A Matter of Perspective: Building a Multi-Perspective Annotated Dataset for the Study of Literary Quality

Yuri Bizzoni*, **Pascale Feldkamp***, **Ida Marie S. Lassen***,
Mads Rosendahl Thomsen†, **Kristoffer L. Nielbo***
†Comparative Literature – School of Communication and Culture, Aarhus University
*Center for Humanities Computing, Aarhus University
yuri.bizzoni@cc.au.dk

Abstract

Studies on literary quality have constantly stimulated the interest of critics, both in theoretical and empirical fields. To examine the perceived quality of literary works, some approaches have focused on data annotated through crowd-sourcing platforms, and others relied on available expert annotated data. In this work, we contribute to the debate by presenting a dataset collecting quality judgments on 9,000 19th and 20th century English-language literary novels by 3,150 predominantly Anglophone authors. We incorporate expert opinions and crowd-sourced annotations to allow comparative analyses between different literary quality evaluations. We also provide several textual metrics chosen for their potential connection with literary reception and engagement. While a large part of the texts is subjected to copyright, we release quality and reception measures together with stylometric and sentiment data for each of the 9,000 novels to promote future research and comparison.

Keywords: Literary quality, stylometry, digital humanities, literary analysis, sentiment analysis, readability

1. Introduction

The advent of computational methods is changing how we analyze and understand literature. From the quantitative assessment of linguistic patterns to the deep, qualitative insights into thematic elements, computational literary studies are redefining the landscape of literary criticism and research. The operationalization of complex concepts in computational linguistics and digital humanities comes with the possibility of deepening our understanding of literary narrative and writing but also involves the difficulty of relying on quantifiable elements. However, comprehensive, well-structured, and curated datasets are indispensable to leverage the full potential of quantitative methods.

In this work, we present a new dataset designed to further the analysis of one of the most complex and controversial concepts of literary theory: *quality* - with a focus, in this case, on the possible relation between textual features and perceived quality at a statistical level. While the study and discussion of literary quality are thousands of years old, extensive datasets to approach the problem from a quantitative and statistical perspective are not abundant.[1] We present a dataset designed to explore the theme of "quality" in computational literary studies, offering a rich array of textual and metadata features and a diverse collection of "quality" or reception proxies. It comprises various literary works spanning multiple genres, periods, and cultural contexts, although mainly confined to the Anglo-Saxon world. It can also be a robust foundation for related research objectives, such as sentiment analysis, stylistic evolution, and literary thematic categorization.[2,3]

The paper is structured as follows. Section 2 provides an overview of the state of the art on literary quality, especially in the more recent context of computational studies. Section 3 offers an overview of the dataset, including its size and origin. Section 4 presents the various metrics for assessing the perceived quality of literary novels that we have collected. In continuation, Section 5 presents the textual metrics we calculated for each novel, and Section 6 briefly explains the metadata fields accompanying the dataset. Section 7 discusses the limitations of our dataset, its advantages, its intended uses and proposes directions for future enhancements.

2. Related works

While the ability to process and analyze large quantities of texts through complex statistical experiments has recently made new ways of study-

[1]Few corpora are available like that compiled by Maharjan et al. (2017), via https://github.com/sjmaharjan/emotion_flow, and tend to index a relatively small amount of texts.

[2]We make both intrinsic and extrinsic features for all novels publicly available at: https://github.com/centre-for-humanities-computing/chicago_corpus

[3]While a large part of the corpus is subject to copyright (so that full texts cannot be released), full text of the pre-1924 novels can be found here: https://artflsrv04.uchicago.edu/philologic4.7/chicago_novel_corpus_pre1923_12-20/.

LREC-COLING 2024, pages 789–800
20-25 May, 2024.

ing literary appreciation possible, the question of how to define literary quality is probably as old as literature. Modeling perceived literary quality or reader appreciation poses a challenge to research on at least two dimensions: the number of features one could explore and the number of potential "judges" one could interrogate. Even if there may be a large consensus on the quality of a particular text, the underlying reasons are usually elusive and not necessarily rooted in the text itself. Setting aside possible biases underlying literary judgments for a moment, text-oriented schools of thought such as Van Peer (2008) have tended to look at the intrinsic textual features of literary works to explore their effectiveness. Still, that alone is a complex endeavor. Through the centuries, there have been many rules and recommendations to write better, supposedly applicable across genres and to both high and low-brow literature. Sherman (1893), for example, proposed that simplicity – i.e. shorter sentences, closer to the way we speak – should be a marker of a "better" literary style. Measuring textual simplicity has often been done via readability indices (gauging, generally, sentence and word length), which have also more recently been valued as creative writing and publishing aids – implemented in editing tools such as the Hemingway or Marlowe applications[4].

