Modeling Multilayered Complexity in Literary Texts

Pascale Feldkamp
CHC / Aarhus University
pascale.moreira@cc.au.dk

Márton Kardos
CHC / Aarhus University
martonkardos@cas.au.dk

Kristoffer L. Nielbo
CHC / Aarhus University
kln@cas.au.dk

Yuri Bizzoni
CHC / Aarhus University
yuri.bizzoni@cc.au.dk

Abstract

We explore the relationship between stylistic and sentimental complexity in literary texts, analyzing how they interact and affect overall complexity. Using a dataset of over 9,000 English novels (19th-20th century), we find that complexity at the stylistic/syntactic and sentiment levels tend to show a linear association. Finally, using dedicated datasets, we show that both stylistic/syntactic features – particularly those relating to information density – as well as sentiment features are related to text difficulty rank as well as average processing time.[1]

1 Introduction

Literary texts exemplify language operating at its most refined and demanding: they are capable of generating an experience – often emotional or evocative (Bizzoni and Feldkamp, 2024) – through the sheer force of words (Starr, 2013; Girju and Lambert, 2021; Miall and Kuiken, 1994). In this domain, language's capacity to evoke emotions, construct worlds, and create experiences is pushed to its limits. To do so, literary texts explore the boundaries of what human language can achieve in terms of expressiveness, depth, and evocative power. It manipulates form and meaning for its effects in a way that seems unmatched in other domains – exhibiting complexity at multiple levels, for example, matching an information-dense style with an unpredictable narrative.

Multidimensional complexity might also be the reason why traditional stylistic metrics for gauging the difficulty of a text – often developed for nonfiction – such as readability formulae, do not adequately capture the level of complexity of literary texts (Dalvean and Enkhbayar, 2018a); and might be a factor in why literary texts are associated with longer human processing times than nonfiction (Zwaan, 1991; Brysbaert, 2019).

This complexity, however, might not manifest uniformly at all levels: a literary story may be emotionally complex while maintaining a simplified syntax. This is why the problem of modeling complexity at different linguistic levels in literary language presents a particularly intriguing challenge. Understanding how linguistic complexity affects reader experience and whether there are trade-offs between formal and emotional aspects is critical in unraveling the cognitive demands and rewards associated with literary reading.

While many recent studies have sought to gauge the effect of stylistic and syntactic features of complexity for forms of reader appreciation (Brottrager et al., 2022; Barré et al., 2023; Wu et al., 2024; Bizzoni et al., 2023b; Wang et al., 2019; Koolen et al., 2020), the sentiment and emotional dimension has been an overlooked aspect of literary complexity. Complexity at this level is difficult to define. While a metric like simple sentiment standard deviation can be used to gauge the width of the 'sentiment palette' that authors are using in a novel, some more sophisticated measures for the complexity of novels' sentiment arcs – i.e., the trajectory of positive and negative valences across a story – have been developed in recent years, like the approximate entropy or the Hurst exponent of sentiment arcs (Bizzoni et al., 2021, 2022).

Very little work has explored the connection between these different levels of complexity: the relation between complexity at the stylistic level and complexity at the sentiment level. Moreover, little work has tested whether sentiment complexity behaves similar to stylistic and syntactic complexity in relation to reader experience. To address this

[1] To ensure reproducibility, all code and raw data are available at: https://github.com/centre-for-humanities-computing/literary_complexity

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 142–158
March 3-4, 2025

gap, we pose two research questions. Firstly:

RQ1: *What is the relationship between complexity features at different textual levels (e.g., stylistic/syntactic, and sentiment levels)?* We hypothesize two possible relationships between different levels of complexity:

H1a: *There is a trade-off between complexity at different levels, where, e.g., increased stylistic and syntactic complexity leads to "simplification" at the sentiment level.*

H1b: *Complexity features at different levels co-occur, so that, e.g., higher stylistic and syntactic complexity is associated with greater sentiment complexity.*[2]

The first two hypotheses carry different consequences. The first hypothesis (H1a) draws from the concept of 'cognitive compensation' observed in other domains, which suggests that optimized communication requires distributing readers' cognitive load across linguistic layers. For example, when lexical complexity increases, syntactic structures may simplify to balance cognitive demands (Degaetano-Ortlieb and Teich, 2022). In this scenario, complexity at one level could functionally balance complexity at another – for instance, syntactic complexity might work alongside sentimental simplicity. In contrast, H1b derives from the idea that aesthetic phenomena function as 'supernormal stimuli', intentionally amplifying complexity across levels to heighten engagement, eliciting amplified responses (Dubourg and Baumard, 2022; Costa and Corazza, 2006). This scenario also carries the interesting possibility that works with high stylistic and syntactic complexity also embrace challenging sentiment profiles. Heightened complexity at multiple levels would impose a higher cognitive load on readers, yet could foster a more compelling aesthetic experience.

Secondly, we seek to probe the relation of each feature level to actual reader experience:

RQ2: *What is the relationship between complexity features at different levels of a text and cognitive load experienced by readers?*

H2a: *Features at the sentiment level behave like stylistic and syntactic features in increasing readers' cognitive load, impacting the reader's ability to process the text.*

H2b: *Features at the sentiment level have an inverse behavior to stylistic and syntactic features, so more complexity at the sentiment level decreases readers' cognitive load.*[3]

Through these questions, we aim to explore how complexity at different linguistic levels might enhance or compromise one another. In a first part of this study, we investigate the relationship between stylistic/syntactic and sentiment complexity (RQ1) in a large corpus of novels. In the second part, we assess whether sentiment complexity mirrors stylistic/syntactic complexity in its impact on readers' cognitive load (RQ2), using dedicated datasets on reading time and novels' difficulty rank.

2 Related Works

Computational literary analyses have long attempted to model textual complexity by analyzing both stylistic and syntactic features. As early as 1893, Sherman used sentence length to study textual complexity. The increasing prominence of Digital Humanities in recent decades has greatly expanded this field. Recent studies have focused on canonical literature (Barré et al., 2023; Brottrager et al., 2022; Wu et al., 2024; Algee-Hewitt et al., 2016), showing that such texts exhibit a higher level of complexity across various dimensions. For example, studies have demonstrated that canonical works tend to have denser nominal styles, lower readability levels, and less predictable sentiment arcs (Wu et al., 2024; Bizzoni et al., 2023b).

Much of the focus on stylistic and syntactic complexity can be traced to formalist literary theory, which emphasizes *stylistic discomfort* as a hallmark of the *literariness* of texts. This theory argues that literary texts slow down reading by creating linguistic unfamiliarity or "foregrounding" (Mukařovský, 1964; van Peer, 1986). While some work has found reader consensus on foregrounding phenomena (van Peer, 1986), no comprehensive taxonomy of such features exists. Still, such features have been implicitly assumed to be formal or stylistic. This aligns with a long-standing debate on formalism in literary analysis, where a superficial focus on form has been claimed to overshadow content (Eagleton, 1983). As an exception, the experimental study of Miall and Kuiken (1994) found that reading times increased with the

[2]The null hypothesis (1) would naturally be that these levels bear no relation to each other, i.e., are independent.

[3]The null hypothesis (2) would naturally be that sentiment features bear no relation to readers' cognitive load.

frequency of foregrounding features, including affective features in their taxonomy.

