Readability and Complexity: Diachronic Evolution of Literary Language Across 9000 Novels

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

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

Ida Marie Lassen
Center for Humanities Computing
Aarhus University, Denmark
idamarie@cas.au.dk

Mads Rosendahl Thomsen
Comparative Literature
School of Communication and Culture
Aarhus University, Denmark
madsrt@cc.au.dk

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

Abstract

Using a large corpus of English language novels from 1880 to 2000, we compare several textual features associated with literary quality, seeking to examine developments in literary language and narrative complexity through time. We show that while we find a correlation between the features, readability metrics are the only ones that exhibit a steady evolution, indicating that novels become easier to read through the 20th century but not simpler. We discuss the possibility of cultural selection as a factor and compare our findings with a subset of canonical works.

1 Introduction

Several textual features have been associated with "good style" or narrative in the stylometric and quantitative literature studies. A recent surge of quantitative studies has used large corpora to investigate whether intra-textual features correlate with perceived literary quality. Average sentence length (Ganjigunte Ashok et al., 2013), type-token ratio (Crosbie et al., 2013), the distribution of parts of speech (van Cranenburgh and Bod, 2017) and level of redundancy (Algee-Hewitt et al., 2018) have been shown to explain literary success partially.

Also measures of readability are often connected to literary success: it is a widespread conception of both readers and publishers that bestsellers are easier to read (Martin, 1996), and readability has recently gained traction in creative writing and publishing, such as in text-editing tools like the Hemingway[1] or Marlowe editors. [2] These applications evaluate texts with simple readability measures and tend to encourage the production of texts that are easier to read, assuming that more readable texts are better.

With the evolution of quantitative methodologies, more sophisticated models of texts have also been explored as possible markers of literary quality: the shape and dynamics of novels' sentiment arcs as a way to approximate their narrative development or the complex way parts of speech alternate throughout a text, influencing readers' experience of the story above or below conscious perception (Bizzoni et al., 2021, 2022b; Mohseni et al., 2022).

Literary evolution may show a progressive convergence towards preferred forms of and styles in narrative, perceived as effective and maintained/further evolved through community feedback (Crocker et al., 2016; Degaetano-Ortlieb and Teich, 2022). Already in the 19th century Sherman (1893) observed an evolution of the language of fiction and suggested a positive selection for simple language in literary language. This idea recurs in a theory where the rise of a mass readership is thought to have prompted the language

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

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 235–247
December 1-3, 2023 ©2023 Association for Computational Linguistics

of Western fiction to become simpler through the 19th and 20th centuries, as it caters to the progressively lower overall literacy and less spare time of the readership (Klancher, 1983; Kimball, 2017; Westin, 2002).

In this study, we first extract multiple textual and stylometric measures connected to perceived literary quality or success from a large collection of English novels, ranging from the most surface-level readability indices to models considering the dynamics of their sentiment arcs. We examine whether any systematic, diachronic trend of these measures can be observed in the period covered by our corpus (1880-2000), as has been noted in theories of slow but continuous change in literary language or narrative style into the 20th century (Underwood, 2019; Underwood and Sellers, 2016; Moretti, 2000). Secondly, considering readability as a measure linked to literary success or quality, we test the correlation between readability and other stylistic measures, as well as two measures with a higher level of abstraction that have previously been used to estimate narrative complexity in relation to reader appreciation: fractality and entropy. These measures are based on the sentiment arcs of novels, which are the sentiment scores (often extracted through dictionaries or machine learning) over the course of a whole novel. Certain shapes or sentiment arc dynamics have been connected to reader appreciation, considering both simple and more complex narratives (Bizzoni et al., 2022a), and Bizzoni et al. (2023) have shown that sentiment features, such as measures of sentiment arc progression, have an effect even compared to the predominantly stylistic features usually employed for this type of task (Koolen et al., 2020; Maharjan et al., 2017). These more complex measures that take into account the sentiment-arc of novels are interesting insofar as they are not direct measures of style, and insofar as they have proven effective in approximating literary quality for different types of quality-standards that may reflect tastes of different reader communities: distinguishing higher-rating works on large user-platforms such as GoodReads (Bizzoni et al., 2021) and telling works of Nobel laureates from those of contemporary authors (Bizzoni et al., 2022b). Finally, we estimate the same measures and diachronic trend for a subsection of the corpus defined through a combination of different "quality resources" chosen to reflect canonicity: authors most often appearing in English Literature syllabi, major literary anthologies and titles defined as "classics" on the large user-platform GoodReads. We show that while there is a correlation between surface-level and arc-based metrics, their change through time is significantly different, with readability metrics being time-dependent, and more sophisticated measures time-independent.

2 Related works

2.1 Readability and text complexity

The connection between text readability and quality has often been implied for non-fiction. Early studies of readability attest to the educational and social concerns in developing measures of readability to improve expository or didactic texts (Chall, 1947). Yet, the role of readability in the quality of *literary* texts is a more complex question, where "opacity" has also been considered a positive trait (Glissant, 1997; White et al., 1981; Moore, 1964).

