Modeling Preference in Fictional Texts Based on Structural Features

Dissertation
(kumulativ)

Zur Erlangung des akademischen Grades doctor philosophiae
(Dr. phil.)

vorgelegt dem Rat der Philosophischen Fakultät
der Friedrich-Schiller-Universität Jena

von Mahdi Mohseni

geboren am 05.09.1983 in Marvdasht

Gutachter/Gutachterin:

1. Herr Prof. Dr. Dr. Christoph Redies, Jena
2. Herr Prof. Dr. Volker Gast, Jena
3. Frau Prof. Dr. Sina Zarrieß, Bielefeld

Tag der öffentlichen Verteidigung: 18.01.2024

ERKLÄRUNG

Ich erkläre,

dass mir die geltende Promotionsordnung der Fakultät bekannt ist;

dass ich die Dissertation selbst angefertigt, keine Textabschnitte eines Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönlichen Mitteilungen und Quellen in meiner Arbeit angegeben habe;

dass mich ausschließlich die folgenden Personen bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts unterstützt haben: Prof. Christoph Redies and Prof. Volker Gast.

dass die Hilfe einer kommerziellen Promotionsvermittlung nicht in Anspruch genommen wurde und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für die Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen;

dass ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe;

dass ich nicht die gleiche, eine in wesentlichen Teilen ähnliche oder eine andere Abhandlung bei einer anderen Hochschule als Dissertation eingereicht habe (wenn doch, bitte Ergebnis angeben).

Jena, ...

ACKNOWLEDGEMENTS

I am deeply grateful to Prof. Redies for his exceptional support throughout the course of my research. His guidance and expertise were invaluable in fulfilling my research goals and pushing me to achieve my best. Beyond his academic mentorship, I am truly grateful for his genuine care and support during challenging times, both academically and personally. I would also like to extend my heartfelt appreciation to Prof. Gast for his valuable contributions to my research project. He has been one of the most hard-working and smartest professors, whom I know. His insightful feedback, dedication, and commitment to my academic growth were instrumental in shaping the direction of my research. I am grateful for his guidance and expertise, which have played a pivotal role in enhancing the quality of my work.

I also express my gratitude to Prof. Zarrieß for her interest in my research work. Given the interdisciplinary nature of my study, her expertise and insights are invaluable in ensuring the quality and rigor of my research. I greatly appreciate her time and support.

Additionally, I would like to sincerely thank all the members of our research lab, Experimental Aesthetics Group. Their collective contributions, discussions, and feedback during our lab presentations created an enriching environment for improving my work. Their presence and insights contributed significantly to the success of my research.

I would also like to take this opportunity to express my gratitude to my friends who were a constant source of support throughout my academic and personal life. Their friendship and presence provided me with the motivation and strength to overcome challenges and pursue excellence in my research. I am truly grateful for their positive impact they had on my overall experience.

ABSTRACT

Computational textual aesthetics is an emerging field that aims to investigate observable differences between aesthetic categories of text. In this study, we explored structural differences between preferred and non-preferred fictional texts. To put our results into perspective, we also analyzed non-fictional texts and compared them with fictional texts. Canonization was used to operationalize preference for texts from the 19^{th} and early 20^{th} centuries, while for contemporary texts, sales figures were regarded as a proxy for readers' preference. Looking for the distinctive structural characteristics of text categories, we represented texts as sequences (series) of text properties and analyzed them using three main approaches: variability, fractality, and predictability analysis.

Our findings revealed that canonical fiction exhibits more variability compared to non-canonical fiction. Fractality analysis showed that long-range correlation patterns are more similar in canonical and non-canonical texts, suggesting that fractality is a universal feature of text, slightly more pronounced in non-fictional texts. Predictability analysis focuses on (ir)regularities and uncertainty within texts. We analyzed different aesthetic categories by applying Approximate Entropy as a measure of surprise in local structures of texts, and Shannon Entropy as a global measure of unpredictability. Our findings demonstrated that preferred texts are less predictable than non-preferred texts, and predictability analysis can reveal structural differences between various categories of text. We further investigated whether structural properties of text, which are potential textual correlates of preference, vary across different time periods. Our findings confirm that design features of text change over time, and they can be utilized to distinguish text categories with varying degrees of preference.

Our thorough investigation and analysis of multiple methodologies contribute significantly to the field of computational textual aesthetics. We comprehensively discuss our findings, including insights from both successful and less successful experiments, which allow us to outline potential avenues for future projects and stimulate further research in the field.

ZUSAMMENFASSUNG

Computergestützte Textästhetik ist ein aufstrebendes Gebiet, das darauf abzielt, beobachtbare Unterschiede zwischen ästhetischen Textkategorien zu untersuchen. In dieser Studie haben wir die strukturellen Unterschiede zwischen bevorzugten und nicht bevorzugten fiktionalen Texten untersucht. Um unsere Ergebnisse einordnen zu können, haben wir auch nicht-fiktionale Texte analysiert und sie mit fiktionalen Texten verglichen. Zur Operationalisierung der Präferenz wurde für Texte aus dem 19. und frühen 20. Jahrhundert die Kanonisierung herangezogen, während für zeitgenössische Texte die Verkaufszahlen als Maß für die Präferenz der Leser diente. Um charakteristische Strukturmerkmale von Textkategorien messen zu können, stellten wir Texte als Sequenzen (Serien) von Texteigenschaften dar und analysierten sie anhand von den drei Merkmalen Variabilität, Fraktalität und Vorhersagbarkeit.

Unsere Ergebnisse zeigen, dass kanonische fiktionale Texte im Vergleich zu nicht-kanonischen fiktionalen Texten eine größere Variabilität aufweisen. Die Fraktalitätsanalyse weist darauf hin, dass die Korrelationsmuster über weite Strecken in kanonischen und nicht-kanonischen Texten ähnlich sind, was darauf hindeutet, dass Fraktalität ein universelles Merkmal von Texten ist, das aber in nicht-fiktionalen Texten etwas stärker ausgeprägt ist. Die Analyse der Vorhersagbarkeit konzentriert sich auf (Un-)Regelmäßigkeiten und Unsicherheit in Texten. Wir analysierten verschiedene ästhetische Kategorien, indem wir die Approximative Entropie als Maß für die Überraschung in lokalen Strukturen von Texten und die Shannon-Entropie als globales Maß für die Unvorhersehbarkeit verwendeten. Unsere Ergebnisse zeigen, dass bevorzugte Texte weniger vorhersehbar sind als nicht- bevorzugte Texte, und dass die Analyse der Vorhersehbarkeit strukturelle Unterschiede zwischen verschiedenen Textkategorien aufdecken kann. Darüber hinaus untersuchten wir, ob strukturelle Eigenschaften von Texten, die potenzielle textuelle Korrelate der Präferenz sind, über verschiedene Zeiträume hinweg variieren. Wir konnten zeigen, dass sich Gestaltungsmerkmale von Texten im Laufe der Zeit verändern und zur Unterscheidung von Textkategorien mit unterschiedlichen Präferenzgraden genutzt werden können.

Unsere grundlegende Untersuchung und die Analyse verschiedener Methoden leisten einen wichtigen Beitrag im Bereich der computergestützten Textästhetik. Wir diskutieren umfassend unsere Ergebnisse, einschließlich der Erkenntnisse aus erfolgreichen und weniger erfolgreichen Experimenten. Dies ermöglicht es uns,

mögliche Wege für zukünftige Projekte zu skizzieren und weitere Forschung auf diesem innovativem Gebiet anzuregen.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

Can texts of difference aesthetic categories be distinguished from each other? What are correlates of aesthetic preference and appreciation in text? Do pleasing and preferred texts share properties with visual and auditory artworks? These questions lie at the heart of the field of Experimental Aesthetics in which the perception, production, and evaluation of objects "that evoke an intense feeling" (Chatterjee, 2011) are investigated by applying a variety of research techniques and using controlled observation. Visual Aesthetics, a sub-field of Experimental Aesthetics, flourished in the last two decades and now is a well-established field of research. Although language and text are the most important ways of communication, textual aesthetics is still understudied, partly due to the lack of methodological and technical tools, the complexity of language structures —the convoluted hierarchy of language components—and the time-distributed nature of text, which makes it hard to measure aesthetic experience in this domain.

