Article

Measuring Literary Quality
Proxies and Perspectives

Pascale Feldkamp[1] (iD)
Yuri Bizzoni[1] (iD)
Mads Rosendahl Thomsen[2] (iD)
Kristoffer L. Nielbo[1] (iD)

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

Abstract. Computational studies of literature use proxies like sales numbers, human judgments, or canonicity to estimate literary quality. However, many quantitative use one such measure as a gold standard without fully reflecting on what it represents. We examine the interrelation of 14 proxies of literary quality in novels published in the US from the late 19th to 20th century, distinguishing between expert-based judgments (e.g., syllabi, anthologies) and crowd-based ones (e.g., GoodReads ratings). We show that works favored in expert-based judgments often score lower on GoodReads, while award-nominated works tend to circulate more widely in libraries. Generally, two main kinds of 'quality perception' emerge as we map the literary judgment landscape: One associated with canonical literature and one with more popular literature. Additionally, prestige in genre literature, reflected in awards like the Hugo Award, forms a distinct category, more aligned with popular than canonical proxies.

1. Introduction

The concept of quality in literature is a fascinating riddle: It would seem that the idiosyncratic nature of reading precludes any objective standard for what constitutes a 'good' book – and yet certain texts seem to have an enduring appeal, being of interest to readers across time and national borders and are consecrated in the institutional canons of different cultures. This paradox lies at the heart of discussions about what literary quality is, as well as of attempts to define, measure, or predict it.[1]

The challenge of defining literary quality is complicated by the diversity of preferences of individual readers and reader types (Riddell and van Dalen-Oskam 2018) and even the tendency of readers to change their opinion about a text (Harrison and Nuttall 2018; Kuijpers and Hakemulder 2018). Moreover, the question of what constitutes literary quality and where it resides (in style, plot, emotional engagement, themes, etc.) quickly

---

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

becomes a complicated matter of its own, one that schools of literary criticism have grappled with in many different ways (Bjerck Hagen et al. 2018).

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

With recent computational inquiry into literary studies and sizeable attempts at measuring 'quality', this disparity is even more apparent. The stricter conditions of quantitative analysis – operationalizing traditional disciplinary concepts – bring the complexity of the idea of 'quality' in literature to the fore. Computational studies of literary preferences have found that reader appreciation or success can, to some extent, be predicted by stylistic features (Maharjan et al. 2017; van Cranenburgh and Bod 2017; van Dalen-Oskam 2023), as well as narrative features such as plot (Jockers 2015), emotional valence and flow (Maharjan et al. 2018; Reagan et al. 2016; Veleski 2020), or the predictability of novels' sentiment-arcs (Bizzoni et al. 2022a,b, 2021) – not to mention text-extrinsic features such as genre, promotion, author visibility, and gender (C. W. Koolen 2018; Lassen et al. 2022; Wang et al. 2019). While such studies point to the existence of certain consensuses, it should be noted that these studies define the concept of success or quality very differently. The first and possibly most complex task of quantitative studies of literary quality is that of defining a 'proxy' of quality itself: From where should we take the judgments we intend to explain?

In computational literary studies, a 'proxy' serves as a formal method for approximating abstract constructs or concepts through operationalization. Proxies bridge qualitative interpretation with quantitative methodologies: They translate constructs or concepts, like 'quality in literature', into measurable variables. A 'quality proxy' thus means

---

2. A very Marxist reader, Leon Trotsky, observed how the historical and aesthetic dimensions of art are utterly independent: "If I say that the importance of the Divine Comedy lies in the fact that it gives me an understanding of the state of mind of certain classes in a certain epoch, this means that I transform it into a mere historical document, for, as a work of art, the Divine Comedy must speak in some way to my feelings and moods... Dante was, of course, the product of a certain social milieu. But Dante was a genius. He raised the experience of his epoch to a tremendous artistic height. And if we, while today approaching other works of medieval literature merely as objects of study, approach the Divine Comedy as a source of artistic perception, this happens not because Dante was a Florentine petty bourgeois of the 13th century but, to a considerable extent, in spite of that circumstance" (Trotsky 1974, 94).

a specific operationalization of appreciation among many. For example, we might differentiate between literary 'fame' and 'popularity', since fame, such as the fame of James Joyce's *Ulysses*, does not necessarily mean it is widely read. These different forms of quality may be measured in dissimilar ways – i.e., through different 'proxies' – for example, by looking at how often a book is the subject of literary scholarship vs. how many copies it sells or how often it is rated on GoodReads.[3]

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

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

To address the question of differences between quality proxies, we collected 14 different possible proxies for literary quality, ranging from popular online platforms to university syllabi and prestigious awards, and used them to annotate a corpus of over 9,000 novels

---

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

(note that we do not analyze the texts themselves in this article).[4] Our central question was whether and to what extent these metrics measure the same thing, with the following hypotheses: If the "quality" measured by, e.g., GoodReads ratings differs from that represented by the number of library holdings, the two metrics are unrelated, that is, "quality" is likely driven by multiple hidden variables. If, instead, there is a statistically reliable overlap – that is, books popular on GoodReads are also acquired by many libraries – they are related, that is, there is a single hidden variable of quality. To the best of our knowledge, this is the first study to compare several judgments of literary quality on a large collection of modern titles and try to understand the relationship between them rigorously.

2. Related Works

Studies have found that there seems to be a consensus among readers about what works are 'classics'. Walsh and Antoniak (2021) tested the relation between GoodReads' Classics, a user-compiled list, and titles included in college English syllabi (as collected by the *Open Syllabus Project*), showing that there is a significant overlap between what is perceived as classics on GoodReads and what appears on college syllabi. Thus, users seem to be replicating a particular perception of the 'canonicity' of titles.

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

Another study by Porter (2018) sought to model differences in popularity and prestige in their corpus, using, on the one hand, GoodReads' average ratings and, on the other hand, the Modern Language Association's database of literary scholarship, counting the number of mentions of an author as the primary subject of a scholarly work. They show a clear difference in the equilibrium between popularity and prestige across genres. Books from genres like Sci-Fi are rated very often on GoodReads but are sparsely represented in scholarly work, while poetry exhibits an opposite tendency. Based on Pierre Bourdieu's conceptualization of the literary field (Bourdieu 1993), they define two axes of literary "success", prestige and popularity, as online popularity (on GoodReads) and prestige among literary scholars (represented in the MLA database) so that their "map" risks to look overly neat. Literary scholars, for example, may not be the primary nor most important actors in processes of literary prestige. Manshel et al. (2019) have shown how literary prizes – appointed by committees who may be either authors themselves, scholars, or lay-readers – appear to have an important role in positively influencing both prestige and popularity.[5]

While only a few studies have tried to measure differences and convergences of literary quality judgments quantitatively, the question of how literary cultures evaluate texts has been central to sociological approaches to literature. Especially the attempts of Bourdieu

---

4. See section 4 for a discussion of this corpus, which, it should be noted, is heavily skewed toward American and Anglophone authors.
5. Using the same definitions of popularity and prestige as Porter (2018), it seems that whether or not books had received a prize significantly raised the probability of both being popular and prestigious (Manshel et al. 2019).
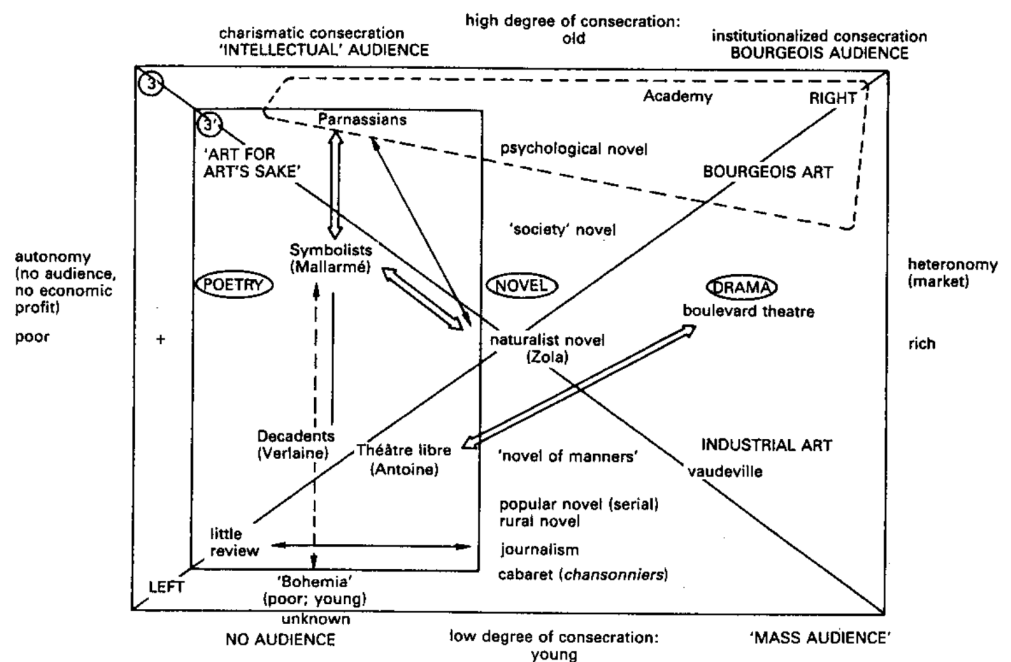
Figure 1: Bourdieu's French literary field of the late 19th century, with audience or popularity on the x-axis and consecration or prestige on the y-axis Bourdieu (1993, 49).

