Perplexing Canon:
A study on GPT-based perplexity
for canonical and non-canonical literary works

Yaru Wu
Uppsala University
yaru.wu.6038@student.uu.se

Pascale Feldkamp Moreira
Center for Humanities Computing
Aarhus University
pascale.moreira@cc.au.dk

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

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

Abstract

This study extends previous research on literary quality by using information theory-based methods to assess the level of perplexity recorded by three large language models when processing 20th-century English works deemed to have high literary quality, recognized by experts as canonical, compared to a broader control group. We find that canonical texts appear to elicit a higher perplexity in the models and we explore which textual features might concur to create such effect. We find that the usage of a more heavily nominal style, together with a more diverse vocabulary, is one leading cause for the difference between the two groups. These traits could reflect "strategies" to achieve a more informationally dense literary style in the canonical groups.

1 Introduction

The question of what "literary quality" is has been at the center of a millennia-long debate in aesthetics and literary studies.While literary judgment is almost by definition subjective, reflecting individual reader preferences, quantitative studies have shown that such preferences tend to converge at the large scale, and that both textual features, like coherence and style (Bizzoni et al., 2023c,a; Koolen et al., 2020; van Cranenburgh and Bod, 2017; Archer and Jockers, 2017), and text-extrinsic features, such as reader or critic demographics (Lassen et al., 2022; Koolen, 2018; Wang et al., 2019), significantly influence appreciation or perceived quality. Most schools of thought in literary research tend to see literary quality as either a perceived quality – an effect of reception and cultural dynamics (Bourdieu, 1993; Casanova, 2007; Guillory, 1995) – or as the effect of certain textual features, such as, among others, authorial strategies of defamiliarization and foregrounding (Shklovsky, 1917; Mukařovský, 1964; Peer, 2008; Attridge, 2004). While consensus on a single gold standard of quality is hard to achieve (Bizzoni et al., 2022), reader preferences and expert valuations can offer a range of measurable levels of appreciation. An often discussed dimension of literary quality is that of the so-called "literary canon", a complex concept generally representing a set of works that have survived or/and (by the same token) remain distinguished in the memory of a literary culture (Bloom, 1995). A community, usually over large periods of time, defines as outstanding and worthy of attention; yet this process is not devolved to any individual authority, which makes the very definition of what is within the canon complex. As such, the canon is often scrutinized in cultural approaches to literary quality. Some schools of thought have seen it as representing nothing but entrenched interests (von Hallberg, 1983) and thus as the cultural capital of ruling classes (Guillory, 1995), while others have considered "canonic" works to excel in terms of intrinsic features (Bloom, 1995), whether stylistic (Brottrager et al., 2022; Barré et al., 2023; Algee-Hewitt et al., 2016) or narrative (Bizzoni et al., 2023d). In his work on the dynamics in the literary field, Bourdieu (1993) placed "popular success" and "consecration" at opposed positions.[1] Recent quantitative studies of the literary field often follow a similar distinction between popularity and prestige (Porter, 2018; Manshel et al., 2019), where more prestigious books and genres are what we could call the more "literary" ones (Porter, 2018;

[1] "There are few fields in which the antagonism between the occupants of the polar positions is more total [than in the literary]" (Bourdieu, 1993, p. 46).

Proceedings of LaTeCH-CLfL 2024, pages 172–184
March 22, 2024 ©2024 Association for Computational Linguistics

Lassen et al., 2023). Supporting the idea that such literariness may be distinguished by certain text-intrinsic features, Koolen et al. (2020) have shown that readers agree more on the "literariness" of books, and that it is easier to model literariness ratings from textual features than overall enjoyment ratings.

In this work, we approximate what can be considered canonical in a large corpus of around 9,000 novels. Based on this corpus, we ask two main questions: (1) Are canonical novels more "perplexing", as measured through different Large Language Models, than a non-canonical control group? (2) If they are, which linguistic and stylistic features might contribute to the difference?

Our reason for using perplexity (see Chapter 4 for the formal definition of perplexity) is at least two-fold. On one hand, canonical works of fiction are often examples of either "virtuous" (Bloom, 1995) or defamiliarizing usage of language (Mukařovský, 1964; Peer, 2008), thus an uncommon usage of language. Such characteristics may make canonical works of fiction more surprising with respect to non-canonical. Even when perplexity is operationalized as a measure of information theory, these works might elicit a higher perplexity on average. On the other hand, perplexity is a central measure of informativity in information theory (Shannon, 1949). Since perplexity is a function of surprisal, more perplexing texts tend to be more informationally dense. A highly specialized scientific paper – like a highly complex and articulate page of James Joyce – is unusually informative in the sense that it constantly relays novel information (highly specialized or new words – neologisms – or words in a new order) to the reader. In theory, a communicative system that manages to be as dense as possible without breaking down or being "too dense" for its own readers indicates elements of a heightened communicative efficiency – a feature that communities might tend to optimize over time (Rubino et al., 2016; Biber and Gray, 2011).