Still, the importance of the "readability" of a literary text in the context of reader appreciation is essentially controversial (Martin, 1996; Garthwaite, 2014). Considering the complexity and internal heterogeneity of what we call "literature". Naturally, features beyond sentence and word length impact the reading experience. Still, studies seeking to predict literary success or perceived literary quality follow the intuitive idea that readers perceive a difference between "difficult" and "easy" reads and tend to approximate some form of stylistic complexity by using textual features related to readability indices, such as sentence-length, vocabulary richness, or 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).

On this intuition, more general "simplicity laws" have been developed by critics and writers alike – for example, Ernest Hemingway's recommendation of a "direct and personal" style in "simple and vigorous" words (Hemingway, 1999). King (2010) offers very concrete advice in *On Writing*, where he advocates, among other things, more "readable" texts (shorter words and sentences) and fewer adverbs. Strunk et al. (1999) influential book *The Elements of Style* advocates very concrete advice, such as using the active voice and putting statements in the positive form, together with vaguer rules such as omitting needless words. Conversely, others have promoted what has been termed "purple prose" (a notion derived from Horace's *Ars Poetica*), characterized as challenging, "rich, succulent and full of novelty" (West, 1985). Indeed, reader preferences regarding the "difficulty" of prose, at least in terms of readability formulas, appear to be audience-specific (Bizzoni et al., 2023a). Studies that seek to model reader appreciation or canonicity have generally looked at stylistic features, 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 measurements like the compressibility of a text file using standard file compressors (Koolen et al., 2020).

Beyond the stylistic level, some work has been done on more underlying narrative features of literary texts, especially with the use of sentiment analysis, even if questions persist on how to measure narratological components or, in the case of sentiment analysis, how to operationalize an affective narratology (Rebora, 2023). Studies have sought to measure the shapes of a text's sentiment arc or to approximate narrative complexity (Maharjan et al., 2018; Reagan et al., 2016; Bizzoni et al., 2022b), on the intuition that readers tend to appreciate certain shapes or a certain balance in the complexity of a narrative flow or arc. Studies have emphasized the potential of sentiment analysis (Alm, 2008; **?**), at the word (Mohammad, 2018), sentence (Mäntylä et al., 2018) and paragraph (Li et al., 2019) level, to uncover meaningful mechanisms in the reading experience (Drobot, 2013; Cambria et al., 2017; Kim and Klinger, 2018; Brooke et al., 2015; Jockers, 2017), usually by drawing scores from human annotations (Mohammad and Turney, 2013) or induced lexica (Islam et al., 2020). While most studies have focused on the shapes of sentiment arcs, Hu et al. (2021) have modeled their persistence, coherence, and predictability by looking at the arcs' entropy or by using fractal analysis (Mandelbrot and Ness, 1968; Mandelbrot, 1982, 1997; Beran, 1994; Eke et al., 2002; Kuznetsov et al., 2013).

Finally, some studies have explored changes in reader preferences as a historical development. For example, a taste for stylistically "easier" books may be an effect of changes in reader demographics with the emergence of mass readership (Klancher, 1983). A more general basic level of literacy across society strata may have led to a consumer demand for more accessible books [5], and an increasing market-logic may have pressed editors to pre-

[4] https://hemingwayapp.com/help.html, https://authors.ai/marlowe/

[5] Notably, the US National Reader Survey of 1993 found that 48 percent of adults have difficulties reading above 5th-grade level texts (Kirsch et al., 1993)

fer more straightforward literary style (Winter and O'Neill, 2022). Similarly, readers might have become younger, for example, with the Young Adult fiction boom in the 1960s (Bach, 2022). Lower reading speed and hermeneutic difficulty may have come to be viewed as a vice rather than a virtue (Steiner, 1978), so authors and publishers have favored more direct prose. Such conjectures of changes in reader demographics do not exclude the existence of a many-tiered literary audience, where an increasing number of readers demand more straightforward texts and different "high culture" readerships favor challenging works. A perspectivist approach to "literary quality", considering many "judges" or audiences, allows insight both into developments in reader demographics and into the multi-faceted phenomenon of literary preference.

2.1. Works using the resource

Some works have already used the presented resource to explore trends related to contemporary English-language literature and the question of literary quality.

Some studies have applied sophisticated measures to gauge shapes and approximate complexity at the narrative level of the books in the corpus, relating these sentiment dynamics to reader appreciation. Bizzoni et al. (2023b) modeled the persistence, coherence, and predictability of arcs through the Hurst coefficient and Approximate Entropy (ApEn) to measure global and local complexity, using them to train classifiers able to gauge the reception and perceived quality of unseen texts. Such measures appear to be applicable for distinguishing between types of literature (e.g., prize-winning novels vs. bestsellers) (Bizzoni et al., 2024). resource has proved valuable to train and test classifiers that try to gauge the reception and perceived quality of unseen novels (Bizzoni et al., 2023b).