Other, more theoretical studies have suggested that the extended processing time associated with literary texts (Zwaan, 1991) is linked to emotional and emphatic engagement (Scapin et al., 2023; László and Cupchik, 1995) and to increased reflection on non-literal meaning distinctive to literary reading (Hakemulder, 2020). In short, the complexity of literary texts may evoke more cognitively demanding affective processes than nonfiction, echoing the idea of literary texts as enhanced stimulus objects (Dubourg and Baumard, 2022). Moreover, recent psycholinguistic research has also emphasized how sentiment and emotional engagement affect readers' cognitive load, showing that negative valence and emotional features can increase reading times and that readers respond rapidly to valence cues (Pfeiffer et al., 2020; Lei et al., 2023; Arfé et al., 2023). These studies suggest that sentiment plays a critical role in reader experience, yet few works have explored the intersection of stylistic, syntactic, and complexity at the sentiment level.

While sentiment analysis (SA) has become a popular method for gauging emotional content in texts (Rebora, 2023), its application in literary analysis remains conceptually and theoretically underdeveloped. Some recent work has applied complexity measures such as approximate entropy and the Hurst exponent to sentiment arcs, suggesting that these measures provide insight into the complexity of narratives at the level of feelings or emotions evoked (Bizzoni et al., 2021, 2022). Yet, the connection between complexity at the stylistic level and the complexity in sentiment trajectories measured by these metrics remains largely unexplored.

We seek to fill this gap by investigating the relationship between stylistic/syntactic complexity and complexity at the sentiment level in literary texts, contributing to the broader understanding of how complexity at different linguistic levels interacts to shape the complexity profile of literature and its readers' cognitive experience.

3 Methods

3.1 Data

The Chicago Corpus

For our investigation on the relation between features, we use the *Chicago Corpus* of novels in English ($n = 9,089$) from the period 1880-2000 (see the distribution of the corpus over time in Fig. 1). The novels in our corpus are predominantly by anglophone authors, selected based on the number of worldwide library holdings,[4] favoring those with broader representation. Since library holdings capture both popular demand and prestigious, curated literature, the corpus spans a diverse range of genres – from Agatha Christie to James Joyce.[5][6]

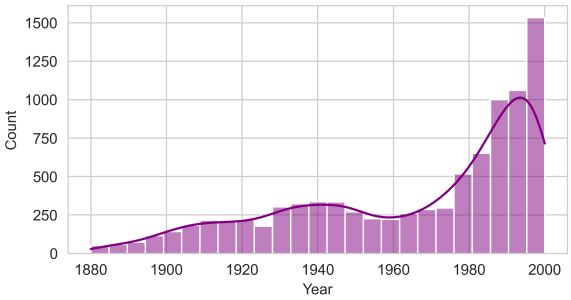


Figure 1: *Chicago Corpus*, temporal distribution of novels.

Beyond the *Chicago Corpus*, we use two dedicated datasets for part II of our study, where we gauge the relation between features at different levels with proxies of perceived complexity – i.e., reading time from the *Natural Stories corpus* and a list of the difficulty rank of novels (Dalvean and Enkhbayar, 2018a).

Natural Stories Corpus

The *Natural Stories corpus* consists of 10 English stories, each approximately 1,000 words long, totaling 485 sentences. These publicly available narratives, which includes tales by the Brothers Grimm, were revised to incorporate low-frequency and psycholinguistically interesting constructions while maintaining fluency. Self-paced reading (SPR) data was collected from 181 native English speakers, recording reaction times (RTs) for each word in a moving window setup. The dataset was filtered for control comprehension questions and outlier RTs ($< 100ms$ or $>$

[4] As indexed in `worldcat.org`

[5] See Bizzoni et al. (2024c) for details on the corpus. Recent studies of literary complexity have also used it, such as Wu et al. (2024).

[6] The feature dataset – though not full texts – is available at: `https://github.com/centre-for-humanities-computing/chicago_corpus`

$3000ms$).[7] Note that our analysis operates at the story level, using average sentence RT, as we examine sentiment features based on broader contexts. Average sentence RT per story was calculated from the word RTs.

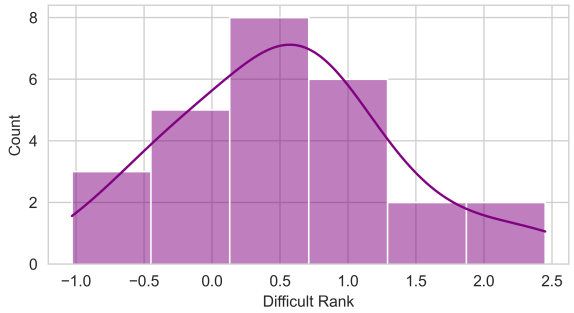


Figure 2: Distribution of Difficulty Rank across the 26 novels.

Difficulty rank of novels

With the aim of matching books to appropriate reader levels, Dalvean and Enkhbayar (2018a) curated a list of 200 novels, each assigned a difficulty rank. This rank is derived from a model trained on a binary prediction task (accuracy 89%) based on 48 linguistic and psycholinguistic features. We use these scores to estimate text complexity for the subset of books extant in the list and in the Chicago corpus, i.e., 26 novels (see Fig. 2). For the titles of the 26 novels, see Table 6 in Appendix B.

3.2 Features

The features utilized in this study have been used in previous works to distinguish textual profiles of different types of literature. The details on each measure can be found in Appendix D (Table 10). We focus on features that supposedly reflect stylistic or syntactic complexity, and have been widely used in recent computational literary studies. Features at the sentiment level were chosen to focus on overall variation and local and global complexity of the sentiment arc (Bizzoni et al., 2023b, 2022).

The sentiment dynamics central to our study are captured by both simple and complex measures. First, sentiment standard deviation (SD) represents the "palette" of sentiment in a novel, quantifying the overall variation in valence scores across sentences to reflect sentiment range. Beyond this, two advanced measures – approximate entropy and the Hurst exponent – are applied to model more nuanced sentiment arcs linearly within a narrative.

Approximate entropy ($ApEn$) assesses the local complexity and unpredictability within sentiment flows, where lower values signal a repetitive, predictable structure, and higher values indicate intricate, less predictable patterns in the narrative (Mohseni et al., 2022). To capture global coherence, we estimate the Hurst exponent (H) with adaptive fractal analysis (AFA) instead of the more commonly used detrended fluctuation analysis (DFA), avoiding the boundary errors and segment discontinuities common to DFA (Hu et al., 2021; Gao et al., 2011). By accounting for nonlinear trends, AFA enables a smooth global trend, with higher H values suggesting sustained narrative coherence and lower values indicating more abrupt sentiment shifts across scales (Hu et al., 2021; Bizzoni et al., 2023d).[8]

For all sentiment features, which are derived from valence scores, we first annotated all novels at the sentence level for sentiment valence (where 1 represents the positive and -1 the negative polarity) using the *Syuzhet* package (Jockers, 2015). This tool was developed explicitly for literary language, and has shown the best performance for English in the literary domain, also compared to transformer-based models (Bizzoni et al., 2023a). We then calculated the standard deviation, $ApEn$, and Hurst exponent of sentiment arcs for all 9,000 *Chicago Corpus* novels, as well as stories of the *Natural Stories* dataset – taking these features to represent the variance, as well as the local and global predictability – in other words, complexity – of novels' sentiment profile.