Few studies have examined readability measures for predicting literary quality or success. Studying a small corpus of bestsellers and more canonical literary works, Martin (1996) found no significant difference in readability using a modified Flesch Reading Ease. In contrast, Garthwaite (2014) found differences in readability between bestsellers and commercially endorsed book-list titles, where endorsed lists of books were more difficult to read. Relying on multiple measures of readability and one measure of literary quality (i.e., GoodReads' average ratings), Maharjan et al. (2017) found that readability was not effective for estimating popularity when comparing it to, for example, character n-grams. Similarly Koolen et al. (2020) preferred other features over readability measures when estimating perceived literary quality. Still, many studies of literary success, popularity, or perceived literary quality have sought to approximate text complexity and have studied textual properties upon which formulae of readability are directly or indirectly based, such as sentence length, vocabulary richness, and text compressibility (Brottrager et al., 2022; van Cranenburgh and Bod, 2017; Crosbie et al., 2013).

2.2 Sentiment arcs

More complex measures based on the linear development of novels – their sentiment arcs – have been used to approximate literary quality for different types of reader standards, and estimate narrative rather than style. The sentiment or emotion-based development of communication is often seen as highly relevant, especially in "artistic" narrative (Drobot, 2013), as it is linked to the central and special tendency of literary texts to evoke, and not only describe experiences and inner states. As Hu et al. (2021) argues, readers engage with the evolution of a story at the emotional level by evocations or "engagement prompts". A sentiment arc is thus a model of the "engagement prompts" in the text, which sentiment analysis models as a primary tool approximate at the word (Mohammad, 2018), sentence (Mäntylä et al., 2018) or paragraph (Li et al., 2019) level (Cambria et al., 2017; Kim and Klinger, 2018; Brooke et al., 2015; Jockers, 2017; Alm, 2008; Jain et al., 2017). Sentiment analysis usually derives its models from human-based resources such as annotated lexica (Mohammad and Turney, 2013) or lists of words induced from labelled documents (Islam et al., 2020). Several studies have also attempted to complement the simplicity of sentiment analysis with systems for textual emotion recognition (Alm et al., 2005), or by developing more complex sentimental tools (Xu et al., 2020). In general, researchers have looked at sentiment arcs in terms of their overall shape (Reagan et al., 2016), but recent works have tried more complex mathematical models to define the arcs' overall level of inner coherence and predictability (Gao et al., 2016). In this study we recur to this last form of series modeling, examining the dynamics of arcs.

2.3 Quality measures

Defining literary quality as one unified standard and formalizing it for quantitative studies is a particularly complex and elusive problem. Studies that seek to predict perceived literary quality from textual features often rely on the provisional proxy of one single gold standard, such as book ratings from large user platforms such as GoodReads (Maharjan et al., 2018; Bizzoni et al., 2021) - usually with the task of predicting high-rated works - or personally as well as institutionally compiled canons (Mohseni et al., 2022), sales-numbers (Wang et al., 2019), or occasionally selections from prestigious awards such as the Nobel prize (Bizzoni et al., 2022b). However, it has been shown that reader preferences are complex and reflect multiple perceptions or standards of quality (Koolen et al., 2020), that are not necessarily based on the same criteria or prompted by the same textual features. For the present work, we use different standards of literary prestige that reflect a particularly "canonical" literary quality as a subset of our corpus to test against the wider set of titles.

3 Methods

3.1 Data

This present study uses the *Chicago corpus*, a collection of over 9,000 novels written or translated into English, spanning from 1880 to 2000. The titles were selected based on the number of libraries holding a copy of the novel (see Table 1).

The collection is rare in terms of its diversity - it represents well-known genres and popular fiction as well as important works from the entire period, including seminal modernist and postmodernist texts as well as Nobel Prize winners and recipients of prestigious literary awards. As such, the Chicago corpus contains a sizeable subsection of prestigious or "canonic" literature.

To estimate the amount of "canonic" literature in the corpus we mark all titles by authors that appear in selected institutional or user-compiled resources indicating literary prestige: in the English and American Norton Anthology (Shesgreen, 2009), two GoodReads user-generated lists, "the GoodReads classics" and the GoodReads "best books of the 20th century" (Walsh and Antoniak, 2021), and among the top thousand most assigned titles in English Literature course syllabi.[3] The amount of these “canonic" titles through time is shown in Fig. 1.

It should be noted that the Chicago corpus contains only works that were either produced or translated into English, exhibiting a clear cultural and geographic bias with a strong over-representation of Anglophone authors. This should also be considered in light of the fact that the readability metrics we use are particularly effective and were developed for the English language.

	Titles	**Authors**
Number	9089	3150
Avg. holdings	535.73	495.1

Table 1: Overall number of titles and authors in the corpus (first line) and average number of library holdings per title and per author (second line).

[3] Based on syllabi collected by the Open-syllabus project: https://opensyllabus.org

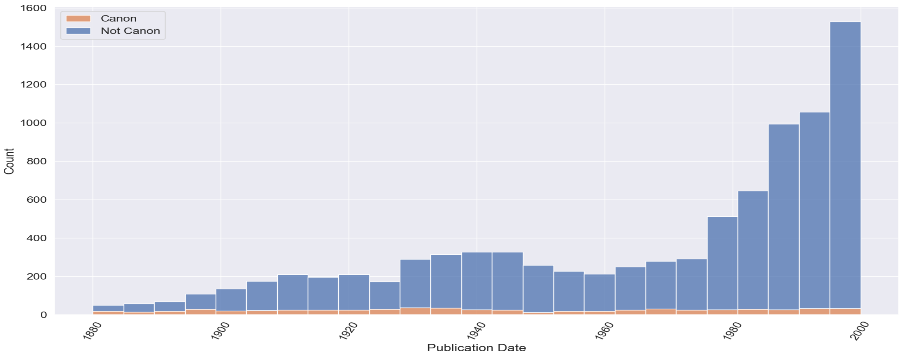


Figure 1: Overall quantity of titles per decade in the corpus, with the number of "canonical" books in orange.

