Good Books are Complex Matters: Gauging Complexity Profiles Across Diverse Categories of Perceived Literary Quality

Anonymous ACL submission

Abstract

In this study, we employ a classification approach to show that different categories of literary "quality" display unique linguistic profiles, leveraging a corpus that encompasses titles from the Norton Anthology, Penguin Classics series, and the Open Syllabus project, contrasted against contemporary bestsellers, Nobel prize winners and recipients of prestigious literary awards. Our analysis reveals that canonical and so called high-brow texts exhibit distinct textual features when compared to other quality categories such as bestsellers and popular titles as well as to control groups, likely responding to distinct (but not mutually exclusive) models of quality. We apply a classic machine learning approach, namely Random Forest, to distinguish quality novels from "control groups", achieving up to 77% F1 scores in differentiating between the categories. We find that quality category tend to be easier to distinguish from control groups than from other quality categories, suggesting than literary quality features might be distinguishable but shared through quality proxies.

1 Introduction

The definition of literary "quality" has long been a subject of debate among scholars, critics, and readers alike. Expert-based quality judgments, such as literary awards, are often set in contraposition to signs of popular appreciation, observed for example in what appears on bestseller lists or has high ratings online (Algee-Hewitt et al., 2016; Porter, 2018; Underwood and Sellers, 2016). An often discussed dimension of literary quality is that of the so-called "literary canon", a complex concept generally denoting a set of works that have survived in the memory of a literary culture (Bloom, 1995). As a collective process of cultural selection, no one individual authority bestows (and can point to the features of) canonicity, which makes the very definition of the canon complex. Canonical literature can be considered a mid-way entity: it is the result of the fine-grained selections of large amounts of people over time, but it is also "curated", disseminated and validated by literary elites (Shesgreen, 2009). Some schools of literary scholars – most notably one side of the "canon-wars" of the canonicity-debate of the 1980s – have held the canon to represent nothing but entrenched interests (von Hallberg, 1983), or the cultural capital of current ruling classes (Guillory, 1995), while others have maintained that "canonic" works excel in terms of some set of intrinsic textual features, though vaguely defined (Bloom, 1995; van Peer, 2008). The quantifiable characteristics that distinguish canonic from non-canonic works, but also from other categories of "literary quality", like bestsellers or prestigious award-winning books, if any, remain elusive, or are framed in vague and undefined terms (*powerful prose*, *great humor*, *smooth development*, etc.). This study seeks to bridge this gap by employing computational techniques to explore the linguistic profiles that differentiate these nuanced categories of literary prestige.

While computational linguistics has made significant strides in text analysis, its application to literary studies has predominantly focused on authorship attribution or genre classification. Moreover, it has often revolved around modelling what can be broadly labelled stylistic features as bags-of-words (Da, 2019; Bode, 2023). There is a notable gap in research that utilizes these and more sophisticated sets of textual and narrative features to investigate literary quality.Specifically, the comparative analysis of "literary quality" as a mutlifaceted category – including canonical works, prestigious award-winning novels, and bestsellers against control groups – in terms of linguistic attributes has been underexplored.

In this work we leverage a large corpus that spans various categories of "quality", by including bags-

of-words stylistic measures as well as linear features of narrative complexity, to study the linguistic profiles of different categories of literary quality, providing empirical evidence to informs the ongoing discourse on literary prestige and merit.

2 Related works

There have been many rules and recommendations about how to write better, supposedly applicable across genres and to both high and low-brow literature, from detail-oriented suggestions about which parts of speech one ought to avoid to funny rituals inducive to writing. Sherman (1893) proposed that simplicity – i.e. shorter sentences – should be a marker of a "better" style. Readability indices have in this regard been thought to hint not only at the accessibility of a text, but implicitly at its "quality", and are widely implemented in more recent creative writing and publishing aids.[1]. Still, the importance of the readability of a literary text in the context of reader appreciation is controversial (Martin, 1996; Garthwaite, 2014). Studies seeking to predict literary success or perceived quality do, however, follow the intuitive idea that readers perceive a difference between "difficult" and "easy" fiction, tending to approximate some form of stylistic complexity by using textual features related to readability (i.a., sentence-length, vocabulary richness, redundancy)(Brottrager et al., 2022; van Cranenburgh and Bod, 2017; Crosbie et al., 2013; Koolen et al., 2020; Maharjan et al., 2017; Algee-Hewitt et al., 2016). In general, the focus on some form of **literary complexity** is not new in western culture. Some "simplicity laws" for literature have traditionally been set forth by critics and writers alike – for example, Hemingway's recommendation of a direct and personal style (Hemingway, 1999). A highly popular if not canonic author, King, advocates more readable texts King (2010); and Strunk et al. (1999)'s influential literary theory book, *The Elements of Style*, advised, i.a., using the active voice and avoiding redundancy. Conversely, others have promoted "purple prose",[2] characterized as a complex and challenging style, "rich, succulent and full of novelty" (West, 1985). However contradictory these positions may seem, both may hold merit for different ways of understanding literature. Regarding the "difficulty" of prose, at least in terms of readability, reader preference appears to be audience-specific (Bizzoni et al., 2023a). In examining the canon, computational literary studies have predominantly followed the same line of modelling stylistic features (Brottrager et al., 2022; Barré et al., 2023). Often, the profile of canonic works has been connected to some form of textual complexity, whether in the form of lower readability (Bizzoni et al., 2023a), textual entropy (Algee-Hewitt et al., 2016) or higher perplexity and cognitive demand on the reader (Bizzoni et al., 2023c). Moreover, features of style are seen to vary across "types" of literature: award-winning works are *less readable*, while *more readable* books appear to score higher on GoodReads (Bizzoni et al., 2023a). Similarly, more prestigious literature appears to elicit higher perplexity (i.e., LLM perplexity) than popular literature (Wu et al., 2024). Computational studies seeking to model reader appreciation and/or canonicity have predominantly focused on **the stylistic level**, modelling distributions of stylistic features in bag-of-words or bag-of-sentences approaches, ranging from the most basic measures of difficulty or complexity, such as sentence length (Maharjan et al., 2017; Mohseni et al., 2022), to more experimental measures like compressibility of a text file, aiming to identify stylistic signatures or markers of literary quality (Archer and Jockers, 2017; Koolen et al., 2020; Wang et al., 2019). Subsequent research expanded into sentiment analysis (SA), examining how emotional dynamics within a narrative – intensity, fluctuations, and trajectory – can influence reader perception and engagement. Much of this work has centered on tracing so-called sentiment arcs, i.e., time-series resulting from sequentially scored words or sentences with SA methods (Jockers, 2014). This focus on the emotional landscape of texts introduced a novel lens for understanding narrative techniques and their impact on the reader experience (Hogan, 2011; Cambria et al., 2017), with potential for moving beyond the stylistic level in modelling perceptions of literary quality (Pianzola et al., 2023). Still, questions persist as to how to operationalize an affective narratology (Rebora, 2023) – that is, for example, are sentiment arcs of novels as derived through SA tools actually palpable to readers? While most studies have focused on the visual shapes of sentiment arcs (Reagan et al., 2016; Jockers, 2015), others have applied more sophisticated measures to gauge their shapes