2 Related works

Studies seeking to predict literary success or perceived literary quality have often followed the intuitive idea that readers perceive a difference between more difficult and easier texts, and approximate some form of stylistic complexity. Such studies use features related to the readability indices developed in linguistics research, such as sentence length, vocabulary richness, or redundancy (Brottrager et al., 2022; van Cranenburgh and Bod, 2017; Crosbie et al., 2013; Koolen et al., 2020; Maharjan et al., 2017; Algee-Hewitt et al., 2016). Additionally, readability formulas find integration in editing tools such as the Hemingway or Marlowe applications,[2] which prioritize more "readable" texts. Yet the relation between stylistic aspects of text complexity and reader appreciation appears complex: while it is suggested that readers prefer more stylistically complex or informationally dense texts (Algee-Hewitt et al., 2016), it is a widespread conception that bestsellers are easier to read (Martin, 1996). In literary studies, reading ease has also been proposed as a marker of "better" style as far back as 1893 (Sherman, 1893). While Martin (1996) and Maharjan et al. (2017) found that readability formulas were weak for predicting reader appreciation, more recent work has shown that preference for the type of text difficulty measured by readability formulas may vary across different audiences: novels with higher readability are preferred by raters on large online platforms, while award-winning novels tend to have lower scores (Bizzoni et al., 2023a). Measures that are more explicitly related to information density or entropy, such as word and bigram entropy (Algee-Hewitt et al., 2016), surprisal (McGrath et al., 2018), and text compressibility (Ehret and Szmrecsanyi, 2016) have also been used to assess the complexity of literary texts.[3] Liddle (2019), for example, shows a diachronic evolution of literary texts towards a greater density of information. Surprisal has been shown to correlate with the cognitive effort of processing words (Hale, 2001; Levy, 2008; Balling and Baayen, 2012) and is as such a measure of the information density of text. The connection of information density or surprisal with a text's relative "quality" (in this case intended as communicative effectiveness) has been linked more explicitly in studies about non-literary domains. For example, Degaetano-Ortlieb and Teich (2022) found that scientific prose has gradually developed informationally denser prose, optimized for expert-to-expert communication.

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

[3] In the latter case, the aim is to approximate Kolmogorov complexity, i.e., the complexity of e.g. a string is defined as the length of the shortest possible description of it, as in Ehret and Szmrecsanyi (2016) and Liddle (2019).

Perplexity, as a closely related measure of the probability of words in context, may be applied as another measure of difficulty or as a measure of the *information density* of a text (Rubino et al., 2016). While perplexity is primarily used as an internal evaluation metric for the performance of language models, it has also been used variously as a descriptive and predictive metric to distinguish between the domain and style of texts, for example between formal and colloquial tweets (Gonzalez) or between speech production by people with dementia and without (Fritsch et al., 2019).

Like surprisal, LLMs' perplexity also shows a relationship to human word processing or perception of text difficulty, for example with gaze duration in reading (Goodkind and Bicknell, 2018), though the similarity of model perplexity to, for example, human reading time may change with larger model size (Oh and Schuler, 2021). Still, the relation between the "difficulty" level of a text and perplexity is not clear-cut, and perplexity seems to capture something different than what can be estimated with traditional readability formulas from linguistics research. Miaschi et al. (2020) show no relation between model perplexity and one readability measure, while Martinc et al. (2021) suggest that models might actually attribute less perplexity to texts aimed at adults compared to texts aimed at younger audiences. Similarly, there seems to be no clear connection of perplexity to stylistic features of texts connected to readability, suggesting that different textual features affect readability and model perplexity (Miaschi et al., 2020).[4] Some work has been done to estimate surprise or narrative coherence in fiction (McGrath et al., 2018; Underwood, 2020; Wu, 2023), still the question of quality or reader appreciation of more or less "surprising" texts remains underexplored. In this context, perplexity may constitute an additional measure easily related to different types of reader appreciation. Notably, in text generation, model perplexity is explored to retrieve more or less diverse output, given that a higher likelihood text (with less perplexity) does not necessarily mean that it is of better quality for human raters (Zhang et al., 2020).

3 Data

3.1 Corpus

We use a corpus spanning 9,089 novels published in the US between 1880 and 2000 (see Table 1 and Figure 1). It is a unique resource both in terms of size [5] and diversity, as it contains relatively recent novels from various genres. It was compiled based on the number of libraries holding each novel with a preference for higher holding numbers, i.e., for more circulated works. As library holdings reflect a diverse demand, the corpus is not homogeneous in terms of genre and features both prestigious and popular works ranging from Mystery to Science Fiction (Long and Roland, 2016).[6]

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

Titles	Authors	Titles per author
9089	3166	2.88

For example, the corpus contains several National Book Award winners (including Don DeLillo, Joyce Carol Oates, and Philip Roth), as well as important works of genre-fiction (i.e., Tolkien or Philip K. Dick), influential authors of mainstream fiction (such as Agatha Christie and Stephen King), and highly canonical names (such as James Joyce and Ernst Hemingway). Books in the corpus vary in length, from 341 words (Beatrix Potter's *The Story of Miss Moppet*) to 714,744 words (Ben A. Williams' *House Divided*), though only 255 books – 2.9% of the corpus – are shorter than 35,000 words – roughly the average length of a novella like Orwell's *Animal Farm*. The total word count of the corpus is 1,060,549,793 words.

3.2 Canonical novels

The definition of "canonical works" used here adheres to a comprehensive principle, relying on the amalgamation of different expert perspectives on canonicity. We considered four main sources (see Fig. 8 for the number of works gathered from each source and the overlap between sources):

[4] Perplexity appears to be estimated consistently across different (and also smaller) models (Goodkind and Bicknell, 2018)

[5] Studies on literary quality often rely on corpora of $<$ 1,000 books (Ashok et al., 2013; Koolen et al., 2020).

[6] While the corpus has no reference publication, several other works have used the same dataset (Underwood et al., 2018; Cheng, 2020). See `https://textual-optics-lab.uchicago.edu/us_novel_corpus` for an overview of the corpus.

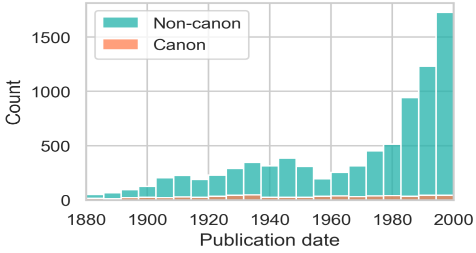

Figure 1: Distribution of canonic titles in the Chicago Corpus over time.

