Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, pages 20–25
Marseille, 20 June 2022

Predicting Literary Quality How Perspectivist Should We Be?

Y. Bizzoni, I.M. Lassen, T. Peura, M.R. Thomsen, K.L. Nielbo
Center for Humanities Computing Aarhus, Aarhus University
yuri.bizzoni@cc.au.dk, idamarie@cas.au.dk, tpeura@cc.au.dk, madsrt@cc.au.dk, kln@cas.au.dk

Abstract

Approaches in literary quality tend to belong to two main grounds: one sees quality as completely subjective, relying on the idiosyncratic nature of individual perspectives on the perception of beauty; the other is ground-truth inspired, and attempts to find one or two values that predict something like an objective quality: the number of copies sold, for example, or the winning of a prestigious prize. While the first school usually does not try to predict quality at all, the second relies on a single majority vote in one form or another. In this article we discuss the advantages and limitations of these schools of thought and describe a different approach to reader's quality judgments, which moves away from raw majority vote, but does try to create intermediate classes or groups of annotators. Drawing on previous works we describe the benefits and drawbacks of building similar annotation classes. Finally we share early results from a large corpus of literary reviews for an insight into which classes of readers might make most sense when dealing with the appreciation of literary quality.

Keywords: Quality assessment, Perspectivism, Literary quality

1. Introduction and Motivation

While literary quality can be considered one of the most subjective fields of evaluation, its perception from large amounts of readers over time does show convergent trends: communities tend to establish and update canons; specific texts and narratives manage to remain popular despite the changing of fashions and political phases; authors' names become eponymous of literary quality in different countries and throughout the social spectrum. This duality has arguably generated two opposed polarities when it comes to the definition of what constitutes literary quality: on one side a highly individualistic, idiosyncratic perspective, that sees quality as a function of either individual or collective, but temporary world views, and as such non-convergent if not for ephemeral artefacts, such as transitory canons (Bloom, 2014). On the other side, a ground-truth inspired perspective, that sees literary value as a sort of universal and underlying quality of texts, that shines through the noise of socio-political or individual differences into broad or long-lasting convergences (Guillory, 2013).

The problem of literary quality's subjective status becomes even more intriguing when we turn to the challenge of its formal or computational assessment. Most works in this direction have, until today, assumed the possibility of one single ground truth by modelling literary quality as a single rating or label assigned to a text. These ratings have been retrieved from various sources: literary critics, book sale numbers, bestseller lists, or crowd-sourced reader opinions. Such approaches have had their limitations. Relying only on experts' judgment (e.g. awards, prestigious reviews) would bias the model to reflect only their preferences, but striving for representativity by crowdsourcing opinions ends up ignoring important differences in the readers' population, as we will discuss in the next sections.

In this paper, we follow the tracks of recent debates in computational linguistics and machine learning about the advantages and limitations of considering different perspectives, what is called "perspectivism" (Basile et al., 2021; Plank et al., 2014). With this in mind, we address the question of how perspectivist we should be when it comes to literary quality. After drawing a spectrum with total subjectivity on one end and the use of a single gold standard on the other, we suggest approaching a middle way, by dividing readers into meaningful classes, each of which can be treated as a single judgment on a literary work. Finally, we present early results on a new corpus of literary reviews validating the feasibility of this approach.

2. Related Work

Several studies have attempted to formally model traits that capture literary quality. The choice of the candidate features for the definition of literary quality has naturally been very broad: some approaches, conflating quality and fame, have focused only on extra-textual features, such as genre and author visibility to predict success in book sales (Wang et al., 2019), or the number of references to a literary work as a measure of canonicality (Ferrer, 2013), whereas others have focused on stylistic features [1], such as syntactic (van Cranenburgh et al., 2019) and semantic (Ashok et al., 2013) complexity, or the emotional flow of a narrative (Maharjan et al., 2018) to predict, for example, the likelihood of a text to become part of a pre-determined literary canon.