In the following first part of this study, we juxtapose stylistic/syntactic and these sentiment features of complexity across all novels, gauging the correlation between them. We then assess the link between stylistic/syntactic and sentiment levels by trying to predict individual sentiment variables using all the stylistic/syntactic features. This is done on the whole set of over 9,000 novels, making it the largest-scale experiment in this study, as well as the most comprehensive diachronically (end of 19^{th} – 20^{th} century).

[7] The *Natural Stories* data is available at: `https://github.com/languageMIT/naturalstories`

[8] See, recently, Bizzoni et al. (2024b) for the details on the computation of these sentiment measures.

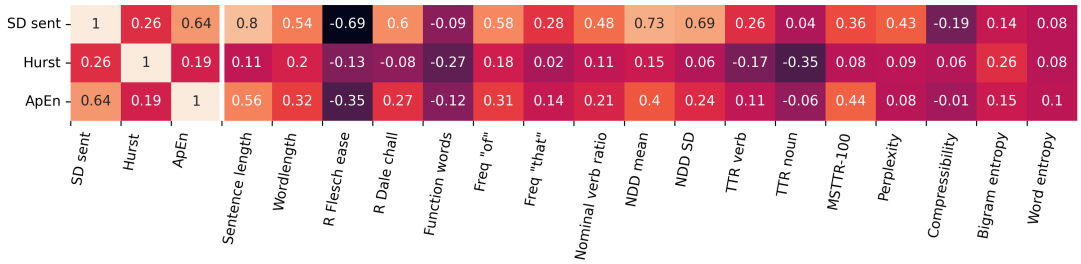


Figure 3: The correlation (Spearman's ρ) between stylistic/syntactic features and sentiment features. See table 10 in Appendix D for details on the computation of these features and for the label explanations.

3.3 Reading time & Difficulty rank

Features such as readability formulae are established indicators of textual complexity, but sentiment-based features are less studied and their impact on reading time remains unclear. Therefore we relate these features to perceived complexity, taking both reading time and text difficulty rank as proxies of perceived complexity associated with increased cognitive load for the reader.

To assess the relationship between the analyzed features and reader processing time, we first evaluate how well these features correlate with reaction times (RTs) from the *Natural Stories* corpus. This initial step provides indicators of how these features may influence cognitive processing and perceived text complexity.

As a second check, we address the absence of RTs for the novels in the Chicago Corpus by using a scoring list of 200 novels (Dalvean and Enkhbayar, 2018a).[9] This list assigns a difficulty rank to 26 Chicago Corpus novels, which serves as a proxy for perceived difficulty. By predicting difficulty rank with our feature sets, we aim to further assess the role of sentiment features in the perceived difficulty of literary texts.

4 Results & Discussion

4.1 Part I: Relations between stylistic/syntactic & sentiment features

In part I of this study, we examined feature relations in the novels. We observe a strong correlation between sentiment-level features and a subset of stylistic/syntactic features, as shown in Fig. 3. Notably, readability formulas, word and sentence length, dependency length, lexical richness ('MSTTR'), indicators of heavy nominal style (e.g., frequency of "of" and nominal verb ratio), and LLM perplexity – all features commonly associated with harder-to-process and information-rich text – show a particularly strong correlation with sentiment standard deviation. Approximate entropy also displays a similar pattern of correlation with these features, while it appears less correlated with LLM-based perplexity. Additionally, the Hurst exponent, which captures global uncertainty, shows a relationship with these complexity metrics – not least do the sentiment features exhibit correlations internally ($.19 < \rho > .64$).

Most correlations across sentiment features align in the same direction; for instance, lower Flesch Ease readability (indicating lesser readability) correlates with higher sentiment arc entropy ($ApEn$) ($\rho = -.35$), higher sentiment standard deviation ($\rho = -.69$), and a tendentially higher Hurst exponent ($\rho = -.13$). For a more comprehensive view of correlation co-directionality, see the visualizations in Appendix A, Fig. 7.

Note that all sentiment features show a correlation with sentence length, which may partly explain their relationship with sentence-length-dependent metrics, such as readability indices (R Flesch Ease and R Dale-Chall). However, sentiment features are also clearly related to features that bear no relation to sentence length, such as the frequency of the use of "of", indicating a more nominal (viz. information dense) writing style (Wu et al., 2024), or average and SD of the dependency length.

Given these strong correlations, we employed a linear regression model to determine whether stylistic/syntactic complexity features could predict sentiment-level complexity, particularly sentiment standard deviation. Results show that textual complexity features are indeed predictive of sentiment complexity (Table 1), with sentiment standard deviation exhibiting the strongest predic-

[9]The list is available in Appendix 2 of Dalvean and Enkhbayar (2018b), and in the repository of our paper.

Feature	F-stat	R^2	adj. R^2
Sentiment SD	1803.0	0.787	0.786
ApEn	364.2	0.427	0.426
Hurst	123.1	0.201	0.2

Table 1: Linear regression of sentiment features **based on stylistic/syntactic** features. Here for all, $p < 0.01$.

Feature	F-stat	R^2	adj. R^2
Flesch Ease Readability	2717.0	0.481	0.481
Dependency Length	4166.0	0.587	0.587
Nominal Ratio	1117.0	0.276	0.275

Table 2: Linear regression **based on sentiment features** to predict a stylistic/syntactic feature. Here, all $p < 0.01$.

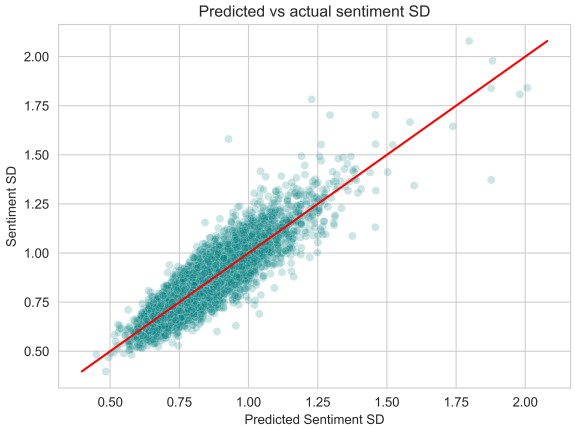


Figure 4: Linear fit between the predicted and actual **sentiment SD** based on the stylistic/syntactic complexity features.

tive relationship (Fig. 4). Interestingly, this relationship is bidirectional: sentiment features also demonstrate predictive power for stylistic and syntactic complexity features, with sentence length, readability formulae, dependency length (avg. & SD) and features like the frequency of "of", indicating nominal style, displaying the strongest predictive relationships. See a few selected features in Table 2, and a full table in Appendix A, Table 5. This finding underscores a tightly coupled relationship between stylistic/syntactic complexity and sentimental variability, reinforcing hypothesis H1b: higher stylistic and syntactic complexity is associated with increased complexity at the sentiment level. This suggests that stylistic and affective dimensions in literary texts are interdependent, potentially amplifying each other's complexity in ways that may shape readers' engagement.