(i) **The Norton Anthology**: This is a leading anthology dedicated to authors considered canonical (Pope, 2019; Ragen, 1992). We consulted both the English and the American Norton Anthology

(ii) **College Syllabi**: The frequency with which college syllabi include an author's work can measure their level of canonization (Barré et al., 2023). We used OpenSyllabus, a database that has compiled 18.7 million college syllabi[7]. Using this data, we tallied all works in our collection from the top 1000 most frequent authors in *English Literature* syllabi.

(iii) **Classics Series**: Numerous major publishers, such as Vintage and Penguin[8], feature a series dedicated to "classic" (e.g. canonic) literature. Given Penguin's status as a leading publisher of English-language literature (Alter et al., 2022), we marked all works in our corpus that featured in this series.

(iv) **Prizes**: We collected long-listed titles (winners and finalists) for prestigious literary awards: The Nobel Prize in Literature, the Pulitzer Prize, the National Book Award. Manshel et al. (2019) have shown that winning an award contributes to the long-term prestige – but also popularity – of titles in academia and on GoodReads. The choices of award-committees seem to be in touch with the general public, but prize-winning books also seem to be connected to disagreement between readers at the large scale (Kovács and Sharkey, 2014).

These sources divide our dataset in two groups: 745 canonical and 8344 non-canonical works. Naturally, we consider this division artificial, as a necessary rule of thumb to make the study possible. In fact, canonicity is not a defined and boolean variable (Barré et al., 2023), but would be best represented as a continuum on several dimensions. To contrast against these proxies, we also collected books in our corpus that are in Publisher's Weekly American 20th century bestseller list.[9]

4 Perplexity

Perplexity is an information-theoretic measurement of how well a probability model predicts a sample (Goldberg, 2022). The perplexity of a well-trained language model on a test text can be interpreted as the exponential of its average level of surprisal (Hao et al., 2020), namely

$$e^{-\frac{1}{N}\sum_{i=1}^{N} ln(P(token_i|tokens_{j<i}))} \tag{1}$$

where N is the number of tokens of the text and $P(token_i|tokens_{j<i})$ is the probability assigned to the ith token after the model has processed the first $i-1$ tokens. Thus, lower perplexities indicate that the model is less uncertain about its predictions. In information theory and linguistics, this measure

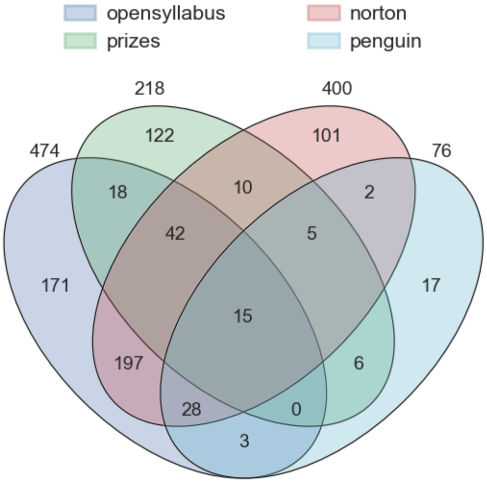

Figure 2: Number and overlap of the canonicity sources used in this study. Note that the largest overlap appears to be between the Norton Anthology and Opensyllabus, indicating the near relation between the two proxies, possibly due to the institutional affiliation of the Norton Anthology.

[7] https://www.opensyllabus.org

[8] https://www.penguin.com/penguin-classics-overview/

[9] Extracted from the database collected 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. Publishers Weekly is a trade news magazine which is published once a week (from 1872) and targeted at agents in the field: publishers, literary agents, booksellers, and librarians. Although based on sales numbers, the full set of selection criteria for the list are unknown.

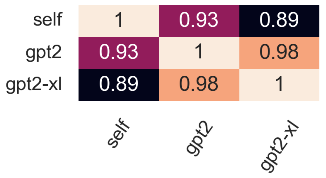

Figure 3: Spearman's correlation between the three GPT-2 based models' mean perplexity scores on each novel of our corpus

is often used to approximate how surprising or complex a text can be for humans as well. As language models are trained on word sequences, perplexity has the benefit of encapsulating lexical, grammatical and syntactic phenomena alike. Strictly in this sense, perplexity-based approaches are able to model a text more holistically than approaches that focus only on one linguistic dimension.

In this study, we calculate the perplexity of three language model on the same corpus. We based all perplexity calculations on the byte pair encoding tokenization (Sennrich et al., 2016) used in the series of gpt2 models. Due to constraints imposed by the maximum input length, we employed a stride-based methodology to gauge perplexity at the text level. This method incorporates a strided sliding window, wherein the context is shifted by fixed-length strides, affording the model an expansive context for making predictions at each iteration. In this specific framework, the window's size is 1024 tokens long, with a stride length of 512 tokens. By doing so, the second half of the previous context window served as the first half of a new context window to calculate perplexity estimates for the remaining 512 tokens (Oh and Schuler, 2023). Therefore, the surprisal for each book is comprised of perplexity values for the initial 1024 tokens, intermediate segments of 512 tokens, and the residual tokens of varying lengths. The aggregate-mean value is designated as the textual perplexity within the confines of this study.

5 Models

We use three alternative, large language models to assess the average perplexity of the novels (see Table 2 for details). For each novel, we thus have three perplexity measures.

The two standard versions of GPT2 models (Radford et al., 2019), namely the gpt2 and the gpt2-xl, are used in this paper since they are based on neural networks with billions of parameters and trained on terabytes of text, achieving very good results both in generating natural text and in estimating the perplexity of unseen texts. However, there is a substantial risk that some of the books of the corpus may have been included within the dataset that these models were trained on, especially when OpenAI has not yet published its dataset until now. Therefore, the main methodology is to train a model of the same architecture as the series of gpt2 models from scratch using a dataset outside the corpus (hereinafter referred to as the self-trained model).

