Effects of Complexity and Publicity in Reader Polarization

Yuri Bizzoni
Center for Humanities Computing
Aarhus University
yuri.bizzoni@cc.au.dk

Kristoffer L. Nielbo
Center for Humanities Computing
Aarhus University
kln@cas.au.dk

Pascale Feldkamp
Center for Humanities Computing
Aarhus University
pascale.feldkamp@cas.au.dk

Abstract

We investigate how Goodreads rating distributions reflect variations in audience reception across literary works. By examining a large-scale dataset of novels, we analyze whether metrics such as the entropy or standard deviation of rating distributions correlate with textual features – including perplexity, nominal ratio, and syntactic complexity. These metrics reveal a disagreement continuum: more complex texts – i.e., more cognitively demanding books, with a more canon-like textual profile – generate polarized reader responses, while mainstream works produce more uniform reactions. We compare evaluation patterns across canonical and non-canonical works, bestsellers, and prize-winners, finding that textual complexity drives rating polarization even when controlling for publicity effects. Our findings demonstrate that linguistically demanding texts, particularly those with higher nominal density and dependency distance, generate divergent reader evaluations. This challenges conventional literary success metrics and suggests that the shape of rating distributions offers valuable insights beyond average scores. We hope our approach establishes a productive framework for understanding how literary features influence reception and how disagreement metrics can enhance our understanding of public literary judgment. Code & data for this paper is available at: https://anonymous.4open.science/r/publicity_complexity_goodreads-873D

1 Introduction

Several computational literary studies estimate literary success using quantitative proxies such as reader evaluation (Koolen et al., 2020), sales data (Wang et al., 2019; Archer and Jockers, 2017), or number of prizes received (Bizzoni et al., 2023). These studies often default to Goodreads' within-platform metrics – such as the number of ratings or the average rating – since the Goodreads platform aggregates opinions from millions of diverse, lay readers, offering a democratic measure of literary judgment (Nakamura, 2013). However, while these metrics capture important aspects of popularity and appreciation, they typically focus on central tendencies. Our study proposes to advance knowledge on reader appreciation by examining the full distribution of ratings, rather than solely relying on the average. Specifically, by analyzing the distribution of ratings via rating entropy and standard deviation, we aim to refine our understanding of literary success, testing three interrelated hypotheses.

First, we hypothesize a positive relationship between rating count and rating distribution entropy (H1), suggesting that books with a higher number of ratings tend to exhibit a broader spread of opinions – a phenomenon we refer to as the "publicity effect", observed in other studies (Kovács and Sharkey, 2014; Maity et al., 2018).

Second, we posit that as a book attracts a more diverse or polarized audience, the relation between average rating and rating count will decouple (H2), resulting in little or no direct correlation between these two metrics. This decoupling implies that popularity (as measured by rating count) does not necessarily equate to higher average appreciation.

Third, and central to our contribution, is our hypothesis regarding textual complexity (H3). Prior studies have observed that highly complex texts tend to be less popular, attracting relatively fewer readers due to their demanding nature (Bizzoni et al., 2023). However, there are notable exceptions where complex texts, often deemed canonical, incite particularly polarized responses among those who do engage with them. This phenomenon may be very similar to the "publicity effect" – where

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 138–150
May 3-4, 2025 ©2025 Association for Computational Linguistics

Kovács and Sharkey (2014) suggest that the popular status of a book leads to more readers, including those not predisposed to like them – in the sense that canonical books, by their canonical status, will find more readers not predisposed to like them – especially if we consider reading assignments in educational settings. We hypothesize that books with more canonical, or more demanding, textual profiles will not only have fewer ratings overall but will also exhibit higher rating distribution entropy or variance. In this sense, a strong textual effect might emerge that runs counter to – or nuances – the "publicity effect". While the effect suggests that increased exposure leads to a wider range of opinions, the textual effect posits that inherent complexity can independently drive polarization, even in a smaller, more select readership.

Furthermore, we propose using rating distribution entropy as an alternative measure of literary judgment. This metric captures not only popularity or general preference but also the uncertainty or divergence in readers' evaluations. By investigating how this measure correlates with a suite of textual features connected to cognitively demanding textual profiles – such as perplexity, nominal ratio, and dependency distance – we seek to determine whether textual complexity itself plays a significant role in shaping reader disagreement. In doing so, our study endeavors to bridge the gap between traditional popularity metrics and nuanced literary analysis, ultimately providing a richer understanding of how textual characteristics influence reader reception.

2 Previous works

Goodreads' average rating has been employed in various studies as a proxy for reader *appreciation* (Maharjan et al., 2018b; Jannatus Saba et al., 2021; Bizzoni et al., 2024a), while rating count is often used to gauge the *popularity* of books (Veleski, 2020; Bizzoni et al., 2023). Prior research has examined aspects such as Goodreads' social function (Nakamura, 2013), its connection to offline literary culture (Walsh and Antoniak, 2021), and cross-platform metrics (Maity et al., 2018).

Several studies suggest that these within-platform metrics capture different forms of appreciation (Feldkamp et al., 2024; Kovács and Sharkey, 2014). For example, Kovács and Sharkey (2014) observed that winning a literary prize can lead to an increase in rating count alongside a decrease in average rating, possibly due to shifts in reader expectations. As such, while avg. rating and rating count usually exhibit a positive relationship (Feldkamp et al., 2024), increases in audience polarization may change the relationship between the two metrics. Similarly, Maity et al. (2018) demonstrated how Amazon bestsellers receive more ratings on Goodreads and have a higher entropy in their rating distributions, indicative of a more polarized audience. Here, we refer to the phenomenon where increased popularity coincides with heightened disagreement as the "publicity effect".

In addition, research into the relationship between textual features and reader responses has shown that books with more difficult or canonical textual profiles tend to be received in a more polarized manner (Bizzoni et al., 2023). Across different forms of appreciation too, canonical books tend to show a more diverse standing. For example, they often secure more literary prizes yet score lower on Goodreads and are less frequently held in libraries (Feldkamp et al., 2024). As studies consistently find that books associated with literary prestige display greater stylistic and syntactic complexity as well as higher information density (Brottrager et al., 2022; Algee-Hewitt et al., 2016; Wu et al., 2024), greater audience disagreement may be an effect of their textual complexity imposing a higher cognitive demand on the reader. For example, Bizzoni et al. (2023) indicates that more challenging novels in terms of readability tend to garner less favorable success on Goodreads.