Moreover, the corpus has been used to explore the relation between different types of reader valuation, as well as the relation of different such proxies to textual characteristics. For example, it has been used to find that features of style vary across "types" of literature: award-winning works are less readable, while more readable books appear to be rated more often on GoodReads (Bizzoni et al., 2023a). Similarly, it has also been the basis for a recent study finding that prestigious literature appears to elicit higher LLM-based perplexity than popular literature (Wu et al., 2024). Finally, beyond the relation between textual features and reader appreciation, the corpus has also been used for tracking stylistic change diachronically (Feldkamp et al., 2023).

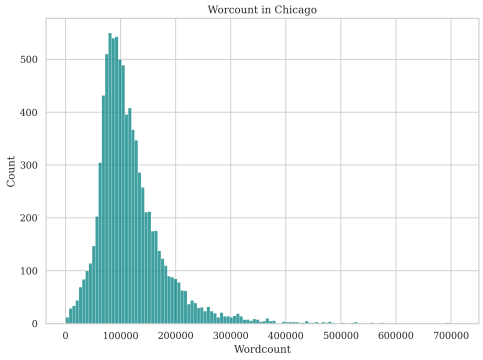

Figure 1: Distribution of wordcount in the corpus

3. Corpus

The corpus of texts from which we constructed our dataset was assembled by Hoyt Long and Richard Jean So; it encompasses 9088 novels published in the United States between 1880 and 2000 and was compiled based on the worldwide number of libraries holding each title[6], favoring works with a higher number of library holdings for their selection.

Because of this selection criteria, the corpus comprises much high-quality fiction from authors who have received prestigious distinctions, such as the Nobel Prize, the National Book Award (including Don DeLillo, Joyce Carol Oates, and Philip Roth), as well as important works of genre-fiction (i.a., Tolkien or Philip K. Dick). Still, library holdings appear to reflect high distinction and mass popularity, as acquisition reflects the average library user's demand and preferences. As such, the corpus also comprises influential novels from mainstream literature (i.a., Agatha Christie), with notable contributions on the broad spectrum of so-called "genre literature", from Mystery to Science Fiction (Long and Roland, 2016).[7].

The corpus has a geographical bias, comprising primarily Anglophone authors (with few exceptions). This bias inevitably situates any analysis of it within the context of a US and "Anglocentric" literary field. Books in the corpus vary in length, from 341 words (Beatrix Potter's *The Story of Miss Moppet*) to 714,744 words(Ben A. Williams' *House Divided*), though only 255 books – 2.9% of the corpus – are shorter than 35,000 words – the length of titles like Orwell's *Animal Farm* or Hemingway's *The Old Man and the Sea*. The total word count of the corpus is 1,060,549,793 words.

We divide the measures that we provide in our datasets into two categories: quality metrics and textual metrics.

[6]Based on the WorldCat catalog.

[7]Previous quantitative literary analyses have employed this corpus, (Underwood et al., 2018; Cheng, 2020; Bizzoni et al., 2022a) [8]

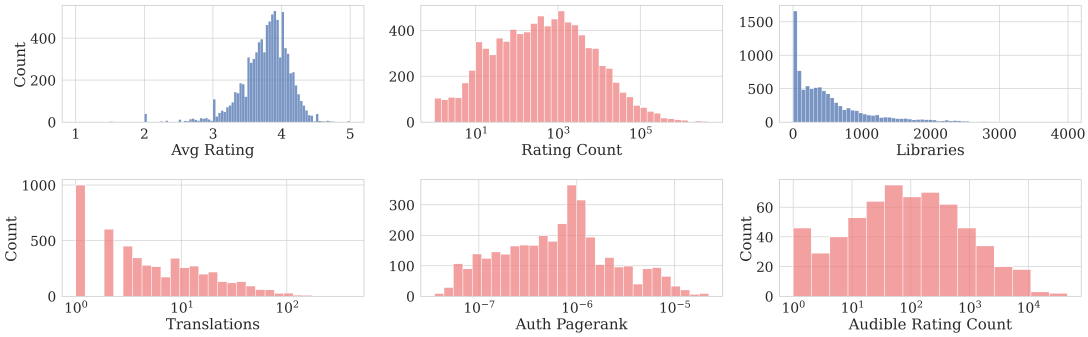

Figure 2: Distribution of continuous quality proxies in our corpus. For each histogram, titles with 0 values for the given proxy were excluded. In terms of audible ratings and rating count, for example, only 629 titles have ratings (see table 3). Note that histograms in red are logarithmically scaled.

Titles	Authors	Titles per author
9088	3166	2.88

Table 1: Number of titles, authors, and average titles per author in the dataset.

4. Quality Metrics

The quality metrics are arguably the rarer of the two categories in literary datasets and the most complicated from a conceptual standpoint. Understanding and quantifying "literary quality" is a complex endeavor (Bizzoni et al., 2022a), often subject to subjective evaluations. However, in the context of this study, we have operationalized this concept by considering a range of metrics commonly used in the academic and public discourse. The quality metrics that we have collected belong to two main types: crowd-based, representing the result of many unfiltered readers, and, on the other hand, expert-based, drawn from prestigious proxies curated by experts, often institutionally affiliated. It should be noted that this distnction is heuristic above all else, as various metrics, such as translation counts, are both subject to expert choice and the taste judgements of a larger readership.