Resource	N. titles
University Syllabi	478
Norton Anthology	402
GoodReads classics	62
GoodReads 20th century	44
Total unique titles	641

Table 2: Number of titles per canonicity resource in the Chicago corpus.

3.2 Measures of readability

While what is "readable" is problematic to define, and clearly varies depending on the reader, the context and the genre (Berlatsky, 2015; Flesch, 1948), readability scores may act as proxy measures for people's reading experience and enable comparison between texts.[4]

To avoid relying on one single interpretation of the readability concept, we compare five different measures of textual readability, chosen for their popularity and interpretability.[5] Since the 1920s, and particularly after the success of Flesch and Dale-Chall formulas in the 1950s, combinations of sentence length, word lengths, and/or number of syllables have been used as proxies for linguistic complexity to gauge the difficulty of a text (Dale and Chall, 1948). In 1980, there were more than 200 distinct readability formulae (Dubay, 2004), and new ones are continually being developed as older ones are refined. Despite their relative simplicity, the measures from what Dubay (2004) refers to as the "classic readability" studies remain the most popular ones and useful in determining text difficulty (Stajner et al., 2012).

The selected readability measures are the following:

- The *Flesch Reading Ease* is a measure of readability based on the average sentence length (ASL), and the average number of syllables per word (ASW). It is calculated as follows:

$$\begin{aligned}\mathrm{RE} = 206.835 - (1.015 \times \mathrm{ASL}) \\ - (84.6 \times \mathrm{ASW})\end{aligned}$$

- The *Flesch-Kincaid Grade Level* is a revised version of the Flesch Reading Ease score. Like the former, it is based on the average sentence length (ASL), and the number of syllables per word (ASW). It is calculated as follows:

$$\mathrm{GL} = (0.4 \times \mathrm{ASL}) + (12 \times \mathrm{ASW}) - 15$$

- The *SMOG Readability Formula* is a readability score introduced by McLaughlin (McLaughlin, 1969). It measures readability based on the average sentence length and number of words with more than 3 syllables (number of polysyllables), applying the formula:

$$\begin{aligned}\mathrm{SMOG} = 1.043 \times \sqrt{polysyllablecount} \\ \times \frac{30}{\mathrm{n_{st}} + 3.1291}\end{aligned}$$

- The *Automated Readability Index* is a readability score based on the average sentence

[4] We use the term "readability" here, since this is properly what readability indices, developed in linguistics, intend to measure. Other terms, like text "simplicity" may be related but are more broadly defined and often measured with a combination of both stylistic and more content-based features (Popović et al., 2022), while "readbaility" is predominantly stylistic.

[5] All readability scores were extracted using the textstat package: https://pypi.org/project/textstat/

length and number of characters per words (word length). It is calculated as follows:

$$\text{ARI} = 4.71\frac{\text{characters}}{\text{words}} + 0.5\frac{\text{words}}{\text{sentences}} - 21.43$$

- The *New Dale–Chall Readability Formula* is a 1995 revision of the Dale-Chall readability score (Chall and Dale, 1995). It is based on the average sentence length (ASL) and the percentage of "difficult words" (PDW) defined as words which do not appear on a list of words which 80 percent of fourth-graders would know (Dale and Chall, 1948).[6] It is calculated as follows:

$$\begin{aligned} \text{DC} &= 0.1579 \times \text{PDW} + 0.0496 \times \text{ASL} \\ \text{If PDW} &> 5\% : \text{Adjusted Score} = \\ &\text{Raw Score} + 3.6365 \end{aligned}$$

We complement these standard readability metrics with three other metrics often used to assess stylistic complexity of texts:

- **Sentence length**. Character-based sentence length is also often integrated into readability measures.

- **Type-token ratio**. A standard index of lexical richness, not used in readability metrics but normally considered indicative of a text's complexity and inner diversity (Torruella and Capsada, 2013).[7]

- **Compressibility** measures to what extent a text can be compressed through a standard compression algorithm. This measure becomes essentially a sign of redundancy or formulaicity: the more a text tends to repeat sequences *ad verbatim*, the more compressible it will be (Benedetto et al., 2002; van Cranenburgh and Bod, 2017).[8]

3.3 Sentiment arcs

To apply more complex, sentiment arc based metrics of the narratives, in this study, we extract sentiment arcs using the VADER model (Hutto and Gilbert, 2014) at the sentence level. Sentiment analysis of a literary text provides a simple and intuitive representation of a narrative's sentimental trajectory, and has been applied as a proxy for meaningful aspects of the reading experience (Drobot, 2013; Cambria et al., 2017; Kim and Klinger, 2018; Brooke et al., 2015; Jockers, 2017). The resulting representation is referred to as a sentiment arc, which a range of studies model to evaluate narratives in terms of genre (Kim et al., 2017), plot archetypes (Reagan et al., 2016), and lastly, reader preference (Bizzoni et al., 2022a). While dictionary-based sentiment analysis remains a popular choice, more recent, transfomer-based methods are more recently explored (Elkins, 2022).