[1] Such as the Hemingway, or Marlowe applications

[2] A notion derived from Horace's *Ars Poetica*; in which "weighty openings and grand declarations" are said to "have one or two *purple patches* tacked on, that gleam / far and wide" (Horace, 2005).

and approximate complexity at **the narrative level** (Maharjan et al., 2018; Bizzoni et al., 2022), on the intuition that readers tend to appreciate certain shapes, or a certain balance in the complexity of narrative flow. Hu et al. (2020) and Bizzoni et al. (2022) have modeled the persistence, coherence, and predictability of arcs through measures like the Hurst coefficient and Approximate Entropy (ApEn) to measure global and local complexity (Bizzoni et al., 2023b). Such measures appear to be applicable for distinguishing between types of literary prestige (Bizzoni et al., 2021, 2023c).This perspective aligns with theories that emphasize the narrative's capacity to engage and challenge readers, proposing narrative or sentiment complexity as a key determinant of literary quality (Hu et al., 2020). Moreover, it draws on studies observing the role of fractal patterns or entropy for aesthetic attraction (Cordeiro et al., 2015; McGavin, 1997) also in other domains, such as in music or the visual arts (McDonough and Herczyński, 2023; Brachmann and Redies, 2017).

3 Methods

Drawing on the insights from literary theory and computational study on features and profiles of literary quality, we focus on narrative complexity in modelling the feature profiles of various categories of perceived literary quality. Our approach not only adopts a multi-level perspective on literary quality itself, but also on literary complexity, examining complexity at the stylistic and syntactic level (including simple features, such as vocabulary richness and deeper features, such as perplexity) and at the level of sentiment or narrative (including simple features, such as mean valence, and deeper features, such as sentiment arc entropy) across a diverse corpus of literary works.

3.1 Corpus

Our corpus comprises a carefully curated selection of 9,089 novels of various genres, published in the US between 1880 and 2000 (see Table 1 and Figure 1). It is a unique dataset both in terms of size [3] and diversity, as the corpus was compiled based on the number of libraries holding each novel, with a preference for more circulated works. Library holdings reflect a diverse demand, therefore the corpus is not homogeneous in terms of genre and lists both prestigious and popular works ranging from Nobel prize winners to Science Fiction classics (Long and Roland, 2016).[4]

Category	Titles	Authors	Titles/Author
All	9089	3166	2.87
Canon	618	163	3.80
Nobel	85	18	4.72
Prizes	144	108	1.33
Bestsellers	228	130	1.75
Rest	7955	2933	2.71

Table 1: Number of titles, authors, and average titles per author in the dataset and for each quality category. Note that "Rest" denotes titles that are included in neither quality category.

3.1.1 Quality categories

We divided titles into different categories of perceived quality (Table 1). We considered novels that bear some mark of perceived quality those that: (i) are canonic in the sense that the they often appear on college syllabi,[5] are included in the most prominent literary anthology,[6] or in a publisher's classics series;[7] (ii) are by Nobel prize-winning authors; (iii) have been long-listed for prestigious literary

[3]Often, studies on reader appreciation rely on $<$ 1,000 books (Ganjigunte Ashok et al., 2013; Koolen et al., 2020).

[4]The corpus has no reference publication, though other studies are based on it (Underwood et al., 2018; Cheng, 2020). See https://textual-optics-lab.uchicago.edu/us_novel_corpus for a corpus description.

[5]We relied on the OpenSyllabus database, which indexes 18.7 million college syllabi: https://www.opensyllabus.org; tallying all works in our collection by the top 1000 most frequent authors in *English Literature* syllabi.

[6]We used the English and American edition of *the Norton Anthology*, which is often referred to as indexing canon (**??**), marking all books by authors indexed.

[7]As one of the – if not *the* – most prominent classics series (Alter et al., 2022), we used the Penguin Classics, marking titles in the corpus that are also printed as part of the series.

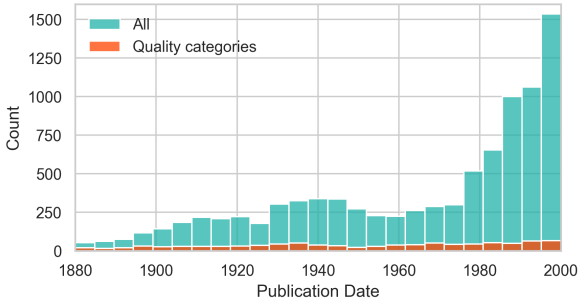

Figure 1: Distribution of titles in categories of perceived quality (Canon, Awards, Nobel, and Bestseller groups) in the Chicago Corpus over time.

awards;[8] (iv) are listed on bestseller lists of the 19th and 20th century.[9] The amount of works that fall inside one of these categories is relatively consistent across decades (Fig. 1).

