panacea for the problem of empirical observations in literary studies. Not just because of the inherent biases (Gitelman 2013; Prescott 2023; Rawson and Muñoz 2016), or the almost complete lack of demographic and social signals in the data, but also because of the difficulties still involved in establishing which concrete signal in novels relates to what type of reaction for which type of reader. This is where we focus our research: we attempt to establish which concrete features of online reviews correlate to which concrete signals in the text

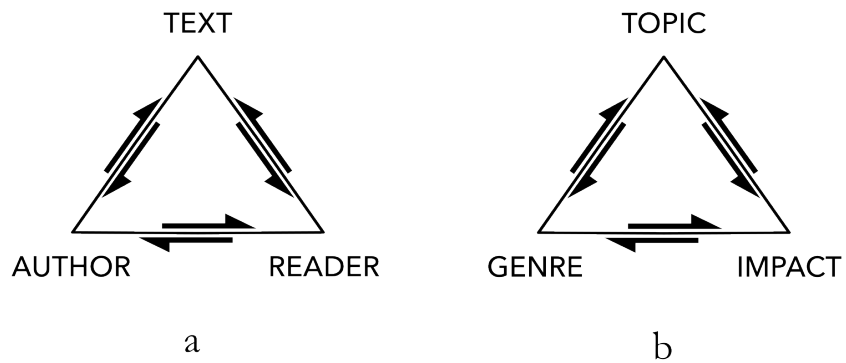


Figure 1: Classic rhetorical model (a) and our operationalization of the text-reader relation (b).

of fiction novels.

In a theoretical sense we are concentrating on the right hand side of the classical rhetorical triangle (cf. Figure 1a) and operationalize the dynamic between text and reader as another triangular relationship between *impact*, *topic*, and *genre*. With “impact” (and the commensurate “reading impact”) we designate expressions of reader experiences identified by some evidence based method (e.g. as reader impact constituents researched by Koolen et al. (2023)). We apply the reader impact model to assign concrete terms to types of reading impact. The concrete text signal that we correlate this impact with are topics mined from a corpus of novels. (As an aside we note that these topics are not to be confused with themes, motives, or aboutness in a literary studies sense, as we will explain later.) A meta-textual property, genre, forms the third measurable aspect of the triangular relationship (see Figure 1b).

Concretely, we link topic models of 18,885 novels in Dutch (original Dutch and translated to Dutch) with the reading impact expressed in 130,751 Dutch online book reviews. We want to know if there is a relationship between aspects of topic in novels, their genre, and the type of impact expressed by readers in their reviews. We extracted expressions for three types of reading impact from the reviews using the previously developed Reading Impact Model for Dutch (Boot and Koolen 2020). The three types of reading impact that we discern are: “general affective impact” which expresses the overall evaluation and sentiment regarding a novel; “narrative impact”, which relates to aspects of story, plot, and characters; and finally “stylistic impact” related to writing style and aesthetics.

We expect that topics in fiction are related to genre. As there is no authoritative source for genre of a novel, nor some general academic consensus about what constitutes genre, we make use of the broad genre labels that publishers have assigned to each published book. Analogous to Sobchuk and Šeĭa 2023, p.2, who define genre as “a population of texts united by broad thematic similarities”, we clustered these genre labels into a set of nine genres. These thematic similarities might be revealed in a topical analysis, e.g. crime novels containing more crime-related topics and romance novels containing more topics related to romance and sex. However, for some genres it might be less obvious whether they are related to topic. For instance, what are the topics one would expect in literary fiction?

It is important to note that, although the name *topic modelling* suggests that what is modelled is *topic*, most topic modelling approaches discern clusters of frequently co-occurring words, regardless

of whether they have a topical connection or not (in the classical sense of “aboutness” in library science). Clusters of words may also reveal a different type of connection, e.g. words from a particular stylistic register. In that sense, genres with less clear thematic similarities may be associated with certain stylistic registers, or any other clustering of vocabulary. Different genres may also attract different types of readers and therefore different types of reviewers, who use different terminology and pay attention to different aspects of novels. It is also plausible that the language and topic of a novel influences how readers write about them in reviews. A novel written in a particularly striking poetic style may consciously or subconsciously lead readers to adopt some of its poetic aspects and register in how they write about their reading experiences. Similarly, topics in novels may be associated with what reviewers choose to mention, again, consciously or subconsciously. A novel on the atrocities of war or on the pain of losing a loved one may lead a reviewer to mention feeling sympathy or sadness during reading, while a story about friendship and betrayal might prompt reviewers to describe their anger at the actions of one of the characters.

Thus, it is clear that the relationship between the three elements – topic, genre and impact – is complex and reciprocal, as expressed in Figure 1b. Our challenge is, of course, to computationally investigate and understand this relationship utilizing the large numbers of full-text novels from different genres and corpora of hundreds of thousands of reviews. We subdivide this overarching aim into several more concrete research questions, namely:

- How are topic and impact related to each other? Do books with certain topics lead to more impact expressed in book reviews? Do different topics lead to different types of impact?
- How are genre and impact related to each other? Do books of different genres lead to different types of impact? Do reviews of different genres use different vocabulary for expressing the same types of impact?
- How are topic and genre related to each other? Are certain topics more likely in some genres than in others?

This paper makes three main contributions to our ongoing research. The first is that it contributes to our understanding of the reading impact model, and through it, of the language of reading impact. We formalize the ability to tell genres apart using the *keyness* of impact terms. Thus, we now have quantitative support to argue that certain impact terms are strongly connected to certain genres and less to others. Second, we find that the topics from novels can be clustered into broader themes that lead to distinct thematic profiles per genre. There is a clear relation between impact terms and genre, but not between impact terms and topic or theme. In the discussion at the end we elaborate on this and provide possible explanations for this finding. The third contribution is the insight that the key impact terms per genre give an indication of the motivation of readers to read a book and how the reading experience relates to their expectations.

2. Background

We are interested in what kind of impression novels leave with their readers. Can we measure this so-called “impact” and how does it relate to features of the actual novel texts? Several studies have tried to link success or popularity of texts to features of those texts. Some studies have related pace, in the sense of how much distance the same length of texts covers in a semantic space, to success; finding that success correlates with higher pacing of narrative (Toubia et al. 2021, Laurino Dos Santos and Berger 2022). It has been argued that songs of which lyrics deviates from a genre’s

usual pattern tend to be more popular (Berger and Packard 2018). Other work relates topic models 92
to surveyed ratings of literariness suggests the same for fiction novels (Cranenburgh et al. 2019). 93
Moreira et al. apply “sentiment arc features [...] and semantic profiling” with some success to 94
predict ratings on Goodreads (Moreira et al. 2023). Taking the number of Gutenberg downloads 95
as a proxy for success Ashok et al. (2013) reach 84% accuracy in predicting popularity based 96
on learning low level stylistic features of the text of novels. Van Zundert et al. (2018) use sales 97
numbers as a proxy for popularity in an machine learning attempt to predict success, concluding 98
that the theme of masculinity is at least one major driver of successful fiction. 99

Common to all these studies is that they target some proxy of success or popularity: Goodreads 100
ratings, sales numbers, download statistics, and so forth. However, to our knowledge no research 101
has tried to link concrete features of fiction narratives to textual features of reviews from readers. 102
We seek to uncover if there is such a relation and if it may be meaningful from a literary research 103
perspective. In our present study we apply a heuristic model for impact features (Boot and Koolen 104
2020) to a corpus of 600,000+ reader reviews mined from several online review platforms. We 105
attempt to relate collocations of impact related terms to genre. Advancing previous research on 106
genre and topic models (Van Zundert et al. 2022) our contribution in this paper is to examine how 107
collocated impact terms relate to genre and genre to topic models of novels, thus offering a first 108
insight into the relation between topics (understood in terms of topic model) and reader reported 109
impact measures. Such work needs to take into account the plethora of problems that surround 110
the application of topic models to downstream tasks. This concerns topics content wise, which is 111
to say that topic models in contrast to their name do not often express much topical information. 112
Rather they may be connected to meta-textual features, such as author (Thompson and Mimno 113
2018), genre (Schöch 2017), or structural elements in texts (Uglanova and Gius 2020). 114

Our current contribution leans more to the side of data exploration than to the side of offering 115
assertive generalizations. We are interested in empirically quantifying the impact that the text 116
of novels has on readers. Any operationalization of this research aim necessarily involves many 117
narrowing choices and, at least initially, the audacious naivety to ignore the stupefying complexity 118
of social mechanisms to which readers are susceptible and thus the mass of confounding text 119
external factors that also drive reader impact. In our setup we assume that there are at least some 120
textual features, such as style, narrative pace, plot, character likability, that may be measured 121
and that can be related to reader impact. We further assume that book reviews scraped from 122
online platforms do serve as a somewhat reliable gauge to measure reader impact. We make these 123
cautioning statements not just proforma, but because we know that our information is selective, 124
biased, and skewed. Thanks to the stalwart experts of the Dutch National Library we do have for 125
our analysis the full text of 18,885 novels in Dutch (both translated and of Dutch origin). We also 126
have 634,614 online reviews, gathered by scraping for platforms such as Goodreads, Hebban¹, 127
and so forth. This corpus is biased. Romance novels comprise only about 3% of the corpus of 128
full texts. This is in stark contrast to its undisputed popularity (cf. Regis 2003, p. xi: “In the last 129
year of the twentieth century, 55.9% of mass-market and trade paperbacks sold in North America 130
were romance novels”). If our book corpus is skewed, our review data is even more so: only 1% of 131
reviews pertain to novels in the romance genre. Obviously we attempt to balance our data with 132
respect to genre and other properties for analysis. Yet, we should remind ourselves of the limited 133
representativeness of our data, which necessitates modesty as to generalizing results. Hence, what 134
follows is more offered as data exploration than as pontification of strong relations. 135

1. See <https://www.hebban.nl/>.

3. Data and Method

136

Our corpus of 18,885 books consists of mostly fiction novels and some non-fiction books in the Dutch language (both originally Dutch and translated). The review corpus boasts 634,614 Dutch book reviews. Obviously we do not have reviews for each book, nor does the set of books fully cover the collection of reviews, but we have upward of 10k books with at least one review.

3.1 Preprocessing

141

Both books and reviews are parsed with Trankit (Nguyen et al. 2021). Reading impact is extracted from the reviews using the Dutch Reading Impact Model (DRIM) (Boot and Koolen 2020).

Topic Modelling For topic modelling of the novels we use Top2Vec (Angelov 2020), and created a model with whole books as documents. We apply multiple filters to select terms that signal topic. Following the advice from previous work (Sobchuk and Šeĭa 2023; Uglanova and Gius 2020; Van Zundert et al. 2022), we focus on content words and select only nouns, verbs, adjectives and adverbs and remove any person names identified by the Trankit NER tagger. Our assumption is that person names have little to no relationship with topic, but are strong differentiating terms that tend to cluster parts of books and book series with recurring characters. Names of locations can have a similar effect, but, at least where the setting reflects the real world, we argue that this setting aspect of stories is more meaningfully related to topic. The book corpus contains 1,922,833,614 tokens including all punctuation and stop words. After filtering, 826,226,855 tokens remain. The next filter is a frequency filter. We remove terms that occur in fewer than 1% of documents or in more than 50% of documents. This leaves 190,607,470 tokens, which is 23% of all content words and just under 10% of the total number of tokens². Books have a mean (median) number of 42,959 (37,940) *content* tokens. The number of tokens is a Poisson distribution, therefore left-skewed, with 68% (corresponding to data within 1 standard deviation from the mean) of all books having between 17,509 and 63,418 tokens. This shows that the books have a high variation in length, but the majority books have a length within a single order of magnitude. After filtering on document frequency, the mean (median) number of tokens is 9,979 (8,325), with 68% having between 3,847 and 14,992 tokens.

Reading Impact Modelling The DRIM is a rule-based model and works at the level of sentences. It has 275 rules relating to impact in four categories: *Affect*, *Aesthetic* and *Narrative* impact, and *Reflection*. Both *Aesthetic* and *Narrative* impact are sub-categories of *Affect*, so rules that identify expressions of the sub-categories are also considered expressions of *Affect* (Boot and Koolen 2020). The rules for *Reflection* were not validated (see Boot and Koolen 2020) so we exclude *Reflection* from our analysis. For our analysis of topic, we expect that *Narrative* is the most directly related category, but we also include general *Affect* in our analysis. Expressions identified by the model consist of at least an impact word or phrase, such as “spannend” (*suspenseful*).³ However, many rules require there to be a book aspect term as well. For instance, the evaluative word “goed” (*good*) by itself can refer to anything. To be considered part of an impact expression it needs to co-occur in one sentence with a word in one of the book aspect categories, e.g. a style-related word

2. Experiments with using different frequency ranges for filtering suggests that the topic modelling process is relatively insensitive with regards to the upper limit. I.e. using 50%, 30% or 10% results in roughly equal numbers of topics that show the same relationship with book genre (see Section 4.1.1 and the following notebook: https://github.com/impact-and-fiction/jcls-2024-topic-genre-impact/blob/main/notebooks/topic_and_genre.ipynb)

3. For all Dutch terms we will consistently provide English translation in italics between parentheses.

like “geschreven” (*written*) to be an expression of *Aesthetic* impact, or a narrative-related word like “verhaal” (*story*) or “plot” to be an expression of *Narrative* impact.

The DRIM identified 2,089,576 expressions of impact in the full review dataset. To identify the key impact terms per genre, we use the full review dataset with all 2.1M impact expressions. To make a clearer distinction between impact expressions of generic affect and affect specific to narrative or aesthetics, we consider as *Affect* only those expressions that are not also categorized as *Narrative* or *Aesthetic*. Of the 2,089,576 expressions, there are 667,672 expressions for *Aesthetic* impact, 690,184 for *Narrative* impact and 731,720 for generic *Affect*.

3.2 Connecting Books and Reviews

A crucial step in relating topic in fiction to reading impact expressed in reviews, we need to connect the books to their corresponding reviews. For this, we rely mostly on ISBN⁴ and author and book title. Note that a particular work may be connected to multiple ISBNs, for instance when reprints or new editions are produced for the same work with a different ISBN. Many mappings between reviews and books, and between multiple ISBNs of the same work were already made by Boot 2017 and Koolen et al. 2020, for the Online Dutch Book Response (ODBR) dataset of 472,810 reviews. We added around 160,000 reviews from Hebban to the ODBR set. To find ISBNs that refer to the same work, we first queried all ISBNs found in reviews using the SRU⁵ service of the National Library of the Netherlands. This SRU service gives access to the combined catalog of Dutch libraries and in many cases links multiple editions of the same work with different ISBNs. Using author and title we resolved another number of duplicated works with different ISBNs. We then mapped all ISBNs of the same work to a unique work ID and linked the reviews via the ISBNs they mention to these work IDs. There are 125,542 distinct works reviewed by the reviews in our dataset. Of the 18,885 books for which we have ePubs, there are 10,056 books with at least one review in our data set. Altogether these 10,056 unique works are linked to 130,751 reviews.

3.3 Connecting Impact and Topic Data

Our goal was to have a comprehensive mapping of the most relevant topics of works to their reviews, the latter analyzed via the DRIM. To create this dataset, we needed to connect the expressions of impact to the topics in our book dataset. To do so, we took the top five dominant topics of each book⁶, and linked those topics to the impact expressions in the reviews of the books for that topic. This resulted in a dataset whereby each entry links specific reviews to the top 5 dominant topics for every book.

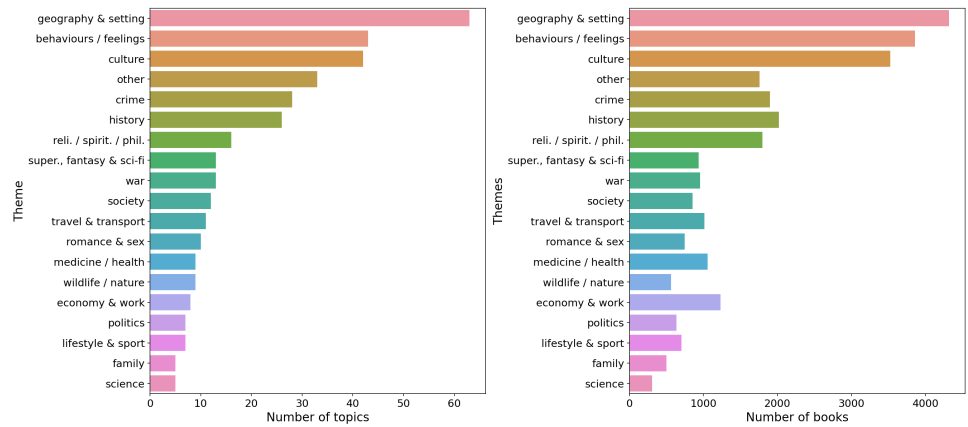
The Top2Vec model gave us a total of 228 topics. We attempted to label each topic with a distinct content label, but found that many topics are thematically very similar, capturing many of the same elements. Therefore, we manually assigned each topic to one or more of 19 broader themes: 1. *geography and setting*, 2. *behaviors/feelings*, 3. *culture*, 4. *crime*, 5. *history*, 6. *religion, spirituality and philosophy*, 7. *supernatural, fantasy and sci-fi*, 8. *war*, 9. *society*, 10. *travel and transport*, 11. *romance and sex*, 12. *medicine/health*, 13. *wildlife/nature*, 14. *economy and work*, 15. *lifestyle and sport*, 16. *politics*, 17. *family*, 18. *science*, 19. *other*. We provide the number of topics grouped per

4. International Standard Book Number, see: <https://en.wikipedia.org/wiki/ISBN>.

5. Search and Retrieval by URL, see: https://en.wikipedia.org/wiki/Search/Retrieve_via_URL.

6. Topc2Vec creates topics by clustering the document vectors and taking the centroid of each cluster as the topic vector. We computed the cosine similarity between the document vector (representing the book) and the topic vectors, and selected the top five closest (i.e., most similar) topics to each book.

Figure 2: The number of topics and books per theme.



theme in Figure 2⁷.

We provide the full list of topics, themes and their respective words in our code repository⁸.

3.4 Book Genre Information

For genre information about books, we use the Dutch NUR classification codes assigned by publishers. As NUR was designed as a marketing instrument to determine where books are shelved in bookshops, publishers can choose codes based not only on the perceived genre of a book but also on marketing strategies related to where they want a book to be shelved to find the biggest audience. Some NUR codes refer to the same or very similar genres. E.g. codes 300, 301, and 302 refer respectively to *general literary fiction*, *Dutch literary fiction*, and *translated literary fiction*, which we group together under *Literary fiction*. Similarly, we group codes 313, 330, 331, 332, and 339 under *Suspense novels*, as they all refer to types of suspense, i.e. *pockets suspense*, *general suspense novels*, *detective novels* and *thrillers* respectively. In total, we select 19 different NUR codes and map them to 9 genres. All remaining NUR codes in the fiction range (300-350) we map to *Other fiction* and the rest to *Non-fiction*. The full mapping is available in our code repository⁹.

3.5 Keyness Analysis on Impact Terms

The goal of this analysis is to determine (i) *which* words readers use in their reviews to describe the impact of a particular book, and (ii) how *characteristic* these words are for a particular genre, compared to another genre. A good candidate to measure both (i) and (ii) is keyword analysis, or keyness (Dunning 1994; Gabrielatos 2018; Paquot and Bestgen 2009).

There is ample literature comparing different keyness measures (Culpeper and Demmen 2015; Du et al. 2022; Dunning 1994; Gabrielatos 2018; Lijffijt et al. 2016), finding that no single measure is perfect.

A commonly used measure is G^2 , which identifies *key* terms that occur statistically significantly more or less often in a target corpus (the reviews for a particular genre) compared to a reference

7. Note that in this paper “theme” should not be taken to coincide with the literary studies sense of theme. Rather we use the term “theme” to clearly distinguish between the topics as identified by Top2Vec and their clustering as done by us.

8. See https://github.com/impact-and-fiction/jcls-2024-topic-genre-impact/blob/main/data/topic_labels.tsv.

9. See https://anonymous.4open.science/r/jcls-2024-topic-genre-impact-EB46/data/nur_genre_map.md.

corpus (reviews of one or more other genres). 236

Lijffijt et al. (2016) showed that Log-Likelihood Ratio (G^2 , Dunning 1994) and several other 237
frequency-based bag-of-words keyness measures suffer from excessively high confidence in the 238
estimates because these measures assume samples to be statistically independent, but words in a text 239
are not independent of each other. Du et al. (2022) compare frequency-based and dispersion-based 240
measures for a downstream task (text classification) to show that for identifying key terms in a 241
sub-corpus compared to the rest of the corpus, dispersion-based measures are more effective. 242

To compare the dispersion of a word or phrase in a target corpus to its dispersion in a reference 243
corpus, Du et al. (2021) introduce *Eta*, which is a variant of the *Zeta* measure by Burrows (2006). 244

They find that *Eta* Du et al. 2021 and *Zeta* Burrows 2006 are among the most effective measures. 245
Both *Eta* and *Zeta* compare document proportions of keywords. The former uses Deviation of 246
Proportions (*DP*) Gries 2008 which computes two sets of proportions. The first are the proportions 247
that the lengths of documents represent with respect to the total number of words in a corpus 248
(e.g. the set of reviews for books of a specific genre) as an expected distribution of proportions of 249
keywords. The second is the set of observed proportions of a keyword across a corpus with respect 250
to the total corpus frequency of that keyword. There are two problems with using *DP* for keyness 251
of impact terms. The first is that some impact terms do not occur in any of the reviews of a specific 252
genre. In such cases, the observed proportions are not properly defined (a proportion of zero is not 253
well-defined), so *DP* cannot be computed. The second is that the frequency distribution of impact 254
terms in reviews is extremely skewed (84% of all impact terms in reviews have a frequency of 1, 255
13% occur twice and the remaining 3% occur three or four times). Although longer reviews have a 256
higher a priori probability of containing a specific impact term than shorter reviews, the frequency 257
distribution of individual impact terms behaves more like a binomial distribution, so length-based 258
proportions are not an appropriate measure of keyness. 259

Because of this, we instead measure dispersion using *document frequencies* (the number of reviews 260
for a book genre in which an impact term occurs) to compute the *document proportion* (the fraction 261
of reviews for a book genre in which an impact term occurs at least once). This gives document 262
proportion $docP(t, G)$ per impact term t and genre G , with the absolute difference *Zeta* between 263
two genres defined as $Zeta(t, G_1, G_2) = abs(docP(t, G_1) - docP(t, G_2))$. 264

To illustrate this approach, we compare the document proportions per genre of the impact terms 265
“stijl” (style) and “schrijfstijl” (writing style). The former has the highest document proportion for 266
reviews of *Literary fiction* (occurring in 3.7% of reviews) and least in those of *Non-fiction* (1.2%), 267
resulting in $Zeta = 0.037 - 0.012 = 0.025$. The latter is most common in reviews of *Romanticism* 268
(14.6%) and least common in those of *Non-fiction* (2.0%), giving $Zeta = 0.146 - 0.02 = 0.126$. 269

4. Results 270

4.1 Topic and Genre 271

Van Zundert et al. (2022) found that the topics identified with Top2Vec are strongly associated with 272
genre as identified by publishers. Similarly, Sobchuk and Šeĭa 2023 find that Doc2Vec – which is 273
used by Top2Vec to embed the documents in the latent semantic space in which topic vectors are 274
identified – is more effective at clustering books by genre than the topic modeling technique LDA 275
(Blei et al. 2003). 276

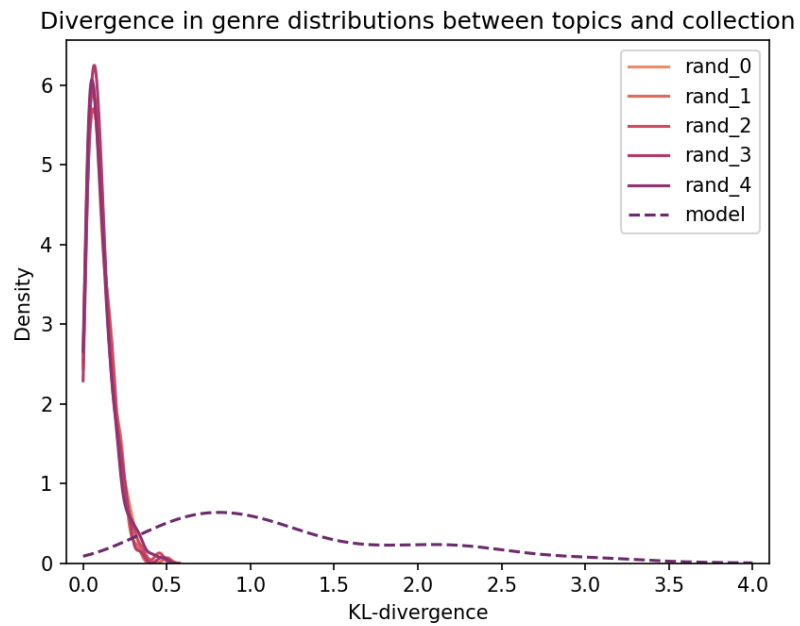


Figure 3: The KL-divergence between the genre distribution per topic and that of the collection for the topic model as well as for five random shufflings of genre labels using the same books per topic.

4.1.1 Genre Distribution per Topic

277

To extent the findings of Van Zundert et al. 2022, we first quantitatively demonstrate that there is a relationship between topic and genre. Each topic is associated with a number of books and thereby with the same number of genre labels. From eyeballing the distribution of genre labels per topic, it seems that for most topics, the vast majority of books in that topic belong to a single genre. But the genre distribution of the entire collection is also highly skewed, with a few very large genres and many much smaller genres. So perhaps the skew in most topics resembles the skew of the genre distribution of the collection.

To measure how much the genre distribution per topic deviates from that of the collection, we compute the KL-divergence between the two distributions. This gives a set of 228 deviations from the collection distribution.

But whether these deviations are small or large is difficult to read from the numbers themselves. For that, we should compare them against a random shuffling of the book genres across books (while keeping the books assigned per topic stable). For large topics (with many books), a random shuffling should have a genre distribution close to that of the collection. For small clusters, the divergence will tend to be higher.

We create five alternative clusterings with books randomly assigned to topics with the same topic size distribution as established by the topic model. The distribution of the 228 KL-divergence scores per model (five random and one topic model) are shown in Figure 3. The five random models have almost identical distributions concentrated around 0.1 with a standard deviation of around 0.075 and a max of around 0.5. The genre distribution of the topic model is very different, with a median score of 1.06 and more than 75% of all scores above 0.68.

From this quantitative analysis, it is clear that there is a strong relationship between topic and genre.

4.1.2 Thematic Distribution per Genre 300

Next, we perform a qualitative analysis of the topics and their relationship to genre. 301

The distribution of topic themes per genre is shown in Figure 4 in the form of radar plots. The 302
genres show distinct thematic profiles. Literary fiction scores high on the themes of *Culture*, 303
Geography & setting and *Behaviors & feelings*, which is perhaps not surprising. Non-fiction scores 304
high on *Religion, spirituality, and philosophy*, *Medicine & health*, *Economy & work*, and *Behaviors* 305
& *feelings*, which are themes that few fiction genres score high on. 306

In Children's fiction, there is relatively little use of the geographical aspect of setting, especially 307
compared to other fiction genres. That is, it seems that children's novels make little explicit reference 308
to geographical places. They score high on *behaviors and feelings* and moderately high on *Culture*, 309
Family and *Supernatural, fantasy & sci-fi*. The main difference between Children's fiction and 310
Young Adult is that the latter scores higher on *Supernatural, fantasy and sci-fi*. On the former 311
theme, Young Adult strongly overlaps with Fantasy novels. Young Adult also adds in a bit of 312
Romance and sex. These observations suggest that Children's fiction and Young Adult by and large 313
treat the same themes but against different 'backgrounds'. Children's fiction is about behaviors 314
and feelings against a backdrop made up of culture and family. Young adult does practically the 315
same, but adds supernatural, fantasy, and sci-fi elements to the story, and opens the stage for some 316
romantic behavior. 317

If one would want to hazard a guess at reader development, it would almost seem as if young 318
readers are invited to pre-sort on the major themes of grown-up literature where *Romance* amplifies 319
the romance and sex encountered in *Young adult* books, while *Literary fiction* and *Literary thrillers* 320
amplify motifs of culture, setting, and crime, and *Fantasy* caters to the interest in the supernatural 321
developed through Young adult fiction. Much more research would be needed, however, to 322
substantiate such a pre-sorting effect. In any case, Romanticism scores high on *Romance and sex* 323
and has medium scores for *Culture* and *Geography and setting*, while Suspense novels score high on 324
Crime, and have medium scores for *Geography and setting* and *War*. 325

We expect that many of these observations coincide with intuitions of literary researchers. This 326
suggests that the grouping of topics by theme makes sense from a literary analytical perspective in 327
any case. The findings also shows where genres overlap and where they differ. For instance, the 328
profile for Literary fiction and Literary thriller are similar, with the main difference being the much 329
higher prevalence of the *Crime* theme in Literary thrillers. Suspense is similar to Literary thrillers 330
in the prevalence of *Crime* as theme, but lower scores for *Culture* and *Geography and setting*. 331

One of the main findings is that, for the chosen document frequency range of mid-frequency terms, 332
there is a clear connection between topic and genre, with thematic clustering of topics leading to 333
distinct genre profiles, but also to thematic connections between certain genres. None of this will 334
radically transform our understanding of genre and topic, but it prompts the question how different 335
parts of the document frequency distribution relate to different aspects of novels. From authorship 336
attribution research we know that authorial signal is mainly found in the high-frequency range, and 337
our work corroborates earlier findings that topics contain genre-signals in mid-range frequencies 338
(Thompson and Mimno 2018; Van Zundert et al. 2022). 339

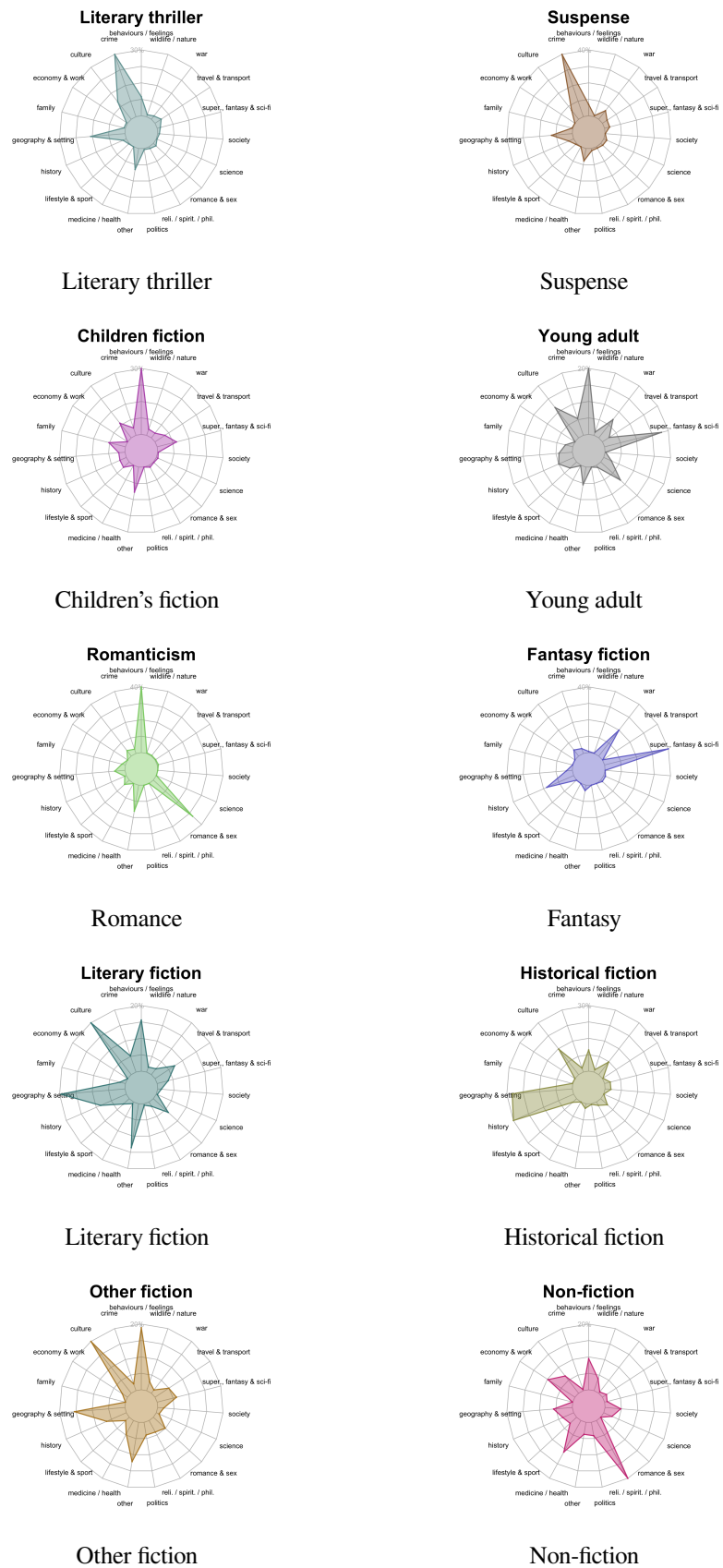
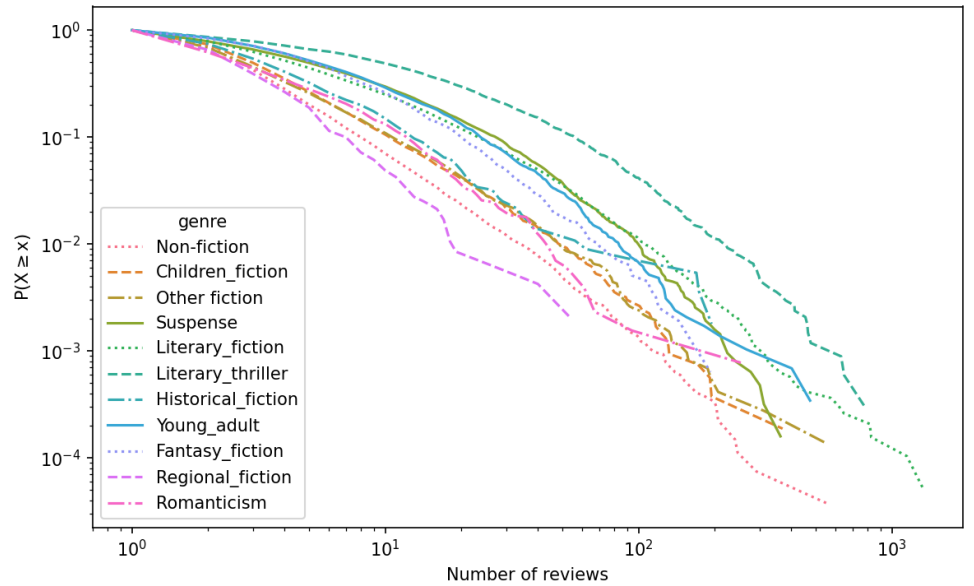


Figure 4: Radar plots showing the relative prevalence of themes in six genres, from left to right, top to bottom: *Literary thrillers*, *Suspense*, *Children's fiction* and *Young adult*, *Romance*, *Fantasy*, *Literary fiction*, *Historical fiction*, *Other fiction* and *Non-fiction*.

Table 1: Reviews per genre and mean number of reviews per book, per genre.

	Reviewed books	Reviews	Mean Reviews/book
Literary fiction	19288	200907	10.4
Literary thriller	3394	77288	22.8
Young adult	2919	30552	10.5
Children fiction	5348	27989	5.2
Suspense	6266	67990	10.9
Fantasy fiction	1571	13739	8.7
Romanticism	1291	6434	5.0
Historical fiction	556	3463	6.2
Regional fiction	472	1528	3.2
Other fiction	7260	37515	5.2
Non-fiction	26884	109158	4.1

**Figure 5:** The cumulative distribution function of the number of reviews per book, on a log-log scale. The Y-axis shows that probability $P(X \geq x)$ that a book has at least x reviews.

4.2 Impact and Genre

340

4.2.1 Reviews per Genre

341

With the genre labels, we can count how many books in each genre have reviews in our dataset, and 342 how many reviews they have (Table 1). It is clear that *Literary fiction* is reviewed most often, with 343 200,907 reviews in our dataset, followed by *Literary thrillers* and *Suspense novels*. *Literary thrillers* 344 have the highest mean number of reviews per book. However, the distribution of the number of 345 reviews per book is highly skewed, with a single review per book being the most likely, and having 346 more reviews being increasingly unlikely (Koolen et al. 2020). The distributions per genre show 347 some differences, but all are close to a power-law. The cumulative distribution function of the 348 number of reviews per book for the different genres are shown in Figure 5, with on the Y-axis the 349 probability $P(X \geq x)$ that a book has at least x reviews.¹⁰ 350

The curves for some of the genres overlap, which makes them difficult to discern, but there are a 351 few main insights. First, *regional fiction* and *non-fiction* have the fastest falling curves, indicating 352 that books in these genres are the least likely to acquire many reviews. Next is a cluster of *children's* 353 *fiction*, *romanticism*, *historical fiction* and *other fiction*, which tend to get a slightly higher number 354 of reviews. Then there is a cluster of *suspense*, *literary fiction*, *young adult* and *fantasy fiction*, 355 which tend to get more reviews than the previous cluster. And finally, clearly above the rest, is the 356 curve of *literary thrillers*, which tend get more reviews than books in any other genre. 357

Thrillers are more often reviewed on the platforms that are in the review dataset. *Romance* novels 358 have fewer reviews but are a very popular genre (Regis 2003, p. 108, see also: Darbyshire 2023). 359 This prompts the question of whether readers of *regional* and *romance* novels have less desire to 360 review these novels or review them on different platforms and in different ways. As there seem 361 to be many video reviews of *romance* novels on TikTok using the tag #BookTok, this would be a 362 valuable resource to add to our investigations. A difference in the number of reviews might be a 363 signal of a difference in impact, but it is also plausible that different genres attract different types of 364 readers who express their impact in different ways linguistically, using different media (e.g. text or 365 video) on different platforms (e.g. GoodReads or TikTok). To that extent, the review dataset may 366 be a biased representation of the impact of books in different genres. Bracketing for a moment 367 the potential skewedness of the number of reviews per genre, and taking number of reviews as a 368 proxy of popularity, it is also interesting to observe that popularity is apparently a commodity that 369 is reaped in orders of magnitude. 370

4.2.2 Key Impact Terms per Genre

371

Correlations between genres First, we compare genres in terms of their impact terms 372 through the percent difference per impact term. For each pair of genres, we compute the Pearson 373 correlation ρ between the %Diff scores of all impact terms. A high positive correlation means 374 that impact terms with high (low) %Diff scores in one genres, tend to also have high (low) %Diff 375 scores in the other genre. 376

The correlations per impact type are shown Figure 6. For *Affect* impact terms (the top correlation 377 table), many of the genre pairs have no correlation ($-0.25 < \rho < 0.25$). There are some weak 378 positive and negative correlations ($0.25 < \rho < 0.50$ and $-0.50\rho < -0.25$ respectively) and 379

10. We show the cumulative distribution instead of the plain distribution, because it produces smoother curves and better shows the trends.

Affect											
	Child. fic	Fantasy	Hist. fic	Lit. fic	Lit. thrill	Non-fic	Oth. fic	Reg. fic	Romance	Suspense	YA
Child. fic	1.00	0.44	-0.15	-0.34	-0.15	0.17	-0.11	-0.09	0.14	-0.12	0.50
Fantasy	0.44	1.00	0.09	-0.42	0.22	-0.20	-0.14	-0.13	0.05	0.24	0.60
Hist. fic	-0.15	0.09	1.00	-0.03	0.07	-0.10	-0.15	0.05	0.22	-0.11	0.18
Lit. fic	-0.34	-0.42	-0.03	1.00	-0.45	-0.18	-0.14	-0.08	-0.09	-0.36	-0.40
Lit. thrill	-0.15	0.22	0.07	-0.45	1.00	-0.38	-0.19	0.21	-0.09	0.61	0.13
Non-fic	0.17	-0.20	-0.10	-0.18	-0.38	1.00	-0.05	-0.09	-0.07	-0.37	-0.15
Oth. fic	-0.11	-0.14	-0.15	-0.14	-0.19	-0.05	1.00	-0.05	-0.08	0.01	-0.17
Reg. fic	-0.09	-0.13	0.05	-0.08	0.21	-0.09	-0.05	1.00	0.39	-0.08	0.08
Romance	0.14	0.05	0.22	-0.09	-0.09	-0.07	-0.08	0.39	1.00	-0.21	0.34
Suspense	-0.12	0.24	-0.11	-0.36	0.61	-0.37	0.01	-0.08	-0.21	1.00	-0.03
YA	0.50	0.60	0.18	-0.40	0.13	-0.15	-0.17	0.08	0.34	-0.03	1.00

Narrative											
	Child. fic	Fantasy	Hist. fic	Lit. fic	Lit. thrill	Non-fic	Oth. fic	Reg. fic	Romance	Suspense	YA
Child. fic	1.00	0.60	-0.12	-0.41	-0.06	0.09	-0.13	0.02	0.07	0.07	0.46
Fantasy	0.60	1.00	-0.02	-0.41	0.09	-0.29	-0.15	-0.10	0.13	0.24	0.48
Hist. fic	-0.12	-0.02	1.00	0.05	-0.01	-0.15	-0.18	0.50	0.46	-0.33	0.24
Lit. fic	-0.41	-0.41	0.05	1.00	-0.44	-0.04	-0.18	0.01	-0.07	-0.35	-0.32
Lit. thrill	-0.06	0.09	-0.01	-0.44	1.00	-0.38	-0.25	-0.12	-0.14	0.32	0.12
Non-fic	0.09	-0.29	-0.15	-0.04	-0.38	1.00	-0.08	0.02	-0.06	-0.30	-0.22
Oth. fic	-0.13	-0.15	-0.18	-0.18	-0.25	-0.08	1.00	-0.14	-0.10	-0.00	-0.23
Reg. fic	0.02	-0.10	0.50	0.01	-0.12	0.02	-0.14	1.00	0.48	-0.30	0.20
Romance	0.07	0.13	0.46	-0.07	-0.14	-0.06	-0.10	0.48	1.00	-0.25	0.44
Suspense	0.07	0.24	-0.33	-0.35	0.32	-0.30	-0.00	-0.30	-0.25	1.00	-0.26
YA	0.46	0.48	0.24	-0.32	0.12	-0.22	-0.23	0.20	0.44	-0.26	1.00

Style											
	Child. fic	Fantasy	Hist. fic	Lit. fic	Lit. thrill	Non-fic	Oth. fic	Reg. fic	Romance	Suspense	YA
Child. fic	1.00	0.16	-0.05	-0.25	-0.30	0.08	-0.07	0.13	0.43	-0.36	0.60
Fantasy	0.16	1.00	0.03	-0.43	0.49	-0.24	-0.46	0.06	0.02	0.19	0.43
Hist. fic	-0.05	0.03	1.00	-0.21	0.24	-0.06	-0.34	0.54	-0.08	0.10	0.15
Lit. fic	-0.25	-0.43	-0.21	1.00	-0.44	-0.36	0.34	-0.34	-0.18	-0.44	-0.49
Lit. thrill	-0.30	0.49	0.24	-0.44	1.00	-0.31	-0.65	0.16	-0.23	0.57	0.13
Non-fic	0.08	-0.24	-0.06	-0.36	-0.31	1.00	0.17	-0.03	-0.04	-0.25	-0.12
Oth. fic	-0.07	-0.46	-0.34	0.34	-0.65	0.17	1.00	-0.10	0.24	-0.35	-0.31
Reg. fic	0.13	0.06	0.54	-0.34	0.16	-0.03	-0.10	1.00	0.39	0.05	0.33
Romance	0.43	0.02	-0.08	-0.18	-0.23	-0.04	0.24	0.39	1.00	-0.23	0.47
Suspense	-0.36	0.19	0.10	-0.44	0.57	-0.25	-0.35	0.05	-0.23	1.00	-0.01
YA	0.60	0.43	0.15	-0.49	0.13	-0.12	-0.31	0.33	0.47	-0.01	1.00

Figure 6: Pearson correlation in the %Diff scores of impact terms between pairs of genres, for *Affect* (top), *Narrative* (middle) and *Style* (bottom).

moderate correlations ($0.50 < \rho < 0.75$ and $-0.75\rho < 0.50$). There are a few clusters of genres with high correlations in %Diff scores, signaling that some genres differ in how impact is expressed and that the DRIM is sensitive to difference between genres. The cluster of Children's fiction, Young adult and Fantasy have weak (0.44) and moderate (0.50 and 0.60) correlations with each other, suggesting that impact terms that are typical for one, are to some extent also typical for the other two. Other clusters are Literary thriller and Suspense novels, with a moderate correlation of 0.61, and Romance and Regional fiction with a moderate correlation of 0.39.

Literary fiction is the one genre with mostly weakly negative correlations, with Children's fiction (-0.34), Fantasy (-0.42), Literary thriller (-0.45), Suspense (-0.36) and Young adult (-0.40). With the remaining three genres, literary fiction has no correlation. In other words, in terms of *Affective* impact, reviews of Literary fiction uses a different register than reviews of other genres.

For *Narrative* impact, we find the same cluster of Children's fiction, Young adult and Fantasy. The cluster of Regional fiction and Romance here also contains Historical fiction, and the two clusters are linked by the moderate correlation of 0.44 between Romance and Young adult. The other genres in the two clusters have no or a negative correlation with each other. Here also the genres of Literary thriller and Suspense novels show a weak correlation (0.32), and Literary fiction has no or at most moderately negative correlations with the other genres. The top impact terms for Thrillers and Suspense novels largely overlap and contain several narrative impact terms relating to plot, e.g. "spannend" (*thrilling* or *suspenseful*), "spanning" (*suspense*), "verrassend", "verrassend" and "onverwacht" (*surprise*, *surprising* and *unexpected* respectively). For Romance and Regional fiction, the top 10 narrative impact terms almost completely overlap, with shared narrative impact terms "romantisch" (*romantic*), "ellende" (*lifelike*), "verdriet" (*sadness*), "leveneucht" (*lifelike*), "fijn" (*nice*), "heerlijk" (*lovely*) and "nieuwsgierig" (*curious*).

Overall, there are more weak negative correlations between pairs of genres than for *Affective* impact were non-existent.

The correlations for *Style* are more different. Children's fiction no longer has a weak positive correlation with Fantasy, but it does with Romance. Children's fiction and Young adult still have a moderately positive correlation and Young adult also have weak correlations with Fantasy and Romance. The biggest shifts are for Romance, which no longer has any correlation with Historical fiction, but now has a weakly positive correlation with Children's fiction. For Literary thrillers there are several weakly and moderately negative correlations with Children's fiction (-0.30), Literary fiction (-0.44), Non-fiction (-0.31) and Other fiction (-0.65). Literary fiction is also in terms of *Style* different from almost all genres apart from Other fiction. A speculative interpretation is that Literary fiction is stylistically distinctive in a similar way to the poetry that is part of the Other fiction genre.

Compared across the different impact types then, it appears that Literary fiction as a genre induces reviews where impact is described in a vocabulary distinct from impact reported in reviews pertaining to other genres. It is tempting to conjecture that Literary fiction attracts an audience of review writers that 'know how to talk' about literature. It is very well possible that these reviewers are acutely aware of the genre of literary review and that they apply conventions of this genre in their own review writing. For now this must remain indeed conjecture as a more focused examination of the vocabulary, style, and structure of these reviews has yet to be undertaken.

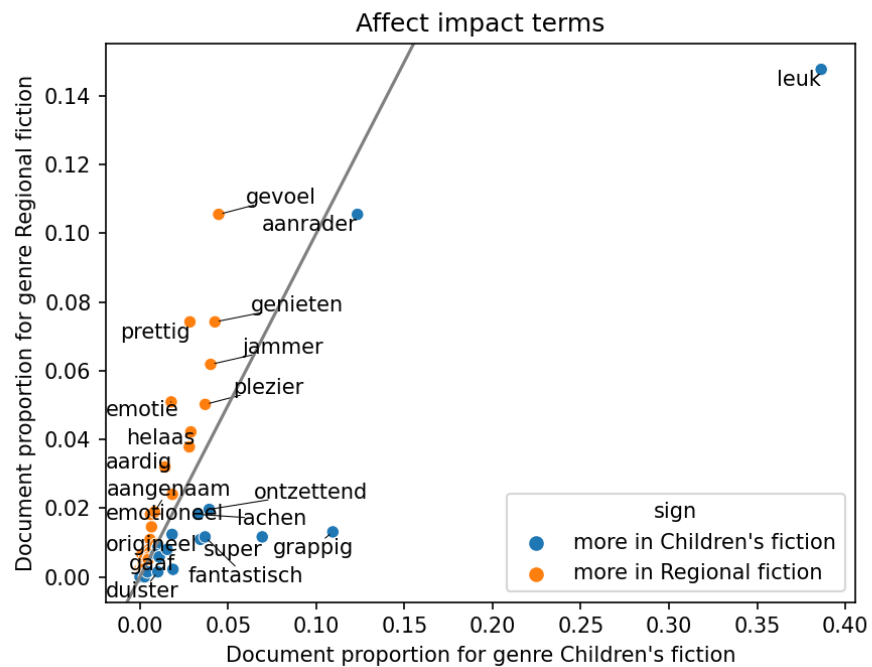


Figure 7: Document proportions of generic *Affect* terms for Children's fiction and Regional fiction.

Vocabulary differences between genres We compute the *Zeta* scores between pairs of genres for all impact terms and sum these scores per impact type to find which pairs of genres have the largest summed difference of *Zeta* scores. For generic *Affect*, Children's fiction is most distinctive as it has high score differences with all other genres. The document proportions for generic *Affect* terms of Children's fiction and *Regional fiction* are shown in Figure 7. The diagonal line shows where terms have equal proportions in both genres. Reviews of children's fiction seem to use a smaller impact vocabulary – almost all document proportions are close to zero – but much higher proportions for the impact term “leuk” (fun or cool). This term is used much less in reviews of other genres

For *Narrative* impact, the biggest summed difference is between Romance and Literary thrillers (see Figure 8). The main differences are found with a handful of terms, “spannend” (thrilling/suspenseful), “spanning” (suspense) and “verrassen” (surprise) are more common in Literary thrillers and “romantisch” (romantic) and “heerlijk” (lovely, wonderful) are more common in Romance novels. These are perhaps somewhat obvious, but show that impact, or at least the language of impact, is related to genre.

For *Aesthetic* impact, the biggest summed difference is between Romance and Historical fiction (see Figure 9). Here, the main differences are again with a few terms. Reviews of Historical fiction more often mention impact terms like “mooi” (beautiful), “beschrijven” (describe), “beschreven” (described) and “prachtig” (beautiful). Reviews of Romance novels more often mention “schrijfstijl” (writing style), “humor” (humor) and “luchtig” (airy). It seems that for Historical fiction, reviewers focus more on descriptions (how evocatively the author describes historical settings, persons or events perhaps), while reviewers of Romance novels focus more on humor and lightness of style. A close reading of some of the contexts in which “schrijfstijl” is mentioned in Romance reviews suggest that reviewers often use it in phrases like “makkelijke schrijfstijl” and “vlotte schrijfstijl” (a writing style that reads easily or quickly respectively).

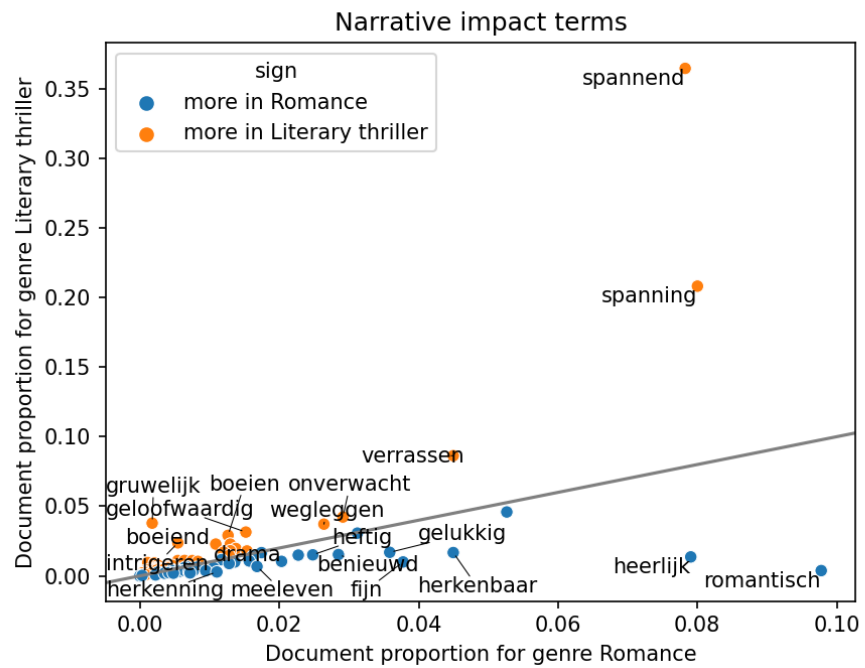


Figure 8: Document proportions of *Narrative* impact terms for Romance and Literary thrillers.

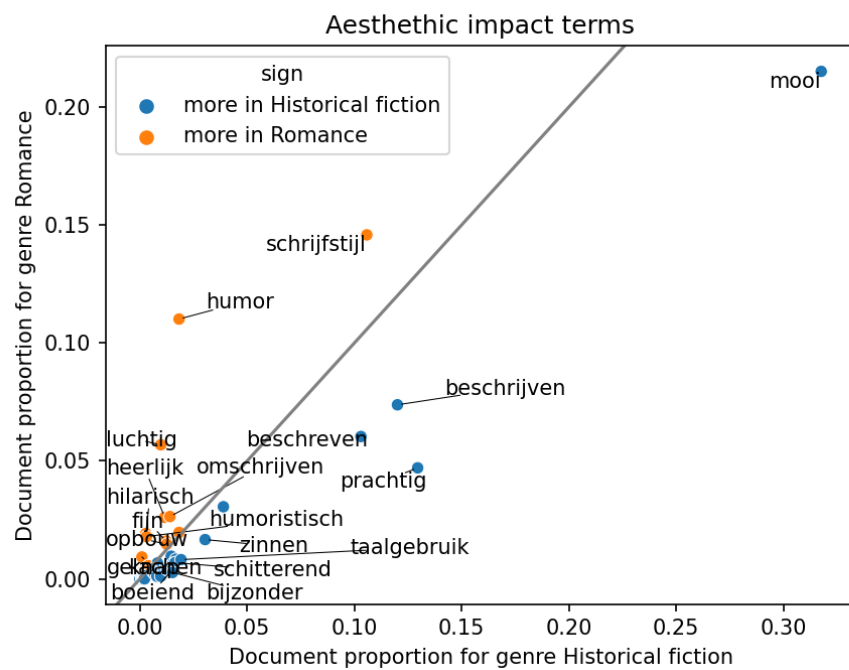


Figure 9: Document proportions of *Aesthetic* impact terms for Historical fiction and Romance.

4.3 Impact and Topic

447

The third link between the three main concepts that are the focus of this paper is between impact and topic.

To study how the use of impact terms differs between reviews of books with different themes, we first need to group the reviews by theme. Because themes are based on topics and some themes share the same topics, some reviews are assigned to multiple themes. We calculated correlations between themes in terms of the %Diff per impact term, just as we did for genre (see Figures 10, 11 and 12 in Appendix C). There are many observations that could be made, but again we limit ourselves to the most salient ones related to the three largest themes (in number of books).

Generic Affect

456

The theme *geography & setting* has a strong correlation for generic Affect with *history* ($\rho = 0.68$) and moderate correlations with *crime* ($\rho = 0.46$) and *war* ($\rho = 0.44$). This is not due to a large overlap in books, as *culture* has the largest overlap with *geography & setting* (sharing 49% and 40% of their books respectively), but a moderately negative correlation ($\rho = -0.41$). With all the other themes, *geography & setting* has no to moderately negative correlations. The connections with *crime*, *history* and *war* make sense, to the extent that for all these themes (we assume), the aspect of place plays an important role. Why this results in similarities of how generic affect is expressed is not immediately clear.

The theme *behaviors / feelings* has moderate correlations for generic Affect with *lifestyle & sport* ($\rho = 0.55$) and *romance & sex* ($\rho = 0.56$). This is partly explained by the latter themes sharing 15% and 22% of their books with *behaviors / feelings*, but it cannot be the only explanation. *Family* shares 65% of its books with *behaviors / feelings* but has no correlation ($\rho = 0.19$).

The theme *culture* has a near perfect correlation with *travel & transport* in terms of generic affect, but no to moderately negative correlations with all other themes. Here the overlap in books is minimal, the two themes sharing respectively 2% and 6% of their books. As mentioned above, With *em geography & setting* it has a moderately negative correlation ($\rho = -0.41$) despite its substantial overlap.

Narrative Impact

474

For Narrative impact, the correlations between *geography & setting* are somewhat different. We again find strong and moderate correlations with *history* ($\rho 0.65$) and *war* ($\rho 0.48$) respectively, but also with *religion, spirituality and philosophy* ($\rho 0.46$) and only a weak correlation with *crime* ($\rho 0.30$).

The theme *behaviors / feelings* only has strong correlation with *culture* ($\rho = 0.67$) but no or weakly negative correlations with all others, despite its overlap with *culture* (sharing 13% and 14% of their books respectively) being similar or lower than with *geography & setting* (sharing 13% and 12%) and with *economy & work* (sharing 12% and 36%). Overlap in books is clearly not the main explanation in overlap in the use of impact terms.

The *culture* theme has the strong correlation with *behaviors / feelings* mentioned above, but no or weakly negative correlations with other themes. Again, books with *em culture* as a theme have a different relationship with how reviewers describe impact than *geography & setting*, despite sharing a substantial number of books.

Aesthetic Impact 488

For *Aesthetic* impact, *geography & setting* has moderate correlations with *crime* ($\rho 0.35$), *culture* (489
($\rho 0.49$), *religion, spirituality and philosophy* ($\rho 0.42$) and *war* ($\rho 0.42$). With *crime*, *culture* and 490
war this could be due to their substantial overlap in books, but again, overlap cannot but the 491
full explanation, as *geography & setting* also substantially overlaps with *history* while having a 492
moderately negative correlation with it ($\rho - 0.41$). 493

The *behaviors / feelings* theme has a strong correlation with *romance & sex* ($\rho = 0.71$) and moderate 494
correlations with *family* ($\rho = 0.48$), *lifestyle & sport* ($\rho = 0.45$) and *science* ($\rho = 0.52$), and no or 495
negatively weak correlations with other themes. As mentioned before, 65% of books in the *family* 496
theme also belong to *behaviors / feelings*, but *science* shares no books with *behaviors / feelings*. 497

Just on these observations alone, it seems that themes have different relationships with how reviewers 498
express the impact of books that cover these themes. 499

5. Discussion & Conclusion 500

In this paper we investigated the relationship between three important concepts in literary studies: 501
genre, topic and impact (more commonly known as “reader response”). We discuss our findings 502
for each pair of concepts in turn. 503

Genre and Topic Our analyzes have corroborated earlier findings on the relationship between 504
genre and topic. By clustering topics identified by topic modelling into broader themes, and 505
measuring the prevalence of these themes in the books of specific genres, we find that topics have 506
a strong relation with genres, and the genres have distinct thematic profiles. These profiles match 507
existing intuitions about the distribution of themes across genres. Potentially these profiles can 508
provide additional insight in genre dynamics (e.g. as to what motivates readers to mix-read genres 509
or not) although much of this aspect remains to be examined. 510

Genre and Impact The Dutch Reading Impact Model (DRIM, Boot and Koolen 2020) 511
identifies sets of words that are to some extent related to genre, and by studying the overlap in key 512
impact terms between genres, we find clusters of genres that are similar in how their impact is 513
described. Of course, this is not entirely surprising. For instance, *Suspense novels* and *Literary* 514
thrillers are more similar in terms of overall impact. However, it is much less obvious or intuitive 515
that these two genres are more similar in terms of stylistic impact than in terms of narrative impact. 516
Neither is it immediately obvious why literary fiction with respect to all types of impact differs 517
most from other genres. 518

It remains unclear for now how we should explain the the relationship between impact and genre. 519
Perhaps this relation signals that reviewers develop and copy conventions of writing about books 520
from a certain genre by adopting what others in a genre-related community do. For instance, in a 521
community of reviewers around crime novels and literary thrillers reviewers might converge on a 522
shared vocabulary for talking about the plot and their reading experiences. It could also be that 523
different types of readers are drawn to different types of genres, with each group having their own 524
characteristics that shape how they write their reviews. Another possibility is that reviewers are 525
influenced by the language used by the authors of the novels they read, and how those authors 526
adopt genre conventions. Finally, depending on how the model was developed, this may also be 527

an artifact of how the rules were constructed. For instance, if reviews per genre were scanned to 528
 identify common expressions of impact. Further analysis is required to establish which, if any, of 529
 these factors contributes to the relationship between fiction genres and reading impact as expressed 530
 in reviews. 531

Topic and Impact For the first two pairs of concepts, there were some expectations, e.g. that 532
 there is a relation between the Romance genre and topics related to the theme of *Romance and* 533
sex, or that typical narrative impact terms in reviews of Young adult novels overlap with those in 534
 reviews of Fantasy novels. For the link between topic and impact, we struggled to come up in 535
 advance with expectations on how the topics in novels are related to impact. Novels discussing 536
 topics such as war and its consequences or living with physical or mental illness might lead to 537
 more reviews mentioning narrative impact. But honest reflection forces us to admit that the results 538
 of topic modelling are still far removed from explaining how authors deal with topics and how 539
 reviewers discuss them. This remove stubbornly persists throughout continued engagements with 540
 our data in several papers. This should give us pause to reflect on our operationalizations that are 541
 by and large still based on bags-of-words approach. Vector modelings are becoming increasingly 542
 more sophisticated. Nevertheless we have not inched significantly closer to answering the question 543
 what features of novel texts relate to what types of reader impact adequately and satisfyingly from 544
 a literary studies perspective. 545

Our reflections tie in with observations and suggestions made in some recent methodological 546
 publications on computational humanities. Bode (2023) argues that humanities researchers applying 547
 conventional methods and those that embrace computational or data-science methods should take a 548
 greater and more sincere interest in each others' work. Rather than addressing research questions by 549
 stretching either method beyond limits, researchers ought to investigate how the different methods 550
 can reinforce and amplify each other. Pichler and Reiter (2022) argue that operationalizations 551
 in computational linguistics and computational literary studies are currently often poor because 552
 we typically fail to express the precise operations that identify the theoretical concept we are 553
 trying to observe. Indeed our operationalizations seem underwhelming in the light of literary 554
 mechanisms. The reason to label a topic as being *about* war is that it contains words directly and 555
 strongly associated with war, and emphasizing the physical aspects of it, such as *war*, *soldier*, 556
bombing, *battlefield*, *wounded*, etc. But novels that readers would describe as being *about* war might 557
 instead focus on more indirect aspects or on aspects that war shares with many other situations, 558
 such as dire living conditions or being cut-off from the rest of the world, feeling unsafe and scared, 559
 or the sense of helplessness or hopelessness. And it is not just that war-related words to describe 560
 these aspects might lead an annotator to label a topic as being about something other than war. It 561
 is also that an author, going by the good practice of "show don't tell" can conjure up images that fit 562
 these words in almost infinitely many ways that are almost impossible to capture by looking at bags 563
 of words. Which means we need infinitely better operationalizations. 564

6. Data Availability 565

Data used for the research can be found at: <https://github.com/impact-and-fiction/jcls-2024-topic-genre-impact>. 566
 567

7. Software Availability

568

All code created and used in this research has been published at: <https://github.com/impact-and-fiction/jcls-2024-topic-genre-impact>.