If the corpus perplexity estimation observed from the self-trained model is in close correlation with results from the series of gpt2 models, then the potential bias risk can be excluded. In this context, a new text generator based on the gpt2 model is trained from the beginning on the "article" content of the CNN Dailymail Dataset[10]. The primary reason for not employing other literary works as the training set is due to the potential bias associated with the selection of these books. Moreover, the CNN Dailymail dataset is chosen for its compilation of approximately one million news stories designed for reading and comprehension tasks (Hermann et al., 2015), offering a narrative consistency more closely aligned with that of novels than other datasets, such as WikiText. Then, we use the Adam optimizer with a learning rate of 5e-5 and a cross-entropy loss criterion to train the model for 10 epochs.

Table 2: Architecture hyperparameters and training set sizes for the three models.

	self_model	**gpt2**	**gpt2-xl**
parameters	117M	117M	1542M
layers	12	12	48
heads	12	12	25
dimensions	768	768	1600
dataset size	535M	40G	40G

The Spearmann Correlation test results presented in Figure 3 show a robust correlation in perplexity values between the self-trained model and the other gpt2 models, indicating that a potential data bias can be excluded at least within this corpus. Therefore, the models forming the final hierarchy can be viewed as a sequential examination on the hypotheses and the consistency of our results across the models of varying sizes, ranging from the smallest version of the self-trained model to the largest version of the gpt2-xl model (see Table 2).

[10] https://huggingface.co/datasets/ccdv/cnn_dailymail

6 Perplexity & the Canon

These three variants of GPT2 models based on the Transformer architecture are employed to calculate perplexity values using the stride-based method across the entirety of the novels in the Chicago Corpus. A first mean of evaluation is to observe whether the mean perplexity changes with the models' size, as we would expect larger models to display lower perplexity. Consistently with our expectations, the mean perplexity values decrease when the model size increases, as delineated in Table 3 [11]. Largest models are likely to be less perplexed by unusual linguistic structures, as they have been trained on much larger datasets and have, in some sense, "seen more". It is also a matter of debate, in this respect, whether larger is always better when it comes to correlations with human intuition. It is possible that very large models are harder to surprise than human readers, and their levels of perplexity may not correspond with human readers' experiences as much as those of smaller models (Oh and Schuler, 2021). In our case, we find that the distinction between canonical and non-canonical works, defined by humans, is most strongly reproduced by the smallest of the three models.

Despite some potential fluctuations, the outcomes exhibit general consistency across the three language models. Notably, the highest and lowest perplexity values are elicited from the same two books, namely *The Graduate* by Charles Richard Webb (the least perplexing novel overall) and *Finnegans Wake* by James Joyce (the most perplexing).

Table 3: An outlook on the perplexity values estimated by the three models.

	self_model	**gpt2**	**gpt2-xl**
min	16.307	8.9058	6.5862
max	998.4872	306.1784	229.1857
mean	67.1944	28.8428	18.2334

Then, the Mann-Whitney test is used to examine the perplexity difference between canonical and non-canonical works. As shown in Table 6 , in terms of perplexity the difference between canonical and non-canonical novels is significant over all of the three models, with canonical books being more perplexing than non-canonical in all cases. This can be in turn surprising: canonical works, due to their status, might influence other works and become more typical. Yet it seems that they retain a unique originality, or a specially distinctive usage of language. Moreover, there seems to be an internal variation within the canon.

Table 4: Correlation between perplexity and readability with GoodReads' rating count and number of libraries' holdings for each novel - proxies for the popularity or circulation of the works.

	GR rating count	Libraries
self_model	-.23	-.31
R Dale-Chall	-.22	-.25

When inspecting works of different types of canonicity (contrasting literary prizes with other types of collections) we find that works judged canonical by experts and that are more closely affiliated to institutions (the Norton Anthology, Opensyllabus, and Penguin Classics) have a higher perplexity (Table 7). If we contrast with bestsellers, we also find that these appear to even have a lower perplexity than non-canon works (Table 7).

It is important here to note here once again that perplexity is not an absolute measure, but is the result of a model's training. Models trained on large enough datasets will capture fundamental regularities and find idiosyncratic uses of language more perplexing, but every model will consider elements closer to its training set as more normal. In this paper we assume the large training sets of the models as representative enough of contemporary English.

7 Correlations with textual features

Table 5: Correlation Matrix of Readability Metrics and PPLs (Spearmann correlation)

	self_model	**gpt2**	**gpt2-xl**
Flesch Ease	-0.530	-0.483	-0.428
Flesch Grade	0.581	0.530	0.470
Smog	0.532	0.480	0.422
ARI	0.636	0.571	0.506
Dale-Chall	0.608	0.603	0.550

Perplexity is a powerful measure of linguistic predictability, as it results from large-scale modelling of word sequences. It can also be, and usually is, the effect of a composition of several factors, so that it is not always easy to understand what elements are driving its values. The richness of a large corpus of narrative fiction only adds to this difficulty. According to all our models, the most perplexing "novel" in the corpus is James

[11] Also see Figure 7 in Appendix

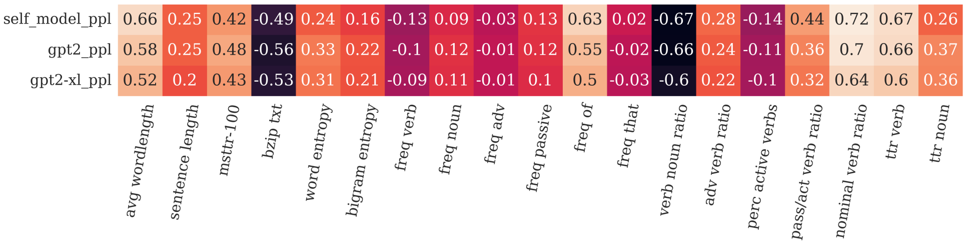

Figure 4: Correlations (Spearman) of perplexity with stylistic and syntactic features.