Our choice of a dictionary-based approach was motivated by a desire for transparency and corpus independence. Among available sentiment analysis tools we selected VADER due to its widespread usage and comprehensive rule set. VADER generates a compound valence score for each sentence, ranging from negative (-1), through neutral (0), to positive (1). Figure 2 serves as a demonstration of the arc extraction process for the first ten sentences of Ernest Hemingway's seminal work *The Old Man and the Sea*. [9] To highlight the efficacy of the annotation on narrative texts, Figure 3 also shows the sentiment arc of *The Old Man and the Sea* with its corresponding narrative events, compared to human annotation of the book.[10]

[6] Contained in the Dale-Chall word-list: `https://countwordsworth.com/download/DaleChallEasyWordList.txt`

[7] 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.

[8] 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.

[9] "He was an old man who fished alone in a skiff in the Gulf Stream and he had gone eighty-four days now without taking a fish. In the first forty days a boy had been with him. But after forty days without a fish the boy's parents had told him that the old man was now definitely and finally salao, which is the worst form of unlucky, and the boy had gone at their orders in another boat which caught three good fish the first week. It made the boy sad to see the old man come in each day with his skiff empty and he always went down to help him carry either the coiled lines or the gaff and harpoon and the sail that was furled around the mast. The sail was patched with flour sacks and, furled, it looked like the flag of permanent defeat. The old man was thin and gaunt with deep wrinkles in the back of his neck. The brown blotches of the benevolent skin cancer the sun brings from its reflection on the tropic sea were on his cheeks. The blotches ran well down the sides of his face and his hands had the deep-creased scars from handling heavy fish on the cords. But none of these scars were fresh. They were as old as erosions in a fishless desert. Everything about him was old except his eyes and they were the same color as the sea and were cheerful and undefeated."

[10] Human annotation of sentiment per sentence was performed by 2 annotators, asked to score individual sentences

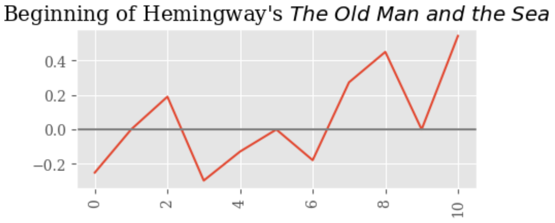


Figure 2: VADER-annotation of Hemingway's *The Old Man and the Sea*, the first 10 sentences. The arc captures the narrative fluctuations of the sentence sequence.

3.4 Sentiment arc-based metrics

To model the underlying complexity of the novels, we use two more complex measures already previously used in studies of literary quality: Hurst's exponent and Approximate Entropy of the novels' sentiment arcs (Bizzoni et al., 2021, 2022b; Mohseni et al., 2022).

Used to detect long-term memory in time series data, the Hurst exponent in our context measures the persistence of sentiment or the long-term memory of sentiment arcs. To estimate Hurst, we combine non-linear adaptive filtering with fractal analysis, specifically adaptive fractal analysis (Gao et al., 2011; Tung et al., 2011). Nonlinear adaptive filtering is used due to the inherent noisiness of story arcs. First, the signal is partitioned into segments (or windows) of length $w = 2n + 1$ points, where neighboring segments overlap by $n + 1$. Then, a polynomial of order D is fitted for each segment. The fitted polynomial for ith and $(i + 1)th$ is denoted as $y^{(i)}(l_1), y^{(i+1)}(l_2)$, where $l_1, l_2 = 1, 2, ..., 2n + 1$. We use the following weights for the overlap of two segments.

$$y^{(c)}(l_1) = w_1 y^{(i)}(l + n) + w_2 y^{(i)}(l), \\ l = 1, 2, \dots, n + 1 \quad (1)$$

where $w_1 = (1 - \frac{l-1}{n}), w_2 = 1 - w_1$ can be written as $(1 - \frac{d_j}{n}), j = 1, 2$, where d_j denotes the distance between the point of overlapping segments and the center of $y^{(i)}, y^{(i+1)}$. Studies have demonstrated the usefulness of adaptive filtering applied to sentiment arcs, especially in the context of estimating dynamics of sentiment arcs (Hu et al., 2021; Bizzoni et al., 2022b).

After nonlinear adaptive filtering, we use the Hurst exponent to measure long-term memory. Assuming that stochastic process $X = X_t : t = 0, 1, 2, ...$, with stable covariance, mean μ and σ^2, the process' autocorrelation function for $r(k), k \geq 0$ is:

$$r(k) = \frac{E\left[X(t)X(t+k)\right]}{E\left[X(t)^2\right]} \sim k^{2H-2}, \text{as} \quad k \to \infty \quad (2)$$

where H is called the Hurst exponent (Mandelbrot, 1982).