1.1 Objectives and Assumptions

The main focus of the present project is on computational textual aesthetics, an interdisciplinary field at the interface of computational linguistics and literary studies. This field of research has a two-fold aim: Identifying statistical properties that reflect the categorization of different text types; and providing a basis to formulate hypotheses for experimental studies by determining potential correlates of aesthetic responses to texts with varying aesthetic claims.

The main question of this project is: What are the characteristics of aesthetically pleasing prose texts? This question will be addressed at two levels: At an 'internal' level, fictional text categories of higher or lower reputation are compared with each other. At an 'external' level, fictional texts are compared with non-fictional texts, to determine inter-genre differences of the two text categories.

The central assumption of this project is that preference has correlates in structural design properties of text. We differentiate between "what is said" and "how it is said". The former could be addressed by taking approaches such as topic analysis or more straightforward lexical methods, e.g. n-grams. However, this type of analyses

is more genre-specific and thus less generalizable. The latter is a structural design problem and differentiates between the content and the organization of a text.

Textual data are one-dimensional and information processing in reading is timely distributed, which makes it hard to experimentally measure the readers' responses. We therefore pursue an observational approach and study "aesthetics" in terms of preference. Specifically, we analyze texts which have been preferred by a community within a culture against texts which have not received such a recognition.

In our study we investigate global structural patterns of preferred fictional texts in comparison with non-preferred fictional texts. We also include non-fictional texts in our analysis to gauge the extent of intra- and inter-genre characteristics, i.e. similarities and differences between the two sub-categories of fictional prose (preferred and non-preferred) with non-fictional prose.

1.2 Experimental Aesthetics

Experimental Aesthetics is a sub-field of psychology that was founded by Gustav Theodor Fechner, who proposed that the aesthetic appeal of visual objects is based on stimulus properties that can be measured in an objective (formalistic) way (Fechner, 1876). This idea was pursued by Bell (1914) a few decades later, who speculated that visual artworks possess a 'significant form', which has the potential to elicit an aesthetic response in beholders across art periods and cultures. The idea that objective properties can characterize artworks has been criticized in some conceptual theories (for example, Danto, 1981; Leder et al., 2004; Gopnik, 2011), which suggest that content and cultural background suffice to explain aesthetic appreciation. Opposite theoretical stances, however, support the notion that formal structures determine how a work of art is perceived (for visual stimuli, see Kandinsky, 1912; Greenberg, 1955; Arnheim, 1974). These ideas constructed the theoretical framework and induced general acceptance that inspired a surge in contemporary aesthetic research especially in the visual domain.

Along with theoretical contemplation, aesthetic studies found their way into a variety of research fields, such as Evolutionary Aesthetics, studying the role of evolutionary processes in shaping aesthetic preferences, Neuroaesthetics, understanding neural mechanism underlying aesthetic perception and experiences, and Developmental Aesthetics, which investigates how aesthetic perception, judgments and skills develop and change in human beings. Various aspects of perception, creation, and evaluation have been scrutinized in Experimental Aesthetics. The focal point in

Experimental Aesthetics is therefore the human beholder, who has an aesthetic experience either as a perceiver or a producer.

1.3 Computational Aesthetics

Computational Aesthetics is a sub-field of artificial intelligence that concerns aesthetic assessment or production of pleasing objects. In his seminal work, George David Birkhoff proposed an "aesthetic measure", M, as a function of "order", O, and "complexity", C, in the form of $M = f(\frac{O}{C})$ (Birkhoff, 1933). His idea was that aesthetic appraisal is a balance between the amount of effort in perceiving an object and aesthetic characteristics of the object. A mathematical formalization of aesthetics was pursued by Max Bense, a German philosopher, and Abraham Moles, a French engineer, in the following decades (Nake, 2012). Ever since, one of the central questions in aesthetics studies has been what aspects of aesthetic stimuli can be linked to objective statistical properties.

In Computational Aesthetics the beholder is absent – at least explicitly. This is in contrast to Experimental Aesthetics, which directly investigates aesthetic experiences of human beholders. Nevertheless, both fields complement each other. Findings in experimental studies can be used to design, evaluate, and analyze artificial computational models. Computational Aesthetics, in return, can analyze objects categorized into different classes and find patterns that may have behavioral or neural correlates.

Computational Aesthetics has been applied to evaluate objects with different degrees of appreciation and to characterize properties of highly and less pleasing objects. Computational techniques have also been used to produce aesthetic objects. More recently, computer-generated art has become more popular thanks to astonishing progress in artificial neural models (Geller et al., 2022; for reviews, see for example, Jing et al., 2020; Santos et al., 2021).

Computational approaches, which allow for a quantitative analysis of aesthetics, have been used in various fields including vision, music, and poetry, among others. However, some fields, e.g. the textual domain, are still understudied compared to others, especially compared to the visual domain.

1.4 Visual Aesthetics

Visual Aesthetics focuses on understanding aesthetic qualities of artworks and non-artwork images. Depending on the method applied, it can be either experimental

or computational. Pictorial elements of visual stimuli are more or less well-defined and distinguishable. Moreover, visual perception is well-studied and many perceptual processes have been analyzed in great detail. As a result, experimental and computational visual aesthetics have flourished during the last decades.

Researchers who applied computational approaches in their studies suggested that visual artworks share measurable image properties which reflect a specific physical structure (more on this below). In categorizing visual properties, local and global image properties can be distinguished from each other. Local properties, such as color and luminance contrast, can be associated with specific positions in an image. Properties that reflect global image structure describe larger parts of an image or an image as a whole. Although local image features may reflect some appreciation value, visual aesthetics has particularly focused on global properties because aesthetic concepts, such as composition (McManus et al., 1985), Gestalt (Arnheim, 1974), or visual rightness (Locher et al., 1999), describe structures of images globally.

Here, we touch upon some of the relevant global properties. Fractality demonstrates recurring similar patterns on coarser and finer scales. Fractal structures have broadly been investigated in images (Mureika et al., 2005; Alvarez-Ramirez et al., 2008; Taylor et al., 2011). Curvature (Bar and Neta, 2006; Bertamini et al., 2016) and distributional features of edge orientations (Koch et al., 2010; Redies et al., 2012) are other image features that correlate with aesthetic preference. Global features can also be measured in the Fourier domain by analyzing regularities in the frequency domain (Graham and Field, 2007; Redies et al., 2007). Variability in image structures has also been studied. Features computed by a Convolutional Neural Network (CNN) were used to show that traditional visual artworks exhibit a high richness and high variability of low-level CNN feature responses (Brachmann et al., 2017).

Research in computational visual aesthetics, as suggested by above-mentioned studies, is more advanced compared to the textual domain. Therefore, we use studies in vision as a point of reference in our present study.

1.5 Text Aesthetics

Aesthetics research in the textual domain is not well-established. The main obstacle of conducting research in this domain is probably the time-distributed nature of text and consequently the incremental processing of information (see, Wallot et al., 2014; Venhuizen et al., 2019) as opposed to the perception of visual works

which is almost instant. This obstacle is especially problematic for analyzing longer texts. Another obstacle is that language constituents cannot be precisely defined or easily determined. Text components at various levels of language structures are intertwined with each other. Small changes at one level, e.g. at the lexical level, may have a profound impact on the comprehension level. This interdependency makes controlled manipulations of text through modifications of language components very difficult, if not impossible. Therefore, it is exceedingly challenging to investigate the aesthetic impacts of different text properties. By comparison, pictorial elements in images can be automatically manipulated and the result is always a 'valid' image, e.g. changing the hue of an image or rearranging object in an image.