4.2 Part II: Relation of features to proxies of perceived complexity

In part II of this study, to examine the relationship between features and perceived complexity, we conducted two experiments. The first used RTs (reading times) from the *Natural Stories* corpus, compared to the same features as before,[10] computed across the dataset's ten stories. The second experiment involved analyzing the difficulty rank of 26 novels from the *Chicago Corpus*. In both cases, we aimed to predict reading time and difficulty rank by exploring correlations between the features and these variables. We employed linear regression based on stylistic/syntactic and sentiment feature sets, using each set separately and then jointly.

Given the relatively small sample sizes in both experiments (10 and 26 data points, respectively), we aimed to strengthen our findings by reducing collinearity in the feature set. To achieve this, we first applied PCA to the entire *Chicago Corpus* to capture the covariance structure and scaling of variables in a larger, more representative dataset. We then applied this PCA model to reduce dimensionality in our smaller dataset, minimizing the risk of overfitting to limited data. Details on this sanity check using PCA for collinearity reduction are presented in Appendix C: for difficulty rank in table 8 and for reading times in table 9.

4.2.1 Reading time

In relating features to reading times, we find that only some stylistic/syntactic and sentiment features exhibit linear correlations with reading time of the stories. These include lexical richness ('MSTTR'), word entropy, and nominal ratio.

This scarcity of correlation might be due to insufficient datapoints. In a setting with augmented datapoints, the mentioned features remain significantly correlated, while we also see the p-value of sentiment SD and compressibility rising above the significance threshold (.05). For the augmented data setting, see Appendix B, Fig. 9. We show the correlation of the original data for lexical richness, nominal ratio and sentiment SD in Fig. 5.

[10] We excluded perplexity, as we could not ensure that publicly available stories were excluded from model training data. For the *Chicago Corpus*, perplexity derives from a self-trained model controlling for overlap (Wu et al., 2024).

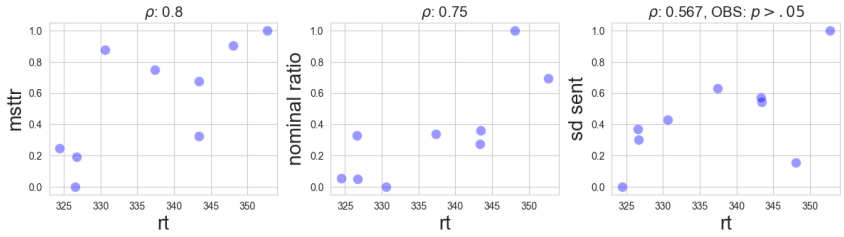


Figure 5: Correlation of selected features with RT, with Spearman's ρ at the top of plots. Note that for sentiment SD, $p > .05$.

Moreover, correlations between features and RTs tend to be nonlinear, as some features, like readability formulae seem to show clustering both in the original and augmented data setting (see Appendix B, Figs. 8 and 9), but no *linear* correlation. Fig. 5 shows the correlation of RT and selected features. Note that while the correlation has $p > .05$, a tendential association of sent SD and RT can be observed. A larger corpus of annotated fiction is required to robustly confirm this tendency.

Features	**F-stat**	R^2	**adj.** R^2	**p-val**
Styl/Synt	15.38	0.902	0.844	< 0.01
Sentiment	2.01	0.547	0.275	0.231
All	28.76	0.945	0.912	< 0.01

Styl/Synt	*Bigram entropy, Nominal ratio, TTR Noun*
Sentiment	All sentiment features used
All	*Nominal ratio, Frequency "of", Sent SD*

Table 3: Linear regression **predicting RTs** of the *Natural Stories* using two feature sets, the three sentiment features, the three selected stylistic/syntactic features, and three selected features among all features. Below, the selected features in each category using RFE.

As the sample was too scarce, linear regression could not be carried out using the full feature set. Instead, we used Recursive Feature Elimination (RFE) to determine 3 features in the stylistic/syntactic category, and 3 out of all features.[11] We thus stay at the number of features corresponding to our number of sentiment features. Results of using linear regression to predict RT are shown in table 3. Notably, RFE leads to selecting sentiment SD as one of the overall top 3 significant features. Considering the scarce data, we consider this a means of comparing feature categories rather than an accurate model, i.e., for predicting RTs on unseen samples.

[11] RFE was performed using sklearn: `https://scikit-learn.org/dev/modules/generated/sklearn.feature_selection.RFE.html`

4.2.2 Difficulty rank

Using the 26 books in Chicago that had an assigned score in the difficulty ranking list, we sought to use different feature categories to predict the score of the novel. Results are shown in table 4. Note that visualizations of the predicted/actual values in Fig. 6 reflect an apparent improvement in our models' predictive power when adding sentiment features to it. As in the reading time experiment, we do not claim any predictive power of this model but observe the effect of adding sentiment features for gauging difficulty rank.

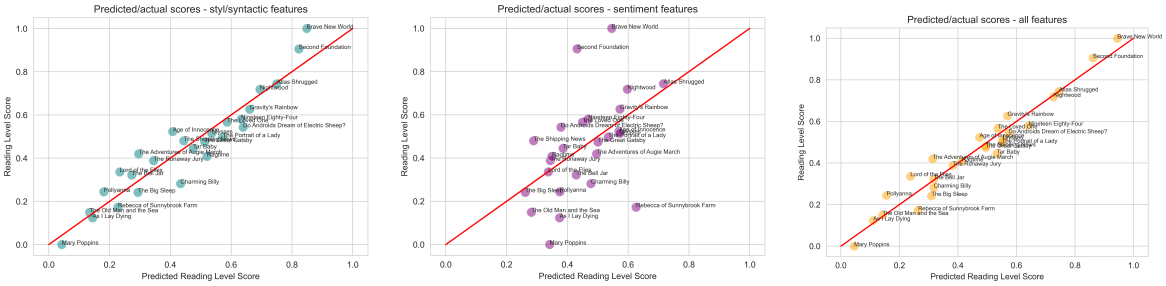
Features	**F-stat**	R^2	**adj.** R^2	**p-val**
Styl/Synt	3.234	0.873	0.603	0.048
Sentiment	2.469	0.252	0.150	0.089
All	3.413	0.932	0.659	0.089

Table 4: Linear regression predicting **difficulty rank** using two feature sets, and all features.

Note that the p-value tends to be high when using all features, probably due to the limited amount of datapoints (table 4). However, predicted and actual difficulty rank in the sentiment-based model still exhibit a relation (Fig. 7(b)) and the model seems to improve when sentiment features are added (Fig. 7(c)). As in the previous experiment with RT, we also selected features with RFE (see Appendix B, table 7). Here, the features: frequency "of", nominal ratio, word entropy, and perplexity appeared to be the most important, without sentiment features showing up among the 3 selected features.

5 Conclusion

Our results pertaining to our first question (RQ1) support H1a. Rather than a balance between dif-



(a) Predicted/actual difficulty rank using stylistic/syntactic features.

(b) Predicted/actual difficulty rank using sentiment features.

(c) Predicted/actual difficulty rank using all features.

Figure 6: Comparison of the predicted vs actual difficulty rankings using different feature sets.

ferent aspects of language, we find that, at least over the whole *Chicago Corpus*, complexity at the stylistic and syntactic level tends to correspond to complexity at the sentimental level.