4.1. Crowd-based Metrics

The main crowd-based metrics that we collected are:

1. **GoodReads' Rating Count**: This metric approximates a book's popularity among a general audience. It is the total Number of ratings a book has received on GoodReads.[9]

2. **GoodReads' Average Rating**: Unlike the rating count, the average rating measures how well GoodReads users received the book on a scale of 1 to 5.

3. **Audible Rating Count and Average Rating**: Rating count and average rating for the titles represented in Audible.[10]

4. **GoodReads Lists**: Users can collectively create and populate lists. Lists like *Best Books of the 20th Century* constitute a crowd-based representation of the concept of high-quality (and often canonical) literature.

5. **WorldCat Holdings**: This metric indicates the number of libraries worldwide holding a particular book, which can indirectly indicate the book's quality and importance.

6. **Wikipedia Author-page Rank**: Using wikipedia page-views, – the number of times visits to an author's page on Wikipedia – is also sometimes used as a proxy for canonicity or literary success (Hube et al., 2017). In Hube et al. (2017)'s (and our) variation of page-rank (a google algorithm) hubs or author-pages on Wikipedia that have the highest number of other pages referencing them have a higher rank, so that more referenced authors rank higher. The Wikipedia page rank thus also measures authors' presence in the popular and cultural sphere, if we consider that Wikipedia-pages may be created and edited by various types of users. It should be noted that ranks refer to authors, so that books by the same author will have the same rank, independently from differences in prestige or popularity between individual titles.

7. **Translation Count**: The *Index Translationum* database collects all translations published in ca. 150 UNESCO member states,

[9] GoodReads ratings and rating counts were collected in December 2022.

[10] Audible ratings and rating counts extend only to a small part of the corpus (see table 3). These ratings were collected in March 2023.

Award	Titles
National book award	108
Pulitzer prize	53
Nobel prize*	85
Scifi awards	163
Hugo award	
Nebula award	
Philip K. Dick award	
(Pope, 2019) J.W. Campbell award	
Prometheus award	
Locus sci-fi award	
Fantasy awards	40
World fantasy award	
Locus fantasy award	
British fantasy award	
Mythopoeic award	
Horror awards	19
Bram Stoker award	
Locus horror award	
Romantic awards*	54
Rita awards*	
RNa awards*	

Table 2: Number of longlisted titles for general fiction and genre-fiction awards, and the specific awards collected. Proxies marked * are author-based: For these, we included all titles extant in the corpus by the author mentioned, either due to the scarcity of awards in the genre or the nature of the award, e.g., the Nobel prize given to authors, rather than to individual titles. All other awards are title-based.

compiled from local bibliographical institutions or national libraries, cataloguing more than 2 million works. Note that the database was created in 1979 and stopped compiling in 2009. As such, the resource lists translations of a particular period, and not the most translated works of all time. The proxy should be interpreted with that in mind. Translation counts not a clear-cut crowd-based metric, as various factors (beyond popular demand) may influence which works are translated.

4.2. Expert-based Metrics

1. **Awards and Prizes**: Winning or being nominated for a prestigious literary award is a significant indicator of literary quality, so prizes can also serve as an expert-based quality metric. We collected long-listed titles (winners and finalists) for both prestigious literary awards: The Nobel Prize in Literature, the Pulitzer Prize, the National Book Award; as well as various genre-based awards (for the full list of awards, see table 2).

2. **Inclusion in Anthologies**: Being included in respected anthologies or literary collections, such as the *Norton Anthology*, a leading literary anthology (Pope, 2019), is another expert-based quality metric and can be seen as a proxy for canonization. For the present study, we marked all titles in our corpus written by authors mentioned in these two series, where the anthology of English Literature is the most widespread (Ragen, 1992).

3. **College Syllabi**: How often an author is assigned on college syllabi can serve as a complementary metric of canonization. We used the resource OpenSyllabus, which has collected 18.7 million college syllabi in an attempt to map the college curriculum.[11] From their data, we count all titles in our corpus by authors who appear as authors of one of the top 1,000 titles assigned in *English Literature* college syllabi.

4. **Classics Series**: Various large publishing houses, like Vintage or Penguin[12], have a type of classics series, while others, like Everyman's library, are entirely devoted to publishing "the classics".[13] As Penguin is arguably one of the biggest publishers of anglophone literature (Alter et al., 2022), we collected their classics series, both individual titles and all titles extant in our corpus by authors included in the series.

Some metrics, like GoodReads' rating count, are continuous, while others, like the Nobel Prize, are binary. This distinction allows for different statistical analyses and comparisons, enabling researchers to approach the question of literary quality from multiple angles. Crowd-based and expert-based metrics are only sometimes in agreement. For example, a high GoodReads rating count does not necessarily correlate with expert recognition (Bizzoni et al., 2023a), suggesting the multi-faceted nature of literary quality.
See Tables 3 and 4 for a complete list of the metrics we include in the dataset.