8. Acknowledgements

571

This project has been supported through generous material and in-kind technical and data-science analytical support from the eScience Center in Amsterdam. We thank the National Library of the Netherlands for providing access to the novels used in this research and for their invaluable technical support.

9. Author Contributions

576

Marijn Koolen: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Writing – original draft

Joris J. Van Zundert: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Visualization, Writing – review & editing

Eva Viviani: Formal analysis, Software, Validation, Visualization

Carsten Schnober: Resources, Software

Willem Van Hage: Methodology, Resources, Software

Katja Tereshko: Writing – original draft, Writing – review & editing

References

587

- Abrams, M.H. (1971). *The Mirror and the Lamp: Romantic Theory and the Critical Tradition*. Oxford etc.: Oxford University Press.
- Angelov, Dimo (2020). “Top2vec: Distributed representations of topics”. In: *arXiv preprint arXiv:2008.09470*.
- Ashok, Vikas Ganjigunte, Song Feng, and Yejin Choi (2013). “Success with Style: Using Writing Style to Predict the Success of Novels”. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington: Association for Computational Linguistics, 1753–1764. <https://api.semanticscholar.org/CorpusID:7100691> (visited on 07/28/2023).
- Berger, Jonah and Grant Packard (2018). “Are Atypical Things More Popular?” In: *Psychological Science* 29.7, 1178–1184. [10.1177/0956797618759465](https://doi.org/10.1177/0956797618759465).
- Berry, David M. (2014). *Critical Theory and the Digital*. Critical Theory and Contemporary Society. New York, London, New Delhi etc.: Bloomsbury Academic.
- Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3, 993–1022.

- Bode, Katherine (2023). "What's the Matter with Computational Literary Studies?" In: *Critical Inquiry* 49.4, 507–529. 603 604
- Boot, Peter (2017). "A Database of Online Book Response and the Nature of the Literary Thriller". 605
In: *Digital Humanities*, 4. 606
- Boot, Peter and Marijn Koolen (2020). "Captivating, splendid or instructive?: Assessing the impact 607
of reading in online book reviews". In: *Scientific Study of Literature* 10.1, 35–63. [https://w 608
ww.jbe-platform.com/content/journals/10.1075/ssol.20003.boo 609](https://www.jbe-platform.com/content/journals/10.1075/ssol.20003.boo)
(visited on 01/22/2024). 610
- Burrows, John (2006). "All the way through: testing for authorship in different frequency strata". 611
In: *Literary and Linguistic Computing* 22.1, 27–47. 612
- Cranenburgh, Andreas van, K.H. van Dalen-Oskam, and Joris J. van Zundert (2019). "Vector 613
space explorations of literary language". In: *Language Resources and Evaluation* 53.4, 625–650. 614
[10.1007/s10579-018-09442-4. 615](https://doi.org/10.1007/s10579-018-09442-4)
- Culpeper, Jonathan and Jane Demmen (2015). "Keywords". In: *The Cambridge handbook of 616
English corpus linguistics*, 90–105. 617
- Darbyshire, Madison (2023). "Hot stuff: why readers fell in love with romance novels". In: *Financial 618
Times*. [https://www.ft.com/content/0001f781-4927-4780-b46c-3a9f1 619
5dffe78](https://www.ft.com/content/0001f781-4927-4780-b46c-3a9f15dffe78) (visited on 01/22/2024). 620
- Du, Keli, Julia Dudar, Cora Rok, and Christof Schöch (2021). "Zeta & eta: An exploration and 621
evaluation of two dispersion-based measures of distinctiveness". In: *Proceedings http://ceur-ws. 622
org ISSN 1613, 0073. 623*
- Du, Keli, Julia Dudar, and Christof Schöch (2022). "Evaluation of measures of distinctiveness. 624
Classification of literary texts on the basis of distinctive words". In: *Journal of Computational 625
Literary Studies* 1.1. 626
- Dunning, Ted (1994). "Accurate methods for the statistics of surprise and coincidence". In: *Com- 627
putational linguistics* 19.1, 61–74. 628
- Fialho, Olivia (2019). "What is literature for? The role of transformative reading". In: *Cogent Arts 629
& Humanities* 6.1. Ed. by Anezka Kuzmicova, 1692532. [10.1080/23311983.2019.16 630
92532. 631](https://doi.org/10.1080/23311983.2019.1692532)
- Gabrielatos, Costas (2018). "Keyness analysis". In: *Corpus approaches to discourse: A critical 632
review*, 225–258. 633
- Gitelman, Lisa, ed. (2013). *"Raw Data" Is an Oxymoron*. Cambridge: The MIT Press. 634
- Gries, Stefan Th (2008). "Dispersions and adjusted frequencies in corpora". In: *International journal 635
of corpus linguistics* 13.4, 403–437. 636
- Hickman, Miranda B. (2012). "Introduction: Rereading the New Criticism". In: *Rereading the 637
New Criticism*. Ed. by Miranda B. Hickman and John D. McIntyre. Columbus: The Ohio 638
State University Press, 1–21. <http://hdl.handle.net/1811/51698> (visited on 639
01/16/2024). 640
- Koolen, Marijn, Peter Boot, and Joris J van Zundert (2020). "Online Book Reviews and the Com- 641
putational Modelling of Reading Impact". In: *Proceedings of the Workshop on Computational 642
Humanities Research (CHR 2020)*. Vol. 2723, 0073. [http://ceur-ws.org/Vol-2723 643
/long13.pdf. 644](http://ceur-ws.org/Vol-2723/long13.pdf)
- Koolen, Marijn, Olivia Fialho, Julia Neugarten, Joris J. van Zundert, Willem van Hage, Ole 645
Mussmann, and P. Boot (2023). "How Can Online Book Reviews Validate Empirical In-depth 646
Fiction Reading Typologies?" In: *IGEL 2023 : Rhythm, Speed, Path: Spatiotemporal Experiences 647
in Narrative, Poetry, and Drama*. Monopoli: NARRNET, IGEL, elit, 1. [conference version](https://igel 648</p>
</div>
<div data-bbox=)

- society.org/events/igel2023/#submission-requirements (visited on 01/16/2024). 649 650
- Laurino Dos Santos, Henrique and Jonah Berger (2022). “The speed of stories: Semantic progression and narrative success.” In: *Journal of experimental psychology. General* 151.8, 1833–1842. [10.1037/xge0001171](https://doi.org/10.1037/xge0001171). 651 652 653
- Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila (2016). “Significance testing of word frequencies in corpora”. In: *Digital Scholarship in the Humanities* 31.2, 374–397. 654 655 656
- Loi, C., F. Hakemulder, M. Kuijpers, and G. Lauer (2023). “On how Fiction Impacts the Self-Concept: Transformative Reading Experiences and Storyworld Possible Selves”. In: *Scientific Study of Literature* 12.1, 44–67. [10.61645/ssol.181](https://doi.org/10.61645/ssol.181). 657 658 659
- Manovich, Lev (2013). *Software Takes Command*. Vol. 5. International Texts in Critical Media Aesthetics. New York, London, New Delhi etc.: Bloomsbury Academic. 660 661
- Miall, David S. and Don Kuiken (1994). “Beyond text theory: Understanding literary response”. In: *Discourse processes* 17.3, 337–352. [10.1080/01638539409544873](https://doi.org/10.1080/01638539409544873). 662 663
- Moreira, Pascale, Yuri Bizzoni, Kristoffer Nielbo, Ida Marie Lassen, and Mads Thomsen (2023). “Modeling Readers’ Appreciation of Literary Narratives Through Sentiment Arcs and Semantic Profiles”. In: *Proceedings of the The 5th Workshop on Narrative Understanding*. Toronto, Canada: Association for Computational Linguistics, 25–35. <https://aclanthology.org/2023.wnu-1.5> (visited on 01/22/2024). 664 665 666 667 668
- Nguyen, Minh Van, Viet Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen (2021). “Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. 669 670 671 672
- Paquot, Magali and Yves Bestgen (2009). “Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction”. In: *Corpora: Pragmatics and discourse*. Brill, 247–269. 673 674 675
- Pichler, Axel and Nils Reiter (2022). “From Concepts to Texts and Back: Operationalization as a Core Activity of Digital Humanities”. In: *Journal of Cultural Analytics* 7.4. 676 677
- Prescott, Andrew (2023). “Bias in Big Data, Machine Learning and AI: What Lessons for the Digital Humanities?” In: *DHQ: Digital Humanities Quarterly* 17.2, 689. <https://www.digitalhumanities.org/dhq/vol/17/2/000689/000689.html> (visited on 01/16/2024). 678 679 680 681
- Rawson, Katie and Trevor Muñoz (July 2016). *Against Cleaning*. Project blog. <http://curatingmenus.org/articles/against-cleaning/> (visited on 09/30/2016). 682 683
- Regis, Pamela (2003). *A Natural History of the Romance Novel*. Philadelphia: University of Pennsylvania Press. 684 685
- Schöch, Christof (2017). “Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama.” In: *DHQ: Digital Humanities Quarterly* 11.2. 686 687
- Sobchuk, Oleg and Artjoms Šeļa (2023). “Computational thematics: Comparing algorithms for clustering the genres of literary fiction”. In: *arXiv preprint arXiv:2305.11251*. 688 689
- Thompson, Laure and David Mimno (2018). “Authorless Topic Models: Biasing Models Away from Known Structure”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 3903–3914. <https://aclanthology.org/C18-1329> (visited on 01/22/2024). 690 691 692 693

- Toubia, Olivier, Jonah Berger, and Jehoshua Eliashberg (2021). “How quantifying the shape of stories predicts their success”. In: *Proceedings of the National Academy of Sciences* 118.26, 1–5. [10.1073/pnas.2011695118](https://doi.org/10.1073/pnas.2011695118).
- Uglanova, Inna and Evelyn Gius (2020). “The Order of Things. A Study on Topic Modelling of Literary Texts.” In: *CHR* 18-20, 2020.
- Van Zundert, Joris, Marijn Koolen, Julia Neugarten, Peter Boot, Willem Van Hage, and Ole Mussmann (2022). “What Do We Talk About When We Talk About Topic?” In: *Proceedings of the Conference on Computational Humanities Research 2022*. Antwerpen: CEUR Workshop Proceedings, 398–410. <https://ceur-ws.org/Vol-3290/> (visited on 11/22/2023).
- Van Zundert, Joris, Marijn Koolen, and Karina Van Dalen-Oskam (2018). “Predicting Prose that Sells: Issues of Open Data in a Case of Applied Machine Learning”. In: *JADH 2018 “Leveraging Open Data”: Proceedings of the 8th Conference of Japanese Association for Digital Humanities*. Tokyo: Center for Open Data in the Humanities, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, 175–177. https://conf2018.jadh.org/files/Proceedings_JADH2018_rev0911.pdf (visited on 11/07/2018).
- Warnock, John (1978). “A THEORY OF DISCOURSE, by James L. Kinneavy. (Review)”. In: *Style* 12.1. Ed. by James L. Kinneavy. Publisher: Penn State University Press, 52–54. <https://www.jstor.org.proxy.uba.uva.nl/stable/45109026> (visited on 01/16/2024).
- Wimsatt, W.K. (1954). “The Intentional Fallacy”. In: *The Verbal Icon: Studies in the Meaning of Poetry*. and two essays written in collaboration with Monroe C. Beardsley. Lexington: The University Press of Kentucky, 3–20.

NUR code	NUR label	Genre label
280	Children's Fiction general	Children's fiction
281	Children's fiction 4 - 6 years	Children's fiction
282	Children's fiction 7 - 9 years	Children's fiction
283	Children's fiction 10 - 12 years	Children's fiction
284	Children's fiction 13 - 15 years	Young adult
285	Children's fiction 15+	Young adult
300	Literary fiction general	Literary fiction
301	Literary fiction Dutch	Literary fiction
302	Literary fiction translated	Literary fiction
305	Literary thriller	Literary thriller
312	Pockets popular fiction	Literary fiction
313	Pockets suspense	Suspense
330	Suspense general	Suspense
331	Detective	Suspense
332	Thriller	Suspense
334	Fantasy	Fantasy fiction
339	True crime	Suspense
342	Historical novel (popular)	Historical fiction
343	Romanticism	Romanticism
344	Regional- and family novel	Regional fiction

Table 2: The selected NUR codes of novels in our dataset of 18,885 novels, and their mapping to genres.

A. Mapping NUR Codes to Genre Labels 717

The complete mapping from NUR codes to genre labels is shown in Table 2. 718

B. Overlap between Themes in Terms of Shared Books 719

The topic modelling process assigns each book to a single topic, but because individual topics 720 can linked to multiple themes, their books are also linked to multiple themes. As a consequence, 721 themes share books and reviews and some pairs of themes may have larger overlap than others. 722 This overlap between themes is shown for pairs of themes where for one theme, at least 25% of 723 books are shared by the other theme. 724

C. Correlations between Themes in Terms of Impact 725

The correlations between themes in terms of the percent difference (%Diff) per impact term for 726 generic *Affect*, *Narrative* and *Aesthetics* is shown respectively in Figures 10, 11 and 12. 727

Theme 1	Share 1	Theme 2	Share 2	Book overlap	Books theme 1	Books theme 2
crime	0.33	geo. & setting	0.14	619	1899	4317
culture	0.49	geo. & setting	0.40	1713	3524	4317
econ. & work	0.36	behav./feelings	0.12	446	1232	3860
econ. & work	0.30	society	0.44	371	1232	851
econ. & work	0.25	politics	0.49	310	1232	634
family	0.65	behav./feelings	0.08	324	498	3860
family	0.30	culture	0.04	151	498	3524
geo. & setting	0.40	culture	0.49	1713	4317	3524
history	0.51	geo. & setting	0.24	1038	2020	4317
history	0.31	war	0.65	622	2020	952
lifest. & sport	0.31	medi./health	0.20	216	702	1058
politics	0.49	econ. & work	0.25	310	634	1232
politics	0.49	society	0.36	310	634	851
society	0.44	econ. & work	0.30	371	851	1232
society	0.36	politics	0.49	310	851	634
war	0.65	history	0.31	622	952	2020

Table 3: Overlap in books between themes, for themes where one theme shares at least 25% of books with the other theme.

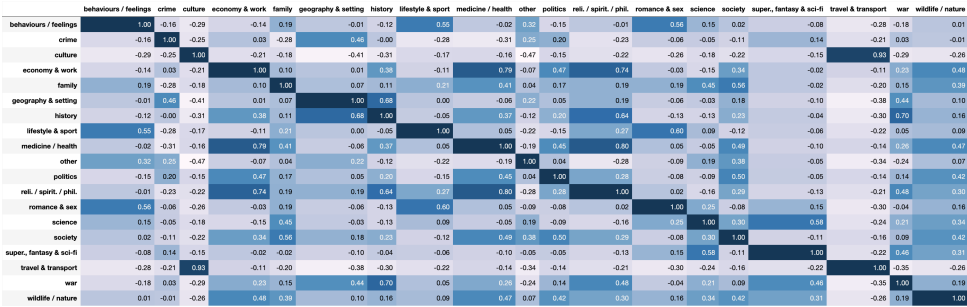


Figure 10: Percent different correlations between themes based on general *Affect* terms.

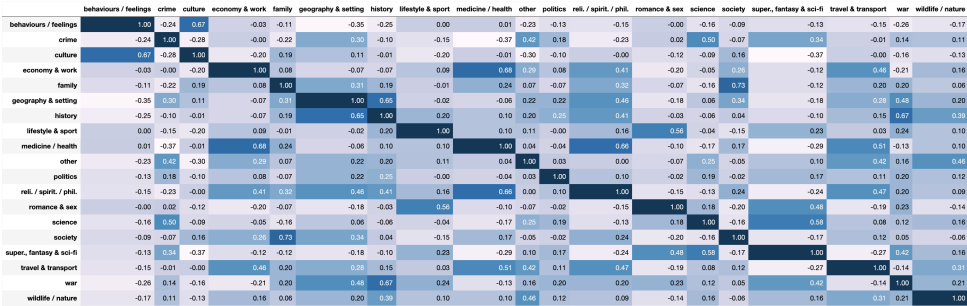


Figure 11: Percent different correlations between themes based on *Narrative* impact terms.





Figure 12: Percent different correlations between themes based on general *Aesthetic* impact terms.

conference version

BookNLP-fr, the French Versant of BookNLP A Tailored Pipeline for 19th and 20th Century French Literature

Frédérique Mélanie-Becquet¹ 
Jean Barré¹ 
Olga Seminck¹ 
Clément Plancq² 
Marco Naguib³ 
Martial Pastor⁴ 
Thierry Poibeau¹ 

1. Lattice UMR 8094, École Normale Supérieure - PSL - CNRS - Université Sorbonne Nouvelle , Montrouge, France.
2. MSH Val de Loire UAR 3501, CNRS - Université de Tours - Université d'Orléans , Tours, France.
3. LISN, Université Paris-Saclay and CNRS , Orsay, France.
4. Centre for Language Studies, Radboud University , Nijmegen, The Netherlands.

Citation

tba (2024). "BookNLP-fr, the French Versant of BookNLP. A Tailored Pipeline for 19th and 20th Century French Literature". In: *CCLS2024 Conference Preprints 3* (1). [10.26083/tuprints-00027396](https://doi.org/10.26083/tuprints-00027396)

Date published 2024-05-28

Date accepted 2024-04-15

Date received 2024-01-25

Keywords

Natural Language Processing, Computational Literary Studies, French Literature, Coreference Resolution, Entity Recognition, Subgenre Classification

License

CC BY 4.0 

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 3rd Annual Conference of Computational Literary Studies at Vienna, Austria, in June 2024.

Abstract. This paper presents BookNLP-fr: the adaptation to French of BookNLP, an existing NLP pipeline tailored for literary texts in English. We provide an overview of the challenges involved in the adaptation of such a pipeline to a new language: from the challenges related to data annotation up to the development of specialized modules of entity recognition and coreference. Moving beyond the technical aspects, we explore practical applications of BookNLP-fr with a canonical task for computational literary studies: subgenre classification. We show that BookNLP-fr provides more relevant and – even more importantly – more interpretable features to perform automatic subgenre classification than the traditional *bag-of-words* approach. BookNLP-fr makes NLP techniques available to a larger public and constitutes a new toolkit to process large numbers of digitized books in French. This allows the field to gain a deeper literary understanding through the practice of distant reading.

1. Introduction

The domain known as Computational Humanities has recently emerged, with the availability of large corpora of literary texts in digitized format, and of transformer-based language models that are quick, robust and (generally) accurate (Devlin et al. 2019; Touvron et al. 2023, e.g.). This situation opened up new opportunities for exploration and analysis. For French, the collection *Literary fictions of Gallica* (Langlais 2021) includes 19,240 public domain documents from the digital platform of the French National Library, enabling researchers to navigate the wide diversity of literature with unprecedented ease.

The sheer volume of digitized texts presents a unique set of challenges. Traditional methods of literary analysis and interpretation are insufficient when confronted with such vast corpora. It is no longer feasible for individuals to manually analyze in close

reading the entirety of these collections. This shift in scale necessitates the development of innovative tools and technologies, particularly Natural Language Processing (NLP). These tools are essential for extracting meaningful insights from digital corpora. They can illuminate patterns, trends, and connections that would be impractical or impossible for humans to discern within the vast amount of text data. This new technical paradigm opens up the possibility of conducting research through distant reading (Moretti 2000; Underwood 2019), enabling scholars to zoom in and out from the literary past, facilitating a more profound comprehension of trends and patterns that delineate the evolution of literature. The knowledge embedded in these digitized literary corpora is crucial not only for literary scholars but also for those interested in cultural analytics, defined as “the analysis of massive cultural datasets and flows using computational and visualization techniques” by Manovich (2018), or more practical applications for example the automatic production of book summaries for catalogs (Zhang et al. 2019). The evolution of literature is intricately tied to the broader shifts in society, and digitized texts offer a unique opportunity to study these transformations.

To make the analysis of such large corpora possible, BookNLP (Bamman 2021) has been proposed as a specialized software solution adapted to literary texts. It includes the analysis of entities, coreference, events, and quotations within textual data. Originally conceived at the University of California, Berkeley in 2014 by David Bamman and his team, BookNLP has undergone continuous enhancements, aligning with the latest advancements in natural language processing. Notably, it has embraced emerging technologies such as integrated embeddings of large language models, more specifically BERT (Devlin et al. 2019) in early 2020.

The ongoing evolution of BookNLP extends beyond its initial scope, as efforts are underway to expand its applicability to five additional languages through the Multilingual BookNLP project (Bamman 2020). However, it’s worth noting that French is not included in this extension. In response to this gap, it was decided in 2021, in coordination with Berkeley, to develop a dedicated French version of BookNLP. The goal is that researchers working with French literary data have access to basic tools required for the structured analysis of fiction. This paper thus presents the French BookNLP project, the related annotated corpus and the pieces of software defined within the project, as well as a specific study illustrating how BookNLP can be used for literary studies.

The structure of the paper is as follows: we start with a literature review in which we specify NLP tools and techniques that are of particular interest in a framework for distant reading (section 2). Special attention will be given to results of the English BookNLP project (subsection 2.2). In section 3, we provide a detailed description of how we elaborated the pipeline of BookNLP-fr: the training data, the annotation process and the software development. In Section 4, we give the evaluation scores of our pipeline on the subtasks of *entity recognition* and coreference resolution. Then, we will present a case study where we used BookNLP-fr for the classification of literary genre (section 5). We finish this article with a discussion about how the use of computational methods and the framework of distant reading using imperfect annotations affects the field of literary studies (subsection 6.1) and its perspectives in the era of *Large Language Models* (subsection 6.2) and finally summarize the paper in the conclusion (section 7).

2. Literature Review

2.1 Computational Methods Applied to Literary Text Analysis

Statistical methods have been used extensively in literary text analysis to identify patterns and trends in large amounts of textual data. Different pieces of software are available for this, for example: Quanteda (Benoit et al. 2018), stylo (Eder et al. 2016), TidyText (Silge and Robinson 2017) or Voyant tools (Rockwell and Sinclair 2016), to cite the most famous. They are available “off the shelf”, which means that they can be used directly by scholars and researchers to analyze texts. These tools can handle raw text directly, or after basic NLP-processes such as lemmatization, part-of-speech-tagging, or other kinds of annotations. They offer various visualizations to interpret the texts, such as dendrograms to represent the ‘distance’ between various books of a corpus or charts that make it visible what type of vocabulary is typical to one author as opposed to another one.

There are clear benefits in using statistical methods to analyze literary texts, such as the ability to process and analyze large amounts of data quickly and efficiently, to identify patterns and trends that might not be apparent through traditional close reading methods, and to generate new research questions and hypotheses. But NLP is needed to better represent the content of the text, i.e. what the text says behind the words used. Natural language processing techniques can be used to annotate literary texts by providing syntactic and semantic annotations. NLP has become an increasingly important tool in the field of literary studies, providing new methods for analyzing and interpreting literary texts. NLP tools (e.g. NLTK (Bird et al. 2019) or Stanford tools (Manning et al. 2014)) have been used to perform a wide range of tasks, including part-of-speech tagging, syntactic analysis, named entity recognition, etc. In the following paragraphs, we will specify the linguistic analyses available by the BookNLP pipeline: entity recognition, coreference resolution, event recognition and quotation detection. The tools mentioned in the paragraph above do not propose these type of semantic analyses, and only use morphological and grammatical linguistic analyses. BookNLP thus occupies a special niche and provides more semantically-oriented annotations.

Entity Recognition. Entity recognition, along with coreference resolution, is of prominent importance, since it makes it possible to track characters, their actions and their relationships over time. Named entity recognition is a well-established task in NLP, referring to the recognition of persons, locations, companies and other institutions, etc. (Maynard et al. 2017) and systems exist for a wide array of languages (Emelyanov and Artemova 2019), with generally good performance, depending of course on the nature of the document to be analyzed and of the gap between training data and target data. Recognizing mentions referring to characters in a novel shares many features with named entity recognition, but is more varied (not all characters have a name, and a character can correspond to an animal, for example). Locations are also of the utmost importance to track the movements of characters (Ryan et al. 2016), but also to detect events. Note that performance may vary greatly depending on the nature of the novel and of the entities to be recognized, for example in the novel *Les Mystères de Paris* written between 1842 and 1843 by Eugène Sue, most characters have names that are similar to noun phrases, such as ‘la Goualeuse’ (meaning *the Street Singer*) or ‘le Chourineur’

(meaning *the Stabber*). Also science fiction, which is full of non-classical proper nouns, can be very challenging for the task (Dekker et al. 2019). A module able to predict, or at least, estimate performance from cues gathered in the text would be useful to process large collections of novels.

Coreference (especially linking together all the mentions in the text of a given character, although the task can involve all kinds of names, or even nouns) is challenging in nature. There is a long tradition of research in coreference resolution in NLP, and modules exist for different languages, with various levels of performance (Poesio et al. 2023). The quality of the different systems is still increasing (through end-to-end models (Lee et al. 2017) and then transformer-based language models (Joshi et al. 2019)), and coreference remains a very active field of research in NLP. The task is more challenging for French or Russian than for English, since the “it” pronoun limits ambiguity in English (whereas all nouns are masculine or feminine in French, not only human beings and are referred to with third person pronouns, as for instance in “*Marie veut qu’on lave la voiture, elle est sale.*” (“*Marie wants that we wash the car, it is dirty.*”), where *elle* refers to *the car*, but could theoretically also refer to *Marie*; there is no ambiguity from a human point of view in this sentence, but the analysis requires semantic information). When applied in literary studies, automatic coreference systems often break long coreference chains due to the fact that they use a fixed-sized sliding window. If a given character does not appear during a certain period of time (i.e. a certain number of pages), it makes it harder to retrieve its antecedent. Literature provides a good test bed for the coreference task, since novels are long, real, and complex texts on which performance can (and should) still improve a lot.

Event Recognition. Event recognition involves the automated identification and extraction of verbs and, more rarely, nouns referring to events. The task is difficult in that there is no clear definition of what an event is, and other features interact with the definition (among others: negation, adverbials and modals), and not all occurrences of verbs should be annotated (e.g. in “*I like to play tennis*”, *play* is an infinitive that refers to something I like, but it is generally considered that there is no event *per se* in the sentence). As for literary texts, there have been initiatives to annotate events (Sims et al. 2019), but most verbs and even some nouns can refer to events (Hogenboom et al. 2016; Sprugnoli and Tonelli 2016), which may lead to a too fine-grained annotation. There is thus a need to redefine the task and provide an intermediate level of annotation, between isolated events and the novel as a whole (Lotman 1977; Schmid 2010a,b), but higher level annotation (like the notion of scene) has also proven difficult to formalize, leading to very low accuracy in practical experiments (Zehe et al. 2021).

Quotation Recognition. Quotation recognition plays a crucial role in enhancing the understanding of textual content by identifying and isolating direct speech instances. This feature is instrumental in extracting and preserving the spoken words of characters, enabling a fine-grained analysis of dialogue patterns and character interactions (Durand et al. 2023; Van Cranenburgh and Van Den Berg 2023). A crucial but complex part of the task consists in establishing what character is at the origin of a given utterance. A recent study has shown that performance on this task are still rather low and would need to improve to be really usable in operational contexts (Vishnubhotla et al. 2023).

2.2 The BookNLP Project

145

BookNLP is a set of natural language processing modules designed specifically for the analysis of novels and other literary texts. Developed by D. Bamman (Bamman 2021; Bamman et al. 2014) and colleagues at the University of Berkeley, BookNLP employs a combination of machine learning and linguistic analysis techniques to extract information from text and perform tasks such as character recognition, coreference resolution, event recognition, and quotation extraction. Note that the Berkeley BookNLP suite currently is based upon BERT (Devlin et al. 2019, e.g.), but this could evolve as better language models continue to appear.

146
147
148
149
150
151
152
153

The annotated files that are available for training constitute the LitBank corpus (Bamman et al. 2020, 2019). This corpus is publicly available (see <https://paperswithcode.com/dataset/litbank>), which makes it possible to regularly retrain the system, as NLP continues to evolve rapidly (especially large language models)

154
155
156
157

Entity Recognition: One of the primary tasks of BookNLP is entity recognition, more specifically characters, locations and vehicles, showing the focus on the actions of characters. This information is used to study how mobile protagonist characters are and what kind of space male and female characters occupy (Soni et al. 2023). Character recognition is often coupled with other information (gender, attributes, relations between characters), that can be useful for sub-stream tasks.

158
159
160
161
162
163

Coreference Resolution: In the context of literature, coreference resolution often involves resolving pronouns and other referring expressions to specific characters or entities. BookNLP employs advanced linguistic analysis to identify and link references to the same entity, and the extra knowledge provided by large language models is especially useful for the task.

164
165
166
167
168

Event Recognition: Event recognition is another essential task performed by BookNLP. It should be crucial for analyzing the development of the storyline and identifying key plot points, but the huge number of verbs supporting actions make the annotation too prolific and not adapted to specific needs. The proper annotation of negation, adverb and modals is also an open problem. This is why event recognition has not been addressed as a priority in the context of the Multilingual BookNLP Project, that rather focus on entity recognition and coreference resolution.

169
170
171
172
173
174
175

Quotation Extraction: BookNLP is equipped with the capability to extract quotations from a text. This involves identifying and isolating the direct speech or quoted passages within the literary work. Accurate quotation extraction is vital for understanding character dialogue, the intentions of characters and develop further analyses. However, quotation recognition without speaker attribution is not so useful and, as we have seen before, speaker attribution remains an open question, as accuracy for the task remains low (Vishnubhotla et al. 2023).

176
177
178
179
180
181
182

The application of BookNLP for the analysis of novels and other literary works aims at providing a deeper understanding of narrative structures, character dynamics, and thematic elements in novels (Piper et al. 2021). The different modules are intended to assist researchers in literary analysis but also in digital humanities and cultural analytics.

183
184
185
186

3. French BookNLP

187

The French BookNLP project endeavors to construct a robust Natural Language Processing (NLP) pipeline specifically tailored for the comprehensive analysis of extensive French literary corpora of the 19th and 20th century. The ongoing MultiLingual BookNLP project (Bamman 2020), coordinated by Berkeley, seeks to update the initial pipeline (Bamman et al. 2014) and extend its capabilities to encompass four additional languages (Spanish, German, Russian and Japanese). In alignment with this initiative – even though we are not part of the Multilingual BookNLP project in itself, in the sense that we are independent from the research grant that the Berkeley’s team obtained – we are actively engaged in the development of the necessary linguistic resources for the French language. Our collaborative efforts with the Berkeley project ensure a coordinated approach to this expansion, by sharing similar annotations and visualization tools, for example.

In line with the Multilingual BookNLP Project, we will mainly focus on entity recognition and coreference resolution. We have seen in the previous sections that annotating events entail a number of problems and may be too general, thus not be useful if it is not done with a specific goal in mind (which may entail some domain-specific annotations, with adapted categories, for example). We have also seen that quotation recognition with no proper speaker attribution algorithm is, for similar reasons, not really useful, but that speaker attribution remains an open problem (Zehe et al. 2021). In what follows, we will thus not address these two tasks (event and quotation recognition) for further investigation and concentrate on entity recognition and coreference resolution.

3.1 The Training Corpus and The Democrat Project

209

The “Democrat” project, led by Frédéric Landragin (2016; 2021) and funded by the French National Research Agency (ANR), aimed to develop an annotated corpus at the level of coreference chains in French. Before the Democrat project, no corpus of this kind existed. The project concluded in 2020.

One of the fundamental aspects of Democrat was the annotation of long texts, in contrast to the Ontonotes corpus (Weischedel et al. 2013) for example, which serves as a standard for English but is predominantly composed of short texts. Additionally, the Democrat project aimed to annotate a wide variety of text types, including chapters from novels, short stories, journalistic pieces, legal documents, encyclopedic entries, technical texts, and more. It also had a diachronic dimension, spanning from medieval French to contemporary French.

For the needs of the BookNLP-fr project, we focused on annotations related to novels and selected the texts spanning from the early 19th century to the early 20th century. Before this period, French is more prone to variation, and for the more recent period, texts are not freely shareable due to copyright issues. Lastly, to keep the annotation task manageable, each text in the Democrat corpus is actually composed of a 10,000-word excerpt (leaving us with 184,137 tokens). In addition to this selection from Democrat, we added two short stories from Balzac, good for 45,238 tokens. Information about these texts and those from Democrat can be found in Table 1.

Year	Author	Title	Source
1830	Honoré de Balzac	La maison du chat qui pelote	Full Text
1830	Honoré de Balzac	Sarrasine	Democrat 10 K
1836	Théophile Gautier	La morte amoureuse	Democrat 10 K
1837	Honoré de Balzac	La maison Nucingen	Full Text
1841	George Sand	Pauline	Democrat 10 K
1856	Victor Cousin	Madame de Hautefort	Democrat 10 K
1863	Théophile Gautier	Le capitaine Fracasse	Democrat 10 K
1873	Émile Zola	Le ventre de Paris	Democrat 10 K
1881	Gustave Flaubert	Bouvard et Pécuchet	Democrat 10 K
1882-1883	Guy de Maupassant	Mademoiselle Fifi, nouveaux contes (1)	Democrat 10 K
1882-1883	Guy de Maupassant	Mademoiselle Fifi, nouveaux contes (2)	Democrat 10 K
1882-1883	Guy de Maupassant	Mademoiselle Fifi, nouveaux contes (3)	Democrat 10 K
1901	Lucie Achard	Rosalie de Constant, sa famille et ses amis	Democrat 10 K
1903	Laure Conan	Élisabeth Seton	Democrat 10 K
1904-1912	Romain Rolland	Jean-Christophe (1)	Democrat 10 K
1904-1912	Romain Rolland	Jean-Christophe (2)	Democrat 10 K
1917	Adèle Bourgeois	Némoville	Democrat 10 K
1923	Raymond Radiguet	Le diable au corps	Democrat 10 K
1926	Marguerite Audoux	De la ville au moulin	Democrat 10 K
1937	Marguerite Audoux	Douce Lumière	Democrat 10 K

Table 1: The texts in the BookNLP-fr corpus.

3.2 Data Preparation and Annotation

229

Entities	#Occurrences
PER - Mentions	32,338
PER - Chain	3,006
FAC	2,325
TIME	1,836
LOC	1,040
GPE	928
VEH	475
ORG	205
TOTAL	39,147

Table 2: The number of occurrences per type of entity.

In the scope of the Democrat project, annotations have been applied to all types of coreference. However, for the BookNLP-fr project, our specific focus lies within a subset of these coreferences, corresponding to certain types of entities: persons, facilities, locations, geo-political entities, vehicles, organizations and denotations of time. Definitions from all these categories except for time are adapted from Bamman et al. (2019).

PER: According to Bamman et al. (2019): “By person we describe a single person indicated by a proper name (*Tom Sawyer*) or common entity (*the boy*); or set of people, such as *her daughters* and *the Ashburnhams*.”. Some examples from our corpus in (1), and (2):

- (1) a. une de ces gentilhommières si communes en Gascogne, et que **les villageois** déco-
rent du nom de château Le Capitaine Fracasse
- b. one of those manors so common in Gascogne, and that **the villagers** deco-
rated by the name of the castle of Captain Fracasse

- (2) a. **Madame François**, adossée à une planchette contre **ses** légumes 242
 b. **Madame François**, who leaning on a board next to **her** vegetables 243

Note that PER mentions are split into three parts to enable more fine-grained analyses, 244
 including proper nouns (PROP), common phrases (NOM), and pronouns (PRON). 245
 Pronouns account for the majority of mentions, specifically 59%, 32%, and 9%, respec- 246
 tively. 247

FAC: We follow Bamman’s (2019) definition: “For our purposes, a facility is defined as a 248
 “functional, primarily man-made structure” designed for human habitation (buildings, muse- 249
 ums), storage (barns, parking garages), transportation infrastructure (streets, highways), and 250
 maintained outdoor spaces (gardens). We treat rooms and closets within a house as the smallest 251
 possible facility.”, see example (3): 252

- (3) a. **Le chemin** qui menait de **la route** à **l’habitation** s’était réduit, par l’en- 253
 vahissement de la mousse et des végétations parasites 254
 b. **The path** that led to **the road** to **the dwelling** was narrowed by the invasion 255
 of moss and parasitic vegetation 256

GPE: We followed Berkeley’s guidelines for this category: “Geo-political entities are single 257
 units that contain a population, government, physical location, and political boundaries.”, see 258
 example (4): 259

- (4) a. Échappé de **Cayenne**, où les journées de décembre l’avaient jeté, rôdant 260
 depuis deux ans dans **la Guyane hollandaise**, avec l’envie folle du retour et 261
 la peur de la police impériale, il avait enfin devant lui **la chère grande ville**, 262
 tant regrettée, tant désirée. 263
 b. Escaped from **Cayenne**, where the December days had thrown him, erring 264
 since two years in **Dutch Guyane**, with a crazy desire of returning and fear 265
 of the imperial police, he finally had before him the **dear big city**, so much 266
 regretted and desired. 267

LOC: As opposed to GPEs, locations are “entities with physicality but without political 268
 organization [...] such as **the sea, the river, the country, the valley, the woods, and the** 269
forest” (Bamman et al. 2019). Two examples from our corpus: 270

- (5) a. des moellons effrités aux pernicieuses influences de **la lune** 271
 b. crumbling rubble masonry under the pernicious influences of **the moon** 272
 (6) a. Poussez-moi ça dans **le ruisseau** ! 273
 b. Push this into **the stream** ! 274

VEH: The definition for a vehicle is a “physical device primarily designed to move an object 275
 from one location to another” (Bamman et al. 2019). An example from our corpus: 276

- (7) a. anciennement **des voitures** avaient passé par là 277

b. before, **carriages** had passed there 278

ORG: "Organizations are defined by the criterion of formal association" (Bamman et al. 2019), 279
for example the church and the army. An example from our corpus: 280

(8) a. et la peur de **la police impériale** 281
b. and fear of **the imperial police** 282

TIME: This category is absent in the annotations of Bamman et al. (2019). We designed 283
it to annotate temporal information, duration indications and moments of the day (*day*, 284
night, *morning*). 285

(9) a. sous **le règne de Louis Xiii**, 286
b. under **the reign of Louis Xiii**, 287

(10) a. **Le soir**, il avait mangé un lapin. 288
b. **At night**, he had eaten a rabbit. 289

As part of the refinement process, the initial annotations required thorough revision 290
and cleaning. We had multiple team discussions about many borderline cases, such as 291
whether Gods and Greek heroes should be annotated as characters, the status of speak- 292
ing animals and the exact distinction between GPE, FAC and LOC. We meticulously 293
documented every choice made during the annotation process. This documentation is 294
publicly available in an annotation guide¹, providing a valuable resource for understand- 295
ing our decisions and methodologies in characterizing entities within the context of the 296
BookNLP project, based on the initial ground provided by the Democrat project. Once 297
the annotation guidelines were finished, the entire corpus was annotated by freshly 298
trained annotators. Their first annotations (comprising 315 tags) produced during 299
their training phase, featured an inter-annotator agreement score of Cohen's kappa 300
= .38, meaning fair and almost moderate agreement (Cohen 1960) but showing that 301
this is no trivial task. With better trained annotators, values between .76 and .75 were 302
reached, which constitutes a reasonable basis for further training models. Most errors 303
were due to forgotten mentions, and uncertainties about difficult cases (plurals, fuzzy 304
expressions, non referential entities). Another look at the annotated files by another 305
trained annotators makes a huge difference so as to get a better and more homogeneous 306
coverage (esp. concerning forgotten entities during the initial annotation stage). 307

After annotation, to facilitate seamless integration with the BookNLP software, the 308
annotations were transformed into a compatible format. We annotated the entity types 309
in TXM (Heiden 2010) because the Democrat corpus is distributed in this format, 310
and later migrated our annotations to brat (Stenetorp et al. 2012), the format used by 311
Berkeley's team. The number of entities in each categorie can be found in Table Table 2. 312

1. See <https://github.com/lattice-8094/fr-litbank>.

3.3 Software Development 313

Large language models play now a prominent role in contemporary natural language processing. Our implementation of BookNLP-fr is built upon the software from the Multi-lingual BookNLP-project. For the two tasks that we perform (entity recognition and coreference resolution), two separated models are developed. Entity recognition is performed before coreference resolution.

Detecting the literary entities, a BiLSTM-CRF model (Bamman et al. 2020; Ju et al. 2018) is fed with contextual embeddings from the CamemBERT model (Martin et al. 2020), which is a BERT (Devlin et al. 2019) based architecture tailored for French.

For the coreference part, a BiLSTM is also fed with the embeddings from CamemBERT. Then, following (Bamman et al. 2020), who in their turn are following Lee (Lee et al. 2017), the BiLSTM architecture is attached to a feedforward network in which the probability of two mentions (detected entities) are coreferent with each other is evaluated. Mentions are linked to their highest scoring antecedent (a null-antecedent is always an option) and coreference chains are defined as the transitive closure of links.

For each model, we split the corpus into training (80%), development (10%) and test (10%) corpus, please see Section section 4 for the results.

While event annotation remains a focal point, challenges persist, primarily due to limitations in performance and the inherently ambiguous nature of defining events. The elusive nature of the concept makes it challenging to generate consistently relevant and usable results. As for quotation identification, we acknowledge the need to integrate speaker recognition for a more comprehensive understanding of textual nuances.

Given these considerations, we have more specifically directed our efforts toward optimizing modules for entity recognition and coreference resolution. This focus allows us to refine and train models that are specifically accurate in identifying and linking entities within a given text, contributing to the effectiveness of BookNLP-fr for downstream tasks (like subgenre classification, see section 5).

4. Results and Evaluation 340

In this section we give the results of our BookNLP-fr modules for entity recognition and coreference resolution on literary texts.

4.1 Named Entity Recognition Evaluation 343

Table 3 reports our results for entity recognition, measured traditionally through precision (the percentage of entities correctly recognized among those recognized) and recall (the percentage of entities correctly recognized among those to be recognized). Please note that ORG is absent from this evaluation, because due to an uneven distribution of this tag in different texts, it was only present 7 times in the test corpus, making estimation of precision and recall unreliable.

When assessing the model's performance, a higher precision relative to recall suggests that the model is more likely to make accurate predictions when identifying literary

	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
GPE	68.2	52.9	59.6

Table 3: Entity recognition evaluation of BookNLP-fr on literary texts.

entities. Precision denotes the percentage of correctly predicted literary entities among all entities predicted by the model. High precision is advantageous, ensuring that the identified literary entities are more likely to be accurate, albeit at the potential cost of missing some relevant entities (lower recall). Prioritizing precision in this context aids in minimizing false positives, thereby enhancing the reliability of the identified literary entities. It is important to highlight that literary entities differ from typical Named Entities in Natural Language Processing (NLP), displaying a much larger range of possibilities. Consequently, the obtained results, though seemingly divergent from NLP standards, represent a pioneering achievement in the analysis of French fiction, as this is the first study of its kind.

Some scores may appear modest in comparison to the state-of-the-art, particularly regarding the recall for TIME expressions. This is due to the extensive diversity of time expressions in our corpus, which is far more varied than in the traditional news corpora typically used in NLP, coupled with the limited number of examples in the training corpus (see below, Table 4 for a comparison with a state-of-the-art system). Nevertheless, we have opted to report these scores for the sake of comprehensiveness. In the near future, we will strive to expand the coverage of our system, aiming to achieve improved recall across various categories beyond PER.

As a baseline, we ran the CamemBERT-NER model², which is a NER model that was fine-tuned from camemBERT on wikiner-fr dataset. Table 4 shows baseline performance in comparison with BookNLP-fr. Results are showing that BookNLP-fr is as good as the fine-tuned model for proper name recognition, but it captures much more by including pronouns and common nouns, which the baseline does not handle at all. The F1 score for the detection of PROP/NOM/PRON mentions reaches 83.13, which is in line with the English BookNLP (88.3).

pos_tag	BookNLP-fr			Camembert-NER		
	precision	recall	F1 Score	precision	recall	F1 Score
PROP	82.5	79.2	80.8	91.85	72.05	80.75
NOM	74.9	74.7	74.8	96.32	14.17	24.70
PRON	86.3	89.5	87.9	100.00	0.10	0.20
ALL	82.39	83.88	83.13	92.58	7.92	14.59

Table 4: Comparison on litbank-fr for PER recognition performance between BookNLP-fr and Camembert-NER.

2. See <https://huggingface.co/Jean-Baptiste/camembert-ner>.

BookNLP-fr thus demonstrates its robustness for the classic task of proper name recognition, but the real value of our model lies in its ability to go beyond this to capture the full spectrum of what constitutes a character in novels. This aligns with Woloch (2003) concept of the character space as “the encounter between an individual human personality and a determined space and position within the narrative as a whole,” allowing for the automatic detection and analysis of the distribution of character mentions throughout the narrative (Barré et al. 2023).

4.2 Coreference Resolution Evaluation

Table 5 presents the evaluation metrics for coreference resolution using BookNLP-fr on our test corpus. Three key metrics, namely MUC , B^3 , and $CEAF_e$, are employed to assess its performance. As coreference chains are complex to modelize, different evaluation metrics are necessary to get a global image of systems performance. We refer to Luo and Pradhan (2016) for a comprehensible explanation of these metrics.

Our average F1 score, calculated as the mean of the three metrics, is presented as 76.4. The reported scores suggest a commendable performance, but the practical utility in the context of literary analysis should be further explored based on the specific goals of the research or application. Note that the English BookNLP yields 79.3 in performance for the same task.

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

Table 5: Coreference resolution evaluation of Fr-BookNLP on literary texts

The challenge of duplication arises when the model detects the same character multiple times within the analyzed text. In some instances, among the top five literary entities identified by the model, there may be cases where two or more main characters share the same name or attributes. While this duplication might raise concerns initially, for example, if one aims to study character networks (Perri et al. 2022) or the overall number of characters in novels, it may not pose a significant issue when the focus is on character characterization. For example, in studies about the representation of male and female characters, the output of BookNLP has been shown to be very useful (e.g. Gong et al. 2022; Hudspeth et al. n.d.; Naguib et al. 2022; Toro Isaza et al. 2023; Underwood et al. 2018; Vianne et al. 2023; Zundert et al. 2023).

Also in the following case study, the primary objective is not to pinpoint unique and distinct characters but rather to establish a proxy for characterization as a whole. Our goal is to capture the prevalence and significance of certain characters across various texts and literary works. Hence, the emphasis lies more on character representation and the overall impact of these characters on the literary landscape, rather than on identifying entirely separate and non-repeating characters.

5. Case Study: Genre Classification Using Booknlp-fr Features 411

5.1 Introduction 413

This case study aims to demonstrate that BookNLP-fr can be of significant assistance in the realm of computational literary studies (CLS). We illustrate this assertion through a canonical issue in CLS: the automatic detection of literary genres. Historically, the division of novels into specific sub-genres has been a classification practice employed by literary stakeholders such as librarians, editors, and critics. This practice is partly justified by a specific textual component that relates to the spatiotemporal framework, characters, themes, or narrative progression.

Genre is a central concept in poetics, defined successively from Aristotle to structuralists, through romantics and Russian formalists (Aristote 1990; Bakhtin 2006; Genette 1986; Schlegel et al. 1996). From our computational standpoint, structuralists have offered intriguing definitions. For example, Schaeffer (1989) defines genericity as an “internalized norm that motivates the transition from a class of texts to an individual text conforming to certain traits of that class”. There could be a set of textual procedures internal to works, and the mission of CLS would be to find the best ways to account for this fact. However, the norms or formal rules of sub-genres cannot be solely boiled down to formal or thematic rules. For instance, the sociological approach, as exemplified by Bourdieu (1979), tends to focus more on the “community of readers” with the study of power dynamics and accompanying aesthetic hierarchies. However, these norms do indeed exist, as they enable a work to align itself with the established and shared usage of a “horizon of expectations” (Jauß 1982) of the audience which might induce the authors to adhere to certain expected norms and styles.

Various studies have devised strategies to automatically identify subgenres. Selected studies have employed methods such as the bag of words (BoW) (Hettinger et al. 2016; Underwood 2019) or topic modeling (Schöch 2017; Zundert et al. 2022) to find subgenre similarities between texts. In addition to these basic features, researchers utilize machine learning techniques in a supervised setting, employing methods such as logistic regression or support vector machines when ground truth is available. However, the challenge often arises from the potential incompleteness or temporal bias of these ground truths. Unsupervised learning approaches and clustering methods have also enabled the exploration of hybrid texts that belong to multiple subgenres, as demonstrated by studies like (Calvo Tello 2021; Sobchuk and Šeja 2023). In our case-study, we will rely on a corpus with predefined labels, while acknowledging the idea that sub-genres are not monolithic categories. Thus, the objective is not so much to demonstrate the validity of sub-genre labels, which are often incomplete or limiting in reality, but rather to show that the interpretability of errors in automatic classification can lead us to a more nuanced and comprehensive understanding of the subgenre phenomenon.

Despite recent advancements in NLP, the bag-of-words approach remains largely unchanged. This is because many tools, including document embeddings, are not easily interpretable and are optimized for short texts. In this context, we present in the next

section a method that aims to find a balance between the use of state-of-the-art methods 453
for literary text processing and their interpretability. 454

5.2 Method 455

5.2.1 Corpus and Subgenre Labels 456

Our case study is built upon one of the largest corpora for fiction in French: the “corpus 457
Chapitres”, a corpus of nearly 3000 French novels (Leblond 2022). The period concerned 458
extends over two centuries of novel production, from the 19th to the 20th century, as 459
can be seen in Figure 1. 460

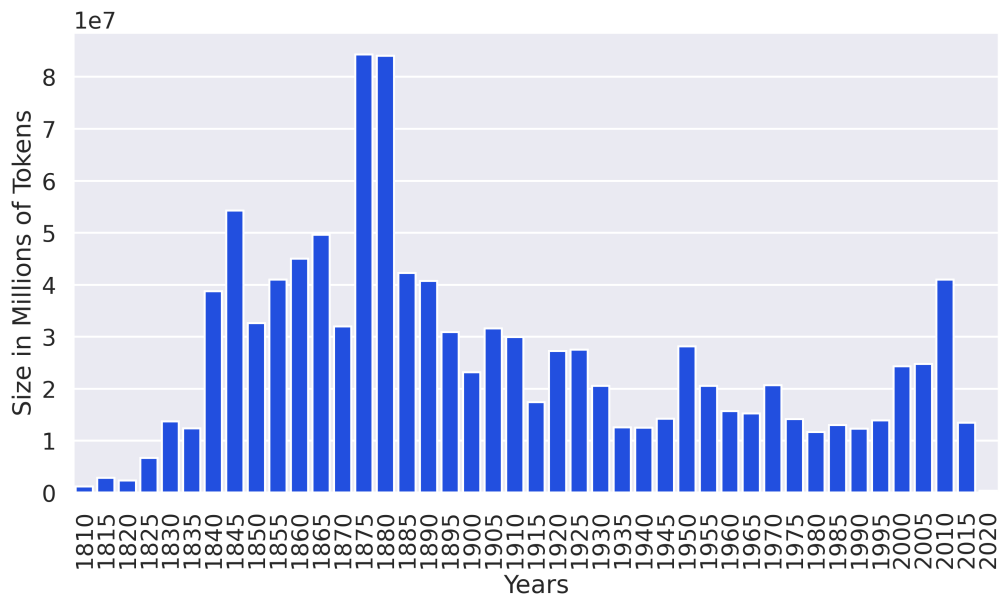


Figure 1: Distribution of the number of tokens over time.

Approximately two-thirds of Chapitres is annotated with sub-genre labels. This an- 461
notation is based on the classification of the French National Library. We choose to 462
concentrate our analysis on the five most prevalent sub-genres within the corpus: adven- 463
ture novels, romance, detective fiction, youth literature, and memoirs. The validity of 464
these labels is not clearly established, as the practices of the BNF for assigning these la- 465
bels have not been systematized nor standardized. Therefore, there is no “Ground Truth” 466
per se, but our supervised approach described in subsection 5.2.3 aims precisely to 467
understand the boundaries of subgenres. 468

5.2.2 Textual Features 469

The BoW method stands out as the default feature extraction technique, as it allows 470
scholars to have an easy task to implement without requiring intensive computational 471
resources (GPU, RAM). Underwood (2019) demonstrated that the BoW approach was 472
highly effective in classifying subgenres such as Gothic, detective stories, and even 473
science fiction. 474

Nevertheless, although this method proves valuable in specific contexts, it is not without 475
two limitations. First, it does not consider the word order within the text. This limitation 476
means that the sequential arrangement of words, which is crucial for capturing the 477

nuances of literary elements like plot and narrative structure, is ignored. Second, there is a risk of overfitting to the idiolects of writers, particularly when emphasizing the most frequent words. Additionally, these tools may inadvertently capture chronolectal aspects, as it is established that the approximate writing date of a book can be predicted based on the prevalence of certain most frequent words (Seminck et al. 2022).

In this paper we rely on two distinct feature extraction approaches: the classic BoW as a control experiment, and the BookNLP-fr one, which we will implement as follows. The idea is based on a previous study (Kohlmeyer et al. 2021) where researchers 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 these findings, we adapted their methodology to evaluate the impact of these features on subgenre classification when contrasted with the traditional BoW approach.

The method runs our BookNLP pipeline on our texts, allowing us to automatically retrieve, on the one hand, information related to space-time, notably with the set of LOC, FAC, GPE, TIME, and VEH. On the other hand, it provides information related to characterization, including all verbs for which characters are patients (PATIENT) or agents (AGENT), as well as the set of adjectives that will characterize them (ADJ).

Thus, two types of features are under consideration:

- For the BoW, we relied on the 600 most frequent lemmas, excluding the first 200, which comprise non-informative stop words not relevant to our subgenre case study. They could have been relevant if we wanted to acknowledge the authors who wrote in a specific subgenre, but it is not our goal here, and we will discuss how we handled this bias in Section 5.2.3.
- For the BookNLP-fr features, we compiled for each novel, lists of words extracted by BookNLP-fr. We then obtained vector representations using a Paragraph Vectors model (Le and Mikolov 2014) (Doc2Vec) trained on a subset of our novel dataset. Two vector embeddings of 300 dimensions were generated: one for characterization (AGENT, PATIENT, ADJ) and one for space-time (LOC, FAC, GPE, TIME, VEH).

Therefore we obtained two datasets for training, one with 600 dimensions representing the 600 most frequent lemmas, and the other with also 600 dimensions representing the two concatenated Doc2Vec vectors, one for the characterization and one for the space and time.

5.2.3 Modeling

We opted for an SVM as it has been demonstrated that these models obtain the best performance in classifying literary texts (Yu 2008), and more specifically literary subgenres (Hettinger et al. 2016). In this paper, we used the implementation of Pedregosa et al. (2011). The SVM doesn't perform multiclassification per se, but it classifies each subgenre against the others in binary classification and then aggregates the results.

Therefore, we don't have a single classification, but rather 518

$$\frac{n_{\text{classes}} \cdot (n_{\text{classes}} - 1)}{2}$$

With our 5 subgenres, this implementation results in 10 different classifications. 519

Considering our task of subgenre classification, we wanted to limit idiolectal bias, 520
 especially for the model trained on the BoW. To do so, we implemented Scikit-learn's 521
 Group strategy. All works by the same author (group) were placed in the same fold. 522
 Thus, each group will appear exactly once in the test set across all folds. Since SVM 523
 models are quite sensitive working with imbalanced classes, we re-balanced the classes 524
 before implementing the classification by randomly taking 130 novels for each subgenre. 525
 We implemented this selection a hundred times and for each resulting sample the model 526
 was run in a 5-fold cross-validation setting. The following results are aggregated from 527
 this process. 528

5.3 Results 529

5.3.1 BoW vs BookNLP-fr features 530

	Precision	Recall	F1-score	Support	Accuracy
Children	0.75	0.75	0.75	130	
Memoirs	0.79	0.82	0.80	130	
Detective	0.67	0.68	0.67	130	
Adventure	0.60	0.65	0.62	130	
Romance	0.84	0.72	0.80	130	
Full Dataset				650	0.72

Table 6: Classification Report for BoW

	Precision	Recall	F1-score	Support	Accuracy
Children	0.65	0.79	0.71	130	
Memoirs	0.78	0.89	0.84	130	
Detective	0.68	0.70	0.70	130	
Adventure	0.73	0.73	0.73	130	
Romance	0.90	0.65	0.75	130	
Full Dataset				650	0.75

Table 7: Classification Report for BookNLP-fr features.

Tables 6 and 7 display the classification report of the models' evaluation on the test set. 531
 Both models achieve good results: 72% for the BoW-based model and the BookNLP- 532
 based model achieves 75% accuracy. This means that our models are capable of correctly 533
 identifying the subgenre three out of four times, whereas a random baseline yields an 534
 accuracy score of 0.2. The main result here is that differences exist among our subgenres, 535
 whether from the perspective of text structure with MFW or from a semantic standpoint 536
 with BookNLP. The fact that the BookNLP-based model obtains an additional 3 points of 537
 accuracy might not be revolutionary, but the primary argument for this type of feature 538
 extraction lies more in the interpretation of features, as discussed in subsection 5.4. 539

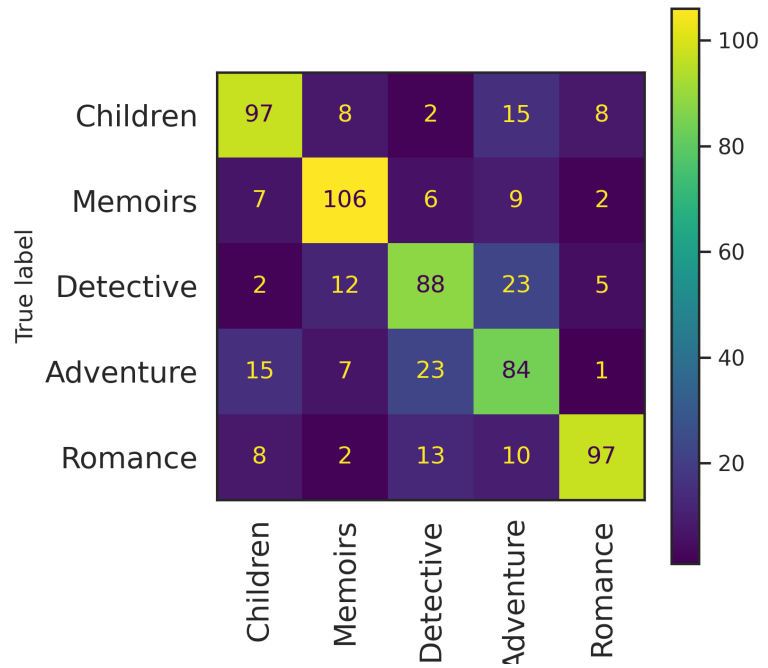


Figure 2: Confusion Matrix for BoW.

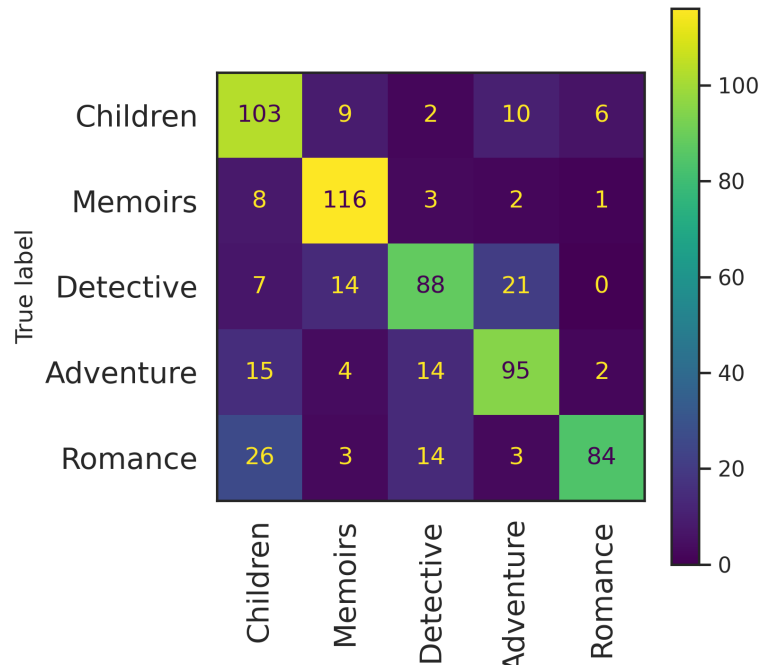


Figure 3: Confusion Matrix for BookNLP-fr features.

To enhance our comprehension of how the models behave and the nature of their errors, we visualize their confusion matrices in Figure 2 and Figure 3. The x-axis represents the predicted subgenre, while the y-axis represents the expected subgenre. A perfect classification would display a diagonal filled with 130 correct predictions for each subgenre.

We observe that both models have quite similar error patterns, and one distinct scenario stands out: Both models predict 'Adventure' instead of 'Detective' (23 errors for BoW, 21

for BookNLP). These common errors are quite understandable since these two subgenres share many similarities, including a penchant for suspense and violent action, which could confuse the models.

Another scenario seemed highly instructive for analysis: The errors made by the models when predicting the label 'Children', but the expected subgenre is 'Romance'. The BoW model performs quite well with 8 errors, but the BookNLP-based model makes 26 errors. The semantic model thus faces more challenges in distinguishing between these two subgenres, which makes sense, as both subgenres are characterized by themes centered around emotions and relationships between characters, common features to both subgenres.

5.3.2 BookNLP-fr Features Accuracy for Subgenre Classification

In this section, the objective is to evaluate, on the one hand, whether specific individual features from BookNLP can classify our subgenres, and on the other hand, we will attempt to interpret the differences in performance for each. Here, each pipeline is trained with a Doc2Vec vector of 300 dimensions for each type of feature.

BookNLP-fr features	Accuracy
LOC	0.45
FAC	0.59
VEH	0.42
GPE	0.47
TIME	0.50
PATIENT	0.52
AGENT	0.62
ADJ	0.50
Baseline	0.2

Table 8: BookNLP-fr features accuracy.