to "map" the literary field is central in this context and has given rise to a string of seminal works on power dynamics in literary cultures (Bennett 1990; Casanova 2007; Guillory 1995; Moretti 2007). Bourdieu's map of the French "literary field" (Figure 1) focuses on literary genres and their interrelation in terms of prestige (and not actors in literary quality judgments *per se*). However, Bourdieu (1993, 46) makes an important distinction between types of audiences and considers "consecration by artists, by institutions of the dominant classes, and by popular success" as distinct axes that are more or less mutually exclusive.[6]

While the relation between these actors is only sketched out (and the present study aims to inspect these more closely), Bourdieu's map can serve as a heuristic conceptualization of types of actors in literary quality judgments. Here, the idea of expert-based and crowd-based literary judgments is apparent at either pole, represented on one side by intellectual and bourgeois audiences, recognized intellectuals such as "Parnassians" and institutions such as *l'Académie Française*, and on the other hand by amateur and mass audience such as the artistic underdogs "Bohemia" and popular media. As Porter (2018, 14) have shown, "on a broad level, real-world data about popularity and prestige appear to confirm Bourdieu's intuitions". In their visualization, the genres "Mystery & Thriller" and "Science Fiction & Fantasy" appear where Bourdieu places the "Popular Novel" (at low consecration and high economic profit), while poetry is in the upper left area of the map, representing high prestige and low popularity. However, the focus of Porter (2018) is on the right-hand part of Bourdieu's map, with prestige defined as institutional or academic consecration: the place of literary works in academia. For a

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

more comprehensive 'map' based on real-world data, various actors, including literary prizes and publishers, should be considered. To this end, the present paper uses a sizeable corpus to examine the interrelation judgments of a type of 'success' in the literary field, including various actors under the general categories of expert-based and crowd-based literary success based on Bourdieu's map. We discuss the selection of various proxies and what they represent before moving on to look at their distribution and interrelation in the Chicago Corpus.

3. Selecting Types of Literary Judgments

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

1. Approaches that seek to approximate literary canonicity or quality in an institutional sense, looking at which works or authors are included in school or university syllabi, literary anthologies, or that win literary awards.

2. Approaches that seek to approximate reader-popularity, basing proxies of literary quality on larger populations, where the selection process appears more 'democratic', seeking the quality perception of 'layman readers' by collecting user-generated data such as ratings from sites like GoodReads, Amazon, or Audible.

3. In-between approaches that seek to measure the market success or market resilience of works, looking at, for example, sales figures.

3.1 Expert-based Quality Proxies

Expert-based proxies for literary quality may, to some extent, be synonymous with canonicity, that is, consecration and institutionalization. Often, quantitative studies of reader appreciation define canonicity or prestige through canon lists compiled by, i.a., individual magazines (Vulture Editors 2018, as in Porter 2018), editors (Karlyn and Keymer 1996, as in Algee-Hewitt et al. 2016), or literary scholars (Bloom 1995, as in Mohseni et al. 2022). However, such lists resemble personal canons that may not have a wide reach, e.g., it is unclear how widely accepted Harold Bloom's chosen canon is among scholars. In this study, we have preferred canonicity proxies that do not depend on the selection of only one critic. To examine expert-based proxies of literary quality and estimate the amount of 'canonic' literature in our dataset, we marked all titles by authors that appear in selected institutional or user-compiled proxies that indicate literary prestige: a literary anthology, the most assigned titles in English Literature course syllabi, literary awards, and a publisher's classics series.

3.1.1 Anthologies

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

	Titles
National book award	108
Pulitzer Prize	53
Nobel Prize*	85
Sci-Fi awards	163
Hugo award	
Nebula award	
Philip K. Dick award	
J.W. Campbell award	
Prometheus award	
Locus Sci-Fi award	
Fantasy awards	40
World Fantasy award	
Locus Fantasy award	
British Fantasy award	
Mythopoeic award	
Romantic awards*	54
Rita awards*	
RNa awards*	
Norton Anthology*	401
Open Syllabus*	477
Penguin Classics Series (titles)	77
Penguin Classics Series*	335
GoodReads' Classics*	62
GoodReads' Best Books of the 20th century*	44
20th century bestsellers (Publishers Weekly)	139
Wikipedia Author-page Rank*	3,558
Translations	5,082
GR avg. rating	8,989
GR rating count	8,989

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

written by authors mentioned in these two series, where the anthology of English literature is the most widespread (Ragen 1992).

3.1.2 Syllabi

While titles assigned on Literature or English syllabi surely vary across colleges and regions, it is possible to find trends and most assigned titles via large collections of data, such as by the *Open Syllabus Project*, which has collected 18.7 million college syllabi in an attempt to map the college curriculum.[7] From this data, we took all titles in our corpus by authors who appear as authors of one of the top 1,000 titles assigned in English Literature college syllabi.

3.1.3 Awards and Longlists

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

---

7. See: `https://www.OpenSyllabus.org`.

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

3.1.4 Classics Series

Various large publishing houses, like Vintage or Penguin[8], maintain a classics series. As Penguin is arguably one of the biggest publishers of Anglophone literature (Alter et al. 2022), we marked all titles or authors in our corpus that appear in their classics series. We looked at both the specific titles (title-based) with matches in our data, and at all titles by authors featured in the series (author-based), keeping these separate in our analysis.

3.2 Crowd-based Quality Proxies

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

3.2.1 GoodReads

GoodReads is a social network or "social catalog site" with links to other social networks (Facebook, Twitter, Instagram, and LinkedIn) designed for readers to discover, review, and share their thoughts. Otis Chandler, GoodReads' co-founder[9], states on the homepage that the idea was to make a social forum akin to looking at the bookshelf at a friend's house: "When I want to know what books to read, I'd rather turn to a friend than any random person or bestseller list". With its 90 million users, GoodReads arguably offers an insight into reading culture "in the wild" (Nakamura 2013, 241), as it catalogs books from a broad spectrum of genres and derives book-ratings from a heterogeneous pool of readers in terms of background, gender, age, native language,

---

8. See: https://www.penguin.com/penguin-classics-overview/.
9. Note that GoodReads was bought by Amazon in 2013.

and reading preferences (Kousha et al. 2017). GoodReads' average ratings represent the average user rating of titles. Ratings range from 0 stars (indicating low appreciation) to 5 stars (indicating high appreciation). The average score provides a general indication of the book's reception, but it is problematic as it conflates types of literary appreciation, i.e., satisfaction, enjoyment, and evaluation, to one scale. While it is important to note that these GoodReads' ratings and number of raters (rating count) do not present an absolute measure of literary quality or even popularity (GoodReads did start with predominantly American users), they do offer a valuable perspective on a work's overall popularity among a diverse population of readers. Beyond ratings, GoodReads also compiles vote-based lists and "shelves", arranged according to the titles most often, either assigned to a particular list or tagged to a particular shelf. These are, for example, "GoodReads' Classics", "Best Books of the 20th Century", "The Worst Books of All Time", etc. For the present study, we used the top 100 of a popular list, the "Best Books of the 20th Century"[10], and a shelf, the "GoodReads' Classics"[11], where titles were *listed* by users 600 to 10,000 times and *shelved* 15,588 to 64,903 times, respectively.

3.2.2 Audible

We use the average rating and number of ratings of titles on Audible, the Amazon audiobook service. Like GoodReads, the site uses a five-star scale for user ratings. However, the number of users and the rating counts are significantly lower for Audible compared to GoodReads: While Dan Brown's *The Da Vinci Code* has 2,259,837 ratings on GoodReads, it has 3,225 ratings on Audible (as of 2024-01-09), and the average Audible rating is inflated in comparison to the GoodReads' average rating for our corpus, which may be an effect of a smaller number of users.

3.3 In-between Quality Proxies

The number of copies sold is often adopted as a reliable standard to estimate the success of novels, for example, to gauge signals that land a book on the bestseller list (Archer and Jockers 2017). It is interesting since a proxy like sales figures appears to occupy a middle ground between the crowd- and expert-based proxies, including a degree of resilience or canonicity of titles (as classics will continue to sell) and popular demand. The NPD BookScan[12], for example, is a popular resource in this regard (as used in Wang et al. 2019), which provides data for the publishing industry regarding genre, prices, and weekly sales figures for all books published in the US since 2003. It is clear that such data is market- and location-specific and is only an option for studies of more contemporary works. As with any other approximation of literary quality, but perhaps especially with sales figures, the issue is that the data pertain to more recent publications, are not readily available, and that contextual factors may influence the data. For book sales, Wang et al. (2019) have shown that marketing, the particular publishing house, and the visibility of the author play a central role in sales numbers.

Instead of sales figures, we may use proxies that also include an aspect of resilience and popular success. Thus, we have used the number of libraries holding a given title on

---

10. See: https://www.GoodReads.com/list/show/6.Best_Books_of_the_20th_Century.
11. See: https://www.GoodReads.com/shelf/show/classics.
12. See: https://www.npd.com/industry-expertise/books/.