5. Textual Metrics

For each title in the collection, we provide several textual metrics.

5.1. Readability

Readability formulae, like Flesch Ease, have used aspects such as sentence length, word lengths, and syllable count to measure linguistic complexity (Dale and Chall, 1948). Despite a multitude of formulae (Dubay, 2004), a handfull of 'classic

[11] https://www.opensyllabus.org
[12] https://www.penguin.com/penguin-classics-overview/
[13] http://www.everymanslibrary.co.uk

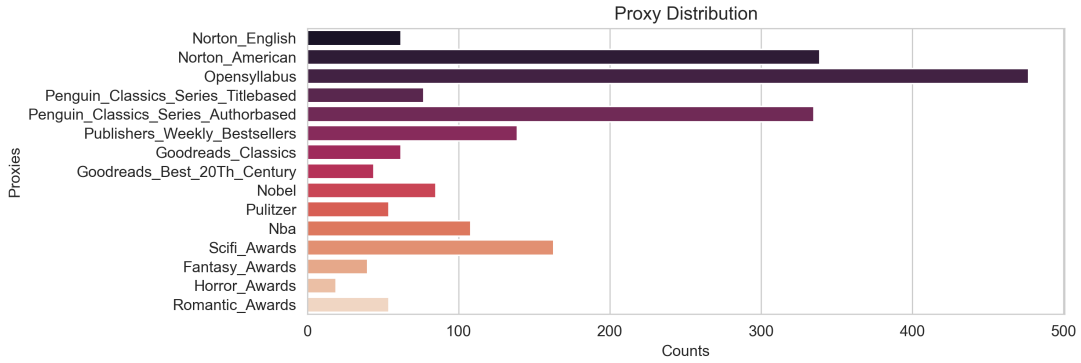

Figure 3: Number of titles in discontinuous quality proxies in our corpus.

	Count	Mean	Std
Translations	5082	11.77	21.47
PageRank*	3558	0.15	0.24
Audible Rat.Avg.	629	4.17	0.50
Audible Rat.Count	629	796.92	3020.15
GR Rat.Avg.	8989	3.77	0.36
GR Rat.Count	8989	14368.39	121551.55

Table 3: Continuous quality proxies. Proxies marked * are author-based. Note that the Wikipedia author page-rank has been multiplied with 100,000 for interpretability.

	Count
Norton English*	62
Norton American*	339
OpenSyllabus*	477
Penguin Classics Series	77
Penguin Classics Series*	335
Publishers Weekly Bestsellers	139
Goodreads Classics*	62
Goodreads Best 20th Century*	44
Nobel*	85
Pulitzer	54
NBA	108
Scifi Awards	163
Fantasy Awards	40
Horror Awards	19
Romantic Awards*	54

Table 4: Discontinuous quality proxies, all literary awards (general and genre-oriented) appear below the line. Note that proxies marked * are author-based.

readability measures that go back to the 1970s remain widely used (Stajner et al., 2012). To avoid relying on one single interpretation of the readability concept, we offer five popular and interpretable formulas for the corpus, all calculated through the textstat package.[14]. These have been shown to be strongly correlated (Bizzoni et al., 2023a). They include the *Flesch Reading Ease* and *Flesch-Kincaid Grade Level*, both based on average sentence length (ASL) and syllable count per word; the *SMOG Readability Formula* that uses ASL and polysyllable count (McLaughlin, 1969); the *Automated Readability Index*, employing ASL and word length; and the *New Dale–Chall Readability Formula*, which uses ASL and a 'difficult words' percentage (PDW), which represents the percentage of words unfamiliar to fourth graders (Chall and Dale, 1995; Dale and Chall, 1948).[15]

5.2. Stylistic Metrics

The stylistic metrics that we provide are:

1. **Lexical Diversity or Type-Token Ratio**: Measures the ratio of unique words to the total number of words in a text. Higher lexical diversity often suggests a richer vocabulary. A standard index of lexical richness, not used in readability metrics but normally considered indicative of a text's complexity and inner diversity (Torruella and Capsada, 2013).[16]

2. **Average Sentence and Word Length**: Average character-based sentence and word length. They both provide, in different ways, a simple yet effective measure of complexity. For example, Kerouac's *The Subterraneans*, a classic example of the "spontaneous" and vernacular prose of Beat Literature (Whaley, 2009), has the longest average sentence length.

3. **Compressibility** Measures how much a text is compressible through a standard compression algorithm. This measure becomes essentially a sign of redundancy and formulaic lan-

[14] https://pypi.org/project/textstat/

[15] https://countwordsworth.com/download/DaleChallEasyWordList.txt

[16] 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.

guage: 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).[17]

4. **Unigram and bigram entropy**: entropy based on unigrams or bi-gram pairs, based on the code[18] and study of Algee-Hewitt et al. (2016) of literary texts. Entropy refers to how much variation or randomness there is in terms of either words or word pairs (bigrams) in a given text. A lower entropy would indicate that words or bigrams recur more often, while a higher entropy would indicate a more significant variation in the vocabulary or the bigrams used. Unigram (word) entropy is, in this sense, similar to vocabulary richness measures.