A first obvious observation is that all our models achieve results at least twice as good as the baseline. The information contained in each of these features is therefore highly relevant from the subgenre perspective. The 'VEH' class lags a bit behind (42% accuracy), which may suggest that vehicles are not decisively discriminating among our subgenres, but it is our least represented class in our texts, and therefore, there may not be enough data. Very good results are obtained for the 'FAC' (0.59) and 'AGENT' (0.62). This indicates that subgenres distinguish well in terms of mentioned buildings or verbs where the character is agentic, meaning that the type of action a character takes is specific to each subgenre.

Interestingly, the misclassifications (see the confusion matrices in the Appendix A for each individual feature), the same pattern emerges (misclassification of 'Adventure' instead of 'Detective' and 'Children' instead of 'Romance'), but the error rates vary depending on the features used. This can provide a lot of information about the differences and similarities between certain subgenres. The next section 5.4 offers an interpretation closely examining these anomalies.

5.4 Interpretability

This section explores the interpretation of the two SVM models, BoW-based and BookNLP-based. It focuses on the misclassifications of ‘Adventure’ instead of ‘Detective’.

One of the advantages of the SVM pipeline is the ability to investigate the statistical inferences of the models when the kernel is in linear mode. The SVM searches for the plane in the latent space of words that best separates our two categories. Each dimension receives a coefficient, with a negative sign if the coefficient is used to predict a specific class and a positive sign for the other. For the BoW-based model, it’s quite straightforward as a coefficient is assigned to each word, as can be seen in Figure 4.

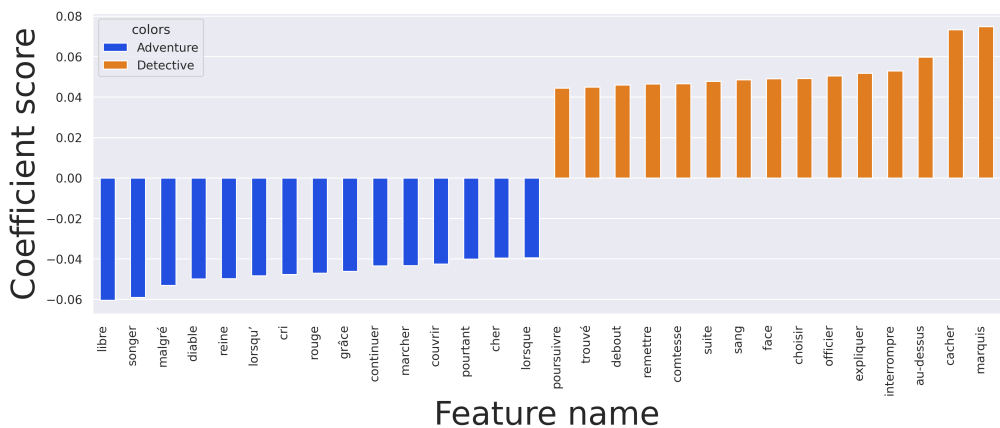


Figure 4: BoW discriminant features for Adventure vs Detective classification.

Looking at the coefficients assigned for the Adventure vs. Detective classification, we find some relevant elements, such as the presence of the word ‘free’ (‘libre’) as the most discriminant word for assigning the Adventure label. Apart from that, with perhaps ‘cry’ (‘cri’), which could signify adventure, few clues remain. Verbs such as ‘dream,’ ‘walk,’ ‘continue,’ or conjunctions like ‘when’ (‘lorsque’), ‘despite’ (‘malgré’), and ‘yet’ (‘pourtant’) are not really characteristic of adventure novels. It is difficult to conclude, except that these less significant coefficients seem to indicate the model’s difficulty in distinguishing between the two sub-genres.

For the BookNLP-based model, it’s a bit more complex since the coefficients are assigned to each dimension of the Doc2Vec vectors. Therefore, we aggregated the coefficients by feature type to gain a more concrete overview of the results. Figure 5 illustrates the sum of all coefficients for each feature extracted by BookNLP-fr. We conducted a t-test to confirm that the difference between the means of the populations is statistically significant. Taking adjectives as an example (T-statistic: 28.7; P-value: 2.25×10^{-180}), we observe that the model relies more on these dimensions to assign the label ‘detective’ compared to ‘adventure’.

This could be explained by the strong emphasis placed on character psychology in detective novels, especially those involving criminals and detectives. For instance, in *Maigret et le tueur* (1969), George Simenon’s beloved detective (*Maigret*) is frequently characterized as ‘wise,’ ‘whimsical,’ or even ‘happy,’ while criminals are ‘suspicious’ or ‘villainous’. This doesn’t imply a lack of characterization in adventure novels but rather suggests that it is not a distinctive feature of the subgenre compared to detective novels.

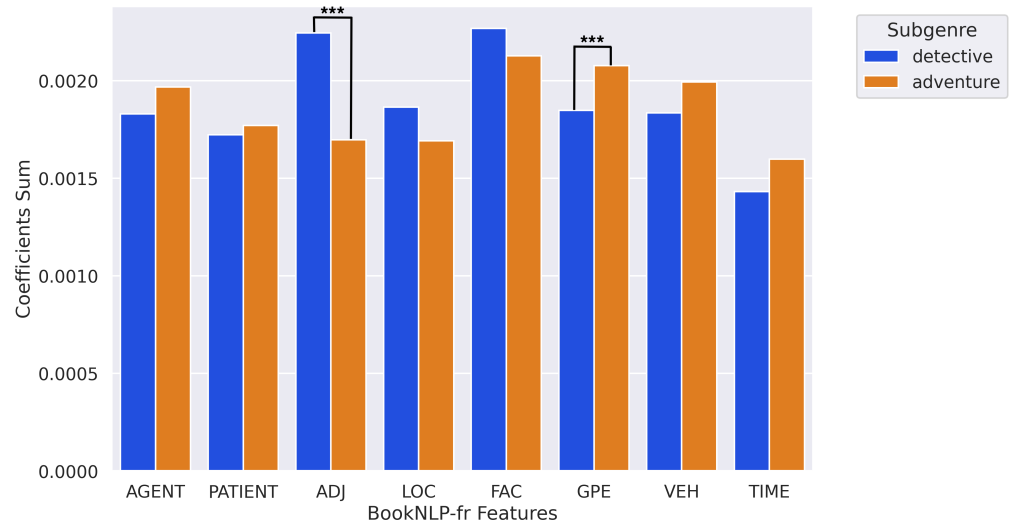


Figure 5: BookNLP-fr discriminant features for Adventures vs Detective classification. '***' meaning $p < 0.001$.

Considering Geo-Political Entities (T-statistic: -21.0 ; P-value: 8.49×10^{-98}), the reasoning is inverse: the model relies slightly more on the dimensions of the GPE vector to assign the adventure label than the detective label. This makes sense when examining GPEs for example in *Les trappeurs de l'Arkansas* by Gustave Aimard (1857): 'Hermosillo', 'America', 'the New World', 'Guadalajara', 'Mexico', etc. The novel heavily emphasizes exotic locations and mentions places in the American or Mexican West for this purpose. GPEs in detective novels are more commonplace, as these novels often take place in France, typically in an urban setting.

Thus the model has learned that certain dimensions of characterization are more strongly associated with a particular subgenre (such as adjectives for detective novels), and that certain dimensions of the GPE or TIME vector are important for assigning the adventure label. Let's now generalize our approach to the entire classification process.

Examining the behavior of the coefficients when aggregated for the 10 classifications, we can observe the graph shown in Figure 6. This graph depicts the model coefficients after training based on the vectors of each facet, using a dataset of 2400 dimensions. We consider this graph as a dive into the model's inferences, where it will assign more weight to certain categories to assign a specific subgenre.

For example, it is observed that the value of 'FAC' is very high for the detective genre, indicating a particular specificity for this sub-genre. Details of locations, crime scenes, investigations in specific places, detective offices, interrogation rooms, etc., are distinguishing elements for this sub-genre. The same applies to 'GPE' for the adventure label, as seen previously, with an emphasis on exoticism that may play a role here, even though 'LOC' and 'FAC' do not show significant differentiation from this perspective. Conversely, for romance and the 'TIME' vector, where the coefficients for these vectors lag behind other sub-genres. Examples of time in romance novels may be used more to describe emotional moments or stages in relationships rather than to highlight complex temporal events. Consequently, the model might perceive that the 'TIME' vector is not as discriminative for this category.

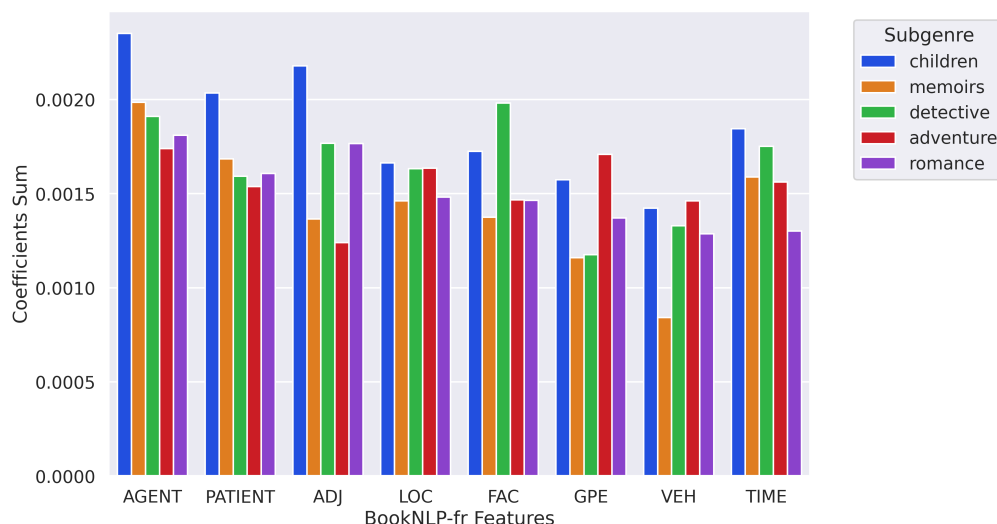


Figure 6: BookNLP-fr discriminant features for the classification.

We have thus demonstrated that the BoW-based classification approach is challenging to interpret, as certain highly discriminating words do not appear to bring about key distinctions between the subgenres. The BookNLP-fr-based method may offer an insightful understanding of the specificities that differentiate one subgenre from another. Both approaches do not completely substitute for each other since we are examining features of different nature (vocabulary vs semantic), but they can complement each other to enhance interpretability.

Diving into the model’s indications, several types of features were observed to interpret the model’s inferences. Many differences among the features were noticed, although we did not have the space to interpret all of them in this article. Much work remains to be done, and new experiments should be considered, for instance going beyond the SVM, including the use of deep neural networks and textual deconvolution saliency Vanni et al. (2018), which could facilitate the return to close reading based on the embeddings derived from BookNLP-fr data.

6. Discussion

6.1 Working with Imperfect Annotations

The utilization of computers for annotating literary texts has profoundly changed the landscape of literary studies, enabling the annotation of vast amounts of texts with unprecedented efficiency. This enables the community to address research questions that were out of reach before, such as a study at scale of characters with disabilities (Dubnick et al. 2018) or the quantitative analysis of characters in fanfiction (Milli and Bamman 2016) and a quantitative, diachronic study of things appearing in fiction (Piper and Bagga 2022). However, this advancement is not without its challenges, particularly in the context of the inherent errors that may accompany automated annotation processes. This poses a twofold challenge for researchers engaged in the field of CLS.

Firstly, ensuring the reliability of studies based on imperfect annotations is a critical concern. Scholars must grapple with the task of guaranteeing that errors, though present,

remain at a marginal level and do not compromise the validity of their research findings. 663
 This necessitates a careful balance between the benefits of computational efficiency and 664
 the maintenance of accuracy in annotations. Researchers are challenged to develop 665
 methodologies and quality control measures that safeguard against the potential pitfalls 666
 introduced by errors in the annotation process. 667

Secondly, the acceptance of computational approaches by literary scholars is not guaran- 668
 teed, as the traditional paradigm within literary studies often revolves around meticu- 669
 lous, supposedly perfect annotations. The shift to working with non-perfect annotations, 670
 even if the errors are marginal, represents a departure from the established norm. This 671
 cultural shift within the academic community poses a psychological barrier, as literary 672
 scholars may be hesitant to fully embrace computational methods if they perceive a 673
 compromise in the level of precision to which they are accustomed. 674

Addressing these challenges requires not only the refinement of computational tools for 675
 annotation but also a broader cultural shift within the academic community. There is 676
 a need for transparent communication about the limitations of automated annotation 677
 processes, the establishment of best practices for mitigating errors, and the development 678
 of strategies to ensure that computational approaches align with the standards expected 679
 both in literary studies and in computer science. 680

6.2 Maintaining Annotations Tools in the Era of Large Language Models 681

The field of computational literary studies is currently grappling with a significant 682
 challenge due to the rapid evolution of natural language processing, particularly with 683
 the proliferation of large language models (LLMs). The continuous emergence of new 684
 LLMs has led to an accelerated pace of research in the domain. While this dynamism 685
 brings about positive outcomes, such as increased research activity, the introduction of 686
 novel tasks, and the generation of new results, it also presents several inherent dangers. 687

One primary challenge lies in the technical aspect of keeping annotation tools up to 688
 date amidst the constant production of new LLMs by the research community and 689
 the industry. There is a delicate balance to strike, ensuring that annotation systems 690
 remain up-to-date, without expending an excessive amount of resources on incessantly 691
 adapting to the latest trends in LLM development. The challenge here is not just about 692
 technological compatibility but also about efficiently managing the resources required 693
 for frequent updates and integrations, and to produce software that is usable by a large 694
 community (i.e. software should not be dependent on an unreasonably heavy computer 695
 infrastructure). 696

A more critical concern revolves around the need to guarantee the reproducibility of 697
 research outcomes. The rapid evolution of LLMs implies that a specific version in use 698
 today may become obsolete or unavailable tomorrow. This raises the risk that crucial 699
 details, such as the corpus utilized, configuration parameters, and hyperparameters 700
 of the model, may not be adequately documented in research reports. Ensuring repro- 701
 ducibility becomes a substantial challenge as the landscape of LLMs continues to evolve, 702
 necessitating a concerted effort to establish standardized practices for reporting model 703
 specifications and associated details. 704

In addressing these challenges, we believe it is crucial to focus not only on technical aspects but also on developing robust frameworks for documentation and reproducibility. Establishing clear guidelines for reporting model specifications, documenting corpus details, and archiving relevant information becomes paramount for the field.

7. Conclusion

In this paper, we introduced the BookNLP-fr pipeline, with a particular emphasis on entity recognition and coreference resolution. Demonstrating its practical utility, we illustrated how this software facilitates the analysis of extensive French literary corpora, relying on semantic features unique to the texts under examination. Through this study, we hope to show the potential of natural language processing in analyzing large literary corpora, to go beyond purely statistical approaches and to overcome bias by taking into account an unprecedented number of texts and not only the reduced set of texts of the literary canon. In concrete terms, we distinguish three research directions, all of which carry the above-described desire for large-scale generalization:

1. Studies on the characteristics of literary genre : BookNLP-en can be used to retrieve textual features of a semantic nature, in particular entities that provide information on the spatio-temporal setting of the story. The latter are very important for determining literary genres. For example, adventure novels have a very specific spatio-temporal setting (the emphasis is on the importance of geographical disorientation), while romance novels take place in a more urban, modern setting. The BookNLP-fr tools could thus be crucial for automatic classification.
2. Characterization: co-reference chains with mentions of a character allow us to recover how each character is portrayed. In this way, we can study the differences between certain types of characters on a large scale. For example, it's possible to report on how men and women have been characterized in literature over time (e.g. Naguib et al. 2022; Vianne et al. 2023) or what role secondary characters actually play in the narrative (Barré et al. 2023). To cite other examples: a tool like BookNLP makes it possible to study how characters with disabilities are presented (Dubnicek et al. 2018) or to carry out a quantitative analysis of characters in fan fiction (Milli and Bamman 2016).
3. Detection of specific scenes: BookNLP could be capable of detecting specific scenes in novels; these could be defined by one or more characters gravitating around a precise location and carrying out particular actions. This scene detection, understood as a minimal narrative unit, could enable us to better understand the workings of the plot by breaking down its layout over the course of the story.

Future work on the BookNLP-fr pipeline will include a renewed exploration of the concepts of events and scenes, aiming to establish an annotation framework that aligns with literary perspectives. Additionally, we plan to address the question of quotation analysis and attribution. Finally, a key focus will be on ensuring that results undergo scientific evaluation and that recent advancements in natural language processing can be continuously integrated, all while preserving the distinctive nature of literary works and literary studies. In that way, BookNLP-fr can play an significant role in the domains

of automatic literary analysis and cultural analysis. Literary questions, one even more exciting and ambitious than the other, can finally be addressed automatically on a large scale.

8. Data Availability

Data can be found here: <https://github.com/lattice-8094/fr-litbank>.

9. Software Availability

Software can be found here: <https://seafile.rlp.net/f/6c9d680114fe4583a89c/?dl=1>.

10. Acknowledgements

This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA0001 (PRAIRIE 3IA Institute).

11. Author Contributions

Frédérique Mélanie-Becquet: Conceptualization, Data Curation, Supervision

Jean Barré: Formal Analysis, Writing – review & editing

Olga Seminck: Formal Analysis, Writing – review & editing

Clément Plancq: Conceptualization, Software

Marco Naguib: Conceptualization, Software

Martial Pastor: Conceptualization, Software

Thierry Poibeu: Conceptualization, Writing – original draft, review & editing, Supervision

References

- Aristote (1990). *Poétique*. Le Livre de poche. Librairie générale française.
- Bachtin, Michail Michajlovič (2006). *Esthétique et théorie du roman*. Collection Tel 120. Gallimard.
- Bamman, David (2020). *Multilingual BookNLP: Building a Literary NLP Pipeline Across Languages*. <https://apps.neh.gov/publicquery/main.aspx?f=1&gn=HAA-271654-2>.
- . Accessed: January 17, 2024.
- (2021). *BookNLP*. <https://github.com/booknlp/booknlp>.

- Bamman, David, Olivia Lewke, and Anya Mansoor (2020). “An Annotated Dataset of Coreference in English Literature”. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: European Language Resources Association, 44–54. <https://aclanthology.org/2020.lrec-1.6>.
- Bamman, David, Sejal Popat, and Sheng Shen (2019). “An annotated dataset of literary entities”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2138–2144. [10.18653/v1/N19-1220](https://doi.org/10.18653/v1/N19-1220).
- Bamman, David, Ted Underwood, and Noah A. Smith (2014). “A Bayesian Mixed Effects Model of Literary Character”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. ACL 2014. Association for Computational Linguistics, 370–379. [10.3115/v1/P14-1035](https://doi.org/10.3115/v1/P14-1035).
- Barré, Jean, Pedro Cabrera Ramírez, Frédérique Mélanie, and Ioanna Galleron (2023). “Pour une détection automatique de l’espace textuel des personnages romanesques”. In: *Humanistica 2023*. Corpus. Association francophone des humanités numériques. Genève, Switzerland, 56–61. <https://hal.science/hal-04105537>.
- Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo (2018). “quanteda: An R package for the quantitative analysis of textual data”. In: *Journal of Open Source Software* 3.30, 774. [10.21105/joss.00774](https://doi.org/10.21105/joss.00774).
- Bird, S., E. Klein, and E. Loper (2019). *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*. <https://www.nltk.org/book/ch00.html>.
- Bourdieu, Pierre (1979). *La distinction: critique sociale du jugement*. Le Sens commun 58. Éditions de Minuit.
- Calvo Tello, José (Dec. 31, 2021). *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Bielefeld University Press. [10.1515/978383839459256](https://doi.org/10.1515/978383839459256).
- Cohen, Jacob (1960). “A coefficient of agreement for nominal scales”. In: *Educational and psychological measurement* 20.1, 37–46.
- Dekker, Niels, Tobias Kuhn, and Marieke van Erp (2019). “Evaluating named entity recognition tools for extracting social networks from novels”. In: *PeerJ Computer Science* 5, e189. <https://doi.org/10.7717/peerj-cs.189>.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, 4171–4186. [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423).
- Dubnick, Ryan, Ted Underwood, and J Stephen Downie (2018). “Creating A Disability Corpus for Literary Analysis: Pilot Classification Experiments”. In: *iConference 2018 Proceedings*.
- Durandard, Noé, Viet Anh Tran, Gaspard Michel, and Elena Epure (2023). “Automatic Annotation of Direct Speech in Written French Narratives”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

- Papers*). Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Association for Computational Linguistics, 7129–7147. [10.18653/v1/2023.acl-long.393](https://doi.org/10.18653/v1/2023.acl-long.393). 822 823
- Eder, Maciej, Jan Rybicki, and Mike Kestemont (2016). “Stylometry with R: a package for computational text analysis”. In: *R Journal* 8.1, 107–121. <https://journal.r-project.org/archive/2016/RJ-2016-007/index.html>. 824 825 826
- Emelyanov, A. and E. Artemova (2019). “Multilingual Named Entity Recognition Using Pretrained Embeddings, Attention Mechanism and NCRF”. In: *Proc. of the 7th Workshop on Balto-Slavic Natural Language Processing*. Florence, Italy, 94–99. 827 828 829
- Genette, Gérard (1986). “Introduction à l’architexte”. In: *Théorie des genres*. Ed. by Gérard Genette and Tzvetan Todorov. Points 181. Éd. du Seuil. 830 831
- Gong, Xiaoyun, Yuxi Lin, Ye Ding, and Lauren Klein (2022). “Gender and power in japanese light novels”. In: *Proceedings http://ceur-ws.org ISSN 1613, 0073*. 832 833
- Heiden, Serge (2010). “The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme”. In: *24th Pacific Asia conference on language, information and computation*. Vol. 2. 3. Institute for Digital Enhancement of Cognitive Development, Waseda University, 389–398. 834 835 836 837
- Hettinger, Lena, Fotis Jannidis, Isabella Reger, and Andreas Hotho (2016). “Significance Testing for the Classification of Literary Subgenres”. In: *ADHO 2016 - Kraków*. <https://dh-abstracts.library.cmu.edu/works/2630> (visited on 01/18/2024). 838 839 840
- Hogenboom, F., F. Frasincar, U. Kaymak, F. de Jong, and E. Caron (2016). “A survey of event extraction methods from text for decision support systems”. In: *Decision Support Systems* 85.c, 12–22. [Doi:10.1016/j.dss.2016.02.006](https://doi.org/10.1016/j.dss.2016.02.006). 841 842 843
- Hudspeth, Marisa, Sam Kovaly, Minhwa Lee, Chau Pham, and Przemyslaw Grabowicz (n.d.). “Gender and Power in Latin Narratives”. In: (). 844 845
- Jauß, Hans Robert (1982). *Toward an aesthetic of reception*. Trans. by Timothy Bahti. Univ. of Minnesota Press. 846 847
- Joshi, Mandar, Omer Levy, Luke Zettlemoyer, and Daniel Weld (2019). “BERT for Coreference Resolution: Baselines and Analysis”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, 5803–5808. [10.18653/v1/D19-1588](https://doi.org/10.18653/v1/D19-1588). 848 849 850 851 852 853
- Ju, Meizhi, Makoto Miwa, and Sophia Ananiadou (June 2018). “A Neural Layered Model for Nested Named Entity Recognition”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, 1446–1459. [10.18653/v1/N18-1131](https://doi.org/10.18653/v1/N18-1131). 854 855 856 857 858 859
- Kohlmeyer, Lasse, Tim Repke, and Ralf Krestel (2021). “Novel Views on Novels: Embedding Multiple Facets of Long Texts”. In: *2021 Association for Computing Machinery*. 860 861
- Landragin, Frédéric (2016). “Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT)”. In: *Bulletin de l’Association Française pour l’Intelligence Artificielle* 92, 11–15. 862 863 864
- (2021). “Le corpus Democrat et son exploitation. Présentation”. In: *Langages* 4, 11–24. 865
- Langlais, Pierre-Carl (May 2021). *Fictions littéraires de Gallica / Literary fictions of Gallica*. Version 1. Zenodo. [10.5281/zenodo.4751204](https://doi.org/10.5281/zenodo.4751204). 866 867

- Le, Quoc V. and Tomás Mikolov (2014). *Distributed Representations of Sentences and Documents*. arXiv: [1405.4053](https://arxiv.org/abs/1405.4053). 868 869
- Leblond, Aude (2022). *Corpus Chapitres*. Version v1.0.0. [10.5281/zenodo.7446728](https://zenodo.org/record/7446728). 870
- Lee, Kenton, Luheng He, Mike Lewis, and Luke Zettlemoyer (2017). “End-to-end Neural Coreference Resolution”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, 188–197. [10.18653/v1/D17-1018](https://arxiv.org/abs/10.18653/v1/D17-1018). <https://aclanthology.org/D17-1018>. 871 872 873 874 875
- Lotman, Yuri (1977). *The Structure of the Artistic Text*. Michigan Univ. Press (Michigan Slavic Contributions No. 7). 876 877
- Luo, Xiaoqiang and Sameer Pradhan (2016). “Evaluation metrics”. In: *Anaphora Resolution: Algorithms, Resources, and Applications*. Springer, 141–163. 878 879
- Manning, C., M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky (2014). “The Stanford CoreNLP Natural Language Processing Toolkit”. In: *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*. Baltimore: ACL. 880 881 882 883
- Manovich, Lev (2018). “The science of culture? Social computing, digital humanities and cultural analytics”. In. 884 885
- Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot (2020). “CamemBERT: a Tasty French Language Model”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 886 887 888 889
- Maynard, D., K. Bontcheva, and I. Augenstein (2017). “Named Entity Recognition and Classification”. In: *Natural Language Processing for the Semantic Web. Synthesis Lectures on Data, Semantics, and Knowledge*. Cham: Springer. 890 891 892
- Milli, Smitha and David Bamman (2016). “Beyond Canonical Texts: A Computational Analysis of Fanfiction”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Ed. by Jian Su, Kevin Duh, and Xavier Carreras. Austin, Texas: Association for Computational Linguistics, 2048–2053. [10.18653/v1/D16-1218](https://arxiv.org/abs/10.18653/v1/D16-1218). <https://aclanthology.org/D16-1218>. 893 894 895 896 897
- Moretti, Franco (2000). “Conjectures on world literature”. In: *New Left Review*. 898
- Naguib, Marco, Marine Delaborde, Blandine Andrault, Anaïs Bekolo, and Olga Seminck (2022). “Romanciers et romancières du XIXème siècle : une étude automatique du genre sur le corpus GIRLS (Male and female novelists : an automatic study of gender of authors and their characters)”. In: *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier TAL et Humanités Numériques (TAL-HN)*. Ed. by Ludovic Moncla and Carmen Brando. Avignon, France: ATALA, 66–77. <https://aclanthology.org/2022.jeptalnrecital-humanum.8>. 899 900 901 902 903 904 905
- Pedregosa, F., 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, and E. Duchesnay (2011). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12, 2825–2830. 906 907 908 909
- Perri, Vincenzo, Lisi Qarkaxhija, Albin Zehe, Andreas Hotho, and Ingo Scholtes (2022). *One Graph to Rule them All: Using NLP and Graph Neural Networks to analyse Tolkien’s Legendarium*. arXiv: [2210.07871](https://arxiv.org/abs/2210.07871) [cs.CL]. 910 911 912
- Piper, Andrew and Sunyam Bagga (2022). “A Quantitative Study of Fictional Things”. In: 268–279. https://ceur-ws.org/Vol-3290/long_paper1576.pdf. 913 914

- Piper, Andrew, Richard Jean So, and David Bamman (2021). “Narrative Theory for Computational Narrative Understanding”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 298–311. 10.18653/v1/2021.emnlp-main.26. <https://aclanthology.org/2021.emnlp-main.26>.
- Poesio, Massimo, Juntao Yu, Silviu Paun, Abdulrahman Aloraini, Pengcheng Lu, Janosch Haber, and Derya Cokal (2023). “Computational models of anaphora”. In: *Annual Review of Linguistics* 9, 561–587.
- Rockwell, G. and S. Sinclair (2016). *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. MIT Press.
- Ryan, Marie-Laure, Kenneth Foote, and Maoz Azaryahu (2016). *Narrating space/spatializing narrative: Where narrative theory and geography meet*. The Ohio State University Press.
- Schaeffer, Jean-Marie (1989). *Qu’est-ce qu’un genre littéraire? Poétique*. Seuil.
- Schlegel, Friedrich, August Wilhelm Schlegel, August Ferdinand Bernhardt, and Wilhelm Dilthey (1996). *Critique et herméneutique dans le premier romantisme allemand : Textes de F. Schlegel, F. Schleiermacher, F. Ast, A.W. Schlegel, A.F. Bernhardt, W. Dilthey*. Trans. by Denis Thouard. Opuscles. Presses universitaires du Septentrion. <https://books.openedition.org/septentrion/95397> (visited on 01/23/2024).
- Schmid, W. (2010a). *Mental Events*. Hamburg University Press.
- (2010b). *Narratology. An introduction*. de Gruyter.
- Schöch, Christof (2017). “Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama”. In: *Digital Humanities Quarterly* 11.2.
- Seminck, Olga, Philippe Gambette, Dominique Legallois, and Thierry Poibeau (2022). “The Evolution of the Idiolect over the Lifetime: A Quantitative and Qualitative Study of French 19th Century Literature”. In: *Journal of Cultural Analytics* 7.3. 10.22148/001c.37588.
- Silge, J. and D. Robinson (2017). *Text Mining with R: A Tidy Approach*. <http://repo.darmajaya.ac.id/5417/> (visited on 09/19/2017).
- Sims, M., J. Ho Park, and David Bamman (2019). “Literary Event Detection”. In: *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 3623–3634.
- Sobchuk, Oleg and Artjoms Šeļa (2023). *Computational thematics: Comparing algorithms for clustering the genres of literary fiction*. arXiv: 2305.11251 [cs.CL].
- Soni, Sandeep, Amanpreet Sihra, Elizabeth F Evans, Matthew Wilkens, and David Bamman (2023). “Grounding Characters and Places in Narrative Texts”. In: *arXiv preprint arXiv:2305.17561*.
- Sprugnoli, R. and S. Tonelli (2016). “Novel Event Detection and Classification for Historical Texts”. In: *Computational Linguistics* 45.2, 229–265.
- Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii (2012). “brat: a Web-based Tool for NLP-Assisted Text Annotation”. In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Frédérique Segond. Avignon, France: Association for Computational Linguistics, 102–107. <https://aclanthology.org/E12-2021>.

- Toro Isaza, Paulina, Guangxuan Xu, Toye Oloko, Yufang Hou, Nanyun Peng, and Dakuo Wang (2023). “Are Fairy Tales Fair? Analyzing Gender Bias in Temporal Narrative Event Chains of Children’s Fairy Tales”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, 6509–6531. [10.18653/v1/2023.acl-long.359](https://doi.org/10.18653/v1/2023.acl-long.359).
- Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. (2023). “Llama 2: Open foundation and fine-tuned chat models”. In: *arXiv preprint arXiv:2307.09288*.
- Underwood, Ted (2019). *Distant horizons: digital evidence and literary change*. The University of Chicago Press, 1–33.
- Underwood, Ted, David Bamman, and Sabrina Lee (2018). “The Transformation of Gender in English-Language Fiction”. In: *Cultural Analytics* Feb 13 2018. [10.22148/16.019](https://doi.org/10.22148/16.019).
- Van Cranenburgh, Andreas and Frank Van Den Berg (2023). “Direct Speech Quote Attribution for Dutch Literature”. In: *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Ed. by Stefania Degaetano-Ortlieb, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz. Association for Computational Linguistics, 45–62. [10.18653/v1/2023.latechcl-fl-1.6](https://doi.org/10.18653/v1/2023.latechcl-fl-1.6).
- Vanni, Laurent, Melanie Ducoffe, Carlos Aguilar, Frederic Precioso, and Damon Mayaffre (2018). “Textual Deconvolution Saliency (TDS) : a deep tool box for linguistic analysis”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, 548–557. [10.18653/v1/P18-1051](https://doi.org/10.18653/v1/P18-1051).
- Vianne, Laurine, Yoann Dupont, and Jean Barré (2023). “Gender Bias in French Literature”. In: *Computational Humanities Research Conference*. CEUR Workshop Proceedings (CEUR-WS. org), 247–262. <https://ceur-ws.org/Vol-3558/paper2449.pdf>.
- Vishnubhotla, Krishnapriya, Frank Rudzicz, Graeme Hirst, and Adam Hammond (2023). “Improving Automatic Quotation Attribution in Literary Novels”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, 737–746. [10.18653/v1/2023.acl-short.64](https://doi.org/10.18653/v1/2023.acl-short.64).
- Weischedel, Ralph, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. (2013). “Ontonotes release 5.0 ldc2013t19”. In: *Linguistic Data Consortium, Philadelphia, PA* 23, 170.
- Woloch, Alex (2003). *The One vs. the Many*. Princeton University Press.
- Yu, B. (Sept. 5, 2008). “An evaluation of text classification methods for literary study”. In: *Literary and Linguistic Computing* 23.3, 327–343. [10.1093/llic/fqn015](https://doi.org/10.1093/llic/fqn015).
- Zehe, Albin, Leonard Konle, Lea Katharina Dümpelmann, Evelyn Gius, Andreas Hotho, Fotis Jannidis, Lucas Kaufmann, Markus Krug, Frank Puppe, Nils Reiter, Annekea Schreiber, and Nathalie Wiedmer (2021). “Detecting Scenes in Fiction: A new Segmentation Task”. In: *Proceedings of the 16th Conference of the European Chapter of the*

- Association for Computational Linguistics: Main Volume*. Ed. by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty. Online: Association for Computational Linguistics, 3167–3177. [10.18653/v1/2021.eacl-main.276](https://doi.org/10.18653/v1/2021.eacl-main.276).
- Zhang, Weiwei, Jackie Chi Kit Cheung, and Joel Oren (2019). “Generating character descriptions for automatic summarization of fiction”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, 7476–7483.
- Zundert, Joris van, Andreas van Cranenburgh, and Roel Smeets (2023). “Putting Dutch coref to the Test: Character Detection and Gender Dynamics in Contemporary Dutch Novels”. In: *CHR 2023: Computational Humanities Research Conference*. CEUR Workshop Proceedings (CEUR-WS.org). Paris, France, 757–771. <https://ceur-ws.org/Vol-3558/paper9264.pdf>.
- Zundert, Joris van, Marijn Koolen, Julia Neugarten, Peter Boot, Willem van Hage, and Ole Mussmann (2022). “What Do We Talk About When We Talk About Topic?” In: *CHR 2022: Computational Humanities Research Conference*. Antwerp, Belgium. https://ceur-ws.org/Vol-3290/short_paper5533.pdf.

A. Appendix: Confusion matrices for BookNLP-fr-based models

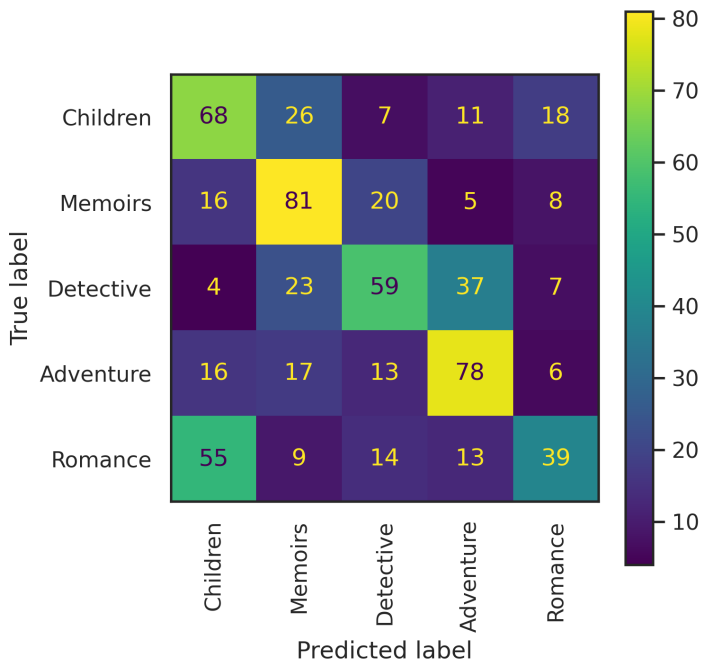


Figure 7: Confusion Matrix for ADJ features

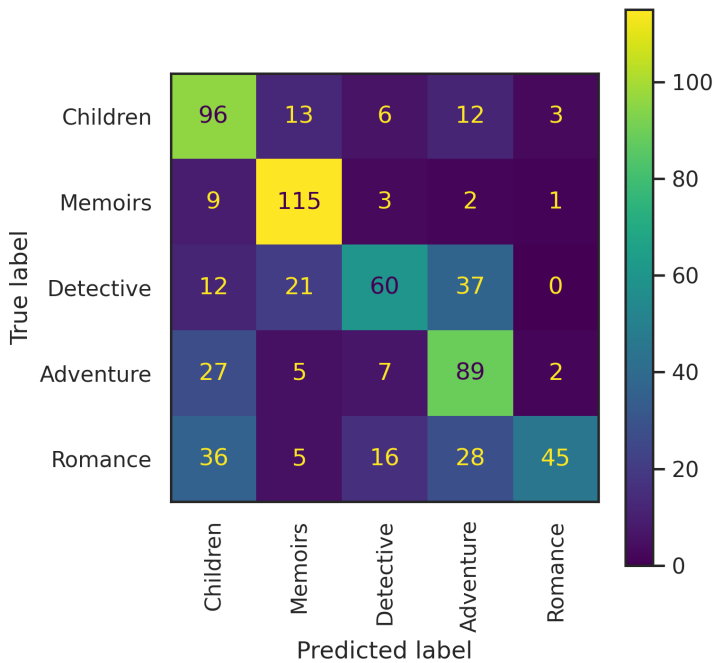


Figure 8: Confusion Matrix for AGENT features

conference version

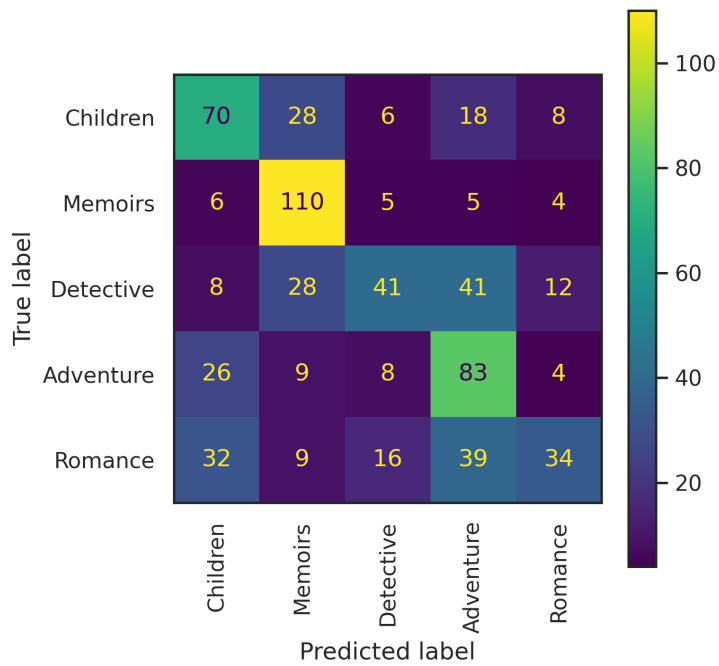


Figure 9: Confusion Matrix for PATIENT features

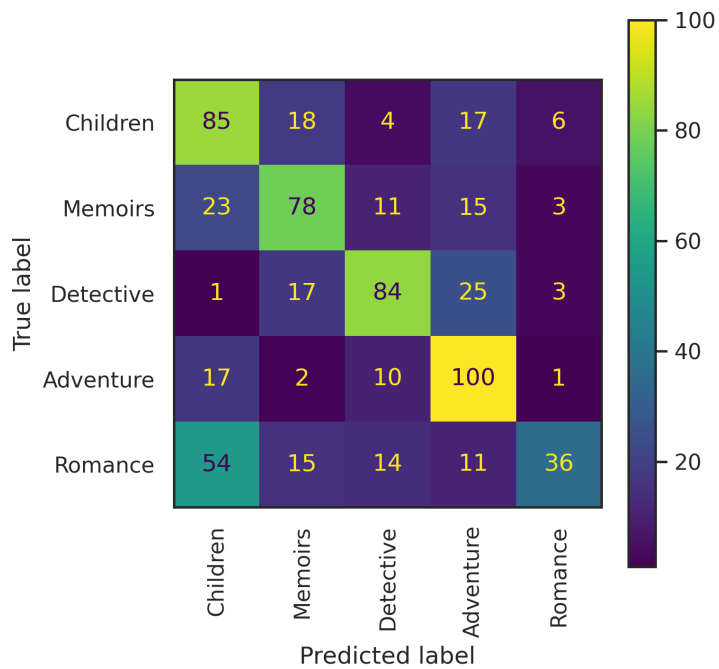


Figure 10: Confusion Matrix for FAC features

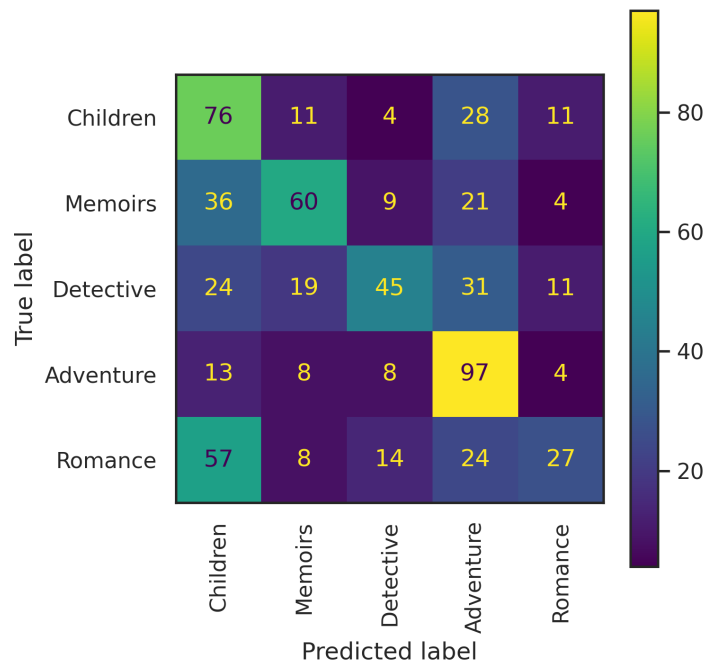


Figure 11: Confusion Matrix for GPE features

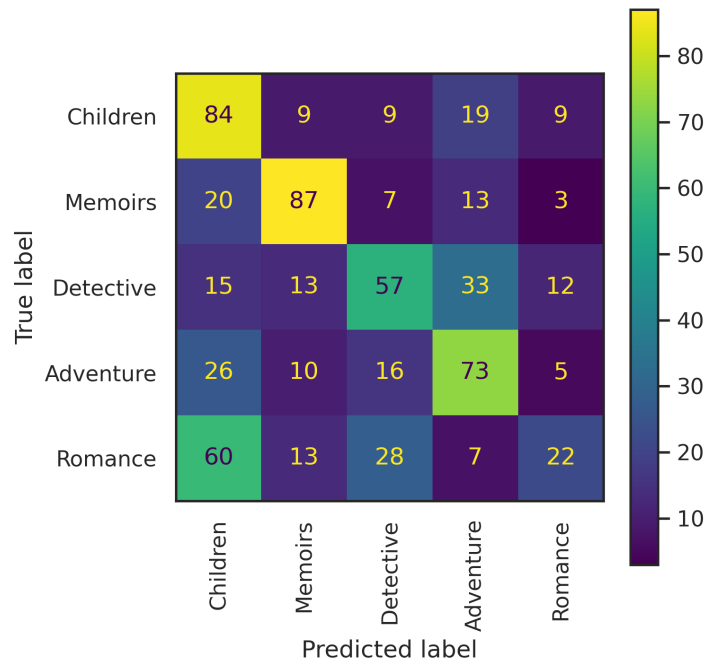


Figure 12: Confusion Matrix for TIME features

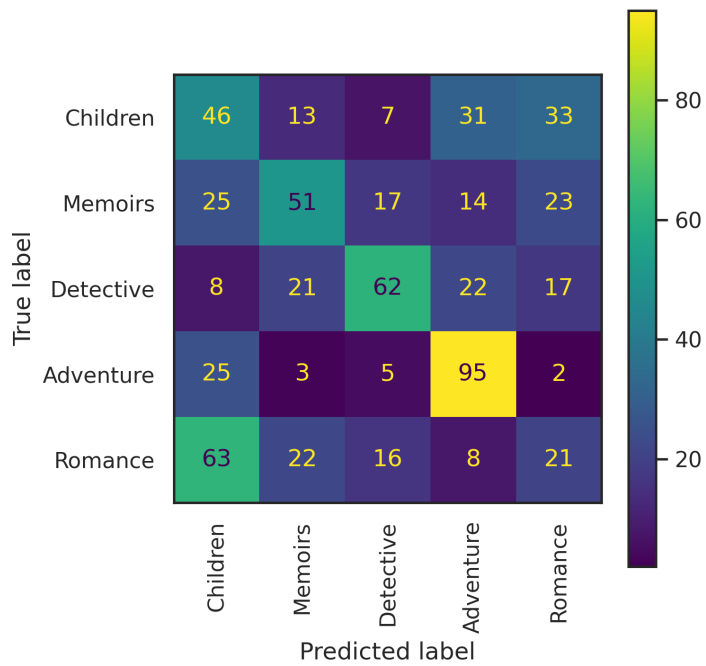


Figure 13: Confusion Matrix for VEH features

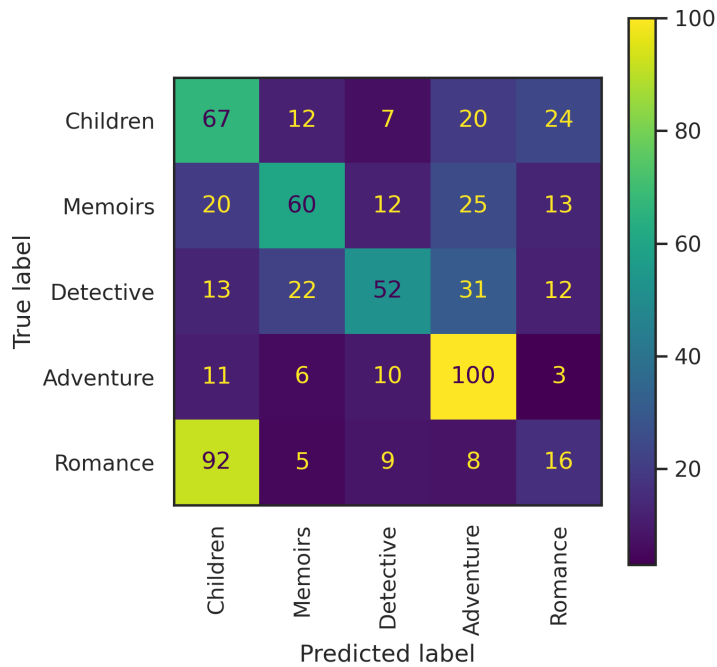








Figure 14: Confusion Matrix for LOC features

Small Worlds

Measuring the mobility of characters in English-language fiction

Matthew Wilkens¹ 
Elizabeth F. Evans²
Sandeep Soni³
David Bamman⁴ 
Andrew Piper⁵ 

1. Information Science, Cornell University , Ithaca, USA.
2. English, Wayne State University , Detroit, USA.
3. Quantitative Theory and Methods, Emory University , Atlanta, USA.
4. School of Information, University of California , Berkeley, USA.
5. Languages, Literatures, and Cultures, McGill University , Montréal, Canada.

Citation

Matthew Wilkens, Elizabeth F. Evans, Sandeep Soni, David Bamman, and Andrew Piper (2024). "Small Worlds. Measuring the Mobility of Characters in English-Language Fiction". In: *CCLS2024 Conference Preprints* (3). [10.26083/tuprints-00027523](https://doi.org/10.26083/tuprints-00027523)

Date published 2024-06-18

Date accepted 2024-04-04

Date received 2024-01-19

Keywords

fiction, mobility, geospatial analysis, narratology

License

CC BY 4.0 

Reviewers

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 3rd Annual Conference of Computational Literary Studies at Vienna, Austria, in June 2024. (Second, updated version)

Abstract. The representation of mobility in literary narratives has important implications for the cultural understanding of human movement and migration. In this paper, we introduce novel methods for measuring the physical mobility of literary characters through narrative space and time. We capture mobility through geographically defined space, as well as through generic locations such as homes, driveways, and forests. Using a dataset of over 13,000 books published in English since 1789, we observe significant "small world" effects in fictional narratives. Specifically, we find that fictional characters cover far less distance than their nonfictional counterparts; the pathways covered by fictional characters are highly formulaic and limited from a global perspective; and fiction exhibits a distinctive semantic investment in domestic and private places. Surprisingly, we do not find that characters' ascribed gender has a statistically significant effect on distance traveled, but it does influence the semantics of domesticity.

1. Introduction

What does it mean for a novel's characters to be mobile? And what effects does spatial mobility have on the novel, the story world it imagines, and the novel's greater cultural significance?

Narrative, especially long narratives, almost always involve a change of location or setting. This is an essential component of what narrative theorists identify as the world-building or world-changing function of narration (Bruner 1991; Herman 2009). Whereas setting was once regarded as the unimportant "background" of fictional narrative, it is now broadly recognized as a vital interface with the material and social world (Evans 2025; Evans and Wilkens 2024; Hones 2022; Ryan et al. 2016; Tally Jr 2012). As Friedman 1998 summarized, "Setting works as symbolic geography, signaling or marking the specific cultural locations of a character within the larger society."

For some genres – the travelogue, the quest narrative, the adventure story, even the Bildungsroman – movement through space is an essential component of the genre’s meaning and identity. The inter-relatedness of space and time in narrative – that the movement through space involves a movement through time – has been influentially theorized by Bakhtin 2010 in the concept of the *chronotope*. For Bakhtin, the space-time nexus has a generative function with respect to narrative.

In this paper, we introduce novel methods by which to measure the physical mobility of characters through narrative space and time. We capture mobility in two distinct ways. First, we define mobility as the movement through geographically-defined space and measure the distance that characters travel between countries, cities, regions, and other mappable places. Second, we examine mobility as movement through the non-geographic semantic spaces of rooms, streets, and other “generic” locations.

The geographic plotting of novels has long been theorized as an important component in the construction of narrative meaning (Moretti 1999; Piatti et al. 2009; Ryan et al. 2016; Wilkens 2013). To take one literary example, the characters of Jack Kerouac’s *On the Road* (1957) travel not only because they want to get from point A to point B (at the novel’s start, New York City to Denver), but also because the road represents to them freedom, discovery, adventure, sex, and, for the narrator, Sal Paradise, creative inspiration. When Sal reflects on his younger self, “I was a young writer and I wanted to take off,” he makes use of the double meaning of “take off” – he wants his writing career to blossom, and he wants to be in motion. The two, and all that being on the road represents to Sal, are necessarily connected: “Somewhere along the line I knew there’d be girls, visions, everything; somewhere along the line the pearl would be handed to me” (Kerouac 2002, 8). For the “girls” Sal and his friends meet along the way, travel is a less-viable choice. While many of them also long for new horizons, women are generally represented by Sal and by the novel as a feature of the landscape, rooted in place, and as lacking in intellectual range as they are in geographic reach. Movement through geographically defined space captures the variety of ideological meanings embedded in mobility, as well as the range of cultural restrictions imposed upon it.

In addition to this focus on geographic space, we also measure movement through what we term “generic space.” For many narratives, mobility may be characterized as a movement between generic spatial entities such as rooms, streets, parks, forests, and homes. In Marilyn Haushofer’s feminist novel *The Wall* (*Die Wand*) from 1963, an invisible wall rises up one day to cut off the unnamed protagonist from the rest of the world. The remainder of the novel involves her moving back and forth between rural hunting lodges and the wall in the Austrian alps. In this case, movement through generic rather than geographically specified space grounds the novel’s reflections on the constraints of female identity, rooting the novel in a more allegorical mode.

Our work is thus tied to prior research in the broader area known as the spatial humanities (Bodenhamer et al. 2010; Roberts et al. 2014). Whether qualitative or computational in nature, this work is grounded in the significance of spatial structures for understanding cultural and narrative meaning. Where prior work often captured space as a static construct (the atlas or map as the principle theoretical frame), the concept of mobility can be a useful addition to this work by taking into account a dimension of narrative time.

Mobility, then, is a way of understanding the world-building function of fictional narratives. How and where characters move through space is integral to the construction of narrative meaning as much as are the specific qualities of the individual places themselves. Modeling mobility at large scale can thus begin to provide insights into the more general chronotopes that shape storytelling across different cultures, genres, and historical time periods.

Questions of narrative mobility – of what mobility is and how we recognize it – also matter when we consider the significance of mobility for human cultures more generally. For Cresswell 2006, “mobility is central to what it is to be human.” Not only do people move from the moment of birth, but cultures blend, splinter, and evolve. And because mobility carries ideological meanings, it also shapes the stories we tell. As Cresswell emphasizes, the modern Western meaning of mobility is not stable: “[m]obility as progress, as freedom, as opportunity, and as modernity, sit side by side with mobility as shiftlessness, as deviance, and as resistance” (1-2). As *On the Road* suggests, the two understandings of mobility can even coexist within a single text. One of the consistent attributes of mobility is its ability to participate in a shifting process of meaning-making. This paper aims to introduce methods for understanding the dynamics of character mobility within literary narratives as part of a broader goal of understanding how mobility has been framed and understood over time.

In the body of our paper, we first describe and validate the model we use to predict narrative mobility derived from prior work (Soni et al. 2023). We then describe a variety of measurements of mobility based on this model as applied to two primary datasets. The first is the CONLIT corpus of contemporary prose, which includes 2,754 works of English prose published since 2001 drawn from twelve different genres. The second is a collection of 10,629 novels by American authors published between 1789 and 2000.

As a way of understanding the function of the different kinds of mobility we are interested in, we examine the relationship between our mobility measurements and particular social categories. These include the effects on character mobility of fictionality (fictional versus nonfictional narratives), prestige (award-winning novels versus bestsellers), audience age-level, and pronoun-signaled character gender.

2. Data and Methods

2.1 Data

We work with a corpus of 13,383 books published between 1789 and 2021. All books are in English; the large majority are works of fiction. The corpus was assembled from a range of sources as described below. The distribution of volumes across subcorpora is shown in table 1.

All subcorpora except CONLIT contain only fiction. As detailed in Piper 2022, CONLIT contains twelve different genres distributed across fiction and nonfiction writing published in the twenty-first century. Nonfiction genres (820 total volumes) are limited to generally narrative forms including biography, memoir, and history. EAF and Wright comprise subsets of the novelistic fiction by US authors cataloged in Wright 1965 and digitized by a consortium of academic libraries (Digital Library Program 2012; Elec-

Collection	Label	Books	Begin	End
Early American Fiction	EAF	488	1789	1850
Wright Bibliography of American Fiction	Wright	1,052	1850	1875
Chicago Novel Corpus I	Chicago I	2,608	1880	1945
Chicago Novel Corpus II	Chicago II	6,481	1946	2000
CONLIT Contemporary Literature	CONLIT	2,754	2001	2021

Table 1: Subdivisions of the research corpus.

tronic Text Center 2000). Chicago I and II include novels by American authors published between 1880 and 2000, sourced from the Chicago Text Lab (Long and So 2020).

Our corpus offers nearly uninterrupted coverage of American fiction over more than 230 years. It is especially rich in twenty-first-century writing, for which it contains extensive metadata concerning fictionality, prestige, and audience type. When we compare fiction to nonfiction, or use metadata facets that are uniquely tabulated for the CONLIT subcorpus, we limit our analysis to that subcorpus. When we analyze fiction alone, we exclude the nonfiction portion of CONLIT. The corpus as a whole does not include a meaningful amount of writing by non-North American authors, nor writing originally published in languages other than English. For this reason, our analysis and conclusions should be understood to apply primarily to the North American, English-language contexts that are well represented in our source collections.

2.2 Methods

2.2.1 Modeling Sequences of Places

From each volume in our corpus, we extract the ordered sequence of locations associated with each of its characters using the method developed in Soni et al. 2023. In brief, we use BookNLP (Bamman 2020, 2021) to identify characters and locations that coöccur within a rolling ten-token window in each source text. The same system performs coreference resolution, consolidates multiple forms of address to single characters, and records pronominally signaled character genders. We then train a BERT-based model to identify possible relationships (including NO RELATION) between each coöccurring character–location pair. From the full set of coöccurrences, we select those that describe a character as occupying the identified location (having relation IN). This method differs significantly from earlier work, in that it allows us both to place characters in specific locations and to trace character movements over narrative sequences.

The locations identified may be geopolitical entities (GPEs), such as nations or cities, facilities (FACs), such as homes or offices, or other locations (LOCs; typically natural settings). In principle, any of these locations might correspond to real, mappable places (England, Mt. Everest) or to imaginary or generic entities (the house, a street corner, Hogwarts). In practice, most GPEs are real, uniquely identifiable, and mappable; most FACs and LOCs are not.¹ We separate our character sequences into GPEs and others. For GPEs, we retrieve detailed geographic information from open and commercial sources as described in Evans and Wilkens 2018. For non-GPEs, we remove stopwords ([the

1. We resolve coreferences to characters, but not to locations. We thus do not attempt to map diectics such as “here” or “there” to any specific place, nor do we identify whether any two instances of a generic term like “house” refer to the *same* house.

house | a house | her house] → house), but do not perform geolocation. 133

After processing, we have two lists of locations (GPEs and others, respectively) that are 134
occupied sequentially by each character in each book. In some of our experiments, we 135
are interested in transitions between locations. We call each case in which a character 136
occupies a location different from the one immediately preceding it a *hop*. For example, 137
a character having the GPE sequence [London, Boston, California] undergoes two hops, 138
London → Boston and Boston → California. If a character occupies the same location 139
multiple consecutive times, we treat that sequence of unchanging locations as single 140
instance. For GPE sequences, we exclude hops for which the distance between locations 141
is conceptually ill-defined, such as London → England or California → USA. 142

2.2.2 Measurements 143

Here we present the primary measures used in our analysis, along with a list of de- 144
pendent variables analyzed in table 5. In most cases, we restrict our calculations to the 145
single most commonly occurring character in each book, which we call the *protagonist*. 146
We condition on protagonists because we observe that the majority of overall mobility 147
in the average book is associated with the most frequently occurring character. 148

Distance: The total geodesic distance (in miles) between sequences of geographic places 149
(GPEs) that are inhabited by the book’s protagonist. This represents the sum of the 150
distances traversed over all valid hops for the character. We exclude a subset of common 151
hop types that are conceptually ill-defined, including hops between cities and the first- 152
level administrative regions (states, provinces, etc.) or nations that contain them, and 153
between first-level regions and the nations to which they belong. We allow hops between 154
any locations at the same administrative level (city to city, state to state) and between 155
different administrative levels when the lower-level location is not contained by the 156
higher-level one (for example, neither Los Angeles → California nor Los Angeles → 157
United States is allowed, but Los Angeles → Iowa is). We make an exception for hops 158
involving continents, which we allow (measuring to the geographic centroid of the 159
continent). 160

GPEs: The count of distinct geographic places inhabited by the main character (e.g., 161
India, Toronto, New York, California). 162

Generics: The count of distinct generic places inhabited by the main character (e.g., 163
room, kitchen, street, yard). These are annotated as LOC and FAC by BookNLP. 164

Semantic distance: The average semantic distance between all sequentially inhabited 165
generic places. Semantic distance is calculated as one minus the cosine similarity 166
between word vectors for each generic place using the Glove 6B Wikipedia pretrained 167
model with 100 dimensions (Pennington et al. 2014). Multi-word phrases average 168
each word’s vector in the phrase. Stop words and punctuation are removed. Semantic 169
distance aims to capture the semantic similarity of places given a general understanding 170
of those terms. 171

Deictics: The frequency of “here” and “there” relative to all generic place names per 172
book. 173

Generic / GPE ratio: The total number of generic locations divided by the total number 174

of GPEs per book. 175

Character count: The count of references to a book's protagonist. 176

Tokens: The total count of tokens per book. 177

Start–finish miles: The direct geodesic distance between the first and last locations inhabited by the protagonist of each book. 178
179

2.2.3 Independent Variables used for CONLIT 180

The number of documents for each class are listed in parentheses. 181

Fictionality: The category designation between FIC (fiction; 1,934 volumes) and NON (nonfiction; 820). 182
183

Prestige: Sub-divided between genre labels PW (prizewinners; 258) for high prestige and BS (bestsellers; 249) for low prestige. 184
185

Youth: Sub-divided between genre labels MID (middle-grade books; 166) and NYT (New York Times reviewed), PW, and BS (926). 186
187

Female: Uses the inferred gender categories “she/her/hers” (744) and “he/him/his” (1,180) for protagonists in fiction. The very small number of other pronominal designations are removed. 188
189
190

2.2.4 Distance Validation 191

The computational pipeline by which we produce our hop sequences and distance measurements is complex and subject to multiple uncertainties. To validate our results, we examined 10,000-word chunks extracted from the beginning of 30 novels sampled at random from the CONLIT subcorpus. For each sample, we annotated by hand the set of true geographic locations occupied by the main character; determined the geographic coordinates of those locations; and calculated the distance traversed by that character. We also labeled each sample's holistic mobility from 1 (lowest mobility) to 5 (highest mobility). We found that our algorithmic distance was linearly correlated with human measurements at $R^2 = 0.525$ ($p \approx 0$ by permutation against a null hypothesis of no relationship between the measurements). We also found that the mean distance traveled by protagonists in high-mobility samples (those with ratings of 4 or 5) was much higher than the mean distance traveled in low-mobility samples (ratings 1 or 2; $\bar{x}_{high}/\bar{x}_{low} = 3.6$; $p < 0.008$ by permutation of the group labels against a null hypothesis of no difference in the group means). We note as well that randomly distributed errors in our pipeline will tend to reduce the observed significance of results derived from our data, hence that we generally understate the statistical significance of our findings (see Spearman [1904] 1987). We are thus confident that our GPE-derived distance measures serve in aggregate as an acceptable class of proxies for character mobility. 192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209

2.2.5 Regression Analysis 210

To evaluate the impact of each social category, which serve as our independent variables, we conducted a linear regression analysis. For this analysis, we incorporated binary dummy variables corresponding to each primary class, namely fiction, prestige, youth, 211
212
213

and female character. Additionally, we introduced control variables to account for potential confounding factors, such as genre, point of view, book length (measured in tokens), and character mention frequency (character count).

The outcomes of this analysis, including the directionality of the effect for each dependent variable and the statistical significance represented by p -values, are summarized in table 5. In our supplementary materials, we present comprehensive results, encompassing sample mean estimates, R^2 values, and the precise p -values obtained from the analysis.