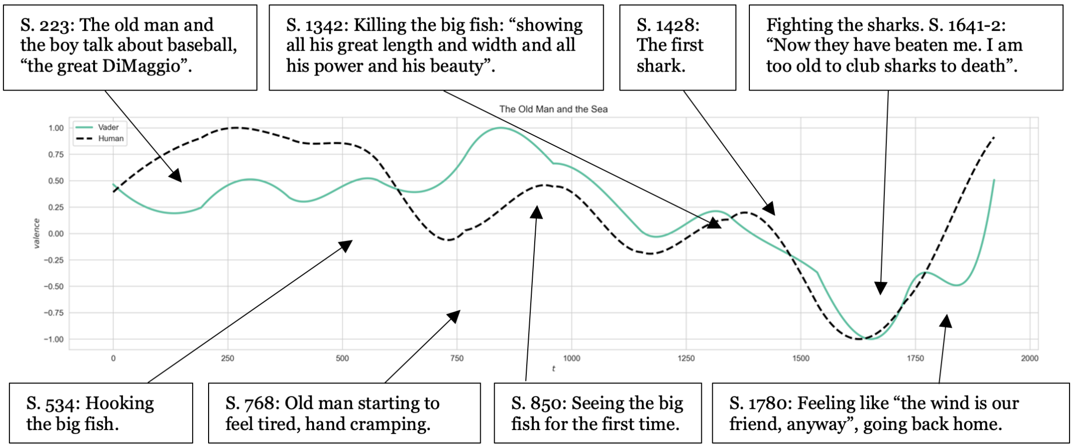
For $0.5 < H < 1$ the story arc is characterized as persistent such that increments are followed by increases and decreases by further decreases. For $H = 0.5$ the story arc only has short-range correlations; and when $H < 0.5$ the story arc is anti-persistent such that increments are followed by decreases and decreases by increments. For the specific application domain (i.e., narratives) persistent story arcs are characteristic of coherent narratives, where the emotional intensity evolves at longer time scales. Story arcs that only show short memory lack coherence and appear like a collection of short stories. Anti-persistent story arcs will appear bland and rigid narratives oscillating around an average emotional state (Hu et al., 2021).

Adaptive fractal analysis consists of the following steps: first, the original process is transformed to a random walk process through first-order integration $u(n) = \sum_{k=1}^{n}(x(k) - \overline{x}), n = 1, 2, 3, ..., N$, where $\overline{x}$ is the mean of $x(k)$. Second, we extract the global trend $(v(i), i = 1, 2, 3, ..., N)$ through the nonlinear adaptive filtering. The residuals $(u(i) - v(i))$ reflect the fluctuations around a global trend. We obtain the Hurst parameter by estimating the slope of the linear fit between the residuals' standard deviation $F^{(2)}(w)$ and w window size as follows:

$$F^{(2)}(w) = \left[\frac{1}{N}\sum_{i=1}^{N}(u(i) - v(i))^2\right]^{\frac{1}{2}} \sim w^H \quad (3)$$

Beyond Hurst exponent, we estimate the approximate entropy (ApEn) of sentiment arcs. ApEn is a measure of the predictability of time-series of data based on Shannon Entropy and introduced by S. Pincus as a measure of physiological system complexity (Pincus, 1991; Pincus et al., 1991). Given a time series X with N data points, ApEn is calculated as follows: a value for m, the length of the comparison segment, and a tolerance value r are chosen. The time series X is then divided

of the book on a 1-10 scale without paying attention to the narrative context.



S. 223: The old man and the boy talk about baseball, "the great DiMaggio".

S. 1342: Killing the big fish: "showing all his great length and width and all his power and his beauty".

S. 1428: The first shark.

Fighting the sharks. S. 1641-2: "Now they have beaten me. I am too old to club sharks to death".

S. 534: Hooking the big fish.

S. 768: Old man starting to feel tired, hand cramping.

S. 850: Seeing the big fish for the first time.

S. 1780: Feeling like "the wind is our friend, anyway", going back home.

Figure 3: The detrended (by adaptive filtering) and normalized sentiment arcs of *The Old Man and the Sea* based on VADER scores and human annotations, shown with main narrative events.

into overlapping segments of length m, such that X_i to X_{i+m-1} represents one segment, where $1 \leq i \leq N - m + 1$. For each segment X_i to X_{i+m-1}, the number of segments X_j to X_{j+m-1} (where $j \neq i$) that are within a distance of r from X_i to X_{i+m-1} is calculated, where r is a real number that specifies a filtering level, essentially defining what constitutes a match. We will call the number of matches $C(i)$. Finally, the probability of observing $C(i)$ matches for a given segment X_i to X_{i+m-1} as can be computed as:

$$p(i) = \frac{C(i)}{N - m + 1} \tag{4}$$

The previous steps are be repeated for increasing values of m and the probabilities are averaged over all segments to obtain the final value:

$$ApEn(m, r) = -\frac{1}{N - m + 1} \sum [\ln (p(i))] \tag{5}$$

The ApEn value for a given time series is determined by the minimum value of $ApEn(m, r)$ for a range of m and r values.

The ApEn value represents the level of randomness or predictability in the time series, with higher values indicating greater randomness and lower values indicating more predictability. ApEn has been used to study the complexity of various types of time series data, i.a., heart-rate (Fleisher et al., 1993), financial (Delgado-Bonal and Marshak, 2019), and narratives (Mohseni et al., 2022). Applied to sentiment arcs, ApEn searches for recurrent patterns in the arc and estimates the (log) likelihood that adjoining sequences of sentences, two in this study, will differ, that is, whether the pattern is predictable. Smaller values of ApEn indicate more recurring patterns and thus higher predictability, while higher values indicate fewer recurring patterns and lower predictability.[11]

4 Results

As we show in Figure 4 and Table 3, the main result of our analysis consists of two series of trends:

1. All measures of readability clearly correlate with the passage of time and point in the same direction: to an increased readability of novels. Sentence length follows this trend, indicating that sentences become on average shorter through the 20th century.