5.3. Sentiment Analysis

At an arguably deeper level, we computed the sentence-based sentiment arcs of the novels, using the nltk's implementation of VADER (Hutto and Gilbert, 2014), arguably one of the most widespread dictionary-based methods. We provide the full version of the arcs and their coarser-grain representation in twenty segments. The detrended sentiment arc[19] based on our VADER scores of Hemingway's *The Old Man and the Sea* can be seen in Fig. 4, compared to a human baseline. Note that the Pearson and Spearman correlations of scores by the two human annotators for this work were robust (0.652, 0.624) but not perfect, reflecting the complexity of the task of assigning valences, as disagreements are considerable also among human annotators. In this light, the relatively straightforward rule-based system VADER appears to perform reasonably (Fig. 4), and has also been shown to have a high consistency across domains (Ribeiro et al., 2016).
Sentiment analysis in our study goes beyond merely categorizing the sentiment or overall valence of the text. We employ several statistical measures to provide a multi-faceted view of sentiment across the document. These measures include:

- **Mean Sentiment**: This is the average sentiment score across all the sentences in the document. It provides an overall sense of the valence of the text.

- **Standard Deviation of Sentiment (Std Sentiment)**: This metric captures the variability in sentiment across the document. A higher standard deviation implies a broader range of emotions expressed.

- **End Sentiment**: Refers to the sentiment score of the concluding part of the document (the last 5% of the sentences). It provides insights into the sentiment with which the document concludes.

- **Beginning Sentiment**: This is the sentiment score for the introductory part of the document (the first 5% of book). It sets the emotional stage for the reader.

- **Difference Ending to Mean**: This is the difference between the end (5%) and the sentiment of the rest of the book. It indicates whether the document ends on a more positive or negative note compared to its overall valence.

- **Hurst Exponent**: Used to detect long-term memory in time series data, the Hurst exponent in our context measures the persistence of sentiment over the document. Values near 0.5 suggest a random walk, while values far from 0.5 indicate trending or mean-reverting sentiment.

- **Approximate Entropy**: This measures the complexity of the sentiment time series. A lower value indicates more regularity in the sentiment, while a higher value suggests a lower predictability.

Each of these metrics serves a specific purpose, and when considered collectively, they provide a comprehensive understanding of the text's sentimental landscape. It is worth noting that these metrics are sensitive to the granularity of text segments (sentence vs. paragraph) and the sentiment lexicon used. Sentiment analysis (relying on word values and rules) provides a rare point of observation for novels, as it stands at the interface between their style and narrative structure. On one hand, the sentiment arcs of the novels represent the fluctuations of the narrative as developed through the novel. On the other hand, it detects the *way* in which the development is portrayed, rather than any judgment that the reader could give on the narrative, allowing us to detect the stylistic and rhetoric features of the text that we would otherwise easily override.

[17] 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.

[18] https://github.com/nan-da/Entropy-for-Bigrams

[19] We detrend arcs using the adaptive filtering technique for nonlinear series proposed by Jianbo Gao et al. (2010), which has previously successfully been applied to sentiment arcs of novels (Hu et al., 2021; Bizzoni et al., 2022c). Arcs are on the second polynomial fit.

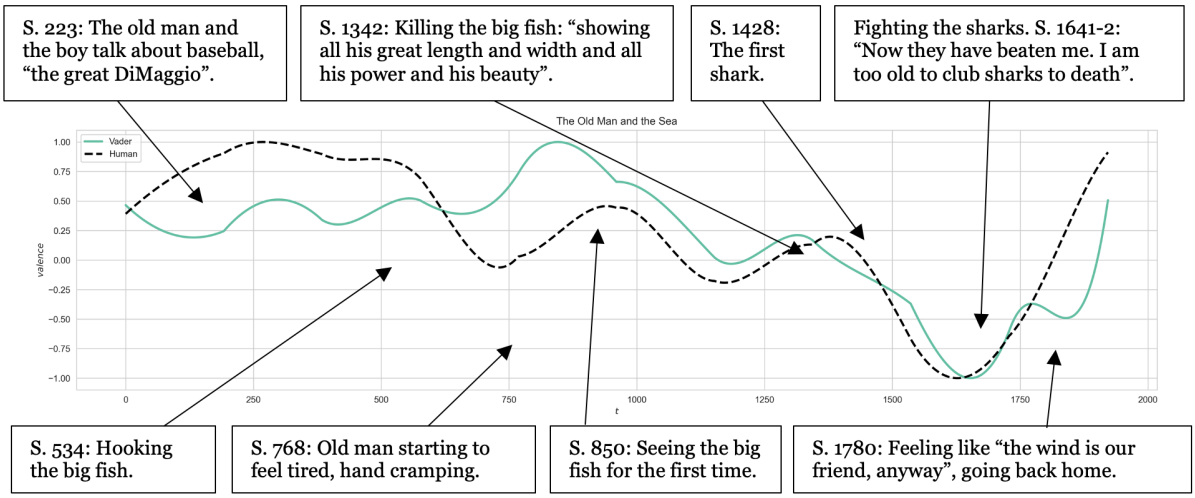

Figure 4: Detrended arcs and manually annotated of *The Old Man and the Sea* based on VADER valences and mean of human annotators (n=2).