It is important to acknowledge the significance of our chosen control variables due to the variability they exhibit in our data. For instance, nonfiction texts exhibit a higher average length compared to fiction, whereas fiction registers a markedly higher average character count, with fictional protagonists being referenced significantly more frequently. Consequently, employing a uniform normalization technique would be inadequate to address the multifaceted disparities inherent in our dataset.

3. Results

Overall Distance. In table 2, we show the mean distance traveled, mean number of unique GPEs, and mean number of unique generic locations in each of our subcorpora.² Figure 1 visualizes the evolution in these quantities over time. As we can see, the average number of unique places, whether GPE or generic, has more than doubled since the nineteenth century, as has the total distance traveled by primary characters.

Collection	Distance	GPEs	Generics	Hops
EAF	13,139	5.9	37.5	5.8
Wright	10,477	5.3	43.8	4.9
Chicago I	21,026	8.4	72.9	9.3
Chicago II	37,023	13.8	113.0	16.3
CONLIT fiction	38,024	13.3	123.9	15.6
CONLIT nonfiction	131,263	35.8	120.8	60.8

Table 2: Means of distance, number of unique GPEs, number of unique generic locations, and number of hops by subcorpus.

Routes Traveled. Figure 2 presents a global map capturing the movement by protagonists between places in fictional narratives. This figure plots the aggregate hops taken by all fictional protagonists over the full corpus; the width of the line connecting each (undirected) origin and destination is proportional to the share of all hops represented by that location pair. While we visualize here only the aggregated results for the full corpus, the supplemental materials provide visualizations by subcorpus and by historical era. There is very little variation in the high-level appearance of this map over historical time. As table 3 further illustrates, the patterns of movement between places within (broadly American) fiction are highly stable and formulaic over historical time.

Gender and Mobility. Previous work has found that novels enriched in she/her charac-

2. Median values of these quantities are lower, since their distributions include a long tail of large values, but the observed historical trends and relationships between subcorpora do not differ meaningfully under that metric. The same is true of the total (as opposed to unique) number of GPEs and generic location mentions. Full results are available in the supplementary material.

Measuring the mobility of characters

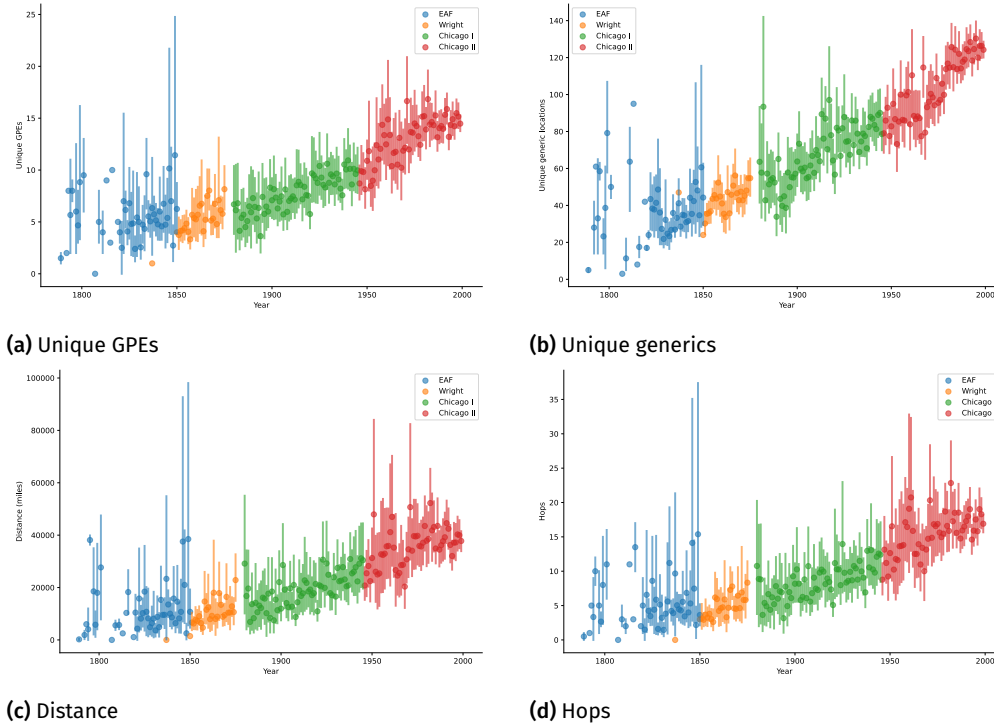


Figure 1: Unique GPEs, unique generic locations, protagonist distance, and hop count over time by subcorpus and year. Markers represent yearly means; bars are 95% confidence intervals.

GPEs	Most frequent hops
New York	America*, Paris, Manhattan*, London, New York City*, Chicago, California, Brooklyn
London	New York, England*, Paris, America, France, Boston
America	New York*, London, England, California*, Paris, China, India
Paris	France*, New York, London, Chicago, England, Europe
California	New York, Los Angeles*, San Francisco*, America*, Chicago, London, San Diego*, Boston
Generics	Most frequent hops
room	house, home, kitchen, bedroom, school
house	room, home, kitchen, living room, bedroom
home	house, room, kitchen, school, apartment
kitchen	house, room, home, living room, bedroom

Table 3: Most frequent inhabited locations in the fiction facet of CONLIT, followed by the most frequent subsequent locations (“hop”) in descending order of frequency. Destinations marked with an asterisk (*) are examples of hops excluded from distance calculations, because their distance from the origin is ill-defined. Such hops are common.

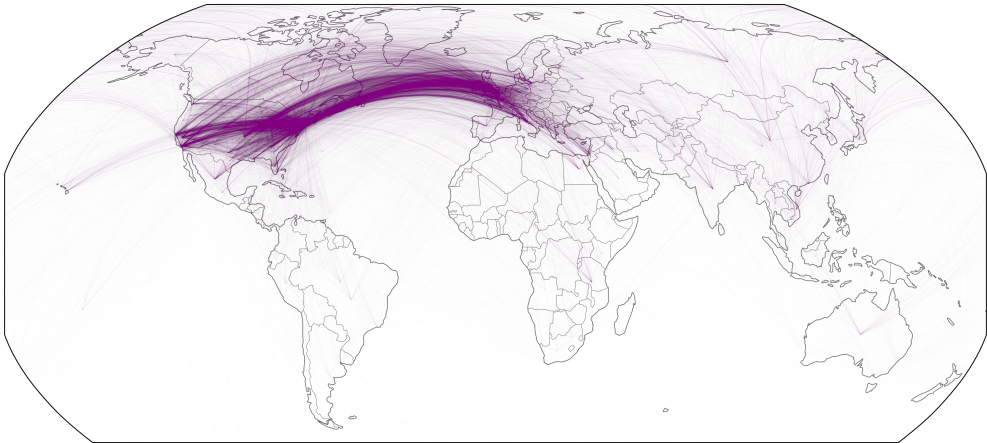


Figure 2: Aggregated character hops in the corpus. Line widths are proportional to the total number of hops between each pair of locations.

ters contain fewer GPEs and that the GPEs in those narratives are less widely separated 244
than are those in he/him-enriched novels (Evans and Wilkens 2024). As shown in table 245
4, we calculate the mean distance traveled and the count of unique GPEs and generics by 246
pronominally indicated character gender. We find over the full corpus that the average 247
male-gendered protagonist in fiction occupies more unique GPEs, fewer unique generic 248
locations, and covers slightly more ground than does the average female-gendered 249
protagonist. But, surprisingly, the difference in distance traveled is not statistically 250
significant either in aggregate or within the individual subcorpora. 251

Feature	she/her	he/him	<i>p</i>
Distance (miles)	29,943	31,134	0.1990
Unique GPEs	11.08	11.85	0.0008 ***
Unique generics	102.0	95.8	0.0008 ***

Table 4: Key mobility metrics by narrativized character gender in fiction in the full corpus.

Social Effects on Mobility. Focusing specifically on the contemporary data, we measure 252
the effects of different social categories on character mobility using the regression 253
models described above. As shown in table 5, we find that both fictionality and intended 254
audience age-level have the strongest negative association with mobility, i.e., both 255
categories significantly lower the distance traveled and the frequency of place names 256
mentioned (both GPE and generic). We also observe a greater reliance on generic place 257
names in both of these categories. Finally, as with the full corpus, we find that, after 258
controlling for genre-related factors, there is no meaningful difference in the distance 259
traveled between differently gendered characters. 260

In addition to our regression analysis, we also seek to identify ways in which mobility 261
may differ *qualitatively* even when overall quantitative levels are similar. We employ the 262
Fightin’ Words method of Monroe et al. 2017 with an informative prior to identify GPEs 263
and generic places that are over- and underrepresented in facets of our corpus (figure 264
3).³ 265

3. Specifically, we use the method described in Monroe et al. 2017, section 3.5.1, equation 23, with an informative Dirichlet prior calculated over all volumes in the corpus.

Measure	Fictionality		Prestige		Youth		Female	
	valence	<i>p</i>	valence	<i>p</i>	valence	<i>p</i>	valence	<i>p</i>
Distance	-	***	+	.	-	***	+	.
GPEs	-	***	-	.	-	***	+	.
Generics	-	***	+	.	-	***	+	***
Semantic distance	-	*	+	***	+	.	-	**
Deictics	+	***	-	***	+	.	-	.
Generic/GPE ratio	+	***	+	.	+	***	+	.

Table 5: Results of regression analysis for each measure across our primary categories in the CONLIT subcorpus. Valence captures whether the estimate for the primary category (e.g. fictionality) is lower or higher than its opposite (e.g. nonfictionality). We provide standard significance codes (*** < 0.001, ** < 0.01, * < 0.05, . ≥ 0.05). Full results, including the estimates and R^2 values, are supplied in the supplementary material.

We observe that contemporary fictional narratives are often enriched in imaginary, extraterrestrial, historical, and otherwise “peripheral” GPEs (Maine, Taos, Sri Lanka) relative to nonfictional narratives, which are themselves enriched in sites of political power and armed conflict. Fiction is also enriched in generic locations that are private and semi-public interior spaces, whereas nonfiction preferentially locates its characters in public sites of power and work.

Within fiction, we find that she/her characters are distinctively located in major and evocative urban localities; he/him characters are assigned preferentially to historical and contemporary sites of power and to those of American political and armed conflict. Generic locations are distributed by gender in ways that resemble their allocation between fiction and nonfiction, she/her characters occupying domestic interiors, he/him characters disproportionately found in public, power-infused sites.

4. Discussion

Our results paint a clear picture of the spatial constraints of fictional worlds. When compared with nonfictional narratives, characters in contemporary fiction travel less distance, visit fewer geographic and generic places, inhabit generic places that are semantically more similar to each other, and rely far more on generic places than on geographic ones. They also utilize deictic markers like “here” and “there” with far greater frequency. Fictional worlds are smaller worlds, both geographically and semantically.

Interestingly, we see little effect on these measures if we examine social categories like prestige or gender. Prizewinning novels do not travel further or utilize more geographic places when compared to more market-driven fiction. They do tend to use fewer deictics and employ more semantic diversity among non-geographic places, suggesting greater sophistication at the level of vocabulary. Books aimed at middle-school audiences generally describe far more limited narrative worlds, as would be expected.

The results concerning character gender are surprising, given our assumption that she/her characters would more likely be associated with social constraints affecting their mobility. This turns out not to be the case. For both the historical and contemporary data, women were no more likely to be associated with diminished levels of mobility after controlling for confounding variables..

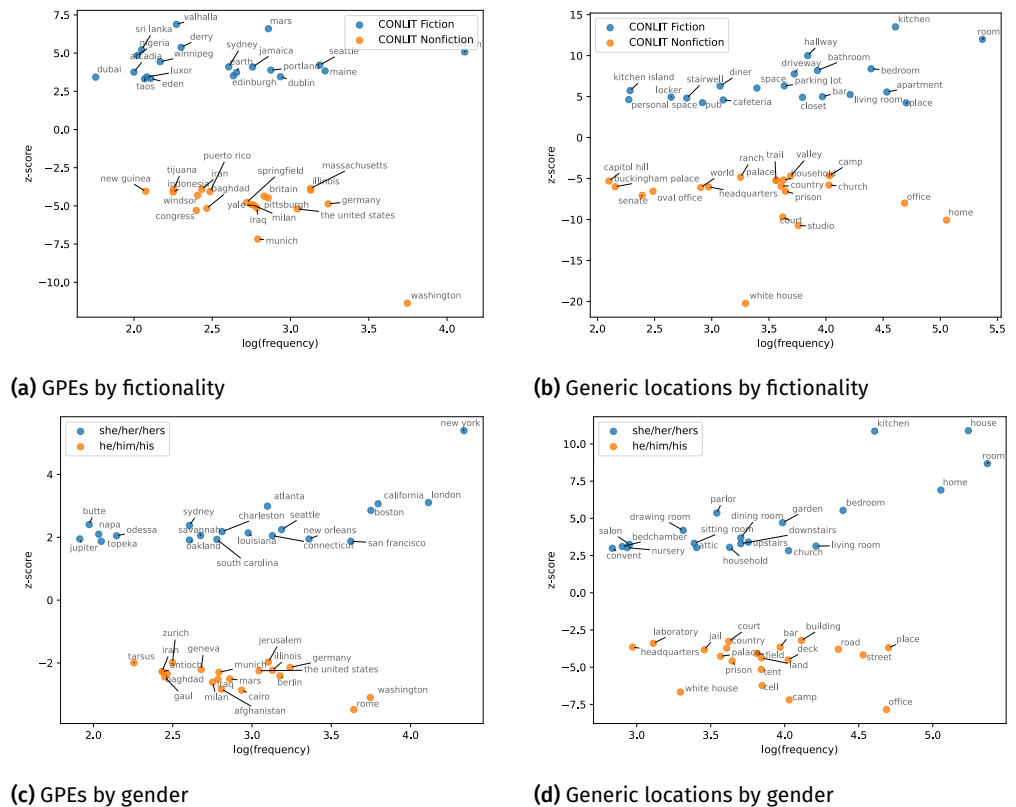


Figure 3: Distinctive location use across fictionality and character gender facets in CONLIT. The x -axis represents the log of the frequency of each term in the indicated corpus; the y -axis represents the z -score of the term in the indicated facet relative to the other facet, informed by a weighted prior calculated over the full corpus.

At the same time, when we examine the distinctive places associated with she/her 296
characters, we do see more expected outcomes. She/her characters are more likely 297
than he/him characters to be associated with domestic, private, and semi-public spaces. 298
If we compare the results for fiction and nonfiction presented in figures 3a and 3b to 299
those for character gender in figures 3c and 3d, we see how the locations distinctively 300
occupied by she/her and he/him characters map closely to those of fiction and nonfiction 301
protagonists, respectively. While we are not yet in a position to assert a blanket spatial 302
homology between fictionality and gender, the resemblance is sufficiently suggestive to 303
merit further investigation. 304

In addition to these small-world effects at the level of physical distance, we also find that 305
the *connections* between geographic places in fictional worlds are remarkably predictable 306
(figure 2). Fictional worlds are “small” not just in the sense of the overall distance 307
characters travel, but also in the diversity of places they move between. We observe 308
a NATO- or grand-tour-driven center surrounded by a much less traveled periphery. 309
Fictional characters spend their time moving around a very small portion of the world. 310

These results accord well with previous work that examined the distribution of named 311
locations (without regard to character associations) in British and American fiction 312
(Wilkins 2016), though there exists some evidence suggesting that British fiction under- 313
went greater evolution of its geographic imagination over the twentieth century than 314
did American (Wilkins 2021). Future work could begin to replicate these methods for 315

more geographically diverse fiction produced around the world to model the spatial
archetypes of mobility. Does every region or national literature have its spatial center
of gravity and its exotic periphery? To what extent are centers and peripheries shared
across nations, languages, and periods? Is every regional literature as constrained as the
North American example, or do other regions have very different network structures of
mobility?

When it comes to changes in mobility over historical time, we see that the distance
traveled by fictional characters has been increasing, as have the number of GPEs and
generic places. One of the drivers of this phenomenon is that fictional narratives have
also been getting longer over time, while the frequency of references to the main character
has been increasing as well.⁴ If we normalize by book length, we still see meaningful
increases over time; if we normalize by character count (that is, by the number of all
character references that pertain to the protagonist), we see slower growth in distance
traveled and essentially zero rise in the count of unique GPEs (figure 4). The same is true
when we compare highly protagonist-centered first-person narratives to more widely
character-dispersed third-person alternatives. What this tells us is that, as books have
become longer and more protagonist-centered, main characters are traveling relatively
further and moving between geographic places more often, but much of this growth can
be accounted for by the sheer increase in character references (allowing for more places
to be counted and thus more distance to be traveled). There does not appear to be an
obvious ceiling on the range or rate of protagonist mobility, even in long books with
potentially saturated story worlds. That said, we are surprised that, over a sustained
period of increasing access to fast, safe, and reliable transportation, we do not observe
more sharply rising distances traveled by protagonists after controlling for narrative
length and protagonist concentration. This fact may suggest narrative constraints on the
density or variety of geographic locations that can be easily accommodated in long-form
fiction.

The final way in which we understand the small-world effect of fiction is through our
examination of the lexical differences between spatial entities in fiction when compared
with nonfiction (figure 3). When we do so, we quickly confirm several differences
that we might have expected, but have not previously quantified. Compared to fiction,
nonfictional narratives overrepresent sites of power, including official political locations
like White House, Oval Office, Senate, Washington, Buckingham Palace (and “palace”
generically), and Capitol Hill; sites of carceral power (court, prison); workplaces (studio,
office, headquarters); and locations of present and historical conflict as experienced
primarily from the United States (Baghdad, Iraq, Iran, Munich, Tijuana). Fiction, by
contrast, overrepresents domestic and semi-public spaces (kitchen, hallway, bedroom,
bathroom, apartment, cafeteria, pub, and many more), driveways, and parking lots. As
has long been theorized, fiction is preëminently occupied with domestic and private
space (Armstrong 1987; McKeon 2006).

On the other hand, the distinctive geographic spaces of fiction are often extremely distant
or otherworldly (Valhalla, Mars, Arcadia, Eden). Fiction compensates for its small-
world effects – either in the real world or through generic private spaces – by investing

4. We note in passing that these measures of average book length and protagonist concentration over nearly 250 years of North American literature are novel in the critical and computational literature. They likely merit future investigation.

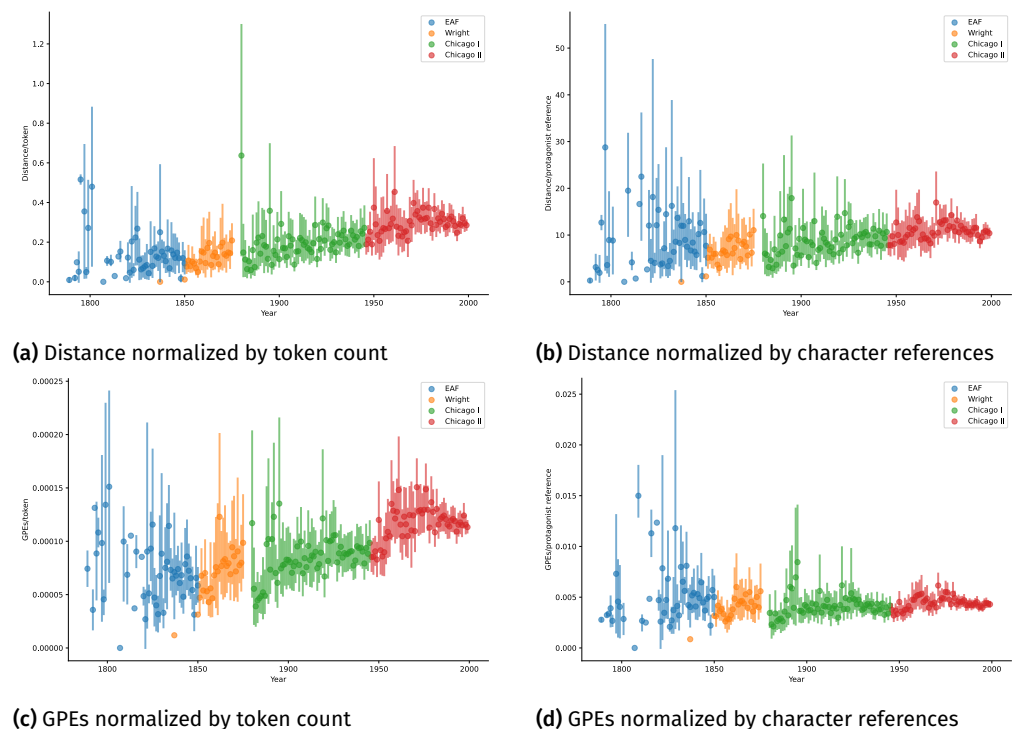


Figure 4: Average fictional protagonist distance and count of unique GPEs by year and subcorpus, normalized by volume length or by count of character references.

at least partially in telling narratives focused on the most distant places imaginable.⁵ 359
It is worth considering what a new genre of fiction might look like that inverted this 360
escapism–power dynamic and focused instead on immersing readers in the central 361
locales of power and punishment rather than the private chambers of imaginary locales. 362

The major limitation of our study, beyond the need for cultural expansion, is that our 363
models cannot account for distances between unreal places or extraterrestrial locations, 364
which are identified by our entity model, but are not easily localizable in terrestrial 365
space. One could argue that the role of genres like fantasy and science fiction is precisely 366
to undo the small-world effects of fiction (Dubourg and Baumard 2022). In simulating 367
vast travel, they reverse the constraints of fictionality. At the same time, the fact that we 368
see these genres still exhibiting lower diversity of generic places and higher semantic 369
constraints between them relative to nonfictional narratives suggests a basic conflict 370
between the expansiveness of space (“to the moon and back”) and the constraints of 371
fictional places that are limited to rooms, vehicles, and home-like structures. 372

5. Conclusion 373

Our project has attempted to add two important methodological dimensions to prior 374
research on literary spaces. First, relying on new models that locate characters in space 375
(Soni et al. 2023), we are able to give a *character-centred* account of fictional spaces. 376
Second, by studying the sequencing of spatial presence we are able to observe the effects 377
of narrative time on the construction of space, for which we employ the term “character 378

5. We say at least partially because these are not the most common locations in contemporary fiction (which are all-too-familiar places like New York, London, and America). Rather, these are the locations that are present at modest rates in fiction and that are virtually absent from works of nonfiction.

mobility.” 379

Applying our models to a large collection of historical and contemporary Anglophone fiction, we make the following key observations concerning the small-world effects of fiction: 380
381
382

1. **Fictional worlds are small in the sense of the distance traveled by characters.** 383
When compared to the movements of nonfictional characters (subjects of memoirs, biography, or historical narratives), fictional protagonists travel less than half the distance of their nonfictional counterparts. Generic places are also much more common and far more semantically similar than is the case in nonfiction. 384
385
386
387
2. **Fictional worlds are small in the constrained routes that characters travel.** Fic- 388
tional characters stick to a very familiar set of pathways that leave much of the world un- or under-explored. 389
390
3. **Fictional worlds are semantically small in the types of generic spaces they foreground.** Fictional characters are much more likely to be located in domestic or private spaces when compared to their nonfictional counterparts. 391
392
393
4. **Fictional worlds have been expanding over historical time.** The distance traveled by fictional characters has doubled since the nineteenth century, but much of this increase can be accounted for by the increased centralization of main characters. 394
395
396
5. **She/her characters do not move less, but they do spend more time in the kitchen.** 397
Insights into the gendered nature of mobility reject assumptions about the spatial limitations of women characters, but support their over-representation within domestic spaces. 398
399
400

We look forward to continuing this work to gain a deeper and more culturally diverse understanding of the relationship between fictional narratives and character mobility. 401
402

6. Data Availability 403

Data and supplementary materials are available at <https://github.com/wilkens/sma-ll-worlds> 404
405

7. Acknowledgements 406

The authors thank Yasmine Chim for her assistance compiling validation data. The research reported in this article was supported by funding from the National Science Foundation (IIS-1942591, to DB) and the National Endowment for the Humanities (HAA-271654-20, to DB; HAA-290374-23, to MW). 407
408
409
410

8. Author Contributions 411

Matthew Wilkens: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, validation, visualization, writing - original draft 412
413

Elizabeth F. Evans: formal analysis, writing – review & editing 414

Sandeep Soni: methods, data analysis, software	415
David Bamman: funding acquisition, methods, resources	416
Andrew Piper: conceptualization, data curation, formal analysis, project administration, investigation, writing – original draft	417 418

References 419



Armstrong, Nancy (1987). <i>Desire and domestic fiction: A political history of the novel</i> . Oxford University Press.	420 421
Bakhtin, Mikhail Mikhailovich (2010). <i>The dialogic imagination: Four essays</i> . University of Texas Press.	422 423
Bamman, David (2020). “LitBank: Born-Literary Natural Language Processing”. In: <i>Computational Humanities</i> . Ed. by Jessica Marie Johnson, David Mimno, and Lauren Tilton. Debates in the Digital Humanities.	424 425 426
— (2021). <i>BookNLP. A natural language processing pipeline for books</i> . https://github.com/booknlp/booknlp . Accessed: 2022-01-30.	427 428
Bodenhamer, David J, John Corrigan, and Trevor M Harris (2010). <i>The spatial humanities: GIS and the future of humanities scholarship</i> . Indiana University Press.	429 430
Bruner, Jerome (1991). “The narrative construction of reality”. In: <i>Critical inquiry</i> 18.1, 1–21.	431 432
Cresswell, Tim (2006). <i>On the move: Mobility in the modern western world</i> . Taylor & Francis.	433
Digital Library Program (2012). <i>Wright American Fiction</i> . Tech. rep. Indiana University Libraries. https://webapp1.dlib.indiana.edu/TEIgeneral/welcome.do?brand=wright .	434 435 436
Dubourg, Edgar and Nicolas Baumard (2022). “Why imaginary worlds? The psychological foundations and cultural evolution of fictions with imaginary worlds”. In: <i>Behavioral and Brain Sciences</i> 45, e276.	437 438 439
Electronic Text Center (2000). <i>Early American Fiction Collection</i> . Tech. rep. University of Virginia Library. https://jti.lib.virginia.edu/eaf/ .	440 441
Evans, Elizabeth F., ed. (2025). <i>Cambridge Critical Concepts: Space and Literary Studies</i> . Cambridge University Press.	442 443
Evans, Elizabeth F. and Matthew Wilkens (2018). “Nation, Ethnicity, and the Geography of British Fiction, 1880-1940”. In: <i>Journal of Cultural Analytics</i> 3.2. 10.22148/16.024.	444 445
— (2024). <i>Gender and Literary Geography</i> . Cambridge University Press.	446
Friedman, Susan Stanford (1998). <i>Mappings: Feminism and the cultural geographies of encounter</i> . Princeton University Press.	447 448
Herman, David (2009). <i>Basic elements of narrative</i> . John Wiley & Sons.	449
Hones, Sheila (2022). <i>Literary geography</i> . Taylor & Francis.	450
Kerouac, Jack (2002). <i>On the Road</i> . Penguin Classics.	451
Long, Hoyt and Richard Jean So (2020). <i>US Novel Corpus</i> . Tech. rep. University of Chicago Textual Optics Lab. https://textual-optics-lab.uchicago.edu/us_novel_corpus .	452 453 454
McKeon, Michael (2006). <i>The secret history of domesticity: Public, private, and the division of knowledge</i> . JHU Press.	455 456

- Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn (2017). "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict". In: *Political Analysis* 16.4, 372–403. [10.1093/pan/mpn018](https://doi.org/10.1093/pan/mpn018).
- Moretti, Franco (1999). *Atlas of the European novel: 1800-1900*. Verso.
- Pennington, Jeffrey, Richard Socher, and Christopher D Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.
- Piatti, Barbara, Hans Rudolf Bär, Anne-Kathrin Reuschel, Lorenz Hurni, and William Cartwright (2009). "Mapping literature: Towards a geography of fiction". In: *Cartography and art*. Springer, 1–16.
- Piper, Andrew (2022). "The CONLIT dataset of contemporary literature". In: *Journal of Open Humanities Data* 8.
- Roberts, Les, Thomas Thevenin, Julia Hallam, Andrew Beveridge, Ruth Mostern, Humphrey Southall, Niall A. Cunningham, Robert M Schwartz, and Elijah Meeks (2014). *Toward spatial humanities: Historical GIS and spatial history*. Indiana University Press.
- Ryan, Marie-Laure, Kenneth Foote, and Maoz Azaryahu (2016). *Narrating space/spatializing narrative: Where narrative theory and geography meet*. The Ohio State University Press.
- Soni, Sandeep, Amanpreet Sihra, Elizabeth Evans, Matthew Wilkens, and David Berman (2023). "Grounding Characters and Places in Narrative Text". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, 11723–11736. [10.18653/v1/2023.acl-long.655](https://doi.org/10.18653/v1/2023.acl-long.655).
- Spearman, Charles [1904] (1987). "The Proof and Measurement of Association between Two Things". In: *The American Journal of Psychology* 100.3/4, 441–471.
- Tally Jr, Robert (2012). *Spatiality*. Routledge.
- Wilkens, Matthew (2013). "The geographic imagination of Civil War-era American fiction". In: *American Literary History* 25.4, 803–840.
- (2016). "The Perpetual Fifties of American Fiction". In: *Neoliberalism and Contemporary Literary Culture*. Ed. by Mitchum Huehls and Rachel Greenwald-Smith. Baltimore: Johns Hopkins UP, 181–202.
- (2021). "'Too isolated, too insular': American Literature and the World". In: *Journal of Cultural Analytics* 6.3. [10.22148/001c.25273](https://doi.org/10.22148/001c.25273).
- Wright, Lyle Henry (1965). *American Fiction, 1851-1875: A Contribution toward a Bibliography*. Revised. The Huntington Library.

A Stylometric Analysis of Seneca's disputed plays

Authorship Verification of *Octavia* and *Hercules Oetaeus*

Paschalis Agapitos¹ 
Andreas van Cranenburgh² 

1. P. M. de Lardizabal 4, Donostia International Physics Center , Donostia/San Sebastian, Spain.
2. Computational Linguistics Department, University of Groningen , Groningen, The Netherlands.

Citation

Paschalis Agapitos and Andreas van Cranenburgh (2024). "A Stylometric Analysis of Seneca's Disputed Plays. Authorship Verification of *Octavia* and *Hercules Oetaeus*". In: *CCLS2024 Conference Preprints* 3 (1). [10.26083/tuprints-00027394](https://doi.org/10.26083/tuprints-00027394)

Date published 2024-05-28

Date accepted 2024-04-04

Date received 2024-01-22

Keywords

Seneca, stylometry, authorship verification, Latin, Stylo

License

CC BY 4.0 

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 3rd Annual Conference of Computational Literary Studies at Vienna, Austria, in June 2024.

Abstract. Seneca's authorship of *Octavia* and *Hercules Oetaeus* is disputed. This study employs established computational stylometry methods based on character n-gram frequencies to investigate this case. Based on a Principal Component Analysis (PCA) of stylistic similarities within the Senecan corpus, *Octavia* and *Phoenissae* emerge as outliers, while *Hercules Oetaeus* only stands out when the text is split in half. Subsequently, applying Bootstrap Consensus Trees (BCT) to a corpus of distractor texts, both disputed plays align with the Senecan cluster/branch. The General Impostors method confidently reports Seneca as the author of the disputed plays under various scenarios. However, upon closer examination of text segments, indications of mixed authorship arise. Based on computational stylometry, it appears that the disputed plays were in large part, but not wholly, written by Seneca.

1. Introduction

Computational stylometry is a quantitative text analysis method mostly concerned with authorship attribution and authorship verification problems. Authorship attribution involves identifying the most likely author of a disputed document from a give set of candidates (Koppel et al. 2007, 1261). Authorship verification concerns the question of whether an author wrote a disputed document (Koppel et al. 2007, 1261; Juola 2015, 1106). The verification task is more challenging than the attribution task, because the verification task involves determining whether an observed similarity in style is sufficient to verify authorship, while the attribution task merely involves picking the most similar author from the given candidates (Potha and E. Stamatatos 2017, 138). It is important to also note that the authorship verification typically involves both close-set and open-set scenarios. In the close-set scenario, the suspected author is one of the candidates provided, whereas in the open-set scenario, the true author may not be among the known candidates.

The main assumption behind computational stylometry is that certain words are chosen unconsciously by the writer, which form a unique, individual fingerprint of an author (Evert et al. 2017, ii4). Since these words are predominantly function words that are used in a way that is hard for the author to control, imitating someone else's writing style is difficult for an impostor. In other words, there is an "immutable signal that authors

emit involuntarily" (Päpcke et al. 2022, 1). The utility of function words in traditional and computation stylometric studies can be condensed into four points: richer dataset because of their high frequency, closeness of the set since function words are limited and fixed, content-independent, and, as mentioned above, unconscious use of them due to their high frequency (Kestemont 2014, 60; Beullens et al. 2024, 393–394).

The aim of this article is to examine whether Seneca the Younger wrote *Octavia* and/or *Hercules Oetaeus* (henceforward: *Oct.* and *H.O.*, respectively), since they are both tragedies of which a plethora of literary scholars have raised concerns about their attribution to Seneca. We aim to contribute to the debate on Seneca's disputed texts by applying a variety of computational stylistic methods and testing several different scenarios. We do this using the Stylo software, an R-package created and developed by Eder et al. (2016).

The ensuing sections of this study are organized as follows. Initially, a concise literature review is provided addressing *Oct.* and *H.O.* (Section 2). Subsequently, Section 3 outlines the rationale for selecting a specific set of impostor texts and acknowledges potential limitations associated with the limited transmission of ancient texts and differences in genre and meter. Section 4 delves into the preprocessing steps and features employed in the study, while also offering a brief explanation of each method utilized in the primary analysis. Section 5 provides a validation of the methods on texts with known authorship. Section 6 presents the main results for the disputed texts and engages in a discussion of these findings. Finally we present our conclusions concerning the findings and outline ideas for future research (Section 7).

2. Literature Review

2.1 Non-quantitative Approaches

The disputed texts considered in this article, *Oct.* and *H.O.*, are Latin tragedies; *Oct.* is the only *fabula praetexta* (i.e., an ancient Roman tragedy that has a Roman historical subject) that survived until today from the corpus of Latin dramas (Ferri 2003, 1), whereas *H.O.* is a *fabula crepidata*, an ancient Roman tragedy with a Greek subject¹.

A lot of arguments have been made over the years by literary scholars to support the idea that Seneca's stylus could not have written *O.* According to Philp (1968, 151–153), the principal manuscript traditions for the Senecan tragedies are the traditions E and A as well as some excerpts and fragments. The A recension is the only one that transmits *Oct.* (Philp 1968, 151; Seneca 2008, 78). Based on the fact that the interest for Senecan tragedies increased at the beginning of the thirteenth century, there is the hypothesis that *Oct.* was included in the A recension at this time (Gahan 1985; Ferri 2014, 525). Moreover, in both recensions, the texts are given in a different order (Marti 1945, 220).² According to Ferri (2003, 31), the resemblance that *Oct.* bears with the other Senecan

1. It should be noted that extant *fabulae crepidatae* are attributed to Seneca's stylus.

2. Manuscript tradition E saves the Senecan plays with the following order: *Hercules (Furens)*, *Troades*, *Phoenissae*, *Medea*, *Phaedra*, *Oedipus*, *Agamemnon*, *Thyestes*, *Hercules (Oetaeus)*; *Octavia* is omitted in tradition E. Manuscript tradition A gives the Senecan plays with the following order: *Hercules furens*, *Thyestes*, *Thebais*, *Hippolytus*, *Oedipus*, *Troades Medea*, *Agamemnon*, *Octavia*, *Hercules Oetaeus*. The order of the plays and their names follow Philp (1968, 151).

plays and the fact that Seneca “participates” as a persona in the play might have been the reason for classifying *Oct.* as a Senecan play.

Concerning the stylistic aspect of *O.*, the same words are repeated a lot, and some poetic phrases seem artificial rather than the inspiration of the author; in other words, a weakening of the literary power is observed (Herington 1961, 24). Even though in the original Senecan plays the rhetorical style of Ovid was a major influence, the author of *Oct.* seems not to care about this aspect (Michalopoulos 2020). Moreover, Carbone (1977, 56) argues that it had been impossible for Seneca to know details about events that took place after his death with such great precision (e.g., the death of emperor Nero). Poe (1989, 435) suggests that *Oct.* is not Seneca's genuine work, but the product of an imitator with limited literary experience and low levels of creativity when it comes to the provision of conclusions among the scenes.

HO also raises some concerns about the attribution of its authorship. As Marshall (2014, 40) points out, referring to Nisbet, the play follows a different approach of play-writing. For example, the length of this tragedy is twice as long as Seneca's other plays, which makes it the longest extant drama to survive from antiquity (Boyle 2009, 220; Star 2015, 255).

However, it has been also argued that *Oct.* and *H.O.* indeed carry the authorial fingerprint of Seneca. Concerning *O.*, in lines 619–621, Agrippina lists some traditional punishments in an effort to predict the tyrant's (i.e., Nero's) imminent death (Seneca, *Oct.* 619–621). In this passage, the demise of Nero appears to be foretold what seems to rule out Seneca as an author. However, some scholars argue that the description of the punishments is not even close to what actually happened to Nero (i.e., suicide) and that it should not be taken as a prophecy that requires knowledge of the historical event of the death of Nero, since the punishments described represent common and mythological punishments (Pease 1920, 390–391).

Furthermore, Pease (1920, 390) supports the idea that the public circulation of *Oct.* is a posthumous event, and that Seneca entrusted the manuscript of the play to friends in order to be published after the death of Nero. This argument – merely a speculation since no additional evidence exists – can explain the inconsistencies in the text which scholars used to argue that *Oct.* is not a Senecan play. If we follow the line of thought of this argument, someone could hypothesize that Seneca is the author of the play but an editor or a ghost author added or edited some segments of *O.*

With respect to *H.O.*, the argument of the late composition is also used in support of the *H.O.* as a genuine Senecan play (Rozelaar, 1985; Nisbet 1995, p. 209–212; as cited in Marshall 2014, 40). If *H.O.* was one of the last tragedies written by Seneca the Younger before his death, this could explain the haste and the anomalies, which might have caused the sheer length of the play in its current form.

2.2 Quantitative Approaches

There is a plethora of papers that apply computational stylistics to Latin texts, therefore the study of the authorial fingerprint of ancient Latin texts is not something new (e.g., Kestemont et al. 2016; Stover et al. 2016; Stover and Kestemont 2016). However, the number of such papers that consider Senecan texts is much smaller, and more so those

that actually consider the authenticity of the two disputed Senecan plays, *Oct.* and *H.O.* 100
per se. 101

Brofos et al. use a machine learning model trained to recognize texts as Senecan or 102
not, namely a “one-class SVM (i.e., Support Vector Machine) with functional n-gram 103
probability features”³. The model predicts that *Oct.* and *H.O.* were not written by Seneca 104
the Younger (Brofos et al. 2014, 8–9). However, their model also makes, as expected, 105
many misclassifications; it classifies some Senecan texts as non-Senecan, and when the 106
model is augmented with prose texts in addition to tragedies, other authors are also 107
classified as Senecan (Brofos et al. 2014, 9). 108

Nolden (2019) examines the authorship of *Oct.* and *H.O.* with a variety of computational 109
stylistics techniques. Nolden (2019) starts with the hypothesis that *Oct.* and *H.O.* were 110
probably not written by Seneca, and evaluates various methods in this light, including 111
type-token ratio, compressibility, and dimensionality reduction. The results present 112
a mixed picture: some methods point to a high similarity between all the ten plays 113
attributed to Seneca (including the disputed ones), while other methods point to *H.O.*, 114
but also *Phoenissae*, as outliers. However, *Phoenissae* is considered Senecan, so this casts 115
doubt on whether these methods are reliable. In the end, no strong conclusions can 116
be drawn as the differences are small and it is not certain whether the mixed results 117
should be explained as unsuitability of particular methods, or uncertainty of Seneca's 118
authorship. 119

Lastly, it is worth mentioning the paper by Cantaluppi and Passarotti (2015). Even 120
though the main aim of their paper is to cluster the works of Seneca and to show that 121
certain statistical methods can be effective at detecting the genre of the text, their insights 122
are useful for some of the limitations of the methods used in authorship attribution 123
studies and in the current study as well (e.g., Principal Component Analysis). For 124
instance, they perform their analysis using the full size of the text and as they show the 125
Principal Component Analysis method can be affected by the topic and the genre of the 126
text (see the clustering and the words that appear next to the filenames in Cantaluppi 127
and Passarotti 2015). 128

2.3 Literature Review Conclusion 129

In conclusion, “the language and style of these two tragedies [*Oct.* and *H.O.*], how- 130
ever, are identical to the language and style of the others; that is why the discussion 131
of whether these two tragedies are genuine has not yet ceased” (Marshall 2014, 74). 132
Moreover, both of the disputed plays can be considered tricky cases because of the 133
small number of extant Roman tragedies and the fact that *Oct.* has no equivalent extant 134
tragedy in its genre. Previous computational approaches seem to hastily design the 135
experiments by not taking into account multiple variables connected to the texts per se or 136
by considering these works as non-Senecan and focusing on the evaluation of authorship 137
attribution/verification methods and software. Trying to fill this research gap, this paper 138
takes into account as many variables as possible, validates the computational methods 139

3. An SVM is a supervised learning algorithm used for classification and regression tasks. It draws a line or a plane that maximizes the space between the data points, in our case the texts. It works both in linear (data points can be separated by a straight line) and non-linear (data points cannot be separated by a straight line) high-dimensional environments.

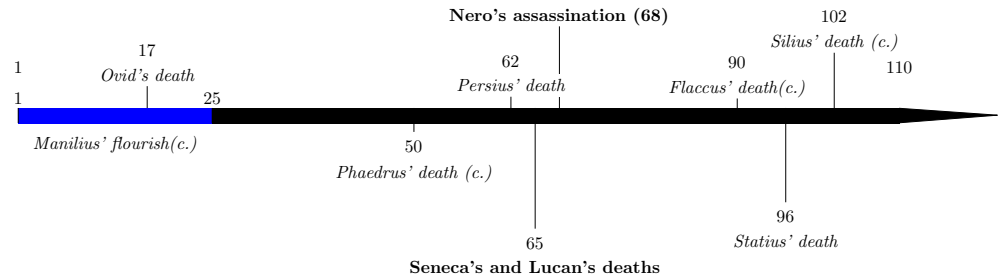


Figure 1: A timeline of the authors used in the dataset, centered around Nero's assassination, Seneca's suicide and Lucan's death. The two extremes in our corpus are Ovid and Silius Italicus.

before it applies them to texts and uses the evaluated methods to contribute and shed new light on the arguments surrounding the authorship of the disputed tragedies. The main research question will be as follows: Were *Oct.* and *H.O.* written by Seneca the Younger or are they, at least in their present form, the product of an imitator or mixed authorship?

3. Dataset

The main dataset employed in this study comprises distractor authors and verse texts that slightly precede and follow the era of Seneca the Younger (c. 4 BCE–65 CE). In the context of computational stylometric approaches, a distractor author, or “impostor”, is utilized for comparison with a disputed text. For clarity, consider a text *X* attributed to author *A*, with distractor authors *B*, *C*, and *D*, known not to be the author of *X*. The soundness of a stylometric method is affirmed by observing significantly higher similarity between *X* and other texts by *A* compared to *B*, *C*, and *D*, confirming *A* as the probable true author or vice versa. In our analysis of Seneca, the dataset includes authors such as Ovid, Manilius, Phaedrus, Persius, Lucan, Valerius Flaccus, Statius, and Silius Italicus (see Table 1). These authors, broadly associated with the literature of the early empire, that wrote within the first century of the Common Era (refer to Figure 1).⁴

In Scenario 5 presented in Table 4 we augment the dataset used by Kestemont et al. Kestemont et al. (2016) with our main corpus (see Table 1, therefore we consider of importance explaining what are the authors and the texts that populate this dataset, as well its main genre. Kestemont's dataset contains 1850 non-overlapping slices of 1000 tokens (for our analysis we split further these texts into non-overlapping slices of 500 tokens). The authors and the text present in the dataset are the following: *Res Gestae A Fine Corneli Taciti* by Ammianus Marcellinus (4th century AD), *Orationum Ciceronis Quinque Enarratio* by Quintus Asconius Pedianus (c. 9 B.C.E. - c. 76 C.E.), *Noctes Atticae* by Aulus Gellius (c. 125 C.E. - after 180), *Declamationes* by Calpurnius Flaccus (2nd century C.E.), *Academica*, *Laelius de Amicitia*, *Pro Archia*, *Brutus*, *Pro Caecina*, *Pro Caelio*, *Cato Maior de Senectute*, *De Divinatione*, *De Fato*, *De Finibus*, *Pro Milone*, *De Natura Deorum*, *De Officiis*, *De Optimo Genere Oratorum*, *Orator*, *De Oratore*, *Paradoxa Stoicorum*, *In Pisonem*, *De Re Publica*, *Topica*, *Tusculanae Disputationes* by M. Tullius Cicero (106 B.C.E. - 43 B.C.E.),

4. Karakasis (2018) suggests Titus Calpurnius Siculus's connection to the reign of Nero, placing him within the Neronian literature. Due to the ongoing debate on Siculus's inclusion in this category, we exclude him from our dataset.

Historiarum Alexandri Magni Libri Qui Supersunt by Quintus Curtius Rufus (1st century C.E.), *Breviarium Historiae Romanae* by Eutropius (4th century C.E.), *Festi Breviarium Rerum Gestarum Populi Romani* by Rufius Festus (c. 370 C.E.), *Epitome De T. Livio Bellorum Omnium Annorum DCC Libri Duo* by Florus (2nd century C.E.), *Historia Apollonii Regis Tyri* by unknown, *Fabulae* by G. Julius Hyginus (c. 64 B.C.E. - 17 C.E.), *Ab Urbe Condita Libri* by Titus Livius (59 B.C.E. - 17 C.E.), *Liber Memorialis* by Lucius Ampelius (c. 2nd century C.E.), *Commentarii in Somnium Scipionis* by Macrobius (flourished 400 C.E.), *Octavius* by M. Minucius Felix (c. 250 C.E.), *Panegyricus Constantino Augusto Dictus* by Nazarius (c. 4th century C.E.), *Epistularum Libri Decem*, and *Panegyricus* by Pliny the Younger (61-2 C.E. - c. 113 C.E.), *De Chorographia* by Pomponius Mela (flourished c. 43 C.E.), *Commentariolum Petitionis* by Quintus Tullius Cicero (102 B.C.E. - 43 B.C.E.), *Declamationes Maiores*, and *Institutiones* by Quintilian (35 C.E. - after 96 C.E.), *Bellum Catilinae*, *Epistola ad Caesarem I & II*, *Bellum Iugurthinum* by Sallustius (c. 86 B.C.E. - 35/4 B.C.E.), *De Beneficiis*, *De Brevitate Vitae*, *De Clementia*, *De Consolatione*, *Epistulae Morales Ad Lucilium*, *De Vita Beata*, *De Ira*, *Quaestiones Naturales*, *De Otio*, *De Providentia*, and *De Tranquillitate Animi* by Seneca the Younger (c. 4 B.C.E. - 65 C.E.), *Controversiae* by Seneca the Elder (c. 55 B.C.E. - 39 C.E.), *De Vitis Caesarum-Augustus*, *De Vitis Caesarum-Gaius*, *De Vitis Caesarum-Divus Claudius*, *De Vitis Caesarum-Domitianus*, *De Vitis Caesarum-Galba*, *De Vitis Caesarum-Divus Iulius*, *De Vitis Caesarum-Nero*, *De Vitis Caesarum-Otho*, *De Vitis Caesarum-Tiberius*, *De Vitis Caesarum-Tiberius*, *De Vitis-Caesaris-Titus*, *De Vitis Caesarum-Divus Vespasianus*, *De Vitis Caesarum-Vitellius* by Suetonius (c. 69 C.E. - after 122 C.E.), *Agricola*, *Annales*, *Historiae*, *Dialogus De Oratoribus* by Tacitus (56 C.E. - c. 120 C.E.), *Factorum Et Dictorum Memorabilium Libri Novem* by Valerius Maximus (flourished 30 C.E.), *De Lingua Latina*, *Rerum Rusticarum De Agri Cultura* by Varro (116 B.C.E. - 27 B.C.E.), *Historiae Romanae* by Velleius Paterculus (c. 19 B.C.E. - after 30 C.E.). Their dataset has mostly historiographical texts since in their paper they compare their corpus with Caesar's writings and it covers a huge time span (from the 4th century B.C.E. up to the 4th century C.E.).

In authorship verification, the challenge of text and author selection inevitably involves some arbitrary or imperfect choices. This section aims to transparently justify our choices. Following Grieve (2007, 255), texts, disputed or not, are inherently tied to their historical era. Consequently, the dataset is designed to narrow the temporal scope, ensuring a more focused linguistic comparison. However, we should highlight two important aspects that complicate the corpus selection.

First, besides the Senecan tragedies, there are no other extant Roman tragedies. Therefore, expanding the timeline is difficult in our case without at the same time increasing the linguistic variation and adding many different genres. Thus, our focus is to run most of the experiments using texts that temporarily are located relatively close to Seneca's the Younger era and of the same kind (in verse)⁵. Second, there is the issue of the varying meter across the texts (e.g., iambic vs hexametric), which constrains the vocabulary available to the author. For computational stylometry, different vocabulary means different features, and therefore dissimilarity between texts. While we cannot completely resolve this issue, we believe that we can limit its influence by considering patterns of frequent character sequences rather than whole words (see subsection 4.1).

5. We do test one scenario where we add historiographical texts in prose that span from the 4th century B.C.E up to the 4th century of C.E (see the description above about Kestemont's dataset (Kestemont et al. 2016)).

In addition to that, prior work on cross-genre and cross-topic stylometry has shown empirically that character-based authorship attribution is robust to such variation (e.g., P. D. Stamatatos et al. 2013, 343). It may be that this robustness also applies to the genre and meter variation in our case. On the other hand, it must be noted that since the disputed plays are compared to Senecan texts in the same genre and meter, while the imposter texts are in a different genre and meter, the likelihood of attributing the disputed plays to Seneca may be increased.

Table 1 provides a complete list of authors and texts included in the dataset variations used for each experiment. All works, with the exception of Manilius's *Astronomica*, were obtained from the Perseus Digital Library (Perseus Digital Library 2024)⁶ because the latter was unavailable from the primary source. Thus, *Astronomica* was sourced from The Latin Library (The Latin Library 2024)⁷.

4. Feature Selection and Methods

The dataset was preprocessed and analyzed using the R package *Stylo* (Eder et al. 2016) and *The Classical Language Toolkit* (CLTK) (Johnson et al. 2021).

4.1 Preprocessing and Feature Selection

Texts were initially tokenized with consideration for the non-differentiation of the letters “v” and “u” in certain text editions. To ensure orthographic consistency, “v” was uniformly converted to “u” where applicable. Pronoun-culling (i.e., eliminating personal pronouns from the text) was then applied to automatically remove frequency information primarily associated with personal pronouns. This step aims to mitigate the impact of genre, topic, author's gender, and narrative perspective on the analysis (Hoover 2004, 480; Newman et al. n.d., 233; Kestemont et al. 2015, 206). Given the varied meter of the texts, even within works by the same author, this approach reduces the “noise” in texts due to the topic or the gender of the author. Both orthographic normalization and pronoun-culling followed the predefined steps of *Stylo* (Eder et al. 2016, 110), with details on the pronoun-culling process outlined in Table 3.

The extraction of relevant features involves character 4-grams in our study, a choice proven effective in cross-genre and cross-topic authorship attribution (Koppel et al. 2009, 12–13; E. Stamatatos 2009, 541–542; Eder 2011, 110; P. D. Stamatatos et al. 2013).⁸ Despite appearing initially inconsequential, character n-grams, particularly of size 4, excel in capturing sub-word level information, including case endings and morphemes (Kestemont 2014, 62–64). In the context of Latin's highly inflected nature, character n-grams preserve details from lower frequency words such as prepositions and determiners (Kestemont 2014, 60–61). Notably, the use of character n-grams eliminates the need for word lemmatization or other normalization, as these features operate below the word level and are language-independent (Daelemans 2013, 4; Kestemont et al. 2015, 206). This approach, utilizing plain inflected surface tokens, has demonstrated increased stability compared to lemma/stem-based methods (Stover and Kestemont

6. Available at: https://github.com/cltk/lat_text_perseus

7. Available at: https://github.com/cltk/lat_text_latin_library

8. For a very simple and informative definition of n-grams see Hagiwara (2021, 53–54).

Author	Text	Filename
Lucan	<i>Pharsalia</i>	luc_phars_{1-10}
Manilius	<i>Astronomica</i>	manil_astro_{1-5}
Ovid	<i>Amores</i>	ovid_am
	<i>Medicamine Faciei Femineae</i>	ovid_medicam
	<i>Ars Amatoria</i>	ovid_ars
	<i>Remedia Amoris</i>	ovid_remed
	<i>Metamorphoses</i>	ovid_meta
	<i>Fasti</i>	ovid_fasti
	<i>Ibis</i>	ovid_ibis
	<i>Tristia</i>	ovid_tristia
	<i>Epistulae ex Ponto</i>	ovid_ponto
	<i>Epistulae or Heroides</i>	ovid_epist
Persius	<i>Saturae</i>	persius_sati_{1-6}
Phaedrus	<i>Fabulae</i>	phaed_fables_{1-6}
Seneca the Younger	<i>Agamemnon</i>	sen_ag
	<i>Hercules Furens</i>	sen_her_f
	<i>Hercules Oetaeus</i> (disputed)	sen_her_o
	<i>Medea</i>	sen_med
	<i>Octavia</i> (disputed)	sen_oct
	<i>Oedipus</i>	sen_oed
	<i>Phaedra</i>	sen_phaed
	<i>Phoenissae</i>	sen_phoen
	<i>Thyestes</i>	sen_thy
	<i>Troades</i>	sen_tro
Silius Italicus	<i>Punica</i>	sil.ita_pun_{1-17}
Statius	<i>Thebaid</i>	stat_theb_{1-12}
	<i>Silvae</i>	stat_silv_{1-5}
	<i>Achilleid</i>	stat_achil
Valerius Flaccus	<i>Argonautica</i>	valflac_argon_{1-8}

Table 1: Authors and texts included in the dataset. All of the texts are written in verse, albeit the only plays are the Senecan tragedies. In total, our corpus comprises 90 texts (including the disputed Senecan plays) and 8 authors to compare against Seneca the Younger.

2016). Slicing words into 4-character packages enhances observations, striking a balance between sparseness and information content (Daelemans 2013, 4–5). In general, character n-grams represent a widely adopted and reliable feature type in stylometry (E. Stamatatos 2009, 541–542; P. D. Stamatatos et al. 2013, 432–433; Eder 2011, 112). In the rest of this paper, we will use the the frequencies of the Most Frequent Character (MFC) n-grams. For example, 2000 MFC refers to the frequencies of the 2000 most common character n-grams.

4.2 Methods

All of the methods we employ estimate the stylistic similarity of texts as the distance between their features (i.e., character n-gram frequencies). For this we pick the Cosine Delta distance metric, based on its effectiveness in various test conditions and particular effectiveness for inflected languages (Jannidis et al. 2015, 6–8; Evert et al. 2017, ii9–

1) que_	2) _et_	3) ere_	4) _in_	5) _qua_
6) ibus_	7) sque_	8) _qu_	9) _bus_	10) usa_
11) _tus_	12) mque_	13) _tis_	14) _qui_	15) pro_
16) per_	17) sin_	18) quo_	19) con_	20) non_

Table 2: Most frequent character 4-grams of the entire corpus (wherever there are less than four characters displayed, the white-spaces are being counted as characters and are displayed using an underscore).

ea	eae	eam	earum	eas	ego
ei	eis	eius	eo	eorum	eos
eum	id	illa	illae	illam	illarum
illas	ille	illi	illis	illius	illo
illorum	illos	illud	illum	is	me
mea	meae	meam	mearum	meas	mei
meis	meo	meos	meorum	meum	meus
mihi	nobis	nos	noster	nostra	nostrae
nostram	nostrarum	nostras	nostri	nostris	nostro
nostros	nostrorum	nostrum	sua	suae	suam
suarum	suas	sui	suis	suo	suos
suorum	suum	suus	te	tibi	tu
tua	tuae	tuam	tuarum	tuas	tui
tuis	tuo	tuos	tuorum	tuum	tuus
vester	vestra	vestrae	vestram	vestrarum	vestras
vestri	vestris	vestro	vestros	vestrorum	vobis
vos					

Table 3: A list of the 98 inflectional forms of 13 pronouns that are removed from every text of the corpus as provided by the software *Stylo* (Eder et al. 2016).

ii10; Eder 2022). Both the validation and main analysis phases utilize the 2000 most frequent character 4-grams (MFCs), a selection supported by studies indicating that the performance of the Cosine Delta plateaus at this threshold for texts in Latin (Jannidis et al. 2015, 6–8; Evert et al. 2017, ii9–ii10).

In general, more MFCs leads to better performance since the features capture more stylistic variation; however, beyond the 2000 MFCs, the character n-grams become more rare and are therefore not as informative. Therefore we consider this point as adequate to capture the necessary amount of authorial fingerprint (Jannidis et al. 2015; Evert et al. 2017; Eder 2022). The frequency distribution plot (see Figure 2) illustrates this diminishing informativeness beyond the 2000th character 4-gram.

The study employs two exploratory analysis methods and one authorship verification method, presented in ascending order of robustness. Firstly, Principal Component Analysis (PCA) is applied. Secondly, the Bootstrap Consensus Tree (BCT) is introduced, followed by the General Impostors (GI) method, each briefly outlined in the subsequent section.

4.2.1 Principal Component Analysis

PCA, a widely used unsupervised algorithm in authorship attribution and verification studies, reduces dimensionality by identifying principal components (eigenvectors) that explain feature variation. In this context, dimensionality refers to the number of features

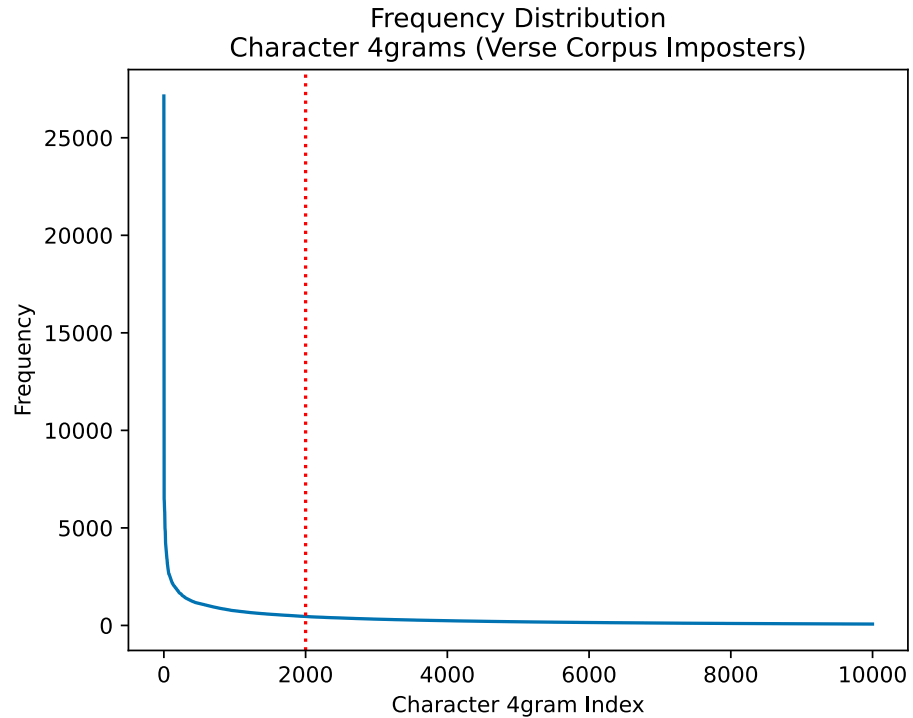


Figure 2: Frequency distribution of the character 4-grams in the whole corpus (i.e., 90 texts including the disputed plays). The vertical line is set to 2000 to show that characters 4-grams after this threshold start to become quite infrequent. The result is what we expect to see since the distribution of the frequency of features in a given text follows Zipf's law (the frequency f of a feature is inversely proportional to its rank r).

or variables initially present in the dataset (in our case the features that are generated 284
by character n -grams). PCA helps reduce this dimensionality by transforming the data 285
into a new set of variables, where each successive variable captures less and less of the 286
total variance in the data. To preserve maximal data variance, PCA zeroes out smaller 287
principal components, employing only those capturing the highest variance (Vander- 288
Plas 2017, 436). These components position texts in a two-dimensional visualization, 289
enhancing readability for human interpretation but at the same time losing some of the 290
variation information (E. Stamatatos 2009, 545). Similarity in frequency distribution 291
correlates with spatial proximity in the PCA plot, indicating text dissimilarity based 292
on vector dissimilarity. Closeness may reflect temporal proximity, common genre, or 293
shared authorship (Manousakis 2020, 171–172). Isolated data points suggest the oppo- 294
site. Applied exclusively to the Senecan corpus, PCA results use a correlation matrix due 295
to its invariance to linear changes in units of measurement, making it suitable for scaled 296
variables like relative frequencies of character 4-grams (Jolliffe and Cadima 2016, 6). 297
The correlation matrix accommodates the varied scale changes within the broad range 298
of 100-2000 most frequent character 4-grams (MFCs). 299

4.2.2 Bootstrap Consensus Tree 300

While the Bootstrap Consensus Tree (BCT) originates from the field of phylogenetics, it 301
was introduced as a method for computational stylometry by Eder (2012) and has since 302
been increasingly used to identify authorial and translator fingerprints (Rybicki 2012; 303
Rybicki and Heydel 2013). The fundamental idea behind bootstrapping is to randomly 304

select a large number of samples with replacement. This process allows us to average the estimates of these samples, thereby enhancing the recurrence of patterns within a document (Jurafsky and Martin 2024, 75–77). Moreover, an assumption of this method is that frequent patterns will reappear many times (robustness), but by increasing the number of iterations and using the consensus strength, we incorporate a larger and thus more diverse number of patterns within a single text (diversity). In other words, a higher number of samples guarantees a greater variety of patterns, making the results more representative of the population.

To clarify some of the concepts mentioned in the previous paragraph: Sampling with replacement involves sampling units returning to the data pool, allowing them to appear in multiple data "snapshots." This facilitates the identification of frequently occurring patterns but also risks letting outliers excessively impact results. To balance the influence of outlier impact, a large number of iterations is usually preferred (Kuhn and Johnshon 2016, 72–73). Moreover, another concept that is being implemented in our approach to further balance the impact of outliers is consensus strength. Consensus strength means that patterns present only in a certain percentage of iterations will be included in the final result. For instance, if we have a consensus strength of 0.5 (i.e., 50%), then only patterns that appeared in at least 50% of the iterations will be included. Unlike a simple dendrogram, a key advantage of BCT lies in its consensus strength, ensuring that more reliable relationships above a specified threshold will influence the final output. Parameters utilized include an MFC n-grams range from 100 to 2000 with a step of 100, and a consensus strength set at 0.5.