3 Data

We used two datasets of literary novels for our analysis: a larger dataset with only metadata and a smaller curated one with access to full texts for the examination of textual features. We restricted our study to the novel (i.e., not considering poetry or short stories) to maximize the comparability of our datapoints.[1]

Goodreads Book Graph Dataset ($n = 809,297$). This dataset indexes the Goodreads data of approximately 2 million titles and was compiled in 2017.[2] We used the metadata (not including shelving and reader interaction) and reduced the dataset significantly by removing anything not

[1] Different literary forms may elicit other reading strategies (Blohm et al., 2022) and employ different communicative strategies (Obermeier et al., 2013).

[2] https://mengtingwan.github.io

tagged as literary, a novel, and by removing titles with less than 10 ratings.[3]

Curated Corpus ($n = 7,939$). To gauge the relation of Goodreads data to textual features, we used a corpus for which we had access to the full texts of novels – a subset of what is known as the *Chicago Corpus*. The corpus indexes 9,089 English-language novels of various genres, published in the US between 1880 and 2000 and covers 3,150 authors (see Table 2, and Bizzoni et al. (2024b) for details). It was compiled based on the number of libraries holding each title, with a preference for higher numbers.

	Mean & SD	**Sum**
Words	$119,776 \pm 65,076$	945,272,857
Rating count	$13,174 \pm 108,959$	104,585,264
Avg. rating	3.77 ± 0.34	

Table 1: Mean/SD and total of wordcount and Goodreads metrics in the curated corpus.

Subsets: To compare groups of novels, we create a *canon* subset. Generally, the *canon* group represents novels that appear in some canonicity indicator: either a novel has received a prestigious prize, is featured in the Norton anthology or Penguin Classics series, or is often assigned on literature syllabi.[4]

Category	**Titles**	**Authors**	**Titles/Author**
Full	7,939	2,909	2.73
Canon	591	223	2.65

Table 2: Overview of the curated corpus, including the number of titles, unique authors, and the average number of titles per author.

3.1 Methods

3.2 Goodreads metrics

From our two datasets, we got the avg. rating and rating count of the book listed on Goodreads, as well as the rating distribution for each title (i.e., how many voted 5, how many voted 3, etc.).[5] We computed the entropy and standard deviation (SD) of the rating distribution for each title. These two metrics reflect how diverse (i.e., entropic) and how varied (around the mean) the ratings received were.

3.3 Textual features

Computational research into literary preferences has indicated that reader appreciation or success can be somewhat predicted by stylistic elements (Koolen et al., 2020; van Cranenburgh and Bod, 2017; Maharjan et al., 2017), as well as by narrative features such as plot (Bizzoni et al., 2024a), emotional tone and flow (Maharjan et al., 2018a; Reagan et al., 2016; Veleski, 2020), or the predictability of a novel's sentiment arcs (Bizzoni et al., 2022). Additionally, factors external to the text, like genre, promotion, and the visibility or gender of the author, may also play a role (Wang et al., 2019; Koolen, 2018; Lassen et al., 2022).

For this condensed study, we chose to examine only intra-textual features that have been recently studied and found related to reader appreciation, canonicity, and cognitive load for readers (see Wu et al. (2024)). Our selection prioritizes features that previous research has demonstrated to be robust indicators of both literary complexity and reader engagement patterns. The features span multiple dimensions of textual analysis, from surface-level stylistic markers to deeper structural and cognitive elements that influence the reading experience. Specifically, we use: word length, sentence length, lexical richness via an overall type-token ratio (TTR), as well as the TTR of all verbs and nouns in a text, compressibility, word- and bigram entropy, readability, frequency of the word "of", the ratio of passive/active verbs, the nominal ratio, perplexity, and dependency distance.[6]

These features collectively capture different dimensions of literary complexity. Word and sentence length provide basic measures of textual den-

[3] We determined this number through sensitivity analysis showing that below 10 ratings, individual outlier ratings skew distribution metrics, with entropy calculations becoming unstable below this threshold.

[4] To tag the canon in our corpus, we follow Wu et al. (2024), using: 1) the Norton Anthology of English and American Literature, (Ragen, 1992), where, if the author was featured, all their titles were tagged *canon*. 2) OpenSyllabus, a resource collecting syllabi; where titles were tagged *canon* if their author featured in the top 1000 entries for English Literature syllabi; and 3) the Penguin Classics Series, where all *titles* featured in the series were tagged *canon* and 4) prizes, i.e., titles that were longlisted (win or nomination) for The Pulitzer Prize or the National Book Award were tagged *canon*.

[5] Note that the Goodreads data was obtained at different times: we used the data contained in the large *Goodreads Book Graphs dataset* (collected in 2017) and collected Goodreads data for the *Curated Corpus* in 2024.

[6] We calculate normalized the mean and SD in dependency length, following the method in Lei and Jockers (2020).

sity, while TTR assesses vocabulary diversity.[7] The compression ratio offers insight into a text's information redundancy, with less compressible texts generally containing more varied and unpredictable content.[8] Word and bigram entropy quantify lexical unpredictability at the local level, measuring how difficult it is to predict the next word or word pair in a sequence. It has been shown to be connected with canonicity (Algee-Hewitt et al., 2016).

The readability formula incorporates both syntactic complexity (sentence length) and vocabulary difficulty (percentage of uncommon words) to estimate cognitive demand.[9] Our syntactic measures extend beyond sentence length to examine specific structural characteristics: passive/active verb ratios and dependency distance capture sentence-level complexity (Bostian, 1983) – with higher levels associated with more canonical literature (Wu et al., 2024). The nominal ratio[10] and frequency of "of" represent aspects of nominal style – a writing approach associated with higher information density and abstraction (Wu et al., 2024; McIntosh, 1975; Bostian, 1983). Perplexity represents perhaps our most sophisticated complexity measure: it uses a large language model to quantify how surprising or unpredictable a text's language patterns are compared to general expectations (Wu et al., 2024).[11] Higher perplexity indicates prose that deviates more significantly from common patterns, requiring greater cognitive effort to process.

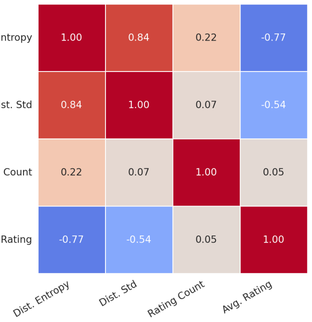


Figure 1: Heatmap of correlations (Spearman's ρ) of Goodreads metrics in the large *Goodreads Book Graph Dataset*. For all correlations $>= 0.1,\ p < .01$.