	Mean	**Std.**
Wordcount	118584.71	64746.05
Sentence Length	86.56	29.44
Wordlength	3.67	0.18
MSTTR-100	0.69	0.02
Bzip	2.92	0.14
Bigram Entropy	14.63	0.55
Word Entropy	9.69	0.30
Flesch Ease	82.70	6.48
Flesch Grade	5.19	1.74
Smog	8.20	1.05
ARI	6.91	2.06
Dale Chall New	5.10	0.33
Mean Sent.	0.03	0.04
Std Sent.	0.35	0.04
End Sent.	0.03	0.07
Beginning Sent.	0.04	0.05
Diff. Ending/Rest	0.01	0.05
Hurst Exponent	0.61	0.04
Approximate Entropy	1.75	0.15

Table 5: Textual measures. From the bottom down: "surface-level" stylometrics, readability formulæ, and measures associated with the novels' sentiment arcs.

6. Metadata

Beyond the textual and quality metrics, we provide metadata. The metadata accompanying our dataset is an essential framework for contextualizing its content. The fields collected for each book in the dataset include:

- **Author Name**: The name of the individual or collective responsible for creating the work. It can be helpful in studies focusing on authorship patterns or historical context.
- **Title**: The book's title, which is instrumental for identification and categorical analysis.
- **Publication Date and Decade**: We provide the exact publication date and the decade to which the book belongs. These temporal markers assist in longitudinal studies and trend analysis.
- **Publishing Location**: This is the geographical location where the book was published, which can be valuable for regional studies and geopolitical analysis.
- **BookID**: A unique identifier assigned to each book in the dataset. It facilitates easy referencing and data manipulation.
- **Author's Gender**: Identifies the gender of the author(s). The distinction is currently binary.
- **Genre Tags**: Genre tags for one or more genres, manually added by a literary scholar. This addition could aid thematic categorization and genre-specific analyses. This information is available only for a subset of 1000 titles in the dataset. This limitation is due to the scope and difficulty of genre annotation.

The metadata fields offer a multi-dimensional lens to understand, segment, and analyze the dataset.

7. Conclusion and Future Works

We have presented a large new dataset designed to study "literary quality" as a compound of several different perspectives. Including crowd-based and expert-based assessments, the dataset allows for several combinations of textual and quality features and the study of continuous and discrete representations of "literary quality". To the best of

our knowledge, this is the largest extant dataset with multiple-perspective literary quality annotations containing extensive textual features.

Naturally, we intend this dataset for scholars and critics to explore the complex interplays between textual and reception metrics. As it is, the dataset can be used to explore simple correlations between different textual metrics (e.g., the correlation between the mean sentiment and the end-sentiment of novels) and between different quality metrics alone (e.g., GoodReads' rating counts correlate more with audible rating counts than with the WorldCat's numbers). It is also essential to consider that binary quality metrics, such as the presence of a novel or an author in a given anthology, are not mutually exclusive. Some titles appear, for example, both in the Norton Anthology *and* in the Penguin Classics Series. This can allow the dataset users to obtain a non-binary metric, scoring higher the texts that appear in more than one proxy and creating a nuanced version of canonicity.

However, the main goal of the dataset is to facilitate the study of the link between textual features and the perceived quality or reader appreciation of a literary text, but a subset of quality proxies can also be used to investigate "canonicity" of literary texts. While we aim to provide a comprehensive range of quality metrics, we acknowledge that no metric can fully capture the nuanced and subjective nature of literary quality. Future work may incorporate additional metrics such as citations in academic work, or social media mentions, as well as a much more comprehensive range of textual features, such as syntactic and semantic profiles of the novels.

8. Bibliographical References

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

Ebba Cecilia Ovesdotter Alm. 2008. *Affect in* text and speech*. Phd thesis, University of Illinois at Urbana-Champaign.

Alexandra Alter, Elizabeth A. Harris, and David McCabe. 2022. Will the biggest publisher in the United States get even bigger? *The New York Times*.

Jacqueline Bach. 2022. Young Adult Boom. In Patrick O'Donnell, Stephen J. Burn, and Lesley Larkin, editors, *The Encyclopedia of Contemporary American Fiction 1980–2020*, 1 edition, pages 1–10. Wiley.

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

Jan Beran. 1994. *Statistics for Long-Memory Processes*, 1 edition. Chapman and Hall/CRC, New York.