2. All other measures we took into consideration, including the "linear" measures of sentiment arcs, do not change meaningfully through time.

The clear trend of all readability measures indicates an overall simplification of the literary prose, beyond the characteristics of the authors' individual styles. Interestingly, the trend can be observed for the corpus at large as well as for the "high prestige" subsection of titles we outline in Table 2. Looking into the relation between the readability measures and Hurst as well as Approximate Entropy (Table 4) we find that they correlate with readability in the sense that more difficult books tend to have a higher Hurst exponent and higher Approximate Entropy. So overall, in our corpus, simpler books have

[11] We used the Neurokit-package to measure ApEn of arcs: https://neuropsychology.github.io/NeuroKit/

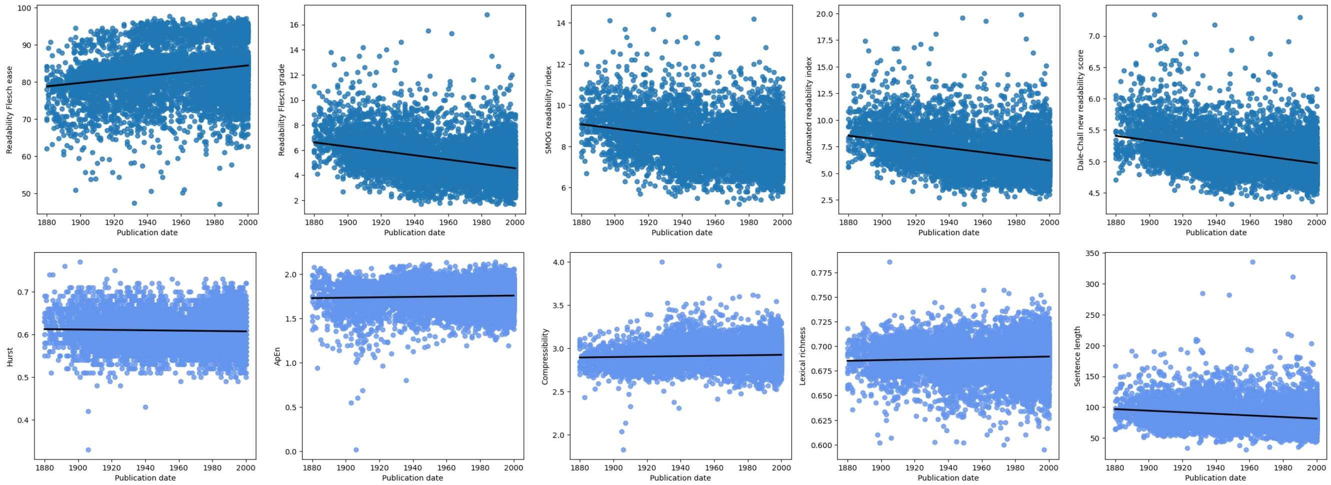


Figure 4: Distribution of readability measures (upper) and other measures (lower row) through time. Note that Flesch Reading Ease shows a score where a lower number means lower readability so that it is inverted with respect to the other readability measures.

less complex arcs. However, we do not see a tendency towards simpler arcs through time: if books become easier to read from 1880 to 2000, they do not become simpler in terms of their sentiment-arc dynamics. The overall level of complexity of the novels' sentiment arcs remains remarkably stable through the corpus - and the titles of our "canon selection" even show a slight tendency towards higher complexity through time. While this points to the fact that readability and arc complexity are only partially correlated (other factors might correlate even more strongly with one or the other), it shows that with time writers might have increasingly favored a kind of prose that strives to keep a non-obvious balance between simplicity of style and complexity of the sentiment arc.

The lack of lasting diachronic changes in the other two stylistic measures considered, type-token ratio and textual compressibility, seems to confirm this picture: if novels become easier in terms of basic readability metrics, they do not lose complexity at many other levels, not becoming overall more repetitive nor lexically poorer. In other words, it might be that there has been a large, overall tendency to favor texts that manage to simplify the most surface level aspect of style, without compromising their linguistic diversity nor their narrative arcs' complexity.

5 Discussion

The different trends we have shown between surface-level readability measures, other metrics of style, and arc complexity through time seem to point towards a large-scale evolution of literary language towards prose that favors increased readability without compromising the novels' linguistic or narrative versatility. As we have seen in Table 4, arc measures (Hurst exponent and Approximate Entropy) and readability are indeed correlated in the corpus, but the development of readability measures shows a tendency to progressively simplify

	Spearman	**Pearson**
Hurst	-0.015 (*-0.1*)	0.036
ApEn	0.032 (*0.1*)	0.05
Lexical richness	0.081 (*0.0*)	0.062
Compressibility	0.042 (*-0.1*)	0.006
Sentence length	-0.185 (*-0.1*)	-0.201
Flesch Ease	0.249 (*0.2*)	0.246
Flesch Grade	-0.316 (*-0.3*)	-0.362
SMOG	-0.287 (*-0.2*)	-0.323
ARI	-0.296 (*-0.3*)	-0.352
Dale Chall	-0.341 (*-0.2*)	-0.383

Table 3: Spearman and Pearson correlations between textual measures and the novels' publication date. For reference, the Spearman correlations of textual meaures and the novel's publication date for *canonic works only* are added in parentheses. All non-null correlations (r>0.1) have p-values $<$ 0.0005.