Joyce's *Finnegans Wake*, while the least perplexing is Webb's *The Graduate*. A look at the first few lines of these books suffices to align our intuition to the models' results.[12] But the reasons for scoring a high perplexity can be different even among those texts that are "clearly" unusual: for example, another high-perplexing novel, Harris' *Nights with Uncle Remus*, often reads as a fairy tale, but is heavily interspersed with heavy use of almost unintelligible eye dialect.[13] While the models' scores clearly pick from the same elements - recording internal Spearman correlations between 0.89 and 0.93 (Fig. 3) - it is not easy to determine which linguistic features have the highest role in determining a given level of perplexity, and, more importantly, in determining which are the perplexing elements in a text that help tell canonical from non-canonical works. In the next sections, we will check the correlation between perplexity and some textual features often considered in the discourse over literary quality and canonicity. We refer to Figure 4 for a summary of the findings.

7.1 Stylometric features

A novel's high perplexity score can be the effect of stylistic complexity. A simple conceptualization of this dimension of style is represented by readability measures, a family of algorithms developed in linguistics that gauge prose difficulty based on simple elements such as sentence and word length, and frequently used in relation to concepts of general literary quality (Bizzoni et al., 2023b; Weigel, 2016; Ashok et al., 2013).[14] The models' perplexity shows robust correlations with all readability measures: books with a higher perplexity are harder to read (Table 5), at least to an extent.

This is not an obvious correlation, as the central elements in readability algorithms, such as sentence length, are not directly factored in the language model's computation of perplexity. Yet, average sentence length alone has $>.2$ correlations with all our models: texts that are challenging at other levels also tend, to an extent, to feature longer sentences. Other features that affect formulae of readability, such as average word length, also show robust correlations with perplexity.It seems to indicate that canonical works present on average a prose that is more difficult to read than non-canonical works. The inverse relation of readability and perplexity with some proxies of mere popularity, as shown in Table 4, additionally indicates that there is at least one "type" of novel that aggregates different strategies of simplicity - unsurprising usage of language, shorter sentences, shorter words etc. - to achieve a higher level of diffusion. While this, too, can be considered a distinct form of quality, it appears that canonical works tend to the opposite stylistic pole.

Another typical metric often associated with more complex and challenging novels is Type-Token Ratio (Kao and Jurafsky, 2012). As TTR shows a significant relation with our perplexity measures, it is likely that more perplexing novels use a more diverse lexicon or more complex lexical structures rather than simpler and more repetitive alternatives, as also shown by the negative correlation with text compressibility, often a proxy for formulaicity or information density (Fig. 4).[15]

7.2 Syntactic features

We looked into selected syntactic and grammatical features often considered in discussions about

[12] **Finnegans Wake**: "riverrun, past Eve and Adam's, from swerve of shore to bend of bay, brings us by a commodius vicus of recirculation back to Howth Castle and Environs.". **The Graduate**: "Benjamin Braddock graduated from a small eastern college on a day in June. Then he flew home. The following evening a party was given for him by his parents."

[13] "Ez ter dat," responded Uncle Remus, "dey mought stan' on one foot an' drap off ter sleep en fergit deyse'f. Deze yer gooses".

[14] Explicating the formulas is out of the scope of this paper, but can be consulted via the package we used to extract readability scores: https://pypi.org/project/textstat/

[15] As it is used in (Ehret and Szmrecsanyi, 2016; Liddle, 2019).

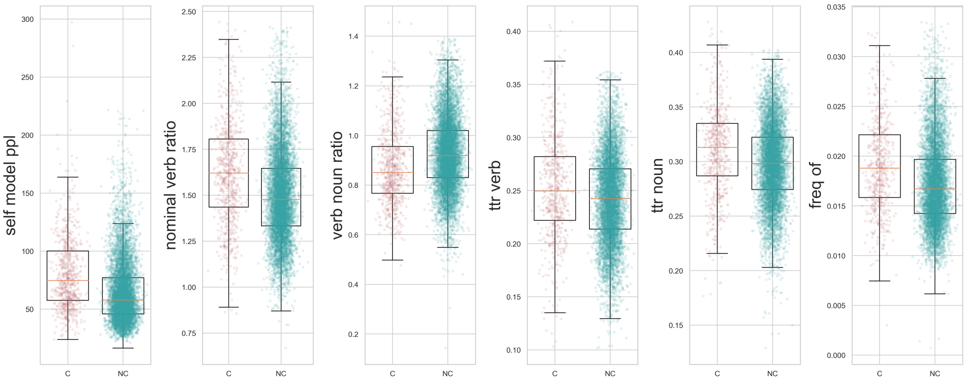

Figure 5: Features distribution for canonical (C) and non-canonical (NC) titles in our corpus. The nominal verb ratio is intended as the ratio of both adjectives and nouns over verbs.

Table 6: Mean and standard deviation for canonical (c) and non-canonical (nc) works regarding features displaying the highest correlations with perplexity. Mann-Whitney's z score and size effect r are reported. * p-value <.005. Numbers in parenthesis report the means, stds, z and r values when running the measures on a corpus where we have randomly selected 1 book per author (thus a smaller corpus of 3153 works of which 200 are canonical).