Worldcat, the number of translations of a work into other languages, and the author's presence on Wikipedia and a bestseller list. The number of library holdings as a proxy is conceptually intermediate between a completely free, crowd-based vote count and an expert-driven single choice, as the list of books held by libraries depends on both popular demand (from library card holders) and expert choice (from librarians). Similarly, a work's translational success shows a degree of market success (if translation is seen as a token of publishers seeking to expand sales of bestselling books outside the national market) and canonicity or resilience (if translation is seen as a token of a work's cultural longevity or durable popularity). Similarly, Wikipedia Author-page Rank and bestseller lists appear conceptually to include a degree of resilience and popular success.

3.3.1 Library Holdings

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

3.3.2 Translations

The *Index Translationum* database[13] collects all translations published in ca. 150 UNESCO member states, compiled from their local bibliographical institutions or national libraries. It catalogs more than 2 million works across disciplines. Note that the database was created in 1979 and stopped compiling in 2009. Thus, we are not looking at the most translated works through time, where the 'classics' may be more frequent, but at a particular period, and the results should be interpreted with that in mind.

3.3.3 Wikipedia Author-page Rank

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

---

13. See: https://www.unesco.org/xtrans/bsform.aspx.

3.3.4 Bestseller Lists

To gauge the commercial success of titles, we also marked titles in our corpus that were also extant in the *Publishers Weekly* American 20th century bestseller list.[14] *Publishers Weekly* is a trade news magazine which is published once a week (from 1872) and targeted at agents within the field: publishers, literary agents, booksellers, and librarians. While sales numbers are considered, the full set of selection criteria for the list is unknown.

4. Dataset: the Chicago Corpus

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

We base our study on the *Chicago Corpus*[15], a corpus of over 9,000 manually compiled novels that were either written or translated into English and published in the US between 1880 to 2000. The corpus was compiled based on the number of libraries holding a copy of the novel, with a preference for novels with more holdings. Beyond responding to the constraints detailed above, the Chicago Corpus allows us to access the number of libraries holding each title in the US. Moreover, the Chicago Corpus has been curated and used by teams of literary scholars and offers access to the full text of all its titles, which makes a study of correlations between quality judgments and textual features possible in the future.

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

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

We previously discussed the essential characteristics of these proxies of literary quality, as well as the outlook on literary judgments that they seem to model or approximate. Some are at the free and vote-counting end of the spectrum, putting equal weight to each user's rating. Resources like the *Norton Anthology* and prestigious literary awards

---

14. Extracted from the database by John Unsworth at the University of Illinois: https://web.archive.org/web/20111014055658/http://www3.isrl.illinois.edu/~unsworth/courses/bestsellers/picked.books.cgi.
15. While there is no accompanying publication, the corpus can viewed at: https://textual-optics-lab.uchicago.edu/us_novel_corpus. It was compiled by the University of Chicago Textual Optics Lab. For more on the corpus, see the paper presenting the curated resource (Bizzoni et al. 2024).
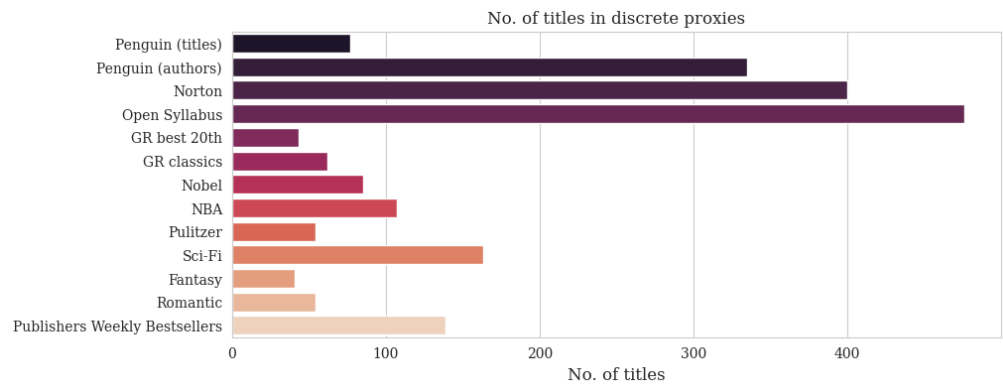
Figure 2: Sizes of discrete proxies in the Chicago Corpus.

arguably fall on the expert-based side of the spectrum, as they are managed by small groups of authoritative readers, usually professional literary critics.

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

In the following sections, we examine the relationship between these proxies, assessing their correlation, their position in a network, and their intersections.

5. Results

5.1 Correlation

Having annotated the titles in our corpus for these proxies, we looked at their correlations to see whether and how they interplay. As some values are discrete and others are not, the correlation matrix is often a measure of overlap: If the correlation coefficient at the intersection of *Penguin Classics* and *Norton Anthology* is a high number, the two proxies have large overlaps. Computing a Spearman or Pearson correlation between two discrete lists means checking whether and to what extent the two lists include the same items (Spearman 2010). Finally, correlations between discrete and continuous values tell us whether there is a sizable change in values when switching from one category to another – for example, whether there is a sizable change in scores between books that were longlisted for a given award and books that were not.[16]

Looking at the correlation matrix resulting from our dataset, we find intriguing correlations between proxies of appreciation. Firstly, we find that there seem to be two

---

16. It is crucial to remember that a correlation between a discrete and a continuous variable is not equivalent to a t-test of significance, as we will discuss later; that is, random samples from the same population could show a valid correlation, and vice versa: Samples from two populations could show no correlation at all.
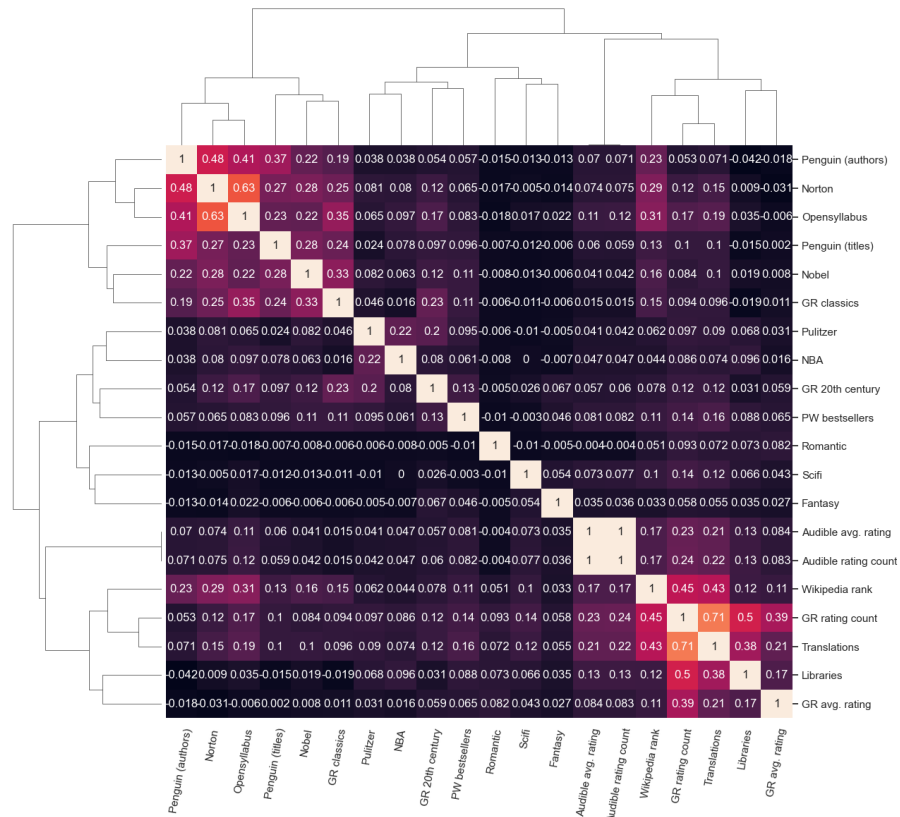
Figure 3: Correlations between discrete and continuous measures of literary quality (Spearman correlation). The matrix shows hierarchical clustering by Ward's method (Ward Jr. 1963).

'islands' of stronger internal correlations: One spans roughly the number of GoodReads and Audible ratings and average ratings along with the library holdings; the other is more or less connecting what we could call 'canon lists' – GoodReads' "Best Books of the 20th Century", GoodReads' "Classics", the Nobel Prize winners, *Open Syllabus*, the *Norton Anthology*, and the *Penguin Classics Series*, and (somewhat surprisingly) the Publishers Weekly bestsellers. Weak correlations happen out of these two areas – Wikipedia Author-page Rank correlates with Sci-Fi awards, but not with the more mediatized Pulitzer Prize, the award which, together with the Nobel Prize, correlates with GoodReads' "Best Books of the 20th Century". However, these do not correlate with each other. Furthermore, the number of ratings of GoodReads and Audible shows correlations with *Open Syllabus*, the *Norton Anthology*, and the *Penguin Classics Series*.