It should be noted that we have sought to make the classification among what we refer to as different – though overlapping – markers of quality difficult. The novels in quality categories do not necessarily stand out in terms of stylistic and narrative quality from those not selected. For example, the corpus contains important works of genre-fiction (i.e., Tolkien or Philip K. Dick) as well as influential authors of popular fiction (such as Agatha Christie and Stephen King). It should be noted that the presence of such other classics and popular titles that do not fall within any of the mentioned categories increases the difficulty of a classifier's tasks. Naturally, we consider the division into categories an artificial, though necessary heuristic to make the study possible. In fact, canonicity is neither defined nor boolean (Barré et al., 2023), but may be best represented as a continuum on several dimensions. Similarly, the overlap of the categories should increase the difficulty of differentiation (see Fig. 2) as in numerous cases a text might be, for example, both canonical and a bestseller. In some sense we challenging the classifier to see whether these categories are representative of distinctive profiles - of which a novel can contain more than one, as it ends up in more than one category.

3.1.2 High/low GoodReads ratings

Beyond the categories of perceived quality, we collected the highest and lowest rated titles on GoodReads, a large online platform for rating and reviewing books. With its 90 million users, GoodReads arguably offers an insight into reading culture "in the wild", cataloguing books from a wide spectrum of genres (Nakamura, 2013). It derives book-ratings from a heterogeneous pool of readers in terms of background, gender, age, native language and reading preferences (Kousha et al., 2017). We distinguished classes at 3.8 average GoodReads rating,[10] where we consider high-rating titles those that are rated above (n=4680), and low-rating those that are rated below or equal to this threshold (n=4387).

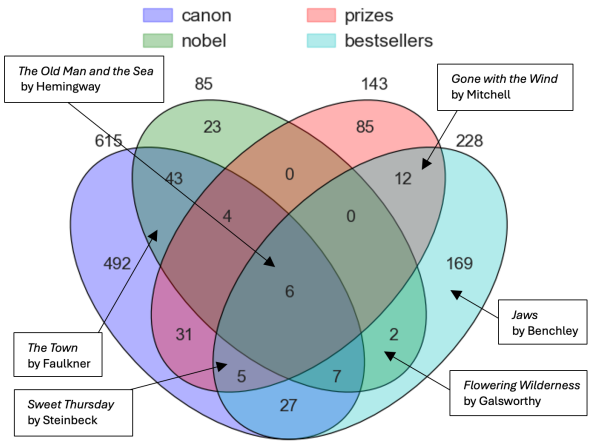

Figure 2: Number and overlap of the quality categories used in this study. The boxes give examples of titles contained in intersecting areas. Note that the largest overlap appears to be between the canon and prizes, indicating the close relation between the two. Still, in terms of percentages, the canon and Nobel categories show the largest overlap.

3.2 Features

To capture the complexity of the literary texts at various levels, we extracted a set of stylistic and narrative features that both approximate some form of complexity and have been known to influence perceptions of literary quality. A description of each feature including reference studies are listed in Table 2. We divide these features into **stylistic features** (with a subcategory of more syntactic features) and **narrative features**, where the former are surface features, calculated using a bag-of-words approach and the latter are higher-level features based on sentiment analysis, where the complexity measures Approximate entropy and the Hurst exponent take the progression of novels into account.

3.3 Model and Evaluation

We employed a "classic" machine learning model to classify novels based on the extracted features: Random Forest (Breiman, 2001). We chose the Random Forest for its robustness to overfitting and ability to handle nonlinear relationships. In each of the following experiments we configured the classifier with 900 trees and trained on 80% of the relevant subset. Model performance was assessed using the accuracy and F1 score, enabling a balanced evaluation of both false positives and false negatives. Additionally, we conducted a feature

[8] We marked all titles extant in the corpus that were long-listed for the Pulitzer Prize and the National Book Award.

[9] Contained in the Publisher's Weekly bestseller list or the New York Times bestseller list.

[10] The threshold is justified by its mid-scale position considering the general positive skew of ratings (see the distribution of ratings in the Appendix), and as we sought to have equally sized low and high rating categories.