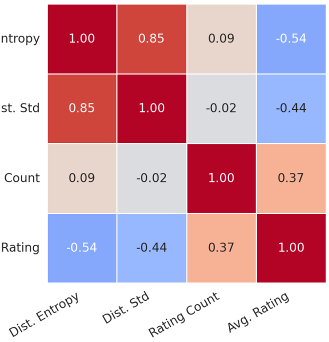


Figure 2: Heatmap of correlations (Spearman's ρ) of Goodreads metrics in *the curated corpus*. For all correlations $>= 0.1,\ p < .01$.

Collectively, these features allow us to examine multiple facets of literary complexity—from surface readability to deeper stylistic and structural characteristics—and their relationship to reader reception patterns. By analyzing correlations between these textual properties and Goodreads metrics, we can better understand how specific aspects of literary craft influence audience engagement, appreciation, and polarization.

4 Text-extrinsic relations

4.1 Relation between Goodreads metrics

We show the correlation between Goodreads metrics in **the large dataset** in Fig. 1. We do not find a correlation between rating count and avg. rating, suggesting that books that are popular in the sense that they are rated more often do not also receive a higher score. This supports H2, i.e., that the relationship between avg. rating and rating count decouples – perhaps as the audiences become more polarized due to a "publicity effect".

[7] For the overall TTR we use the Mean Segmental Type-Token Ratio (MSTTR) to gauge lexical richness. This splits the text into sequential chunks, usually a fixed set where a length of 100 words has been used as a standard (Torruella and Capsada, 2013), of which the mean TTR is then taken. For TTR within each of the two Parts-of-Speech categories, we use the mean TTR of the first 1500 sentences for each text.

[8] We use bzip2, a standard file compressor, to get a compression ratio (original bit-size/compressed bit-size) of texts. The ratio is not sensitive to length as we take only the first 1500 sentences of each text. This measures how compressible, i.e., redundant, a text is: the more a text tends to repeat sequences *ad verbatim*, the more compressible it will be (Benedetto et al., 2002; van Cranenburgh and Bod, 2017).

[9] We chose the *New Dale–Chall Readability Formula* among few different classic formulas that remain widely used (Stajner et al., 2012) – also seeing these formulas have been shown comparable for literary texts (Bizzoni et al., 2023). The formula is based on the average sentence length and the percentage of "difficult words", defined as words that do not appear on a list of words that 80% of fourth-graders would know (Dale and Chall, 1948).

[10] We here use a ratio of nouns + adjectives over verbs to gauge the nominality of the prose style, as in Wu et al. (2024).

[11] Perplexity is the predictability of the prose as indicated by the perplexity output of a large language model. Higher values indicate greater complexity or unpredictability. We use the specific GPT2 model trained by Wu et al. (2024), namely a model that has shown comparable results, but is exclusively trained on data which excludes works of the corpus that we use to apply it on.

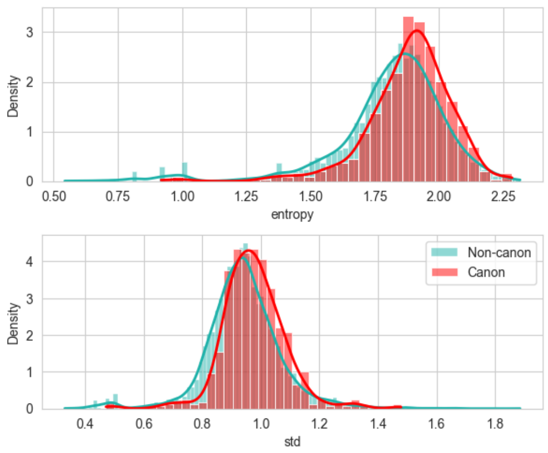


Figure 3: Distribution of titles by rating distribution metric – entropy & SD – per group (canon/non-canon).

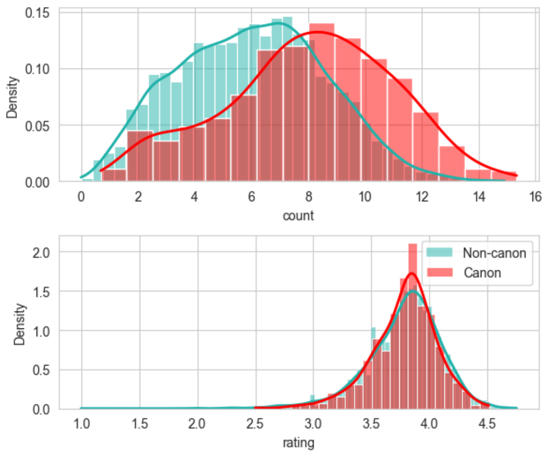


Figure 4: Distribution of average rating and rating log count across canon and non-canon groups. The rating count is log-transformed to account for its heavy-tailed distribution.

In fact, we do see a moderate correlation between the entropy of the rating distribution of books and the number of ratings (ρ .22). In other words, books that are rated more often – i.e., are more disseminated or popular – also have a higher diversity in the rating they receive, suggesting a larger but more uncertain audience, in support of H1. More rated books also tend to have a more uncertain reception, speaking for a "publicity effect". Moreover, we see that avg. rating has a robust negative correlation with rating distribution entropy (ρ $-.77$), suggesting that raters seem to agree more on high values (for the distribution of both entropy and avg. rating, see Appendix, Figs. 7-8).

For **the curated corpus** (Fig. 2), we see a similar correlation between rating distribution entropy and avg. rating (ρ $-.54$). However, we do not see a correlation – or a very weak one – between the entropy of rating distribution and rating count (ρ .09). This lack of correlation suggests that a "publicity effect" may not be as visible in a highly curated corpus, where all books may be above a certain threshold of popularity already.

Moreover, we see another discrepancy observed between the correlations of the large dataset and the curated corpus, namely that we here do see a correlation between rating count and avg. rating (ρ .37), suggesting that the amount of ratings given is often accompanied by higher scores.

4.2 Uncertainty & categories

When comparing different canon/non-canon groups of novels, we observe notable variations in rating distribution metrics. Canonical works consistently exhibit the highest levels of rating entropy and standard deviation, suggesting that these texts elicit the most polarized reactions (Fig. 3). Both a t-test and a Mann-Whitney Rank Test showed a significant difference ($p < 0.01$) between the groups in terms of rating distribution entropy and SD.

The canon group also exhibits an overall higher rating count, without this being followed by a higher avg. rating (Fig. 7). This canonical status effect bears similarity with the proposed "publicity effect" here, where higher ratings are connected with higher audience uncertainty for the canon (supporting H1) and where the relationship between rating count and avg. rating decouples (in support of H2). As such, while H1 – a positive relationship between rating count and rating entropy – is not confirmed in the curated corpus as a whole (Fig. 2), we do find that the canonical type of book is connected to this rating behavior.