4.2.3 General Impostors Method

The GI method, initially introduced by Koppel and Winter (2014), has won for two consecutive years (i.e., 2013 and 2014) the first places in the PAN competitions for shared tasks in authorship verification (Seidman 2013; Khonji and Iraqi 2014). Since then it has proven effective in authenticating disputed writings attributed to Julius Caesar, attributing the text *Compendiosa expositio* to Apuleius, and identifying the author behind the pseudonym Elena Ferrante, and (Kestemont et al. 2016; Stover and Kestemont 2016; Savoy 2020).

In the context of the GI method, authentication involves determining whether a text is consistently attributed to an author across many comparisons and quantifying the confidence in this determination. Unlike many other authorship attribution methods, the GI method handles open-set authorship verification problems, allowing for scenarios where the actual author may or may not be among the candidates.

The GI method verifies authorship based on the document's similarity to the purported author's writings and dissimilarity with impostors. The process is akin to a witness identifying a suspect from a police lineup. Multiple iterations using different subsets of the 2000 most frequent character n-grams enhance the robustness of the results (Eder and Rybicki 2013). In each iteration, 50% of each impostor's text and features are randomly selected for analysis, enabling consideration of numerous feature combinations and outlier detection, leading to more reliable outcomes (Eder et al. 2016). The method produces a score between 0 and 1 for each author in the lineup, indicating the proportion of times an author was identified. A higher score reflects greater confidence that the

author wrote the disputed text (Eder 2018). This score not only gauges stylistic similarity 349
but also assesses how consistently an author is identified with respect to the imposters. 350

5. Validation 351

The methods described were assessed across multiple validation sub-corpora (detailed in 352
respective subsections) to measure their efficacy for authorship attribution/verification 353
tasks. Utilizing the Cosine Delta distance metric and a frequency band of the top 2000 354
MFCs 4-grams, no culling parameter was applied to ensure an adequate feature set.⁹ 355

5.1 PCA (Validation) 356

To validate PCA, a sub-corpus was created from the initial dataset, consisting of works 357
by four authors: Ovid, Lucan, Persius, and Statius (refer to Table 1). These authors 358
were chosen due to their temporal proximity to Seneca's work, despite differences in 359
genre; while Lucan, Ovid, and Statius wrote epic poems, Persius focused on satires. 360
Including Persius's works in this validation corpus was based on their relatively smaller 361
size compared to the other works, posing a potential challenge for PCA analysis. 362

Demonstrating the method's emphasis on text variance over author names, three texts 363
had their author names replaced with "unknown." The filenames were adjusted to 364
unknown_amores for Ovid's *Amores*, unknown_theb_1 for Statius' first book of *Thebaid*, 365
and unknown_sati_4 for Persius' fourth *Satura*. The first two texts were randomly chosen, 366
while the last, due to its small size (392 tokens, including pronouns), posed a challenge 367
for PCA. 368

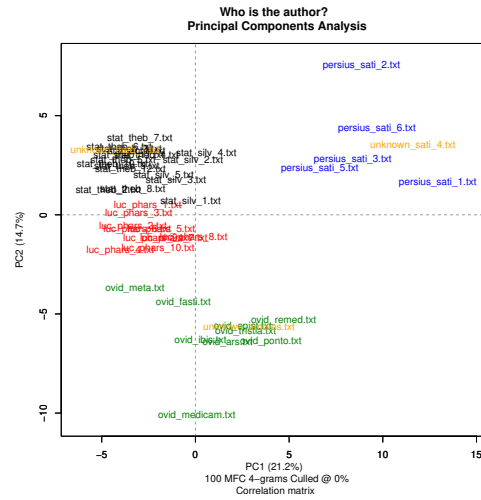
Figure 3 presents PCA results using the correlation matrix, showcasing the impact 369
of different frequency bands (100 MFC 4-grams in Figure 3a and 2000 MFC 4-grams 370
in Figure 3b). Observation reveals a consistent attribution in both cases, with larger 371
frequency bands showing less distinct clusters. Notably, in Figure 3b, Persius' fourth 372
Satura and Ovid's text *Medicamina Faciei Femineae* exhibit some movement outside their 373
relevant clusters. This deviation could be attributed to the small size of these texts 374
relative to others in the corpus, as text size may influence authorship attribution or 375
verification tasks (Luyckx and Daelemans 2011, 52; Eder 2013, 180). 376

5.2 BCT (Validation) 377

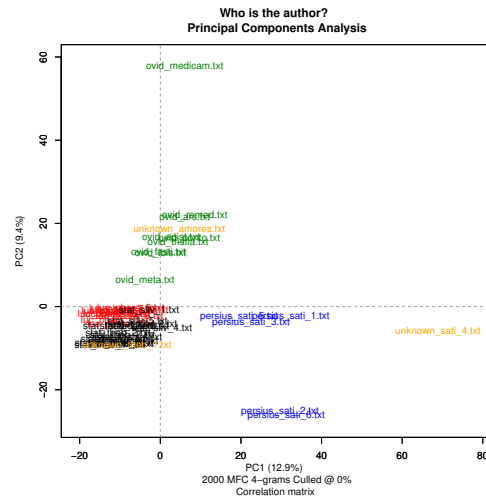
At this point, it is crucial to note that the Bootstrap Consensus Tree (BCT) functions as 378
a consensus, capturing more dimensions and information than PCA due to the robust 379
patterns observed across different iterations (see above subsection 4.2.2). 380

In this validation, the corpus is slightly changed, and file names were altered again to 381
demonstrate the independence of the final result (unrooted tree and branches) from file 382
names. Due to its very small size, this time instead of *Amores* we use *Medicamina Faciei* 383
Femineae as part of the unknown texts by converting its filename to to unknown_medicam. 384

9. Culling, with a ratio of 20, involves including only words occurring in at least 20% of documents in a corpus. While enhancing result comparability, especially with balanced corpora, it introduces a drawback. In unbalanced corpora like ours, with varying document lengths, culling may lead to insufficient features, resulting in an indistinguishable authorial fingerprint for some authors.



(a) 100 MFC 4-grams.



(b) 2000 MFC 4-grams.

Figure 3: PCA using the correlation matrix to visualize the results. **Figure 3a** demonstrates how the attribution works given a small frequency band (i.e., 100 MFCs 4-grams). On the other hand, **Figure 3b** (on the right) demonstrates the authorship attribution given a larger frequency band (i.e., 2000 MFCs 4-grams).

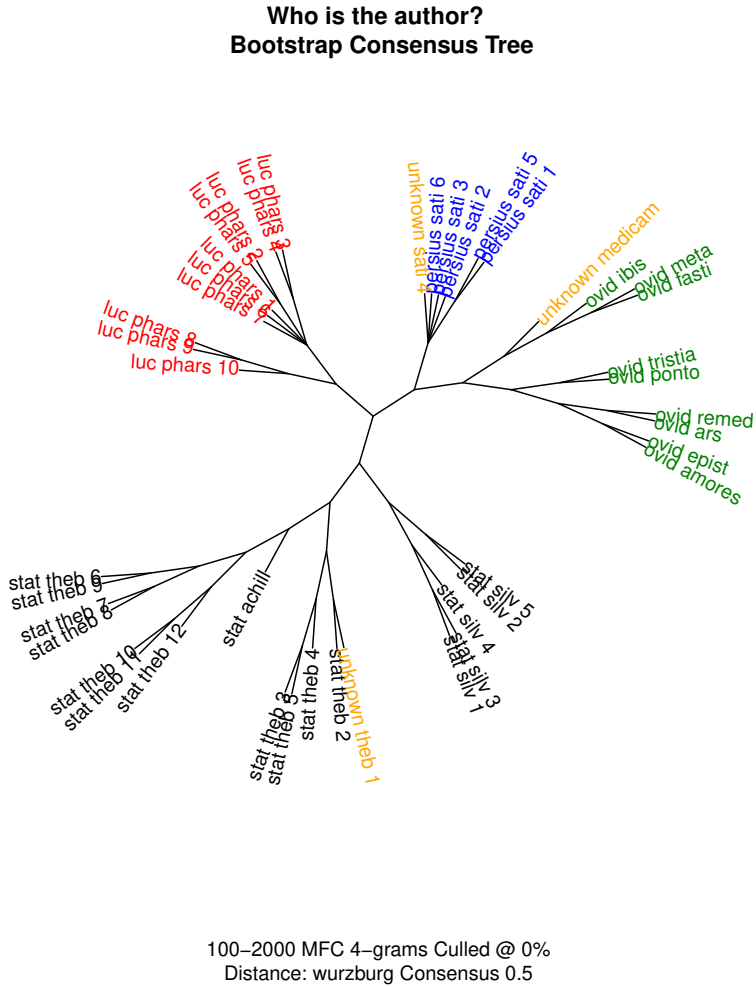


Figure 4: A Bootstrap Consensus Tree that was generated using the top 100-2000-100 (start-end-step) MFC 4-grams and Cosine Delta as distance metric (no culling set); pronoun culling was applied and a consensus strength of 0.5 was used.

The rest of the “unknown” texts remain consistent as in the previous validation test (see 385
above subsection 5.1). 386

All texts in the test set are accurately attributed to their respective authors using BCT 387
(see Figure 4). Notably, the texts renamed as “unknown,” which presented challenges 388
in PCA (i.e., Ovid’s *Medicamina Faciei Femineae* and Persius’ 4th Satura), are handled 389
adeptly by BCT, emphasizing the robustness of BCT in authorship attribution tasks 390
regardless of text size (refer to subsubsection 4.2.2 for further details). 391

5.3 GI Method (Validation) 392

The GI method was validated using all known texts in our corpus, excluding the two 393
disputed Senecan plays (O and *H.O.*), resulting in a total of 88 texts for validation. The 394
Cosine Delta served as the distance metric, and frequency bands ranged from the top 395
100 to 2000 Most Frequent Character (MFC) 4-grams. The method is applied for 100 396
iterations per run to enhance performance. No culling parameter was set, and consistent 397
preprocessing steps were applied, including orthographic normalization (see subsec- 398

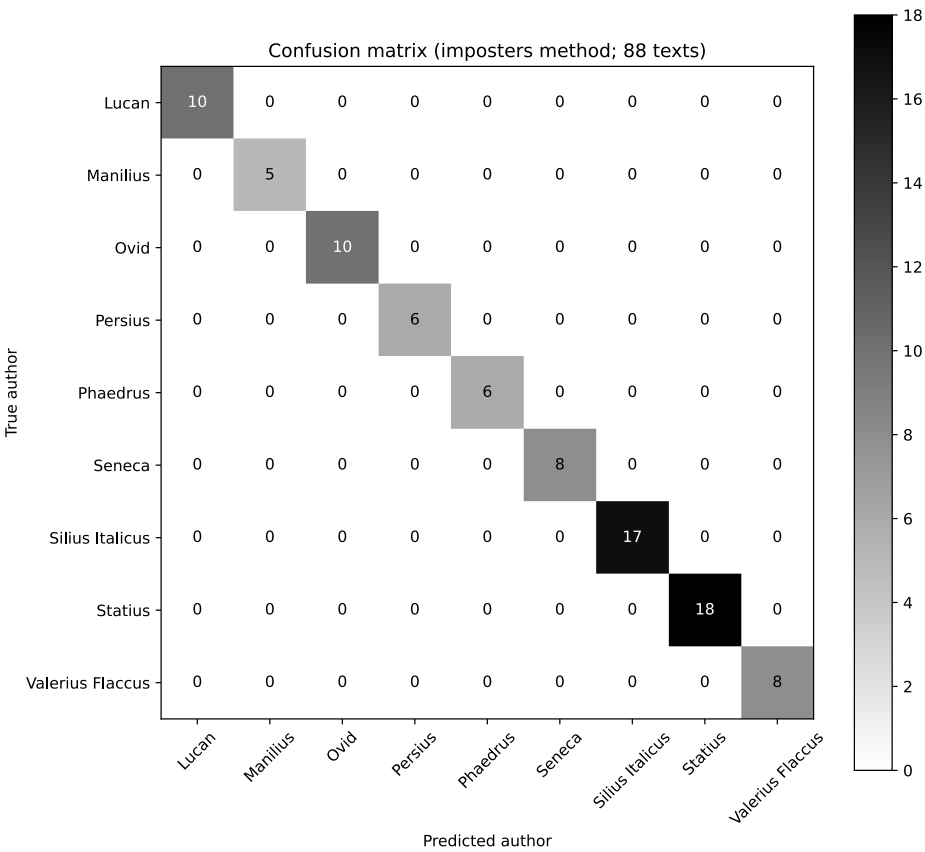


Figure 5: Confusion matrix that shows the results of the GI method on the validation dataset. P1 value = 0.35 and P2 value = 0.64. The result is based on the author that returned the highest score for a given text. The two disputed plays, *Oct.* and *H.O.*, by Seneca the Younger are excluded from the validation set.

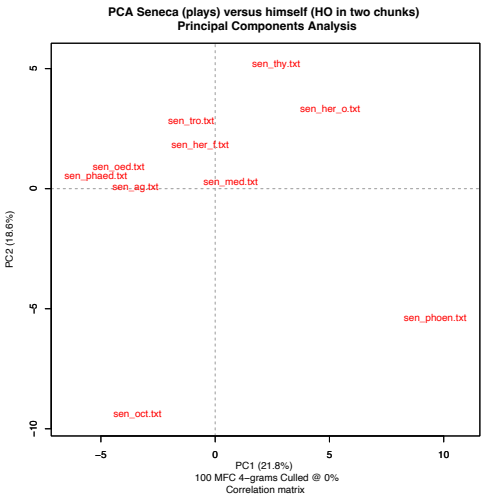
tion 4.1), tokenization and lower-casing, along with pronoun-culling. Subsequently, the GI method was applied to each text in the validation corpus.

5.4 Validation Findings

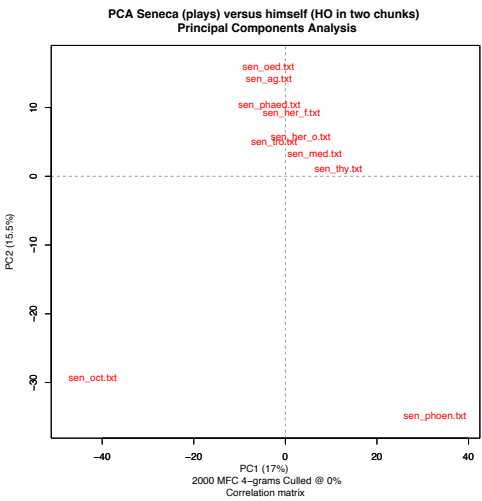
The validation indicates effective performance for all methods on the texts within the corpus, with PCA showing limitations for short texts (Figure 3). The BCT method demonstrates robust recognition of authorial fingerprints across varied text lengths, owing to their bootstrapping techniques, culminating in a consensus from multiple iterations (see Figure 4). Similarly, the GI method reports a perfect accuracy for attributing the 88 texts (see Figure 5). These findings suggest that the selected frequency band (top 100 to 2000 Most Frequent Character 4-grams) is informative for capturing authorial fingerprints, yielding high success rates in each validation scenario. Consequently, the main analysis phase will replicate this process, with a focus on the disputed texts.

6. Results and Discussion

We first explore the stylometric properties of the Senecan plays using PCA, to see how they relate to each other. When treating the plays as a whole, it can be observed that

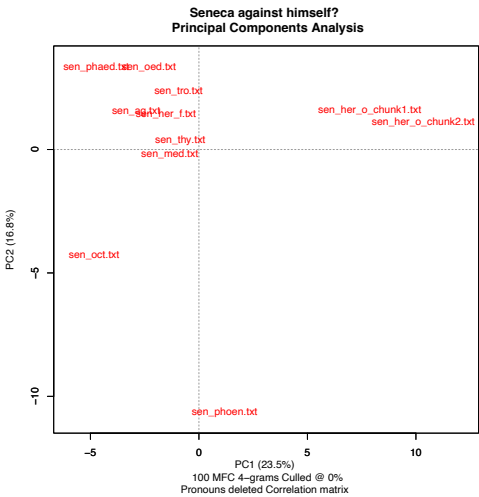


(a) 100 MFC 4-grams.

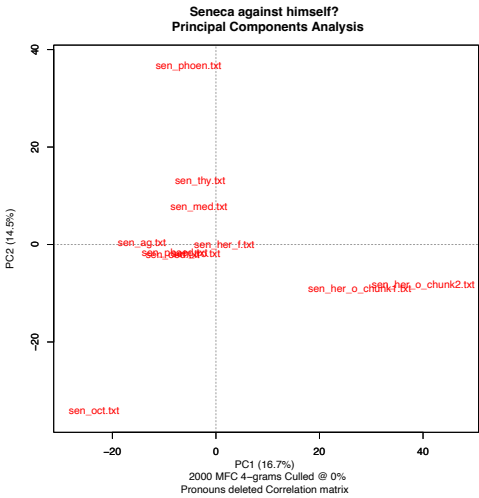


(b) 2000 MFC 4-grams.

Figure 6: PCA correlation matrix of the Senecan corpus of plays (disputed or not). The texts seneca_oct and seneca_her_o correspond to *Oct.* and *H.O.* respectively. In both cases, regardless of the size of the frequency band, *Oct.* and *Phoenissae* behave as outliers within the Senecan corpus, whereas *H.O.* is placed among the Senecan plays. It's important to highlight that the percentage shown in PC1 and PC2 varies in each plot because the principal components capture different amounts of variance each time.



(a) 100 MFC 4-grams.



(b) 2000 MFC 4-grams.

Figure 7: PCA correlation matrix of the Senecan corpus of plays (disputed or not), this time with *H.O.* split in half. *H.O.* starts to behave as outlier and *Oct.* remains among the outliers. It's important to highlight that the percentage shown in PC1 and PC2 varies in each plot because the principal components capture different amounts of variance each time

from the two disputed texts, only *Oct.* behaves as outlier within the Senecan corpus of plays (see Figure 6). However, *H.O.* consists of 11.1147 tokens which, compared to the average size of a Senecan play (excluding *Oct.*) in terms of tokens, is almost double the size (average size of a Senecan play is 6192.5 tokens). When *H.O.* is divided into two halves to align its size more closely with the average size of a Senecan play, it shifts away from the cluster of Senecan texts (refer to Figure 7). Meanwhile, *Oct.* consistently remains outside the cluster of Senecan plays. A possible explanation of why *Oct.* and *H.O.* behave as outliers is the fact that when considering the works of a single author using a PCA, the genre-related signal tends to become stronger than the author-related signal (Stover and Kestemont 2016, 659).

In addition to that, it should be stressed that in all of the PCA plots *Phoenissae* also behaves as an outlier within the Senecan corpus, while its authorship is not disputed. An explanation for this behavior could be that *Phoenissae* is an unfinished play and the shortest text in the Senecan corpus of plays. Furthermore, the aforementioned play has a lot of issues in terms of structure and unity; based on the number of innovations that were attempted in the text, Frank (2018, 1–2) points out that this might be the reason why this text was abandoned by Seneca when he realized the difficulty of this venture.

Figure 8 shows a Bootstrap Consensus Tree (BCT) for the Senecan plays alongside two selected authors from the literature of the early empire, Lucan and Statius. Statius is included to test the hypothesis of Ferri (2003, 17–27), suggesting a temporal connection between the composition of *Oct.* and Statius. The BCT exhibits distinct branches for each author, placing both the disputed plays in proximity to the Senecan works, but *Oct.* is slightly gravitating towards the center of the unrooted tree. This again highlights the special nature of this specific text. On the other hand, *H.O.* remains among the Senecan cluster of plays.

Regarding the GI method, we test 5 different scenarios. However, since GI returns a confidence score as the final output we need to pick thresholds in order to reject or accept the verification of an author. Stylo provides a method to automatically determine such thresholds using cross-validation (the `stylo.optimize()` method). For Scenario 1, 2, and 3 (see Table 4), this gives thresholds of 0.25 and 0.74 (i.e., under 0.25, Seneca is definitely not the author; above 0.74, Seneca is verified as the author; when the score is in between, no determination can be made). Unfortunately, the cross-validation method is too expensive to run with the larger datasets we use in the rest of our experiments (see scenarios 4 and 5 in Table 4) due to the nested loops and the bootstrapping that takes place which results to an increase of the time complexity of the algorithm. Therefore we will use a conservative threshold of 0.9 for all our experiments.

With the GI method, Scenario 1 and 2 confidently attribute Seneca the Younger as the author of the disputed plays (see Table 4). Next, in Scenario 3, we consider the cento-argument by Ferri (2014, 48).¹⁰ We do this by identifying and removing lines from the disputed texts resembling those in the Senecan corpus of plays. We operationalize sentence similarity using Tf-Idf (term frequency, inverse document frequency) vectors of the character 4-grams for each sentence, and cosine similarity as the metric for the similarity of pairs of sentences. We identify and exclude all sentences with a similarity

10. A basic definition of a cento would describe it as a composition largely comprised of quotations from the works of other authors.

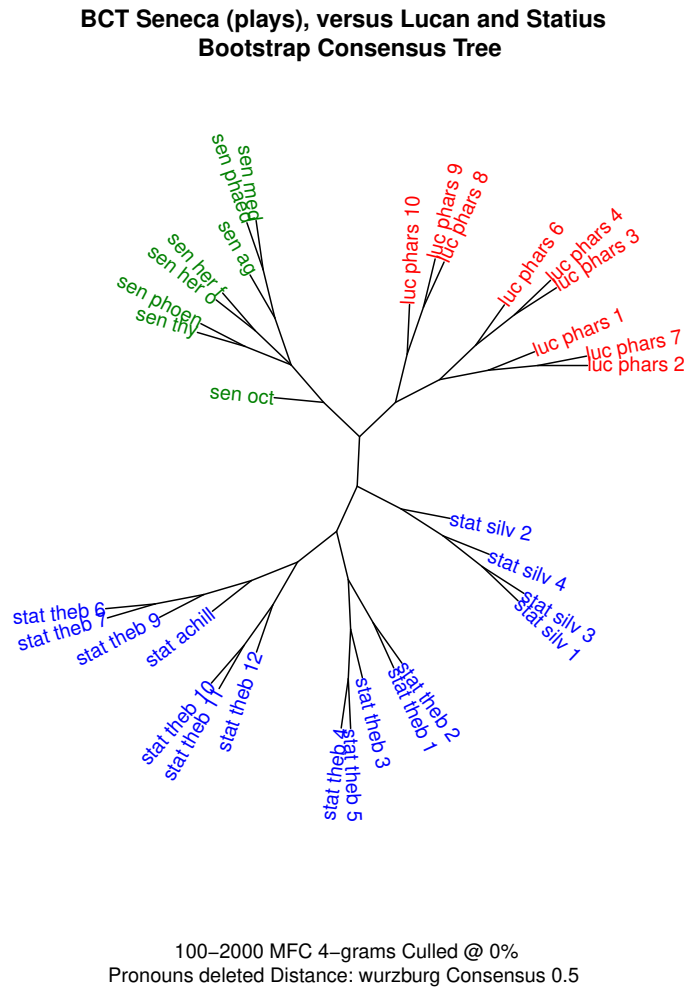


Figure 8: BCT of texts from Statius (*Achilleid*, *Thebaid*, *Silvae*), Lucan (*Pharsalia*), and Seneca (plays). The texts ‘seneca_oct’ and ‘seneca_her_o’ correspond to Oct. and H.O. respectively.

Scenario	Dataset	Results
Scenario 1: The GI method used against the disputed texts (no changes were applied to the texts per se)	90 text samples in verse written by authors that lived slightly before and after Seneca the Younger (see Figure 1 and 1).	<i>Octavia</i> : 1.0 <i>Hercules Oetaeus</i> : 1.0
Scenario 2: The GI method is applied to <i>H.O.</i> split into two chunks.	Same as Scenario 1, but <i>H.O.</i> split into two chunks.	<i>Hercules Oetaeus</i> chunk 1: 1.0 <i>Hercules Oetaeus</i> chunk 2: 1.0
Scenario 3: The GI method is applied to the two disputed texts. <i>Oct.</i> and <i>H.O.</i> are cleaned by removing sentences that are above the similarity threshold (i.e., 0.6) in terms of cosine similarity.	Same as Scenario 1, but <i>Oct.</i> and <i>H.O.</i> are cleaned from similar lines with the rest of the Senecan corpus of plays.	<i>Octavia</i> : 1.0 <i>Hercules Oetaeus</i> : 1.0
Scenario 4: The GI method is applied to the two disputed texts (i.e., <i>Oct.</i> and <i>H.O.</i>). Each text in the corpus is split into non-overlapping chunks of 500 words if their length is above 500 tokens. This addresses a possible length bias due to shorter or longer texts. In addition, it enables checking for mixed authorship throughout the disputed texts.	The main corpus, but the texts are divided into chunks of 500 tokens, resulting in 1257 text samples.	For the scores for each chunk, see Figure 9 and 11
Scenario 5: The GI method is applied to the chunks of the two disputed plays. This time the texts are compared with texts in prose (the dataset is the one used by Kestemont et al. (2016) but augmented with the chunks of our impostors dataset). The total size of this dataset including the disputed plays is 3061 text samples.	A larger dataset of mostly historiographical texts written in prose (a small number are in verse), augmented with the 500 token chunks of our main impostors dataset, resulting in 3051 text samples. This dataset includes texts written by Seneca the Younger in prose (e.g., <i>De Ira</i> , <i>De Providentia</i> , etc.)	For the score for each chunk, see Figure 10 and 12

Table 4: All the scenarios tested using the GI method, a brief description of the results, and the P1 & P2 values for each scenario. The interpretation of the P1 and P2 values is as follows: any score below P1 suggests a negative answer to the question, "Can author A be confirmed as the author of disputed document X?" Conversely, any score above P2 indicates a positive answer to the same question. Between P1 and P2 lies a 'grey area' where no definitive conclusions should be drawn.

Play	Line	Score
<i>Phoenissae</i> O	scelus in propinquo est nihil in propinquos temere constitui decet	0.40
<i>Agamemnon</i> HO	eheu quid hoc est quid hoc	0.52
<i>Phaedra</i> HO	anime quid segnis stupes quid stupes segnis furor	0.60
<i>Medea</i> O	Profugere dubitas? Parere dubias?	0.64
<i>Thyestes</i> HO	Viduam relinques? Vitam relinques?	0.71
<i>Phoenissae</i> O	Et hoc sat est nec hoc sat est	0.74
<i>Phaedra</i> HO	quam bene excideram mihi quam bene excideras dolor	0.77
<i>Agamemnon</i> HO	scelus occupandum est scelus occupandum est	1

Table 5: Lines from Senecan and disputed plays with cosine similarity scores. The first two rows are examples of sentences that did not pass the threshold (< 0.6).

exceeding a threshold of 0.6. The cosine similarity metric measures directional similarity between vectors, irrespective of magnitude or scale (Singhal et al. 2001, 2–3). The presented methodology, when integrated with specific preprocessing procedures including the conversion to lowercase, elimination of punctuation marks (with the understanding that an editor may subsequently reintroduce punctuation marks), and the utilization of character 4-grams as distinctive features, exhibits the capability to discern similarities. This capability is exemplified in Table 5, wherein similarities are identified not only among various declensions of identical terms but also amid permutations in word order. For *Oct.* from a total 422 sentences, we identified and thus removed 2 (i.e., 0.46%) sentences above the similarity threshold (i.e., 0.6), whereas for *H.O.*, from a total of 1149 sentences we identified and removed 33 (i.e., 2.87%) sentences.

To address potential length bias and investigate possible mixed authorship throughout the disputed texts, in Scenario 4 each text exceeding 500 tokens is divided into non-overlapping chunks of 500 tokens. This approach, inspired by Rolling Stylometry (Eder 2016), simplifies the process by using non-overlapping segments instead of overlapping ones. Note that, Rolling Stylometry works by analyzing text in sequential segments to track stylistic patterns and changes over time within a document or corpus. The results for Scenario 4 (Figure 9 and Figure 11) reveal a nuanced internal composition, uncovering authorship diversity within the disputed plays. Although Seneca's authorship dominates, specific segments warrant attention, as highlighted in Figure 9 and 11.

For *Oct.* we observe a declining pattern in some text segments, especially for chunks 1, 3, 6, and 8 (Figure 9 and Table 6). However, excluding chunk 6 and 8 (score of 0.77), the rest of the scores are very close to 0.9 and thus the most prudent inference is that they remain of Senecan origin. Concerning chunk 6 (467–553) and chunk 8 (lines 634–733) the playwright condenses the time in a way that seems unnatural for Seneca the Younger in

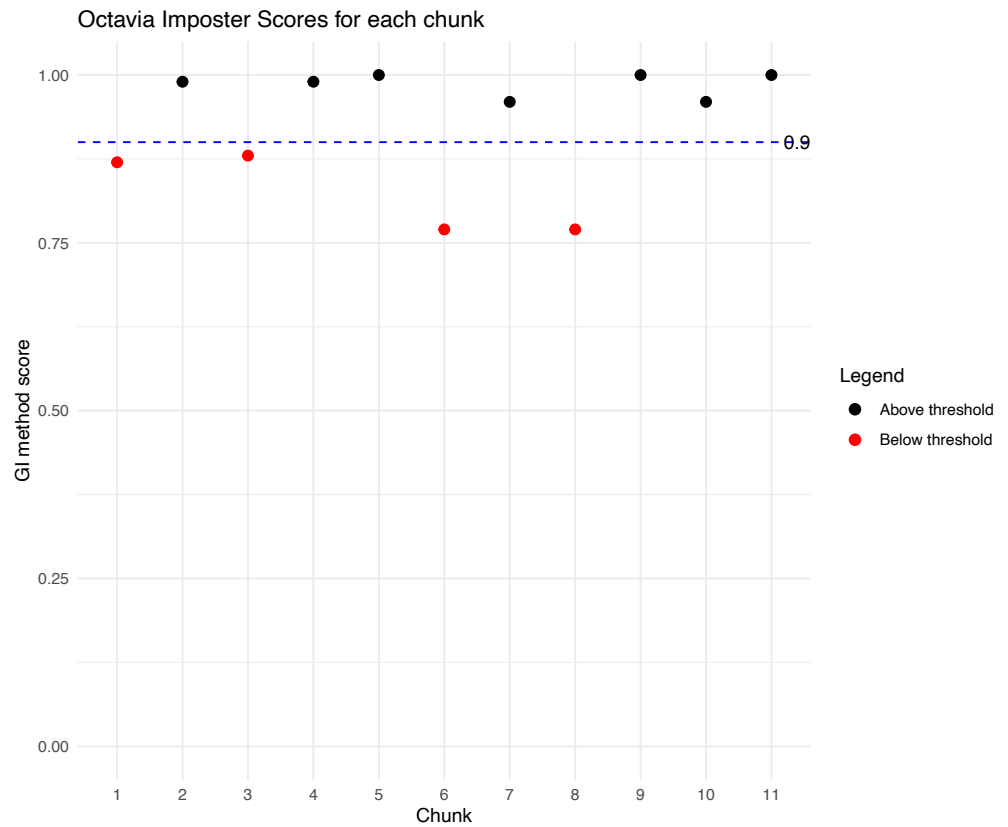


Figure 9: Results of the GI method for O's chunks (Scenario 4).

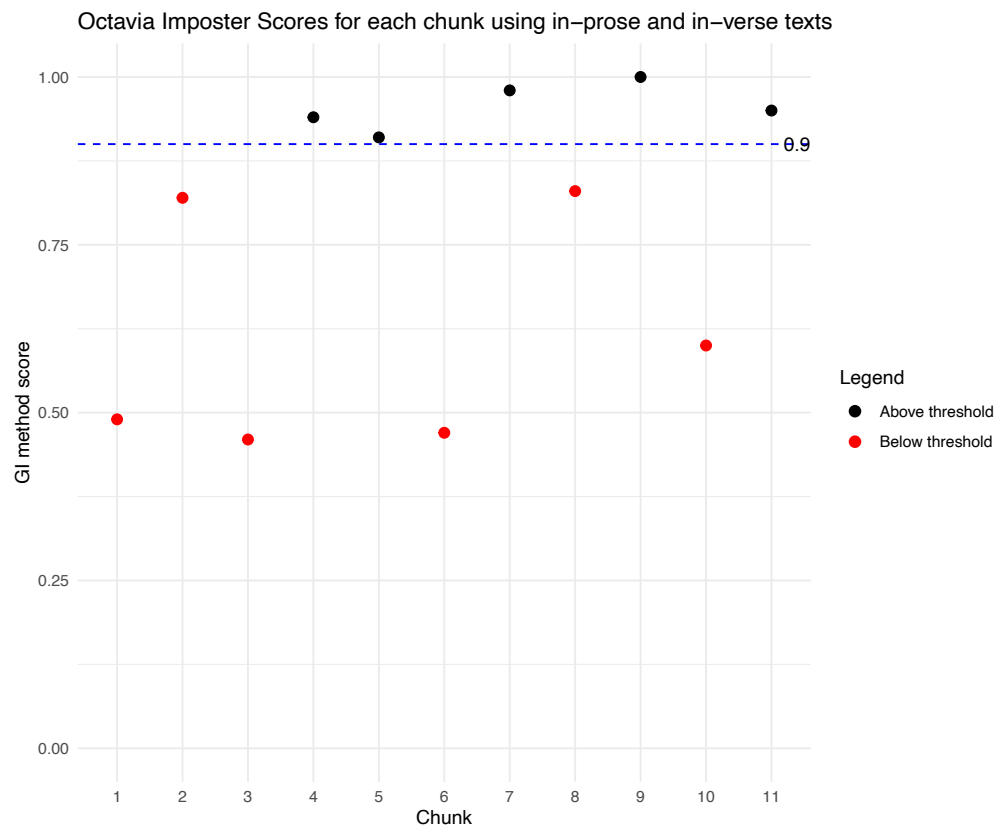


Figure 10: Results of the GI method for O's chunks using the dataset of Kestemont et al. (2016) (Scenario 5).

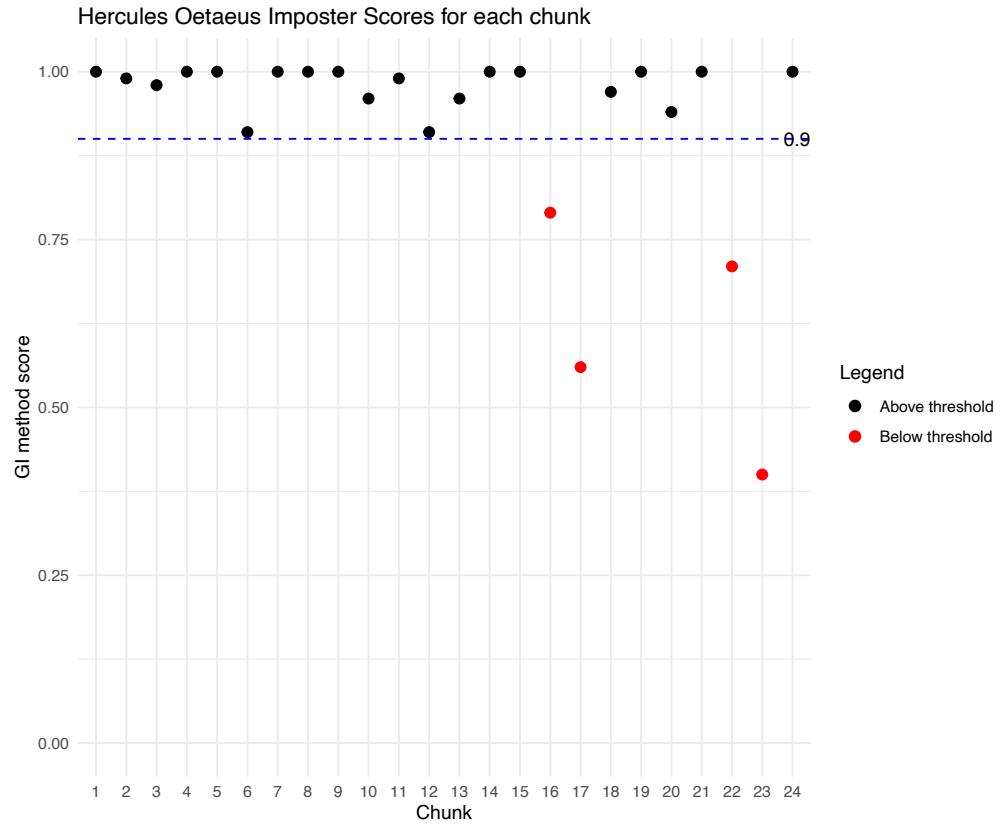


Figure 11: Results of the GI method for *H.O.*'s chunks (Scenario 4).

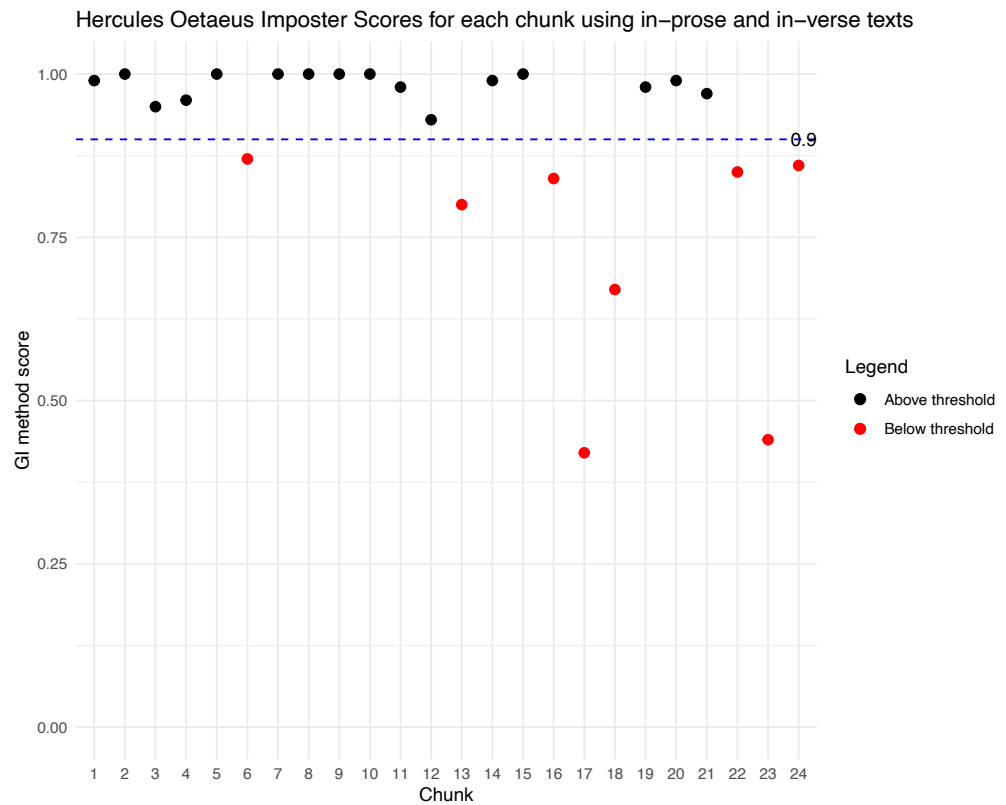


Figure 12: Results of the GI method for *H.O.*'s chunks using the dataset of Kestemont et al. Kestemont et al. (2016) (Scenario 5).

order to present a large number of events in a small amount of time (Ferri 2014, 307–309). 482
 Moreover, in both of the chunks the direct critique to Nero's reign in this passage can be 483
 considered as a task that is difficult to perform by someone (i.e., Seneca) who is working 484
 as the advisor of the emperor. 485

Furthermore, building upon the earlier discoveries, Figure 11 illustrates a noteworthy 486
 pattern within the *H.O.* text (see Table 7). Beyond chunk 16 (i.e., line 1297 and onwards), 487
 there is a small number of chunks with scores below the specified threshold of 0.9, 488
 indicating that they might have not been written by Seneca. This observation to some 489
 extent aligns with the hypothesis positing that the first half of the text originates from 490
 Seneca, while the remainder was finished by someone else (Tarrant 2017, 97). However, 491
 according to our results, most of the chunks in the second half were written by Seneca, 492
 which suggests that the second half is a case of mixed authorship, rather than having 493
 been completely written by someone else. 494

Lastly, in Scenario 5 we consider the dataset used by Kestemont et al. (2016) which 495
 mainly consists of historiographical texts that span from the 4th century B.C.E. until 496
 the 4th century C.E.. We augment their corpus with our current corpus of impostors 497
 resulting in 3015 text samples and a mix of texts in prose and verse. Notably, the corpus 498
 also contains additional texts by Seneca (in prose). In this scenario, the texts are more 499
 dissimilar in terms of genre and chronology. On the other hand, the number of impostor 500
 authors is larger (in total 35 authors), should make it more difficult to pick out the right 501
 author and increase the reliability of the result (similar to picking out a subject from a 502
 larger police lineup). The results for *Oct.* (Figure 10) are highly similar to the results 503
 of Scenario 4 (Figure 9), where the dataset contains only texts in verse but the chunks 504
 that indicate mixed authorship grow in number (chunks 1, 2, 3, 6, 8, 10 (see Table 8)). 505
 Concerning *H.O.* (Figure 12 and Table 9), when compared against Kestemont's dataset, 506
 the signal for mixed authorship is becoming stronger too, especially after chunk 13 507
 (lines 1027ff.). However, it should be noted again that chunks 6, 16, 22, and 24 still fall 508
 very close to the threshold of 0.9, therefore most likely remain of Senecan origin. 509

Chunk no.	Lines	Score
Chunk 1	l. 1-102	0.87
Chunk 3	l. 184-276	0.88
Chunk 6	l. 467-553	0.77
Chunk 8	l. 634-733	0.77

Table 6: Chunks of *Oct.* that return a score below the threshold of 0.9 using the main corpus split into non-overlapping chunks of 500 tokens. The lines correspond to their online version in the Perseus Digital Library.

Chunk no.	Lines	Score
Chunk 16	l. 1319-1398	0.79
Chunk 17	l. 1398-1480	0.56
Chunk 22	l. 1819-1917	0.71
Chunk 23	l. 1918-1996	0.40

Table 7: Chunks of *H.O.* that return a score below the threshold of 0.9 using the main corpus split into non-overlapping chunks of 500 tokens. The lines correspond to their online version in the Perseus Digital Library.

Chunk no.	Lines	Score
Chunk 1	l. 1-102	0.49
Chunk 2	l. 102-185	0.79
Chunk 3	l. 185-276	0.46
Chunk 6	l. 467-553	0.47
Chunk 8	l. 634-733	0.62
Chunk 10	l. 825-914	0.59

Table 8: Chunks of *Oct.* that return a score below the threshold of 0.9 using Kestemont’s corpus. The lines correspond to their online version in the Perseus Digital Library.

Chunk no.	Lines	Score
Chunk 6	l. 430-508	0.89
Chunk 13	l. 1027-1149	0.78
Chunk 16	l. 1319-1398	0.83
Chunk 17	l. 1398-1480	0.42
Chunk 18	l. 1480-1573	0.69
Chunk 22	l. 1819-1917	0.88
Chunk 23	l. 1918-1996	0.45
Chunk 24	l. 1970-end	0.88

Table 9: Chunks of *H.O.* that return a score below the threshold of 0.9 using Kestemont’s corpus. The lines correspond to their online version in the Perseus Digital Library.

7. Conclusions

510

Our findings underscore the complexity of the authorship verification problem, particularly evident in the case of the disputed Senecan plays, *Oct.* and *H.O.*. Across experimental runs, varying results highlight the intricate nature of this challenge in computational stylometry.

511

512

513

514

Paraphrasing Stover and Kestemont (2016, 647), our aim is not to replace existing modes of analysis but rather to illuminate longstanding issues by shedding new light through the application of innovative tools grounded in traditional methods. This analysis underscores the importance of considering genre and meter variations in our conclusions. As previously noted, these two factors can introduce complexities due to their influence on vocabulary. It is impossible to completely remove the influence of variation in meter and genre, thus to mitigate their impact on the final results, we employ preprocessing techniques.

515

516

517

518

519

520

521

522

Through the validation phase, we demonstrate the effectiveness of these techniques for our task. Consequently, we apply these techniques consistently to generate uniform features for each method. Notably, in the case of the two exploratory methods—PCA and BCT—*Oct.* and *H.O.* emerge as intriguing examples of texts concerning their authorship among the Senecan corpus of plays. In certain instances, they exhibit clustering with the broader set of Senecan plays, while in other instances, they do not. For instance, when using only the Senecan plays, the genre and thus the meter seems to win over the authorial fingerprint and variables like the size of the plays (see the cases of *Phoenissae* and *H.O.* in Figure 6).

523

524

525

526

527

528

529

530

531

The initial two scenarios of the GI method confidently verify Seneca as the author with a high degree of confidence ($=1.0$). Moreover, after removing from both disputed plays lines that are similar to lines from other Senecan plays, the GI method still verifies Seneca as the author of the disputed plays. Therefore, the stylistic similarity of the disputed plays with the works of Seneca cannot be explained by borrowed phrases. Nevertheless, the fourth scenario highlights segments in *Oct.* and *H.O.* that are likely not attributable to Seneca, implying the involvement of a distinct author or editor. By concentrating on the fourth GI scenario for *H.O.* (refer to Figure 9 and 11) and observing a diminishing trend in confidence after the 13th chunk, though remaining proximate to the average scores for each chunk, we posit that an editor of the text may have edited or added certain portions to the original play, even though it was primarily authored by Seneca. Lastly, the results hold up when the disputed plays are compared with a larger corpus of prose texts, suggesting that our findings are robust.

Against this algorithmic confidence, two objections can be made. First, we cannot rule out a highly skilled imitator; however, this seems implausible given the advanced nature of modern stylometry, of which an imitator could not have been aware. Second, the distractor texts differ in genre and meter from the Senecan texts. Unfortunately, it is impossible to construct a perfect distractor corpus, due to limitations of extant texts. Therefore, while our empirical findings cannot positively confirm Seneca as the author of the disputed plays, our main contribution is that, perhaps contrary to expectation given the consensus against Seneca's authorship, most of the text of the disputed plays is highly stylistically similar to Seneca's writings. This means that Seneca cannot be ruled out as the author of the disputed plays based on stylometry. Moreover, our results provide evidence for mixed authorship in specific parts of the disputed plays.

8. Further Research

Deciphering the authorial fingerprint of the Senecan disputed plays requires further investigation and consideration of study limitations. Future work could take a closer look at the specific text chunks diverging from Seneca the Younger's style. Employing Rolling Stylometry or using the General Imposters method with overlapping text segments (Eder 2016;Beullens et al. 2024), in collaboration with close reading approaches, could enable identification of authorship at the sentence level and enhance understanding of why these segments differ from Seneca's style. Moreover, exploring the impact of prosody in ancient languages (e.g., Latin or ancient Greek) on stylometric methods is another avenue for investigation. Controlled experiments using authors that wrote in different meters would make it possible to quantify its effect on the stylometric profile of texts. Furthermore, while the GI method has been shown to be robust and reliable in previous studies, including for Latin (Kestemont et al. 2016), it would be useful to examine and empirically test whether an imitator can successfully deceive the GI method. The Ferrante case shows that the pseudonym of an author who is highly motivated to hide his identity can be unmasked by pinpointing the gender, age, region and city of the author profile (Mikros 2018). A potential improvement would be to use a large language model, which could also detect paraphrases by taking into account semantic similarity.

9. Data Availability 575

Data and code: https://github.com/PaschalisAg/seneca_stylometry 576

10. Software Availability 577

Data and code: https://github.com/PaschalisAg/seneca_stylometry 578

11. Acknowledgements 579

We extend our sincere gratitude to all the anonymous reviewers for their invaluable feed- 580
back. Their insightful comments illuminated aspects of the paper that might otherwise 581
have escaped our notice. Special thanks are due to Vasileios Dimoglidis, a PhD student 582
in Classics at the University of Cincinnati (UC), whose initial inspiration sparked this 583
study. Additionally, we are grateful to Associate Professor Vasileios Pappas and Pro- 584
fessor Helen Gasti from the University of Ioannina (UIO) for generously providing an 585
extensive bibliography to support our research into the non-quantitative approaches 586
examined in this study. 587

12. Author Contributions 588

Paschalis Agapitos: Conceptualization, Writing – original draft 589

Andreas van Cranenburgh: Formal Analysis, Writing – review & editing 590

References 591

- Beullens, Pieter, Wouter Haverals, and Ben Nagy (Apr. 2024). "The Elementary Particles: 592
A Computational Stylometric Inquiry into the Mediaeval Greek-Latin Aristotle". In: 593
Mediterranea. International Journal on the Transfer of Knowledge 9, 385–408. [https://jo](https://journals.uco.es/mediterranea/article/view/16723) 594
[urnals.uco.es/mediterranea/article/view/16723](https://journals.uco.es/mediterranea/article/view/16723). 595
- Boyle, A. J. (2009). *Tragic Seneca: An Essay in the Theatrical Tradition*. Routledge. 596
- Brofos, James, Ajay Kannan, and Rui Shu (2014). "Automated Attribution and Intertext- 597
tual Analysis". In: *arXiv*. [10.48550/ARXIV.1405.0616](https://arxiv.org/abs/10.48550/ARXIV.1405.0616). 598
- Cantaluppi, Gabriele and Marco Passarotti (2015). "Clustering the Corpus of Seneca: 599
A Lexical-Based Approach". In: *Advances in Latent Variables: Methods, Models and* 600
Applications. Ed. by Maurizio Carpita, Eugenio Brentari, and El Mostafa Qannari. 601
Springer International Publishing, 13–25. [10.1007/10104_2014_6](https://doi.org/10.1007/10104_2014_6). 602
- Carbone, Martin E. (1977). "The "Octavia": Structure, Date, and Authenticity". In: 603
Phoenix 31.1, 48–67. [10.2307/1087155](https://doi.org/10.2307/1087155). 604
- Daelemans, Walter (2013). "Explanation in Computational Stylometry". In: *Proceedings* 605
of the 14th International Conference on Computational Linguistics and Intelligent Text 606
Processing - Volume 2. 14th International Conference on Computational Linguistics 607
and Intelligent Text Processing. Vol. 2. CICLing'13. Springer, 451–462. [10.1007/978-](https://doi.org/10.1007/978-3-642-37256-8_37) 608
[3-642-37256-8_37](https://doi.org/10.1007/978-3-642-37256-8_37). 609

- Eder, Maciej (2011). "Style-Markers in Authorship Attribution A Cross-Language Study of the Authorial Fingerprint". In: *Studies in Polish Linguistics; Issue 1*. ISSN: 1732-8160. <https://www.ejournals.eu/SPL/2011/SPL-vol-6-2011/art/1171/>.
- (2012). "Computational stylistics and Biblical translation: How reliable can a dendrogram be?" In: *The Translator and the Computer*. The Translator and the Computer. Ed. by Tadeusz Piotrowski and Łukasz Grabowski. Wrocław: Wyższa Szkoła Filologiczna we Wrocławiu.
- (Nov. 2013). "Does size matter? Authorship attribution, small samples, big problem". In: *Digital Scholarship in the Humanities* 30.2. eprint: <https://academic.oup.com/dsh/article-pdf/30/2/167/21517531/fqt066.pdf>, 167–182. ISSN: 2055-7671. [10.1093/llc/fqt066](https://doi.org/10.1093/llc/fqt066).
- (Sept. 1, 2016). "Rolling stylometry". In: *Digital Scholarship in the Humanities* 31.3, 457–469. [10.1093/llc/fqv010](https://doi.org/10.1093/llc/fqv010).
- (2018). *Authorship verification with the package stylo*. Computational Stylistics. <https://computationalstylistics.github.io/docs/imposters>.
- (2022). "Boosting word frequencies in authorship attribution". In: *arXiv e-prints*. [10.48550/arXiv.2211.01289](https://arxiv.org/abs/2211.01289).
- Eder, Maciej and Jan Rybicki (June 1, 2013). "Do birds of a feather really flock together, or how to choose training samples for authorship attribution". In: *Literary and Linguistic Computing* 28.2, 229–236. [10.1093/llc/fqs036](https://doi.org/10.1093/llc/fqs036).
- Eder, Maciej, Jan Rybicki, and Mike Kestemont (2016). "Stylometry with R: A Package for Computational Text Analysis". In: *The R Journal* 8.1, 107–121. [10.32614/RJ-2016-007](https://doi.org/10.32614/RJ-2016-007).
- Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt (Dec. 1, 2017). "Understanding and explaining Delta measures for authorship attribution". In: *Digital Scholarship in the Humanities* 32 (suppl_2), ii4–ii16. [10.1093/llc/fqx023](https://doi.org/10.1093/llc/fqx023).
- Ferri, Rolando (2003). *Octavia: A Play Attributed to Seneca*. Cambridge Classical Texts and Commentaries. Cambridge University Press.
- (Jan. 1, 2014). "Octavia". In: Brill, 521–527. [10.1163/9789004217089_043](https://doi.org/10.1163/9789004217089_043).
- Frank, M. (July 17, 2018). *Seneca's Phoenissae: Introduction and Commentary*. Brill. [10.1163/9789004329430](https://doi.org/10.1163/9789004329430).
- Gahan, John J. (1985). "Seneca, Ovid, and Exile". In: *The Classical World* 78.3, 145–147. [10.2307/4349723](https://doi.org/10.2307/4349723).
- Grieve, Jack (Sept. 1, 2007). "Quantitative Authorship Attribution: An Evaluation of Techniques". In: *Literary and Linguistic Computing* 22.3, 251–270. [10.1093/llc/fqm020](https://doi.org/10.1093/llc/fqm020).
- Hagiwara, M. (2021). *Real-World Natural Language Processing: Practical Applications with Deep Learning*. Manning.
- Herington, C. J. (1961). "Octavia Praetexta: A Survey". In: *The Classical Quarterly* 11.1, 18–30. [10.1017/S0009838800008351](https://doi.org/10.1017/S0009838800008351).
- Hoover, David L. (Nov. 1, 2004). "Delta Prime?" In: *Literary and Linguistic Computing* 19.4, 477–495. [10.1093/llc/19.4.477](https://doi.org/10.1093/llc/19.4.477).
- Jannidis, Fotis, Steffen Pielström, Christof Schöch, and Thorsten Vitt (2015). "Improving Burrows' Delta – An empirical evaluation of text distance measures". In: *Book of Abstracts of the Digital Humanities Conference 2015*. ADHO. UWS. http://dh2015.org/abstracts/xml/JANNIDIS_Fotis_Improving_Burrows_Delta_An_empirical/JANNIDIS_Fotis_Improving_Burrows_Delta_An_empirical.html.



- Johnson, Kyle P., Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly (2021). "The Classical Language Toolkit: An NLP Framework for Pre-Modern Languages". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 20–29. [10.18653/v1/2021.acl-demo.3](https://doi.org/10.18653/v1/2021.acl-demo.3).
- Jolliffe, Ian T. and Jorge Cadima (Apr. 13, 2016). "Principal component analysis: a review and recent developments". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065, 20150202. [10.1098/rsta.2015.0202](https://doi.org/10.1098/rsta.2015.0202).
- Juola, Patrick (Dec. 1, 2015). "The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions". In: *Digital Scholarship in the Humanities* 30 (suppl_1), i100–i113. [10.1093/llc/fqv040](https://doi.org/10.1093/llc/fqv040).
- Jurafsky, Dan and James H. Martin (2024). "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition". 3rd ed. draft. <https://web.stanford.edu/~jurafsky/slp3/> (visited on 05/03/2024).
- Karakasis, Evangelos (2018). *T. Calpurnius Siculus: A Pastoral Poet in Neronian Rome*. Vol. 35. Trends in Classics. De Gruyter. 335 pp. [10.33776/ec.v24i0.5007](https://doi.org/10.33776/ec.v24i0.5007).
- Kestemont, Mike (2014). "Function Words in Authorship Attribution. From Black Magic to Theory?" In: *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*. 3rd Workshop on Computational Linguistics for Literature (CLFL). Association for Computational Linguistics, 59–66. [10.3115/v1/W14-0908](https://doi.org/10.3115/v1/W14-0908).
- Kestemont, Mike, Sara Moens, and Jeroen Deploige (June 1, 2015). "Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux". In: *Digital Scholarship in the Humanities* 30.2, 199–224. [10.1093/llc/fqt063](https://doi.org/10.1093/llc/fqt063).
- Kestemont, Mike, Justin Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans (Nov. 30, 2016). "Authenticating the writings of Julius Caesar". In: *Expert Systems with Applications* 63, 86–96. [10.1016/j.eswa.2016.06.029](https://doi.org/10.1016/j.eswa.2016.06.029).
- Khonji, Mahmoud and Youssef Iraqi (2014). "A Slightly-modified GI-based Author-verifier with Lots of Features (ASGALF)". In: *CLEF (Working Notes)*, 977–983. <http://ceur-ws.org/Vol-1180/CLEF2014wn-Pan-KonijEt2014.pdf>.
- Koppel, Moshe, Jonathan Schler, and Shlomo Argamon (Jan. 1, 2009). "Computational methods in authorship attribution". In: *Journal of the American Society for Information Science and Technology* 60.1, 9–26. [10.1002/asi.20961](https://doi.org/10.1002/asi.20961).
- Koppel, Moshe, Jonathan Schler, and Elisheva Bonchek-Dokow (2007). "Measuring Differentiability: Unmasking Pseudonymous Authors." In: *Journal of Machine Learning Research* 8.6.
- Koppel, Moshe and Yaron Winter (Jan. 1, 2014). "Determining if two documents are written by the same author". In: *Journal of the Association for Information Science and Technology* 65.1, 178–187. [10.1002/asi.22954](https://doi.org/10.1002/asi.22954).
- Kuhn, Max and Kjell Johnshon (2016). "Over-Fitting and Model Tuning". In: *Applied Predictive Modelling*. 5th ed. Springer, 600.
- Luyckx, Kim and Walter Daelemans (Apr. 1, 2011). "The effect of author set size and data size in authorship attribution". In: *Literary and Linguistic Computing* 26.1, 35–55. [10.1093/llc/fqq013](https://doi.org/10.1093/llc/fqq013).

- Manousakis, Nikos (2020). ›Prometheus Bound‹ - A Separate Authorial Trace in the Aeschylean Corpus. De Gruyter. 10.1515/9783110687675. 703
704
- Marshall, C.W. (2014). "The Works of Seneca the Younger and Their Dates". In: Brill, 33–44. 10.1163/9789004217089_003. 705
706
- Marti, Berthe (1945). "Seneca's Tragedies. A New Interpretation". In: *Transactions and Proceedings of the American Philological Association* 76, 216–245. 10.2307/283337. 707
708
- Michalopoulos, Andreas N. (2020). "Seneca quoting Ovid in the Epistulae morales". In: *Intertextuality in Seneca's Philosophical Writings*. 1st ed. London: Routledge, 130–141. 709
710
- Mikros K., George (2018). "Blended Authorship Attribution: Unmasking Elena Ferrante Combining Different Author Profiling Methods". In: *Drawing Elena Ferrante's profile*. Padova University Press, 85–96. 711
712
713
- Newman, Matthew L., Carla J. Groom, Lori D. Handelman, and James W. Pennebaker (n.d.). "Gender Differences in Language Use: An Analysis of 14,000 Text Samples". In: *Discourse Processes* 45.3 (), 211–236. 10.1080/01638530802073712. 714
715
716
- Nolden, Luuk (July 19, 2019). "Finding Seneca in Seneca: using Text Mining techniques of Hercules Oetaeus and Octavia". Bachelor Thesis. Leiden, The Netherlands: Leiden Institute of Advanced Computer Science (LIACS). <https://theses.liacs.nl/pdf/2018-2019-NoldenLSJ.pdf>. 717
718
719
720
- Päpcke, Simon, Thomas Weitin, Katharina Herget, Anastasia Glawion, and Ulrik Brandes (Aug. 9, 2022). "Stylometric similarity in literary corpora: Non-authorship clustering and Deutscher Novellenschatz". In: *Digital Scholarship in the Humanities*, fqac039. 721
722
723
724
- Pease, Arthur Stanley (1920). "Is the "Octavia" a Play of Seneca?" In: *The Classical Journal* 15.7. Publisher: The Classical Association of the Middle West and South, 388–403. 725
726
727
- Perseus Digital Library (2024). Ed. Gregory R. Crane. Tufts University. <https://www.perseus.tufts.edu/hopper/> (visited on 05/14/2024). 728
729
- Philp, R. H. (1968). "The Manuscript Tradition of Seneca's Tragedies". In: *The Classical Quarterly* 18.1. Publisher: [Classical Association, Cambridge University Press], 150–179. <http://www.jstor.org/stable/637696>. 730
731
732
- Poe, Joe Park (1989). "Octavia Praetexta and Its Senecan Model". In: *The American Journal of Philology* 110.3, 434–459. 10.2307/295219. 733
734
- Potha, Nektaria and Efstathios Stamatatos (2017). "An improved impostors method for authorship verification". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings* 8. Springer, 138–144. 735
736
737
738
- Rybicki, Jan (2012). "The great mystery of the (almost) invisible translator: Stylometry in translation". In: *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research*. Ed. by Michael P. Oakes and Meng Ji. Studies in Corpus Linguistics. John Benjamins Publishing Company, 231–248. 10.1075/scl.51.09ryb. 739
740
741
742
743
- Rybicki, Jan and Magda Heydel (Dec. 1, 2013). "The stylistics and stylometry of collaborative translation: Woolf's Night and Day in Polish". In: *Literary and Linguistic Computing* 28.4, 708–717. 10.1093/lc/fqt027. 744
745
746
- Savoy, Jacques (2020). "Elena Ferrante: A Case Study in Authorship Attribution". In: *Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling*. Springer International Publishing, 191–210. 10.1007/978-3-030-53360-1_8. 747
748
749

- Seidman, Shachar (2013). "Authorship verification using the impostors method". In: *CLEF 2013 Evaluation labs and workshop—Working notes papers*, 23–26. 750
- Seneca (Apr. 17, 2008). *Octavia: Attributed to Seneca*. Place: Oxford Publisher: Oxford University Press. <http://oxfordscholarlyeditions.com/view/10.1093/actrade/9780199287840.book.1/actrade-9780199287840-book-1>. 751
- Singhal, Amit et al. (2001). "Modern information retrieval: A brief overview". In: *IEEE Data Eng. Bull.* 24.4, 35–43. 752
- Stamatatos, Efstathios (Mar. 1, 2009). "A Survey of Modern Authorship Attribution Methods". In: *Journal of the American Society for Information Science and Technology* 60, 538–556. [10.1002/asi.21001](https://doi.org/10.1002/asi.21001). 753
- Stamatatos, Ph D et al. (2013). "On the robustness of authorship attribution based on character n-gram features". In: *Journal of Law and Policy* 21.2, 7. 754
- Star, Christopher (Jan. 1, 2015). "Roman Tragedy and Philosophy". In: Brill, 238–259. [10.1163/9789004284784_013](https://doi.org/10.1163/9789004284784_013). 755
- Stover, Justin and Mike Kestemont (2016). "Reassessing the Apuleian Corpus: A Computational Approach to Authenticity". In: *The Classical Quarterly* 66.2. Edition: 2017/01/30 Publisher: Cambridge University Press, 645–672. [10.1017/S0009838816000768](https://doi.org/10.1017/S0009838816000768). 756
- Stover, Justin, Yaron Winter, Moshe Koppel, and Mike Kestemont (Jan. 1, 2016). "Computational authorship verification method attributes a new work to a major 2nd century African author". In: *Journal of the Association for Information Science and Technology* 67.1, 239–242. ISSN: 2330-1635. [10.1002/asi.23460](https://doi.org/10.1002/asi.23460). 757
- Tarrant, Richard (2017). "Custode rerum Caesare: Horatian Civic Engagement and the Senecan Tragic Chorus". In: *Interactions, Intertexts, Interpretations*. Ed. by Martin Stöckinger, Kathrin Winter, and Andreas T. Zanker. De Gruyter, 93–112. [doi:10.1515/9783110528893-005](https://doi.org/10.1515/9783110528893-005). 758
- The Latin Library* (2024). <http://www.thelatinlibrary.com/> (visited on 05/13/2024). 759
- VanderPlas, Jake (2017). "In Depth: Principal Component Analysis". In: *Python Data Science Handbook*. O'Reilly Media, Inc., 433–445. 760

Repetition and Innovation in Dramatic Texts

An attempt to measure the degree of novelty in character's speech

Botond Szemes¹ 
Mihály Nagy² 

1. Institute for Literary Studies, HUN-REN Research Centre for the Humanities, Budapest, Hungary.
2. Doctoral School of History, Eötvös Loránd University, Budapest, Hungary.