Secondly, suppose we disregard the Nobel Prize, which correlates with "canon" proxies such as the *Open Syllabus*. In that case, the awards do not overlap much and do not display strong correlations with other categories. Beyond the mentioned correlations of the Pulitzer and Nobel with the GoodReads' list of Best Books of the 20th century, awards, especially genre awards, do not appear to correlate with other proxies. This lack of correlation is relevant, especially as it means that longlisted works of genre literature appear to have no strong presence in resources like the *Norton Anthology* or the GoodReads' Classics list, indicating the strong presence of general fiction in these resources. However, it is still possible that the awards elicit a particular range of ratings in terms of GoodReads' ratings or library holdings without eliciting a detectable correlation. Also, not surprisingly, genre-fiction awards do not overlap with more

literary awards (such as the Pulitzer, National Book Award, and the Nobel). At the same time, the Pulitzer Prize and the National Book Award do converge. The awards for Romantic fiction and Fantasy are the most removed, showing little convergence with other proxies.

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

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

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

Thus, no clear distinction between these two clusters can be based on the method of selection (expert-based versus crowd-based), but rather on the form of perceived canonicity or literariness that distinguishes the second group from the first. In other words, what we see might be two different 'faces' of the concept of literary quality perceived by the same reader. An observation supporting the idea that there should be two main 'perceptions' of quality is that GoodReads users do not seem to give the highest ratings to the titles of the *Norton Anthology*. Still, when GoodReads users constitute lists of "classics" and "20th century best", they converge with the anthology on similar ground.

5.2 Network

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

Indeed, the interrelation between different proxies may be complex to gauge when looking at correlation coefficients and visualizations. Proxy interrelations are better visualized in the literary quality standard landscape when visualized as a network,
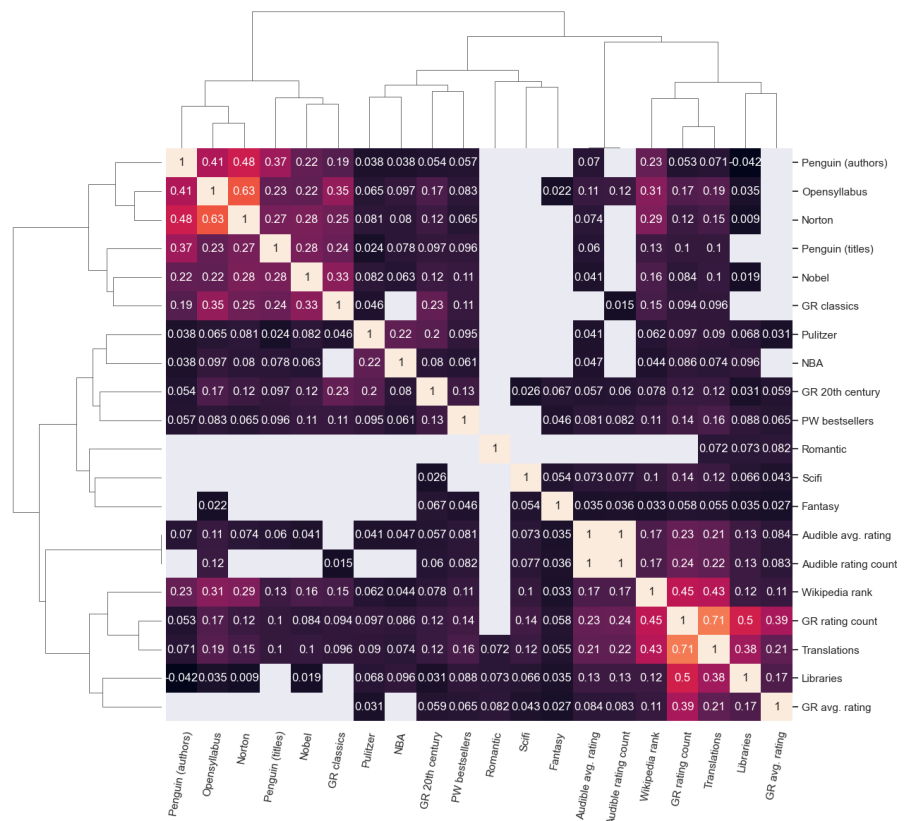
Figure 4: Again, correlations between discrete and continuous measures of literary quality (Spearman correlation), this time with non-significant correlations masked (*p*-value < 0.05).
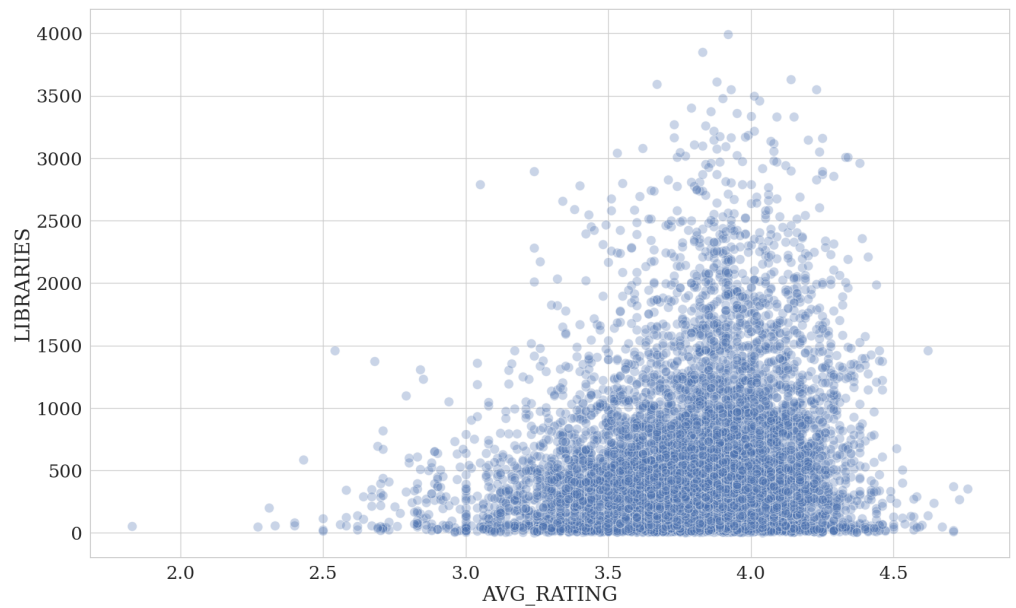

Figure 5: Scatterplot of library holdings vs. average rating of all titles with a threshold of five ratings.
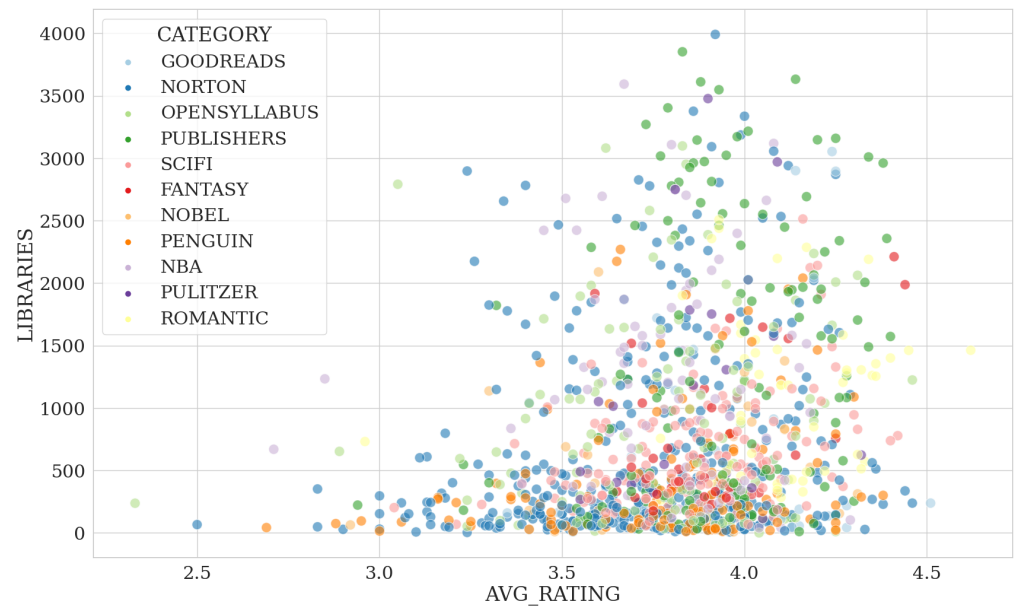
Figure 6: Scatterplot of library holdings vs. average rating of titles contained in one of the quality proxies.

where each node represents one proxy and each edge the correlation (i.e., for discrete lists, the overlap) between proxies.

As was also apparent in the correlation matrix (Figure 3), longlists of genre-fiction awards tend to be far removed from other proxies, with a slight correlation between Fantasy and Sci-Fi awards, which might be explained by the thematic overlap between these genres. The disconnection between more "literary proxies" like the *Penguin Classics Series* and the *Norton Anthology* may also be affected by the relabelling of genre-fiction in literary markets. Genre tags may act as implicit quality judgments themselves: Prestigious horror is often relabeled "gothic" or "literary fiction" and doesn't even run for genre awards (think of, i.a., Bram Stoker and Mary Shelley). Genre labeling is a complex issue, involving various cultural factors and market forces. For example, works by women authors are often labeled or relabeled in less prestigious genres, such as 'Romantic fiction' over 'literary novel' (Groos 2000).

In our network, books listed in the *Index Translationum* show a strong correlation with authors in our Wikipedia Author-page Rank data, and also have a large actual overlap: 52.7 percent of translated books are books by authors in our Wikipedia Author-page Rank data, and 75.3 percent of books by authors in our Wikipedia Author-page Rank data are also in the *Index Translationum* list of translated works.