Interestingly, within the curated corpus, canonical works also show a stronger correlation between textual complexity and reader *dis*agreement than non-canonical works. This implies that the reception of complex texts is shaped not only by their intrinsic features but also by their cultural positioning: canonical texts, often associated with prestige and social endorsement, may invite readers to approach them with heightened expectations or preconceptions, which can amplify the strength of their disappointment (see Fig. 6).

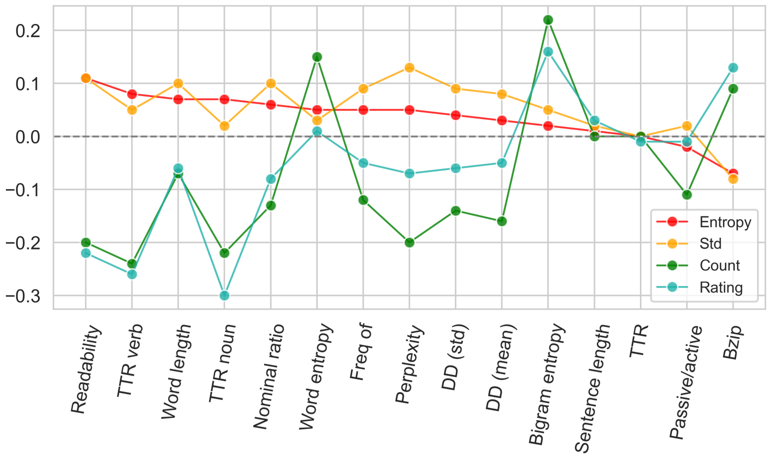


Figure 5: Spearman correlations between textual features and Goodreads metrics in the curated corpus. On the y-axis: the strength of the Spearman correlation between a Goodreads metric across features (x-axis). Note that the features have been ordered by the strength of the correlation with rating distribution entropy (descending). It does not reflect a linear development but aims to give a sense of how the distribution-based metrics – entropy and std (red, yellow) – coalesce with the count-based Goodreads metrics – rating count and avg. rating (green, blue). For the exact correlation strength of features with Goodreads metrics, see Appendix A, Figs. 10-11.

5 Text-intrinsic relations

Our analysis reveals a complex interplay between intrinsic textual features and reader responses as captured by Goodreads metrics. In particular, we find that measures of stylistic and syntactic complexity are strongly associated with the variability of readers' evaluations, thereby offering insight into the underlying cognitive and interpretive processes [12] involved in literary appreciation. We highlight some of the relationships between text complexity and varied reception observed in Figs. 5- 6.

Note that we might expect a diachronic change here, i.e., older books could be more challenging for modern readers and language models, potentially affecting human scoring and perplexity computed by LMs. We checked for a difference in the correlations by comparing the full corpus to a smaller set of the last 50 years of the corpus (1950-2000, $n = 5,591$). The correlations between features of textual complexity and Goodreads metrics remain similar in both sets (full and recent set), i.e., correlations observed in the full set either remain or increase in the set of more recent novels. Perplexity even shows an increase in its correlation with rating distribution entropy and SD, so we might assume that a recency bias of the model does not significantly impact our results. Results of the more recent subset of novels can be found in Appendix A (Figure 12).

5.1 Role of perplexity

Among the features examined, **perplexity** stands out as a particularly salient indicator. As a metric derived from language models, perplexity quantifies the unpredictability or complexity of a text (Wu et al., 2024). Higher perplexity scores indicate that a text is less predictable, often due to richer vocabulary, more intricate syntax, or unconventional narrative structures. Our results show that higher perplexity is correlated with increased SD ($\rho = .13$) in rating distributions. This suggests that when readers encounter texts that challenge their expectations, they tend to form more divergent opinions. In canonical works this correlation is even more pronounced, with a correlation between perplexity and SD ($\rho = .26$), and perplexity and entropy ($\rho = .19$), pointing to a potential cognitive load effect where complex texts elicit a wider range of interpretations and, consequently, more polarized ratings.

5.2 Role of nominality

In addition to perplexity, other textual features also contribute significantly to audience disagreement. The **nominal ratio** – which reflects the prevalence of nouns and adjectives relative to verbs – serves

[12] In the rest of the paper, we use 'interpretive effort' or 'interpretive strategies' in the basic cognitive sense of mental processing required to comprehend linguistic structures, not in the literary-critical sense of subjective meaning-making. This refers specifically to the cognitive load of unpacking syntactic and semantic relationships rather than higher-order interpretive activities.

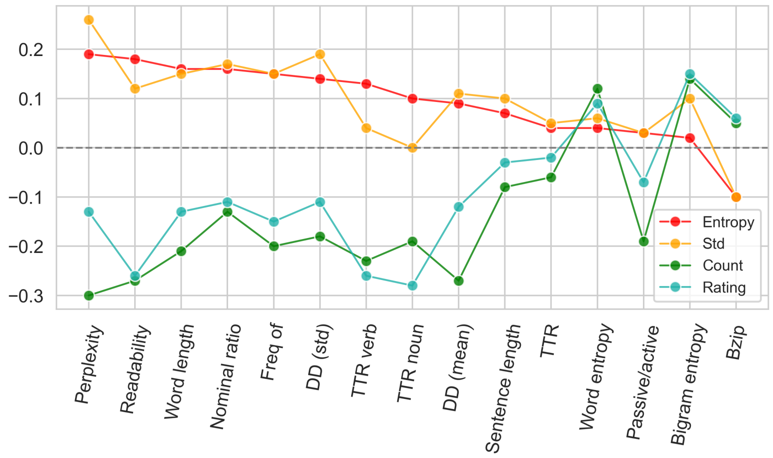


Figure 6: Spearman correlations between textual features and Goodreads metrics in the **canon subset** ($n = 591$). On the y-axis: the strength of the Spearman correlation between a Goodreads metric across features (x-axis). Features have been ordered by the strength of the correlation with rating distribution entropy (descending).

as a proxy for the degree of nominalization in a text. A higher nominal ratio, often associated with denser prose (McIntosh, 1975; Wu et al., 2024), appears to amplify rating variability: nominal ratio correlates with SD ($\rho = .1$ overall, and $\rho = .16$ in the canon set) and entropy of the rating distribution ($\rho = .1$ overall, and $\rho = .17$ in the canon set). This is likely because such texts demand greater interpretive effort, causing some readers to appreciate the prose while others may find the prose opaque or overly challenging.