These limitations have caused researchers to take descriptive approaches by applying computational methods to textual collections (see the references in Section 1.5.2 and also in Section 1.5.1). Exploratory studies provide a basis to design experimental investigations and to explore potential correlates of aesthetic experience in the textual domain.

Even though there is no elaborate tradition in text aesthetics, an exploration to find objective properties of texts that reflect aesthetic values has been conducted in various studies, especially on poetry, in which the units of analysis, i.e. poems, are usually short, and poetry properties, e.g. meter and rhyme, can be formally defined, and are measurable as a result.

1.5.1 Poetry

Among textual works, a poem is the most obvious text type which is expected to possess features that elicit aesthetic responses in the reader. Results from studies of rhythm, rhyme and other poetic techniques (Jakobson, 1960; Leech, 1969; Jacobs, 2015; Jacobs et al., 2016; Vaughan-Evans et al., 2016; König and Pfister, 2017; Egan et al., 2020; Menninghaus and Wallot, 2021) promote the assumption that poetic composition reflects properties that are measurable and can be studied experimentally (see, for example, Obermeier et al., 2013; Menninghaus et al., 2017). It has also been suggested that composition features of poetry, such as meter and rhyme, enhance aesthetic appreciation because they increase ease of processing (Obermeier et al., 2016) according to the theory of cognitive fluency (Reber et al., 2004). This idea is in accordance with Birkhoff's generic model of "aesthetic measure" (Birkhoff, 1933), which, as previously described, assumes a balance between aesthetic qualities of an object and the amount of effort in perceiveing that object.

Relevant studies of computational analysis of poetry can be traced back to the 1990s to the work of Simonton (1990), who analyzed the more popular and the lesser known sonnets of Shakespeare in terms of the vocabulary used. His analyses revealed a correlation between lexical diversity and the "aesthetic success" of Shakespeare's sonnets. Forsyth (2000) compared more popular and less popular poems with each other using word features, lexical diversity and the frequency distribution of syntactic tags. He showed that there were differences between the vocabularies and the distribution of specific part-of-speech tags, such as personal pronoun, for the preferred and non-preferred poems. Based on the assumption that professional and amateur writers apply language components with a different distribution, Kao and Jurafsky (2012) analyzed poems of the two groups of writers using style and content features. They showed that professional poets refer to natural objects more frequently and to abstract concepts less often compared to amateurs, who also use a richer vocabulary; nevertheless, their vocabulary incorporates a higher number of more ordinary and common words.

Affective analysis, a term closely related to aesthetic investigation, was studied in a collection of German poems and it was shown that phonological structures can explain affective ratings to some extent (Aryani et al., 2016). In a multimodal setting, melodic recurrence in time series of recited poems were studied using autocorrelation coefficients and spectral exponents. Results showed a positive correlation between the two acoustic measures and aesthetic ratings (Scharinger et al., 2022).

What is surprisingly missing in these studies is that poetic properties of composition, such as rhythm und rhyme, have either no or a minor role in the analyses, even though it is assumed that poetic techniques can be formally defined and measured in texts.

1.5.2 Prose

Aesthetic aspects of prose texts are less distinct and have been discussed more implicitly compared to poetry texts in previous research. "Readability", for example, has been used to reflect quality of text or proficiency in writing. However, it should be noted that readability is not equivalent to "quality". Traditional readability metrics, such as the Gunning FOG Index (Robert, 1952) and the Flesch-Kincaid Grade Level (Kincaid et al., 1975), use shallow features of words, sentences and documents to determine the difficulty –from the perspective of the reader– or proficiency –from the perspective of the writer– of a text. Later on, automatic readability assessment became popular by using language models (Si and Callan, 2001; Collins-Thompson

and Callan, 2004) and thereafter in combination with grammatical information (Schwarm and Ostendorf, 2005; Heilman et al., 2007; Heilman et al., 2008). Second language acquisition is a related research and application field, in which lexical features, such as vocabulary richness, have been used to assess the proficiency of writers (Laufer and Nation, 1995; Zareva et al., 2005; Yu, 2009).

Researchers have also investigated the quality of writing in news articles (Pitler and Nenkova, 2008) and scientific texts (Louis and Nenkova, 2013) using a variety of lexical, grammatical and discourse features. One of earliest studies related to literary prose analysis is the work by Louwerse et al. (2008), who used distributions of uni- and bi-grams, Latent Semantic Analyses (LSA) and a hierarchical clustering algorithm to distinguish literary from non-literary texts with a high accuracy.

Lexical, grammatical and semantic features have also been used to analyze preferred and less-preferred literary texts. McIntyre and Lapata (2009) designed a story generation system, in which an "interest model" was designed based on ratings by a group of participants on very short fairy tales with the size of a few sentences. They showed a correlation between word and POS-tag frequencies and the interestingness of stories. Ashok et al. (2013) carried out a study based on data from the website of Project Gutenberg. They assumed that download counts reflect the preference of readers, which in turn shows how successful a text has been. Using grammatical and sentiment features, they classified successful and less successful novels with an acceptable performance. Maharjan et al. (2017) used readers' ratings from the website Goodreads, a social network for commenting, recommending and rating books, as a measure of likability for each book. They explored a wide variety of lexical, syntactical and readability features, and classified texts from 8 different genres with a relatively highly accuracy. The same dataset was used in another study to predict the success of books using semantic features (Saba et al., 2021). A concept model for each book was defined using the distance between vector embeddings of words in each book and concepts in Roger's thesaurus. The Filter Method, a feature selection technique, was applied to improve the performance. Even though the best-performing concept model for each genre was different from the others, the F1-scores, which are calculated as the harmonic means of percision and recall, were averaged and a high score was reported in the end. In a survey-based study, participants were asked to rate contemporary Dutch novels texts by the degree of their "literariness", a vague term that was left to interpretation. Frequency distributions of bigrams and very simple features were used to predict the human ratings

(Cranenburgh and Koolen, 2015). In the follow-up studies, frequency distribution of lexical and syntactic features (Cranenburgh and Bod, 2017), and summary statistics derived from a topic modeling and neural paragraph embeddings (Cranenburgh et al., 2019) were used to model human ratings of "literariness". The inherent vagueness of the term "literariness", upon which these studies have been based, makes it difficult to draw conclusions from the results about aesthetics and to establish a clear connection with the preference of readers.

Literary works can also be affective and evoke emotions. Applying computer-assisted methods to analyze emotion in literary texts was initiated by Anderson and McMaster (1982) and was continued by analyzing fairy tales and novels in other research (Alm and Sproat, 2005; Francisco and Gervás, 2006; Kakkonen and Galić Kakkonen, 2011; Mohammad, 2011; Reagan et al., 2016; for a review on emotion analysis in literary studies, see, Kim and Klinger, 2021). In the study by Maharjan et al., 2018a, the flow of emotions was used to predict the success (popularity) of a text. It was also shown that if the flow of emotions is combined with genre information, the performance of likeability prediction of fictional texts can be improved (Maharjan et al., 2018b).

Scale-invariant and fractal patterns are global structural properties that can be revealed using long-range correlation analysis. Fractal analysis techniques has been widely applied to images (e.g. Wendt and Abry, 2007; Li et al., 2009; Wendt et al., 2009; Ji et al., 2013) and music (e.g. Hsü, 1993; Bigerelle and Iost, 2000; Levitin et al., 2012; Teixeira Borges et al., 2019). It has also been shown that fractal patterns are observable in text structures (see, for example, Drożdż et al., 2016; Yang et al., 2016). However, fractal analysis has not been applied in computational textual aesthetics in the sense of modeling textual appraisal and preference. An exception is recent research on fractality of sentiment arcs that were predictive of reader's appreciation of fairy tales (Bizzoni et al., 2021) and could locate works by Nobel-prize winners, representative of high quality texts, in a specific range on the spectrum of fractal values.