Feature	Description	Type	Reference
Type-Token Ratio	Measures lexical diversity by comparing the variety of words (types) to the total number of words (tokens) in a text, indicating a text's vocabulary complexity and inner diversity (Torruella and Capsada, 2013).[a]	Stylistic	Forsyth (2000)*, Kao and Jurafsky (2012)*, Algee-Hewitt et al. (2016), Maharjan et al. (2017), Koolen et al. (2020), Brottrager et al. (2022), Jacobs and Kinder (2022), Bizzoni et al. (2023b)
Readability	Estimate reading difficulty based variously on sentence length, syllable count and word length/difficulty. Assessed using five different classic formulas that remain widely used (Stajner et al., 2012).[b]	Stylistic	Martin (1996), Garthwaite (2014), Maharjan et al. (2017), Febres and Jaffe (2017), Zedelius et al. (2019)*, Berger et al. (2021)*, Brottrager et al. (2022), Bizzoni et al. (2023a)
Compressibility	Measures the extent to which the text can be compressed, serving as an indirect indicator of redundancy and lexical variety (Ehret and Szmrecsanyi, 2016).[c]	Stylistic	van Cranenburgh and Bod (2017), Koolen et al. (2020), Bizzoni et al. (2023b)
Passive/active ratio	Quantifies the number of active against passive verbs in the text, associated to a better style (King, 2010).	Stylistic/ Syntactic	Hye-Knudsen et al. (2023), Wu et al. (2024)
Nominal style ratio	Quantifies the proportion of nouns and adverbs (over verbs) in the text, reflecting the nominal tendency in style, which is often associated with complex linguistic structures, denser communicative code, expert-to-expert communication (McIntosh, 1975; Bostian, 1983).	Stylistic/ Syntactic	Charney and Rayman (1989)*, Crossley et al. (2014)*, Wu et al. (2024)
"Of"/"that" frequencies	Frequency of these function words have been seen to indicate, in the case of "of", a more nominal prose, and in the case of "that", a more declarative and verb-centered prose. a more declarative or nominal style.	Stylistic/ Syntactic	Wu et al. (2024)
Perplexity	Represents the predictability of the prose through three different large language models (GPTs).[d] Higher values indicate greater complexity or unpredictability.	Stylistic/ Syntactic	Sheetz (2018), Wu (2023), Wu et al. (2024)
Mean valence	Represents the average sentiment of the text (positivity or negativity).[e]	Narrative/ Sentiment	Veleski (2020), Pianzola et al. (2020)*, Berger et al. (2021)*, Jacobs and Kinder (2022), Pianzola et al. (2023), Bizzoni et al. (2023b)
Valence std.	Represents the average variability in sentiment, indicating the range of sentiment within the narrative.[e]	Narrative/ Sentiment	Berger et al. (2021)*, Bizzoni et al. (2023b)
Hurst exponent	Quantifies the long-term auto-correlation of the sentiment arc,[e] with higher values suggesting a more complex, self-similar structure across different scales.[f]	Narrative/ Sentiment	Mohseni et al. (2021), Bizzoni et al. (2021), Bizzoni et al. (2023c)
Approximate entropy	Assesses the predictability of sequences of the sentiment arc,[e] with lower values indicating greater regularity or simplicity.[f]	Narrative/ Sentiment	Hu et al. (2020), Mohseni et al. (2022), Bizzoni et al. (2023b)

Table 2: **Features related to stylistic and narrative complexity**. "References" refer to studies that have included the given feature and shown some relation between the feature and reader appreciation, success, or canonicity. Note that this table only includes features chosen for this study. * Denotes studies on objects connected to cultural success, however in relevant domains other than *established prose fiction* (e.g., online stories, movies).

[a] We used a common method insensitive to text length: the Mean Segmental Type-Token Ratio (MSTTR). MSTTR-100 represents the overall average of the local averages of 100-word segments of each text.
[b] Flesch Reading Ease, Flesch-Kincaid Grade Level, SMOG Readability Formula, Automated Readability Index, and New Dale–Chall Readability Formula.
[c] We calculated the compression ratio (original bit-size/compressed bit-size) for the first 1500 sentences of each text using bzip2, a standard file-compressor.
[d] All perplexity calculations were via gpt2 models, done on the byte pair encoding tokenization used in the series of gpt2 models. To get the mean perplexity per novel, we used a sliding window due to maximum input length. For details on the computation, see Wu et al. (2024).
[e] All sentiment analysis was performed using nltk's VADER implementation on a sentence-basis (compound score per sentence). For complexity measures (Hurst and ApEn, we used both VADER and the widely used Syuzhet dictionary to extract the sentiment arcs on which these measures are based.
[f] For details on the measure, please refer to Bizzoni et al. (2023c).

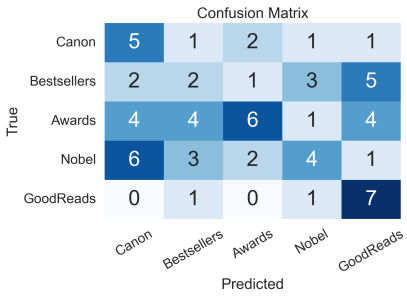

Figure 3: Confusion matrix of the multiclass experiment.

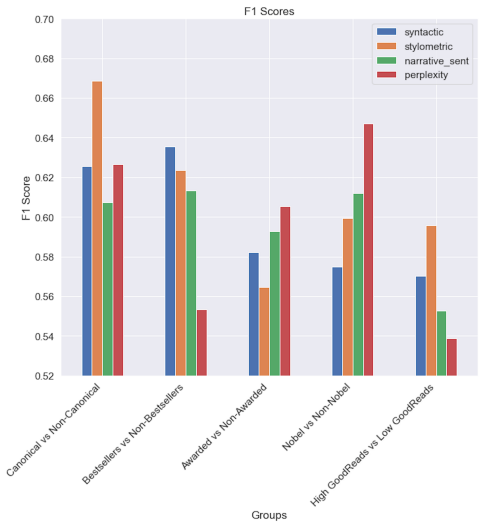

Figure 4: Performance for each category per features set in isolation.

ablation study to understand the impact of removing specific features (e.g., stylometric features, perplexity) on the classification accuracy, providing insights into the relative importance of different complexity measures in predicting literary quality.

4 Results

4.1 Performance

4.1.1 Sampling

We used random subsampling for balancing the dataset. To mitigate the risk of aleatory results, in the rest of the paper all reported results will be averaged over ten independent runs, each run training and testing on a new subset where the majority class was randomly subsampled. All classifications are run on balanced classes.

4.1.2 Binary classification

In binary classification tasks, we evaluated the performance of our models using different subsets of features, achieving balancing through repeated random subsampling. The variation in precision, recall, and F1 scores across different feature sets (see Fig. 4) indicates the differential predictive power of the features. The highest F1 score was achieved when all proposed features were included (Table 3), reinforcing the hypothesis that a multifaceted approach to textual analysis is crucial for accurate classification.