This is further supported by the observation that the **frequency of the function word "of"**, also shows a correlation with increased polarization among readers, particularly in the canon subset ($\rho = .15/.15$ for SD and entropy). The frequency of the word "of" is tied to nominal constructions, creating dense informational structures that compress multiple concepts into compact syntactic units. As such, the cognitive challenge of unpacking such compressed prose creates divergent experiences.

5.3 Readability and dependency distance

Additionally, **readability** is a case in which metrics on either side – the standard Goodreads metrics, avg. rating and rating count, as well as our derived SD and entropy – show the strongest correlations (Fig. 5). While negatively correlated with popularity (i.e., lower rating counts and average ratings for more complex texts, $\rho = -.2$), readability also shows a nuanced relationship with rating distribution entropy: more complex texts attract a smaller readership, yet the opinions of those who do engage with them are increasingly uneven with reading difficulty. Both SD and entropy correlate with readability ($\rho = .11/.11$) – an effect that for entropy becomes even stronger in the canon subset ($\rho = .18$). Similarly, **dependency distance** shows stronger correlations with rating variability within the canon subset ($\rho = .19$). Longer dependency distances suggest more complex sentence structures, which again might lead to divergent reader responses depending on individual cognitive and interpretive capacities.

5.4 Comparative insights from canonical vs full corpus

When comparing the full curated corpus to the canonical subset, we observe that the correlations between textual features and rating distribution metrics tend to either remain or become stronger in the canonical subset. For example, features such as word length, readability, nominal ratio, and perplexity exhibit more robust associations with both the entropy and SD of ratings among canonical works. This suggests that while our so-called "publicity effect" implies that broader exposure leads to more varied opinions, the intrinsic qualities of the text itself can independently drive polarization. In canonical literature, where texts are generally at a more challenging level (Wu et al., 2024), this effect is even more salient, implying that a **textual effect** might be at work – a counterpoint to the general "publicity effect" observed across the bigger dataset (Fig. 1).

5.5 Implications for literary judgment

These findings underscore the idea that literary complexity does not merely influence the volume of ratings (i.e., popularity) but also shapes the nature of reader responses. High-complexity texts, as evidenced by higher perplexity and related metrics, seem to generate greater disagreement among readers. This divergence in opinion may reflect the varied interpretive strategies and differing cognitive loads experienced by readers. In platforms like Goodreads, where a heterogeneous audience converges, such textual features help explain why canonical works might be both less popular and more polarizing – highlighting the dual effect of text complexity to tend toward small or niche audiences as well as divided reception. Generally, our study highlights that capturing the polarizing effect of literary complexity requires moving beyond aggregate metrics like average ratings or raw counts, instead considering measures that reflect disagreement, dispersion, or interpretive diversity in reception.

6 Discussion & conclusion

Evidence of the publicity effect and rating patterns

Our analysis reveals important relationships between Goodreads metrics, audience reception patterns, and textual features, showing how different dimensions of literary appreciation interact. At a large scale (Book Graph Dataset), we observe the “publicity effect” suggested in previous studies (Kovács and Sharkey, 2014; Maity et al., 2018), confirming our hypothesis (H1): Books with higher rating counts consistently demonstrate more diverse audience opinions, as measured by increased entropy in their rating distributions (Fig. 1). In other words, books with greater visibility encounter more heterogeneous evaluation. The lack of correlation between average rating and rating count in the large dataset confirms our second hypothesis (H2), indicating that books with higher visibility don’t necessarily receive higher average scores. This decoupling suggests that popularity and appreciation represent distinct dimensions of literary reception. Still, this pattern shifts in the smaller, curated corpus, where we observe a positive correlation between rating count and avg. rating, as well as a slighter correlation between rating count and avg. rating, likely reflecting the already-established status of works in this more curated corpus.

Canonicity and rating polarization

When comparing literary categories, canonical works exhibit the highest rating distribution entropy, receiving more ratings (Fig. 4) but generating polarized responses (Fig. 3). This polarization reflects the dual nature of canonical reception: these works are both cultural artifacts worthy of respect (with a higher rating count) and personal reading experiences subject to individual taste. This tension contributes to the uneven distribution of ratings for the canonical subset, akin to a “publicity effect”. However, rather than being driven solely by visibility, this may also show a *canonicity effect*, which is not only driven by the cultural status of these works but also by their generally higher textual complexity, as shown in previous works (Wu et al., 2024; Bizzoni et al., 2024a; Brottrager et al., 2021).

Textual complexity and reader disagreement

Our analysis of textual features reveals relationships with rating patterns, confirming, in part, our third hypothesis (H3). Several markers of literary complexity show positive correlations with rating distribution entropy, particularly within the canon subset. Perplexity emerges as the strongest predictor of rating polarization (for entropy/SD, $\rho = 0.5/.13$ for the whole corpus, increasing to $\rho = .19/.26$ in the canon subset). This suggests that linguistic unpredictability contributes to varied reader responses. Nominal writing style, associated with perplexity (Wu et al., 2024), also correlates with rating entropy. This kind of prose, characterized by an informationally dense style, appears to divide reader opinions rather than diminish appreciation uniformly. Similarly, complexity measured by dependency distance and readability shows an increased correlation with rating entropy, especially in our canonical subset. More unreadable and complex sentence structures appear to generate more divergent responses among readers. Texts requiring a higher cognitive effort don’t simply receive lower ratings but provoke diverse evaluations.

Notably, some complexity markers, such as passive/active verb ratio (linked to lower reading speed (Bostian, 1983)), impact average rating and popularity without increasing rating dispersion. This suggests that certain textual features function as *bottlenecks*, limiting general appreciation without necessarily provoking more polarized reception.

Theoretical and practical implications

Rather than viewing complexity as merely a barrier to appreciation – which it is *not only* in most cases (pace passive/active ratio) – our findings suggest that complexity functions as a polarizing force, widening the spectrum of reader responses. This polarization may, in fact, constitute *a form of success in itself* for certain literary works/authors that aim to challenge readers or introduce innovative techniques. The relationship between complexity and polarization appears bidirectional: complex texts may generate diverse experiences due to their cognitive demands, while books positioned as complex or canonical may attract readers with varied motivations – from reading assignments to aspirational reading – leading to divergent evaluations. For publishers, authors, and literary platforms, these findings carry practical implications: rating distribution entropy provides valuable insights beyond average scores, potentially indicating a work's capacity to generate meaningful engagement and discussion. Highly complex works could expect more polarized reception, which doesn't necessarily indicate failure, but rather a different mode of success. Additionally, the relationship between textual features and reception patterns suggests opportunities for more nuanced recommendation systems that consider not just predicted ratings, but also the likelihood of polarized reception.