Entropy metrics are capable of measuring uncertainty and irregularity. They thus measure unpredictability as a structural property of the system. Entropy has been extensively used to analyze distributional laws of linguistics, e.g. word order (Montemurro and Zanette, 2011; Montemurro and Zanette, 2016; Futrell et al., 2015; Koplenig et al., 2017) and word length (Piantadosi et al., 2011; Mahowald et al., 2013; Ferrer-i-Cancho et al., 2015; Kanwal et al., 2017), or for a comparison of

languages in terms of ordering preferences and complexity (Bentz et al., 2015; Kalimeri et al., 2015; Ehret and Szmrecsanyi, 2016; Hernández-Gómez et al., 2017; Bentz et al., 2017). Nevertheless, entropy measures have been rarely used to investigate aesthetic preference of fictional prose. In an analysis of literary texts written by English and Spanish Nobel laureates and non-Nobel laureates, Febres and Jaffe (2017) used entropy, along with other features, and showed that there is a correlation between entropy values of texts and the two categories of authors.

The aforementioned studies presented evidence that there are correlates of readers' preference in text. However, the studies mostly focused on statistical and distributional linguistics properties and did not relate aesthetic preference to structural linguistic properties. In our study, we rather focus on structural properties that reflect the global organization of text and are more generalizable to different genres and time periods. We analyze global structural properties in preferred fictional text and compare them with those found in non-preferred fictional texts.

1.6 Structure of the Dissertation

In the current chapter, which serves as an introduction to this dissertation, we established a connection between our research, previous studies and other relevant fields, especially visual aesthetics. In the following chapters, we first describe our analysis methods comprehensively in Chapter 2 by providing technical details and visualizations of some examples. Chapter 3 explains our first experiment of modeling textual preference using fractality and variability. Chapter 4 introduces the terms predictability and surprise in the field of textual aesthetic. Experiments in Chapter 4 were extended in Chapter 5 to include texts from two different time periods in order to analyze potentially time-dependent and time-invariant correlates of preference. Our findings are discussed in Chapter 7, which is followed by an outlook on our ongoing projects in Chapter 8.

References

Alm, Cecilia Ovesdotter and Richard Sproat (2005). "Emotional Sequencing and Development in Fairy Tales". In: *Affective Computing and Intelligent Interaction*. Ed. by Jianhua Tao, Tieniu Tan, and Rosalind W. Picard. Berlin, Heidelberg: Springer, pp. 668–674. DOI: 10.1007/11573548_86.

Alvarez-Ramirez, Jose, Juan Echeverria, and Eduardo Rodriguez (2008). "Performance of a high-dimensional R/S method for Hurst exponent estimation". In: *Physica A: Statistical Mechanics and its Applications* 387, pp. 6452–6462. DOI: 10.1016/j.physa.2008.08.014.

Anderson, CW and GE McMaster (1982). "Computer assisted modeling of affective tone in written documents". In: *Computers and the Humanities* 16, pp. 1–9.

Arnheim, Rudolf (1974). *Art and Visual Perception: A Psychology of the Creative Eye*. Berkeley: University of California Press.

Aryani, Arash, Maria Kraxenberger, Susann Ullrich, Arthur M Jacobs, and Markus Conrad (2016). "Measuring the basic affective tone of poems via phonological saliency and iconicity." In: *Psychology of Aesthetics, Creativity, and the Arts* 10.2, pp. 191–204. DOI: 10.1037/aca0000033.

Ashok, Vikas, S. Feng, and Y. Choi (2013). "Success With Style: Using Writing Style to Predict the Success of Novels". In: *Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 1753–1764.

Bar, Moshe and Maital Neta (2006). "Humans Prefer Curved Visual Objects". In: *Psychological Science* 17.8, pp. 645–648. DOI: 10.1111/j.1467-9280.2006.01759.x.

Bell, Clive (1914). *Art*. London: Chatoo & Windus.

Bentz, Christian, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i-Cancho (2017). "The Entropy of Words—Learnability and Expressivity across More than 1000 Languages". In: *Entropy* 19.6. DOI: 10.3390/e19060275.

Bentz, Christian, Annemarie Verkerk, Douwe Kiela, Felix Hill, and Paula Buttery (2015). "Adaptive Communication: Languages with More Non-Native Speakers Tend to Have Fewer Word Forms". In: *PLoS ONE*, p. 0128254. DOI: 10.1371/journal.pone.0128254.

Bertamini, Marco, Letizia Palumbo, Tamara Nicoleta Gheorghes, and Mai Galatsidas (2016). "Do observers like curvature or do they dislike angularity?" In: *British Journal of Psychology* 107.1, pp. 154–178.

Bigerelle, M. and A. Iost (2000). "Fractal dimension and classification of music". In: *Chaos, Solitons & Fractals* 11.14, pp. 2179–2192. DOI: 10.1016/S0960-0779(99)00137-X.

Birkhoff, George David (1933). *Aesthetic Measure*. Cambridge, Massachusetts: Harvard University Press.

Bizzoni, Yuri, Telma Peura, Mads Rosendahl Thomsen, and Kristoffer Nielbo (2021). "Sentiment Dynamics of Success: Fractal Scaling of Story Arcs Predicts Reader Preferences". In: *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*. NIT Silchar, India: NLP Association of India (NLPAI), pp. 1–6.

Brachmann, Anselm, Erhardt Barth, and Christoph Redies (2017). "Using CNN features to better understand what makes visual artworks special". In: *Frontiers in Psychology* 8, p. 830. DOI: 10.3389/fpsyg.2017.00830.

Chatterjee, Anjan (2011). "Neuroaesthetics: A Coming of Age Story". In: *Journal of cognitive neuroscience* 23, pp. 53–62. DOI: 10.1162/jocn.2010.21457.

Collins-Thompson, Kevyn and James P. Callan (2004). "A Language Modeling Approach to Predicting Reading Difficulty". In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*. Boston, Massachusetts, USA: Association for Computational Linguistics, pp. 193–200.

Cranenburgh, Andreas van 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*. Valencia, Spain: Association for Computational Linguistics, pp. 1228–1238.

Cranenburgh, Andreas van, Karina van Dalen-Oskam, and Joris van Zundert (2019). "Vector space explorations of literary language". In: *Lang Resources & Evaluation 53*, pp. 625–650. DOI: 10.1007/s10579-018-09442-4.

Cranenburgh, Andreas van and Corina Koolen (2015). "Identifying Literary Texts with Bigrams". In: *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver, Colorado, USA: Association for Computational Linguistics, pp. 58–67. DOI: 10.3115/v1/W15-0707.

Danto, Arthur Coleman (1981). *The Transfiguration of the Commonplace: A Philosophy of Art*. Cambridge: Harvard University Press.

Drożdż, StanisŁaw et al. (2016). "Quantifying Origin and Character of Long-Range Correlations in Narrative Texts". In: *Information Sciences* 331, pp. 32–44. DOI: 10.1016/j.ins.2015.10.023.

Egan, Ciara, Filipe Cristino, Joshua S. Payne, Guillaume Thierry, and Manon W. Jones (2020). "How alliteration enhances conceptual–attentional interactions in reading". In: *Cortex* 124, pp. 111–118. DOI: 10.1016/j.cortex.2019.11.005.

Ehret, Katharina and Benedikt Szmrecsanyi (2016). "An Information-Theoretic Approach to Assess Linguistic Complexity". In: *Complexity, Isolation, and Variation*. Ed. by Raffaela Baechler and Guido Seiler. De Gruyter, pp. 71–94. DOI: 10.1515/9783110348965-004.