[1] For a review of computational stylistics, see Hermann et al. (2020)

What is often less discussed in many of these studies is the problem of defining a ground truth for literary quality: most of the existing literature in automatic quality prediction of narrative texts relies on a single gold standard, adopting what is still today the mainstream approach to machine learning (Basile et al., 2021). Some works, for example, use the Nobel Prize for Literature as a way to assess the quality of an author (Hu et al., 2021), while others draw the average rating for a book from a large-scale reader platform as a ground truth for a text's appreciation (Bizzoni et al., 2021; Maharjan et al., 2018). The number of copies sold is often adopted as a reliable golden label to rank novels, based on the assumption that there is a distinct, overarching set of signals that has predictive power for whether or not a book ends up on the bestseller list (Archer and Jockers, 2016; Wang et al., 2019). Finally, some works have attempted to use the guide of prestigious literary periodicals or references in academic literature in order to create their ground truth for literary value (Ferrer, 2013; Underwood and Sellers, 2016).

Studies questioning the limitations of one or the other approach have appeared: Porter (2018) questions the conflation of quality and prestige, pointing out that the deviation within a single canon might be broader than between canonical and non-canonical works , while van Cranenburgh et al. (2020) designed a new set of experiments to try and tease contextual from textual factors in readers' evaluation of a literary piece (van Cranenburgh and Koolen, 2020), still in general, the existence of one single average representing a text's quality seems to have been preferred by the community.

A different line of research, with a less prominent predictive vocation, has instead focused on the demographic and individual differences between reading preferences. Touileb et al. (2020) explored Norwegian book reviews and found differences in the literary preferences and the expression of sentiments of female and male reviews, depending on genre (Touileb et al., 2020). A similar analysis of Goodreads reviews pointed to the same direction: there are differences between female and male readers, and it is possible to find evidence for it on a larger scale (Thelwall, 2019). Readers' communities and readers' status also seem to influence the way different groups of people perceive a literary text: Squires (2020) discusses the importance of reviews and reviewers in shaping the book circulation and reading practices, while the increasing availability and popularity of social reading platforms allows for the creation of like-tasted reader groups in a way that has not been possible before (Rebora et al., 2021).

3. Between ground truths and relativism: mild perspectivism?

If quality is absolute, why do readers disagree on the quality of a text? Even an individual reader can change their own idea on the literary value of a text over time. If quality is entirely idiosyncratic, how come there are texts that survive the most drastic historical and cultural changes with an almost unfaltering status? *The Aeneid* remained appreciated as a canonical masterpiece in western Europe from the Roman Empire down to modern times.

A similar question has been discussed in other so-called "subjective annotation" tasks in computational linguistics and machine learning (Davani et al., 2022): the attempt at attaining one single meaningful value in similar contests risks to back-fire, creating an artificial representation of the phenomenon one is trying to model. Some researchers have advocated for a new paradigm, "perspectivism" (Basile et al., 2021), to deal with similar problems by considering a plurality of different points of view on the same data, either by building an average from several individual values (weak perspectivism) or attempting to maintain the inter-annotator differences in the dataset and try to model their diversity (strong perspectivsm) (Checco et al., 2017; Cabitza et al., 2020; Akhtar et al., 2020).

When it comes to literary quality, applying either an non-perspectivist approach (such as having one ground truth or a gold standard), or a strong perspectivist approach gives rise to difficulties, and there seem to be apparent limitations to both approaches when brought to their ultimate consequences.

A non-perspectivist approach suggests assessing the appreciation of literary quality through a single gold standard. Such a gold standard can be approximated by aggregating perspectives of different readers in one value (a rating score, the number of copies sold, an average review sentiment, etc.). A weak perspectivist approach is probably ingrained in any such attempt at modelling and evaluating literary quality: even the works that have tried to reduce it to a single number have relied on majority votes from several readers (average ratings from a crowdsourced annotation task; the number of copies sold; the number of ratings; presence in one or more canons; and so forth). Most literary awards are assigned by a committee composed of several individuals, so even when relying on such institutions to define literary quality, a text's value is approximated by collecting and averaging over several points of view. This form of weak perspectivism essentially treats literary quality as an objective measure to be approximated through many individual measurements (Basile et al., 2021). Taking many imperfect measures of the length of a table will bring us closer to its exact length; taking many personal assessments of the quality of a text will bring us closer to its real value. This take on the stance can help us clarify whether, despite the subjective nature of the task, a common ground of convergence does exist on the topic.