	Hurst	**ApEn**
Hurst	1.000	0.366
ApEn	0.366	1.000
Flesch Ease	-0.162	-0.404
Flesch Grade	0.172	0.431
SMOG	0.153	0.412
ARI	0.172	0.428
Dale Chall	-0.043	0.104

Table 4: Spearman correlations between linear metrics and readability measures (all statistically significant).

the prose of all novels, including those with complex arcs (Table 3).

Regarding the trends towards readability alone, it is reasonable to exclude that they are the effect of an overall change of the English language. Similar tendencies towards simplification have been found in narrative (Sherman, 1893; Liddle, 2019) but is not as obvious in other domains (Säily et al., 2017). Moreover, scientific and journalistic prose has even shown an opposed trend, with texts becoming more difficult to read (Plavén-Sigray et al., 2017; Danielson et al., 1992).

If this trend is not an effect of language change, its presence in literature can give way to intriguing hypotheses. The emergence of what scholars have called mass readership (Klancher, 1983) and a widening of the alphabetized population might have pushed the success of easier books,[12] while the increasingly pressing market logic applied to the editorial world might have helped shaping literary style into simpler and simpler forms, easier to consume in a shorter time (Winter and O'Neill, 2022). It is also possible that in the last century, difficulty of reading has shifted from a virtue to a vice in the view of the English writing world, with novelists and publishers alike slowly favoring more direct or transparent prose.

A central question to ask is whether we are seeing an actual transformation of literary prose, or whether we are witnessing an effect of cultural selection. In theory, there might have been no evolution at all through the 120 years in question, but less appreciated exemplars might have been progressively lost or overlooked. If texts are undergoing a constant process of selection, it is possible that the "books worth keeping" maintain more complex stylistic features, while the larger number of more easily readable novels is progressively forgotten, leading to the illusion of a historical change through survivor bias. A look at the absolute number and percentage of "canonic" books in our corpus, as defined by various indicators of prestige (see Section 3.1), seems to point to this competing view: while the number of texts in the corpus increases with each decade, the percentage of canonic titles decreases drastically through time.

However, when looking at the canonic subset alone, we see changes similar to those that we observe in the whole corpus: what we have defined as canonical literature has also become more readable from 1880 to 2000 (Table 3). Moreover, the canonic subset seems to tend even more toward a disentanglement of surface readability and arc complexity, with the latter showing even a slight increase in complexity – Hurst and ApEn – through time.

6 Conclusion and Future Works

In our analysis of a curated corpus of 9000 English-language novels published between 1880 and 2000, we employed specific readability metrics – i.a., the Flesch-Kincaid Readability Score – as well as complexity indices like the Hurst exponent of sentiment arcs and classic stylometrics, i.a., type-token ratio and compressibility. Our data indicates some clear trends: most readability scores have increased through time, displaying Spearman correlations of up to 0.34 with the publication year, signifying a gradual simplification of the overall narrative language. In contrast, richness and complexity metrics remained relatively unchanged over the same period. These divergent trends might suggest that authors are increasingly focusing on making their works more accessible while maintaining a consistent level of narrative complexity and lexical diversity, which we might interpret as a literary strategy to engage a broader audience without sacrificing depth or complexity. It's worth noting that our study does not account for genre-specific trends and is based on available works, thus introducing potential selection bias.

Future research could expand upon these findings by exploring how these trends vary across different genres and cultures. Naturally, exploring this further would require properly discussing and deploying a system of judgments for literary quality, an undertaking beyond the scope of this work. In the future, we would like to both conduct qualitative analyses to assess these results on individual work level and repeat the experiment on different and possibly larger literary data sets. We also plan to collect more textual features, such as model perplexity, as well as develop more sophisticated models for the Sentiment Analysis that underlies measures of arc dynamics (Hurst, ApEn), such as using LMs, and examine the change through time of these features with more sophisticated mathematical models.

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

Limitations

The Chicago Corpus serves as a valuable resource for our study, as it encompasses an expansive and representative sample of widely read Anglophone literature over a century, allowing for a robust analysis. Still, it is worth noting that the corpus has a geographical bias: most authors are of US origin and few are non-Anglophone. This bias inevitably situates the entire analysis within the context of a well-defined “Anglocentric" literary field. Moreover – perhaps also due to an inherent skew in this literary field – 36% of authors are women.

While these imbalances do not inherently undermine our experiments, it is crucial to bear them in mind when interpreting the results, and we advise against extrapolating our findings to the context of a wider or global literary field. Moreover, when estimating the canonicity of works in the corpus we have relied on external lists that are, to an extent, characterised by similar biases, for example the *Norton Anthology*, which is similar both in terms of most predominantly selecting among Anglophone authors and in terms of its gender biases (Pope, 2019). We trust that any interpretation of our findings will have these limitations in mind.

References

Mark Algee-Hewitt, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser. 2018. Canon/archive : large-scale dynamics in the literary field.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 579–586.

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

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

Noah Berlatsky. 2015. Readability is a myth. Section: Culture.

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

Yuri Bizzoni, Telma Peura, Kristoffer Nielbo, and Mads Thomsen. 2022a. Fractal sentiments and fairy tales-fractal scaling of narrative arcs as predictor of the perceived quality of andersen's fairy tales. *Journal of Data Mining & Digital Humanities*.