Febres, Gerardo and Klaus Jaffe (2017). "Quantifying Structure Differences in Literature Using Symbolic Diversity and Entropy Criteria". In: *Journal of Quantitative Linguistics* 24.1, pp. 16–53. DOI: 10.1080/09296174.2016.1169847.

Fechner, Gustav Theodor (1876). *Vorschule der Ästhetik*. Leipzig: Breitkopf and Härtel.

Ferrer-i-Cancho, Ramon, Chris Bentz, and Caio Seguin (2015). "Compression and the Origins of Zipf's Law of Abbreviation". In: *ArXiv* abs/1504.04884, pp. 1–36.

Forsyth, Richard S. (2000). "Pops and Flops: Some Properties of Famous English Poems". In: *Empirical Studies of The Arts* 18, pp. 49–67. DOI: 10.2190/E7Q8-6062-K6H4-XFRW.

Francisco, Virginia and Pablo Gervás (2006). "Exploring the compositionality of emotions in text: Word emotions, sentence emotions and automated tagging". In: *Workshop on Computational Aesthetics: Artificial Intelligence Approaches to Beauty and Happiness*, pp. 1–6.

Futrell, Richard, Kyle Mahowald, and Edward Gibson (2015). "Quantifying Word Order Freedom in Dependency Corpora". In: *Proceedings of the Third International Conference on Dependency Linguistics*. Uppsala, Sweden: Uppsala University, Uppsala, Sweden, pp. 91–100.

Geller, Hannah Alexa, Ralf Bartho, Katja Thömmes, and Christoph Redies (2022). "Statistical image properties predict aesthetic ratings in abstract paintings created by neural style transfer". In: *Frontiers in Neuroscience* 16, p. 999720. DOI: 10.3389/fnins.2022.999720.

Gopnik, Blake (2011). "Aesthetic Science and Artistic Knowledge". In: *Aesthetic Science: Connecting Minds, Brains, and Experience*. Oxford University Press. DOI: 10.1093/acprof:oso/9780199732142.003.0036.

Graham, Daniel and David Field (2007). "Statistical regularities of art images and natural scenes: Spectra, sparseness and nonlinearities". In: *Spatial Vision* 21.1-2, pp. 149–164. DOI: 10.1163/156856807782753877.

Greenberg, Clement (1955). "American-type painting". In: *Partisan Review* 22, pp. 179–196.

Heilman, Michael, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi (2007). "Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts". In: *Conference of the North American Chapter of the Association for Computational Linguistics*. Rochester, New York: Association for Computational Linguistics, pp. 460–467.

Heilman, Michael, Kevyn Collins-Thompson, and Maxine Eskenazi (2008). "An Analysis of Statistical Models and Features for Reading Difficulty Prediction". In: *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*. Columbus, Ohio, USA: Association for Computational Linguistics. DOI: 10.3115/1631836.1631845.

Hernández-Gómez, Candelario, Rogelio Basurto-Flores, Bibiana Obregón-Quintana, and Lev Guzmán-Vargas (2017). "Evaluating the Irregularity of Natural Languages". In: *Entropy* 19, p. 521. DOI: 10.3390/e19100521.

Hsü, Kenneth J. (1993). "Fractal Geometry of Music: From Bird Songs to Bach". In: *Applications of Fractals and Chaos*. Berlin, Heidelberg: Springer, pp. 21–39.

Jacobs, Arthur M. (2015). "Neurocognitive poetics: methods and models for investigating the neuronal and cognitive-affective bases of literature reception". In: *Frontiers in Human Neuroscience* 9, p. 186. DOI: 10.3389/fnhum.2015.00186.

Jacobs, Arthur M., Jana Lüdtke, Arash Aryani, Burkhard Meyer-Sickendieck, and Markus Conrad (2016). "Mood-empathic and aesthetic responses in poetry reception A model-guided, multilevel, multimethod approach". In: *Scientific Study of Literature* 6, pp. 87–130. DOI: 10.1075/ssol.6.1.06jac.

Jakobson, Roman (1960). "Linguistics and poetics". In: *Style in Language*. Cambridge, MA: MIT Press, pp. 350–377.

Ji, Hui, Xiong Yang, Haibin Ling, and Yong Xu (2013). "Wavelet domain multifractal analysis for static and dynamic texture classification". In: *IEEE Transactions on Image Processing* 22.1, pp. 286–299. DOI: 10.1109/TIP.2012.2214040.

Jing, Yongcheng, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song (2020). "Neural Style Transfer: A Review". In: *IEEE Transactions on Visualization and Computer Graphics* 26.11, pp. 3365–3385. DOI: 10.1109/TVCG.2019.2921336.

Kakkonen, Tuomo and Gordana Galić Kakkonen (2011). "SentiProfiler: Creating Comparable Visual Profiles of Sentimental Content in Texts". In: *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*. Hissar, Bulgaria: Association for Computational Linguistics, pp. 62–69.

Kalimeri, Maria, Vassilios Constantoudis, Constantinos Papadimitriou, Konstantinos Karamanos, Fotis K. Diakonos, and Harris Papageorgiou (2015). "Word-length Entropies and Correlations of Natural Language Written Texts". In: *Journal of Quantitative Linguistics* 22.2, pp. 101–118. DOI: 10.1080/09296174.2014.1001636.

Kandinsky, Wassily (1912). *Über das Geistige in der Kunst, insbesondere in der Malerei*. München: Piper.

Kanwal, Jasmeen, Kenny Smith, Jennifer Culbertson, and Simon Kirby (2017). "Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication". In: *Cognition* 165, pp. 45–52. DOI: 10.1016/j.cognition.2017.05.001.

Kao, Justine 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*. Montréal, Canada: Association for Computational Linguistics, pp. 8–17.

Kim, Evgeny and Roman Klinger (2021). "A Survey on Sentiment and Emotion Analysis for Computational Literary Studies". In: *Zeitschrift für digitale Geisteswissenschaften*. DOI: 10.17175/2019_008.

Kincaid, J Peter, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom (1975). *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for navy enlisted personnel*. Tech. rep. Naval Technical Training Command Millington TN Research Branch.

Koch, Michael, Joachim Denzler, and Christoph Redies (2010). "$1/f^2$ Characteristics and isotropy in the fourier power spectra of visual art, cartoons, comics, mangas, and different categories of photographs". In: *PLoS ONE* 5.8, pp. 1–11. DOI: 10.1371/journal.pone.0012268.

König, Ekkehard and Manfred Pfister (2017). *Literary Analysis and Linguistics*. Berlin: Erich Schmidt Verlag.

Koplenig, Alexander, Peter Meyer, Sascha Wolfer, and Carolin Müller-Spitzer (2017). "The Statistical Trade-off Between Word Order and Word Structure – Large-Scale Evidence for the Principle of Least Effort". In: *PLoS ONE* 12.3, pp. 1–25. DOI: 10.1371/journal.pone.0173614.

Laufer, Batia and Paul Nation (1995). "Vocabulary size and use: Lexical richness in L2 written production". In: *Applied Linguistics* 16.3, pp. 307–322. DOI: 10.1093/applin/16.3.307.

Leder, Helmut, Benno Belke, Andries Oeberst, and Dorothee Augustin (2004). "A model of Aesthetic Appreciation and Aesthetic Judgments". In: *British journal of psychology* 95.4, pp. 489–508.

Leech, Geoffrey N (1969). *A Linguistic Guide to English Poetry*. London: Harlow, Longmans.

Levitin, Daniel J., Parag Chordia, and Vinod Menon (2012). "Musical rhythm spectra from Bach to Joplin obey a 1/f power law". In: *Proceedings of the National Academy of Sciences* 109.10, pp. 3716–3720. DOI: 10.1073/pnas.1113828109.