Regarding our second question (RQ2), our findings support H2a: it seems that both the stylistic/syntactic and the sentimental complexity impact the cognitive load of the readers, not only when measuring whole novels but even within much shorter stories. We should thus assume that the novels that push both levels to higher complexity are indeed asking more from the readers, and are providing a more challenging experience. It's not obvious that features like overall variance in sentiment (sentiment SD) and the local and global linear dynamics of the sentiment arc ($ApEn$, Hurst) would relate to perceived complexity, making this finding particularly intriguing.

The question then remains as to why these two levels of complexity are tendentially intensified together, rather than showing a trade-off. In other words, why do works that offer a wider sentimental palette or a less predictable story arc also have, in general, a higher noun-to-verb ratio and a wider vocabulary?

Such literature – complex on multiple levels – may offer higher-quality reading experiences by amplifying both emotional and stylistic profiles. In this way, our findings suggest that literature may function as a "supernatural stimulus", where every element is intensified simultaneously – a phenomenon that, while possibly engaging in fiction, would be counterproductive in nonfiction or more didactic texts, where clarity and ease are often prioritized. This distinction potentially sets literary texts apart from other domains, though future studies should more rigorously test differences between literary and nonliterary texts with regard to multidimensional complexity.

The possibility of a trade-off between these dimensions of language is not off the table: it might occur within specific groups of texts with varying degrees of difficulty; and it is also possible that specific works of literature strike a balance differently, depending on their intended audience and the author's specific style. But in general, our findings suggest that rather than being independent dimensions, style and "content" – taking sentiment here as a semantic element – might have a strong relation in literary texts. The style of the texts might have to align with its semantics, at least at the sentimental level. In this sense, the relevance of linguistic traits associated with the "nominal style" is particularly intriguing. Degaetano-Ortlieb and Teich (2022) has shown that this style is developed and applied in scientific and technical language to convey semantic information more efficiently while requiring a higher degree of concentration and preparation from the reader. This "optimal" strategy of linguistic communication might not be limited to technical prose but be exploited, despite their completely different aims, by literary works as well. In other words, it is possible that some aspects of complexity at the stylistic level are necessary for most works of art to *convey* the complexity of the sentimental level in a manner that is most effective to the creation of a powerful reading experience.

In the future, we intend to explore the relationship between these levels of complexity in literary language further, better formalizing the relation and role of each of the selected components. We would also examine the relationship between the perceived complexity or difficulty of a text and these features in an experiment setting.

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.

Barbara Arfé, Pablo Delatorre, and Lucia Mason. 2023. Effects of negative emotional valence on readers' text processing and memory for text: an eye-tracking study. *Reading and Writing*, 36(7):1743–1768.

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

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

Yuri Bizzoni and Pascale Feldkamp. 2024. Below the sea (with the sharks): Probing textual features of implicit sentiment in a literary case-study. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 54–61, Malta. Association for Computational Linguistics.

Yuri Bizzoni, Pascale Feldkamp, Ida Marie Lassen, Mia Jacobsen, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2024a. Good Books are Complex Matters: Gauging Complexity Profiles Across Diverse Categories of Perceived Literary Quality. ArXiv:2404.04022 [cs].

Yuri Bizzoni, Pascale Feldkamp, and Kristoffer Laigaard Nielbo. 2024b. Global Coherence, Local Uncertainty: 2024 Computational Humanities Research Conference, CHR 2024. In *Proceedings of the Computational Humanities Research Conference 2024*, Aarhus Denmark. CEUR Workshop Proceedings. Publisher: CEUR-WS.org.

Yuri Bizzoni, Pascale Feldkamp Moreira, Emily Öhman, and Kristoffer L. Nielbo. 2023a. Comparing Transformer and Dictionary-based Sentiment Models for Literary Texts: Hemingway as a Case-study. In *NLP4DH (forthcoming)*, Tokyo, Japan.

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, Tórshavn, Faroe Islands. University of Tartu Library.

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, Ida Marie S. Lassen, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2024c. A matter of perspective: Building a multi-perspective annotated dataset for the study of literary quality. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 789–800, Torino, Italia. ELRA and ICCL.

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.

Yuri Bizzoni, Telma Peura, Kristoffer Nielbo, and Mads Thomsen. 2022. Fractality of sentiment arcs for literary quality assessment: The case of nobel laureates. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 31–41, Taipei, Taiwan. Association for Computational Linguistics.

Yuri Bizzoni, Telma Peura, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2021. Sentiment dynamics of success: Fractal scaling of story arcs predicts reader preferences. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 1–6, NIT Silchar, India. NLP Association of India (NLPAI).

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

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.

Marc Brysbaert. 2019. How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, 109. Place: Netherlands Publisher: Elsevier Science.

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

Marco Costa and Leonardo Corazza. 2006. Aesthetic Phenomena as Supernormal Stimuli: The Case of Eye, Lip, and Lower-Face Size and Roundness in Artistic Portraits. *Perception*, 35(2):229–246. Publisher: SAGE Publications Ltd STM.

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.

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

Michael Dalvean and Galbadrakh Enkhbayar. 2018a. Assessing the readability of fiction: A corpus analysis and readability ranking of 200 English fiction texts. *Linguistic Research*, 35:137–170.

Michael Dalvean and Galbadrakh Enkhbayar. 2018b. A New Fiction Text Complexity Metric for Ranking Fiction Texts. pages 1–29.

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.

Edgar Dubourg and Nicolas Baumard. 2022. Why and How Did Narrative Fictions Evolve? Fictions as Entertainment Technologies. *Frontiers in Psychology*, 13. Publisher: Frontiers.

Terry Eagleton. 1983. *Literary Theory: An Introduction*, later printing edition edition. Univ of Minnesota Pr.

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

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

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

Jianbo Gao, Jing Hu, and Wen-wen Tung. 2011. Facilitating joint chaos and fractal analysis of biosignals through nonlinear adaptive filtering. *PLoS ONE*, 6(9).

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

Roxana Girju and Charlotte Lambert. 2021. InterSense: An Investigation of Sensory Blending in Fiction. ArXiv:2110.09710 [cs].

Frank Hakemulder. 2020. Finding Meaning Through Literature. *Anglistik*, 31(1):91–110. Publisher: Universitätsverlag WINTER GmbH Heidelberg.

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

Qiyue Hu, Bin Liu, Mads Rosendahl Thomsen, Jianbo Gao, and Kristoffer L Nielbo. 2021. Dynamic evolution of sentiments in *Never Let Me Go* : Insights from multifractal theory and its implications for literary analysis. *Digital Scholarship in the Humanities*, 36(2):322–332.

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

Matthew L Jockers. 2015. Syuzhet: Extract sentiment and plot arcs from text. *Matthew L Jockers blog*.

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

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.

Anqi Lei, Roel M. Willems, and Lynn S. Eekhof. 2023. Emotions, fast and slow: processing of emotion words is affected by individual differences in need for affect and narrative absorption. *Cognition and Emotion*, 37(5):997–1005. Publisher: Routledge _eprint: https://doi.org/10.1080/02699931.2023.2216445.

Lei Lei and Matthew L. Jockers. 2020. Normalized Dependency Distance: Proposing a New Measure. *Journal of Quantitative Linguistics*. Publisher: Routledge.

J. László and Gerald Cupchik. 1995. The role of affective processes in reading time and time experience during literary reception. *Empirical Studies of the Arts*, 13:25–37.