Citation

Botond Szemes and Mihály Nagy (2024). "Repetition and Innovation in Dramas. An attempt to measure the degree of novelty in character's speech". In: *CCLS2024 Conference Preprints 3* (1). [10.26083/tuprints-000273](https://doi.org/10.26083/tuprints-000273) 95

Date published 2024-05-28

Date accepted 2024-04-03

Date received 2024-01-25

Keywords

computational drama analysis, information theory, innovation, sentence embedding, Shakespeare

License

CC BY 4.0 

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 3rd Annual Conference of Computational Literary Studies at Vienna, Austria, in June 2024.

Abstract. In the following, we develop a method to study dramas as information networks. We examine how innovative characters are in relation to each other, i.e. whether they tend to repeat the utterances of others or introduce new information to the discourse of the play. Our method captures the role of characters in this discourse, and through pairwise comparisons, we can also construct networks that represent character relationships in a new way compared to existing approaches. By examining some of Shakespeare's plays, we also identify general patterns regarding the structural differences of the networks and gender roles in comedies and tragedies/non-comedies.

1. Introduction

In dramatic works, the flow of information maintained by the speech acts of the characters is particularly important. In terms of the *internal communication system*, the flow (or the withholding) of information between characters is the driving force of the plot (Andresen et al. 2022, 2024); in terms of the *external communication system*, the audience/readers gain access to the storyworld also mostly through the dialogues (for theoretical description of the two types of systems, see Pfister 1988). Accordingly, co-presence or co-occurrence networks (Trilcke 2013; Trilcke et al. 2015), which have become increasingly popular in recent years, are also often interpreted from the perspective of the internal information flow, although usually implicitly, as in the case of using betweenness centrality as a metric to infer the mediating, even "conspiratorial" role of characters (e.g. Algee-Hewitt 2017; Szemes and Vida 2024). Benjamin Krautter, however, points out that knowledge networks, which represent the transfer of knowledge between characters, and which may well show a different arrangement than co-presence networks, are more helpful and theoretically better grounded in such an investigation of information flow (Krautter 2023, see also Andresen et al. 2022).

In contrast to these approaches, the present study analyses the information value of characters' speeches in Shakespeare's works from the perspective of the *external communication system*, i.e. from the perspective of the recipient. Andresen et al. 2022 also took this aspect into account in their research, albeit in less detail and focusing on just a specific type of knowledge transmission. Furthermore, we do not follow Manfred Pfister's theory (Pfister 1988) strictly in our analysis as they did. That is, we do not only consider

utterances when a character conveys specific knowledge to the audience;¹ rather, we consider all utterances according to the extent to which they add new meanings to the storyworld. When in *Hamlet*, for example, Claudius raises the idea of Hamlet's exile, the information value of the speech is increased by the mentioning of England (and its relationship to Denmark) for the first time in the play – the horizon of the storyworld is literally expanded. However, Denmark's foreign policy relations (with Norway) have been discussed before, so the difference from the earlier discourse is not that great. Equally, it can be informative if a character speaks in a new register, different from previous ones, since this shows that such ways of speaking are in fact possible in the represented world, and that these as contexts influence the interpretability of other utterances as well. Consider, for example, the differences between the royal speech at the beginning of *Hamlet*'s second scene and the sentences exchanged between Horatio and his companions in the first scene, or the dialogue of the Gravediggers in Act 5. The tensions between the royal propaganda and the friendly or humorous remarks create the framework in which the tragedy unfolds. The Gravediggers' sentences about Hamlet's exile are less novel, however, as this is already mentioned earlier in the play (see the comparison of sentences from these characters in Appendix 2.) Together, we refer to these types of differences from the previous discourse as *semantic difference*, which according to our experiments can be captured well with the use of BERT-based language models. The term indicates a focus on the content of the dialogues, but also a consideration of the semantic components of style (for example, a highly metaphorical utterance is usually more distinct from sentences that elaborate the meaning less metaphorically.)

In light of this, we are interested in the role that a character plays in shaping the storyworld. Two general functions can be distinguished according to the extent to which they contribute to the creation of new meanings by often deviating from what has been said before, or to the extent that they repeat and thus reinforce an already established discourse. *Innovative characters* are responsible for the elaboration of new (semantically distinct) meanings, while *repeaters* or *maintainers* contribute to the development of the central themes and the general ways of speaking in the drama. There is, of course, also a duality of innovation and repetition within each individual character. This can also be detected with our method, since we calculate the semantic difference between each sentence and its preceding discourse for each character, which makes it possible to examine the distribution of both functions in the cast separately. This sentence-level approach can also help us to answer the question of what the innovative function of a character means in a specific case beyond the broad definition. In this paper, we argue that Shakespeare's innovative characters can be divided into two groups: those who are in fact responsible for transmitting knowledge, and those who speak in a different way from the dominant discourse in the drama, usually expressing uncertainty and/or emotion, or using metaphorical language. Our results, furthermore, provide a novel way of describing the difference between comedies and tragedies (or more precisely "non-comedies"²). Namely that female characters in Shakespeare's comedies are more likely to have innovative functions and be repeated by others compared to tragedies.

1. Pfister's example is Prospero's speech to Ariel in the beginning of *The Tempest* (1/ii, 250-293), which is more informative for the audience, since Ariel already knew everything that was in the speech.

2. Dramas labelled as „comedy“ are those that are listed as such in the First Folio (1623). All others are labelled as „non-comedy“ or sometimes in the paper as „tragedy“ for the sake of simplicity. For the structural similarities of the „non-comedies“ (and their resemblance to tragedies) see Szemes and Vida 2024

Finally, the paper also addresses the question of the network representation of character relations. Benjamin Krautter has pointed out that the interpretability of networks is significantly affected by the type of relations they represent – different methods lead to different conclusions (Krautter 2023). In the following, we present a new method intended to complement already existing ones. It is based on defining the innovativeness of a character’s speech along pairwise comparisons, i.e. comparing characters with each other separately. On the one hand, this makes it possible to measure the similarities between two characters at sentence level. On the other hand, it allows us to represent the relationships on a directed graph, showing which character in the pairwise comparison is more likely to repeat the other. Similarly to Andresen et al. 2022, we attempt to use “a more content-based form of character networks [...] to chart a path to better integrate quantitative analysis and interpretative reading.” In the resulting networks, the role played in the whole discourse of the drama and the relationship between two characters can be examined simultaneously.

2. Related Works

The paper draws from previous research within information theory that has likewise attempted to measure innovation and repetition in different communicative situations. However, these studies differ not only in their methods, but also in their theoretical assumptions. As well as in their understanding of the terms ‘information’, ‘novelty’, or ‘innovation’. Therefore the paper must be situated within previous research and define its subject of measurement – i.e. how it considers the concept of ‘innovation’ to be operationalised in the study of dramas.

South et al. 2022 analyzed repeated linguistic elements to detect the flow of information between Twitter accounts of news organizations. They assume that when more words exist in the same order across two texts, the degree of novelty between them is lower, and vice versa that previously unused phrases and novel word order make a text innovative. Accordingly, their method is based on the identification of the longest repeated sequences of words. This approach functions well in the case of Twitter posts, however, when applied to less homogenous and considerably more poetic dramatic texts, it is less useful. This is because in such texts, repeating sequences almost in all cases are conventionalised expressions (e.g.: ‘there are’, ‘good morning’). Therefore, the results would not primarily indicate semantic similarity.

Sims and Bamman 2020 also set out to explore recurring linguistic elements when determining the role of characters in a novel’s social and information networks. Beyond considering the mere frequency of words, they also examined POS tags and grammatical relations. Using a selection of verbs that describe the most important events of a plot, they identified ‘Subject – Verb – Object’ triples (e.g.: ‘Thomas – left – Vienna’) – if a triple is mentioned by two characters, we can say that they refer to the same event so that the former has an *informational impact* on the later. The challenges of the method include inaccuracies in co-reference resolution (which assigns each utterance to the corresponding character, although this is much simpler in dramatic works) and in dependency analysis, as well as the somewhat arbitrary selection of the group of verbs to be considered. Whereas Sims and Bamman 2020 sought to explore the direct effect between characters

(internal communication system), we interpret innovation and repetition in relation to the entire discourse preceding an utterance (external communicational system): even though we make pairwise comparisons, we do not assume that the similarity of two characters' utterances indicates a direct causal relation; we just examine the extent to which the content of an utterance is similar to what was said before.

The same question was asked by Barron et al. 2018, who measured whether speeches by members of the Parliament during the French Revolution had raised new themes or contributed to maintaining previous ones. Their approach applies Kullback–Leibler Divergence (KLD), a measure often used in similar contexts due to its strong foundation in information theory. In short, with KDL the difference between the vector representation of texts is not calculated through the spatial metaphor of distance (how far one text is from another in a vector space), but through a model of *experience* (how surprising a text is when conditioned on prior knowledge - see Chang and DeDeo 2020). Barron et al. 2018 first determined the distribution of different topics across parliamentary speeches, then compared these distributions with the help of KLD. A similar attempt was made by Piper et al. 2023 who, on the other hand, used a simple distribution of word frequencies of equal-length chunks to calculate their divergence, through which they could measure the process of narrative revelation.

Since the comparison of texts in this study is based on their semantic relations, neither the consideration of the longest recurring sequences nor word frequency distributions proved to be useful approaches. Similarly, doing topic modelling like Barron et al. 2018 also proved impractical, because in the case of a drama, the utterances are usually too short to effectively identify themes in them. Nor does one drama provide enough data to distinguish the characters efficiently according to the distribution of themes. Therefore, we use Large Language Models (LLMs) to determine the position of each sentence of a drama within a vector space representing the semantic field of the given language. The embedding process is driven by the SBERT (Sentence-BERT) algorithm, which can quantitatively capture the meaning of larger units, such as sentences, compared to the word-level embeddings of previous BERT models (Reimers and Gurevych 2019). The vector representation of separate sentences makes their semantic comparison possible, which can be utilized in our research to examine the character speeches based on their content. *Semantic similarity* refers mainly to thematic similarities, but also includes the style of the sentences (e.g. terms belonging to the same style/register are semantically more similar). In light of this, we can say that semantically the less similar a sentence is to its predecessors, the greater the degree of information it conveys (innovativeness). Conversely, the more similar a sentence is to its predecessors, the more it contributes to the repetition of an already existing discourse.

This was the approach also used by Dubourg et al. 2023 in their study measuring the innovation of movie plots. Converting the plot summaries of over 19,000 films into vectors with the help of the SBERT algorithm, they calculated the cosine similarity between a summary and all preceding film summaries and averaged them to determine a film's Innovation Score, i.e. the average distance of the current embedding from previous ones. Our method compares the sentences spoken by characters in a similar way. It is important to note because Dubourg et al. 2023 also evaluated the method and found their results to be positively correlated with results from text mining of viewer

reviews (see Luan and Kim 2022). In our case such a comparison is not possible due to the lack of other results and because, as we have seen, the procedures mentioned so far cannot be adapted without problems to answer our research question.

Indeed, so far in the field of quantitative drama analysis, there have not yet been any attempts to answer such a question relating to repetition and innovation in a character's speech. Most of the previous research investigated primarily the structural characteristics of plays (for an overview: Szemes and Vida 2024); while other, more language-oriented investigations have mostly experimented with topic-modelling of larger corpora (and explore genre differences - see Schöch 2017), and regarding Shakespeare's works most attention has been paid to authorial style and keyword analysis (Craig and Kinney 2009), or uncovering changes in word use in the oeuvre (Hope and Witmore 2014). The closest to the research is that of Andresen et al. 2022 and Krautter 2023, with the differences already mentioned in the *Introduction*. It is also important to refer to the research of Šeĉa et al. 2024, in which they used stylometric methods developed for authorship attribution to calculate the difference between characters' speeches. However, their focus was not on the semantic content of the texts and their degree of innovation, but exclusively on their stylistic differences. We hope, therefore, that our study will provide new perspectives to the field, and at the same time enrich the interpretability of certain plays.

3. Method

For our study, we used dramas from Shakespeare in TEI-XML format provided by the Drama Corpus Project (Fischer et al. 2019).³ As a first step we created a tabular representation of all the individual sentences from a play. We assigned to each sentence 1) the name of the character, 2) a timestamp representing the position of the spoken text within the whole drama (from 1 to the last sentence), 3) the number of the act in which the sentence is spoken, and 4) the embedding score provided by a language model. Regarding the last point, the selection of the right model is a primary concern. Using example sentences taken from the corpus, we experimented with several state-of-art best-performing SBERT models.⁴ We selected sentences with similar and dissimilar meanings (at this stage we judged similarity intuitively and the selection was made manually), and calculated their cosine similarity in a pairwise manner. Subsequently, we calculated the standard deviation of the similarities. Although there was a minimal variation between the models, we chose to use the popular 'all-MiniLM-L6-v2', as its results showed the highest standard deviation, which means that the distribution among similar and dissimilar meanings are the largest in this case. See the experiment details and the performance of the chosen model in the project's GitHub repository (*Software availability*) where the performance can also be evaluated manually by looking at the most/least similar sentence pairs of the plays (see also the *Appendix* and the *Results* sections for further manual evaluation.) Regarding the most similar sentences, for example, character names seem to have a strong influence on sentence similarity. The names could have been therefore filtered out during the pre-processing stage, but it was considered worth keeping them because of their role in the creation of meaning. At the same time, sentences with fewer than four words (e.g., "Yes, sir") were excluded, as they

3. <https://dracor.org/shake>

4. See the list of best-performing models: https://www.sbert.net/docs/pretrained_models.html

are less likely to convey relevant meaning, but are rather conventionalised expressions. 195

We then created pairs from the most frequent speakers (i.e. the main characters⁵) in 196
a specific order: the first member of the pair became the *Source*, and the second the 197
Target character. During their comparison, we calculated the cosine similarity between a 198
Target-sentence and all the preceding Source-sentences. In contrast to the method of 199
Dubourg et al. 2023, we did not take the average of these similarities but only selected 200
the largest of them to characterize semantic proximity. Thus, for each sentence of the 201
Target character, we assigned a number indicating *how semantically similar it is to the most* 202
similar of the previous sentences of Source (Maximum Cosine Similarity - MCS). It can 203
be assumed that the higher the number, the less innovative the meaning of the sentence 204
since it repeats previous content. 205

There are several arguments for using the Maximum Cosine Similarity instead of the 206
average. Firstly, if a Source character speaks on many different topics in many different 207
registers before the current Target-sentence, then on average this Target-sentence will 208
be less similar, even if the Source character has spoken the same sentence before. MCS 209
avoids this by focusing on the maximum value, however, this also means that the result 210
does not report on *how often* the Source character has elaborated similar meanings. 211
Secondly, MCS values can be used to find the most similar sentence pairs between 212
Source and Target, contributing to the overall interpretability of the results. Thirdly, the 213
average cosine similarity (as Dubourg et al. 2023 also point out) is strongly influenced 214
by temporality: the later the utterance, the more similar it is on average to the earlier 215
discourse (see Fig 1a). Therefore, by using the average cosine similarity, we would 216
measure more the time in the plot at which a character speaks, than the novelty of his or 217
her sentences. The MCS is also exposed to temporality, but to a much lesser extent (Fig 218
1b), and the effect can be compensated for by weighting/adjusting the values (Fig 1c). 219
To do this, we first calculated the average MCS value for each act and for the drama as a 220
whole, and then used the difference between the values for the acts and for the drama 221
to weigh the scores according to the act in which the sentence was uttered. For example, 222
the sentences in the first act were weighted by the difference between the average MCS 223
for the first act and the drama as a whole. At the same time, a high degree of variation 224
can be seen in the dataset: sentences with high MCS values can be found in the first act 225
just as much as low ones at the end of a drama. 226

In the next step, we assigned the average of the weighted MCS scores to each Source- 227
Target pair and performed network normalization on the dataset following the method- 228
ology developed by South et al. 2022. The key consideration here is that if character 229
"B" frequently repeats character "A", but character "A" also repeats other characters, 230
then character "B" is indirectly connected to such other characters as well. To conduct 231
our network normalization, we determined the average score of a given character as 232
Target, and then divided all similarity scores by this number where this character was 233
the Source. 234

Finally, we calculated the differences for character pairs depending on which character 235

5. Main characters are considered those with more than 30 long sentences for shorter plays (less than 1000 long sentences), more than 40 for plays with medium length (number of long sentences between 1000 and 1700), and more than 50 for longer plays. Occasionally, individual considerations may also come into play, for example if a character speaks a lot but only in one scene (e.g. the Gravediggers in *Hamlet*).

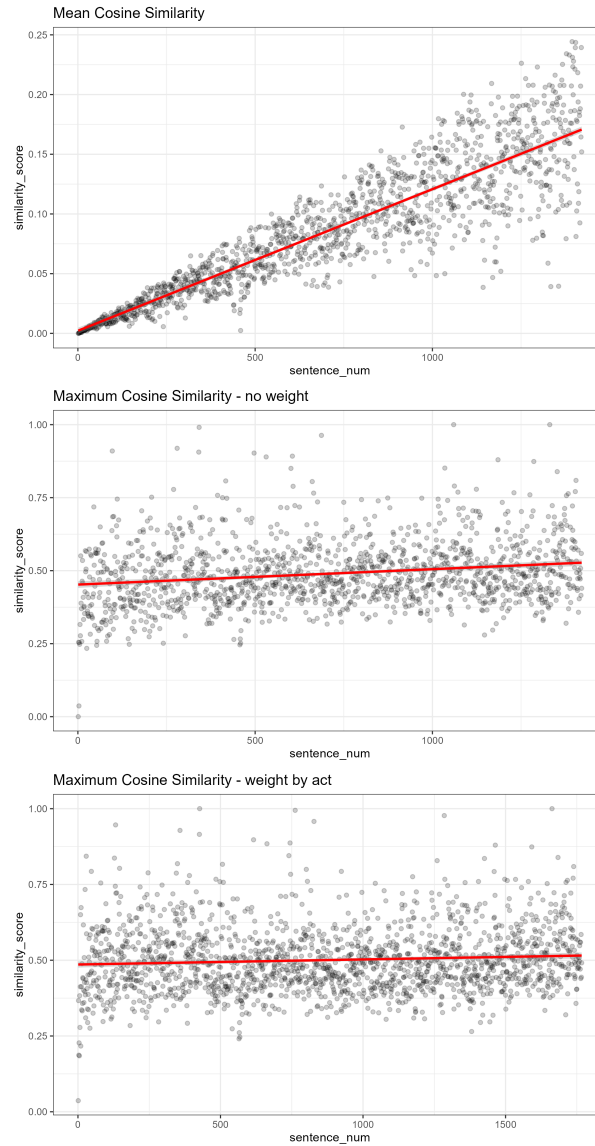
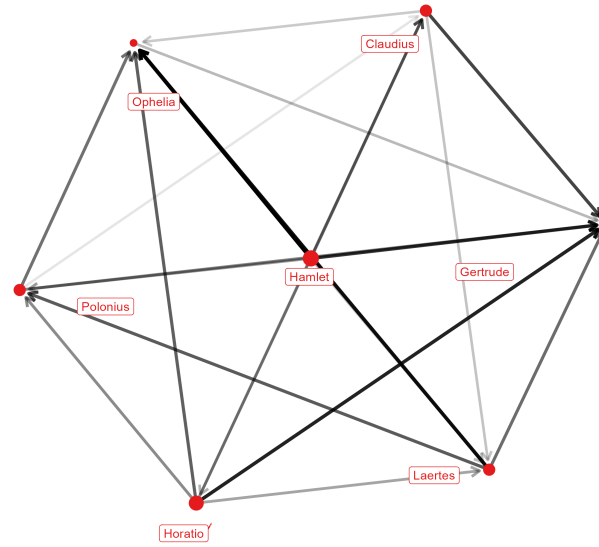


Figure 1: The relationship between time of utterance and similarity score in *Hamlet*. Up: Mean Cosine Similarity, Middle: Maximum Cosine Similarity - without weight, Down: Maximum Cosine Similarity - weight by act.

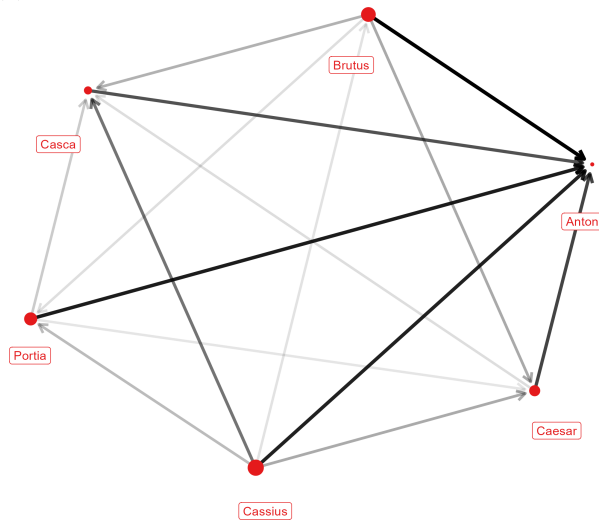
is listed as the Source or Target (e.g. Hamlet-Claudius vs. Claudius-Hamlet). If the difference is positive, then the Target character's sentences are more likely to develop a similar meaning to the Source character's earlier sentences than vice versa - i.e. the Source character is considered more innovative in their relationship. As a final result, only these positive values were retained and used for network visualization.

4. Results

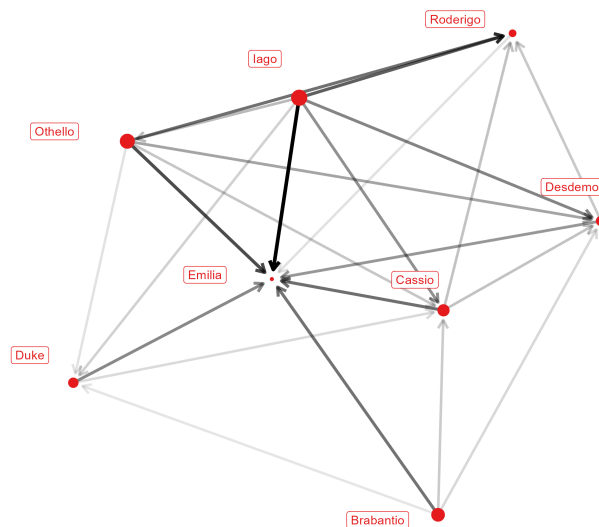
The results allow us to visualize the relationships between characters in terms of repetition and innovation as a network. In the example networks seen in Figure 2, the arrows go from Source to Target (indicating which character is more likely to repeat the other), their thickness is determined by the degree of similarity/repetition, and the size of the nodes as an innovation score indicates how often the character is listed as Source, i.e.



(a) *Hamlet*

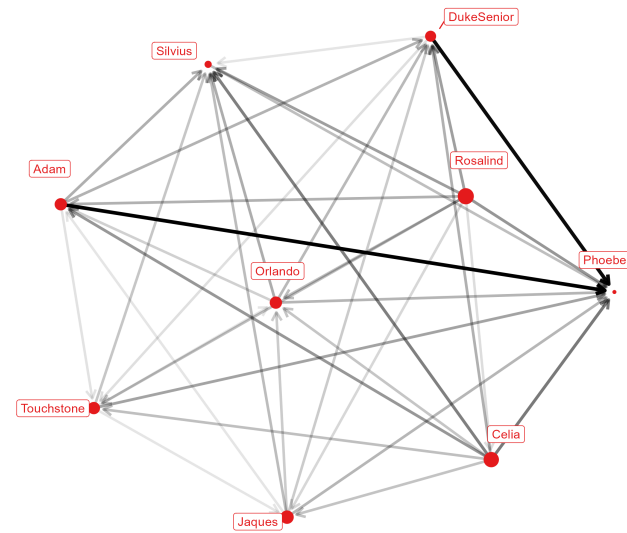


(b) *Julius Caesar*

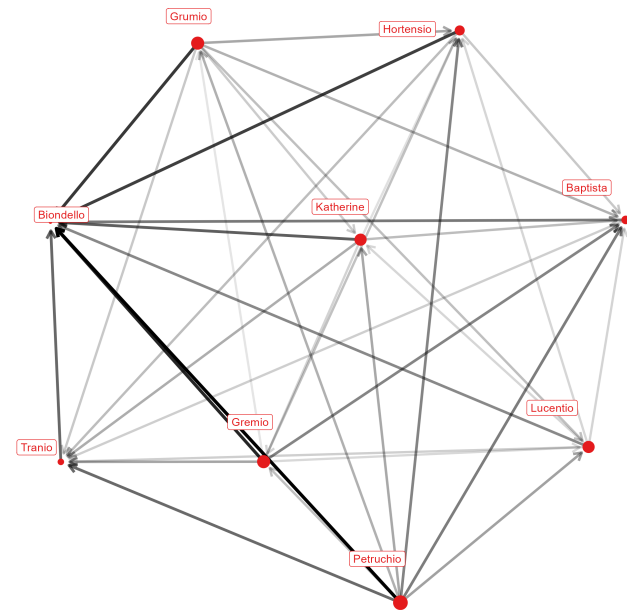


(c) *Othello*

Figure 2: Networks of Shakespeare's plays.

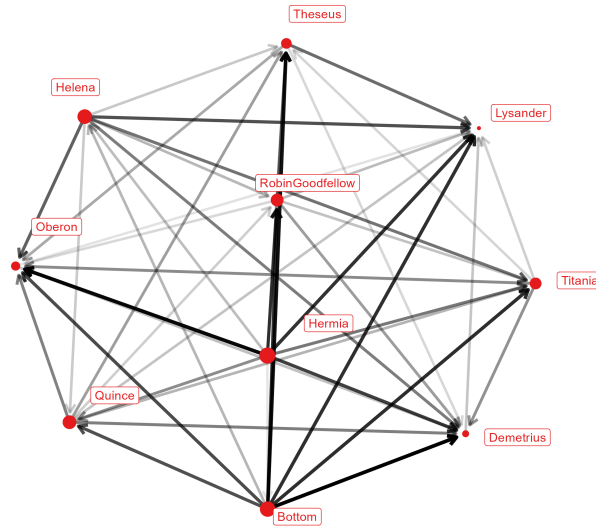


(d) *As You Like It*



(e) *The Taming of the Shrew*

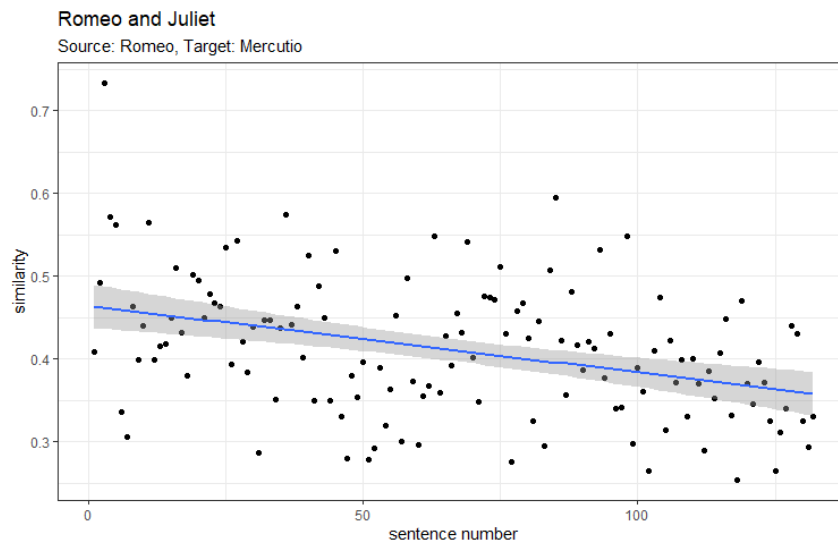
Figure 2: Networks of Shakespeare's plays.



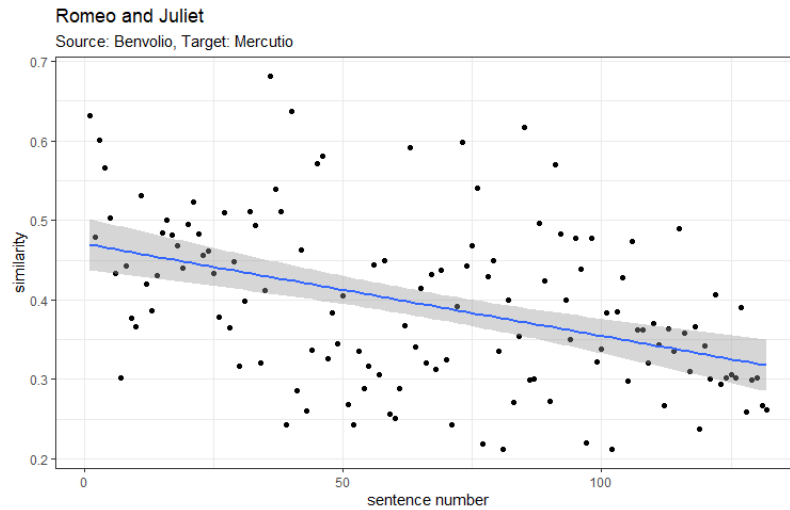
(f) *A Midsummer's Night Dream*

Figure 2: Networks of Shakespeare's plays. The arrows go from Source to Target (indicating which character is more likely to repeat the other), their thickness is determined by the degree of similarity/repetition, and the size of the nodes indicates how often the character is considered innovative in pairwise comparisons.

how often it is considered innovative in pairwise comparisons. The latter is influenced 247
by both the number of observed sentences and partly the time of utterance: the chance 248
of a character being novel is increased by speaking both earlier, and on more occasions. 249
Even though we applied the above-mentioned weighting method, characters that speak 250
mainly in the second half of the plot generally received lower innovation points (e.g. 251
Antonius in *Julius Caesar* or Emilia in *Othello*). We do not see this as a measurement bias 252
but as a characteristic of a character type. This is supported by the fact that there are 253
also examples where as the plot progresses one character becomes increasingly different 254
from another, such as Mercutio, the character with the highest innovation score in *Romeo* 255
and Juliet, compared to both Romeo and Benvolio, the characters with the second and 256
third highest scores, respectively (Figure 3). 257



(a) Target = Mercutio, Source = Romeo



(b) Target = Mercutio, Source = Benvolio

Figure 3: Changes in maximum cosine similarity over time between the most innovative characters in *Romeo and Juliet*. Mercutio's sentences become less similar to others.

The overall examination of Shakespeare's dramas shows that the relationship between characters is in most cases hierarchical (i.e. the characters can be ordered hierarchically according to their innovation scores). This is particularly true for tragedies/non-comedies, where the characters with the highest innovation scores can almost always be arranged in a hierarchical way, and only at lower levels can equal scores be found. Equal scores mean that there is a degree of circularity in the dramas: character "A" tends to repeat "B", "B" repeats "C", whereas "C" repeats "A" etc. At a higher level, this happens mainly in comedies (among non-comedies, in *Cymbeline*, *Macbeth* and *Pericles*, a play with much debated genre). For example, in *The Taming of the Shrew* Grumio and Gremio, and also Lucentio and Katharine; in *As You Like It* Orlando, Adam and Touchstone; in *Measure for Measure* Duke, Lucio and Angelo take on the same values. This difference between genres is in line with previous results based on co-occurrence networks, which show that comedies are characterized by a denser system of relationships, while tragedies by one or two characters with a connecting function who control the social relations (more hierarchical distribution of node degrees). This also means that in comedies there are many misunderstandings and parallelisms (two characters connected by different paths) during the interactions, however, for the same reason such networks are "protected" from falling apart when a certain piece of information is revealed to be untrue. In contrast, information flow is effective and fast in tragedies, but the networks themselves are fragile, as the failure of a connecting character can lead to the disintegration of the whole system (cf. Szemes and Vida 2024).

All of this is further nuanced by another distinction between genres based on our measures. It is striking that in the 23 non-comedies the characters most repeated by others are males (except Imogen in *Cymbeline* and Lady Macbeth who is as innovative as Macbeth and Banquo), while in comedies, female characters are more likely to be the most innovative (six times out of 14). In *As You Like It* Rosalinda (and Celia in the second place) has the highest score; in *All's Well That Ends Well* the Countess (and Helen in the second place), in *The Comedy of Errors* Adriana; in *A Midsummer Night's Dream* Hermia (and Helena in the third place, while their counterparts, Lysander and

Demetrius have the lowest innovation scores among the main characters); in *Much Ado* 287
About Nothing Beatrice, and maybe most surprisingly in *The Tempest* Miranda ahead of 288
Gonzalo and Prospero. We can say, that in the two kinds of communities, those who 289
thematise the discourse (or at least who is repeated more than he or she repeats others) 290
appears to differ, although not exclusively, in terms of gender. Women are more likely 291
to play that role in the protected networks of the comedies, and men in the effective but 292
vulnerable tragedies. 293

It is also worth looking at the results of pairwise comparisons in more detail and 294
identifying the most and least similar sentences between characters. In addition to a 295
qualitative evaluation of the method, this can also contribute to a close reading of the 296
dramas and a deeper understanding of the characters. As an example, in *Hamlet*, the 297
model grasps exactly the essential duality of the main character: he is striving to define 298
himself and others but, at the same time, is constantly doubting such identifications. 299
Hamlet's sentences which are most similar to the earlier utterances of the other characters, 300
are often about defining his own and others' identity; while his most different and 301
innovative sentences report doubt and uncertainty, often in a conditional or interrogative 302
mood (Table 1; see our GitHub repository for all the sentences and their most/least 303
similar pairs from other characters).⁶ 304

High similarity, low innovation	Low similarity, high innovation
This is I, Hamlet the Dane.	I doubt some foul play.
The King is a thing -	I would I had been there.
O God, Horatio, what a wounded name, Things standing thus unknown, shall I leave behind me!	Do they hold the same estimation they did when I was in the city?
If Hamlet from himself be ta'en away, And when he's not himself does wrong Laertes, Then Hamlet does it not; Hamlet denies it.	The time is out of joint.
Here comes the King, The Queen, the courtiers.	These foils have all a length?

Table 1: Examples of the least and most innovative sentences spoken by Hamlet as Target (Hamlet)

Hamlet's speech is most similar to the discourse of the court when he names or identifies 305
someone/something, and most divergent when he questions or is uncertain. Since he is 306
considered the most innovative in the drama, we can say that his sentences about doubt 307
are predominant, and they give the essence of his character – but it is also important to 308
see his statements in the opposite direction. Conversely, the most innovative sentences by 309
Horatio, the second most innovative character in the drama, do not express uncertainty. 310
He is rather the one who brings news to others and often speaks as an *eyewitness* – in 311
this sense, he really creates new information, not just develops semantically divergent 312
meanings (Table 2). These sentences illustrate well his dramaturgical function of linking 313
events and communities (cf. Moretti 2011). 314

6. The example sentences reported here have been hand-picked for interpretation from the 10 sentences with the highest and lowest cosine distance in the pairwise comparisons. The selection is therefore somewhat arbitrary: it is analogous to a researcher trying to make sense of the output of keyword analysis or topic modelling. The full list is given in the project's GitHub repository.

Low similarity, high innovation

Not when I saw 't.

My lord, I think I saw him yesternight.

Indeed, I heard it not.

It was as I have seen it in his life,
A sable silvered.

It would have much amazed you.

Table 2: Examples from the most innovative sentences spoken by Horatio (*Hamlet*)

Utterances expressing doubt, reflecting on either mental states like emotions or the out- 315
side world appear as most divergent in other characters from other dramas as well. One 316
example is Hermia in *A Midsummer Night's Dream* (Table 3), who is the most innovative 317
character in the drama precisely because of questioning the nature of things around 318
her (even compared to Bottom who appears in a subplot separate from the majority of 319
the cast and, therefore often speaks about something else). Furthermore, the duality 320
observed in *Hamlet* is also characteristic of Brutus in *Julius Caesar*. His most similar 321
sentences to the previous discourse are predominantly about the murder; whereas the 322
least similar ones are about doubts and emotions (Table 4). It is worth comparing this 323
with the utterances of Caesar, who only briefly expresses doubt, specifically about going 324
to the Senate (his most innovative utterances), and instead accepts his death to maintain 325
the conventional image of the emperor. This is shown by the fact that he often speaks of 326
himself in the singular third person: "Caesar shall forth."; "Danger knows full well/ 327
That Caesar is more dangerous than he." etc. 328

Characters with connecting functions like Horatio can be found also in other plays, 329
whose novelty lies in their reports about specific events. Such is Cassius in *Julius Caesar*, 330
who can be seen as an innovator even compared to Brutus. His sentences with the 331
highest/lowest MCS score show an opposite pattern to Brutus: he repeats the others 332
when he uses terms referring to emotions and inner values, while his sentences about 333
concrete events differ the most (Table 5). Cassius is in charge of moving the plot forward, 334
bringing news and argument – he also recruits the wavering Brutus into the conspiracy. 335
Part of it is that when Cassius speaks of emotions, he is not talking about himself, but 336
about others. On the other hand, the sentences of Brutus that mark specific events, refer 337
not to the conspiracy but to the murder itself; they are often retrospective and thus less 338
novel. Until the murder takes place, or until he is determined to commit it, he speaks of 339
more abstract topics, demonstrated by one of his most divergent sentences relative to 340
Caesar: „Between the acting of a dreadful thing/ And the first motion, all the interim 341
is/ Like a phantasma or a hideous dream.” 342

Low similarity, high innovation

Who is 't that hinders you?

Then I well perceive you are not nigh.

I understand not what you mean by this.

Too high to be enthralled to low.

Nothing but "low" and "little"?

Table 3: Examples of the most innovative sentences spoken by Hermia (*A Midsummer Night's Dream*)

High similarity, low innovation	Low similarity, high innovation
Mark Antony, here, take you Caesar's body.	I would not, Cassius, yet I love him well.
And for Mark Antony, think not of him, For he can do no more than Caesar's arm When Caesar's head is off.	That you do love me, I am nothing jealous.
I killed not thee with half so good a will.	If I have veiled my look, I turn the trouble of my countenance Merely upon myself.
Hold, then, my sword, and turn away thy face While I do run upon it.	But if these – As I am sure they do - bear fire enough To kindle cowards and to steel with valor The melting spirits of women, then, countrymen, What need we any spur but our own cause To prick us to redress?
But, alas, Caesar must bleed for it.	Enjoy the honey-heavy dew of slumber.

Table 4: Examples of the most and least innovative sentences spoken by Brutus (*Julius Caesar*)

High similarity, low innovation	Low similarity, high innovation
Yet I fear him, For in the engrafted love he bears to Caesar -	The clock hath stricken three.
Well, Brutus, thou art noble.	The morning comes upon 's.
I blame you not for praising Caesar so.	And I do know by this they stay for me In Pompey's Porch.
Caesar doth bear me hard, but he loves Brutus.	When went there by an age,] since the great flood, But it was famed with more] than with one man?
I know that virtue to be in you, Brutus, As well as I do know your outward favor	No, it is Casca, one incorporate To our attempts.

Table 5: Examples of the most and least innovative sentences spoken by Cassius (*Julius Caesar*)

Finally, it is worth highlighting *Othello*, in which Iago is associated with the highest innovation score. This is not surprising as he increasingly controls the discourse as the plot develops, and in some cases even makes others, especially Othello, repeat his sentences (e.g. “Men should be what they seem” [Iago], “Certain, men should be what they seem.” [Othello]; “Or to be naked with her friend in bed/ An hour or more, not meaning any harm?” [Iago], “Naked in bed, Iago, and not mean harm?” [Othello]). The sentences of Othello that differ most from Iago’s previous utterances are at the end of the drama. In these, he describes his situation using more abstract language, which may indicate that by the end of the plot, he will be able to view events from an external and broader perspective (Iago’s mastery of always focusing his attention on the concrete signs). However, this may also indicate that he is still incapable of introducing novel information about the concrete storyworld, and thus becomes innovative compared to Iago just when he refrains from naming things, as Iago does it instead of him. This is exemplified by one of Othello’s less similar sentences said to Desdemona: “Let me not name it to you, you chaste stars.”

5. Conclusion

Comparing sentence-level embeddings of character utterances can be useful both for interpreting specific dramas and for identifying general patterns in bigger corpora. According to the method proposed in the paper, characters whose sentences are the most semantically different from the previous sentences of other characters can be considered innovative. In this case, the degree of difference is measured by Maximum Cosine Similarity of embedding scores of a language model (how similar the most similar sentence is), rather than the average distance from all the previous sentences. The networks resulting from pairwise comparisons present the relationships between characters and provide at the same time a new way of describing the difference between Shakespeare’s comedies and non-comedies. While in non-comedies that are more hierarchical in terms of the distribution of innovation scores, the male protagonists’ speeches are repeated by others, whereas in more circular comedies, female characters are more likely to thematise the discourse of the play.

When analyzing the sentence pairs with the highest/lowest similarity scores, two types of characters seem to be distinguishable in Shakespeare’s plays, both of which can be considered innovative. On the one hand, some characters often introduce new information into the discourse and report on events distant in time or space. For example, Horatio in *Hamlet* as an eyewitness to various events functions as a link between groups; Cassio in *Julius Caesar*, the main organizer of the conspiracy; and Bottom in *A Midsummer Night’s Dream*, who also connects a subplot with the main characters. Others don’t bring new information into the discourse in the traditional sense, i.e. they do not talk about something different, but in a *different way*. This may be the result of the doubt in the established relations and identities (for example, Hamlet on the question of identity, Hermia on the perception and interpretation of the outside world), the predominance of emotions (Brutus), or the use of puns and a language with erotic connotations (Mercutio). In this context, the difference between abstract and concrete sentences also seems to be a general pattern: the more poetic and abstract an utterance is, the more innovative it appears.

6. Appendix - Cosine Similarity Scores 387

6.1 Similar and Dissimilar Sentences from Hamlet Used to Model Comparison 388 389

- Sentences: 390
1. How now, what noise is that? 391
 2. Alack, what noise is this? 392
 3. Exchange forgiveness with me, noble Hamlet. 393
 4. O Hamlet, speak no more! 394
 5. To die, to sleep—\No more—and by a sleep to say we end\The heartache and the thousand natural shocks\That flesh is heir to—’tis a consummation\Devoutly to be wished. 395 396
 6. This gentle and unforced accord of Hamlet\Sits smiling to my heart, in grace whereof\No jocund health that Denmark drinks today\But the great cannon to the clouds shall tell,\And the King’s rouse the heaven shall bruit again,\Respeaking earthly thunder. 397 398 399
 7. To be or not to be, that is the question:\Whether ’tis nobler in the mind to suffer\The slings and arrows of outrageous fortune,\Or to take arms against a sea of troubles And, by opposing, end them. 400 401 402
 8. Though yet of Hamlet our dear brother’s death\The memory be green, and that it us befitted\To bear our hearts in grief, and our whole kingdom\To be contracted in one brow of woe,\Yet so far hath discretion fought with nature\That we with wisest sorrow think on him\Together with remembrance of ourselves. 403 404 405 406
 9. Ay, truly, for the power of beauty will sooner transform honesty from what it is to a bawd thanthe force of honesty can translate beauty into his likeness. 407 408
 10. Could beauty, my lord, have better commerce than with honesty? 409
 11. Rest, rest, perturbed spirit! 410
 12. Their residence,both in reputation and profit, was better both ways. 411

Similarity scores: 412

2	0.85											
3	0.04	0.04										
4	0.11	0.09	0.59									
5	0.05	0.09	0.36	0.34								
6	0.12	0.13	0.52	0.47	0.54							
7	-0.04	-0.01	0.39	0.33	0.40	0.32						
8	-0.03	-0.04	0.53	0.53	0.53	0.55	0.39					
9	-0.05	-0.07	0.26	0.19	0.30	0.31	0.22	0.25				
10	-0.06	-0.09	0.26	0.14	0.19	0.28	0.21	0.18	0.72			
11	0.10	0.09	0.23	0.18	0.42	0.36	0.19	0.27	0.20	0.14		
12	0.04	-0.03	0.16	0.01	-0.02	0.09	0.10	0.05	0.07	0.24	-0.03	
	1	2	3	4	5	6	7	8	9	10	11	

6.2 Similar and Dissimilar Sentences from Hamlet – Examples from the First Scene, the King’s Speech and the Gravediggers’s Dialogue 414 415

- Sentences: 416
1. He shall with speed to England\For the demand of our neglected tribute. 417

2. It was that very day that young Hamlet was born — he that is mad, and sent into England. 418
 3. Th' ambassadors from Norway, my good lord,\Are joyfully returned. 419
 4. Therefore our sometime sister, now our queen,\Th' imperial jointress to this warlike state,\Have 420
 - we (as 'twere with a defeated joy,\With an auspicious and a dropping eye,\With mirth in funeral 421
 - and with dirge in marriage,\In equal scale weighing delight and dole)\Taken to wife. 422
 5. I think it be no other but e'en so. 423
 6. Is not this something more than fantasy? 424
 7. It harrows me with fear and wonder. 425
 8. I like thy wit well, in good faith. 426
 9. Cudgel thy brains no more about it, for your dull ass will not mend his pace with beating. 427
- Similarity scores: 428

2	0.34							
3	0.27	0.22						
4	0.35	0.28	0.31					
5	0.10	0.12	0.15	0.19				
6	0.05	0.12	0.03	0.19	0.16			
7	0.19	0.23	0.09	0.29	0.19	0.17		
8	0.06	0.17	0.23	0.21	0.14	0.09	0.18	
9	0.26	0.23	0.08	0.20	0.10	0.10	0.23	0.20
	1	2	3	4	5	6	7	8

430

7. Data Availability

431

Data can be found here: <https://github.com/dracor-org/shakedracor>

432

8. Software Availability

433

Software can be found here: <https://anonymous.4open.science/r/innovation-drama/>

435

9. Acknowledgements

436

Botond Szemes was supported by the ÚNKP-23-4 New National Excellence Program 437
of the Ministry for Culture and nnovation (Hungary) from the source of the National 438
Research, Development and Innovation Fund. 439

The authors are grateful for the help of Zsombor Komán in application of LLMs. 440

10. Author Contributions

441

Botond Szemes: Conceptualization, Methodology, Visualization, Writing - original 442
draft 443



Mihály Nagy: Preprocessing, Methodology - LLM, Writing – editing 444



References

- Algee-Hewitt, Mark (2017). "Distributed Character: Quantitative Models of the English Stage, 1550–1900". In: *New Literary History* 4.48, 751–782. <https://doi.org/10.1353/nlh.2017.0038>.
- Andresen, Melanie, Benjamin Krautter, Janis Pagel, and Nils Reiter (2022). "Who Knows What in German Drama? A Composite Annotation Scheme for Knowledge Transfer. Annotation, Evaluation, and Analysis". In: *Journal of Computational Literary Studies* 1. <https://doi.org/10.48694/jcls.107>.
- (2024). "Knowledge Distribution in German Drama". In: *Journal of Open Humanities Data* 1.10, 1–7. doi:10.5334/johd.167.
- Barron, Alexander T. J., Jenny Huang, Rebecca L. Spang, and Simon DeDeo (2018). "Individuals, institutions, and innovation in the debates of the French Revolution". In: *PNAS* 18.115, 4607–4612. <https://doi.org/10.1073/pnas.171772911>.
- Chang, Kent K. and Simon DeDeo (2020). "Individuals, institutions, and innovation in the debates of the French Revolution". In: *Journal of Cultural Analytics* 2.5, 4607–4612. <https://doi.org/10.22148/001c.17585..>
- Craig, Hugh and Arthur F. Kinney (2009). *Shakespeare, Computers and the Mystery of Authorship*. New York: Cambridge University Press.
- Dubourg, Edgar, Andrej Mogoutov, and Nicolas Baumard (2023). "Is Cinema Becoming Less and Less Innovative With Time? Using neural network text embedding model to measure cultural innovation". In: *Proceedings of the Computational Humanities Research Conference 2023 Paris, France, December 6-8, 2023*. Ed. by Artjoms Šeļa, Fotis Jannidis, and Iza Romanowska. CEUR-WS. <https://ceur-ws.org/Vol-3558/paper7806.pdf>.
- Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, and Peer Trilcke (2019). "Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama". In: *Proceedings of DH2019: "Complexities"*. Utrecht University. <https://doi.org/10.5281/zenodo.4284002>.
- Hope, Jonathan and Michael Witmore (2014). "Quantification and the language of later Shakespeare". In: *Actes des congrès de la Société française Shakespeare* 31, 123–149. <https://doi.org/10.4000/shakespeare.2830..>
- Krautter, Benjamin (2023). "Kopräsenz-, Koreferenz- und Wissens-Netzwerke. Kan-tenkriterien in dramatischen Figurennetzwerken am Beispiel von Kleists Die Familie Schrockenstein (1803)". In: *Journal of Literary Theory* 2.17, 261–289. 10.1515/jlt-2023-2012.
- Luan, Yingyue and Yeun Joon Kim (2022). "An integrative model of new product evaluation: A systematic investigation of perceived novelty and product evaluation in the movie industry". In: *PloS One* 3.17. 10.1371/journal.pone.0265193.
- Melanie, Andresen and Nils Reiter, eds. (2024). *Computational Drama Analysis*. Berlin: De Gruyter.
- Moretti, Franco (2011). "Network Theory, Plot Analysis". In: *Stanford Literary Lab Pamphlets* 2. <https://litlab.stanford.edu/LiteraryLabPamphlet2.pdf>.
- Pfister, Manfred (1988). *The Theory and Analysis of Drama*. Trans. by John Halliday. Cambridge: Cambridge University Press.
- Piper, Andrew, Hao Xu, and Eric D. Kolaczyk (2023). "Modeling Narrative Revelation". In: *Proceedings of the Computational Humanities Research Conference 2023 Paris, France*,

- December 6-8, 2023. Ed. by Artjoms Šeļa, Fotis Jannidis, and Iza Romanowska. CEUR-WS. <https://ceur-ws.org/Vol-3558/paper6166.pdf>.
- Reimers, Nils and Iryna Gurevych (Nov. 2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT- Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Ed. by Sebastian Padó and Ruihong Huang. Hong Kong, China: Association for Computational Linguistics. <https://aclanthology.org/D19-1410.pdf>.
- Schöch, Christoph (2017). "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama". In: *Digital Humanities Quarterly* 2.11, 4607–4612. <https://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html>.
- Šeļa, Artjoms, Fotis Jannidis, and Iza Romanowska, eds. (2023). *Proceedings of the Computational Humanities Research Conference 2023 Paris, France, December 6-8, 2023*. CEUR-WS.
- Šeļa, Artjoms, Ben Nagy, Joanna Byszuk, Laura Hernández-Lorenzo, Botond Szemes, and Maciej Eder (2024). "From Stage to Page: Stylistic Variation in Fictional Speech". In: *Computational Drama Analysis*. Ed. by Andresen Melanie and Nils Reiter. Berlin: De Gruyter.
- Sims, Matthew and David Bamman (Nov. 2020). "Measuring Information Propagation in Literary Social Networks". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, 642–652. <https://aclanthology.org/2020.emnlp-main.47>.
- South, Tobin, Bridget Smart, Matthew Roughan, and Lewis Mitchell (2022). "Information flow estimation: A study of news on Twitter". In: *Online Social Networks and Media* 31, 100231. [10.1016/j.osnem.2022.100231](https://doi.org/10.1016/j.osnem.2022.100231).
- Szemes, Botond and Bence Vida (2024). "Tragic and Comical Networks- Clustering Dramatic Genres According to Structural Properties". In: *Computational Drama Analysis*. Ed. by Andresen Melanie and Nils Reiter. Berlin: De Gruyter.
- Trilcke, Peer (2013). "Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft". In: Ajouri, Philip, Katja Mellmann, and Christoph Rauen. *Empirie in der Literaturwissenschaft*. Leiden, The Netherlands: Brill | mentis, 201–247.
- Trilcke, Peer, Frank Fischer, and Dario Kampkaspar (2015). "Digital Network Analysis of Dramatic Texts". In: *Digital Humanities 2015: Global Digital Humanities. Book of Abstracts*. Ed. by Anne Baillet, Toma Tasovac, Walter Scholger, and Georg Vogeler. University of Western Sydney.

The Anxiety of Prestige in Stephen King's Stylistics

Erik Ketzan¹ 
Martin Paul Eve² 

1. Department of Digital Humanities, King's College London , London, United Kingdom.
2. School of Creative Arts Culture and Communication, Birkbeck University of London , London, United Kingdom.

Citation

Erik Ketzan and Martin Paul Eve (2024). "The Anxiety of Prestige in Stephen King's Stylistics". In: *CCLS2024 Conference Preprints* 3 (1). 10.26083/tuprints-000273 92

Date published 2024-05-28

Date accepted 2024-04-03

Date received 2024-01-17

Keywords

Stephen King, prestige, computational literary studies

License

CC BY 4.0 

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 3rd Annual Conference of Computational Literary Studies at Vienna, Austria, in June 2024.

Abstract. This paper introduces a term, *the anxiety of prestige*, to examine thematic or stylistic textual commentaries by generally considered "popular" fiction authors on issues of literary prestige, with Stephen King as a case study. While, thematically, an anxiety of prestige has been obvious in many of King's works for decades, we suggest a novel approach: unearthing latent evidence of an anxiety of prestige in King's stylistics, through corpus query of specific stylistic features suggested by King's own writing advice book, namely adverbs, the passive voice, and "Swifties". Through close and distant reading, we interpret these stylistic features as evidence of King's textual responses to perceptions of "low" and "high" literature, and suggest that the anxiety of prestige can be investigated in larger popular fiction corpora in future work.

1. Introduction

Twentieth-century literary history can often seem enmeshed in an oscillating dialectics of "high" and "low" culture. Horkheimer and Adorno's *Culture Industry* (1947) and Pierre Bordieu's *La Distinction* (1984) are only two of many notable works in the "Great Divide", a term popularized by Andreas Huyssen as "discourse which insists on the categorical distinction between high art and mass culture" (1986, vii). Huyssen framed modernism, a paragon of high culture, as displaying an "obsessive hostility to mass culture", but as modernism ceded to (or merged with) postmodernism, the relationship between "modernism, avantgarde, and mass culture" came to be described in terms of "a new set of mutual relations and discursive configurations" (1986, vii, x). Postmodernism is generally described as embracing "popular," "mass," or "kitsch" culture through a variety of ironic strategies, especially pastiche and parody; the "postmodern paradox," as Linda Hutcheon put it, in which "to parody is both to enshrine the past and to question it" (1988, 126). While every aspect of postmodernism, including "its very existence," has "been a matter of fierce controversy," per Brian McHale, the "term and concept 'postmodernism' began to lose traction around the beginning of the new millennium", and by 2015, "postmodernism, it is generally agreed, [was] now 'over'" (2015, 5) as both an active aesthetic movement and a useful discriminative term. Meanwhile, sociologists have devoted extensive study to a new phenomenon which has emerged since at least the 1980's: highbrow "snobbery" being replaced by omnivorousness cultural consumption by elites (Richard A Peterson and Simkus 1992, Richard A. Peterson and Kern 1996,

Ollivier 2008). As de Vries and Reeves (2022) summarize, “The distinction between ‘elite’ and ‘mass’ consumers once dominated theories of cultural consumption [...]. However, over the last quarter century the ‘elite-mass’ hypothesis has fallen out of favour in the sociological literature, largely supplanted by Richard Peterson’s ‘omnivore’ hypothesis”.

Distinctions between “high” and “low” are crumbling not only among readers, but academics, as well. It is now recognized that notions of canonicity and what is considered “literary fiction,” by whom, and when, are highly complex dynamics of social and economic (Bourdieu 1979), gender (Light 2013, 6) and racial (So 2021) concerns. Richard Jean So writes that, “Today, scholars are more interested in studying the porousness and interchangeability of these categories [of high and low], rather than their imagined difference or hierarchy,” and that “The categories of ‘high’ and ‘low’ are still important to cultural scholars; it’s just that the imagined space between them has contracted or at least become altered, shaping the way works of literature are judged and received” (2021, 105).

But a major gap exists in many of our narratives about both the Great Divide — discourse based on a categorical distinction of “high” and “low” literature — and the new omnivorousness in cultural consumption which followed: how did popular fiction authors and texts respond to these discourses? While literary modernism and postmodernism basked in prestige throughout most of the twentieth century, how did the so-called mass, popular, or kitsch authors of thrillers, science fiction, romances, horror, comic books, and pulp fiction — unfairly implied as an undistinguished mass by Horkheimer and Adorno’s term, Culture Industry — respond to the dismissal, exclusion, and derision by literary fiction and its attendant gatekeepers of critical acclaim and the canon? Despite the rise of popular culture and popular fiction studies, this story remains largely fragmentary. Ken Gelder writes that “Literary fiction is ambivalent at best about its industrial connections and likes to see itself as something more than ‘just entertainment’, but popular fiction generally speaking has no such reservations” (2004, 1). We suspect that this is far from the whole story, however; that many popular fictions have responded to issues of The Great Divide and now culture omnivorousness in a variety of textual ways.

We suggest a new term to explore such commentaries in popular fiction: *the anxiety of prestige*. We propose the definition: thematic or stylistic textual, paratextual, and metatextual commentaries by generally considered “popular” fiction authors on issues of literary prestige, which can include critical or parodic portrayals of literary prestige and its gatekeepers, or explicit or implicit attempts by the popular fiction author to attain or achieve higher literary prestige for themselves, either by adopting stylistic features of “high” fiction, or asserting the value of “popular” fiction. This definition, while broad, provides us with a starting point to examine a wide variety of textual responses by generally-considered popular authors to issues of literary prestige, often through ambivalent or sometimes even contradictory means: retorts and responses by popular fiction to The Great Divide or the new cultural omnivorousness, which we suggest remains a largely untold story in literary history.

We suggest that digital humanities can help illuminate the anxiety of prestige, especially through its ability to distant read large corpora; as the term “mass” fiction suggests,

the corpus of popular fiction is certainly massive. Digital humanities can locate textual evidence more easily, through query of, for instance, thematic portrayal of literary prestige's gatekeepers, such as literature professors, literary critics, literary awards, and so on. But corpus query can also unearth less obvious textual evidence of the anxiety or prestige through query and modelling of style and change of style, for instance corpus stylistics (Wynne 2006), which can unearth patterns in latent, formal, quantifiable stylistic features. This inquiry can be aided by, and aspire to add to, a growing body of digital humanities studies on the relations between formal textual features and perceptions of literary quality (Verboord 2003, Hakemulder 2004, Van Peer 2008, Archer and Jockers 2016, Knoop et al. 2016, Piper and Portelance 2016, Underwood and Sellers 2016, Van Cranenburgh et al. 2019, Cranenburgh and Koolen 2019, Underwood 2019, Van Cranenburgh and Ketzan 2021, Van Dalen-Oskam 2023), as well as canon (Algee-Hewitt and McGurl 2015, Porter 2018), genre classification (Rybicki and Eder 2011, Schöch 2017, Underwood 2019), and linguistic criticism of the writing advice genre (e.g. Pullum 2004 and Pullum 2015). We note that while recent work on literary quality is employing sophisticated computational methods that quantify dozens or hundreds of textual features at once (often features which are undefined to the scholar within a "black box" of machine learning), we apply less sophisticated corpus query methods that have the benefit of allowing close reading of definable textual features.

Our term, anxiety of prestige, is coined with a nod to Harold Bloom's *anxiety of influence* (1997), and our choice of term is somewhat tongue-in-cheek, as Bloom himself was a vociferous critic of popular fiction, as well as of popular American author Stephen King (1947-), the subject of this paper. We suggest King as a major figure in inquiries into the anxiety of prestige, as King began his best-selling career (over 350 million copies sold, per Heller 2016) derided and dismissed by high literary critics, but is now firmly established as a critically-acclaimed American author. King exemplifies, and perhaps contributed to, the current cultural omnivorosity. The writer once so dismissed by high literary critics such as Bloom has been contributing to *The New Yorker*, a leading arbiter of literary prestige, since 1994, and King won the National Book Award Medal for Distinguished Contribution to American Letters in 2003.