Measure	**Mean_c**	**Std_c**	**Mean_nc**	**Std_nc**	**z**	**r**
Perplexity (self)	**81.57** (80.17)	53.27 (76.70)	65.01 (64.32)	74.92 (32.26)	3.1* (2.7*) (m)	.33 (.27)
Perplexity (gpt2)	**34.21** (34.33)	15.17 (22.74)	28.00 (28.39)	9.39 (7.91)	3.1* (2.8*) (m)	.38 (.33)
Perplexity (gpt-xl)	**21.81** (22.06)	10.70 (16.97)	17.74 (17.91)	7.03 (4.87)	3.2* (2.8*) (m)	.39 (.35)
Adj+Noun/Verb Ratio	**1.62** (1.60)	0.29 (0.28)	1.50 (1.50)	0.24 (0.25)	3.0* (2.6*) (m)	.27 (.22)
Verb/Noun Ratio	0.874 (0.881)	0.159 (0.153)	**0.929** (0.924)	0.151 (0.15)	1.8* (1.7*) (m)	.23 (.17)
Adverb/Verb Ratio	**0.406** (0.399)	0.069 (0.067)	0.386 (0.378)	0.066 (0.065)	2.7* (2.5*) (m)	.17 (.18)
TTR verbs	**0.253** (0.252)	0.052 (0.053)	0.242 (0.245)	0.042 (0.043)	2.5* (2.3) (m)	.17 (.10)
TTR nouns	**0.312** (0.312)	0.046 (0.053)	0.298 (0.299)	0.037 (0.038)	2.7* (2.4*) (m)	.12 (.15)

the quality of literary (and general) writing: frequency of passive voice and adverbs (Strunk Jr and White, 2007) and relative ratios of Parts-of-Speech, especially looking for traces of so-called "nominal style" (McIntosh, 1975; Bostian, 1983).

The frequency of the passive voice has a faint positive correlation with perplexity, and the active voice a slight negative correlation, suggesting that the passive is slightly more unusual than the active voice. While the percentage of adverbs and verbs plays no role in the perplexity of the novels, the adverbs-to-verb ratio does show a positive correlation.

The verb/noun ratio of each novel displays a very robust correlation with the texts' perplexity. This effect is even more pronounced when we compute the ratio of nouns plus adjectives against verbs. It displays one of the strongest correlations with model perplexity along the three models (see Figures 4 and 6) and delineates a significant difference between the canonical and non-canonical groups (Table 6). We also checked for the relative frequency of the "function words" *of* and *that*: the first is associated with the presence of more nominal phrases, while the latter is typical of more declarative and verb-centered prose. The fact that *of* has a stronger correlation with perplexity than *that*, and is more frequent in the canonical group (Figure 5), is another hint to the larger presence of nominal phrases in more perplexing works. Interestingly, these differences can be extended to subcategories *within* the canonical category: longlisted novels exhibit less perplexity, reading difficulty and traces of nominal style than the rest of the canon group (more "classically" canonical) but more so than the non-canon group, bestsellers included (Table 7).

Nominal style is often considered "heavier" (Huckin, 1993). Several studies on linguistic and information theory also found that non-fiction domains tend to optimize their communication strategies by increasingly relying on nominal phrases – a strategy that works for "expert" audiences (Rubino et al., 2016; Degaetano-Ortlieb et al., 2019; Juzek et al., 2020; Bizzoni et al., 2020). It is possible that the canons' higher perplexity is partly due to including more cognitively demanding, "heavily nominal" texts. One example is Jack Kerouac's

Table 7: Means and standard deviations (in parentheses) for measures across proxies. Note how bestsellers are closer to non-canonical works than canonicals in terms of overall perplexity.

	Non-canon	**Bestsellers**	**Prizes**	**Canon lists**
Perplexity (self model)	67.85 (71.44)	64.1 (24.69)	73.08 (28.82)	**85.17** (54.17)
Nominal/verb ratio	1.51 (0.27)	1.55 (0.25)	1.56 (0.28)	**1.64** (0.29)
Verb/noun ratio	**0.92 (0.15)**	0.91 (0.15)	0.90 (0.15)	0.87 (0.16)
TTR verb	0.24 (0.04)	**0.25** (0.05)	**0.25** (0.05)	**0.25** (0.06)
TTR noun	0.30 (0.04)	0.30 (0.04)	**0.31** (0.05)	**0.31** (0.05)
Dale-Chall readability	5.10 (0.32)	5.01 (0.29)	5.15 (0.35)	**5.29** (0.46)

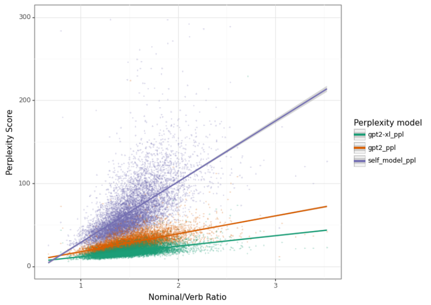

Figure 6: Correlation between perplexity scores and nominal/verb ratio of texts.

Doctor Sax which is one of the top books in our corpus in terms of perplexity and also of nominal/verb ratio. Its prose is rich with adjectives and nouns, sometimes skipping verbs altogether, as in: "not as if idiot but as if sensual or senseless and bitter with venoms of woe". Combined with its frequent neologisms this work offers a good example of text-intrinsic features of perplexity.

8 Conclusions

We have explored some features of canonical vs non-canonical works based on a corpus of 9,000 novels from the late 19$^{\text{the}}$ and the 20$^{\text{th}}$ century. We first found that canonical novels seem to elicit higher perplexity scores based on three LLMs, with the difference remaining significant across different model sizes. Perplexity seems to reflect a higher complexity in style of canonical novels, compared to that of non-canonical works that enjoy a vast readership.