Future research directions

In the future, we intend to expand our analysis to include metrics beyond Goodreads, as well as datasets encompassing different literary genres and linguistic traditions. Longitudinal analyses tracking how ratings evolve would also provide an important dimension of publicity effects and readers' interaction with complexity. Additionally, incorporating reader demographic information could help disentangle the multiple factors contributing to rating polarization.

7 Limitations

This study has several limitations. First, our analysis is constrained by the availability of full texts, leading to a focus predominantly on anglophone literature, particularly by male authors, which is limited to novels. This bias may affect the generalizability of our findings, especially when considering the relationship between reception polarization and textual features in other genres like poetry, where the level and effect of perceived reading complexity may differ significantly.

Second, canonicity is inherently vague and open to interpretation. Our canon definition and our binary classification of canonical works may oversimplify a concept that may be better represented as a continuous variable (Brottrager et al., 2022). With a more nuanced canonicity measure – such as a 0-1 scale – we might be able to better understand how canonicity related to publicity effects and how feature levels of works above a certain threshold of textual complexity (where we here considered our canonical works to place) relates to audience polarization.

Additionally, Goodreads, initially a platform predominantly of anglophone users, does not represent the global reader base, further influencing the generality of our results.

Finally, while we focused on Goodreads metrics, other textual and extra-textual features likely play significant roles in shaping reader appreciation and should be explored in future work. Specifically, extra-textual factors, such as author and reviewer gender, are known to impact rating behavior (Lassen et al., 2022) and were not directly addressed in our analysis.

References

Mark Algee-Hewitt, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser. 2016. *Canon/Archive. Large-scale Dynamics in the Literary Field*. Stanford Literary Lab.

Jodie Archer and Matthew Lee Jockers. 2017. *The Bestseller Code*. Penguin books, London.

Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto. 2002. Language Trees and Zipping. *Physical Review Letters*, 88(4):1–5.

Yuri Bizzoni, Pascale Feldkamp, Ida Marie Lassen, Mia Jacobsen, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2024a. Good Books are Complex Matters: Gauging Complexity Profiles Across Diverse Categories of Perceived Literary Quality. ArXiv:2404.04022 [cs].

Yuri Bizzoni, Pascale Moreira, Nicole Dwenger, Ida Lassen, Mads Thomsen, and Kristoffer Nielbo. 2023. Good reads and easy novels: Readability and literary quality in a corpus of US-published fiction. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 42–51, Tórshavn, Faroe Islands. University of Tartu Library.

Yuri Bizzoni, Pascale Feldkamp Moreira, Ida Marie S. Lassen, Mads Rosendahl Thomsen, and Kristoffer

Nielbo. 2024b. A matter of perspective: Building a multi-perspective annotated dataset for the study of literary quality. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 789–800, Torino, Italia. ELRA and ICCL.

Yuri Bizzoni, Telma Peura, Kristoffer Nielbo, and Mads Thomsen. 2022. 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*, pages 31–41, Taipei, Taiwan. Association for Computational Linguistics.

Stefan Blohm, Stefano Versace, Sanja Methner, Valentin Wagner, Matthias Schlesewsky, and Winfried Menninghaus. 2022. Reading Poetry and Prose: Eye Movements and Acoustic Evidence. *Discourse Processes*, 59(3):159–183.

Lloyd R. Bostian. 1983. How active, passive and nominal styles affect readability of science writing. *Journalism quarterly*, 60(4):635–670.

Judith Brottrager, Annina Stahl, and Arda Arslan. 2021. Predicting Canonization: Comparing Canonization Scores Based on Text-Extrinsic and -Intrinsic Features. In *CEUR Workshop Proceedings*, pages 195–205, Antwerp, Belgium. CEUR.

Judith Brottrager, Annina Stahl, Arda Arslan, Ulrik Brandes, and Thomas Weitin. 2022. Modeling and predicting literary reception. *Journal of Computational Literary Studies*, 1(1):1–27.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28.

Pascale Feldkamp, Yuri Bizzoni, Mads Thomsen, and Kristoffer Nielbo. 2024. Measuring Literary Quality. Proxies and Perspectives. *Journal of Computational Literary Studies*, 3(1).

Syeda Jannatus Saba, 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*, pages 38–51, Online. Association for Computational Linguistics.

Corina Koolen, Karina van Dalen-Oskam, Andreas van Cranenburgh, and Erica Nagelhout. 2020. Literary quality in the eye of the Dutch reader: The national reader survey. *Poetics*, 79:1–13.

Cornelia Wilhelmina Koolen. 2018. *Reading beyond the female: the relationship between perception of author gender and literary quality*. Number DS-2018-03 in ILLC dissertation series. Institute for Logic, Language and Computation, Universiteit van Amsterdam, Amsterdam.

Balázs Kovács and Amanda J Sharkey. 2014. The paradox of publicity. *Administrative Science Quarterly*, 1:1–33.

Ida Marie Schytt Lassen, Yuri Bizzoni, Telma Peura, Mads Rosendahl Thomsen, and Kristoffer Laigaard Nielbo. 2022. Reviewer Preferences and Gender Disparities in Aesthetic Judgments. In *CEUR Workshop Proceedings*, pages 280–290, Antwerp, Belgium.

Lei Lei and Matthew L. Jockers. 2020. Normalized Dependency Distance: Proposing a New Measure. *Journal of Quantitative Linguistics*. Publisher: Routledge.

Suraj Maharjan, John Arevalo, Manuel Montes, Fabio A. González, and Thamar 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 Linguistics: Volume 1, Long Papers*, pages 1217–1227, Valencia, Spain. Association for Computational Linguistics.

Suraj Maharjan, Sudipta Kar, Manuel Montes, Fabio A. González, and Thamar Solorio. 2018a. Letting emotions flow: Success prediction by modeling the flow of emotions 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*, pages 259–265, New Orleans, Louisiana. Association for Computational Linguistics.

Suraj Maharjan, Sudipta Kar, Manuel Montes, Fabio A. González, and Thamar Solorio. 2018b. Letting Emotions Flow: Success Prediction by Modeling the Flow of Emotions 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)*, pages 259–265, New Orleans, Louisiana. Association for Computational Linguistics.

Suman Kalyan Maity, Abhishek Panigrahi, and Animesh Mukherjee. 2018. Analyzing Social Book Reading Behavior on Goodreads and how it predicts Amazon Best Sellers. ArXiv:1809.07354 [cs].

Carey McIntosh. 1975. Quantities of qualities: Nominal style and the novel. *Studies in Eighteenth-Century Culture*, 4(1):139–153.