2. Stephen King's Anxiety of Prestige

King's fiction contains a prodigious amounts of commentary on literary prestige, some of which is too salient to miss, but much of which has so far not been the subject of sustained attention from scholars. Perhaps the most obvious example is *Misery*, in which the writer Paul Sheldon, who "wrote novels of two kinds, good ones and best-sellers", has finished his best-selling "series of romances about sexy, bubbleheaded, unsinkable Misery Chastain" and jubilantly resumed his ambitions to write serious literary fiction, despite his audience's protests: "He could write another [...] *The Sound and the Fury*; it wouldn't matter. They would still want Misery, Misery, Misery." (1987a, 36). Sheldon revels in the completion of his new, ambitiously literary novel, but Sheldon's aspirations of literary prestige are thwarted when he is kidnapped by superfan Annie Wilkes, who literally chains Sheldon to a typewriter and, under threat of death, forces him to write a new genre novel about her beloved character Misery. Many more examples from King's long oeuvre could be named, especially as King made a rather conscious turn to attempt

more “literary fiction” in the early 1990’s, most notably with *Dolores Claiborne* (1992a). 111
 And questions of literary prestige are abundant in King’s fiction to this day. In *Rat* (in *If It Bleeds*, 2020), college English professor Drew Larson, a failed high literary novelist 112
 known to “steer clear of popular fiction,” is suddenly seized by the inspiration to write a 113
 commercial pulp Western novel. In *Fairy Tale*, King lightly parodies academia by having 114
 his teenage narrator reveal that he went on to become an academic: “I am considered 115
 quite the bright spark, mostly because of [...] an essay I wrote as a grad student. It was 116
 published in *The International Journal of Jungian Studies*. The pay was bupkes, but the 117
 critical cred? Priceless” (2022, 591). 118
 119

The issue of King’s literary prestige, or lack of it, also abounds in King reception. Earlier 120
 critics opined on whether King is or is not “literature,” whether he is a “mere” horror 121
 or “genre” writer or somehow more “literary” than this label might suggest. The most 122
 hyperbolic of such statements came from Harold Bloom, who introduced his edited 123
 volume of scholarly essays on King with the sentiment that “King has replaced reading” 124
 and that “King’s books [...] are not literary at all, in my critical judgment” (2007, 2). 125
 Further, a 2012 scholarly monograph on King’s magnum opus is titled *Respecting The* 126
Stand (Paquette 2014, as though 190 pages of literary criticism were required to show 127
 why the novel should be respected. The same volume’s publisher description opens with 128
 the assertion that “[a]cademics dismiss Stephen King as a genre writer who appeals 129
 to the masses but lacks literary merit”. Scholars often cannot approach any topic in 130
 King studies without some discussion of King’s literary quality, which likewise read 131
 as disclaimers or justifications for the scholarly study itself. James Arthur Anderson, 132
 for instance, writes that “[i]t is my hope that my application of these theories will [...] 133
 show that [King] is more than just a horror writer, more than just the creator of ‘popular 134
 fiction’” (2017, 8). This attention to King’s literariness or prestige – or otherwise – can 135
 also stand in the way of other close readings. For instance, King’s early novel, *The* 136
Long Walk (1979), holds up well as an allegory of the Vietnam War, a fact that can be 137
 obscured when appraisals of literary value displace textual attention (see Texter 2007, 138
 47). King’s retorts to these decades of criticism may be read in his paratextual interviews 139
 and prefaces, for instance telling a *Guardian* journalist that “I have outlived most of my 140
 most virulent critics. It gives me great pleasure to say that” (Xan 2019). 141

More clues to King’s anxiety of prestige may be read in *On Writing: A Memoir of the Craft* 142
 (2000), which combines reminiscences of King’s career as a writer with prescriptive 143
 writing advice for would-be authors. According to King, adverbs, passive verbs, and 144
 adverbially modified dialogue attribution should be avoided, for instance. King is hardly 145
 alone in offering such writing advice to aspiring authors, which is arguably a tradition 146
 as old as writing itself; Plato himself discouraged the reader from writing at all (Plato 147
 2005, 63)! And writing advice books today could even be considered its own genre 148
 (Steve Evans 2005). The writing advice in William Strunk Jr. and E. B. White’s *Strunk* 149
 and *White* 1999, a prescriptive style and grammar guide, has sold over 10 million copies 150
 and achieved, per Geoffrey Pullum, “a vice-like grip on educated Americans’ views 151
 about grammar and usage” (2010, 34). The path that King treads in issuing such advice 152
 has been well travelled by other authors and his advice is typical of the genre. 153

3. Research Aims and Methods

154

A traditional scholar could easily fill a monograph by close-reading the anxiety of prestige in King's voluminous fiction (over 60 novels and over 200 short stories, as of 2024), paratexts such as author interviews and King's commentaries on style in *On Writing*. But in this paper, we suggest less obvious avenues for unearthing evidence of King's anxiety of prestige, which, while King-specific in method, could inspire future work in larger popular fiction corpora.

We explore how the anxiety of prestige may be interpreted by comparing King's writing advice with his own published fiction. These provide small contributions to, specifically, King studies; how did King's stylistics change over a 50+ year career, and did King actually follow his own advice? But we also hope that our corpus stylistic experiments, applying a mixed-methods approach of close and quantitative or distant reading (Hermann 2017), may provide models for the study of the anxiety of prestige in popular fiction more broadly.

We first examine the frequencies of word patterns based on King's advice for writers to avoid: first adverbs, then "Swifties" (adverbially modified dialogue attribution), then the passive voice, all queried in King's own fiction and comparison corpora. The methods are simple corpus query via regular expressions using two widely-used corpus query platforms that pre-process texts by adding part of speech and lemma tags: LancsBox 6.0 (2020) and TXM 0.8.1 (2010). Both have implemented part of speech tagging using TreeTagger (Schmid 1999), while LancsBox was used in the third experiment because it contains a built-in regular expression for passive constructions (as discussed in more detail in Experiment 3, below). Manual inspection and cleanup of all query results was performed, and visualizations of frequencies were created in Google Sheets.

We note here in the methods section that our query of words and linguistic patterns which King attributes to "good" and "bad" writing cannot necessarily be naively equated with "high" and "low" literary style, but we attempt to interpret these connections. King has been consistently vocal in his advocacy of popular fiction, even if many of his fictions clearly aim for, or achieve, high literary merit; King made a conscious attempt at more literary fiction in the early 90s, especially with *Dolores Claiborne* (1992), but such efforts to write more "literary" novels has never been consistent in King's career, and more straightforwardly entertaining fictions by King have sometimes followed more literary ones, and vice versa. One could certainly interpret King's specific elements of writing advice as genre- or prestige-neutral; advice for writers to simply write better, regardless of literary aim. But we argue below that King's writing advice can sometimes be read as exhortations to write in an implicitly more "high" literary way, or that King's own implementation of his own writing advice can be interpreted as evidence of King's own high literary aspirations. Tracing King's writing advice against his own works, then, can provide evidence for interpretations of the anxiety of prestige in King's texts. If the reader is critical of our comparison of King's notions of "good" and "bad" writing with "high" and "low" literary writing, we agree that the connection is interpretive and far from unambiguous, and return to this question a number of times below.

4. Corpora

196

We assembled all 73 novels and novellas solely authored by Stephen King up to 2020. 197
 We also separated out “Misery’s Return,” a 9,000 word story-within-a-story pastiche 198
 of intentionally “bad” genre writing from King’s *Misery*, which we treat as a distinct 199
 comparator text. Exploring questions about King’s distinctiveness meant that we also 200
 needed comparison corpora. For these we selected The Brown Corpus of Standard 201
 American English as a snapshot of US English from 1961 (Francis. and Kučera 1979) 202
 and The Freiburg-Brown corpus of American English (FROWN) as a snapshot of 1992 203
 (Mair 1992). We also assembled a Stephen King Fanfiction corpus containing the first 204
 5,000 tokens from all King-inspired stories on Fanfiction.net exceeding 5,000 words 205
 (91 stories in total; 455,000 word tokens); the 5,000 word cut off is arbitrary, and is 206
 intended to separate fanfictions which evidence a serious attempt at fiction from the 207
 short, sometimes free-form fanfictions on the website. While comparing an author to 208
 his/her amateur literary imitators is a useful foil, a second fanfiction comparison corpus 209
 was also desirable for reference (Sigelman and Jacoby 1996). We thus also compiled a 210
 corpus of *Harry Potter* Fanfiction (91 texts, first 5,000 word tokens each), chosen simply 211
 as a well-known popular fiction which has inspired many fanfictions. As a final baseline 212
 comparison, we assembled a corpus of National Book Award-winning novels from 213
 1974–2020 as our high literary fiction corpus (Appendix I). We attempted to control for 214
 diachronic change in English by selecting only American authors of roughly the same 215
 age (within 10 years) as King, nineteen novels total. 216

5. Experiments

217

5.1 Experiment 1: “The Road to Hell is Paved with Adverbs”

218

King emphatically warns his readers to avoid adverbs, which he sees as a sign of timid 219
 writing: “[t]he adverb is not your friend” and “the road to hell is paved with adverbs” 220
 (2000, 138-39). Such prescriptions against adverbs are common in the writing advice 221
 genre, which has drawn the ire of Pullum (2015). Assertions to “avoid adverbs” are 222
 also problematic, as So has shown that one of the core stylistic characteristics shared by 223
 bestselling and prize-winning fiction is a “syntactical preference” for adverbs, when 224
 compared to a corpus of black writing that was excluded from these canons (2021, 129). 225
 Given that King’s work is bestselling, then, we would expect his adverbial prevalence to 226
 be similar to other bestselling and prizewinning works. 227

It turns out that, despite King’s pronouncements, this is indeed the case. Ben Blatt 228
 has already made a first contribution to this question; noting King’s advice about 229
 adverbs, Blatt queried adverbs in a large corpus of contemporary fiction, including a 230
 King corpus of 51 novels, reporting that King scores average in a selection of authors 231
 from Hemingway to E. L. James (2017). We expand this inquiry with a larger King 232
 corpus and present data per King novel, to trace diachronic adverb frequency, and trace 233
 more of the stylistic devices discussed in *On Writing*. As shown in Figures 1 and 2, there 234
 is statistically significant, but not major variation between the reference corpora, King’s 235

Adverbs

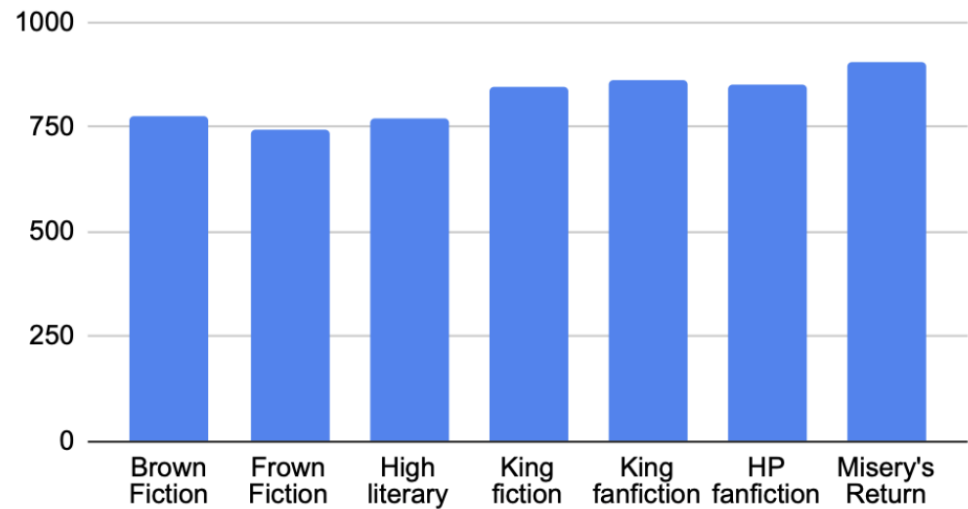


Figure 1: Relative frequency of adverbs (per 10,000 word tokens).

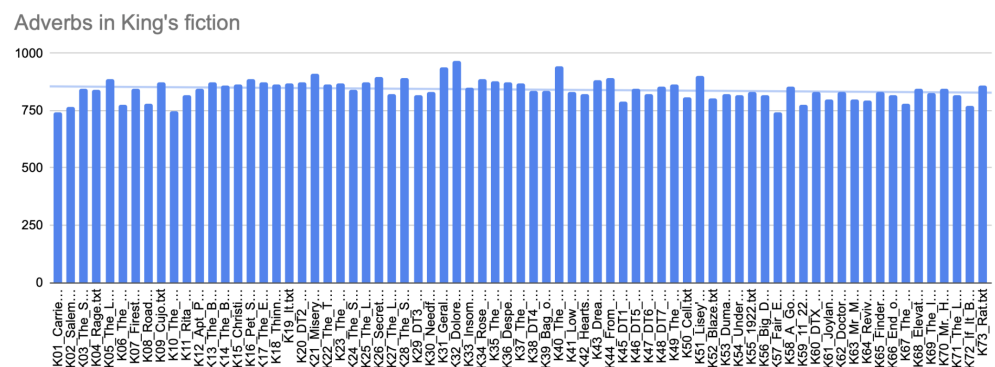


Figure 2: Relative frequency of adverbs in King's texts chronologically (per 10,000 word tokens).

texts, high literary, and, surprisingly, fanfiction,¹ and little variation in adverb usage throughout King's career. Perhaps ironically, King's lowest frequency of adverbs is in his first published novel, *Carrie* (1974), while the highest use of adverbs is King 1999, published just one year before *On Writing*. This seems inconsistent with King's opinion that "the road to hell is paved with adverbs".

However, these initial results are misleading. As noted by Blatt, when King proscribes adverbs, King actually means adverbs ending in <-ly>, e.g. *totally*, *completely*, and *modestly*. This then excludes temporal adverbs and various locative forms. The number of adverbs that are excluded in such filtering vary by author, but Blatt proposes that approximately 10% to 30% of all adverbs are of the <-ly> type (2017, 12-12). In Figures 3 and 4 we show the same query confined to <-ly> adverbs.

The data for Figure 3 confirm one of Blatt's findings: that <-ly> adverbs are significantly

1. King's fiction compared with Brown: 128.16 LL, $p < 0.0001$. King's fiction compared with Frown: 7.44 LL $p < 0.01$. King's fiction compared with high literary: 1210.58 LL, $p < 0.0001$. Calculated using Rayson's Log Likelihood calculator.

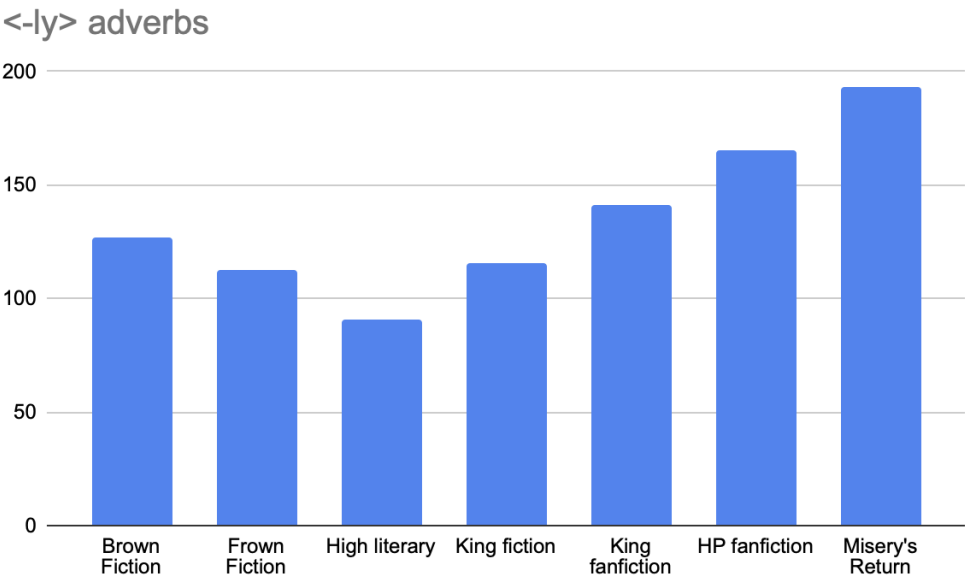


Figure 3: Relative frequency of <-ly> adverbs (per 10,000 word tokens).

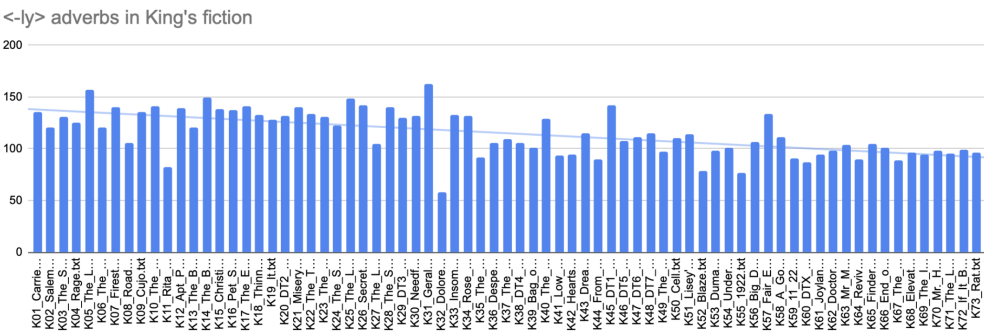


Figure 4: Relative frequency of <-ly> adverbs in King's texts chronologically (per 10,000 word tokens).

more frequent in fanfiction (2017, 27), suggesting that King's and others' distaste for <-ly> adverbs can be distinctions of "good" vs. "amateur" (or "bad") writing. Consistent with this, <-ly> adverbs are lowest in our "high literary" corpus. Although van Cranenburgh and others cast doubt on the correlation of single stylistic features with literariness measures, this is some evidence that <-ly> adverbs may be a textual marker of low literariness.

Figure 4 also yields new insights into diachronic changes in King's style: <-ly> adverbs significantly decline over the course of his career, consistent with his advice. It is possible that the changes exhibited over King's style reflect a broader shift in American fiction or the generic movements with which King is associated. Jack Elliott (2015), for instance, has documented declining adverb usage within a corpus of romance novels over time. However, rather than moving outwards to entire genre study, these results instead also allow us to delve more closely into King's own anxiety of prestige, specifically in his intentional parody of bad writing: "Misery's Return."

In King's *Misery*, the violent kidnapper character Annie Wilkes forces author Paul Sheldon to write a new genre story starring her beloved character, Misery, and Sheldon produces "Misery's Return," selections of which are spread throughout *Misery*. Even a cursory first reading of these sections shows a marked increase of egregiously florid or unnecessary <-ly> adverbs: a "stuporously warm West Country kitchen", "[s]he stood lightly poised," and "[h]e honked mightily into [the handkerchief]" (132, 161, emphasis ours). Thus, when King parodies bad writing, he augments a great many verbs with an adverbial modifier. King parodying genre writing in this way expresses an anxiety of prestige, with King implicitly placing Sheldon's true potential as a writer, and King's own, as above badly written mass fiction.

Hypothesizing why some texts are outliers in adverbial usage should be approached with caution. But it is notable that King 1992a, King's nineteenth novel, is the text with the lowest number of <-ly> adverbs. This novel was a serious stylistic departure for King and a significant attempt at more literary writing, as discussed below. *Dolores Claiborne*, the bestselling US novel of 1992, deploys a great deal of phonetic dialect and is written from a single narrative perspective, an unusual feature for King (Smythe 2015). We suggest that here, again, is a marker of King's anxiety of prestige. Having associated the <-ly> adverb with low, King's eschews it most in one of his most intentionally literary works.

5.2 Experiment 2: "Swifties," he dismissed quickly

Related to <-ly> adverbs, King urges would-be writers to avoid the "Tom Swiftie": dialogue attribution with an excessive, absurd, or "purple" (meaning excessive or extravagant) adverb, which eventually took the form of a pun or parody of bad writing. An example of a true, punning Tom Swiftie might be: "'Pass me the fish,' Tom whispered, crabably". King broadens the purview, though, to include all adverbially modified dialogue attribution: "I can be a good sport about adverbs, though. Yes I can. With one exception: dialogue attribution. I insist that you use the adverb in dialogue attribution only in the rarest and most special of occasions" (2012, 140). King illustrates this with:

"Put it down!" she *shouted menacingly*.

Direct discourse attribution with <-ly> adverb, e.g. "said quietly"

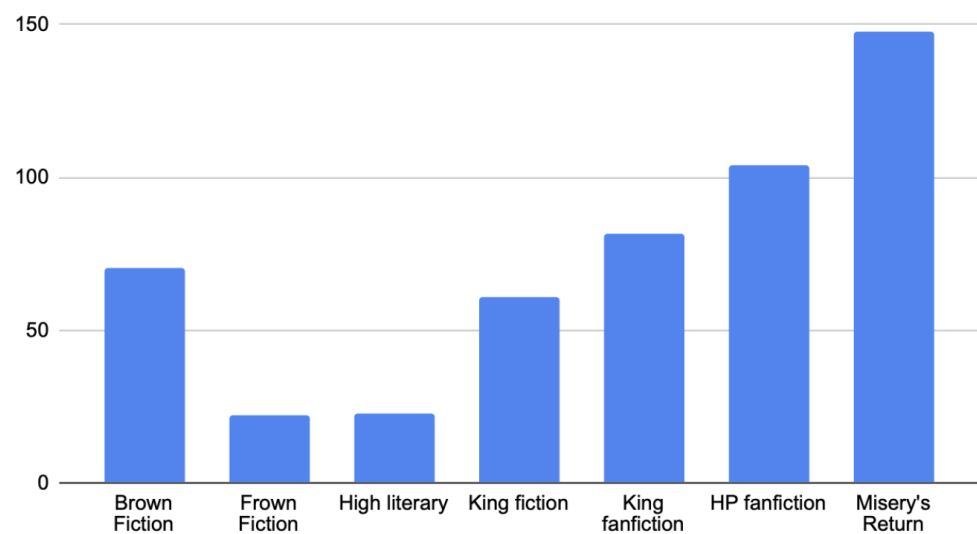


Figure 5: Relative frequency (per 100,000 word tokens) of the Swiftie construction.

"Give it back," he *pleaded abjectly*, "it's mine." 291

"Don't be such a fool, Jekyll," Utterson *said contemptuously*. (2000, 140-41, 292
emphasis added) 293

Query reveals that King has avoided these specific phrases almost entirely in his own 294
writing.² Having decried such adverbial modification under most circumstances, King 295
nonetheless admits that he still occasionally uses the form: 296

And here's one I didn't cut . . . not just an adverb but a Swiftie: "Well," 297
Mike *said heartily* . . . But I stand behind my choice not to cut in this case, 298
would argue that it's the exception which proves the rule. "Heartily" has 299
been allowed to stand because I want the reader to understand that Mike is 300
making fun of poor Mr. Olin. Just a little, but yes, he's making fun. (2000, 301
344, emphasis in original) 302

As a next step, we wished to query Swifties in King's texts, which could be opera- 303
tionalized in a number of ways. Lessard 1992 designed a Swiftie-generating computer 304
program. *litovkina_swifties* writes that more recent examples of Swifties do not strictly 305
require an adverb. While canonical Swifties contain an element of humor, we simply 306
query the basic adverbial construction that King decries. All of King's examples follow 307
a precise word order: Direct Speech → Noun/Pronoun of the speaker → Attribution 308
Verb → <-ly> adverb. The frequency of this form is shown in Figure 5. 309

These results are consistent with King's perception of the Swiftie — adverbially modified 310
direct discourse attribution — as a marker of bad writing: King's fiction and Brown 311
score similarly, the high literary texts use the construction far less frequently, while 312
fan fiction displays a high prevalence. As with adverbs, "Misery's Return" scores the 313
highest. Certainly, in King's case, the use or avoidance of the Swiftie construction can 314

2. The phrase "said contemptuously" appears in King's second novel, King 1975, as well as the 2010 novella *Big Driver*.

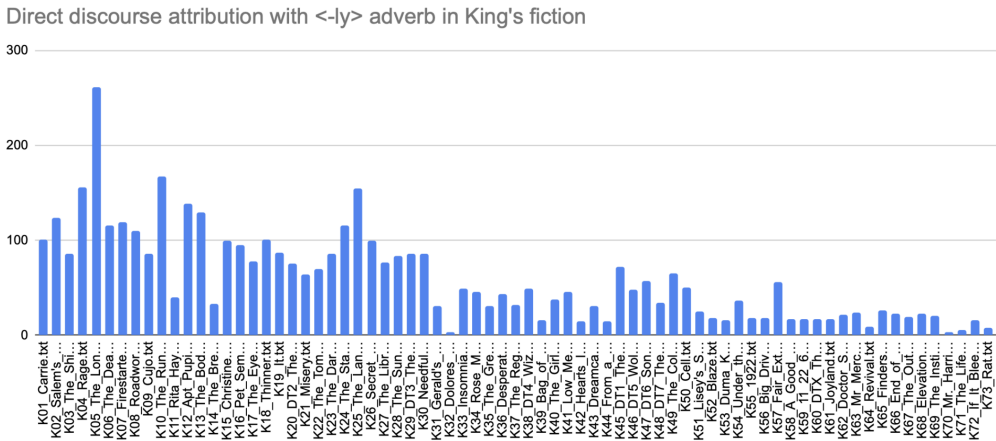


Figure 6: Relative frequency (per 10,000 word tokens), of the Swiftie construction in King's texts.

be considered a marker of the anxiety of prestige. 315

A closer inspection of this Swiftie construction in the comparison corpora underscores 316
its association with prestigious, high literature. A number of the National Book Award 317
winners eschew the construction entirely, perhaps an indication that these writers 318
have absorbed the collective (if questionable) stylistic wisdom of the writing guide 319
genre. While examples from fanfiction would raise the ire of many a writing teacher — 320
“Vernon boomed happily,” “Carlos yammered ecstatically” — the majority of Swiftie 321
constructions are mostly, by themselves, aesthetically inoffensive and found in many 322
professional comparison texts; it is rather the high frequency of them in fanfiction that 323
correlates with low prestige. 324

Within King's oeuvre, this Swiftie construction clearly decreases over the course of his 325
career (Figure 6). King's earlier, journeyman works employed this Swiftie construction 326
far more frequently, but this decreased over time as he developed the stylistic aesthetics 327
eventually expressed in *On Writing*. Interestingly, the highest result, *The Long Walk*, 328
was King's fifth published novel but first written novel, begun in 1966–67 during his 329
freshman year at the University of Maine (King 2000, 428–32), bolstering the impression 330
that King as a younger man dabbled in the Swiftie, but quickly decreased its usage. 331
The next highest result, *The Running Man* (1982), was also written before King's first 332
published novel, *Carrie*. The Swifties in these early works are, for the most part, not 333
purple prose — e.g. “said casually”, “said cheerfully”, “thought bitterly” — it is again the 334
frequency which is notable. Some of the Swifties do, however, read as what many would 335
consider bad prose. Twice in *The Long Walk*, direct speech is introduced by “shrewishly”: 336
“Barkovitch screamed shrewishly” and “Garraty said shrewishly”. Similarly, in *The Long* 337
Walk, King broke his own rule against the use of pretentious vocabulary, writing that 338
“McVries said sententiously”; a word that query reveals King never used again. All of 339
this suggests that King formed his disdain for this kind of Swiftie (adverbially modified 340
discourse attribution) very early in his career. 341

For the use of Swiftie constructions, Figure 6 shows that there is a distinct point of 342
division in his works. The break occurs in 1992 with the publication of *Gerald's Game* 343
(May 1992b) and the aforementioned *Dolores Claiborne* (November 1992a). These novels, 344

importantly, were attempts by King to move away from the (inaccurate) label of horror 345
 genre writer and write more prestigious, literary works. Although King had previously 346
 written works that were narrated in omniscient third-person and that followed a number 347
 of characters' thoughts in each novel via free indirect discourse (with occasional first- 348
 person narration for stories within stories, diary entries, etc.), *Gerald's Game* and *Dolores* 349
Claiborne were attempts by King to follow a single character's voice. *Gerald's Game* 350
 features a woman who is handcuffed to a bed and must escape, alone with her thoughts, 351
 narrated in the third person and eventually first person. *Dolores Claiborne* goes a step 352
 further, with the entire novel narrated in the first-person voice of the eponymous Dolores, 353
 a 65-year old widow. In this text, King phoneticizes the speech of the narrator throughout 354
 (e.g. "he ast me" for "he asked me"), uses frequent contractions (dropped 'g's in <-ing> 355
 words: "'lookin'", "'givin'"), and vernacular exclamations of "Gorry!". This "single 356
 point of view is a huge change for King," observes James Smythe, who notes "the semi- 357
 phonetic nature of the text" (Smythe 2015). These novels from 1992 also mark a turning 358
 point in King's characterization and portrayals of women. Carol Senf (1998), for instance, 359
 has praised the realist psychological portraits of female characters in these novels. Heidi 360
 Strengell further writes that "since the publication of *Carrie* (1974), King has been 361
 blamed for depicting women characters as stereotypes," but notes that, "especially since 362
Gerald's Game (1992), he has more consciously concentrated on women, the emphasis 363
 shifting from child characters to women characters" (2005, 16). Senf, in a feminist 364
 analysis of the two novels, writes that she finds herself "applauding King for the risks 365
 he has taken in *Gerald's Game* and *Dolores Claiborne*" and praises his "shift in perspective 366
 and his ability to create strong, plausible women characters" (Senf 1998, 105). 367

The low prevalence of the Swiftie construction in *Gerald's Game* and *Dolores Claiborne* 368
 and the subsequent decline in this form over the remainder of King's career can be read 369
 as an indication of King's intensified literary ambitions in these particular novels, and 370
 the anxiety of prestige. On the other hand, it could be hypothesized that *Gerald's Game* 371
 and *Dolores Claiborne* feature a lowered number of Swiftie constructions because, being 372
 single-character studies, they have only a small quantity of direct speech. If there is 373
 little quoted dialogue, it would follow that fewer Swifties would emerge. But this is not 374
 necessarily the case. We estimated the quantity of direct speech in King's fiction via a 375
 simple query: word tokens between left and right quotation marks (Figure 7).³ By this 376
 estimate, *Gerald's Game* does indeed have the lowest volume of direct speech (4.23%) 377
 of any of King's novels, which makes sense, as much of the dialogue in this novel is 378
 presented indirectly in the memories, fantasies, and hallucinations of its protagonist, 379
 who is trapped alone in a bedroom. *Dolores Claiborne*, however, while on the low end 380
 of dialogue by volume (10.86%), is slightly higher than a number of other earlier King 381
 novels — *The Eyes of the Dragon* (1984), *The Tommyknockers* (1987b) — and is only 1% 382
 lower than *Cujo* (1981). This suggests that the number of Swiftie constructions in a text 383
 by King cannot necessarily be directly correlated merely with lower quantities of direct 384
 speech. 385

This new evidence — low Swifties in novels aiming to be high and literary, and the low 386

3. The limitation of this query is that quoted word tokens may also indicate not only direct speech, but direct thought and direct writing, as well. This method also captures single words and phrases that are quoted for emphasis, rather than attribution (e.g. "the Democrat had stopped doing its yearly 'oldest resident' interview with him three years previous"; so-called "scare quotes"). For more on such direct speech query see e.g. Liberman 2017.

Estimate of direct discourse as % of novel

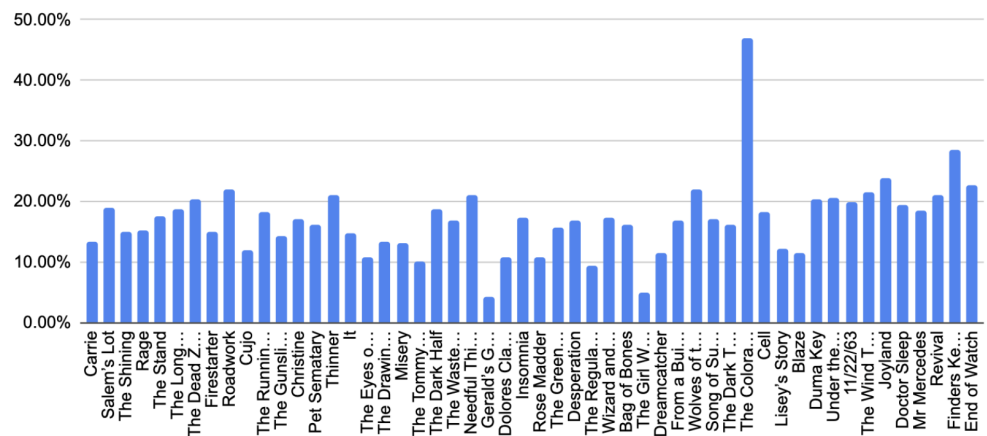


Figure 7: Estimate of direct discourse word tokens as percentage of novel, using regular expressions and quotation marks.

Swiftie query not explainable by low amount of direct speech alone — underscores the 387
close reading impression that Swifties in “Misery’s Return” appear stark and deliberate. 388
The overbaked adverbially modified speech attributions in “Misery’s Return” — e.g. “he 389
whispered strengthlessly” — also do not appear anywhere else in King’s writing. 390

The question remains, though, as to the extent that King associates such “bad” writing 391
with genre fiction, whether the two are separable, and thus, whether our queries truly 392
reveal an anxiety of prestige, or merely an anxiety of King’s notions of good and bad 393
writing, that are distinguishable from the style of high, prestigious literature. First, 394
in *On Writing*, King frames his disdain of Swifties by noting their historical origin in 395
juvenile genre fiction and dime novels (2000, 125-26). Second, it is at a point where 396
King veers away from his own generic stylings that the Swiftie construction declines, 397
giving evidence of a conjunction of high prose style with new high literary genre modes. 398
This is complicated, though, by the fact that even when King later returns on occasion 399
to generic horror writing after 1992, the Swiftie construction is nonetheless used less 400
and less often. The conclusion that we draw is that while King initially and historically 401
associates Swifties with “bad” writing within generic moods, after 1992, even when 402
returning to various genres, King aims for a higher literary prose style. 403

5.3 Experiment 3: The Passive Voice Should Be Avoided 404

In *On Writing*, King exhorts the would-be writer to avoid passive verbs, which he 405
contends are “weak”, “circuitous”, and “frequently tortuous, as well” (2000, 122). As 406
with his warning against adverbs, King hedges this advice, specifying that he “won’t say 407
there’s no place for the passive tense. Suppose, for instance, a fellow dies in the kitchen 408
but ends up somewhere else. The body was carried from the kitchen and placed on 409
the parlor sofa is a fair way to put this, although ‘was carried’ and ‘was placed’ still irk 410
the shit out of me” (Ibid.). Nonetheless, King’s opinion is clear: overuse of the passive 411
voice is characteristic of bad writing. 412

Such warnings against passive verbs are a staple of 20th-century writing advice, from 413
Edwin Woolley in 1907 via George Orwell through William Strunk (Zwicky 2006). 414

However, as Pullum notes, “there is rampant confusion about what ‘passive’ means linguistically”, as “contrary to popular belief, passives do not always contain be and do not always contain a past participle” (2014). Pullum sternly admonishes writing advice authors for their “extraordinary level of ignorance of simple facts” and laments that “the state of the general public’s education regarding the notion ‘passive voice’ is nothing short of disastrous” (2014, 64, 67). King at least provides correct examples of passive verbal phrases, unlike many of the writing advice offenders castigated by Pullum. But King, like most of his writing advice forebears, means *be verbal phrases* when stating “avoid the passive”, and his examples of bad passive phrases in *On Writing* fall into two categories: future tense (e.g. “the meeting *will be held* at seven o’clock”) and past simple (e.g. “the body *was carried* from the kitchen”). Querying and classifying the tense of passive verb forms in the Brown Fiction corpus suggests that past simple passive verbs make up the large majority of passive verbs found in fiction, and that future tense passive verbal phrases are rare (Table 1).⁴

Passive verb forms	Brown Fiction
Present Simple	63
Present Continuous	0
Present Perfect	34
Past Simple	700
Past Continuous	1
Past Perfect	154
Future	0
Future Perfect	0
Total	952

Table 1: Passive Verb Forms in Brown Fiction corpus

As a next step in investigating whether the types of passive verbal phrases that King warns against display variance in King’s fiction and are observably higher elsewhere, we queried passive *be*-verb constructions in the corpora (Figure 8) and the trend over the course of King’s writing career (Figure 9).

These results show a low variance in use of *be* passive phrases in texts as disparate as National Book Award winners and *Harry Potter* fanfiction, suggesting that despite the common advice to “avoid passives”, they remain a widespread feature of English writing, as Pullum suggests, and a poor indicator of differential literariness. Furthermore, although there is a steady and marked decline in *be* passive use over the course of King’s career, it is hardly substantial, and some of the later texts feature significantly more passives than a number of the earlier books. This is all to say that passives, in general, do not seem to serve as good indicators of high and low literary language.

6. Conclusion and Future Work

This paper has introduced a term, the anxiety of prestige, along with a proposed definition, above, to serve as a starting point in the analysis of a still largely unexamined

4. These data were derived from the 1,093 passive verb forms detected by the LancBox query `PASSIVES — or _VB. (R.*){0,3}V.N/` — sorted by simple regular expressions to detect the canonical forms of passive verbs: present simple (am/are/is + past participle); present continuous (am/are/is being + past participle); present perfect (have/has been + past participle); past simple.

Passive verbal phrases with word forms of "be"

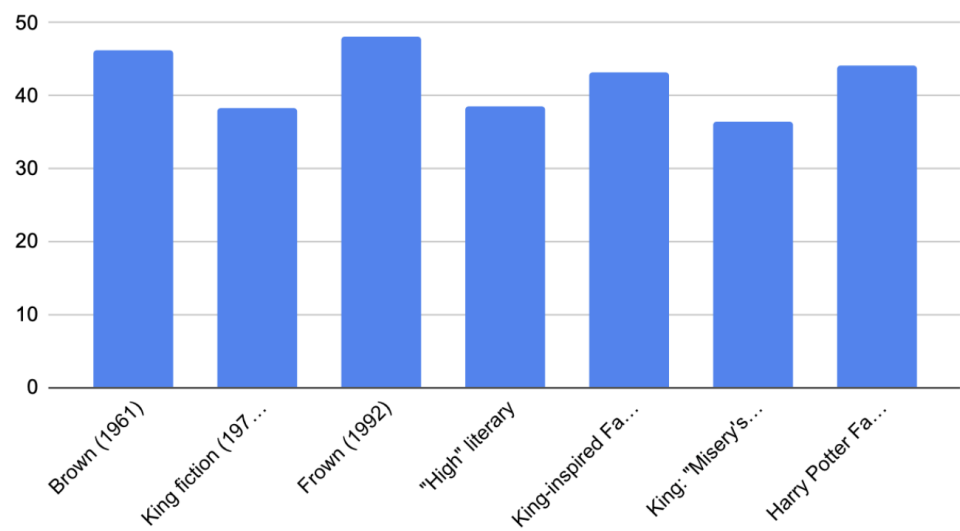


Figure 8: Passive verbal phrases (with word forms of be), per 10k tokens.

Passive verbal phrases with word forms of "be"

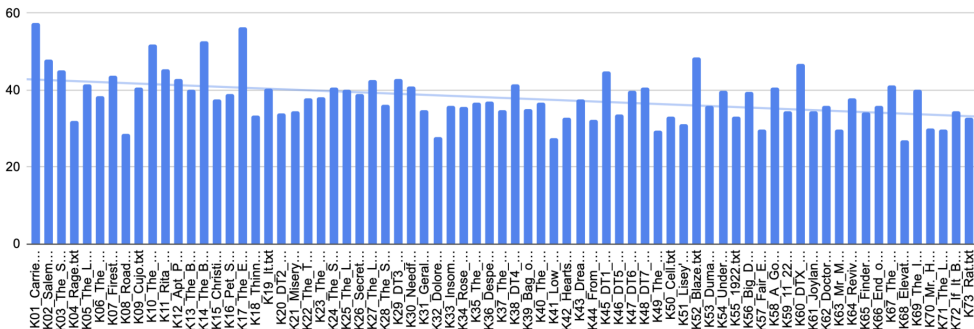


Figure 9: Passive verb forms in King corpus, per 10,000 word tokens.

phenomenon in literary history: textual responses by widely-considered “popular” fiction authors to issues of literary prestige. Our experiments provide contributions to King studies in particular, but also hope to contribute to future investigations of the anxiety of prestige in popular fiction broadly. Digital humanities may be well suited to this task, most simply in the location of textual thematic evidence in larger corpora, but also, as we have attempted to show, through corpus stylistics. Future work could also attempt to locate veiled or explicit antagonism to the act of criticism itself (Eve 2016) within popular fiction, perhaps through suggestions by narrators or characters that books should not be “dissected” through critical theory, but merely enjoyed.

7. Data Availability

Due to copyright restrictions, the full corpus cannot be made available publicly. Frequencies and results of queries can be accessed at https://github.com/erikannotatio/ns/King_data.

8. Acknowledgements

The authors would like to thank the editors and peer reviewers for their many insightful comments.

9. Author Contributions

Erik Ketzan: Conceptualization, Writing

Martin Paul Eve: Writing

References

- Algee-Hewitt, Mark and Mark McGurl (2015). *Between canon and corpus: six perspectives on 20th-century novels*. 2164-1757. Stanford Literary Lab.
- Anderson, James Arthur (2017). *The Linguistics of Stephen King: Layered Language and Meaning in the Fiction*. McFarland.
- Archer, Jodie and Matthew L Jockers (2016). *The bestseller code: Anatomy of the blockbuster novel*. St. Martin's Press.
- Blatt, Ben (2017). *Nabokov's Favorite Word is Mauve: What the Numbers Reveal about the Classics, Bestsellers, and Our Own Writing*. Simon and Schuster.
- Bloom, Harold (1997). *The anxiety of influence : a theory of poetry*. Second edition. New York: Oxford University Press.
- (2007). “Introduction”. In: *Stephen King*. Ed. by Harold Bloom. Updated ed. 1 online resource (vii, 228 pages). Vols. Bloom's modern critical views. New York: Chelsea House Publishers, 1–3. <https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=302687>.
- Bourdieu, Pierre (1984). *Distinction: A Social Critique of the Judgment of Taste*. Harvard University Press.

- Brezina, Vaclav, Pierre Weill-Tessier, and Anthony McEnery (2020). *LancsBox v. 5.1.2*. 480
<http://corpora.lancs.ac.uk/lancsbox>. 481
- Cranenburgh, Andreas van and Corina Koolen (2019). "The Literary Pepsi Challenge: 482
 intrinsic and extrinsic factors in judging literary quality". In: Digital Humanities 483
 2019 Conference. Utrecht University. 484
- De Vries, Robert and Aaron Reeves (June 2022). "What Does it Mean to be a Cultural 485
 Omnivore? Conflicting Visions of Omnivorosity in Empirical Research". In: *Socio-* 486
logical Research Online 27.2, 292–312. 10.1177/13607804211006109. 487
- Elliott, Jack (Sept. 3, 2015). "Whole Genre Sequencing". In: *Digital Scholarship in the* 488
Humanities, fqv034. 10.1093/llc/fqv034. (Visited on 05/15/2024). 489
- Francis, W. N. and H. Kučera (1979). *A Standard Corpus of Present-Day Edited American* 490
English, for Use with Digital Computers. Providence, RI. 491
- Gelder, Ken (Dec. 17, 2004). *Popular Fiction: The Logics and Practices of a Literary Field*. 492
 2nd ed. Routledge. 10.4324/9780203023365. 493
- Hakemulder, Jemeljan F. (Sept. 2004). "Foregrounding and Its Effect on Readers' Per- 494
 ception". In: *Discourse Processes* 38.2, 193–218. 10.1207/s15326950dp3802_3. 495
- Heiden, Serge (2010). "The TXM platform: Building open-source textual analysis soft- 496
 ware compatible with the TEI encoding scheme". In: 24th Pacific Asia conference 497
 on language, information and computation. Vol. 2. Issue: 3. Institute for Digital 498
 Enhancement of Cognitive Development, Waseda University, 389–398. 499
- Heller, Karen (2016). "Meet the Writers Who Still Sell Millions of Books. Actually, 500
 Hundreds of Millions". In: *The Washington Post*. [https://www.washingtonpost.com](https://www.washingtonpost.com/lifestyle/style/meet-the-elite-group-of-authors-who-sell-100-million-books-or-350-million/2016/12/20/db3c6a66-bb0f-11e6-94ac-3d324840106c_story.html) 501
[/lifestyle/style/meet-the-elite-group-of-authors-who-sell-100-million-b](https://www.washingtonpost.com/lifestyle/style/meet-the-elite-group-of-authors-who-sell-100-million-books-or-350-million/2016/12/20/db3c6a66-bb0f-11e6-94ac-3d324840106c_story.html) 502
[ooks-or-350-million/2016/12/20/db3c6a66-bb0f-11e6-94ac-3d324840106c_sto](https://www.washingtonpost.com/lifestyle/style/meet-the-elite-group-of-authors-who-sell-100-million-books-or-350-million/2016/12/20/db3c6a66-bb0f-11e6-94ac-3d324840106c_story.html) 503
[ry.html](https://www.washingtonpost.com/lifestyle/style/meet-the-elite-group-of-authors-who-sell-100-million-books-or-350-million/2016/12/20/db3c6a66-bb0f-11e6-94ac-3d324840106c_story.html). 504
- Herrmann, J Berenike (2017). "In a test bed with Kafka. Introducing a mixed-method 505
 approach to digital stylistics". In: *Digital Humanities Quarterly* 11.4. [https://www.di](https://www.digitalhumanities.org/dhq/vol/11/4/000341/000341.html) 506
[gitalhumanities.org/dhq/vol/11/4/000341/000341.html](https://www.digitalhumanities.org/dhq/vol/11/4/000341/000341.html). 507
- Horkheimer, Max and Theodor W Adorno (1947). "Dialektik der Aufklärung: Philosophis- 508
 che Fragmente [Dialectic of Enlightenment: Philosophical Fragments]". In: *Amster-* 509
dam, the Netherlands: Querido. 510
- Hutcheon, Linda (1988). *A Poetics of Postmodernism: History, Theory, Fiction*. New York: 511
 Routledge. 512
- Huyssen, Andreas (1986). *After the Great Divide*. London: Palgrave Macmillan UK. 10.1 513
 007/978-1-349-18995-3. 514
- King, Stephen (1974). *Carrie*. New York: Doubleday. 515
- (1975). *Salem's Lot*. New York: Doubleday. 516
- (1979). *The Long Walk*. New York: Pocket Books. 517
- (1981). *Cujo*. New York: Viking Press. 518
- (1982). *The Running Man*. New York: Signet Books. 519
- (1984). *The Eyes of the Dragon*. New York: Viking. 520
- (1987a). *Misery*. New York: Viking. 521
- (1987b). *The Tommyknockers*. New York: Putnam. 522
- (1992a). *Dolores Claiborne*. New York: Viking. 523
- (1992b). *Gerald's Game*. New York: Viking. 524
- (1999). *The Girl Who Loved Tom Gordon*. New York: Scribner. 525
- (2000). *Stephen King on Writing: A Memoir on the Craft*. Simon and Schuster. 526

- King, Stephen (2020). *If It Bleeds*. New York: Scribner. 527
- (2022). *Fairy tale*. New York: Scribner. 528
- Knoop, Christine A., Valentin Wagner, Thomas Jacobsen, and Winfried Menninghaus (June 2016). "Mapping the aesthetic space of literature "from below"". In: *Poetics* 56, 35–49. 10.1016/j.poetic.2016.02.001. 531
- Lessard, Greg (1992). "Computational modelling of linguistic humour: Tom Swifities". In: Selected Papers from the 1992 Association for Literary and Linguistic Computing (ALLC) and the Association for Computers and the Humanities (ACH) Joint Annual Conference. Oxford University Press, 175–178. 532
- Liberman, Mark (Dec. 29, 2017). *Proportion of Dialogue in Novels*. <http://languagelog.ldc.upenn.edu/nll/?p=35968>. 533
- Light, Alison (Aug. 21, 2013). *Forever England*. oth ed. Routledge. 10.4324/9780203713518. 534
- Mair, Christian (1992). *The Freiburg-Brown Corpus*. Freiburg im Breisgau. 535
- McHale, Brian (June 25, 2015). *The Cambridge Introduction to Postmodernism*. Cambridge University Press. 10.1017/CB09781139108706. 536
- Ollivier, Michèle (Apr. 2008). "Modes of openness to cultural diversity: Humanist, populist, practical, and indifferent". In: *Poetics* 36.2, 120–147. 10.1016/j.poetic.2008.02.005. 537
- Paquette, Jenifer (2014). *Respecting The Stand: A critical analysis of Stephen King's apocalyptic novel*. McFarland. 538
- Peterson, Richard A and Albert Simkus (1992). "How musical tastes mark occupational status groups". In: *Cultivating differences: Symbolic boundaries and the making of inequality* 152. 539
- Peterson, Richard A. and Roger M. Kern (Oct. 1996). "Changing Highbrow Taste: From Snob to Omnivore". In: *American Sociological Review* 61.5, 900. 10.2307/2096460. 540
- Piper, Andrew and Eva Portelance (2016). "How cultural capital works: Prizewinning novels, bestsellers, and the time of reading". In: *Post45* 10. 541
- Plato (2005). *Phaedrus*. Translated by Christopher Rowe. London: Penguin. 542
- Porter, J. D. (2018). *Popularity/Prestige*. 17. Stanford Literary Lab. 543
- Pullum, Geoffrey K. (Feb. 18, 2004). *Those Who Take the Adjectives from the Table*. <http://itre.cis.upenn.edu/~myl/languagelog/archives/000469.html>. 544
- (June 2010). "The Land of the Free and *The Elements of Style*". In: *English Today* 26.2, 34–44. 10.1017/S0266078410000076. 545
- (July 2014). "Fear and loathing of the English passive". In: *Language & Communication* 37, 60–74. 10.1016/j.langcom.2013.08.009. 546
- (Mar. 21, 2015). *Awful Book, so I Bought It*. <https://languagelog.ldc.upenn.edu/nll/?p=18345>. 547
- Rybicki, J. and M. Eder (Sept. 1, 2011). "Deeper Delta across genres and languages: do we really need the most frequent words?" In: *Literary and Linguistic Computing* 26.3, 315–321. 10.1093/lc/fqr031. 548
- Schmid, Helmut (1999). "Improvements in part-of-speech tagging with an application to German". In: *Natural language processing using very large corpora*. Springer, 13–25. 549
- Schöch, Christof (2017). "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama." In: *DHQ: Digital Humanities Quarterly* 11.2. 550

- Senf, Carol A (1998). "Gerald's Game and Dolores Claiborne: Stephen King and the Evolution of An Authentic Female Narrative Voice". In: *CONTRIBUTIONS TO THE STUDY OF POPULAR CULTURE* 67, 91–110.
- Sigelman, Lee and William Jacoby (1996). "The not-so-simple art of imitation: Pastiche, literary style, and Raymond Chandler". In: *Computers and the Humanities* 30.1, 11–28.
- Smythe, James (Feb. 5, 2015). "Rereading Stephen King, Chapter 31: Dolores Claiborne". In: *The Guardian*. <https://www.theguardian.com/books/2015/feb/05/rereading-stephen-king-chapter-31-dolores-claiborne>.
- So, Richard Jean (Dec. 31, 2021). *Redlining Culture: A Data History of Racial Inequality and Postwar Fiction*. Columbia University Press.
- Steve Evans, Jeri Kroll (Apr. 28, 2005). "How to Write a 'How to Write' Book: The Writer as Entrepreneur". In: *TEXT* 9.1.
- Strengell, Heidi (2005). *Dissecting Stephen King: from the Gothic to literary naturalism*. Popular Press.
- Strunk Jr., William and E. B. White (1999). *The Elements of Style*. 4th. London: Pearson.
- Texter, Douglas W. (Jan. 1, 2007). "A Funny Thing Happened on the Way to the Dystopia: The Culture Industry's Neutralization of Stephen King's *The Running Man*". In: *Utopian Studies* 18.1, 43–72.
- Underwood, Ted (2019). *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.
- Underwood, Ted and Jordan Sellers (Sept. 2016). "The *Longue Durée* of Literary Prestige". In: *Modern Language Quarterly* 77.3, 321–344.
- Van Cranenburgh, Andreas and Erik Ketzan (2021). "Stylometric Literariness Classification: the Case of Stephen King". In: *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. Punta Cana, Dominican Republic (online): Association for Computational Linguistics, 189–197.
- Van Cranenburgh, Andreas, Karina Van Dalen-Oskam, and Joris Van Zundert (Dec. 2019). "Vector space explorations of literary language". In: *Language Resources and Evaluation* 53.4, 625–650.
- Van Dalen-Oskam, Karina (June 26, 2023). *The Riddle of Literary Quality: A Computational Approach*. Amsterdam University Press.
- Van Peer, Willie (2008). *The quality of literature: Linguistic studies in literary evaluation*. Vol. 4. John Benjamins Publishing.
- Verboord, Marc (June 2003). "Classification of authors by literary prestige". In: *Poetics* 31.3, 259–281.
- Wynne, M. (2006). "Stylistics: Corpus Approaches". In: *Encyclopedia of Language Linguistics (Second Edition)*. Ed. by Keith Brown. Second Edition. Oxford: Elsevier, 223–226.
- Xan, Brooks (Sept. 7, 2019). "Stephen King: 'I Have Outlived Most of My Critics. It Gives Me Great Pleasure.'" In: <http://www.theguardian.com/books/2019/sep/07/stephen-king-interview-the-institute>.

Zwicky, Arnold (July 22, 2006). *How Long Have We Been Avoiding the Passive, and Why?* 617
Language Log. [http://itre.cis.upenn.edu/~myl/languagelog/archives/003380](http://itre.cis.upenn.edu/~myl/languagelog/archives/003380.html) 618
.html. 619

Neither Telling nor Describing Reflective Passages and Perceived Reflectiveness 1700-1945

Benjamin Gittel¹ 
Florian Barth² 
Tillmann Dönicke³ 
Luisa Gödeke⁴ 
Thorben Schomacker⁵ 
Hanna Varachkina⁴ 
Anna Mareike Weimer⁴ 
Anke Holler⁴ 
Caroline Sporleder³ 

1. Trier Center for Digital Humanities, Trier University, Trier, Germany.
2. Göttingen State and University Library (SUB) / Göttingen Centre for Digital Humanities, Göttingen University, Göttingen, Germany.
3. Göttingen Centre for Digital Humanities, Göttingen University, Göttingen, Germany.
4. German Department, Göttingen University, Göttingen, Germany.
5. Computer Science Department, Hamburg University of Applied Sciences, Hamburg, Germany.

Citation

Benjamin Gittel, Florian Barth, Tillmann Dönicke, Luisa Gödeke, Thorben Schomacker, Hanna Varachkina, Anna Mareike Weimer, Anke Holler, and Caroline Sporleder (2024). "Neither Telling nor Describing. Reflective Passages and Perceived Reflectiveness 1700-1945". In: *CCLS2024 Conference Preprints 3* (1). 10.26083/tuprints-00027390

Date published 2024-05-28

Date accepted 2024-04-04

Date received 2023-12-16

Keywords

annotation, reflective passages, narratology, literary change, literary reception, neural classifiers

License

CC BY 4.0 

Note

This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 3rd Annual Conference of Computational Literary Studies at Vienna, Austria, in June 2024.

Abstract. The paper analyses within-fiction reflections in 250 years of literary history. To this end, we formalised the concept of "reflective passage", demonstrate how our annotation categories are deduced from literary theory and derive three subphenomena – COMMENT, GENERALISATION, and NON-FICTIONAL SPEECH – that constitute literary reflection. A collaborative annotation serves (a) as basis for the training of a neural classifier and (b) as dataset for a reception experiment leading to the calculation of a "reflection score", a measurement for the perceived reflectiveness of a textual passage. The classifier is applied to a diachronic corpus of German-language literary fictions derived from the KOLIMO corpus through extensive metadata enrichment and filtering. The results suggest three boom periods of reflective passages: around 1755, 1835 and 1920 and show effects of text length, canonisation status and authors' sex.

1. Introduction

In 1795, Friedrich Schiller, in his famous poetological treatise "On Naïve and Sentimental Poetry", claims that "ancient" and "modern" poetry differ in their degree of reflection. While the naïve poet moves us by imitating nature, "by sensuous truth, by living presence" (Schiller 1985[1795], 194),¹

"[t]he case is quite otherwise with the sentimental poet. He *reflects* upon the impression that objects make upon him, and only in that reflection is the emotion grounded which he self experiences and which he excites in

1. The German original reads: „durch sinnliche Wahrheit, durch lebendige Gegenwart" (Schiller 2004[1795], 717).

us.” (Schiller 1985[1795], 196)²

This poetological distinction is linked in Schiller’s treatise with a philosophy of history in such a way that naïve poetry is possible in the present, but “latently anachronistic” (Prill 1994, 521): under the conditions of modernity, in which a “correspondence between [...] feeling and thinking” is hardly possible any more,³ poetry must increasingly become sentimental poetry, that is, a poetry that is moved “through ideas” (Schiller 1985[1795], 194, 197).⁴

More than 220 years after Schiller formulated this influential thesis, which has found a diverse echo especially in discourses on the “reflexivity” of the modernist novel (see Beebe 1976, Orr 1981), computational philological methods offer the possibility to study inner-literary reflections on a broad empirical basis. Using the example of German-language narrative fiction, the present paper will investigate whether literature indeed became more and more “sentimental” – as Schiller has it –, that is, whether it exhibits an increasing degree of reflectiveness.

Of course, the concept of “literary reflectiveness” or – maybe more wide-spread – “literary reflexivity” is till today a very complex one and there is no direct route from Schiller’s concept of sentimental (reflective) poetry, which is embedded in an entire anthropology and philosophy of history, to an annotation based and narratologically underpinned approach like ours. The concept of “literary” or “narrative reflexivity” (Williams 1998) belongs to a whole semantic field of (often interchangeably used) ‘big concepts’ like “metatextuality”, “metafiction”, “self-reflexivity” on the one hand (see Julie Tanner 2022) and rather text-passage oriented concepts like “authorial intrusions”, “commentary” or “digression” on the other hand. This may be one of the reasons why there is little consensus about the historical development of literary reflectiveness: While it is evident from a number of case studies that at least some early-modern works of literature exhibit significant traits of reflectiveness (see Zapf et al. op. 2005, 8, Henke op. 2005), it is by no means clear how this phenomenon developed in the context of a rapidly growing book market in the 19th century and a mass market in the 20th century.

Our approach aims at measuring the degree of reflectiveness of a narrative by identifying so-called “reflective passages”. In the next section, we will introduce our concept of a **reflective passage** and illustrate how we collaboratively annotated three different subtypes of reflective passages. Section 3 will present a questionnaire that was used to empirically assess the contribution of each of these subtypes (and their interplay) to readers’ perception of a textual passage being a reflection. Based on the statistical analysis of the results of this questionnaire we introduce the notion of **perceived reflectiveness** of a given text passage, which is measured by the **reflection score**. Section 4 will describe two neural classifiers: a multi-label and a binary classifier for identifying reflective passages. In section 5, we will present a diachronic analysis of reflective passages as well as perceived reflectiveness in German fiction based on these two classifiers, that allows for evaluating the hypothesis of a gradual increase of reflectiveness in the

2. The German original reads: „Ganz anders verhält es sich mit dem sentimentalischen Dichter. Dieser reflektiert über den Eindruck, den die Gegenstände auf ihn machen, und nur auf jene Reflexion ist die Rührung gegründet, in die er selbst versetzt wird und uns versetzt.“ (Schiller 2004[1795], 720)

3. The German original reads: „Übereinstimmung zwischen [...] Empfinden und Denken“ (Schiller 2004[1795], 717).

4. The German original reads: „durch Ideen“ (Schiller 2004[1795], 717).

modern period. Finally, we will summarise our results and sketch prospects for future research. 49 50

2. Reflective Passages and their Annotation 51

When speaking of reflective passages in the context of fictional literature, one may think of various things. Without a doubt, fictional narrative texts regularly stimulate reflections in readers. Authors of such texts also often engage in extensive reflection before or during writing. Reflective passages, in contrast, refer to those reflections that are present on the surface of the text in fictional narrative texts (Gittel 2022). The broad and complex field of the phenomenon of reflective passages becomes clear from the fact that they are referred to in research by many terms that are by no means synonyms, such as "authorial intrusion" (Dawson 2016), "commentary" (Chatman 1980, 226–252), "digression" (Esselborn 2007), "factual discourse" (Konrad 2017), "serious speech acts in fictional works" (Klauck 2015), "gnomic statement" (Mäkelä 2017), "narrator's comment" (Zeller 2007), or *Sentenz* ('aphorism', Reuvekamp 2007). Although reflective passages have been much discussed recently in connection with their specific manifestations in essayistic and encyclopaedic narrative (Ercolino 2014; Gittel 2015; Herweg et al. 2019), they are not a clearly delimited phenomenon either in narratology or in literary history. For a definition of **reflective passage**, however, one can draw on considerations of two more established terms in literary theory – 'comment'/'commentary' and 'non-fictional speech' – and one in linguistics, namely 'generalisation'. We consider a **reflective passage** as a textual passage that is either a comment, non-fictional speech, a generalisation or a combination of these three phenomena. Reflective passages greatly differ regarding their length, ranging from one clause to several sentences or whole paragraphs. The minimal length of a reflective passage being a clause, we will focus in our quantitative diachronic analysis (see section 5) on **reflective clauses** as the minimal unit of a reflective passage. Since the details of our annotation of these phenomena can be found elsewhere (cf. Barth et al. 2021, Gödeke et al. 2022, Weimer et al. 2022, Barth et al. 2022) we will introduce these phenomena by means of examples in the following and use the corresponding tags COMMENT, NON-FICTIONAL SPEECH, and GENERALISATION henceforth. 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77

"Comment" is listed in narrative theory alongside "report", "description" and "speech" as a fourth so-called "narrative mode" (Bonheim 1975, 329, see also Bonheim 1982). These four modes, which can overlap, are sufficient for a classification of all passages in a narrative text according to Bonheim. Comments express an evaluative attitude of the speaker towards diegetic state of affairs, illuminate his relationship to the diegesis, or the representation of the events. Thus, they can reveal the narrator's attitude towards characters or events or his interpretations and explanations of them, as well as his relation to the concrete representation respectively to narration/fictionality in general. To illustrate what this main type of within-fiction reflections may look like, we may take a look at the beginning of Goethe's "Elective Affinities" (square brackets are used here and in the following to highlight relevant passages; original wording of all examples can be found in the appendix): 78 79 80 81 82 83 84 85 86 87 88 89

- (1) Eduard - [let that be the name we give to a wealthy baron in the best years of his life]_{COMMENT} - Eduard had spent the loveliest hours of an April afternoon in his 90 91

nursery grafting young trees with shoots newly arrived for him. (J. W. v. Goethe 2008, 3) 92 93

The account of Edward's April afternoon is interrupted here by a (metafictional) comment that identifies the speaker as an entity that exercises power of designation over the entities of the narrated world. Overall, however, comment is a relatively heterogeneous class. In research, for example, comments on the story, which can have an interpretive, judgemental or generalising character, are distinguished from comments on the discourse (Chatman 1980, 226–252, see also the term „non-mimetic judgements“ in Martinez-Bonati and Silver 1981, esp. 32–33). Because of this heterogeneity, two criteria are often involved in the identification of comments, one formal and one content-related: According to the formal criterion, comments are those passages of text that are neither speech, report nor description. Like descriptions, they belong to the static mode according to Stanzel and are accompanied by narrative pauses (Stanzel 1988, 66, Martínez and Scheffel 2007, 46). One often speaks of "pure comment" in reference to such ex negativo identifiable passages (Bonheim 1975, 337). According to the criterion of content, these are passages that express an evaluative attitude of the speaker, his relationship to the event or the representation of the event. If this criterion is taken as a basis, comments can also occur within descriptions, character speech or narrator's report, so-called "integral comments" (ibid.). The following dialogue in Theodor Fontane's "The Stechlin" can serve as an example, in which Woldemar, the son of the old Stechlin, expresses his astonishment: 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112

(2) "Erratics?" "Yes, erratics," repeated Woldemar. "But if that word bothers you, you can call them monoliths too. [It's really remarkable, Czako, how extremely discriminating you get about phrases when you're not the one doing the talking at the moment] COMMENT..." (Fontane 2013, 10) 113 114 115 116

Please note that COMMENT is a relatively heterogeneous category that comprises different sub-phenomena: ATTITUDE is annotated whenever the speaker comments on fictional events, characters, objects or itself. INTERPRETATION is annotated when explanations or interpretations are provided in a passage through which the diegesis can be understood anew. METACOMMENT is annotated whenever the narrator comments on the fictionality of the story or the process of writing or telling the story. 117 118 119 120 121 122

In addition to comment, there is a second phenomenon relatively well described in literary theory that can be used to formalise the concept of reflective passages: the phenomenon of non-fictional speech in fictional texts. According to many theorists, fictional texts consist not only of fictional speech, which - according to a common characterisation - serves to construct the fictional world but also of non-fictional speech (Searle 1975, Klauk 2015).⁵ The typical case of non-fictional speech with an assertive character (in the speech act theoretical sense) is relevant to the question of reflections in literature. Characteristic of this phenomenon is that (1) an assertion/hypothesis about the real world is suggested in a clearly delimitable text passage and (2) the propositional content of the assertion/hypothesis can be read off from this text passage 123 124 125 126 127 128 129 130 131 132

5. Konrad also assumes the possibility of "fictional-factual text passages" (Konrad 2014, 447). Without being able to discuss this in detail here: Insofar as these fictional-factual passages have an assertive character, they also fall under the term "non-fictional speech" introduced in the following.

itself.⁶ Corresponding examples are the following: 133

(3) [All happy families resemble one another, but each unhappy family is unhappy in 134
its own way]_{NON-FICTIONAL SPEECH}. (Tolstoy 2017, 1) 135

(4) [Every country has its Samarkand and its Numancia]_{NON-FICTIONAL SPEECH}. That night, 136
both places were here with us on the Morava. [Numancia, located in the Iberian 137
highlands, had at one time been the last refuge from and bulwark against the 138
Roman Empire, while Samarkand, whatever it may have represented in history, 139
became and remains legendary, and will still be legendary when history is no 140
more]_{NON-FICTIONAL SPEECH}. (Handke 2016, 3) 141

Example (4) – more precisely the third sentence of the Handke quote – demonstrates 142
that NON-FICTIONAL SPEECH does not always have to take the form of GENERALISATION, even 143
though this is the case most often discussed in research (e.g. Vesper 2014). 144

Third, the phenomenon of GENERALISATION may be regarded as a subtype of reflective 145
passages in its own right. Although GENERALISATION is considered to be an indicator for 146
'non-fictional speech' and 'comment' (see Chatman 1980; Vesper 2014), its appearances 147
in narrative fiction are much less explored than 'comment' and 'non-fictional speech' (see 148
Gödeke et al. 2022 for a first attempt). AS GENERALISATION we annotate any statements 149
not made about specific objects, individuals, time periods, or spaces, but about whole 150
classes or groups of entities. 151

(5) Naphta responded, with disagreeable composure: "My good sir, [there is no such 152
thing as pure knowledge]_{GENERALISATION}." (Mann 1969, 397) 153

As in this example, non-fictional speech often co-occurs with generalisation. However, 154
generalisations can be about all sort of entities (characters, spaces, events) in the fictional 155
world as well. Generalisations and non-fictional speech (as comments) can also occur 156
within characters' speech: characters can make statements about whole classes or groups 157
of entities and characters can suggest in a clearly delimited text passage an hypothesis 158
about the real world whose propositional content (e.g. "there is no pure knowledge") 159
can be read off from this text passage itself. 160

Having examined the three reflection constituting phenomena, we will give a brief 161
overview of our annotation results. Our annotation corpus consists of 34 texts with 162
16893 sentences covering the time period from 1616 to 1942 (cf. <https://gitlab.gwdg.de/mona/korpus-public/-/releases/v5.1> and data publication). In general, 163
the first approximately 400 sentences of each text were annotated by two annotators 164
with a background in German Philology. 2–3 experts (authors of this paper) created 165
gold standards for all texts collaboratively adjudicating (i.e. review, accept, correct or 166
delete) the initial annotations. We compute inter-annotator agreement on clause-level 167
based on Fleiss' Kappa (κ , Fleiss 1971) and Mathet's Gamma (γ , Mathet et al. 2015), 168
cf. table 1. κ calculates agreement based on the differences for each clause while γ 169
respects the individual annotated passages as units in a continuum, and also partial 170
overlapping passages are compared as units instead of disjointed clauses. We, therefore, 171
consider that γ better represents the errors made by annotators for a category with 172
173

6. It should be noted that there is nothing attached to the term "non-fictional speech", which is particularly controversial among narratologists. One could also use another term, such as "passages with an assertive character", for the passages that fall under the above definition.

rather long passages such as reflection.⁷ Using Landis and Koch 1977's guideline for interpreting the results of κ , we achieve moderate values for COMMENT and substantial for both, GENERALISATION and NON-FICTIONAL SPEECH (see table 1) for κ . In our perception, γ generally tends to yield more conservative values compared to κ .