4.1.3 Multi-class

The results of the multi-class classification task are summarized in Fig 3. The matrix reveals the model's performance in classifying texts into the five categories: canonical works, awarded works, Nobel works, bestsellers, and high/low GoodReads ratings. Notably, the model demonstrates a strong ability to distinguish awarded texts, with a substantial number of true positives. However, there is some confusion between canonical works and bestsellers, indicating areas where the feature set may not fully capture the distinguishing characteristics between these two categories.

4.2 Feature impact

The analysis of feature impact demonstrates the intricate nature of literary quality and the necessity of a multilevel approach to textual analysis. Each feature contributes uniquely to the model's ability to discern among categories of literary quality, and the combined use of stylistic and narrative features enriches the classification process.

4.2.1 Stylistics

The so-called stylistic features alone, including TTR, compressibility and readability scores, had a noticeable impact on all models' performance, suggesting that this level of stylistic complexity - lexical diversity and a composition of sentence length and word complexity - is a significant marker of literary quality through most considered dimensions. This category is especially useful in distinguishing canonical novels and GoodReads higher-rated books from their relative control groups, and its absence brings the bestsellers classifier to its lowest performance. As we show in Fig. 5, bestsellers exhibit a higher TTR, suggesting a wider range of vocabulary usage compared to other textual categories like long-listed novels, which – perhaps surprisingly – do not display a high TTR. Still, canonical novels predominantly tend to have a systematically higher level of readability scores, characterizing a more complex language usage, while

	Canon/not	Awards/not	Nobel/not	Bestseller/not	High/low GR	Multiclass
N. samples	1236	288	170	456	8774	5755
F1	.77 (.02)	.7 (.03)	.76 (.07)	.7 (.05)	.63 (.009)	.36(.08)
Accuracy	.75 (.02)	.65 (.02)	.76 (.06)	.68 (.04)	.64 (.006)	.41 (.07)

Table 3: F1 score and accuracy per category and for the multiclass classification. Values in parenthesis are the standard deviation. Note that GR stands for GoodReads Rating.

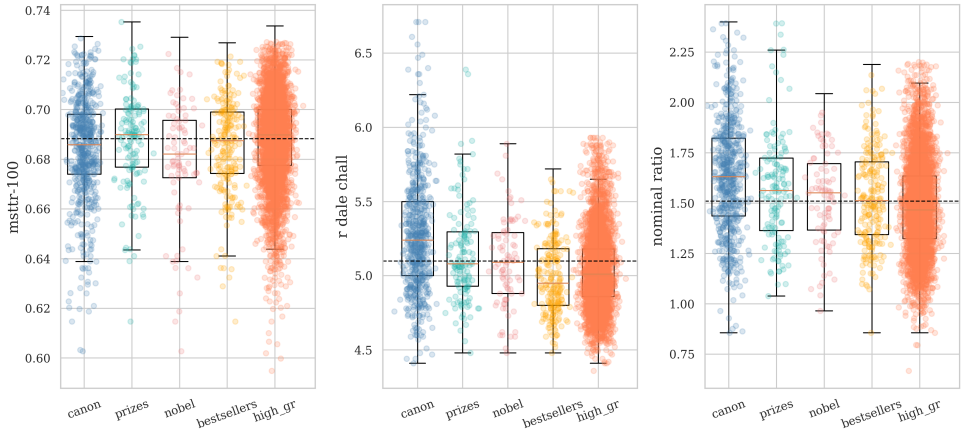

Figure 5: Boxplots indicating – from left to right – the levels of TTR, Readability, and Nominal Ratio per quality category. The black dashed line indicates the corpus mean value per feature.

bestsellers tend to have lower readability scores, reflecting simpler language and sentence structure, and both the award group and the Nobel group also show higher scores than the other categories. Overall, canonical texts appear to be the most demanding in terms of readability, in alignment with a previous study (Bizzoni et al., 2023a).

4.2.2 Syntactic features

The syntactic features we selected appear very important on their own – especially in differentiating between bestsellers and non-bestsellers (Fig. 4) – and their absence harms the performance of the classifier for the awards, the bestsellers and the GoodReads categories (Tab. 4).When combined with other features, they still indicate the importance of syntactic complexity also in distinguishing canonic and non-canonic literature.

4.2.3 Perplexity

Perplexity, as a measure of predictability of text, shows a strong impact on all classifications, and especially on the differentiation of canonical, award and Nobel groups from control-groups (Fig. 4). Perplexity appears lower than average in bestsellers and in the high GR rating group, suggesting a higher degree of predictability and simplicity in their language (Fig. 6). In contrast, canonical novels and Nobel texts show the highest perplexities, alluding to more complex language usage that requires greater cognitive effort to process (Fig. 6). This finding aligns with another recent study (Wu et al., 2024), and together with the higher nominal style of canonic texts suggests that there is a particular "canonic profile" of works, which uses language less expectedly and manages to reach a particularly high information density. A similar mechanism seems to be at work for the Nobel texts.

4.2.4 Narrative features

The sentiment features' predictive power was improved the performance of at least some categories. The variability of sentiment (valence std.) seems more pronounced in two usually opposed metrics of "literary quality", canonical novels and bestsellers (Fig. 6). Canonical texts have a particularly high valence std., showcasing the ability to move frequently through a broader emotional range. The Hurst exponent is the highest for bestsellers (Fig. 6), suggesting a more self-similar and less complex narrative structure over various scales. Canonical, Nobel and long-listed texts, on the other hand, show Hurst exponents that are lower than average, indicating a higher complexity and less self-similarity in their narrative structures. While these features appear to pick on a weaker signal than the others

	Canon/not	Awards/not	Nobel/not	Bestseller/not	High/low GR
- Stylistic	.71	.67	.68	.68	.60
- Perplexity	.74	.63	.68	.69	.61
- Styl./Syntactic	.76	.64	.75	.65	.61
- Narrative/Sentiment	.75	.69	.71	.74	.63

Table 4: F1 score per category against control for the **ablation experiment**. Each row represents the features that were *removed* before performing the classification.