Yuri Bizzoni, Pascale Feldkamp, Ida Marie Lassen, Mia Jacobsen, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2024. Good books are complex matters: Gauging complexity profiles across diverse categories of perceived literary quality.

Yuri Bizzoni, Ida Marie Lassen, Telma Peura, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2022a. Predicting Literary Quality How Perspectivist Should We Be? In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 20–25, Marseille, France. European Language Resources Association.

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, Telma Peura, Kristoffer Nielbo, and Mads Thomsen. 2022b. 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, Kristoffer Nielbo, and Mads Thomsen. 2022c. 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.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2015. Gutentag: an nlp-driven tool for

digital humanities research in the project gutenberg corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 42–47.

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.

Jeanne S. Chall and Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books.

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

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.

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

Irina-Ana Drobot. 2013. Affective narratology. the emotional structure of stories. *Philologica Jassyensia*, 9(2):338.

William Dubay. 2004. *The Principles of Readability*. Impact Information.

A. Eke, P. Herman, L. Kocsis, and L. R. Kozak. 2002. Fractal characterization of complexity in temporal physiological signals. *Physiological Measurement*, 23(1):R1.

Pascale Feldkamp, Yuri Bizzoni, Ida Marie S. Lassen, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2023. Readability and complexity: Diachronic evolution of literary language across 9000 novels. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 235–247, Tokyo, Japan. Association for Computational Linguistics.

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

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

Qiyue Hu, Bin Liu, Mads Rosendahl Thomsen, Jianbo Gao, and Kristoffer L Nielbo. 2021. 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.

Christoph Hube, Frank Fischer, Robert Jäschke, Gerhard Lauer, and Mads Rosendahl Thomsen. 2017. World literature according to Wikipedia: Introduction to a DBpedia-based framework.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

SM Mazharul Islam, Xin Dong, and Gerard de Melo. 2020. Domain-specific sentiment lexicons induced from labeled documents. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6576–6587, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jianbo Gao, H. Sultan, Jing Hu, and Wen-Wen Tung. 2010. Denoising Nonlinear Time Series by Adaptive Filtering and Wavelet Shrinkage: A Comparison. *IEEE Signal Processing Letters*, 17(3):237–240.

Matthew Jockers. 2017. Syuzhet: Extracts sentiment and sentiment-derived plot arcs from text (version 1.0. 1).

Evgeny Kim and Roman Klinger. 2018. A survey on sentiment and emotion analysis for computational literary studies. *arXiv preprint arXiv:1808.03137*.

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

Irwin S. Kirsch, United States, Educational Testing Service, and National Center for Education Statistics, editors. 1993. *Adult literacy in America: a first look at the results of the National Adult Literacy Survey*, 2nd ed edition. Office of Educational Research and Improvement, U.S. Dept. of Education, Washington, D.C.

Jon P. Klancher. 1983. From "crowd" to "audience": The making of an english mass readership in the nineteenth century. *ELH*, 50(1):155–173.

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.

Nikita Kuznetsov, Scott Bonnette, Jianbo Gao, and Michael A. Riley. 2013. Adaptive Fractal Analysis Reveals Limits to Fractal Scaling in Center of Pressure Trajectories. *Annals of Biomedical Engineering*, 41(8):1646–1660.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41, Hong Kong, China. Association for Computational Linguistics.

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.

Benoit Mandelbrot. 1982. *The Fractal Geometry of Nature*. Times Books, San Francisco.

Benoit B. Mandelbrot. 1997. *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk. Selecta Volume E*, 1997 edition edition. Springer, New York.

Benoit B. Mandelbrot and John W. Van Ness. 1968. Fractional Brownian Motions, Fractional Noises and Applications. *SIAM Review*, 10(4):422–437.

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

Harry G. McLaughlin. 1969. Smog grading: A new readability formula. *Journal of Reading*, 12(1):639–646.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif Mohammad and Peter Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2:1–234.

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

Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. 2018. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. 27:16–32.

Colin Pope. 2019. We need to talk bout the canon: Demographics in 'The Norton Anthology'. *The Millions*.

Brian Abel Ragen. 1992. An uncanonical classic: The politics of the "Norton Anthology". *Christianity and Literature*, 41(4):471–479.

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).

Filipe N. Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29. Number: 1 Publisher: SpringerOpen.

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

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.

George Steiner. 1978. On Difficulty. *The Journal of Aesthetics and Art Criticism*, 36(3):263–276. Publisher: [Wiley, American Society for Aesthetics].

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.

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. *The quality of literature: Linguistic studies in literary evaluation*, volume 4. John Benjamins Publishing.

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

Preston Whaley. 2009. *Blows Like a Horn: Beat Writing, Jazz, Style, and Markets in the Transformation of U.S. Culture*. Harvard University Press.

Marna K. Winter and Kristen O'Neill. 2022. An exploration of prevalence and usage of hi-lo texts in today's classrooms. *Reading & Writing Quarterly*, 0(0):1–15.

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.