Li, Jian, Qian Du, and Caixin Sun (2009). "An improved box-counting method for image fractal dimension estimation". In: *Pattern Recognition* 42.11, pp. 2460–2469. DOI: 10.1016/j.patcog.2009.03.001.

Locher, Paul J, Pieter Jan Stappers, and Kees Overbeeke (1999). "An empirical evaluation of the visual rightness theory of pictorial composition". In: *Acta Psychologica* 103.3, pp. 261–280. DOI: 10.1016/S0001-6918(99)00044-X.

Louis, Annie and Ani Nenkova (2013). "What Makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain". In: *Transactions of the Association for Computational Linguistics* 1, pp. 341–352. DOI: 10.1162/tacl_a_00232.

Louwerse, Max, Nick Benesh, and Bin Zhang (2008). "Computationally discriminating literary from non-literary texts". In: *Directions in empirical literary studies: In honor of Willie Van Peer*. Amsterdam: John Benjamins, pp. 175–191.

Maharjan, Suraj, John Arevalo, Manuel Montes, Fabio 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*. Valencia, Spain: Association for Computational Linguistics, pp. 1217–1227. DOI: 10.18653/v1/E17-1114.

Maharjan, Suraj, Sudipta Kar, Manuel Montes, Fabio A. González, and Thamar Solorio (2018a). "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)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 259–265. DOI: 10.18653/v1/N18-2042.

Maharjan, Suraj, Manuel Montes, Fabio A. González, and Thamar Solorio (2018b). "A Genre-Aware Attention Model to Improve the Likability Prediction of Books". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 3381–3391. DOI: 10.18653/v1/D18-1375.

Mahowald, Kyle, Evelina Fedorenko, Steven T Piantadosi, and Edward Gibson (2013). "Info/Information Theory: Speakers Choose Shorter Words in Predictive Contexts". In: *Cognition* 126, pp. 313–318. DOI: 10.1016/j.cognition.2012.09.010.

McIntyre, Neil and Mirella Lapata (2009). "Learning to Tell Tales: A Data-driven Approach to Story Generation". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore: Association for Computational Linguistics, pp. 217–225.

McManus, IC, D Edmondson, and J Rodger (1985). "Balance in pictures". In: *British Journal of Psychology* 76.3, pp. 311–324. DOI: 10.1111/j.2044-8295.1985.tb01955.x.

Menninghaus, Winfried, Valentin Wagner, Eugen Wassiliwizky, Thomas Jacobsen, and Christine A. Knoop (2017). "The emotional and aesthetic powers of parallelistic diction". In: *Poetics* 63, pp. 47–59. DOI: 10.1016/j.poetic.2016.12.001.

Menninghaus, Winfried and Sebastian Wallot (2021). "What the eyes reveal about (reading) poetry". In: *Poetics*, p. 101526. URL: https://doi.org/10.1016/j.poetic.2020.101526.

Mohammad, Saif (2011). "From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales". In: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Portland, OR, USA: Association for Computational Linguistics, pp. 105–114.

Montemurro, Marcelo A. and Damián H. Zanette (2011). "Universal Entropy of Word Ordering Across Linguistic Families". In: *PLoS ONE* 6.5, pp. 1–9. DOI: 10.1371/journal.pone.0019875.

Montemurro, Marcelo A. and Damián H. Zanette (2016). "Complexity and Universality in the Long-Range Order of Words". In: *Creativity and Universality in Language*. Cham: Springer, pp. 27–41. DOI: 10.1007/978-3-319-24403-7_3.

Mureika, Jonas, Charles Dyer, and Gerald Cupchik (2005). "Multifractal structure in nonrepresentational art". In: *Physical review. E, Statistical, nonlinear, and soft matter physics* 72, p. 046101. DOI: 10.1103/PhysRevE.72.046101.

Nake, Frieder (2012). "Information Aesthetics: An heroic experiment". In: *Journal of Mathematics and the Arts* 6.2-3, pp. 65–75. DOI: 10.1080/17513472.2012.679458.

Obermeier, Christian, Sonja Kotz, Sarah Jessen, Tim Raettig, Martin Koppenfels, and Winfried Menninghaus (2016). "Aesthetic appreciation of poetry correlates with ease of processing in event-related potentials". In: *Cognitive, Affective & Behavioral Neuroscience* 16. DOI: 10.3758/s13415-015-0396-x.

Obermeier, Christian et al. (2013). "Aesthetic and Emotional Effects of Meter and Rhyme in Poetry". In: *Frontiers in Psychology* 4. DOI: 10.3389/fpsyg.2013.00010.

Piantadosi, Steven T., Harry Tily, and Edward Gibson (2011). "Word Lengths Are Optimized for Efficient Communication". In: *Proceedings of the National Academy of Sciences* 108.9, pp. 3526–3529. DOI: 10.1073/pnas.1012551108.

Pitler, Emily and Ani Nenkova (2008). "Revisiting Readability: A Unified Framework for Predicting Text Quality". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, pp. 186–195.

Reagan, Andrew J, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds (2016). "The Emotional Arcs of Stories Are Dominated by Six Basic Shapes". In: *EPJ Data Science* 5.1. DOI: 10.1140/epjds/s13688-016-0093-1.

Reber, Rolf, Norbert Schwarz, and Piotr Winkielman (2004). "Processing Fluency and Aesthetic Pleasure: Is Beauty in the Perceiver's Processing Experience?" In: *Personality and Social Psychology Review* 8.4, pp. 364–382. DOI: 10.1207/s15327957pspr0804_3.

Redies, Christoph, Seyed Ali Amirshahi, Michael Koch, and Joachim Denzler (2012). "PHOG-Derived Aesthetic Measures Applied to Color Photographs of Artworks, Natural Scenes and Objects". In: *Computer Vision – ECCV 2012. Workshops and Demonstrations. Lecture Notes in Computer Science*. Vol. 7583. Berlin: Springer-Verlag, pp. 522–531. DOI: 10.1007/978-3-642-338-63-2_54.

Redies, Christoph, Jens Hasenstein, and Joachim Denzler (2007). "Fractal-like image statistics in visual art: similarity to natural scenes". In: *Spatial Vision* 21.1-2, pp. 137–148. DOI: 10.1163/156856807782753921.

Robert, Gunning (1952). *The technique of clear writing*. New York: McGraw-Hill.

Saba, Syeda Jannatus, Biddut Sarker Bijoy, Henry Gorelick, Sabir Ismail, Md Saiful Islam, and Mohammad Amin (2021). "A Study on Using Semantic Word Associations to Predict the Success of a Novel". In: *The Tenth Joint Conference on Lexical and Computational Semantics*. Bangkok, Thailand (online): Association for Computational Linguistics, pp. 38–51. DOI: 10.18653/v1/2021.starsem-1.4.

Santos, Iria, Mª Castro Pena, Nereida Rodriguez-Fernandez, Alvaro Torrente-Patiño, and Adrián Carballal (2021). "Artificial Neural Networks and Deep Learning in the Visual Arts: a review". In: *Neural Computing and Applications* 33, pp. 1–37. DOI: 10.1007/s00521-020-05565-4.

Scharinger, Mathias, Valentin Wagner, Christine A Knoop, and Winfried Menninghaus (2022). "Melody in poems and songs: Fundamental statistical properties predict aesthetic evaluation". In: *Psychology of Aesthetics, Creativity, and the Arts*, pp. 163–177. DOI: 10.1037/aca0000465.

Schwarm, Sarah and Mari Ostendorf (2005). "Reading Level Assessment Using Support Vector Machines and Statistical Language Models". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 523–530. DOI: 10.3115/1219840.1219905.

Si, Luo and Jamie Callan (2001). "A Statistical Model for Scientific Readability". In: *Proceedings of the Tenth International Conference on Information and Knowledge Management*. CIKM '01. Atlanta, Georgia, USA: Association for Computing Machinery, pp. 574–576. DOI: 10.1145/502585.502695.