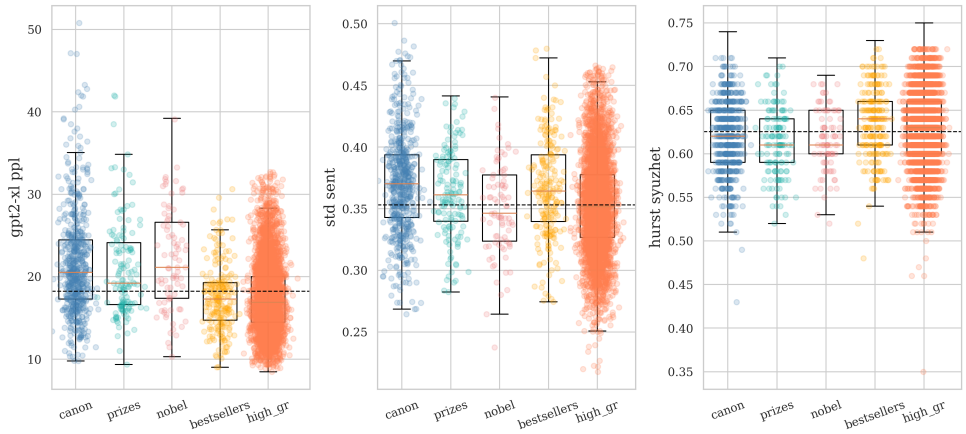

Figure 6: Boxplots indicating – from left to right – the levels of Perplexity, Valence Std., and Hurst exponent per quality category. The black dashed line indicates the corpus mean value per feature.

(Fig. 4), a decrease in performance is observed when they are removed from the feature set (compare with full results in Table 4), highlighting the importance of these high-level complexity metrics in capturing an aspect of the narrative structure that is not grasped by the other features.

5 Discussion & Conclusion

The results of this study elucidate the intricate relationship between textual complexity and perceived literary quality. Canonical works, bestsellers, Nobel laureates' and award-winning works, and high-rated novels on GoodReads each exhibit unique profiles with respect to the "control populations" represented by the rest of our corpus across various stylistic and narrative dimensions, and could be positioned on a multi-dimensional "complexity" continuum. At the same time, the difficulty of telling them apart in a multi-class classification experiment shows that they also represent overlapping profiles (partly explained by their de facto overlap, seen Fig. 2). We found canonical texts to have the most distinctive profile across all dimensions and to be the easiest to classify in the binary classification task. These have are more perplexing and have a denser nominal style and lower readability scores while maintaining a less predictable sentimental line. Such complexity, which is held to require greater cognitive effort (McIntosh, 1975), may be one contributing factor to the lasting impact and classification of these works as 'canonical'. It appears to be also partly shared by the long-listed novels and the books of Nobel laureates. In the multi-class classification, these three groups are easily confused with each other. On the other hand, bestsellers, characterized by a somewhat opposite profile, display an increased readability, lower perplexity, and a higher Hurst exponent. Together with the group of novels more highly praised on GoodReads, yet to a higher extent, they seem to employ a more accessible and predictable language, which could account for their mass appeal and commercial success. For these works, easier is better (Sherman, 1893; King, 2010). Finally, it is worth noting how binary classification tends to report higher results than multi-class. While this is partly to be expected from the nature of the experiment, it might also suggest that "quality profiles" are indentifiable but shared through different quality proxies, pointing to a more universal perspective on what has high quality in literary works. Future research should aim to expand the corpus, integrate more diverse (non-Anglophone) literary traditions, and explore the temporal dynamics of literary quality.

Limitations

The selection of texts, while extensive, is not exhaustive and may reflect biases inherent in the compilation of canonical and award-winning lists. One important limitation of our corpus of novels is its strong Anglophone and American tilt: there are few non-American and non-Anglophone authors, which inevitably situates the entire analysis within the context of an Anglophone literary field.

Regarding the proxies of reader appreciation used in this study, it is hard to control the demographics of each proxy for literary quality and reception. Generally, sources like GoodReads are more diverse and represent a more comprehensive demographic selection than awards committees or anthologies' editorial boards, which are also susceptible to quick changes. Still it should be noted that the majority of GoodReads users from the beginnings of GoodReads in 2007 were native English speakers, which may affect the way users value non-Anglophone literary productions. Additionally, it is likely that there is a correlation between reviews on GoodReads and the quality categories suggested in this study, but as with any proxy measurement, it is difficult to concretely distinguish popularity, success, and quality.

Finally, the interpretation of complexity and its relation to quality is culturally and temporally situated and may change with both shifting literacy standards and literary norms.

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.

Alexandra Alter, Elizabeth A. Harris, and David McCabe. 2022. Will the Biggest Publisher in the United States Get Even Bigger? *The New York Times*.

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

Jean Barré, Jean-Baptiste Camps, and Thierry Poibeau. 2023. Operationalizing Canonicity: A Quantitative Study of French 19th and 20th Century Literature. *Journal of Cultural Analytics*, 8(3).

Jonah Berger, Yoon Duk Kim, and Robert Meyer. 2021. What Makes Content Engaging? How Emotional Dynamics Shape Success. *Journal of Consumer Research*, 48(2):235–250.

Yuri Bizzoni, Pascale Moreira, Nicole Dwenger, Ida Lassen, Mads Thomsen, and Kristoffer Nielbo. 2023a. 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 Moreira, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2023b. Sentimental matters - predicting literary quality by sentiment analysis and stylometric features. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 11–18, Toronto, Canada. Association for Computational Linguistics.