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.

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

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

David S. Miall and Don Kuiken. 1994. Foregrounding, defamiliarization, and affect: Response to literary stories. *Poetics*, 22(5):389–407.

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

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

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.

Willie van Peer. 1986. *Stylistics and Psychology*. Croom Helm.

Christian Pfeiffer, Nora Hollenstein, Ce Zhang, and Nicolas Langer. 2020. Neural dynamics of sentiment processing during naturalistic sentence reading. *NeuroImage*, 218:116934.

Simone Rebora. 2023. Sentiment analysis in literary studies. A critical survey. *Digital Humanities Quarterly*, 17(2).

Giulia Scapin, Cristina Loi, Frank Hakemulder, Katalin Bálint, and Elly Konijn. 2023. The role of processing foregrounding in empathic reactions in literary reading. *Discourse Processes*, 60(4-5):273–293. Publisher: Routledge _eprint: https://doi.org/10.1080/0163853X.2023.2198813.

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

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

Sanja Stajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity? In *Proceedings of Workshop on natural language processing for improving textual accessibility*, pages 14–22, Istanbul, Turkey. Association for Computational Linguistics.

G. Gabrielle Starr. 2013. *Feeling Beauty: The Neuroscience of Aesthetic Experience*. The MIT Press.

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

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.

Yaru Wu, Yuri Bizzoni, Pascale Moreira, and Kristoffer Nielbo. 2024. Perplexing canon: A study on GPT-based perplexity of canonical and non-canonical literary works. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 172–184, St. Julians, Malta. Association for Computational Linguistics.

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

Rolf A. Zwaan. 1991. Some parameters of literary and news comprehension: Effects of discourse-type perspective on reading rate and surface structure representation. *Poetics*, 20(2):139–156.

A Relation between features

We attach the visualization of some correlations of stylistic/syntactic features with all three sentiment features (Fig. 7).

Additionally, the results of the extended linear regression are presented in table 5, where we sought to predict each stylistic/syntactic feature individually by sentiment features.

Styl/synt. feature	**F-stat**	R^2	**adj.** R^2
Sentence length	6862.0	0.7	0.7
Dependency SD	4295.0	0.594	0.594
Dependency Length	4166.0	0.587	0.587
Flesch Ease Readab.	2717.0	0.481	0.481
Dale-Chall Readab.	2655.0	0.475	0.475
"Of" Frequency	1651.0	0.36	0.36
Word length	1326.0	0.311	0.311
Nominal Verb Ratio	0.612	0.276	0.275
MSTTR	754.5	0.204	0.204
TTR Noun	494.0	0.144	0.144
TTR Verb	442.2	0.131	0.131
"That" Frequency	249.0	0.078	0.078
Bigram Entropy	225.9	0.071	0.071
Compressibility	166.0	0.054	0.053
Perplexity	147.3	0.048	0.047
Function words	146.0	0.047	0.047
Word Entropy	35.65	0.012	0.012

Table 5: Linear regression **based on sentiment features** to predict a stylistic/syntactic feature. The table is ordered by decreasing R^2. Here for all, $p < 0.01$.

B Reading time & difficulty rank

Here we present the full results of our analysis on the relationship between features and both reading times (RTs) and difficulty rank.

For the **reading time (RT)** experiment, additional correlation coefficients, including stylistic and syntactic feature levels, with RTs from the Natural Stories corpus are provided and visualized in Fig. 8. To increase data points, we further split the stories with a 90% overlap between segments, effectively duplicating the data points. This approach retains as much of the global structure of the stories as possible – a crucial factor for features like the Hurst exponent, which is sensitive to structural changes. A visualization of these correlations is shown in Fig. 9.

For relating features to **difficulty rank (DR)**, we took the overlap of titles between the list of novels in Dalvean and Enkhbayar (2018a) and the *Chicago Corpus*. These are listed in table 6.

Author	**Title**	**DR**
Aldous Huxley	*Brave New World*	2.45
Isaac Asimov	*Second Foundation*	2.12
Ayn Rand	*Atlas Shrugged*	1.56
Djuna Barnes	*Nightwood*	1.47
Thomas Pynchon	*Gravity's Rainbow*	1.15
George Orwell	*Nineteen Eighty-Four*	0.99
Evelyn Waugh	*The Loved One*	0.94
Philip K. Dick	*Do Androids Dream of Electric Sheep?*	0.86
Edith Wharton	*The Age of Innocence*	0.79
James Joyce	*Ulysses*	0.76
Henry James	*The Portrait of a Lady*	0.70
Annie Proulx	*The Shipping News*	0.64
F. Scott Fitzgerald	*The Great Gatsby*	0.62
Toni Morrison	*Tar Baby*	0.52
Saul Bellow	*The Adventures of Augie March*	0.43
E.L. Doctorow	*Ragtime*	0.39
John Grisham	*The Runaway Jury*	0.32
William Golding	*Lord of the Flies*	0.14
Sylvia Plath	*The Bell Jar*	0.09
Alice McDermott	*Charming Billy*	-0.05
Eleanor H. Porter	*Pollyanna*	-0.18
Raymond Chandler	*The Big Sleep*	-0.19
Kate Douglas Wiggin	*Rebecca of Sunnybrook Farm*	-0.43
Ernest Hemingway	*The Old Man and the Sea*	-0.51
William Faulkner	*As I Lay Dying*	-0.60
P.L. Travers	*Mary Poppins*	-1.03

Table 6: difficulty rank (DR)(not normalized) of 26 novels in the Chicago Corpus. difficulty rank descending.

As in the RT experiment, we carried out linear regression with Recursive Feature Elimination (RFE) for predicting DR, these results are presented in table 7.

Features	**F-stat**	R^2	**adj.** R^2	**p-val**
Styl/Synt	8.908	0.548	0.487	< 0.01
Sentiment	2.469	0.252	0.150	0.09
All	7.955	0.52	0.455	< 0.01
Styl/Synt	*Freq "of", Perplexity, Word Entropy*			
Sentiment	All sentiment features used			
All	*Freq "of", Nominal Ratio, Word Entropy*			

Table 7: Linear model **predicting difficulty rank** of novels using two feature sets, the three sentiment features, three selected stylistic/syntactic features, and three selected features among all features. Below, the selected features in each category using RFE.

C Collinearity reduction

To avoid overfitting our feature selection method to the small datasets in the regression models above, we fitted a PCA on the *Chicago Corpus* and projected features in the smaller regression datasets to its first 3 principal components. PCA also helps us avoid the curse of collinearity in regression models, therefore the reported statistics might be more representative of the features' true

predictive strength. For reading times, results are in table 8; for difficulty rank in table 9.