Simonton, Dean Keith (1990). "Lexical Choices and Aesthetic Success: A Computer Content Analysis of 154 Shakespeare Sonnets". In: *Computers and the Humanities* 24.4, pp. 251–264. DOI: 10.1007/BF00123412.

Taylor, Richard, Branka Spehar, Caroline Hagerhall, and Paul Van Donkelaar (2011). "Perceptual and physiological responses to Jackson Pollock's fractals". In: *Frontiers in Human Neuroscience* 5, p. 60. DOI: 10.3389/fnhum.2011.00060.

Teixeira Borges, Ana Filipa, Mona Irrmischer, Thomas Brockmeier, Dirk Smit, Huibert Mansvelder, and Klaus Linkenkaer-Hansen (2019). "Scaling behaviour in music and cortical dynamics interplay to mediate music listening pleasure". In: *Scientific Reports* 9, pp. 1–15. DOI: 10.1038/s41598-019-54060-x.

Vaughan-Evans, Awel, Robat Trefor, Llion Jones, Peredur Lynch, Manon W. Jones, and Guillaume Thierry (2016). "Implicit Detection of Poetic Harmony by the Naïve Brain". In: *Frontiers in Psychology* 7, p. 1859. DOI: 10.3389/fpsyg.2016.01859.

Venhuizen, Noortje J., Matthew W. Crocker, and H. Brower (2019). "Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience". In: *Discourse Processes* 56.3, pp. 229–255. DOI: 10.1080/0163853X.2018.1448677.

Wallot, Sebastian, Beth A O'Brien, Anna Haussmann, Heidi Kloos, and Marlene S Lyby (2014). "The Role of Reading Time Complexity and Reading Speed in Text Comprehension". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 40 (6), pp. 1745–1765. DOI: 10.1037/xlm0000030.

Wendt, H. and P. Abry (2007). "Multifractality Tests Using Bootstrapped Wavelet Leaders". In: *IEEE Transactions on Signal Processing* 55.10, pp. 4811–4820. DOI: 10.1109/TSP.2007.896269.

Wendt, Herwig, Stéphane G. Roux, Stéphane Jaffard, and Patrice Abry (2009). "Wavelet leaders and bootstrap for multifractal analysis of images". In: *Signal Processing* 89.6, pp. 1100–1114. DOI: 10.1016/j.sigpro.2008.12.015.

Yang, Tianguang, Changgui Gu, and Huijie Yang (2016). "Long-Range Correlations in Sentence Series from A Story of the Stone". In: *PLoS ONE* 11.9, pp. 1–11. DOI: 10.1371/journal.pone.0162423.

Yu, Guoxing (2009). "Lexical diversity in writing and speaking task performances". In: *Applied Linguistics* 31.2, pp. 236–259. DOI: 10.1093/applin/amp024.

Zareva, Alla, Paula Schwanenflugel, and Yordanka Nikolova (2005). "Relationship between lexical competence and language proficiency: Variable sensitivity". In: *Studies in Second Language Acquisition* 27.4, pp. 567–595. DOI: 10.1017/S0272263105050254.

Chapter 2

METHODOLOGY

In this chapter, we introduce ways of analyzing text that we utilize in the following chapters. Our point of departure are studies in the visual domain, in which there is a long-standing tradition of experimental and descriptive research in aesthetics (Fechner, 1876; Arnheim, 1974; Chatterjee and Vartanian, 2014; Jacobs, 2015; Redies, 2015). In vision, a variety of global structural properties of artworks and non-artwork images has been studied. Examples of global structural properties that relate to preference of visual stimuli are fractal (self-similar) properties (Taylor et al., 2011), entropy, i.e. the amount of irregularity, in edge orientation distribution (Redies et al., 2017), and variability in the low-level neural representation of pictorial elements (Brachmann et al., 2017). In our study, we also use methods which center around analyzing variability, fractality and predictability in underlying structures of texts. The association of these properties with visually pleasing stimuli further highlights their potential significance in analyzing the structural composition and organization of texts across various categories.

2.1 Representation of Texts

The central hypothesis of our study is that text categories of different aesthetic quality can be characterized by measurable textual correlates, which describe global structural properties of text. Speaking of textual correlates, various linguistic and structural language units could potentially reflect aesthetic preference. At the lower level of the language hierarchy lie basic units such as words, phrases, sentences, which are associated with syntactic classes of part-of-speech (POS), grammatical constituents and syntactic structures. At the higher level, in which comprehension occurs, there are phenomena that can be less straightforwardly measured as they are less formally definable, compared to language units at the lower-levels, and they are more prone to variation according to assumptions and definitions. Topic, plot and diversity in lexicon are example of higher-level properties. Our goal is to analyze features that reflect global structure of text. We thus measure several properties along text and represent each text by sequences (series) of lower- and higher-level text property values.

The text properties which we used in our experiments will be explained in the

following chapters. For the purpose of this chapter, which is to introduce analysis methods in detail and to support technical details with examples, we use only one text property, i.e. sentence length, to represent texts. We count the number of tokens per sentences along the texts and convert each text to a series of sentence length values. We first describe each method thoroughly and then apply it to sentence length series of some sample texts. The texts are selected from the JEFP corpus (for more information about the corpus, see, Chapters 3 and 4 / Mohseni et al., 2021; Mohseni et al., 2022). To sentencize and tokenize texts, we used the `Stanza` package for python (Qi et al., 2020).

2.2 Variability

Variability analysis has proven to be a successful approach in studies related to the field of visual aesthetics (e.g. Brachmann et al., 2017; Geller et al., 2022). Variability can be determined using various statistics, e.g. range, interquartile range and variance. Variance is the most often used measure of dispersion and shows the amount of deviation from the mean of the population. The advantage of using variance, or its square root, i.e. the standard deviation, is that the two metrics take all samples into computation. Variance is computed as:

$$\mathcal{V}(X) = E[(X - \mu)^2] \tag{2.1}$$

in which $E[.]$ denotes the expected value and μ is the mean of X, $\mu = E[X]$. The more values of a random variable deviate from the mean, the higher the variance of the population will be. For example, if we measure the length of sentences as the number of tokens per sentence in a text, the amount of dispersion of sentence lengths can be demonstrated by variance, which is obtained by Eq. 2.1.

To show examples we selected two texts from the JEFP corpus. The first text is *The Golden Bowl* by *Henry James*, which is a canonical prose text initially published in 1904. The second text is *The Puppet Crown* by *Harold MacGrath*, which belongs to the non-canonical category in the corpus and was firstly published in 1901. Figure 2.1 visualizes the distribution of sentence lengths for the two texts. *The Golden Bowl* has a very high variance of 499, while the variance of *The Puppet Crown* is only 74.

2.3 Fractality and Long-Range Correlations

Fractality quantifies dynamically fluctuating variability of complex systems through multi-scale analyses and provides insights into underlying structures of objects under

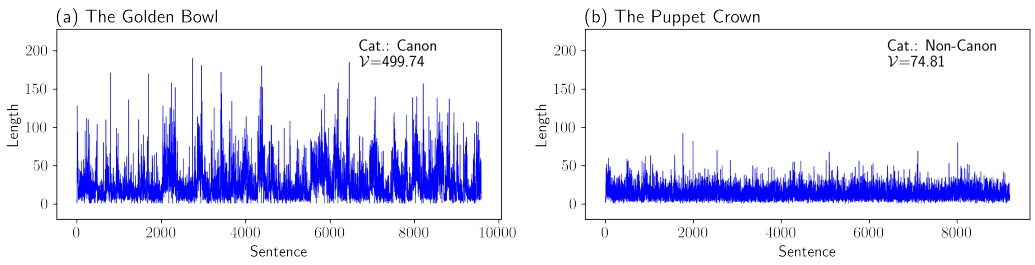

Figure 2.1: Sentence length series of (a) *The Golden Bowl* by *Henry James* and (b) *The Puppet Crown* by *Harold MacGrath*. The category (Cat.) and the variance, $\mathcal{V}$, of each text are shown inside each panel.

study. While variability, as described in the previous section, shows the amount of dispersion of property values, fractality and long-range correlation analysis demonstrates how fluctuations in values are distributed along a text. Probably the most widely used methods to analyze long-range correlations are Detrended Fluctuation Analysis (DFA; Peng et al., 1994) and Multi-Fractal Detrended Fluctuation Analysis (MFDFA; Kantelhardt et al., 2002; Oświecimka et al., 2006), which is an extension of DFA.