While the literary awards, the National Book Award and the Pulitzer, show some overlap, the cluster of most related proxies seems to be the more expert-based type of proxy: especially *Open Syllabus*, *Norton Anthology*, and the *Penguin Classics Series* form a distinct triangle in the network. Books that are in one of these three proxies also tend to be in the other, which is particularly interesting in this case because the underlying selection mechanisms of these three proxies seem distinct, split between institutional and commercial affiliations. Nevertheless, their selection still converges on a common perception of the quality of titles. Furthermore, the divergence of awards from the remaining proxies, as well as the divergence between general (National Book Award, Pulitzer) and genre
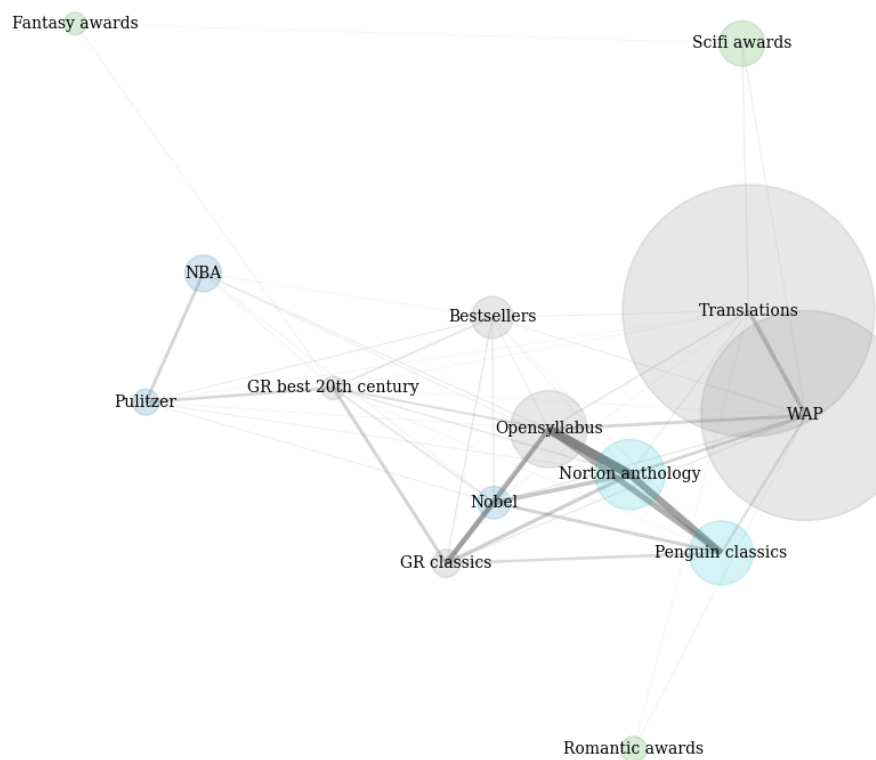
Figure 7: Network of literary quality proxies with edge-width and opacity based on the correlation coefficient between proxies (Spearman correlation), excepting the corpus-wide categories of GoodReads ratings. We apply a coefficient threshold of 0.05 for edges being visualized. Positions are likewise determined by correlation between proxies, using the Fruchterman-Reingold force-directed algorithm for positioning (Fruchterman and Reingold 1991). The sizes of the nodes are determined by the number of titles in each proxy. Colors are used to indicate similar types of awards: literary awards, genre-fiction awards, book series/anthologies.

	GR avg. rating	GR rating count	Library holdings	Translations	Wikipedia Author-page Rank
Corpus average	3.75	14,246.36	535.74	6.58	0.000058
Open Syllabus	3.78	109,831.81	738.05	25.22	0.000423
Penguin Classics (authors)	3.72	57,105.42	463.54	16.18	0.000334
Penguin Classics (titles)	3.76	194,615.08	496.74	43.14	0.000418
Norton Anthology	3.74	74,424.81	687.75	22.09	0.000402
GoodReads Classics	3.82	4,307,090.65	501.37	57.11	**0.000869**
GoodReads Best Books of the 20th Century	4.04	**992,225.89**	998.41	**98.02**	0.000439
Nobel Prize	3.81	119,078.32	811.09	32.04	0.000558
National Book Award	3.83	62,071.08	1,266.10	17.28	0.000111
Pulitzer Prize	3.91	135,290.26	**1,498.77**	33.98	0.000176
Sci-Fi awards	3.88	73,716.60	701.81	13.81	0.000135
Fantasy awards	3.92	164,753.12	804.28	18.27	0.000158
Romantic awards	**4.09**	31,595.07	1,078.24	11.69	0.000037
Bestsellers	3.94	120,453.92	1,290.56	43.03	0.000222

Table 2: Intersectional values: mean continuous quality-measure per discontinuous proxies. Bold font indicates the highest mean within the selection of proxies. Note that the Wikipedia Author-page Rank has been multiplied by 100 because of the generally low values. The abbreviation 'GR' stands for 'GoodReads'.

fiction award types is even more apparent in the network, while the Nobel Prize shows stronger convergences with the aforementioned triad of more canonical, expert-based proxies, indicating its difference from the other prestigious awards.

5.3 Intersection

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

When visualizing the distribution of titles of different categorical proxies in terms of our continuous proxies (rating count, translations, etc.), we see that titles included in categorical quality proxies generally have a longer tail and may have different distributions than titles not contained in any quality proxy ("None" in Figure 8). Looking at the GoodReads average rating and library holdings, books included in categorical proxies seem to have smoother slopes in comparison to the rest of the corpus ("None"). In contrast, in terms of rating count, Wikipedia Author-page Rank, and translations, we see a much larger number of titles in each proxy with very low values, with a long tail of few outliers at very high values. Measures such as rating count tend to exhibit a logarithmic distribution.

Moreover, different categorical proxies peak at different values within the continuous proxies. For example, the distribution of books that have won a Romantic literary award seems to peak at a higher value of the GoodReads average rating, having also the highest
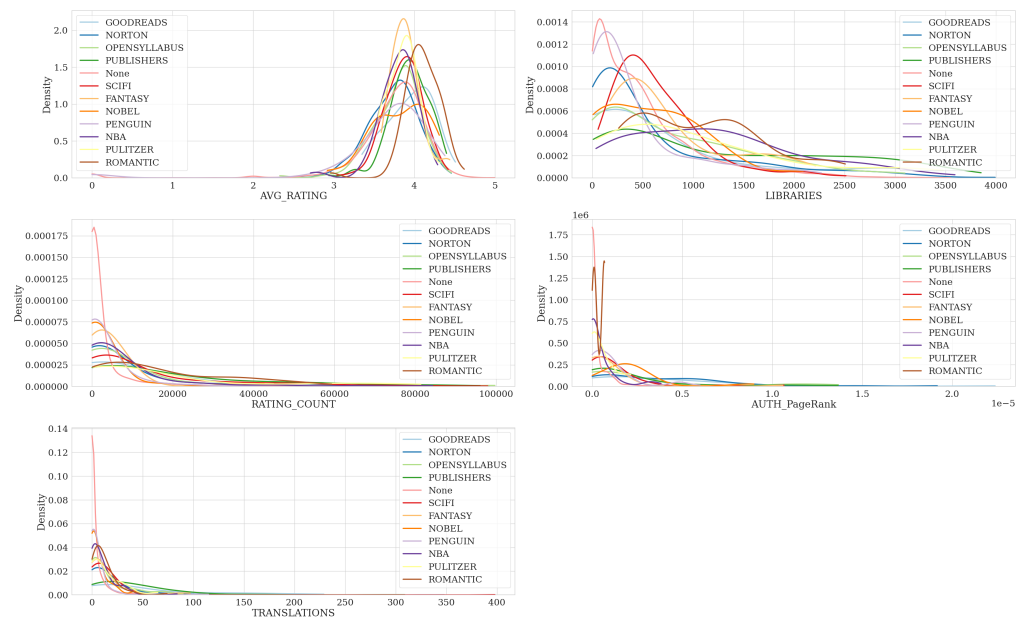
Figure 8: Kernel density estimate (KDE) plots of the distributions of measures per quality proxy. Note that rating count values above 100,000 have been filtered out for the purpose of visualization. "None" represents titles that are not in either of the proxies.

mean average rating of all the proxies (Table 2).[17] Titles in GoodReads Classics, Nobel Prize, *Open Syllabus*, and *Norton Anthology* are represented more evenly across values of Wikipedia Author-page Rank, which may be expected as we also saw that these proxies seem to be closely related in our network (Figure 7). This indicates that these base their selection on some shared perception of quality, which may also prompt their authors to have more prominent Wikipedia pages. Interestingly, the plot showing distributions over library holdings shows a somewhat opposite tendency: here, genre-fiction tends to place at higher values, so Sci-Fi, Fantasy, and Romantic fiction, for example, peak at higher values and have high mean library holdings numbers (Table 2). In general, the two quality 'islands' detected in our correlation matrix (Figure 3) can be observed in the colors that peak in the different quadrants, genre fiction in some, what we might call 'higher brow' or canonical literature in others.

Visualizing the mean values of each discrete proxy in terms of continuous proxies further aids in gauging the differences between these quality perspectives (see Figure 9, Figure 10, Figure 11, Figure 12, Figure 13).