Lisa Nakamura. 2013. "Words with friends": Socially networked reading on Goodreads. *PMLA*, 128(1):238–243.

Christian Obermeier, Winfried Menninghaus, Martin Von Koppenfels, Tim Raettig, Maren Schmidt-Kassow, Sascha Otterbein, and Sonja A. Kotz. 2013. Aesthetic and Emotional Effects of Meter and Rhyme in Poetry. *Frontiers in Psychology*, 4.

Brian Abel Ragen. 1992. An Uncanonical Classic: The Politics of the "Norton Anthology". *Christianity and Literature*, 41(4):471–479. Publisher: Sage Publications, Ltd.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):1–12.

Sanja Stajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity? In *Proceedings of Workshop on natural language processing for improving textual accessibility*, pages 14–22, Istanbul, Turkey. Association for Computational Linguistics.

Joan Torruella and Ramon Capsada. 2013. Lexical statistics and tipological structures: A measure of lexical richness. *Procedia - Social and Behavioral Sciences*, 95:447–454.

Andreas van Cranenburgh and Rens Bod. 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*, pages 1228–1238, Valencia, Spain. Association for Computational Linguistics.

Stefan Veleski. 2020. Weak negative correlation between the present day popularity and the mean emotional valence of late victorian novels. In *Workshop on Computational Humanities Research (CHR)*, pages 32–43. CEUR Workshop Procceedings.

Melanie Walsh and Maria Antoniak. 2021. The goodreads 'classics': A computational study of readers, amazon, and crowdsourced amateur criticism. *Journal of Cultural Analytics*, 4:243–287.

Xindi Wang, Burcu Yucesoy, Onur Varol, Tina Eliassi-Rad, and Albert-László Barabási. 2019. Success in books: Predicting book sales before publication. *EPJ Data Science*, 8(1):31.

Yaru Wu, Yuri Bizzoni, Pascale Moreira, and Kristoffer Nielbo. 2024. Perplexing canon: A study on GPT-based perplexity of canonical and non-canonical literary works. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 172–184, St. Julians, Malta. Association for Computational Linguistics.

A Appendix

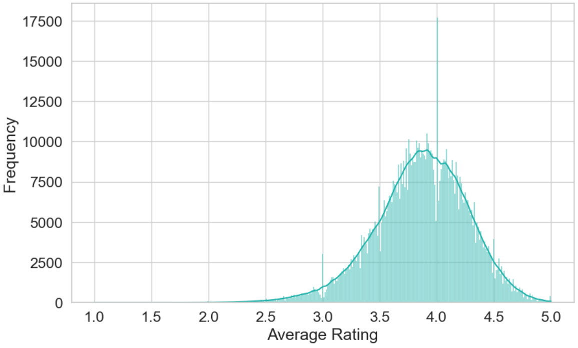


Figure 7: Distribution of avg. rating in the Goodreads Book Graph Dataset.

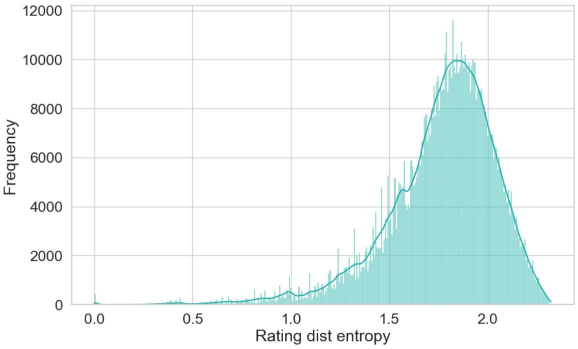


Figure 8: Distribution of entropy in the Goodreads Book Graph Dataset.

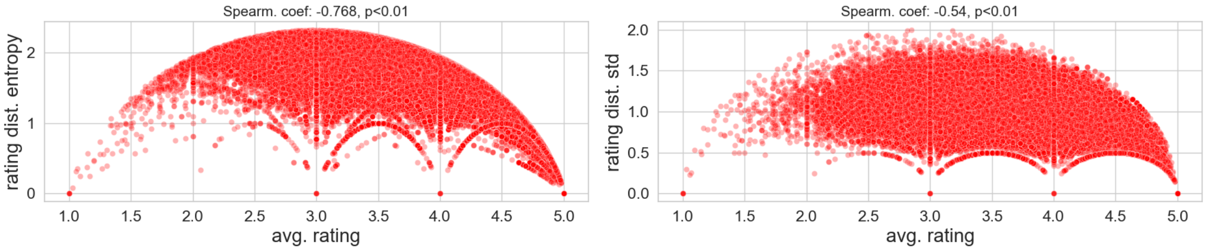
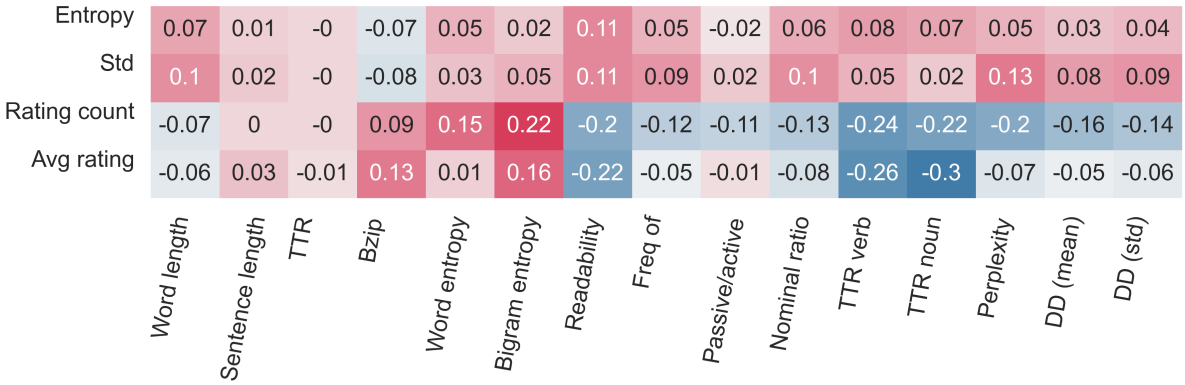
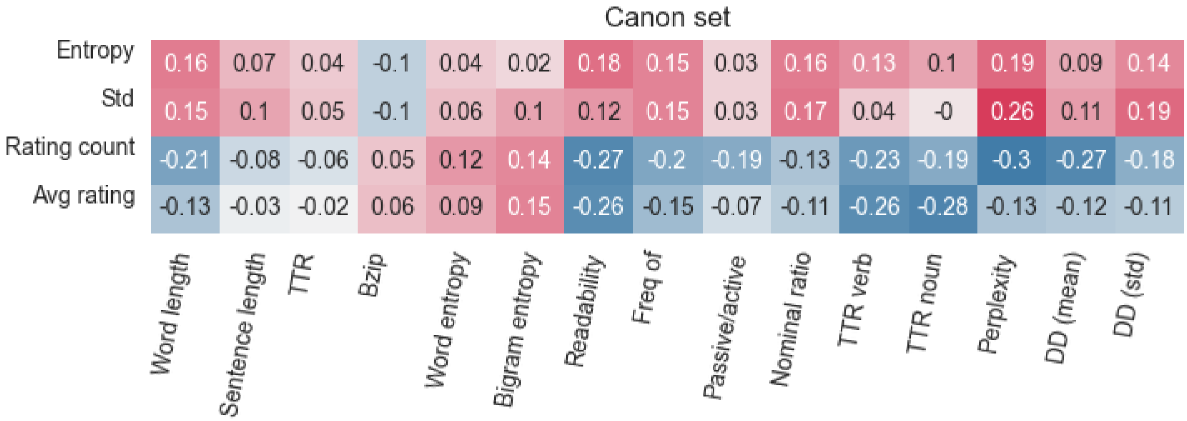