	κ	(σ)	γ	(σ)
GENERALISATION	.65	(.19)	.63	(.16)
COMMENT	.52	(.25)	.46	(.21)
NON-FICTIONAL SPEECH	.74	(.21)	.61	(.17)

Table 1: Clause-level inter-annotator agreement for each phenomenon, averaged over all texts (standard deviations in parentheses).

So far, we have presented the theoretical background for and our operationalisation of 'reflective passages' and the associated phenomena of 'comment', 'non-fictional speech' and 'generalisation' as well as our annotation results. We stipulated that whenever at least one of these three phenomena is present, such a passage is a **reflective passage**. In the following section, we will introduce the second central term for the envisioned diachronic analysis: perceived reflectiveness as represented by the "reflection score".

3. Survey and Reflection Score

We tested the perception of reflectiveness in a reception experiment conducted via a survey. In particular, we were interested in the contribution of individual phenomena (GENERALISATION, COMMENT, NON-FICTIONAL SPEECH) to the overall reflectiveness of a text passage and whether the passages that were not annotated with any of the above mentioned phenomena can be perceived as reflective. Our objective is to quantify the contribution of the three phenomena and their combinations to the perception of a textual passage as reflective.

The survey was designed as follows: First, we extracted passages from our corpus, more precisely, from texts after 1850 (because we assumed that our participants would more readily understand the language in these more modern texts than in many of the earlier texts). The extracted passages consisted of one sentence and were annotated with the tags GENERALISATION, COMMENT, NON-FICTIONAL SPEECH or their combinations.

Second, we manually chose ten sentences for each of the following groups:

- COMMENT only
- GENERALISATION only
- NON-FICTIONAL SPEECH only
- COMMENT + GENERALISATION + NON-FICTIONAL SPEECH
- COMMENT + NON-FICTIONAL SPEECH
- GENERALISATION + NON-FICTIONAL SPEECH
- COMMENT + GENERALISATION

Additionally, we extracted passages that do not carry any of these tags as negative examples. Altogether there were 100 passages in the survey.

7. This assessment was already given in a similar form in Weimer et al. 2022.

★Trifft folgende Aussage Ihrer Meinung nach zu?
"In der markierten Textpassage wird über etwas reflektiert."

Den Rotschimmel ließ ich natürlich auch zurück; ich brauchte ihn nicht mehr. Wir alle waren der Ansicht, daß meine Abwesenheit nur eine kurze sein werde. Es sollte aber anders kommen. Wir befanden uns, was ich noch gar nicht erwähnt habe, weil es auf die bisher erzählten Ereignisse keinen Einfluß gehabt hatte, mitten im Bürgerkriege.

Bitte wählen Sie eine der folgenden Antworten:

☐

trifft zu

☐

trifft eher zu

☐

teils-teils

☐

trifft eher nicht zu

☐

trifft nicht zu

Figure 1: Example question from the survey

For the better understanding of the passage, we provide the survey participants with the context of one sentence before and one sentence after the passage. The passage in question is highlighted (see Figure 1). We attach the following question to each of the passages with the corresponding answer options on the scale from 1 to 5:

- In your opinion, is the following statement true: "In the highlighted text passage, something is reflected upon"?⁸
- 1: false

2: somewhat false

3: neither true nor false

4: somewhat true

5: true

For our experiment, we used the web-based survey tool LimeSurvey (LimeSurvey 2023). It allows us to give the participants 30 randomly selected passages. We chose 30 passages as a good trade-off between obtaining a sufficient coverage for each passage in the survey while at the same time limiting the experimentation time for the participants. In total, we received 118 complete answers, in which the participants provided their assessments for all 30 passages.

For a statistical analysis, we averaged the ratings from all participants for each passage. When we speak of "reflection ratings", we refer to these averages. The left column in Table 2 shows that all three phenomena correlate with the reflection ratings, but to a varying degree. Using Dancey and Reidy 2004's naming convention, the correlation is weak for NON-FICTIONAL SPEECH and GENERALISATION, and moderate for COMMENT. This illustrates that none of our phenomena is perfectly congruent to (perceived) reflection.

In a next step, we created a logistic regression model to get insights into the interplay between the phenomena. As features, we used the three phenomena as main effects as well as all combinations as interaction effects. We ran both forward selection and backward elimination to determine the best model in terms of the Akaike information criterion (AIC), both leading to the same result: a model that uses all main effects and the interaction effect GENERALISATION*COMMENT. The model's coefficients are shown in the right column of Table 2. Note that the regression coefficients of the main effects sort

8. The survey was conducted in German.

	corr. (p)	coef. (p)
COMMENT	.61 (.000)	1.29 (.000)
GENERALISATION	.35 (.000)	.72 (.000)
NON-FICTIONAL SPEECH	.29 (.003)	.34 (.023)
GENERALISATION*COMMENT	–	–.61 (.039)
const.	–	–.72 (.000)

Table 2: Spearman’s correlation coefficient (left) and logistic regression weights (right) for the three phenomena (main effects) and the only significant interaction effect. p -values are shown in parentheses.

in the same way as their correlation coefficients.

Using the regression coefficients we can calculate a reflection score r for any passage with known labels for GENERALISATION, COMMENT OR NON-FICTIONAL SPEECH as follows:

$$r = \sigma([1.29 \cdot f_{\text{COMMENT}}] + [0.72 \cdot f_{\text{GEN.}}] + [0.34 \cdot f_{\text{NON-FICT. SPEECH}}] - [0.61 \cdot f_{\text{COMMENT}} \cdot f_{\text{GEN.}}] - 0.72)$$

$\sigma(x)$ denotes the logistic sigmoid function $\frac{1}{1+e^{-x}}$. This means that, for example, a passage that is annotated as COMMENT but neither GENERALISATION nor NON-FICTIONAL SPEECH receives the following reflection score:

$$r = \sigma([1.29 \cdot 1] + [0.72 \cdot 0] + [0.34 \cdot 0] - [0.61 \cdot 1 \cdot 0] - 0.72) = \sigma(1.29 - 0.72) = 0.64$$

The value of r lies between 0 and 1. Since $0.64 > 0.5$, the reflection score for COMMENT-only passages can be interpreted as “reflective”. Table 3 shows that:

- passages that feature none of our phenomena or only non-fictional speech are not perceived as reflective,
- passages that feature only generalisation are equally often perceived as reflective or non-reflective,
- while passages that contain both non-fictional speech and generalisation as well as passages that contain comment are perceived as reflective.

Generally, the presence of each of our phenomena increases the reflection score.

r	phenomena
.33	–
.41	NON-FICTIONAL SPEECH
.50	GENERALISATION
.58	GENERALISATION & NON-FICTIONAL SPEECH
.64	COMMENT
.66	COMMENT & GENERALISATION
.71	COMMENT & NON-FICTIONAL SPEECH
.73	COMMENT & GENERALISATION & NON-FICTIONAL SPEECH

Table 3: Reflection scores for all label combinations

While further research would be necessary to understand why certain combinations tend to be perceived as reflective more often than others, another question is, whether the perception of a reflective passage actually triggers reflection on the part of the reader. We have to leave such intriguing questions for (psychological) researchers, but may emphasize two more general insights from our experiment: On the one hand, we can assume that our 'flexible' operationalization of a "reflective passage" captures basic intuitions about what it is "to reflect upon something". On the other hand, this results in a hierarchisation of the subphenomena we examined, which have a varying degree of influence on whether a certain passage is perceived as reflective.

4. Neural Classifier for Reflection

So far, we developed a basic definition of "reflective passage" and a more complex reflection score in order to analyse literary reflection. Since both rely on the identification of the three reflective subphenomena (GENERALISATION, COMMENT and NON-FICTIONAL SPEECH), we trained two neural classifiers for the automatic tagging of these phenomena: one multi-tagger and, additionally, one binary tagger (reflective vs. non-reflective passage). To our knowledge this has not been tried before. Each classifier takes a text span of three sentences as input, where one clause of the inner sentence is marked, and was trained to predict the categories of the marked clause.⁹ We split our corpus text-wise into training, development and test set so that the distribution of GENERALISATION, COMMENT and NON-FICTIONAL SPEECH is similar in all sets. Wieland's "The History of Agathon" and Seghers' "The Seventh Cross" are held out for the evaluation of the models, and Fontane's "The Stechlin" and Mann's "The Magic Mountain" serve as development set, while the other texts are used for training.¹⁰ The classifiers are available through the software package (Dönicke et al. 2022).¹¹

We followed the approach of Schomacker et al. 2022. The multi-label classifier has three output neurons, where each neuron corresponds to one tag (GENERALISATION, COMMENT, NON-FICTIONAL SPEECH), and the binary classifier has one (REFLECTIVE). Both classifiers are based on a large BERT model, that was pre-trained on German data (Chan et al. 2020),¹² and were trained for 20 epochs with a batch size of 8. To increase the convergence speed, we used the LAMB optimiser with a learning rate of 10^{-4} (You et al. 2020). Furthermore, we set the hidden dropout to 0.3 and the attention dropout to 0.0.

Table 4 shows Precision, Recall and Fscore of our classifiers on the test texts (cf. Sokolova and Lapalme 2009). For GENERALISATION, the multi-label reflection classifier performs with 61% F1 like the binary GENERALISATION-only classifier from Schomacker et al. 2022, which illustrates that the other two phenomena can be learned in addition without performance loss. The same classifier achieves with 69% F1 the best results for COMMENT, and hereby outperforms the statistical COMMENT-only classifier from Weimer et al. 2022 by 10%. Overall, the multi-label reflection classifier achieves a micro-averaged F1 score of 66% and the binary reflection classifier adds 3% on top of that. While the multi-label

9. The clauses are detected within our NLP pipeline MONAPipe (cf. <https://gitlab.gwdg.de/mona/pipy-public> and software publication) using our own algorithm for clause segmentation (Dönicke 2020).

10. We also excluded Kleist's "Michael Kohlhaas" from the training set, because the annotated text part does not contain one of our phenomena (non-fictional speech).

11. See <https://gitlab.gwdg.de/mona/pipy-public> and software publication.

12. <https://huggingface.co/deepset/gbert-large>.

classifier achieves a similar performance on both test texts ($\pm 7\%$), the binary classifier 291
shows a greater variation in F1 ($\pm 18\%$). 292

		GENERALISATION			COMMENT			NON-FICTIONAL SPEECH			micro-avg.		
		P	R	F	P	R	F	P	R	F	P	R	F
NN-multi	all texts	.52	.74	.61	.79	.61	.69	.78	.53	.63	.68	.63	.66
	└ Wieland	.53	.74	.62	.80	.68	.74	.70	.50	.59	.68	.66	.67
	└ Seghers	.52	.73	.61	.75	.38	.51	1.00	.59	.74	.68	.53	.60
NN-binary	all texts	—	—	—	—	—	—	—	—	—	.77	.62	.69
	└ Wieland	—	—	—	—	—	—	—	—	—	.77	.69	.73
	└ Seghers	—	—	—	—	—	—	—	—	—	.80	.42	.55

Table 4: Clause-level Precision (P), Recall (R) and Fscore (F) of our neural models for classifying clauses according to reflection in the test texts.

5. Diachronic Analysis 293

This section will first introduce our diachronic corpus “KOLIMO-selection” (1700-1945, 294
see 5.1). In a second step, we report the results of our diachronic corpus analysis (see 295
5.2). In addition to the “reflection score”, we analysed the presence of the three subtypes 296
of reflective passages (COMMENT, GENERALISATION, NON-FICTIONAL SPEECH), that according 297
to our initial definition constitute a “reflective passage”. In a third step, we took into 298
account potential covariates that may relate to the distribution of reflective passages in 299
literary history, like text length, canonisation status and author’s sex (see 5.3). 300

5.1 Corpus Building, Metadata Enrichment and Data Cleaning 301

For our analyses, we used a subset of the “German Corpus of Literary Modernism” 302
(KOLIMO, Herrmann 2023), which comprehends more than 41k texts and spans the 303
period mainly from 1500-1930. We filtered KOLIMO to obtain a subcorpus (“KOLIMO- 304
selection”) which fulfils the following criteria: 305

- only German fiction 306
- no translations into German 307
- only first editions 308
- only works with known first publication year 309
- no duplicates 310
- being balanced in the sense of single authors not being overrepresented 311
- minimum text length of 10 sentences 312

Concretely, we proceeded as follows. For each step either an annotation is performed or 313
a filtering is applied (see table 5): 314

1) Metadata enrichment: We identified texts with metadata on first publication years, 315
and enriched the corpus with data on the canonisation status (see Brottrager et al. 2021) 316
and data on the authors’ sex (relying on publicly available data on German first names, 317
Neumann 2018). We also relied on metadata concerning publication years from the 318
Corpus d-Prose (Gius et al. 2021, a metadata-enriched subset from KOLIMO which 319
covers the period from 1870-1920 only. 320

2) Author annotation: We manually annotated at the author-metadata level “predomi- 321

nantly fiction-author" vs. "predominantly non-fiction-authors". We filtered KOLIMO 322
and excluded a) texts without author or title, b) duplicates, c) works from overrepre- 323
sented authors (>500 texts) and d) works from predominantly non-fiction-authors such 324
as Kant, Freud, or Hegel. The threshold of more than 500 texts is a qualitatively explored 325
boundary set to exclude artifacts of highly productive authors that (apparently) have 326
been created by adding texts from text collections or chapters/paragraphs from books 327
as separate texts from one author/ editor to KOLIMO. This left us with 9467 texts. 328

3) Neural classifier: We applied the neural classifier for the corpus, which tags reflective 329
clauses. Some texts (196) could not be processed by the classifier due to artefacts in the 330
text file such as unexpected character encodings etc. These texts were dropped. 331

4) Publication year annotation: We manually annotated the first publication year of 332
texts without publication year relying on the following digitally available databases and 333
multi-volume reference works: Arend et al. 2022, Arnold 2020, Kühlmann 2012, and 334
only as last resort GoogleBooks. Annotators were also asked to mark non-German, non 335
narrative, non-fictional and translations into German. Based on this data, we filtered 336
our corpus a second time, which left us with 6218 texts. 337

5) Fiction status annotation: Since we observed that our corpus still contains non- 338
fictional narrative texts, we undertook a further annotation: We manually annotated 339
the fictionality status (fiction/ non-fiction / unclear) of texts that contained more than 340
9.94 percent non-fictional speech at clause-level (the 75-percent quantile) according 341
to the results of our multi-label classifier, thereby using a disproportionately high 342
share of non-fictional speech as a heuristic to identify remaining non-fiction in our 343
corpus. Subsequently, we removed texts that have been identified as non-fiction by our 344
annotators from our corpus. 345

6) Data cleaning: In a last step, we removed outliers regarding the proportion of re- 346
flective clauses per text (interquartile range method), that are partly due to wrong or 347
incomplete texts being part of the KOLIMO corpus (e.g. novel-prefaces instead of the 348
novel itself). The resulting subcorpus ("KOLIMO-selection", 1700-1945) contains 5209 349
original German language fictions with known first publication year. 350

Table 5 provides an overview of the filtering process and Figure 2 of the resulting 351
KOLIMO-selection corpus. 352

Step	Dropped	Remaining
1) Metadata enrichment	0	41,382
2) Author annotation		
Texts without author and without title	340	41,042
Texts without author-classification	23	41,019
Duplicates	924	40,095
Texts from non-fiction authors	15,740	24,355
Overrepresented authors (>500 texts)	12,789	11,566
Texts from non-German writing authors	2,099	9,467
3) Neural classifier		
Texts with exceptions during processing	196	9,271
4) Publication year annotation		
Texts without first publication*	2,633	6,639
Translations	44	6,595
Non-german language texts	0	6,595
Non-fictional texts	192	6,403
Non-narrative texts	4	6,399
Texts with less than 10 sentences	181	6,218
5) Fiction status annotation		
Non-fiction or texts with unclear fiction status	360	5,858
6) Data cleaning		
Texts before 1700	167	5,691
Texts after 1945	134	5,557
IQR-based outliers (> 61.68% reflective clauses)	348	5,209

Table 5: Overview of filtering the KOLIMO corpus; * at this step we additionally excluded 463 texts from one over-represented author with the same publication year

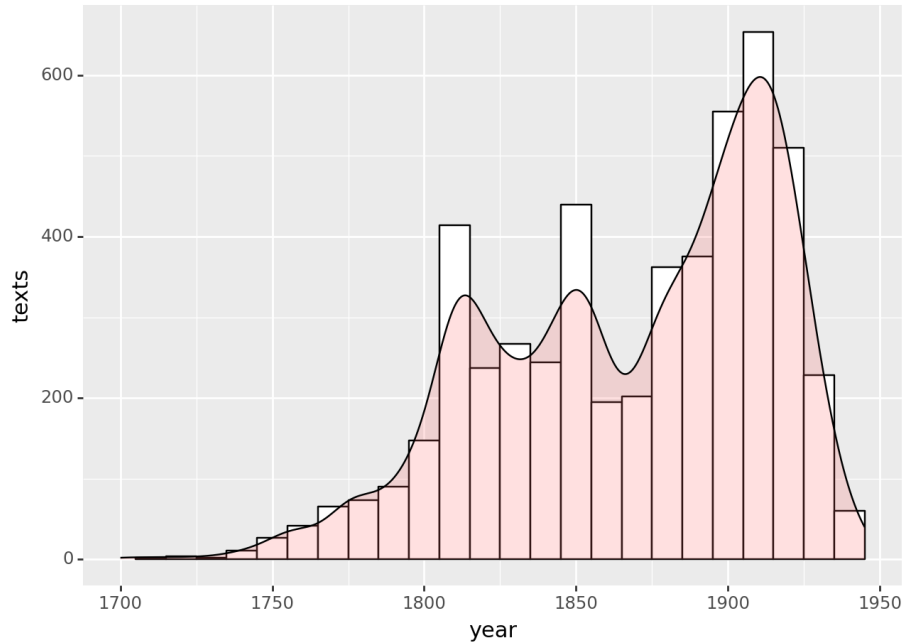


Figure 2: Distribution of texts in KOLIMO-selection corpus over time

5.2 Reflective Passages and Perceived Reflectiveness

353

Since the reader is by now familiar with our diachronic corpus and the assumptions 354
built into it, we can start with the intended analysis of the development of reflective 355

passages in 250 years of literary history. In a first step, we take a look at the reflection score, which represents the perceived reflectiveness of a text as explained above. Figure 3 shows the annual mean of the reflection score.

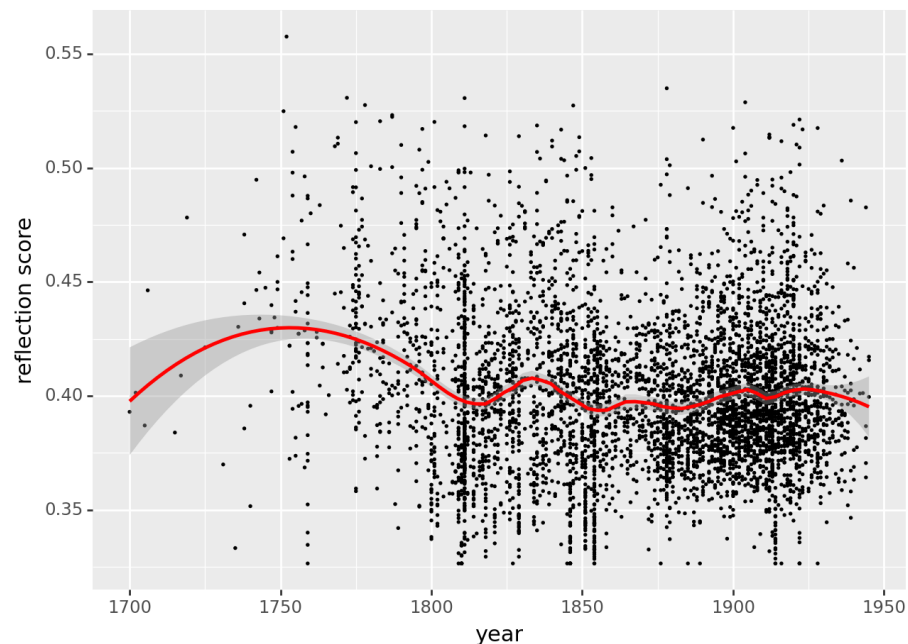


Figure 3: Perceived reflectiveness from 1700 to 1950

It can be observed that the average perceived reflectiveness is relatively stable (between 0.38 and 0.43) over time. Keeping in mind that the baseline reflection score, that means where none of our three phenomena is present, is 0.33 ($\sigma(-0.72)$) (cf. Table 3 above), this is very plausible: The average German fiction contains some reflections. A second interesting result are the three local maxima around 1755, 1830 and 1920. The first maximum may explain how Schiller, when he wrote "On Naïve and Sentimental Poetry" in 1795, arrived at his initially cited claim, that literature is becoming more and more reflective: In fact, Schiller looked back on a period in which fiction was more reflective than before. Although, in his famous essay, he mainly cites examples from antiquity – Homer as naïve and Horaz as sentimental (reflective) poet – he does mention "the sentimental poets of the French, and the Germans, [...], of the period from 1750 to about 1780", who seemed long time more appealing to him than 'the naïve Shakespeare'. (Schiller 1985[1795], 191). Figure 3 seems to confirm Schiller's subjective impression. The local peak around 1920 (which forms a saddle with the local peak shortly after 1900) dovetails nicely with the research thesis that there was a boom in essayism in the beginning of the 20th century that describes one aspect of the general trend toward the "dissolution of the boundaries of forms" (Kiesel 2004, p. 153): on the one hand, fictional essays emerged, and on the other, essayistic passages increasingly found their way into fiction, especially into the novel (see Ercolino 2014; Jander 2008; Just 1960; Müller-Funk 1995). However, the increase of perceived reflectiveness is less pronounced as one might have expected from the amount of research that exists on the phenomenon of essaysim in that period. The peak around 1835 is an interesting finding, which may relate to a politicisation of literature during the *Vormärz* period. However, further research beyond

the scope of this paper is needed to underscore such an hypothesis.

In a next step, we take a closer look at the frequency of reflective passages and their subtypes. Please recall that reflective passages greatly differ regarding their length, ranging from one clause to several sentences or whole paragraphs. For that reason, we carry out the following analyses at the clause level and speak of **reflective clauses**. Figure 4 represents the proportion of reflective clauses over time. Please note that we count a clause as reflective –according to our initial definition–, if at least one of our three phenomena (COMMENT, GENERALISATION, NON-FICTIONAL SPEECH) is present. The confidence intervals, here as in the following, are calculated with Python’s ggplot2 implementation “plotnine” employing LOESS smoothing with a span parameter of 0.3.

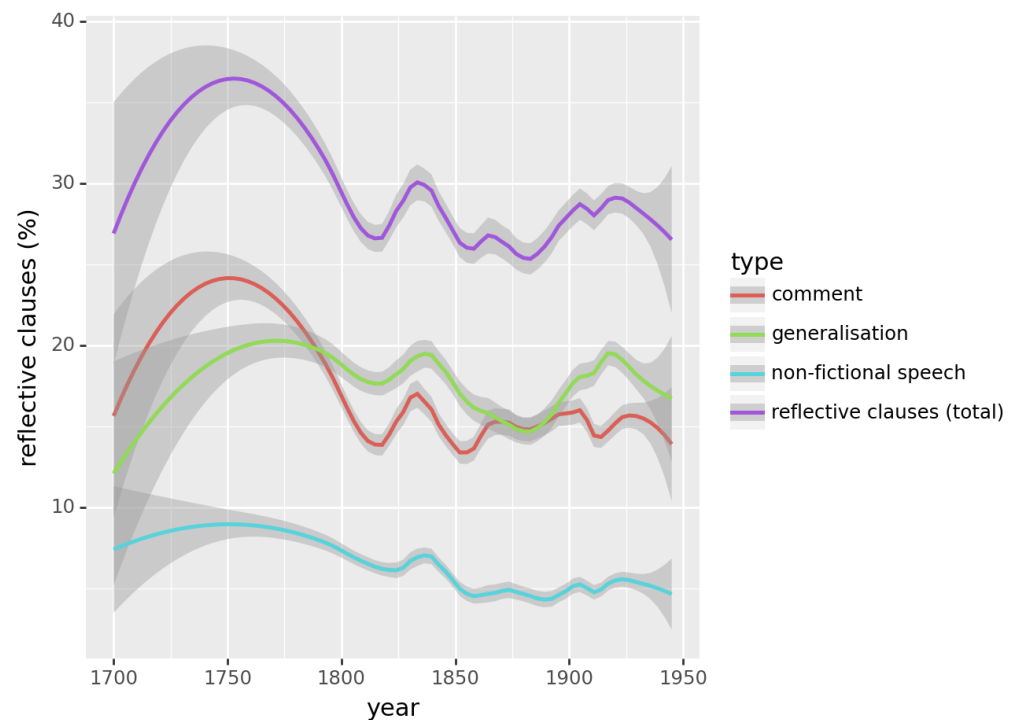


Figure 4: Reflective clauses and their subtypes over time

One may observe four things: 1) The proportion of reflective passages (violet graph) is high over the 18th century ($>30\%$), drops below 30% in 1800, reaches a local peak 1830 and another 1920. However, these local peaks in the 19th and 20th century never reach the level of the 18th century. The period of realism forms a tale, in which literary reflections are less widespread. 2) The shape of the graphs are very (or for COMMENT: relatively) similar one to another and to the reflection score graph in Figure 3. This indicates that the three phenomena do indeed co-evolve and represent different aspects of the overall phenomenon of reflection in fiction. 3) Only two graphs intersect: GENERALISATION (green) and COMMENT (red). In the end of the 18th century COMMENT loses its position as most common subtype to GENERALISATION, which it more or less keeps till 1945. Only during the period of realism, GENERALISATION is less predominant, its “pole position” being contested by COMMENT again. 4) As one might expect, NON-FICTIONAL SPEECH is the least frequent subtype. Interestingly, its development can be cut into two halves: Between 1700 and 1840 it has a significant share between 7.5 and 10%, but after

1850 its proportion is more or less stable around 5%. 406

5.3 Effects of Text Length, Canonisation Status and Sex 407

This section is dedicated to the analysis of three factors that plausibly may correlate 408
with fictions' degree of reflectiveness: text length, canonisation status and authors' sex. 409
For example, the fact that the phenomenon of within-fiction reflections has attracted 410
attention primarily in novel research might indicate that reflective passages occur more 411
often in novels than in shorter texts. To scrutinise this hypothesis, we calculated quantiles 412
in the distance of 25% based on text length in tokens separating our corpus in four parts: 413
very short, short, long and very long texts. Very long texts have more than 58k tokens (i.e. 414
> 4800 sentences based on an estimate of 12 tokens per sentence). Since our diachronic 415
corpus contains almost only prose fiction, this category can be interpreted as "novels". 416
Table 6 shows the proportion of reflective passages grouped by text length. 417

	Mean	SD	SEM
Text length			
Very short	26.63	14.70	0.41
Short	27.63	11.88	0.33
Long	29.26	9.88	0.27
Very long	29.22	9.46	0.26

Table 6: Proportion of reflective clauses (%) and text length

Longer texts tend to be more reflective than shorter texts, although differences are 418
delicate, overall. There is almost no difference between long texts (e.g. novellas) on 419
the one hand and very long texts (e.g. novels) on the other hand. A further analysis 420
revealed that long and very long texts contain on average more COMMENT passages 421
(almost 18%) than very short and short texts (12% resp. 14.6%), while the values for 422
the other subtypes are very similar. 423

Another plausible hypothesis is that canonical texts are more reflective than others, 424
because complexity is often seen as a text-related standard that may favour canonisation 425
(see Winko 2002, pp. 21-22). Therefore, we added information on the canonisation 426
status (the so-called "canonisation score" based inter alia on work-mentions in literary 427
histories and anthologies as proposed by Brottrager et al. 2021), of 357 texts that we 428
were able to identify in our KOLIMO-selection. Table 7 compares these texts against all 429
other (non-canonical) texts. 430

	Mean	SD	SEM
Canonisation status			
Canonical	30.71	11.16	0.59
Non-canonical	28.00	11.74	0.17

Table 7: Proportion of reflective clauses (%) and canonisation

The group difference presented here is statistically significant as a *t*-test reveals: Canon- 431
ical texts contain on average 2.7% more reflective passages than non-canonical texts 432
($t(5207) = 4.22, p < .001, d = 0.23$). However, the relation between the degree of reflec- 433
tiveness and canonisation is more complex as Figure 5 reveals. It represents the relation 434

between canonisation score (highest degree of canonisation, values from 0 to 1) and the proportion of reflective clauses of a text (taking only the 357 texts with canonisation score into account).

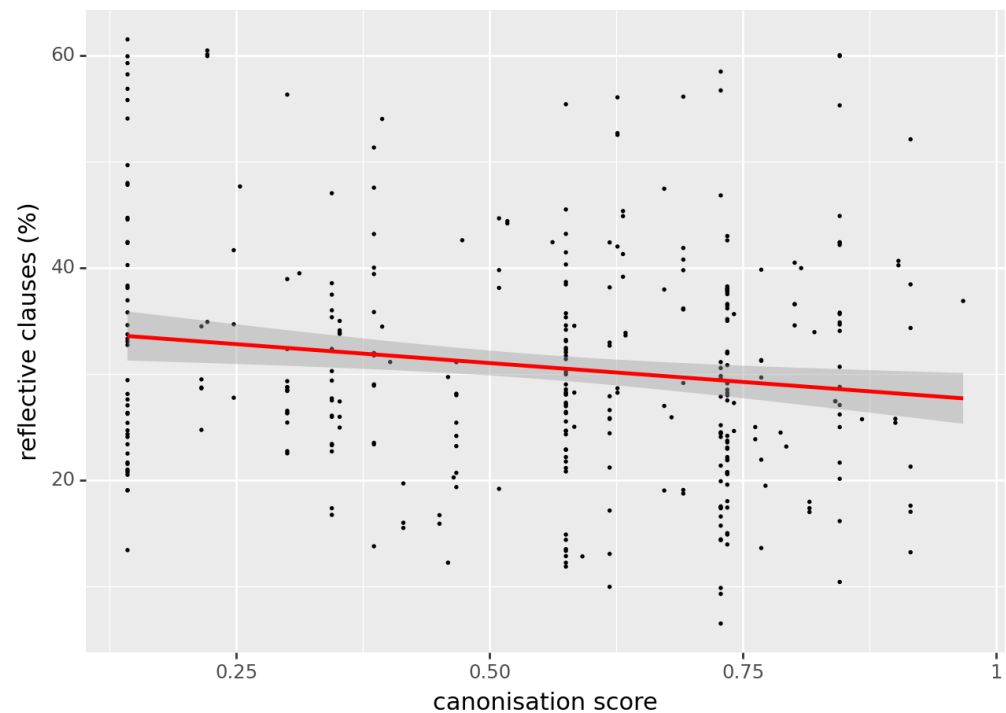


Figure 5: Proportion of reflective clauses in function of canonisation status, $n = 357$

One observes that the relation is negative: the *less* reflective clauses a text contains, the more canonised the text is. Taking this result together with the previous one (that canonised texts contain on average more reflection), this seems to suggest that a *moderately* increased degree of reflectiveness favours canonisation. We intentionally formulate this hypothesis in cautious terms, because there are many other factors involved about which we have no information. However, there is one aspect of the complex relationship we can explore: the diachronic dimension (see Figure 6). The restricted temporal coverage is due to the fact that there are no canonical works before 1750 in our corpus.

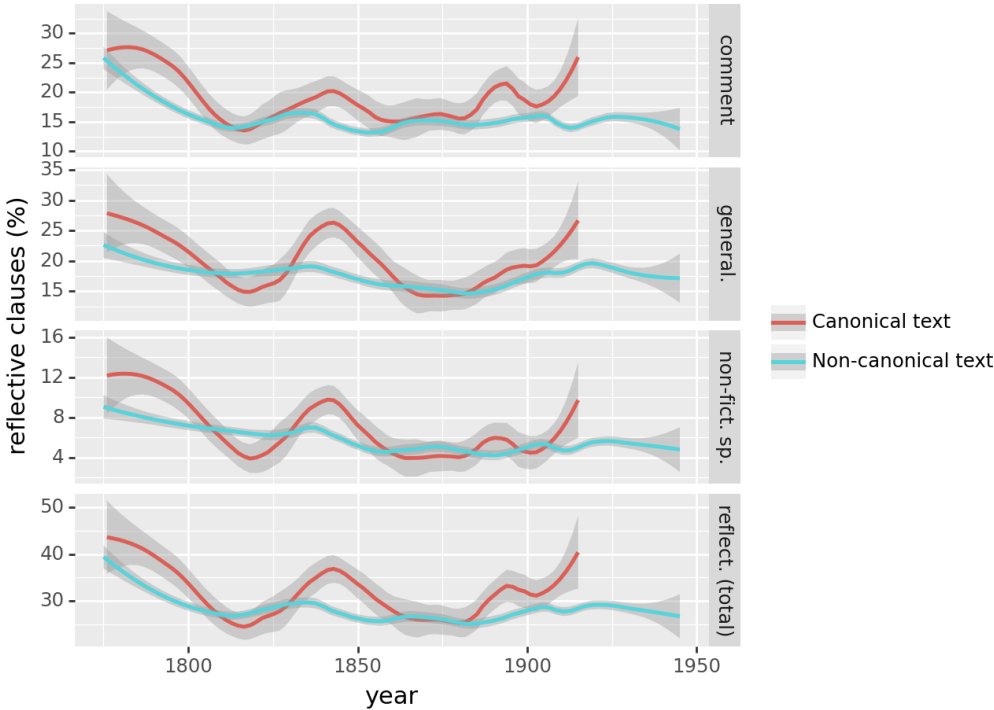


Figure 6: Proportion of reflective clauses and canonisation status over time

Figure 6 reveals several things: 1) The observed mean difference for reflective clauses 446
between canonical and non-canonical texts is due to relatively specific time periods, 447
especially in the middle and in the end of the 19th century and in the beginning of the 448
20th century. 2) There is a remarkably steep increase for COMMENT and NON-FICTIONAL 449
SPEECH for canonical texts in the beginning of the 20th century. For canonical texts, one 450
may indeed witness the boom of reflection that one could have expected given the above 451
mentioned research. This underscores how much traditional research is driven by its 452
attention to relatively few more or less canonical texts; the ratio between canonical texts 453
and non-canonical texts in our KOLIMO-selection being 1 to 13,6 (357 to 4852 texts). 454

As a third factor for analysis, we selected the authors' sex. From 5.2k texts more than 455
1.4k texts are from female authors. Table 8 shows that there is an association with the 456
mean proportion of reflective clauses: Male authors tend to use reflective passages on 457
average more often than female authors. 458

	Mean	SD	SEM
Authors' sex			
Female	26.28	12.25	0.33
Male	28.66	11.49	0.20

Table 8: Proportion of reflective clauses (%) and authors' sex

This finding is confirmed by a *t*-test ($t(4561)=6.23$, $p<0.001$), which reveals a small 459
effect ($d=0.20$). However, this is only a very general result in the light of the highly 460
varying presence of female authors in literary history. For this reason, Figure 7 enables 461
the reader to take a closer look on the interrelations of reflective clauses and authors' 462
sex over time. 463

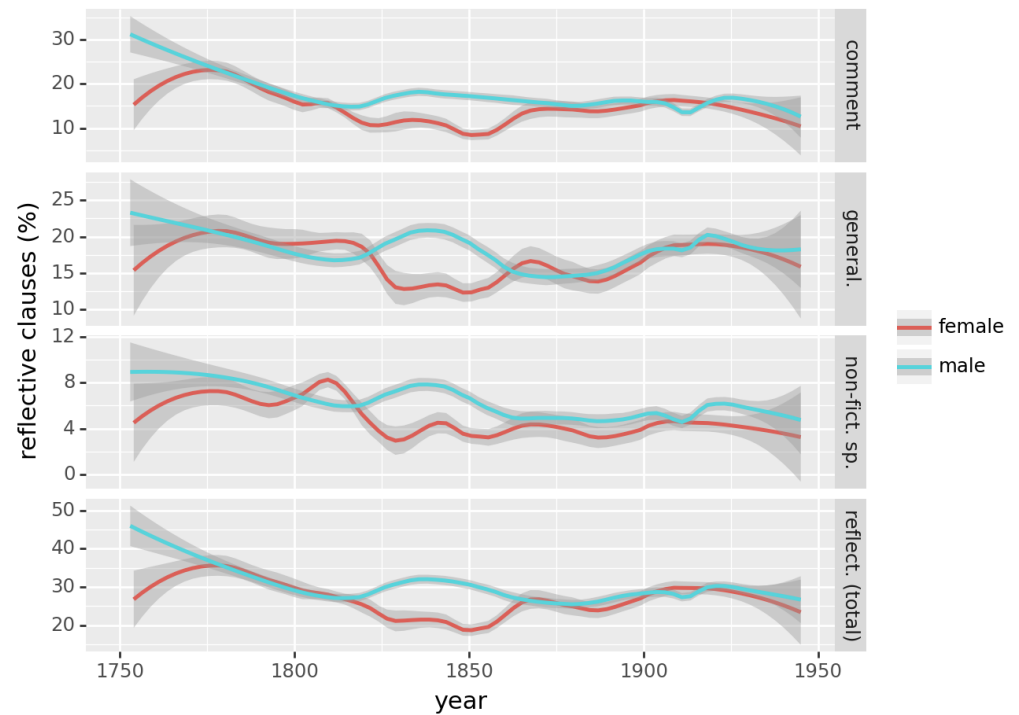


Figure 7: Proportion of reflective clauses and authors' sex over time

From Figure 7 it becomes clear that the more frequent usage of reflective passages by male authors is mainly due to developments before 1875, where female authors – with one exception in the beginning 19th century – reflect less often on average in their fictions. From 1875 onward female authors use reflective passages on average as often as their male counterparts. Only in the 1920s, a new discrepancy seems looming regarding NON-FICTIONAL SPEECH, which tends to be used less often by female authors.

6. Summary

A so far unfulfilled promise of Computational Literary Studies is to write a more empirically saturated history of literature. Our aim in this paper was to contribute to this new literary history through a diachronic analysis of the narratological phenomenon of reflective passages. Our approach illustrates how many different elements have to come together to get closer to this goal: After 1) a resource-intensive annotation of more than 16k sentences for the phenomenon of reflection, we were able 2) to build a multi-label and a binary classifier for reflective passages. 3) We studied how different types of reflective passages are perceived by actual readers and introduced the reflection score as a measure for perceived reflectiveness of a textual passage. 4) Through a complex filtering process, we build an suitable diachronic corpus of 5.2k original German language fictions from the much larger KOLIMO corpus and 5) enriched their metadata regarding fictionality status, canonisation status and authors sex. Finally, we were able to analyse the frequency of reflective passages over 250 years of literary history. Our findings suggest three boom periods of reflective passages: around 1755, 1835 and 1920. GENERALISATION is the most common phenomenon (M=17.6% of all clauses), COMMENT the second common (M=15.6%), while NON-FICTIONAL SPEECH is rather rare (M=5.6%).

In terms of perceived reflectiveness, all sub-phenomena contribute to a textual passage's reflectiveness, while COMMENT is the best indicator, GENERALISATION plus NON-FICTIONAL SPEECH also indicate reflectiveness. Important covariates of the proportion of reflective clauses are text length, canonisation status and authors' sex. On average, longer texts, canonised texts, and texts from male authors contain more reflective clauses than their respective counterparts. Since our diachronic corpus itself is only a (small) sample from the literary production in German language (cf. Gittel 2021, 5), and —due to limited metadata— does allow to control only a few potential covariates that steer literary production, our results should be regarded as motivation for further quantitative research in the future. Nevertheless, our research represents a step forward towards an empiricisation of literary studies. It demonstrates that quantitative research can underpin existing hypotheses in literary studies (like the one from a boom of essayism in the beginning of the 20th century) and set new questions on the agenda (e.g. about the nature of the boom of reflection in the *Vormärz* period). To answer such questions, Computational Literary Studies and hermeneutic research need to go hand in hand in our opinion. Quantitative research may in the future shed light on the thematic contents of the different subtypes of reflection and their combinations – a question deliberately put aside in the present paper – and hermeneutic research may formulate justified hypotheses about the functions of different types of reflective passages in specific contexts. In this way, literary studies may advance towards an empirically saturated functional literary history.

7. Appendix: Examples in Original Wording

- (1') Eduard – [so nennen wir einen reichen Baron im besten Mannesalter]_{COMMENT} – Eduard hatte in seiner Baumschule die schönste Stunde eines Aprilmittags zugebracht, um frisch erhaltene Pflanzfreier auf junge Stämme zu bringen. (J. W. Goethe 2021[1809], 7)
- (2') „Findlinge?“ „Ja, Findlinge,“ wiederholte Woldemar. „Aber wenn Ihnen das Wort anstößig ist, so können Sie sie auch Monolithe nennen. [Es ist merkwürdig, Czako, wie hochgradig verwöhnt im Ausdruck Sie sind, wenn Sie nicht gerade selber das Wort haben]_{COMMENT} ...“ (Fontane 2015[1897/98], 17)
- (3') [Все счастливые семьи похожи друг на друга, каждая несчастливaя семья несчастлива по-своему]_{NON-FICTIONAL SPEECH}. (Толстой 1998[1878], 7)
- (4') [Jedes Land hat sein Samarkand und sein Numancia]_{NON-FICTIONAL SPEECH}. In jener Nacht lagen die beiden Stätten hier bei uns, hier an der Morava. [Numancia, im iberischen Hochland, war einst die letzte Flucht- und Trutzburg gegen das Römerreich gewesen; Samarkand, was auch immer der Ort in der Historie darstellte, wurde und ist sagenhaft; wird, jenseits der Geschichte, sagenhaft sein]_{NON-FICTIONAL SPEECH} (Handke 2008, 7)
- (5') Naphta erwiderte mit unangenehmer Ruhe: "Guter Freund, [es gibt keine reine Erkenntnis]_{GENERALISATION}." (Mann [1924] 1991, 207)

8. Data Availability 527

Data can be found here: <https://zenodo.org/records/10246193>, and here: <https://doi.org/10.5281/zenodo.11164190> 528
529

9. Software Availability 530

Software can be found here <https://doi.org/10.5281/zenodo.11163719>, and here 531
<https://doi.org/10.5281/zenodo.11164036> 532

10. Acknowledgements 533

This work is funded by Volkswagen Foundation (Weimer, Dönicke, Gödeke, Holler, 534
Sporleder, Gittel), and by the Deutsche Forschungsgemeinschaft (DFG, German Re- 535
search Foundation) – 424264086 (Barth, Varachkina, Holler, Sporleder, Gittel). In ad- 536
dition to our funders, we cordially thank our research assistants: Friederike Altmann, 537
Annika Labitzke, Jan Lau, Jonas Lipski, Nele Martin, Thorben Neitzke, Evelyn Ovsjan- 538
nikov, Benita Pangritz, Lennart Speck, Janina Schumann, Noreen Scheffel, Ruben van 539
Wijk, and Marina Wurzbacher. 540

11. Author Contributions 541

Benjamin Gittel: Supervision, Funding acquisition, Conceptualization, Formal analysis, 542
Visualization, Writing – original draft, Writing – review & editing 543

Florian Barth: Project administration, Data curation, Formal analysis, Resources, Soft- 544
ware, Writing – original draft, Writing – review & editing 545

Tillmann Dönicke: Data curation, Formal analysis, Methodology, Resources, Software, 546
Writing – original draft, Writing – review & editing 547

Luisa Gödeke: Conceptualization, Investigation, Writing – original draft, Writing – 548
review & editing 549

Thorben Schomacker: Software, Writing – original draft, Writing – review & editing 550

Hanna Varachkina: Writing – review & editing 551

Anna Mareike Weimer: Conceptualization, Investigation, Writing – original draft, 552
Writing – review & editing 553

Anke Holler: Funding acquisition, Supervision, Writing – review & editing 554

Caroline Sporleder: Funding acquisition, Methodology, Supervision, Writing – review 555
& editing 556

References

557

- Arend, Stefanie, Jahn Bernhard, Jörg Robert, Robert Seidel, Johann Anselm Steiger, Stefan Tilg, and Friedrich Vollhardt (2022). *Verfasserlexikon – Frühe Neuzeit in Deutschland 1620-1720*. Berlin: de Gruyter. [10.1515/vdbo](https://doi.org/10.1515/vdbo).
- Arnold, Heinz Ludwig, ed. (2020). *Kindlers Literatur Lexikon (KLL)*. Stuttgart: J.B. Metzler.
- Barth, Florian, Tillmann Dönicke, Benjamin Gittel, Luisa Gödeke, Anna Mareike Weimer, Anke Holler, Caroline Sporleder, and Hanna Varachkina (2021). *MONACO: Modes of Narration and Attribution Corpus*. <https://gitlab.gwdg.de/mona/korpus-public>.
- Barth, Florian, Hanna Varachkina, Tillmann Dönicke, and Luisa Gödeke (2022). "Levels of Non-Fictionality in Fictional Texts". In: *Proceedings of ISA-18 Workshop at LREC2022*, pages 27–32 Marseille, 20 June 2022.
- Beebe, Maurice (1976). "Reflective and Reflexive Trends in Modern Fiction". In: *The Bucknell Review* 22 (2), 83–94.
- Bonheim, Helmut (1975). "Theory of Narrative Modes". In: *Semiotica* 14.4, 329–344.
- (1982). *The Narrative Modes: Techniques of the Short Story*. Cambridge.
- Brottrager, Judith, Annina Stahl, and Arda Arslan (2021). "Predicting Canonization: Comparing Canonization Scores Based on Text-Extrinsic and -Intrinsic Features". In: *CEUR Workshop Proceedings*, 195–205. https://ceur-ws.org/Vol-2989/short_paper21.pdf.
- Chan, Branden, Stefan Schweter, and Timo Möller (2020). "German's Next Language Model". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 6788–6796. [10.18653/v1/2020.coling-main.598](https://doi.org/10.18653/v1/2020.coling-main.598).
- Chatman, Seymour Benjamin (1980). *Story and Discourse: Narrative Structure in Fiction and Film*. Ithaca, NY: Cornell Univ. Press.
- Dancey, Christine P and John Reidy (2004). *Statistics Without Maths for Psychology: Using SPSS for Windows*. London: Pearson Education.
- Dawson, Paul (2016). "From Digressions to Intrusions: Authorial Commentary in the Novel". In: *Studies in the Novel* 48.2, 145–167.
- Dönicke, Tillmann (2020). "Clause-level Tense, Mood, Voice and Modality Tagging for German". In: *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, 1–17.
- Dönicke, Tillmann, Florian Barth, Hanna Varachkina, and Caroline Sporleder (Dec. 2022). "MONAPipe: Modes of Narration and Attribution Pipeline for German Computational Literary Studies and Language Analysis in SpaCy". In: *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*. Potsdam, Germany: KONVENS 2022 Organizers, 8–15. <https://aclanthology.org/2022.konvens-1.2>.
- Ercolino, Stefano (2014). *The Novel-Essay, 1884 - 1947*. Studies in European Culture and History. New York: Palgrave Macmillan.
- Esselborn, Hartmut (2007). "Digression". In: *Reallexikon der deutschen Literaturwissenschaft: Neubearbeitung des Reallexikons der deutschen Literaturgeschichte*. Ed. by G. Braungart, H. Fricke, K. Grubmüller, J. D. Müller, F. Vollhardt, and K. Weimar. Vol. 1. Berlin, Boston: de Gruyter, 363–364.
- Fleiss, Joseph L (1971). "Measuring Nominal Scale Agreement Among Many Raters." In: *Psychological Bulletin* 76.5, 378–382.

- Fontane, Theodor (2013). *The Stechlin*. Trans. by William L. Zwiebel. Rochester, NY: Camden House.
- (2015[1897/98]). *Der Stechlin: Roman*. 3. Auflage. Vol. / herausgegeben in Zusammenarbeit mit dem Theodor-Fontane-Archiv ; editorische Betreuung Christine Hehle ; 17. Große Brandenburger Ausgabe Das erzählerische Werk. Berlin: Aufbau.
- Gittel, Benjamin (2015). "Essayismus als Fiktionalisierung von unsicheres Wissen prozessierender Reflexion". In: *Scientia Poetica* 19.1, 136–171. [10.1515/scipo-2015-0106](#).
- (2022). "Reflexive Passagen in fiktionaler Literatur. Überlegungen zu ihrer Identifikation und Funktion am Beispiel von Wielands ‚Geschichte des Agathon‘ und Goethes ‚Wahlverwandtschaften‘". In: *Euphorion* 116.2, 175–191.
- Gius, Evelyn, Svenja Guhr, and Benedikt Adelman (June 2021). *d-Prose 1870-1920*. Version 2.0. [10.5281/zenodo.5015008](#).
- Gödeke, Luisa, Florian Barth, Tillmann Döncke, Anna Mareike Weimer, Hanna Varachkina, Benjamin Gittel, Anke Holler, and Caroline Sporleder (2022). "Generalisierungen als literarisches Phänomen. Charakterisierung, Annotation und automatische Erkennung". In: *Zeitschrift für digitale Geisteswissenschaften* 7. https://zfdg.de/2022_010.
- Goethe, Johann Wolfgang (2021[1809]). *Die Wahlverwandtschaften. Ein Roman*. Ditzingen: Reclam.
- Goethe, Johann Wolfgang von (2008). *Elective Affinities: A Novel*. Trans. by David Constantine. Oxford: Oxford University Press.
- Handke, Peter (2008). *Die morawische Nacht: Erzählung*. 1. Aufl. Frankfurt am Main: Suhrkamp.
- (2016). *The Moravian Night: A Story*. Trans. by Krishna Winston. New York: Farrar, Straus and Giroux.
- Henke, Christoph (op. 2005). "Self-Reflexivity and Common Sense in A Tale of a Tub and Tristram Shandy: Eighteenth-Century Satire and the Novel". In: *Self-reflexivity in literature*. Ed. by Hubert Zapf, Werner Huber, and Martin Middeke. Text & Theorie. Würzburg: Königshausen & Neumann, 13–38.
- Herrmann, J. Berenike (2023). *digital resources* — jberenike.github.io. https://jberenike.github.io/dig_res.html. [Accessed 08-12-2023].
- Herweg, Mathias, Johannes Klaus Kipf, and Dirk Werle (2019). *Enzyklopädisches Erzählen und vormoderne Romanpoetik (1400-1700)*. Wolfenbütteler Forschungen. Wiesbaden: Harrassowitz.
- Jander, Simon (2008). *Die Poetisierung des Essays: Rudolf Kassner, Hugo von Hofmannsthal, Gottfried Benn*. Heidelberg: Winter.
- Julie Tanner (2022). "The legacy of literary reflexivity; or, the benefits of doubt". In: *Textual Practice* 36.10, 1712–1730. [10.1080/0950236X.2021.1972038](#).
- Just, Klaus Günther (1960). "Die Geschichte des Essays in der europäischen Literatur". In: *Anstöße. Berichte aus der evangelischen Akademie Hofgeismar* 3, 83–94.
- Kiesel, Helmuth (2004). *Geschichte der literarischen Moderne: Sprache, Ästhetik, Dichtung im zwanzigsten Jahrhundert*. München: C.H. Beck.
- Klauk, Tobias (2015). "Serious Speech Acts in Fictional Works". In: *Author and Narrator*. Ed. by Dorothee Birke and Tilmann Köppe. Berlin/Boston: de Gruyter, 187–222.
- Konrad, Eva-Maria (2014). *Dimensionen der Fiktionalität: Analyse eines Grundbegriffs der Literaturwissenschaft: Zugl.: Regensburg, Univ., Diss., 2013. Explicatio*. Münster: Mentis.

- Konrad, Eva-Maria (2017). "Signposts of Factuality: On Genuine Assertions in Fictional Literature". In: *Art and Belief*. Ed. by Ema Sullivan-Bissett, Helen Bradley, and Paul Noordhof. Oxford: Oxford University Press, 42–62.
- Kühlmann, Wilhelm, ed. (2012). *Killy Literaturlexikon Autoren und Werke des deutschsprachigen Kulturraums. Begründet von: Walther Killy*. Berlin, Boston: de Gruyter. 10.1515/9783110220292.
- Landis, J. Richard and Gary G. Koch (1977). "The Measurement of Observer Agreement for Categorical Data". In: *Biometrics* 33.1, 159–174.
- LimeSurvey, Limesurvey GmbH / (2023). *An Open Source Survey Tool*. Hamburg. <http://www.limesurvey.org> (visited on 07/24/2023).
- Mäkelä, Maria (2017). "The Gnomonic Space: Authorial Ethos Between Voices in Michael Cunningham's "By Nightfall"". In: *Narrative* 25.1, 113–137.
- Mann, Thomas (1969). *The Magic Mountain / Der Zauberberg*. Trans. by H. T. Lowe-Porter. New York: Vintage Books.
- [1924] (1991). *Der Zauberberg. Roman*. 24th ed. Frankfurt a. Main: Fischer.
- Martínez, Matías and Michael Scheffel (2007). *Einführung in die Erzähltheorie*. 7th ed. München: C.H. Beck.
- Martinez-Bonati, Félix and Philip W. Silver (1981). *Fictive Discourse and the Structures of Literature: A Phenomenological Approach*. Ithaca: Cornell Univ. Press.
- Mathet, Yann, Antoine Widlöcher, and Jean-Philippe Métivier (2015). "The Unified and Holistic Method Gamma (γ) for Inter-annotator Agreement Measure and Alignment". In: *Computational Linguistics* 41.3, 437–479.
- Müller-Funk, Wolfgang (1995). *Erfahrung und Experiment: Studien zu Theorie und Geschichte des Essayismus*. Berlin: Akademie.
- Neumann, Felix (2018). *German Prenames as CSV Data*. <https://github.com/fxnn/vornamen>. [Accessed 10-Jul-2023].
- Orr, Leonard (1981). "Vraisemblance and Alienation Techniques: The Basis for Reflexivity in Fiction". In: *The Journal of Narrative Technique* 11.3, 199–215. <http://www.jstor.org/stable/30225027> (visited on 08/02/2023).
- Prill, Meinhard (1994). "Über naive und sentimentalische Dichtung". In: *Hauptwerke der deutschen Literatur*. Ed. by Rudolf Radler. München: Kindler, 520–521.
- Reuvekamp, Silvia (2007). "Sentenz". In: *Reallexikon der deutschen Literaturwissenschaft: Neubearbeitung des Reallexikons der deutschen Literaturgeschichte*. Ed. by G. Braungart, H. Fricke, K. Grubmüller, J. D. Müller, F. Vollhardt, and K. Weimar. Berlin, Boston: de Gruyter, 425–427.
- Schiller, Friedrich (1985[1795]). "On Naive and Sentimental Poetry". In: *Winckelmann, Lessing, Hamann, Herder, Schiller, Goethe*. Ed. by Hugh Barr Nisbet. German aesthetic and literary criticism. Cambridge: Cambridge Univ. Press, 180–232.
- (2004[1795]). "Über naive und sentimentalische Dichtung". In: *Erzählungen - Theoretische Schriften*. Ed. by Peter-André Alt et al. Friedrich Schiller - Sämtliche Werke. Cambridge: Hanser, 694–780.
- Schomacker, Thorben, Tillmann Döncke, and Marina Tropmann-Frick (Sept. 2022). *Automatic Identification of Generalizing Passages In German Fictional Texts Using BERT With Monolingual and Multilingual Training Data*. Extended abstract submitted and accepted for the KONVENS 2022 Student Poster Session. 10.5281/zenodo.6979858.
- Searle, John (1975). "The Logical Status of Fictional Discourse". In: *New Literary History* 6.2, 319–332.

- Sokolova, Marina and Guy Lapalme (2009). "A Systematic Analysis of Performance Measures for Classification Tasks". In: *Information Processing & Management* 45.4, 427–437. [10.1016/j.ipm.2009.03.002](https://doi.org/10.1016/j.ipm.2009.03.002).
- Stanzel, Franz Karl (1988). *A Theory of Narrative*. Paperback ed. Cambridge: Cambridge Univ. Press.
- Tolstoy, Leo (2017). *Anna Karenina*. Trans. by Louise Maude and Aylmer Maude. London: Macmillan Collector's Library.
- Толстой, Лев Николаевич (1998[1878]). *Анна Каренина. Романъ*. Москва: Типография Рис.
- Vesper, Achim (2014). "Literatur und Aussagen über Allgemeines". In: *Wahrheit, Wissen und Erkenntnis in der Literatur*. Ed. by Christoph Demmerling and Ingrid Vendrell Ferran. Deutsche Zeitschrift für Philosophie / Sonderband. Berlin: de Gruyter, 181–196.
- Weimer, Anna Mareike, Florian Barth, Tillmann Dönicke, Luisa Gödeke, Hanna Varachkina, Anke Holler, Caroline Sporleder, and Benjamin Gittel (2022). "The (In-)Consistency of Literary Concepts. Operationalising, Annotating and Detecting Literary Comment". In: *Journal of Computational Literary Studies* 1.1. [10.48694/jcls.90](https://doi.org/10.48694/jcls.90).
- Williams, Jeffrey (1998). *Theory and the novel: Narrative reflexivity in the British tradition*. Cambridge: Cambridge University Press. ISBN: 0521120853.
- Winko, Simone (2002). "Literatur-Kanon als ‚invisible hand‘-Phänomen". In: *Literarische Kanonbildung*. Ed. by Heinz Ludwig Arnold and Hermann Korte. München: edition text + kritik, 9–24.
- You, Yang, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh (2020). "Large Batch Optimization For Deep Learning: Training BERT In 76 Minutes". In: <https://openreview.net/forum?id=Syx4wnEtvH> (visited on 12/16/2022).
- Zapf, Hubert, Werner Huber, and Martin Middeke, eds. (op. 2005). *Self-reflexivity in Literature*. Vol. Bd. 6. Text & Theorie. Würzburg: Königshausen & Neumann.
- Zeller, Rosmarie (2007). "Erzählerkommentar". In: *Reallexikon der deutschen Literaturwissenschaft: Neubearbeitung des Reallexikons der deutschen Literaturgeschichte*. Ed. by G. Braungart, H. Fricke, K. Grubmüller, J. D. Müller, F. Vollhardt, and K. Weimar. Berlin, Boston: de Gruyter, 505–506.