GoodReads "Best Books of the 20th Century" appear to have the highest average GoodReads ratings, closely followed by Hugo and Pulitzer titles. In contrast, the *Norton* and *Open Syllabus* titles record the lowest average ratings (Figure 9). Overall, *Open Syllabus* and *Norton Anthology* titles score consistently lower than any other category regarding their GoodReads average ratings and number of libraries holdings (Figure 10).

GoodReads "Best Books of the 20th Century" is the only proxy that stands out in terms of GoodReads rating *count* (Figure 11). Note that rating count is a problematic proxy because of its non-normal distribution, with very few titles at very high values, which

---

17. Note that the odd distribution of Romantic titles in the plots with library holdings and Wikipedia Author-page Rank may be an effect of the small number of titles. One author with higher canonicity may be responsible for the peak at the higher end in both plots.
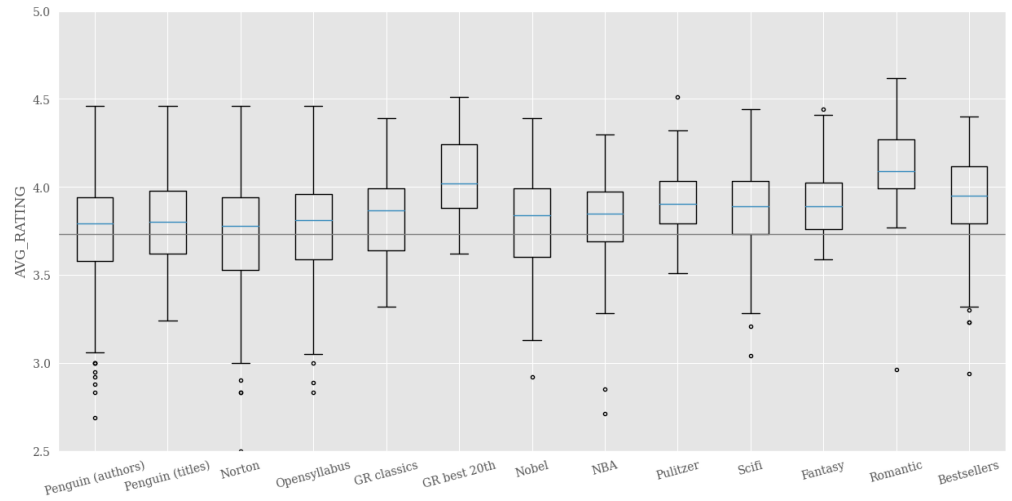
Figure 9: Boxplot of average GoodReads rating for discrete categories. The grey line indicates the corpus average rating.
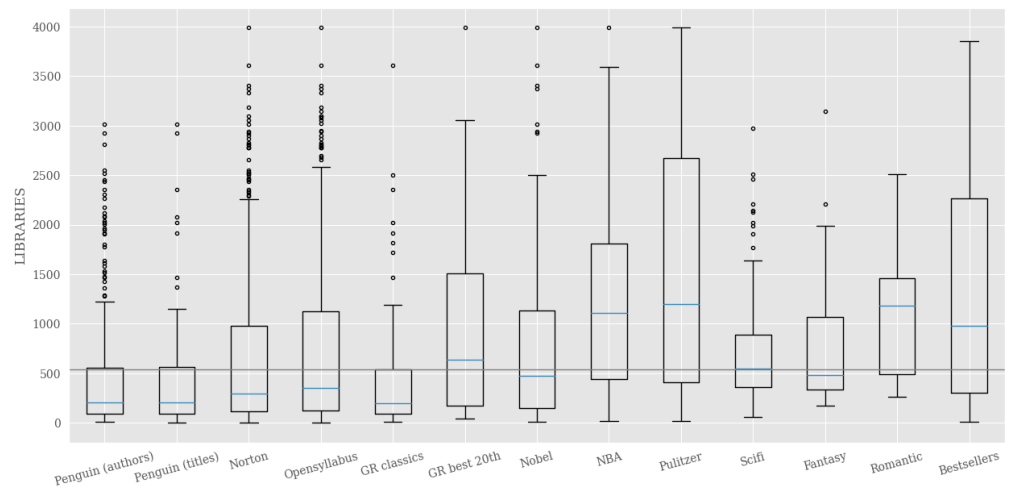

Figure 10: Boxplot of average number of library holdings for discrete categories. The grey line indicates the corpus average holdings.
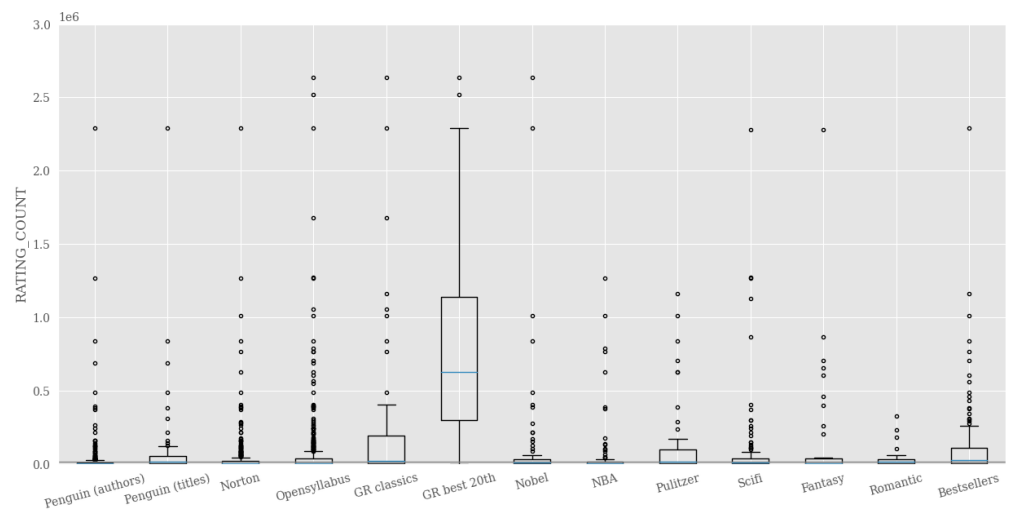

Figure 11: Boxplot of rating count of discrete categories. The grey line indicates the corpus average rank.
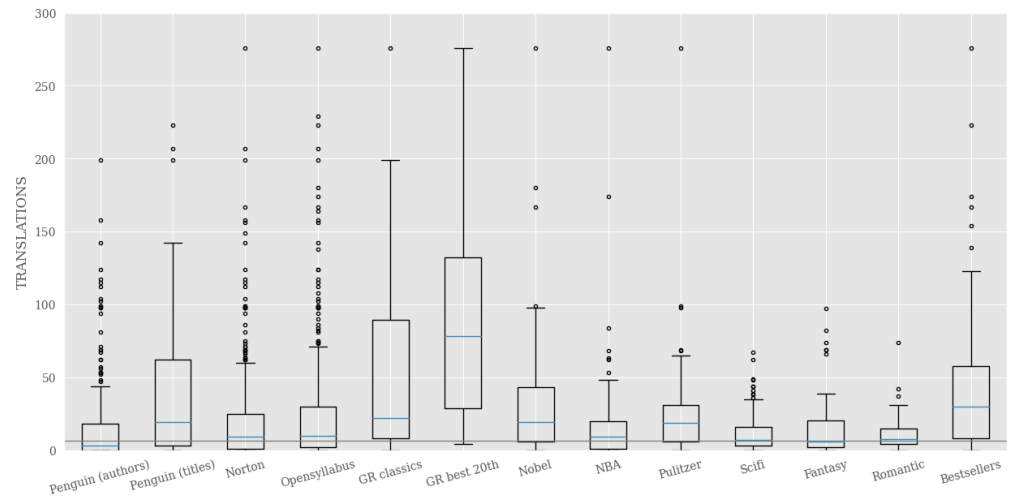
Figure 12: Boxplot of average translation numbers for discrete categories. The grey line indicates the corpus average number.
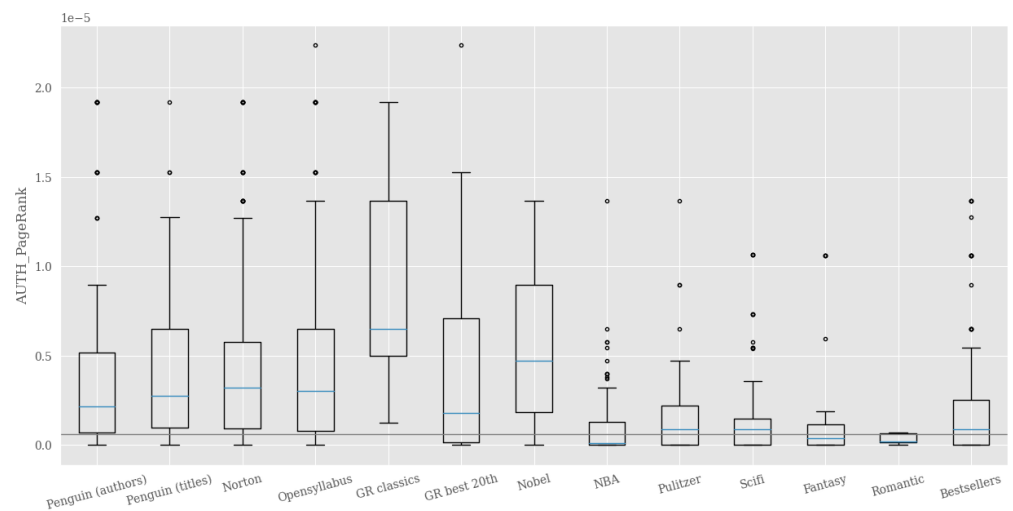

Figure 13: Boxplot of average Wikipedia Author-page Rank for discrete categories. The grey line indicates the corpus average rank.

is why we see a low corpus mean with many outliers for each proxy as well as long whiskers for the GoodReads "Best Books of the 20th Century" category.