Yuri Bizzoni, Telma Peura, Kristoffer Nielbo, and Mads Thomsen. 2022b. Fractality of sentiment arcs for literary quality assessment: The case of nobel laureates. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, 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).

Julian Brooke, Adam Hammond, and Graeme Hirst. 2015. Gutentag: an nlp-driven tool for digital humanities research in the project gutenberg corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 42–47.

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

Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. 2017. Affective computing and sentiment analysis. In *A practical guide to sentiment analysis*, pages 1–10. Springer.

Jeanne S. Chall. 1947. This business of readability. *Educational Research Bulletin*, 26(1):1–13.

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

Matthew W. Crocker, Vera Demberg, and Elke Teich. 2016. Information Density and Linguistic Encoding (IDeaL). *KI - Künstliche Intelligenz*, 30(1):77–81.

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

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

Wayne A. Danielson, Dominic L. Lasorsa, and Dae S. Im. 1992. Journalists and novelists: A study of diverging styles. *Journalism Quarterly*, 69(2):436–446.

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.

Alfonso Delgado-Bonal and Alexander Marshak. 2019. Approximate Entropy and Sample Entropy: A Comprehensive Tutorial. *Entropy*, 21(6):541.

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

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

Katherine Elkins. 2022. *The Shapes of Stories: Sentiment Analysis for Narrative*. Cambridge University Press.

Lee Fleisher, Steve Pincus, and Stanley Rosenbaum. 1993. Approximate entropy of heart rate as a correlate of postoperative ventricular dysfunction. *Anesthesiology*, 78(4):683—692.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233.

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, Seattle, Washington, USA. Association for Computational Linguistics.

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):e24331.

Jianbo Gao, Matthew L Jockers, John Laudun, and Timothy Tangherlini. 2016. A multiscale theory for the dynamical evolution of sentiment in novels. In *2016 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)*, pages 1–4. IEEE.

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

Édouard Glissant. 1997. *Poetics of relation*. University of Michigan Press, Ann Arbor.

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.

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

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

Swapnil Jain, Shrikant Malviya, Rohit Mishra, and Uma Shanker Tiwary. 2017. Sentiment analysis: An empirical comparative study of various machine learning approaches. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 112–121, Kolkata, India. NLP Association of India.

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

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

Evgeny Kim, Sebastian Padó, and Roman Klinger. 2017. Investigating the relationship between literary genres and emotional plot development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 17–26, Vancouver, Canada. Association for Computational Linguistics.

Courtney Kimball. 2017. Sweeney todd's dreadfuls and mass readership. *The Journal of Publishing Culture*, 7:1–12.

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

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

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

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

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

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

Suraj Maharjan, Sudipta Kar, Manuel Montes, Fabio A. González, and Thamar Solorio. 2018. Letting emotions flow: Success prediction by modeling the flow of emotions in books. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Volume 2, Short Papers*, pages 259–265, New Orleans, Louisiana. Association for Computational Linguistics.

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

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

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

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

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

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

Arthur K. Moore. 1964. The case for poetic obscurity. *Neophilologus*, 48(1):322–340.

Franco Moretti. 2000. The slaugtherhouse of literature. *MLQ: Modern Language Quarterly*, 61(1):207–227.

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

Steve Pincus. 1991. Approximate entropy (apen) as a complexity measure. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 5(1):110–117.

Steve Pincus, Igor Gladstone, and Richard Ehrenkranz. 1991. A regularity statistic for medical data analysis. *Journal of Clinical Monitoring*, 7(4):335–345.

Pontus Plavén-Sigray, Granville James Matheson, Björn Christian Schiffler, and William Hedley Thompson. 2017. Research: The readability of scientific texts is decreasing over time. *eLife*, 6:e27725.

Colin Pope. 2019. We Need to Talk About the Canon: Demographics in 'The Norton Anthology'.

Maja Popović, Sheila Castilho, Rudali Huidrom, and Anya Belz. 2022. Reproducing a Manual Evaluation of the Simplicity of Text Simplification System Outputs. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 80–85, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. 5(1):1–12.

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

Sean Shesgreen. 2009. Canonizing the canonizer: A short history of the norton anthology of english literature. *Critical Inquiry*, 35(2):293–318.

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.

Tanja Säily, Arja Nurmi, Minna Palander-Collin, and Anita Auer, editors. 2017. *Exploring Future Paths for Historical Sociolinguistics*, volume 7 of *Advances in Historical Sociolinguistics*. John Benjamins Publishing Company, Amsterdam.

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

Wen-wen Tung, Jianbo Gao, Jing Hu, and Lei Yang. 2011. Detecting chaos in heavy-noise environments. *Physical Review E*, 83(4).

T. Underwood. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.

Ted Underwood and Jordan Sellers. 2016. The *Longue Durée* of Literary Prestige. *Modern Language Quarterly*, 77(3):321–344.

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.

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

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.

I. Westin. 2002. *Language Change in English Newspaper Editorials*. Language and computers : studies in practical linguistics. Rodopi.

Allon White, Lecturer in English Allon White, and White Allon. 1981. *The Uses of Obscurity: The Fiction of Early Modernism*. Routledge & Kegan Paul.

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

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2020. DomBERT: Domain-oriented language model for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1725–1731, Online. Association for Computational Linguistics.