Naturally, this approach is at odds with a subjective view of quality assessment and aesthetic deliberation, and reducing a variety of individual opinions to one score is very helpful in some studies, but is bound to

leave out important variation.

The opposite approach is to apply a strong perspectivist angle and to keep all of the different appreciations of a book in their diversity, without trying to reduce them to an average. If we believe in the irreducibility of readers' preferences to a meaningful average, and if the perception of literary quality is entirely idiosyncratic, it makes sense to model each reader independently. However, considering each reader's appreciation as an irreducible perspective to keep independent from the others risks confusing, or at best diluting, the very scope of this kind of research: finding out whether, *beyond individual variations*, there can be features that define something like an underlying, universally perceived quality in a text.

A third approach is to take a middle way between the two extremes. This will be outlined in the following section.

4. Looking at readers' classes

Instead of relying on either one gold standard or treating all reader perspectives independently, one possibility is to model readers in different classes and have a majority approach for each class. In the study of canonization and literary fame, some differences between readers have been discussed: for example, readers' gender and ethnicity have been posited as playing important roles in the perception of texts (Keen, 2013), and the challenges minorities might face to enter mainstream literature (Berkers et al., 2014).

Another relevant difference is to be found between lay-readers and professional critics. In the debate, the former are often highlighted as inclined to be fooled by cheap reads, and the latter as incline to inaccessible literature. But even between the 'critics' and 'laymen', we can disentangle important subgroups: an occasional Goodreads user and the maintainer of a book blog are both not literary critics in the most canonical sense, but the latter is probably more dedicated to the art of reading and reviewing than the former. A professional literary critic who writes for a local newspaper and one who writes for a specialized niche magazine might belong to two quite distinct categories in terms of sensibility and severity. There are middle ground identities as well: the work by De Greve and Martens (2021) studies the emerging role of social media and argues that 'lay critics' also act as cultural transmitters, challenging the traditional gatekeepers role of professional critics (Greve and Martens, 2021). These differences neither mean that one of these groups' judgment is more correct than another nor that there is no variation or outliers within these groups – but they indicate classes of what we call sensitivity convergence that are likely to display a higher degree of inner agreement than outer.

Hence, with the sensitivity toward groupings of different readers, the approach of aggregating reader perspectives can be applied in a more fine-grained manner.

Dataset overview	
Nr of reviews	57 369
Male reviewer	18 958
Female reviewer	28 984
Unknown	9 427
Nr of different titles	14 647
Male author	8 056
Female author	6 591
Nr of reviews by media type	
Newspapers	22 131
Blogs	16 791
Online media	10 635
Blog-like websites	3 456
Regional newspapers	2 622
Weekly magazines	1 566
Professional magazines	168

Table 1: An overview of the dataset presented in this paper. The category Online media includes (literary) sites that fall between online newspapers and personal blogs.

Instead of relying on *one* gold standard for an overall literary appreciation, we suggest letting the aggregation and statistical means depend on these reader classes, allowing for multiple points of view. It has been argued that computational methods allow for capturing readers' preferences (Walsh and Antoniak, 2021), what we will next discuss from a Danish perspective.

5. Exploring the classes of Danish readers

As a preliminary study in this direction, we have analysed a large dataset of book reviews published in Danish media, such as newspapers and blogs, from 2010 to 2021. [2]. The composition of the dataset is presented in Table 1. The grades of the reviews are fitted to a shared 6-point scale through a linear transformation. In addition, the dataset contains metadata of the publications, such as publisher house, year of publication, etc.

This dataset is unprecedented in terms of dimension, annotation quality, and diachronic extension for the contemporary Danish scene over several platforms, and offers a unique viewpoint to determine in which classes readers - at least those readers who write reviews - most clearly tend to cluster. Since newspapers, blogs, and other online media are the dominant platforms in the dataset, we focus on these to gain a more informed insight into reading preferences within

[2]The dataset is retrieved from the web page bog.nu, a platform that collects book reviews published in Danish media. Reviews without a numeric rating were attributed a rating by the site administrator. We see the same trends in the data with the ratings retrieved from the original reviews ($> 75\%$) as in the data relying on the estimated rating.