Figure 9: The Relation between Goodreads avg. rating and Rating Distribution Entropy and SD in the Goodreads Book Graph Dataset.



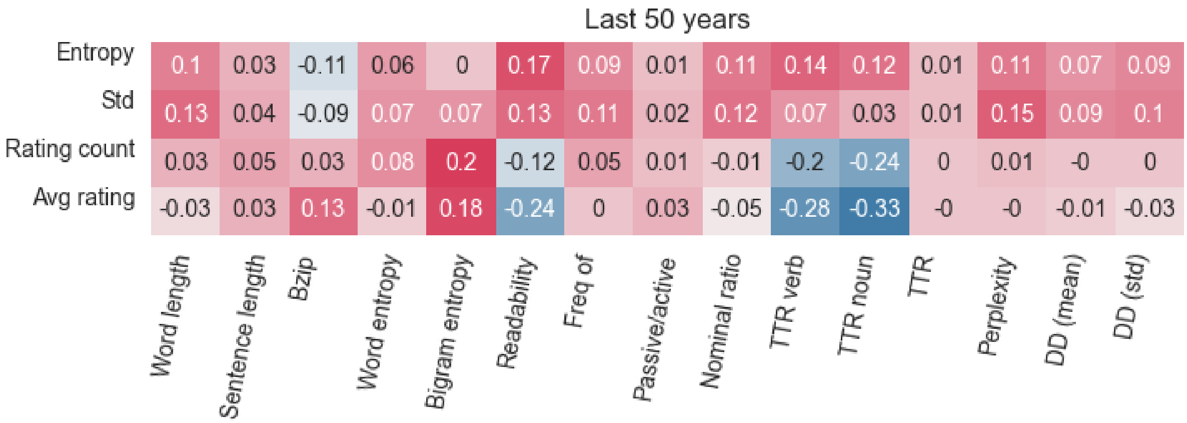
	Word length	Sentence length	TTR	Bzip	Word entropy	Bigram entropy	Readability	Freq of	Passive/active	Nominal ratio	TTR verb	TTR noun	Perplexity	DD (mean)	DD (std)
Entropy	0.07	0.01	-0	-0.07	0.05	0.02	0.11	0.05	-0.02	0.06	0.08	0.07	0.05	0.03	0.04
Std	0.1	0.02	-0	-0.08	0.03	0.05	0.11	0.09	0.02	0.1	0.05	0.02	0.13	0.08	0.09
Rating count	-0.07	0	-0	0.09	0.15	0.22	-0.2	-0.12	-0.11	-0.13	-0.24	-0.22	-0.2	-0.16	-0.14
Avg rating	-0.06	0.03	-0.01	0.13	0.01	0.16	-0.22	-0.05	-0.01	-0.08	-0.26	-0.3	-0.07	-0.05	-0.06

Figure 10: Spearman correlations between Goodreads metrics and textual features in the curated corpus ($n = 7,939$). For all $\rho > .1$, $p < .01$.



	Word length	Sentence length	TTR	Bzip	Word entropy	Bigram entropy	Readability	Freq of	Passive/active	Nominal ratio	TTR verb	TTR noun	Perplexity	DD (mean)	DD (std)
Entropy	0.16	0.07	0.04	-0.1	0.04	0.02	0.18	0.15	0.03	0.16	0.13	0.1	0.19	0.09	0.14
Std	0.15	0.1	0.05	-0.1	0.06	0.1	0.12	0.15	0.03	0.17	0.04	-0	0.26	0.11	0.19
Rating count	-0.21	-0.08	-0.06	0.05	0.12	0.14	-0.27	-0.2	-0.19	-0.13	-0.23	-0.19	-0.3	-0.27	-0.18
Avg rating	-0.13	-0.03	-0.02	0.06	0.09	0.15	-0.26	-0.15	-0.07	-0.11	-0.26	-0.28	-0.13	-0.12	-0.11

Figure 11: Spearman correlations between Goodreads metrics and textual features in the *canon subset* ($n = 591$). For all $\rho > .1$, $p < .01$.



Last 50 years

	Word length	Sentence length	Bzip	Word entropy	Bigram entropy	Readability	Freq of	Passive/active	Nominal ratio	TTR verb	TTR noun	TTR	Perplexity	DD (mean)	DD (std)
Entropy	0.1	0.03	-0.11	0.06	0	0.17	0.09	0.01	0.11	0.14	0.12	0.01	0.11	0.07	0.09
Std	0.13	0.04	-0.09	0.07	0.07	0.13	0.11	0.02	0.12	0.07	0.03	0.01	0.15	0.09	0.1
Rating count	0.03	0.05	0.03	0.08	0.2	-0.12	0.05	0.01	-0.01	-0.2	-0.24	0	0.01	-0	0
Avg rating	-0.03	0.03	0.13	-0.01	0.18	-0.24	0	0.03	-0.05	-0.28	-0.33	-0	-0	-0.01	-0.03

Figure 12: Spearman correlations between Goodreads metrics and textual features in *the last 50 years of the corpus, 1950-2000* ($n = 5,591$). Compared with the full set (Fig. 10), we see that correlations either persist or increase – for example *perplexity* – showing that the correlation with textual features does not seem to be an effect of modern readers reading (much) older texts. For all $\rho > .1$, $p < .01$.