Translation numbers and Wikipedia Author-page Rank are the two continuous measures that appear similar in that titles longlisted for awards tend to score low compared to, for example, GoodReads Classics titles. Again, there is a difference between general fiction awards (National Book Award, Pulitzer) and genre fiction awards, where titles longlisted for genre fiction awards tend to rank lower. It is interesting that for these two plots (see Figure 12 and Figure 13), the user-generated lists of GoodReads Classics and Best Books of the 20th Century score high, with a subtle difference between the two plots. When looking at translation numbers, we see that GoodReads Best 20th Century Books score higher than GoodReads Classics and that bestsellers are also one of the proxies with higher mean translation numbers. Conversely, when looking at the Wikipedia Author-page Rank, we see that GoodReads Classics have a higher mean than the Best 20th Century Books and that the Nobel titles, as well as the more expert-based measures

that showed the strongest affinities in our network (Figure 7), also have a higher mean in comparison to when looking at translation numbers. Considering each of these boxplots together, the following patterns emerge:

1. Titles longlisted for awards, both general fiction and genre awards, tend to have higher average GoodReads ratings and library holdings.

2. The proxies we found to be strongly correlated in the 'island' of our correlation matrix representing more 'canonical' fiction (Figure 3), *Open Syllabus*, *Norton*, and *GoodReads Classics* tend to have lower average GoodReads ratings and library holdings.

3. There is a partial convergence between vote-based continuous scores and discrete categories. While translation numbers and Wikipedia Author-page Rank seem to ascribe higher values to more 'canonical' fiction, GoodReads users and library holdings seem to have a higher appreciation for awards and genre fiction and a lower appreciation for the canon.

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

Looking at the different types of awards, we seem to confirm Bourdieu's intuition that the literary field is polarized: Our genre-award proxies appear far removed from other proxies (including more general literary awards, see Figure 7). Yet they have higher average GoodReads ratings and library holdings than, for example, the more institutionally oriented *Norton Anthology*. These characteristics would situate titles of genre awards roughly at the place of the "popular novel" in Bourdieu's map of the literary field, which also aligns with the study of the prestige versus popularity of genre fiction by Porter (2018). In contrast, a proxy like the *Norton Anthology* may be situated more toward the 'intellectual' and 'bourgeois' poles of Bourdieu's map, considering it is part of the interlinked triangle of proxies observed in our network (see Figure 7) of which *Open Syllabus* has an institutional status. The apparent divergence between proxies like the *Norton Anthology* and genre fiction awards may be explained by differences in style and topic of books. Still, studies have also suggested that different types of audiences appreciate books at varying levels of readability (Bizzoni et al. 2023). Thus, the divergence may also have to do with socio-cultural factors like population literacy, where more 'readable' works are preferred at the level of a larger audience, and more institutionally acclaimed works, such as those included in the *Norton Anthology*, less so, partly because of difficulty at the sentence level.

Following Bourdieu, we might contrast actors behind the general fiction award proxies as "intellectual audiences" against those behind genre fiction awards as a "mass audience" (Figure 1). However, it is important to note that we do not find audiences to be as polarized or distinct as Bourdieu suggested. Rather, proxies seem to transverse their actor type affiliations. For instance, while bestsellers and *Open Syllabus* have dissimilar

actors underlying them – institutional versus market-oriented – bestsellers had the strongest correlation with *Open Syllabus*, as seen in Figure 3. These findings imply the potential existence of two overarching types of 'quality perception', which overlay and interlink proxies underpinned by divergent actors or audiences. This insight emerges from the observation of two 'islands' when looking at the correlations (Figure 3), but also from looking at the differential favoring of each of the continuous measures contained in the first 'island'. When exploring the discrete proxies in terms of the continuous ones, we saw that GoodReads ratings and library holdings on one side, and translation numbers and Wikipedia Author-page Rank on the other, were more similar in the way they evaluate, for example, longlisted titles for genre awards. This suggests that actor-based or audience-based distinctions might not fully capture the intricate dynamics of appreciation judgments in the literary field.

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

6. Conclusion and Future Work

Generally, we seem to observe two types of 'quality perception' – or two faces of the concept of quality – that emerge from the differences and surprising convergences of the host of proxies considered in this study.

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

This duality is reminiscent of several similar dichotomies theorized in previous works: C. Koolen et al. (2020)'s distinction of literariness and enjoyability, Porter (2018) and Manshel et al. (2019)'s distinction between prestige and popularity, and naturally of Bourdieu (1993)'s two axes of institutionalized vs. popular art. Yet, the duality that emerges from our data is nuanced and does not represent a polar opposition but rather

fuzzy islands between different proxies. Bestseller lists agree with canonical groups and with GoodReads metrics, and the distinctness of titles included in longlists for genre awards might even indicate a possible third – or many – different perceptions of quality, which may be connected to various extra- and intra-textual features.

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

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

Various limitations inhere to the selection of quality proxies and the quality proxies themselves, and it should be noted that various other proxies could be collected, among others, sales figures. Moreover, different literary cultures may vary in their ways of assessing quality, while this study is situated in an Anglophone and American context. In terms of challenges in evaluating the quality proxies themselves, for example, GoodReads may represent a contemporary audience so that canonical literature, assessed over decades or centuries, does not precisely align with their tastes. In future studies, we suggest a closer inspection of possible biases, such as the publication dates of titles and gender or race biases influencing literary judgments. We also suggest a stronger focus on the interplay between textual features and different types of quality proxies. For example, assessing the importance of readability for different types of proxies, which is an often underrated metric that is likely to explain, among other things, the decline of certain avant-garde works over time, as well as differences in preferences between types of audiences.

7. Data Availability

Data can be found here: https://github.com/centre-for-humanities-computing/chicago_corpus.

8. Software Availability

All code created and used in this research has been published at: `https://github.com/PascaleFMoreira/measuring_literary_quality`. It has been archived and is persistently available at: `https://doi.org/10.5281/zenodo.13960503`.

9. Author Contributions

Pascale Feldkamp: Formal analysis, Writing – review & editing

Yuri Bizzoni: Formal analysis, Writing – review & editing

Mads Rosendahl Thomsen: Methodology, Writing – review & editing, Project administration

Kristoffer L. Nielbo: Methodology, Project administration

References

Algee-Hewitt, Mark, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser (2016). "Canon/Archive: Large-scale Dynamics in the Literary Field". In: *Pamphlets of the Stanford Literary Lab* 11. `https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf` (visited on 10/16/2024).

Alter, Alexandra, Elizabeth A. Harris, and David McCabe (2022). "Will the Biggest Publisher in the United States Get Even Bigger?" In: *The New York Times*. `https://www.nytimes.com/2022/07/31/books/penguin-random-house-simon-schuster-antitrust-trial.html` (visited on 09/13/2024).

Archer, Jodie and Matthew L. Jockers (2017). *The Bestseller Code*. Penguin Books.

Bennett, Tony (1990). *Popular Fiction: Technology, Ideology, Production, Reading*. Routledge.

Bizzoni, Yuri, Pascale Moreira, Nicole Dwenger, Ida Lassen, Mads Thomsen, and Kristoffer Nielbo (2023). "Good Reads and Easy Novels: Readability and Literary Quality in a Corpus of US-published Fiction". In: *Proceedings of the 24th Nordic Conference on Computational Linguistics* (*NoDaLiDa*), 42–51. `https://aclanthology.org/2023.nodalida-1.5` (visited on 09/13/2024).

Bizzoni, Yuri, Pascale Feldkamp Moreira, Ida Marie S. Lassen, Mads Rosendahl Thomsen, and Kristoffer Nielbo (2024). "A Matter of Perspective: Building a Multi-Perspective Annotated Dataset for the Study of Literary Quality". In: *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (*LREC-COLING 2024*), 789–800. `https://aclanthology.org/2024.lrec-main.71` (visited on 10/16/2024).

Bizzoni, Yuri, Telma Peura, Kristoffer Nielbo, and Mads Thomsen (2022a). "Fractal Sentiments and Fairy Tales-Fractal Scaling of Narrative Arcs as Predictor of the Perceived Quality of Andersen's Fairy Tales". In: *Journal of Data Mining & Digital Humanities*. `10.46298/jdmdh.9154`.

— (2022b). "Fractality of Sentiment Arcs for Literary Quality Assessment: The Case of Nobel Laureates". In: *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, 31–41. `https://aclanthology.org/2022.nlp4dh-1.5` (visited on 01/17/2023).

Bizzoni, Yuri, Telma Peura, Mads Rosendahl Thomsen, and Kristoffer Nielbo (2021). "Sentiment Dynamics of Success: Fractal Scaling of Story Arcs Predicts Reader Preferences". In: *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*. NLP Association of India (NLPAI), 1–6. `https://aclanthology.org/2021.nlp4dh-1.1` (visited on 01/17/2023).