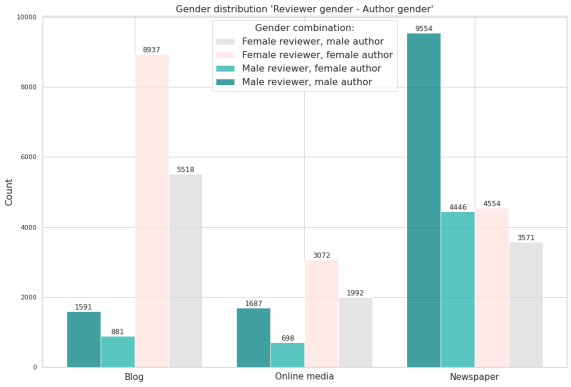

Figure 1: A histogram of the number of reviews shows that male and female reviewers are not equally distributed among the different media types. Blogs and blog-like websites are added together and so are newspaper and regional newspapers.

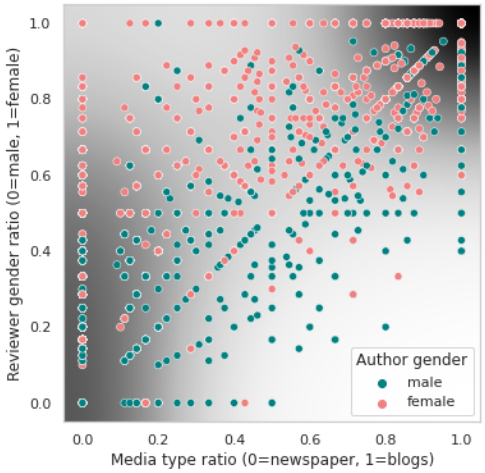

Figure 2: A scatter plot of books reviewed at least 5 times, showing the relative proportion of blog (1) and newspaper (0) reviews on the x axis plotted against the relative proportion of reviewer gender (0 = male, 1 = female) for each title, colored by author gender. As many of the points overlap, the heatmap in the background illustrates where the highest density areas fall.

these media types. Figure 1 shows the gender distribution across three media types - newspaper reviews, online literary reviews and personal blogs - and we see a sharp distinction of reviewer's gender as well as author's gender: male reviewers are more likely to review male authors in newspapers whereas the blogosphere is dominated by female reviewers reviewing female authors.

In this analysis, we are working with a binary understanding of gender and have used a gendered name list to retrieve the gender feature [3]. This method is not an ideal way of approaching gender variables, and we are aware of the problems with this method and how it rules out other gender identities (Dev et al., 2021). However, we find it useful to apply this method in this preliminary study.

Between newspapers and blogs, we can furthermore show a difference in grading. The grades given in newspapers are significantly lower, with an average of 4.1/6, compared to those given in blogs which have an average of 4.5/6. This indicates a difference in review culture, which may be due to blogs being a medium where the emphasis is placed on positive experiences, rather than being professional critics that do not choose the reviewed works according to their preferences. In addition, the social nature of blogs makes it a place to discuss leisure readings with like-minded readers, whereas newspaper reviews tend to be more one-directional. These divided reader profiles also support our argument for modeling reader appreciation in subclasses instead of working with one single gold standard that would apply to all readers.

Figure 2 shows the polarization of the book reviewing scene in Denmark. The ratios on the axes show how books are read between the two genders and across the two media types, the middle point (0.5,0.5) corresponding to works reviewed equally by both genders and on both media platforms. Only titles with five or more reviews were included in this analysis. The heatmap indicates that most titles fall near the two extremes: men seem to dominate the newspaper venues, and the dominance of women in the blogosphere is even stronger. Moreover, the coloring of the plot by author gender reveals that the polarization applies to author gender, too. Along the y axis, we see a clear split at 0.5, showing that books read mostly by female reviewers are also mostly written by female authors, and vice versa. These observations imply that female and male readers read different books, and each groups seems to prefer books written by their own gender.

To obtain a deeper understanding of this polarization, we examined which books had received the highest ratings in each category. When looking at books reviewed by both genders and on both media platforms, the titles that received the best average rating fell in diverse categories. This overarching top includes, among others, Nordic classics, Stoner by John William[4], more modern international bestsellers such as The Goldfinch by Donna Tartt and a graphic novel by Karoline Stjernfelt.