Features	**F-stat**	R^2	**adj.** R^2	**p-val**
Styl/Synt	13.0	0.629	0.581	< 0.01
Sentiment	6.050	0.441	0.368	< 0.01
All	12.74	0.624	0.575	< 0.01

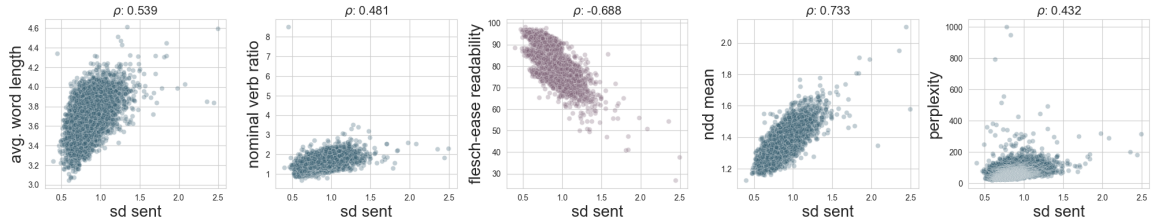
Table 8: Linear model **predicting difficulty rank** of novels using feature sets reduced for collinearity by fitting it to the *Chicago Corpus* PCA (3 components).

Features	**F-stat**	R^2	**adj.** R^2	**Prob. F-stat**
Styl/Synt	84.56	0.977	0.965	< 0.01
Sentiment	18.81	0.904	0.856	< 0.01
All	78.24	0.975	0.963	< 0.01

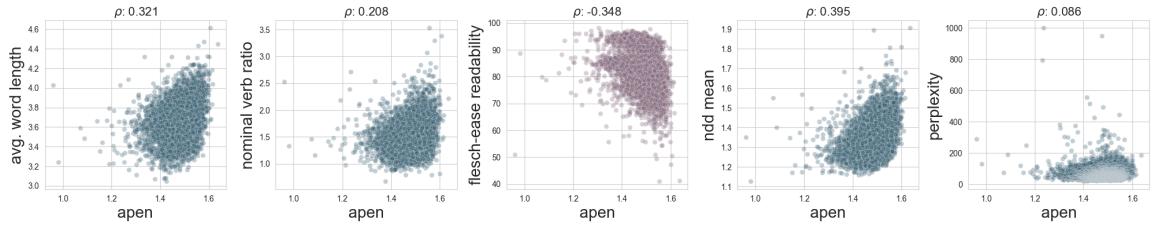
Table 9: Linear model **predicting reading time** of stories using feature sets reduced for collinearity by fitting it to the *Chicago Corpus* PCA (3 components).

D Features

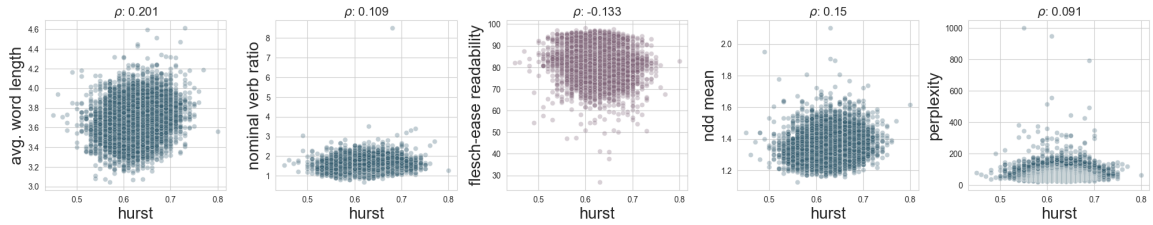
The full set of features with corresponding labels is indexed in table 10.



(a) Correlation between Sentiment SD and stylistic/syntactic features.

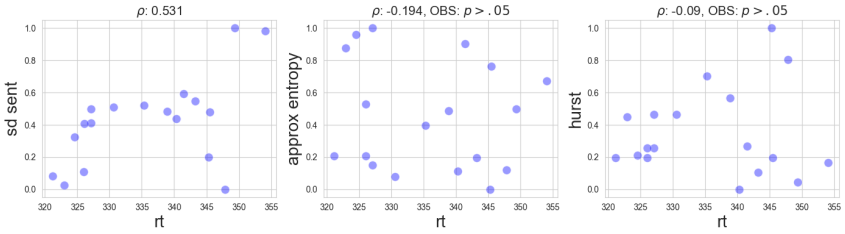


(b) Correlation between $ApEn$ and a few stylistic/syntactic features.

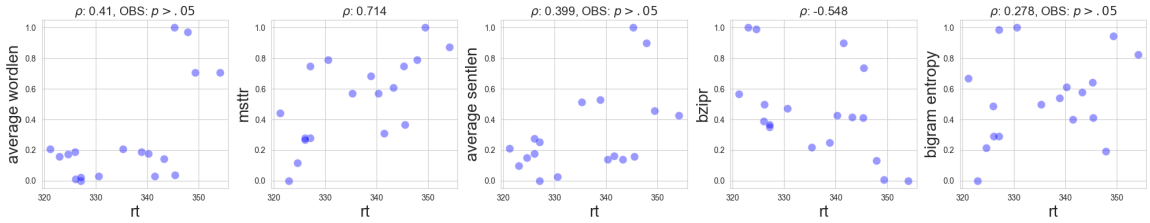


(c) Correlation between Hurst exponent and a few stylistic/syntactic features.

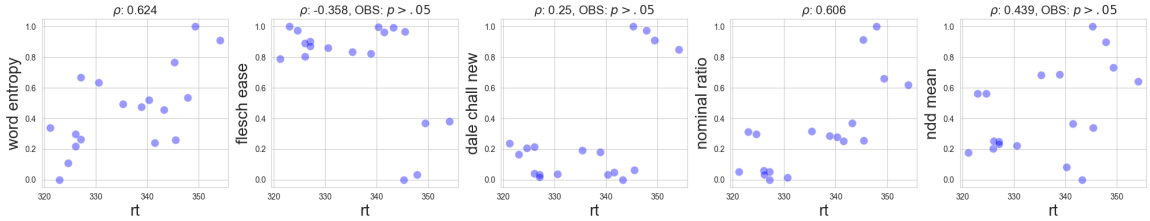
Figure 7: Correlation between **sentiment complexity features** and a few **stylistic or syntactic complexity features**. Note Spearman's ρ at the top of plots.



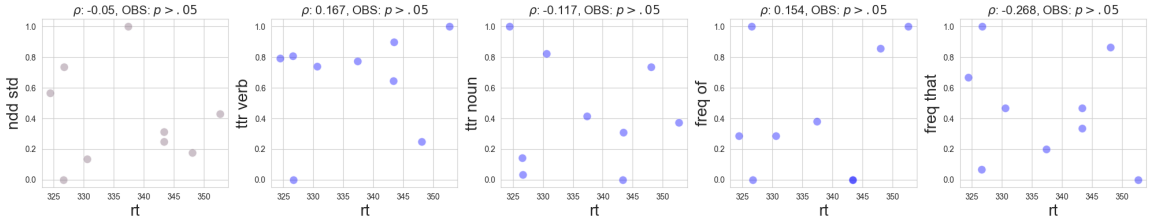
(a) Correlation between RT and **sentiment** features.



(b) Correlation between RT and stylistic/syntactic features.

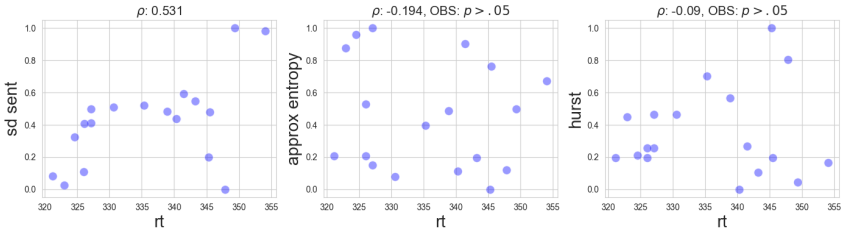


(c) Correlation between RT and stylistic/syntactic features.