We have then explored some specific features that might contribute to this effect. Based on our collection, the higher perplexity of the canonical group is linked to different distributions of grammatical constructions: heavier use of nominal phrases, paired with average longer sentences, words, and a higher lexical diversity. Specifically, the presence of a more marked nominal style might be an important cause for the difference in perplexity between the two groups, although it is clear that the overall effect is a result of an ensemble of features at the syntactic, stylistic, and semantic level. The idea that "canonical" novels, on average, are more challenging for readers than non-canonical ones, while the opposite holds for widely spread but non-canonical texts (such as texts rated very often on GoodReads), mirrors existent findings (Bizzoni et al., 2023b).[16] The characteristics of this difference are of particular interest as they seem to be at least partly linked to the communication efficiency observed in expert-domain prose for other fields. A heavily nominal style has been linked with the development of refined and diverse vocabulary, a higher cognitive load for the reader, and more effective communication of information, as nouns can be highly specific and diverse, bringing a higher amount of information at the cost of higher decoding effort. It is a hypothesis worth considering that canonical works might achieve a lower communicative *immediacy* in favor of a higher communicative *efficiency*. What this means for a literary work and what it implies for the reader's experience – given the unique communicative functions of literary texts – remains an open question to explore in future studies.

Finally, it is important to consider that in this work we have intentionally ignored the dimension of time: we are interested in which features distinguish canonical, awarded, best-selling texts etc., independently from their distribution within the corpus. This means that we are treating our models as an abstract "contemporary reader" who, in front of the corpus, reacts to the canonical group differently from the bestseller group and so on. In future, we intend to observe how the features we have studied in this work correlate with time.

[16] Naturally, there are important overlaps, as these are intermingling categories. *The Hobbit* features in our list of canonical works, and so does Twain's *The Prince And The Pauper*.

References

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

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

Jodie Archer and Matthew Lee Jockers. 2017. *The bestseller code*. Penguin books, London.

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1753–1764.

Derek Attridge. 2004. *The Singularity of Literature*. Routledge, London; New York.

Laura Winther Balling and R. Harald Baayen. 2012. Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, 125(1):80–106.

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

Douglas Biber and Bethany Gray. 2011. The historical shift of scientific academic prose in english towards less explicit styles of expression. *Researching specialized languages*, 47(11).

Yuri Bizzoni, Stefania Degaetano-Ortlieb, Peter Fankhauser, and Elke Teich. 2020. Linguistic variation and change in 250 years of english scientific writing: A data-driven approach. *Frontiers in Artificial Intelligence*, 3:73.

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

Yuri Bizzoni, Pascale Moreira, Nicole Dwenger, Ida Lassen, Mads Thomsen, and Kristoffer Nielbo. 2023a. Good Reads and Easy Novels: Readability and Literary Quality in a Corpus of US-published Fiction. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 42–51, Tórshavn, Faroe Islands. University of Tartu Library.

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

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

Yuri Bizzoni, Pascale Feldkamp Moreira, Mads Rosendahl Thomsen, and Kristoffer L. Nielbo. 2023d. The Fractality of Sentiment Arcs for Literary Quality Assessment: the Case of Nobel Laureates. *Journal of Data Mining & Digital Humanities*, NLP4DH:11406.

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

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

Pierre Bourdieu. 1993. *The field of cultural production: essays on art and literature*. Columbia University Press, New York.

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

Pascale Casanova. 2007. *The World Republic of Letters*. Convergences: Inventories of the Present. Harvard University Press, Cambridge, MA.

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

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

Stefania Degaetano-Ortlieb, Katrin Menzel, and Elke Teich. 2019. Typical linguistic patterns of english history texts from the eighteenth to the nineteenth century. *Writing History in Late Modern English: Explorations of the Coruña Corpus*, pages 58–81.

Stefania Degaetano-Ortlieb and Elke Teich. 2022. Toward an optimal code for communication: The case of scientific english. *Corpus Linguistics and Linguistic Theory*, 18(1):175–207.

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

Julian Fritsch, Sebastian Wankerl, and Elmar Noth. 2019. Automatic Diagnosis of Alzheimer's Disease Using Neural Network Language Models. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5841–5845, Brighton, United Kingdom.

Yoav Goldberg. 2022. *Neural network methods for natural language processing*. Springer Nature.

Meritxell Gonzalez. An Analysis of Twitter Corpora and the Differences between Formal and Colloquial Tweets. In *"Proceedings of the Tweet Translation Workshop 2015"*, pages 1–7, Alicante, Spain. CEUR-WS.org.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

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

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–86, Online. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Thomas N Huckin. 1993. Stylistic prescriptivism vs. expert practice. *Discourse and Writing/Rédactologie*, 11(2):17–Jan.

Tom S Juzek, Marie-Pauline Krielke, and Elke Teich. 2020. Exploring diachronic syntactic shifts with dependency length: the case of scientific english. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 109–119.

Justine Kao and Dan Jurafsky. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature*, pages 8–17.

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

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

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

Ida Marie S Lassen, Pascale Feldkamp Moreira, Yuri Bizzoni, Mads Rosendahl Thomsen, and Kristoffer L. Nielbo. 2023. Persistence of Gender Asymmetries in Book Reviews Within and Across Genres. In *CEUR Workshop Proceedings*, pages 14–28, Paris, France.

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

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Dallas Liddle. 2019. Could Fiction Have an Information History? Statistical Probability and the Rise of the Novel. *Journal of Cultural Analytics*, page 22.

Hoyt Long and Teddy Roland. 2016. Us novel corpus. Technical report, Textual Optic Labs, University of Chicago.

Suraj Maharjan, John Arevalo, Manuel Montes, Fabio A. González, and Thamar Solorio. 2017. A multi-task approach to predict likability of books. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1217–1227, Valencia, Spain. Association for Computational Linguistics.

Alexander Manshel, Laura B McGrath, and J.D. Porter. 2019. Who Cares about Literary Prizes?