The titles rated the highest by either gender, shows another division: men preferred more canonical books - Herman Melville, Roberto Bolaño, and Victor Hugo being in the top 5 - whereas women preferred reading genre literature, their top-rated books including romance, crime/thriller, and fantasy novels. A similar

[3] We have used the API genderize.io that gives the probability of a name being either male or female, based on a dataset of 250.000 names

[4] Stoner was translated into Danish in 2014, which might explain its sudden occurrence in the dataset.

division was found between the best-rated books in newspaper and blog reviews, although blog reviews were even more dominated by romance books compared to the books most appreciated by female reviewers overall. These observations imply that newspapers, a more established venue dominated by men, focus on canonical works, whereas the constantly evolving blogosphere, dominated by women, seems to seek more leisurable or genre-specific reading.

From these early results it seems that the motivation behind reading, reader status and the gender distribution of authors and readers are valid candidate classes to cluster individual literary perspectives. Thus, as a mild perspectivist approach, we propose taking the degree of professional expertise and the effect of gender into account when assessing literary quality.

6. Discussion

In this article, we have addressed the question of how perspectivist we should be in measuring literary quality. While it has become clear that one literary canon or one gold standard based on e.g., sales numbers cannot capture the variety of aspects readers appreciate in literature, the relevance of a traditional literary canon is reflected in our observations; some works seem to have reached a status that cannot be ignored. However, this literary canon is not a ground truth for quality, and non-canonical popular works might have other features that make them beloved by readers.

Therefore, the problem of literary quality can - and should - be explored from different angles within the same project. Applying strong perspectivism in the future can still be a relevant and viable option to contrast the classes we have divided the Danish readers into.

Furthermore, the division proposed here is not perfect. The division of gender was binary, excluding other gender identities from the current analysis, that need to be considered in the future. Similarly, the contrast of professional and amateur readers is not as absolute as the division into two categories here might suggest. Indeed, some bloggers can be seen as tastemakers that have gained what Driscoll (2019) calls 'readerly capital', and form a lively environments for readers to interact, contributing to a diverse literary space (Driscoll, 2019; Rebora et al., 2021).

In light of the investigated review venues, we can only infer what readers voluntarily reveal about their literary preferences, while they also might have hidden preferences not shown in this data. That could be approximated through a different kind of dataset, such as library loans. With the current method, we are still not capturing all types of readers. Nevertheless, the current findings support the claim that it is not trivial what kind of reader profiles we consider and value when studying literary quality.

7. Conclusion and Future Works

Literary quality is a complex topic, and it remains a challenge for both strong and weak perspectivist stances. In this paper we have tried to consider the pros and cons of both approaches and what adopting them implies. We have, then, suggested a middle way between the two extremes, by dividing readers into meaningful classes that would represent different perspectives on the same text, without holding each individual rating as a independent judgment. Through the analysis of over 57.000 book reviews in Danish media we have shown that some features of the reviewers – especially gender and whether they write for a blog or a newspaper – appear to significantly predict a shift in the assessment of a text, and thus allow for a meaningful clustering of readers into perspective classes.

Naturally, we have much left to do to further explore the relevance of this approach for literary quality modeling. In future we intend to use the existing classes as labels for quality prediction to see whether they can yield a more informative picture of the judgments a literary text is likely to elicit. We would also like to look for subtler differences between the reviewers and compare these findings with other existing resources for literary quality. Another important question we would like to address in the future is whether and when a preference becomes a bias: for example, in what situations controlling for gender preferences should be used to "correct" a system's output rather than just inform it.

Overall, the complexity of the problem and its midway status between objectivity and subjectivity remains a topic for debate both within and beyond computational linguistics, and leaves large room for future developments.

8. References

Akhtar, S., Basile, V., and Patti, V. (2020). Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 151–154.

Archer, J. and Jockers, M. L. (2016). *The bestseller code: Anatomy of the blockbuster novel*. St. Martin's Press.

Ashok, V. G., Feng, S., and Choi, Y. (2013). Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1753–1764.