Yuri Bizzoni, Pascale Feldkamp Moreira, Mads Rosendahl Thomsen, and Kristoffer L. Nielbo. 2023c. The fractality of sentiment arcs for literary quality assessment: the case of nobel laureates. *Journal of Data Mining & Digital Humanities*, NLP4DH.

Yuri Bizzoni, Telma Peura, Kristoffer Nielbo, and Mads Thomsen. 2022. Fractal sentiments and fairy tales-fractal scaling of narrative arcs as predictor of the perceived quality of Andersen's fairy tales. *Journal of Data Mining & Digital Humanities*, NLP4DH.

Yuri Bizzoni, 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*, pages 1–6, NIT Silchar, India. NLP Association of India (NLPAI).

Harold Bloom. 1995. *The Western Canon: The Books and School of the Ages*, first riverhead edition edition. Riverhead Books, New York, NY.

Katherine Bode. 2023. What's the Matter with Computational Literary Studies? *Critical Inquiry*, 49(4):507–529.

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

Anselm Brachmann and Christoph Redies. 2017. Computational and experimental approaches to visual aesthetics. *Frontiers in Computational Neuroscience*, 11:102.

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45(1):5–32.

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.

Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. 2017. Affective computing and sentiment analysis. In *A Practical Guide to Sentiment Analysis*, pages 1–10. Springer.

Davida H. Charney and Jack R. Rayman. 1989. The Role of Writing Quality in Effective Student Résumés. *Journal of Business and Technical Communication*, 3(1):36–53. Publisher: SAGE Publications Inc.

Jonathan Cheng. 2020. Fleshing out models of gender in English-language novels (1850–2000). *Journal of Cultural Analytics*, 5(1):11652.

João Cordeiro, Pedro R. M. Inácio, and Diogo A. B. Fernandes. 2015. Fractal beauty in text. In Francisco Pereira, Penousal Machado, Ernesto Costa, and Amílcar Cardoso, editors, *Progress in Artificial Intelligence*, Lecture Notes in Computer Science, pages 796–802. Springer.

Tess Crosbie, Tim French, and Marc Conrad. 2013. Towards a model for replicating aesthetic literary appreciation. In *Proceedings of the Fifth Workshop on Semantic Web Information Management*, SWIM '13, pages 1–4, New York, NY, USA. Association for Computing Machinery.

Scott A. Crossley, Rod Roscoe, and Danielle S. McNamara. 2014. What Is Successful Writing? An Investigation Into the Multiple Ways Writers Can Write Successful Essays. *Written Communication*, 31(2):184–214. Publisher: SAGE Publications Inc.

Nan Z. Da. 2019. The computational case against computational literary studies. *Critical inquiry*, 45(3):601–639.

Katharina Ehret and Benedikt Szmrecsanyi. 2016. An information-theoretic approach to assess linguistic complexity. In *Complexity, Isolation, and Variation*, pages 71–94. De Gruyter.

Gerardo Febres and Klaus Jaffe. 2017. Quantifying literature quality using complexity criteria. *Journal of Quantitative Linguistics*, 24(1):16–53. ArXiv:1401.7077 [cs].

Richard S. Forsyth. 2000. Pops and flops: Some properties of famous english poems. *Empirical Studies of the Arts*, 18(1):49–67.

Vikas Ganjigunte Ashok, 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*, pages 1753–1764, Seattle, Washington, USA. Association for Computational Linguistics.

Craig L. Garthwaite. 2014. Demand spillovers, combative advertising, and celebrity endorsements. *American Economic Journal: Applied Economics*, 6(2):76–104.

John Guillory. 1995. *Cultural Capital: The Problem of Literary Canon Formation*. University of Chicago Press, Chicago, IL.

Ernest Hemingway. 1999. *On Writing*. Touchstone, New York.

Patrick C. Hogan. 2011. *Affective Narratology: The Emotional Structure of Stories*. University of Nebraska Press.

Horace. 2005. Ars poetica, espistula iii.

Qiyue Hu, Bin Liu, Mads Rosendahl Thomsen, Jianbo Gao, and Kristoffer L Nielbo. 2020. Dynamic evolution of sentiments in Never Let Me Go: Insights from multifractal theory and its implications for literary analysis. *Digital Scholarship in the Humanities*, 36(2):322–332.

Marc Hye-Knudsen, Ross Deans Kristensen-McLachlan, and Mathias Clasen. 2023. How Stephen King writes and why: Language, immersion, emotion. *Orbis Litterarum*, 78(5):353–367.

Arthur M. Jacobs and Annette Kinder. 2022. Computational analyses of the topics, sentiments, literariness, creativity and beauty of texts in a large Corpus of English Literature. ArXiv:2201.04356 [cs].

Matthew Jockers. 2014. A novel method for detecting plot.

Matthew Jockers. 2015. Revealing sentiment and plot arcs with the syuzhet package.

Justine Kao and Dan Jurafsky. 2012. A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 8–17, Montréal, Canada. Association for Computational Linguistics.

Stephen King. 2010. *On Writing: A Memoir of the Craft*, anniversary edition. Scribner, New York.

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.

Kayvan Kousha, Mike Thelwall, and Mahshid Abdoli. 2017. Goodreads reviews to assess the wider impacts of books. 68(8):2004–2016.

Hoyt Long and Teddy Roland. 2016. US Novel Corpus. Technical report, Textual Optic Labs, University of Chicago.

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. 2018. 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.

Claude Martin. 1996. Production, content, and uses of bestselling books in quebec. *Canadian Journal of Communication*, 21(4).

John McDonough and Andrzej Herczyński. 2023. Fractal patterns in music. *Chaos, Solitons & Fractals*, 170:113315.

Lee H. McGavin. 1997. Creativity as Information: Measuring Aesthetic Attractions. *Nonlinear Dynamics, Psychology, and Life Sciences*, 1(3):203–226.