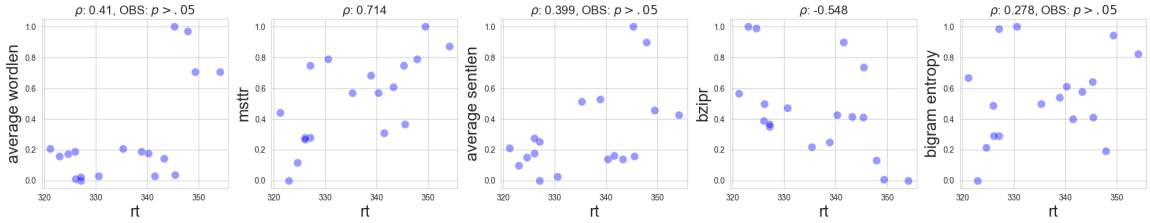


(d) Correlation between RT and stylistic/syntactic features.

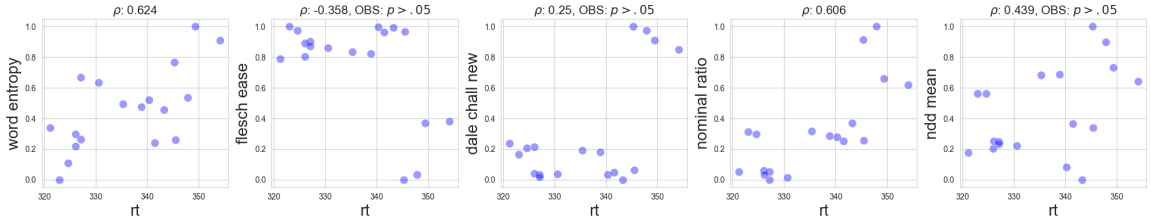
Figure 8: Full visualization of the correlation of features and RTs (10 stories).



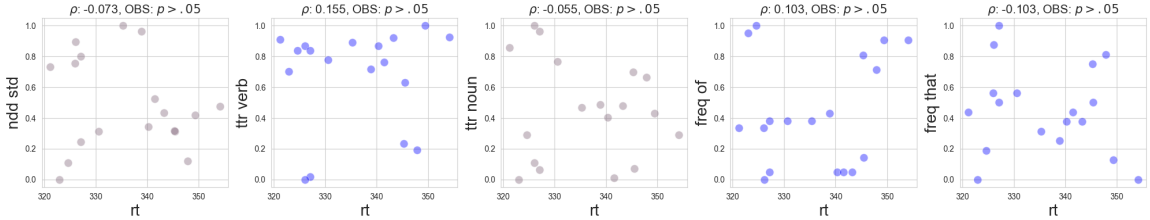
(a) Correlation between RT and **sentiment** features.



(b) Correlation between RT and stylistic/syntactic features.



(c) Correlation between RT and stylistic/syntactic features.



(d) Correlation between RT and stylistic/syntactic features.

Figure 9: Correlation features and RT, **augmented datapoints**. We split stories in two with a 90% overlap. This duplication of datapoints serve to show that the scarcity of correlations between features and RT may be due to a low number of datapoints (10 stories).

Feature	Description	Type	Reference
Type-Token Ratio (MSTTR-100), TTR Noun, TTR Verb	Measures lexical diversity by comparing the variety of words (types) to the total number of words (tokens), indicating a text's vocabulary complexity and inner diversity. A high TTR represents a richer prose: a higher diversity of elements and a lower lexical redundancy (Torruella and Capsada, 2013). TTR of nouns or of verbs quantifies the diversity within these Parts-of-Speech categories.[a]	Stylistic	Forsyth (2000)*, Kao and Jurafsky (2012)*, Algee-Hewitt et al. (2016), Maharjan et al. (2017), Koolen et al. (2020), Brottrager et al. (2022), Jacobs and Kinder (2022), Bizzoni et al. (2023c)
Readability (R Flesch Ease, R Dale Chall)	Estimate reading difficulty based variously on sentence length, syllable count, and word length/difficulty. Assessed using five different classic formulae that remain widely used (Stajner et al., 2012).[b]	Stylistic	Martin (1996), Garthwaite (2014), Maharjan et al. (2017), Febres and Jaffe (2017), Zedelius et al. (2019)*, Berger et al. (2021)*, Brottrager et al. (2022), Bizzoni et al. (2023b)
Compressibility	Measures the extent to which the text can be compressed, serving as an indirect indicator of redundancy and lexical variety (Ehret and Szmrecsanyi, 2016).[c]	Stylistic	van Cranenburgh and Bod (2017), Koolen et al. (2020), Bizzoni et al. (2023c)
Word and bigram entropy	Measures the unpredictability in word choices and combinations, with higher entropy indicating greater variety and stylistic complexity.	Stylistic	Algee-Hewitt et al. (2016)
Normalized Dependency Distance, mean & SD (NDD Mean, NDD STD)	Quantifies the mean and SD in dependency length, following the procedure proposed in Lei and Jockers (2020) .	Stylistic/ Syntactic	Lei and Jockers (2020)
Nominal verb ratio	Quantifies the proportion of nouns and adverbs (over verbs) in the text, reflecting the nominal tendency in style, which is often associated with complex linguistic structures, denser communicative code, expert-to-expert communication (McIntosh, 1975; Bostian, 1983).	Stylistic/ Syntactic	Charney and Rayman (1989)*, Crossley et al. (2014)*, Wu et al. (2024)
"Of"/"that" frequencies	Frequency of these function words have been seen to indicate, in the case of "of", a more nominal prose, and in the case of "that", a more declarative and verb-centered prose.	Stylistic/ Syntactic	Wu et al. (2024)
Function words	Frequency of function words (normalized for text length), suggesting a more information-rich prose when lower.	Stylistic/ Syntactic	Bizzoni et al. (2024a)
Perplexity	Represents the predictability of the prose through a self-trained large language models (GPT), as outlined in Wu et al. (2024).[d] Higher values indicate greater complexity or unpredictability.	Hybrid	Sheetz (2018), Wu et al. (2024), Wu et al. (2024)
Sentiment SD (SD Sent)	Represents the average variability in sentiment, indicating the range of sentiment within the narrative.[e]	Narrative/ Sentiment	Berger et al. (2021)*, Bizzoni et al. (2023c)
Hurst exponent	Quantifies the long-term auto-correlation of the sentiment arc,[e] with higher values suggesting a more complex, self-similar structure across different scales.[f]	Narrative/ Sentiment	Mohseni et al. (2021), Bizzoni et al. (2021), Bizzoni et al. (2023d)
Approximate entropy (APEN)	Assesses the predictability of sequences of the sentiment arc,[e] with lower values indicating greater regularity or simplicity.[f]	Narrative/ Sentiment	Hu et al. (2020), Mohseni et al. (2022), Bizzoni et al. (2023c)

Table 10: **Used features related to stylistic and sentiment complexity**. "References" refer to studies that have used the complexity feature showing some relation between it and reader appreciation. * Denotes studies in domains other than *established prose fiction* (e.g., online stories, movies).

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