Basile, V., Cabitza, F., Campagner, A., and Fell, M. (2021). Toward a perspectivist turn in ground truthing for predictive computing. *Conference of the Italian Chapter of the Association for Intelligent Systems (ItAIS 2021)*.

Berkers, P., Janssen, S., and Verboord, M. (2014). Assimilation into the literary mainstream? the classification of ethnic minority authors in newspaper reviews in the united states, the netherlands and germany. *Cultural Sociology*, 8(1):25–44.

Bizzoni, Y., Peura, T., Thomsen, M. R., and Nielbo,

K. (2021). Sentiment dynamics of success: Fractal scaling of story arcs predicts reader preferences.

Bloom, H. (2014). *The western canon: The books and school of the ages*. Houghton Mifflin Harcourt.

Cabitza, F., Campagner, A., and Sconfienza, L. M. (2020). As if sand were stone. new concepts and metrics to probe the ground on which to build trustable ai. *BMC Medical Informatics and Decision Making*, 20(1):1–21.

Checco, A., Roitero, K., Maddalena, E., Mizzaro, S., and Demartini, G. (2017). Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*.

Davani, A. M., Díaz, M., and Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Dev, S., Monajatipoor, M., Ovalle, A., Subramonian, A., Phillips, J. M., and Chang, K. (2021). Harms of gender exclusivity and challenges in non-binary representation in language technologies. *CoRR*, abs/2108.12084.

Driscoll, E. (2019). Book blogs as tastemakers. *Participations. Journal of Audience & Reception Studies*, 16:280–305.

Ferrer, C. (2013). Canonical values vs. the law of large numbers: The canadian literary canon in the age of big data. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 5(3):81–90.

Greve, L. D. and Martens, G., (2021). *The Audience (Dis)Agrees: Studying the Impact of Award-Winning Books on Lay Literary Value Judgements Using Social Media Data*, volume 9, pages 85–130. Barkhuis.

Guillory, J. (2013). *Cultural Capital: The Problem of Literary Canon Formation*. University of Chicago Press.

Herrmann, J. B., Jacobs, A. M., and Piper, A. (2021). Computational stylistics. *Handbook of Empirical Literary Studies*, page 451.

Hu, Q., Liu, B., Thomsen, M. R., Gao, J., and Nielbo, K. L. (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.

Keen, S. (2013). Empathy in reading. *Anglistik*, 24(2).

Maharjan, S., Kar, S., Montes, M., González, F. A., and Solorio, T. (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, June. Association for Computational Linguistics.

Plank, B., Hovy, D., and Søgaard, A. (2014). Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511.

Porter, J. D. (2018). *Popularity/Prestige*. Literary Lab.

Rebora, S., Boot, P., Pianzola, F., Gasser, B., Herrmann, J. B., Kraxenberger, M., Kuijpers, M. M., Lauer, G., Lendvai, P., Messerli, T. C., and Sorrentino, P. (2021). Digital humanities and digital social reading. *Digital Scholarship in the Humanities*, 36(Supplement 2):230–250, 11.

Squires, C., (2020). *The Review and the Reviewer*, pages 117–132. Routledge. Num Pages: 16.

Thelwall, M. (2019). Reader and author gender and genre in goodreads. *Journal of Librarianship and Information Science*, 51(2):403–430.

Touileb, S., Øvrelid, L., and Velldal, E. (2020). Gender and sentiment, critics and authors: a dataset of Norwegian book reviews. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 125–138, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Underwood, T. and Sellers, J. (2016). The longue durée of literary prestige. *Modern Language Quarterly*, 77(3):321–344.

van Cranenburgh, A. and Koolen, C. (2020). Results of a single blind literary taste test with short anonymized novel fragments. *arXiv preprint arXiv:2011.01624*.

van Cranenburgh, A., van Dalen-Oskam, K., and van Zundert, J. (2019). Vector space explorations of literary language. *Language Resources and Evaluation*, 53(4):625–650.

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

Wang, X., Yucesoy, B., Varol, O., Eliassi-Rad, T., and Barabási, A.-L. (2019). Success in books: predicting book sales before publication. *EPJ Data Science*, 8(1):1–20.