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

Mahdi Mohseni, Volker Gast, and Christoph Redies. 2021. Fractality and variability in canonical and non-canonical english fiction and in non-fictional texts. *Frontiers in Psychology*, 12.

Mahdi Mohseni, Christoph Redies, and Volker Gast. 2022. Approximate entropy in canonical and non-canonical fiction. *Entropy*, 24(2):278.

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

Federico Pianzola, Simone Rebora, and Gerhard Lauer. 2020. Wattpad as a resource for literary studies. Quantitative and qualitative examples of the importance of digital social reading and readers' comments in the margins. *PLOS ONE*, 15(1). Publisher: Public Library of Science.

Frederico Pianzola, Srishti Sharma, and Frank Tsiwah. 2023. A computational analysis linking the emotion arcs of books and reader response.

J.D. Porter. 2018. *Stanford Literary Lab Pamphlet 17: Popularity/Prestige*. Stanford Literary Lab.

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.

Simone Rebora. 2023. Sentiment Analysis in Literary Studies. A Critical Survey. *Digital Humanities Quarterly*, 17(2).

Emily Sheetz. 2018. Evaluating Text Generated by Probalsitic Language Models.

Lucius A. Sherman. 1893. *Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry*. Athenaeum Press. Ginn.

Sean Shesgreen. 2009. Canonizing the canonizer: A short history of the norton anthology of english literature. *Critical Inquiry*, 35(2):293–318.

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.

William Strunk, E. B. White, and Roger Angell. 1999. *The Elements of Style*, 4th edition edition. Pearson, New York, Munich.

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

Ted Underwood, David Bamman, and Sabrina Lee. 2018. The transformation of gender in english-language fiction. *Journal of Cultural Analytics*, 3(2):11035.

Ted Underwood and Jordan Sellers. 2016. The *Longue Durée* of literary prestige. *Modern Language Quarterly*, 77(3):321–344.

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.

Willie van Peer. 2008. Ideology or aesthetic quality? In Willie van Peer, editor, *The Quality of Literature: Linguistic Studies in Literary Evaluation*, pages 17–29. John Benjamins Publishing.

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.

Robert von Hallberg. 1983. Editor's Introduction. *Critical Inquiry*, 10(1):iii–vi.

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.

Paul West. 1985. In Defense of Purple Prose. *The New York Times*.

Yaru Wu. 2023. Predicting the Unpredictable – Using Language Models to Assess Literary Quality. Master's thesis, Uppsala University, Uppsala.

Yaru Wu, Pascale Feldkamp Moreira, Kristoffer L. Nielbo, and Yuri Bizzoni. 2024. Perplexing Canon: A study on GPT-based perplexity for canonical and non-canonical literary works. In *To appear in: Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, St. Julians, Malta. Association for Computational Linguistics.

Claire M. Zedelius, Caitlin Mills, and Jonathan W. Schooler. 2019. Beyond subjective judgments: Predicting evaluations of creative writing from computational linguistic features. *Behavior Research Methods*, 51(2):879–894.

A Appendix

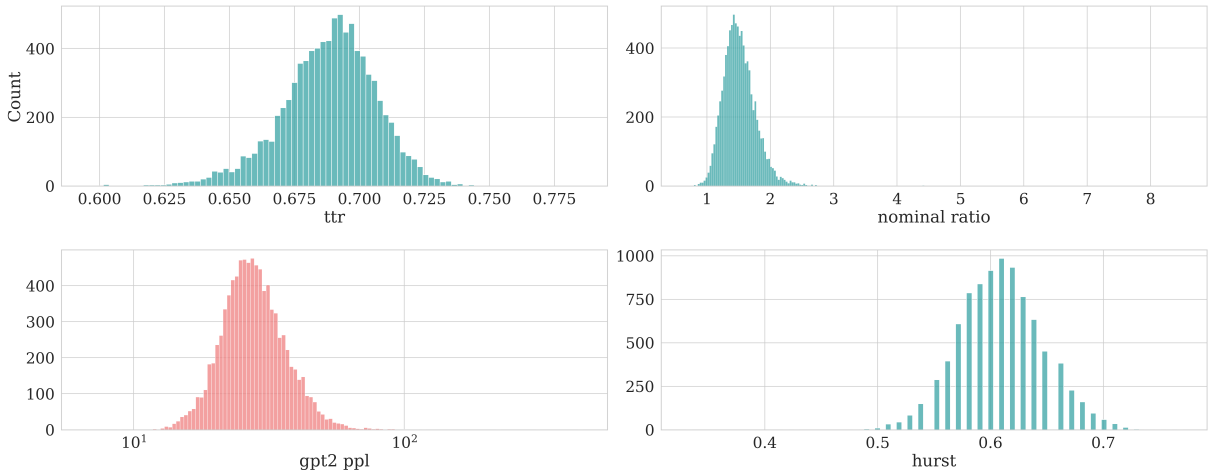

Figure 7: Distribution of selected features from each feature-type (Stylistic, Stylistic/syntactic, and narrative/sentiment).

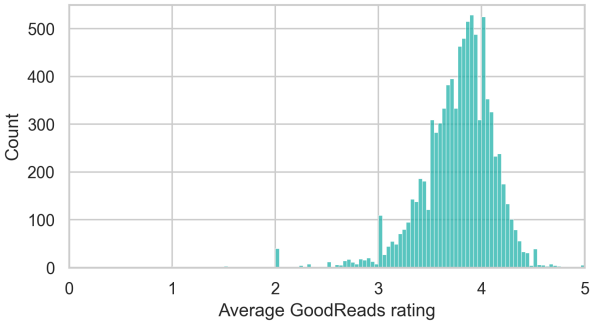

Figure 8: Distribution of average GoodReads ratings in our corpus. Note the noticeable positive skew of ratings.