Bjerck Hagen, Eric, Christine Hamm, Frode Helmich Pedersen, Jørgen Magnus Sejersted, and Eirik Vassenden (2018). "Literary Quality: Historical Perspectives". In: *Contested Qualities*. Ed. by Knut Ove Eliassen, Jan Hovden, and Øyvind Prytz. Fagbokforlaget, 47–74.

Bloom, Harold (1995). *The Western Canon: The Books and School of the Ages*. First Riverhead Edition. Riverhead Books.

Bourdieu, Pierre (1993). *The Field of Cultural Production: Essays on Art and Literature*. Ed. by Randal Johnson. Columbia University Press.

Casanova, Pascale (2007). *The World Republic of Letters*. Harvard University Press.

Febres, Gerardo and Klaus Jaffe (2017). "Quantifying Literature Quality Using Complexity Criteria". In: *Journal of Quantitative Linguistics* 24 (1), 16–53. `10.1080/09296174.2016.1169847`.

Fruchterman, Thomas M. J. and Edward M. Reingold (1991). "Graph Drawing by Force-directed Placement". In: *Software: Practice and Experience* 21 (11), 1129–1164. `10.1002/spe.4380211102`.

Groos, Marije (2000). "Wie schrijft die blijft? Schrijfsters in de literaire kritiek van nu". In: *Tijdschrift voor Genderstudies* 3 (3).

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

Harrison, Chloe and Louise Nuttall (2018). "Re-reading in Stylistics". In: *Language and Literature* 27 (3), 176–195. `10.1177/0963947018792719`.

Hube, Christoph, Frank Fischer, Robert Jäschke, Gerhard Lauer, and Mads Rosendahl Thomsen (2017). "World Literature According to Wikipedia: Introduction to a DBpedia-Based Framework". In: *arXiv*. `10.48550/arXiv.1701.00991`.

Jannatus Saba, Syeda, Biddut Sarker Bijoy, Henry Gorelick, Sabir Ismail, Md Saiful Islam, and Mohammad Ruhul Amin (2021). "A Study on Using Semantic Word Associations to Predict the Success of a Novel". In: *Proceedings of SEM 2021: The 10th Joint Conference on Lexical and Computational Semantics*, 38–51. `10.18653/v1/2021.starsem-1.4`.

Jockers, Matthew L. (2015). "Syuzhet: Extract Sentiment and Plot Arcs from Text". In: *Matthew L. Jockers Blog*. `https://www.matthewjockers.net/2015/02/02/syuzhet/` (visited on 10/16/2024).

Karlyn, Danny and Tom Keymer (1996). *Chadwyck-Healey Literature Collection*. `http://collections.chadwyck.com/marketing/products/about_ilc.jsp?collection=ncf` (visited on 09/13/2024).

Koolen, Corina, Karina van Dalen-Oskam, Andreas van Cranenburgh, and Erica Nagelhout (2020). "Literary Quality in the Eye of the Dutch Reader: The National Reader Survey". In: *Poetics* 79, 1–13. `10.1016/j.poetic.2020.101439`.

Koolen, Cornelia Wilhelmina (2018). *Reading Beyond the Female: The Relationship between Perception of Author Gender and Literary Quality*. ILLC Dissertation Series. Institute for Logic, Language and Computation, Universiteit van Amsterdam. `https://eprints.illc.uva.nl/id/eprint/2152/1/DS-2018-03.text.pdf` (visited on 10/16/2024).

Kousha, Kayvan, Mike Thelwall, and Mahshid Abdoli (2017). "GoodReads Reviews to Assess the Wider Impacts of Books". In: *Journal of the Association for Information Science and Technology* 68 (8), 2004–2016. `10.1002/asi.23805`.

Kovács, Balázs and Amanda J. Sharkey (2014). "The Paradox of Publicity". In: *Administrative Science Quarterly* 1, 1–33. `10.1177/0001839214523602`.

Kuijpers, Moniek M. and Frank Hakemulder (2018). "Understanding and Appreciating Literary Texts Through Rereading". In: *Discourse Processes* 55 (7), 619–641. `10.1080/0163853X.2017.1390352`.

Lassen, Ida Marie Schytt, Yuri Bizzoni, Telma Peura, Mads Rosendahl Thomsen, and Kristoffer Laigaard Nielbo (2022). "Reviewer Preferences and Gender Disparities in Aesthetic Judgments". In: *CEUR Workshop Proceedings*, 280–290. `10.48550/arXiv.2206.08697`.

Maharjan, Suraj, 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*, 1217–1227. `https://aclanthology.org/E17-1114` (visited on 10/16/2024).

Maharjan, Suraj, 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*, 259–265. `10.18653/v1/N18-2042`.

Manshel, Alexander, Laura B. McGrath, and J. D. Porter (2019). *Who Cares about Literary Prizes?* `https://www.publicbooks.org/who-cares-about-literary-prizes/` (visited on 04/20/2023).

Mohseni, Mahdi, Christoph Redies, and Volker Gast (2022). "Approximate Entropy in Canonical and Non-Canonical Fiction". In: *Entropy* 24 (2), 278. `10.3390/e24020278`.

Moretti, Franco (2007). *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso.

Nakamura, Lisa (2013). "'Words with Friends': Socially Networked Reading on GoodReads". In: *PMLA* 128 (1), 238–243. `10.1632/pmla.2013.128.1.238`.

Pope, Colin (2019). "We Need to Talk About the Canon: Demographics in 'The Norton Anthology'". In: *The Millions*. `https://themillions.com/2019/04/we-need-to-talk-about-canons-picturing-writerly-demographics-in-the-norton-anthology-of-american-literature.html` (visited on 09/13/2024).

Porter, J. D. (2018). "Popularity/Prestige: A New Canon". In: *Pamphlets of the Stanford Literary Lab* 17. `https://litlab.stanford.edu/LiteraryLabPamphlet17.pdf` (visited on 09/13/2024).

Ragen, Brian Abel (1992). "An Uncanonical Classic: The Politics of the 'Norton Anthology'". In: *Christianity and Literature* 41 (4), 471–479. `https://www.jstor.org/stable/44312103` (visited on 10/16/2024).

Reagan, Andrew J., Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds (2016). "The Emotional Arcs of Stories Are Dominated by Six Basic Shapes". In: *EPJ Data Science* 5 (1), 1–12. `10.1140/epjds/s13688-016-0093-1`.

Riddell, Allen and Karina van Dalen-Oskam (2018). "Readers and Their Roles: Evidence from Readers of Contemporary Fiction in the Netherlands". In: *PLOS ONE* 13 (7). Ed. by K. Brad Wray. `10.1371/journal.pone.0201157`.

Shesgreen, Sean (2009). "Canonizing the Canonizer: A Short History of 'The Norton Anthology of English Literature'". In: *Critical Inquiry* 35 (2), 293–318. `10.1086/5966` `44`.

Spearman, Charles (2010). "The Proof and Measurement of Association between Two Things". In: *International Journal of Epidemiology* 39 (5), 1137–1150. `10.1093/ije/dyq` `191`.

Trotsky, Leon (1974). *Class and Art: Problems of Culture under the Dictatorship of the Proletariat*. New Park.

van Cranenburgh, Andreas 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*, 1228–1238. `https://aclantholo` `gy.org/E17-1115` (visited on 10/16/2024).

van Dalen-Oskam, Karina (2023). *The Riddle of Literary Quality*. Amsterdam University Press.

van Peer, Willie (2008). "Canon Formation: Ideology or Aesthetic Quality?" In: *The Quality of Literature: Linguistic Studies in Literary Evaluation*. Ed. by Willie van Peer. John Benjamins Publishing, 17–29. `10.1075/lal.4`.

Veleski, Stefan (2020). "Weak Negative Correlation between the Present Day Popularity and the Mean Emotional Valence of Late Victorian Novels". In: *Proceedings of the Workshop on Computational Humanities Research*, 32–43. `http://ceur-ws.org/Vol-27` `23/long44.pdf` (visited on 10/16/2024).

Vulture Editors (2018). "A Premature Attempt at the 21st Century Literary Canon". In: *Vulture*. `https://www.vulture.com/article/best-books-21st-century-so-far.h` `tml` (visited on 10/16/2024).

Walsh, Melanie and Maria Antoniak (2021). "The Goodreads 'Classics': A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism". In: *Post45*. `https://post45.org/2021/04/the-goodreads-classics-a-computational-s` `tudy-of-readers-amazon-and-crowdsourced-amateur-criticism/` (visited on 10/16/2024).

Wang, Xindi, Burcu Yucesoy, Onur Varol, Tina Eliassi-Rad, and Albert-László Barabási (2019). "Success in Books: Predicting Book Sales before Publication". In: *EPJ Data Science* 8 (1). `10.1140/epjds/s13688-019-0208-6`.

Ward Jr., Joe H. (1963). "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58 (301), 236–244. `10.1080/01621459.19` `63.10500845`.

Wellek, René (1972). "The Attack on Literature". In: *The American Scholar* 42 (1), 27–42. `https://www.jstor.org/stable/41207073` (visited on 09/13/2024).