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

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics*, 47(1):141–179. Place: Cambridge, MA Publisher: MIT Press.

Laura McGrath, Devin Higgins, and Arend Hintze. 2018. Measuring Modernist Novelty. *Journal of Cultural Analytics*, 3(1).

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

Alessio Miaschi, Chiara Alzetta, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Is Neural Language Model Perplexity Related to Readability? In *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020*, pages 303–309. Accademia University Press.

Jan Mukařovský. 1964. Standard language and poetic language. In Paul L. Garvin, editor, *A Prague School Reader on Esthetics Literary Structure, and Style*, pages 17–30. 1932. Georgetown University Press.

Byung-Doh Oh and William Schuler. 2021. Contributions of propositional content and syntactic category information in sentence processing. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 241–250, Online. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Willie van Peer, editor. 2008. *The quality of literature: linguistic studies in literary evaluation*. Number v. 4 in Linguistic approaches to literature. John Benjamins Publishing.

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

J.D. Porter. 2018. Popularity/prestige: A new canon. *Stanford Literary Lab*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Raphael Rubino, Stefania Degaetano-Ortlieb, Elke Teich, and Josef van Genabith. 2016. Modeling diachronic change in scientific writing with information density. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 750–761, Osaka, Japan. The COLING 2016 Organizing Committee.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Claude E Shannon. 1949. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21.

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

Viktor Shklovsky. 1917. Art as technique. In J Rivkin and M Ryan, editors, *Literary Theory: An Anthology*, pages 15–21. Blackwell Publishing Ltd.

William Strunk Jr and Elwyn Brooks White. 2007. *The Elements of Style Illustrated*. Penguin.

Ted Underwood. 2020. How predictable is fiction?

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

Andreas van Cranenburgh and Rens Bod. 2017. A data-oriented model of literary language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1228–1238, Valencia, Spain. Association for Computational Linguistics.

Robert von Hallberg. 1983. Editor's Introduction. *Critical Inquiry*, 10(1):iii–vi. Publisher: The University of Chicago Press.

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

Sigrid Weigel. 2016. Literature, literary criticism and the historical index of the readability of literary texts. *Social Sciences in China*, 37(3):175–185.

Yaru Wu. 2023. Predicting the unpredictable–using language models to assess literary quality. Master's thesis, Uppsala University.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2020. Trading off diversity and quality in natural language generation.

A Appendix

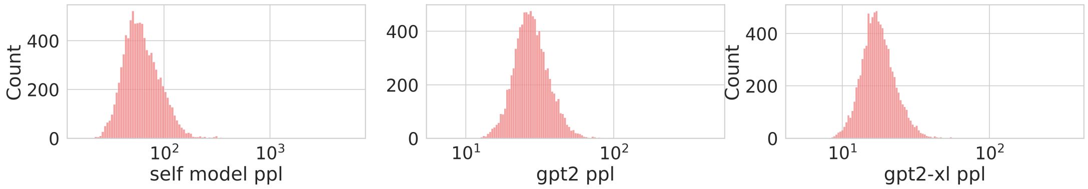

Figure 7: Histogram of the distribution of perplexity per model in our corpus. Note that perplexity has a log normal distribution.

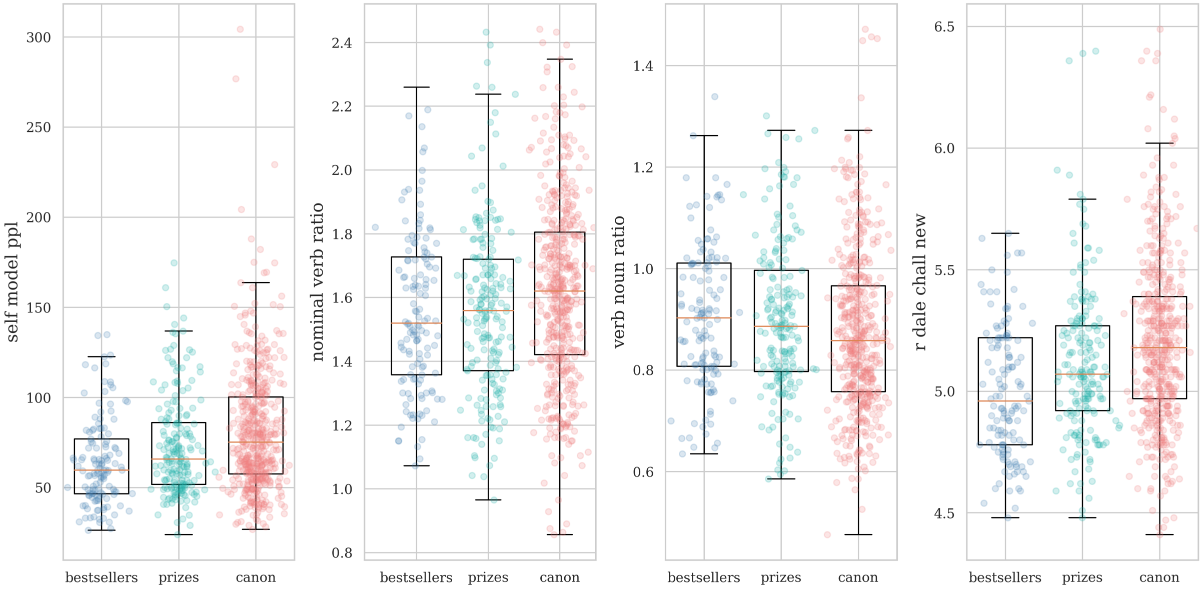

Figure 8: Nominal style features for each canon-type: Bestselling books, Prizewinning books, and books contained in one of the Canon lists. Note that outliers (points beyond the 99.5th percentile of our data) have been removed for this visualization.