In what follows, we outline the MFDFA procedure, according to Kantelhardt et al. (2002), followed by a detailed account of each step and an elaboration on how multifractal characteristics are computed.

Given a series $X = x(1), x(2), \cdots, x(N)$, MFDFA processes the series as follows:

1. Build the profile of the series by subtracting the mean and computing the cumulative sum:

$$Y(i) = \sum_{k=1}^{i} [x(k) - \langle X \rangle], i = 1, \cdots, N \tag{2.2}$$

   in which $\langle X \rangle$ is the mean of X.

2. Divide the profile of the series into $N_s = N/s$ windows for different values of s, which is the size of windows. As the length of the series, N, may not be always divisible by s and a portion of the series in the end may be excluded from the computation, the windowing procedure is repeated starting from the end. As a result, the number of windows increases to $2 \times N_s$.

3. Detrend the values in each window v, $v = 1, \cdots, 2 \times N_s$, by subtracting the best fitting line, Y', and calculate the mean square fluctuation of residuals:

$$F^2(s, v) = \frac{1}{s} \sum_{i=1}^{s} [Y(s \times (v-1) + i) - Y'(s \times (v-1) + i)]^2 \tag{2.3}$$

4. Calculate the qth order of the mean square fluctuations:

$$F_q(s) = \{\frac{1}{2 \times N_s} \sum_{v=1}^{N_s} [F^2(s, v)]^{q/2}\}^{1/q} \tag{2.4}$$

5. Compute the growth factor of fluctuations, $h(q)$, using a log-log regression on $F_q(s)$ values, i.e. $\log F_q(s) \sim h(q) \times \log s$

2.3.1 Creation and Segmentation of the Profile of Series

The procedure of MFDFA is based on a *random walk* (Pearson, 1905), which is a mathematical random process to model dynamics of systems. In a random walk, the variance of the process is time-dependent and it increases as more steps are taken. For illustration, suppose that an integer variable is initially set to 0 and its value increases or decreases by 1 point, i.e. either +1 or -1, according to a probability model. One can compute the mean and the variance of the values at each step. If the model has no preference for either direction, neither increase nor decrease, the expected value of the variable, i.e. the mean, remains fixed at zero. However, the variance of the values increases proportional to the number of increases and decreases. This implies that the standard deviation increases at a rate of the square root of the number of increases and decreases as standard deviation is the square root of variance. Consequently, the scaling factor of such a process, which exhibits no long-range correlations, is computed at 0.5. This behavior is typical of white noise, which exhibits no long-range correlations.

If a system has some persistency and chooses values similar to its recent choices, the scaling factor increases, indicating the presence of long-range correlations in the series of the values. In this so-called persistent system, the big events tend to be succeeded by big events, while small events are more likely followed by small events. In the opposite case, if a system shows an anti-persistent behavior, the standard deviation of the series drops to values below 0.5. In an anti-persistent system, there is a higher probability that small events follow big events than following small events, and vice versa.

As a real example from the field of text processing, suppose that we measure lengths of sentences in a text in terms of the number of tokens. If the length of a sentence is independent from the length of preceding sentences, the sentence length series exhibits no long-range correlations, resulting in a scaling factor of 0.5. However, if longer (shorter) sentences tend to be followed by longer (shorter) sentences, the sentence length series is persistent, which indicates the presence of long-range correlations in the series and, as a result, the scaling factor is > 0.5. Conversely, if there is a higher likelihood of longer (shorter) sentences to follow shorter (longer) sentences, the series is anti-persistent. In this case, the scaling factor is < 0.5.

In the first step of the MFDFA procedure, where the profile of the series is created, the series is converted into a random walk. According to the preceding discussion, we are interested to measure the growth rate of the standard deviation as the length of the series increases. This is why the profile of the series goes into the windowing procedure in step 2 in order to compare the standard variation of sub-sequences of different sizes with each other. In our experiments, we select the size of windows from the list 16, 24, 32, 48, 64, ..., up to a point where the series can still be segmented into at least 3 non-overlapping windows. This list is a realization of the sequence s_i:

$$s_i = \begin{cases} s_0 &= 16, \\ s_i &= s_{i-1} + 2^{\lfloor \log(s_{i-1}) - 1 \rfloor}, \quad \text{for} \quad i \geq 1 \quad \text{and} \quad s_i \leq \lfloor N/3 \rfloor \end{cases} \tag{2.5}$$

2.3.2 Detrending

In step 3, the sequence is first detrended by subtracting the best linear fit in each window before calculating the amount of dispersion in that window. A trend is an imposed changes to a system that is not raised from the intrinsic properties of the system but rather external factors. This undesired variation may affect statistics of observations, such as variance.

It is not trivial to determine the source of trends in a complex system. Nevertheless, we can speculatively formulate some guesses to convey our intended meaning. For example, if some specific structure is commonly used in a specific genre, variations in distributional text properties may not be the result of the author's choice, but rather dictated by requirements of the genre. As another example, suppose an author decides which discourse modes to use and how to switch between them. Although the author has control over how to utilize discourse modes, nevertheless, alterations to distributional text properties are not completely within the author's control as

distributions of syntactic word classes necessarily differ for various discourse modes. Our interest is in the authors' preferences of textual structures during the process of text composition. Detrending eliminates the effect of artifacts that are beyond the author's control.

In step 3, what remains after detrending are residuals, which are entered into the computation instead of the initial values. However, it is expected that if the data exhibits no discernible trend, the mean square fluctuations of the initial profile values and those of the residuals are closely similar. Note that it is possible to detrend the sequence using polynomial fits, in case a system is under influence of more complex trends.

The mean square fluctuations, which are calculated in step 3, resemble the mean variance of windows with specific size. Recall from the outset of our discussion that our goal is to determine the growth rate of fluctuations, which is directly proportional to the number of steps in a random walk. That is why, after creation of the profile of the series and segmentation of it into windows the mean square fluctuations for windows of different sizes are calculated and compared in the following steps.

2.3.3 Multiple Scaling Factors and Multifractality

The difference between DFA and MFDFA lies in steps 4 and 5. In DFA as proposed by Peng et al. (1994) only one growth factor is calculated for a time series. However, the long-range correlation paradigm of a time series may be too complex to be explained by a single value. Kantelhardt et al. (2002) thus proposed MFDFA as an extension to DFA in order to capture various scaling patterns by calculating the qth order of the mean square fluctuations. Let us focus first on $q = 2$. This setting turns Eq. 2.4 to a form, which resembles the mean standard deviation for windows with size s. This is in fact the scaling factor, which is computed by DFA. The resulting exponent in step 5, $h(2)$, is an important measure in fractal analysis. It equals to the Hurst Exponent (Mandelbrot and Wallis, 1968) for stationary series. In the present study, we refer to $h(2)$ as the *degree of fractality* and represent it with $\mathcal{D}$.

By changing the value of q, the procedure puts emphasis on larger or smaller fluctuations. If $q > 2$, it accentuates larger values, while if $q < 2$, it emphasizes smaller variations. Given q, the scaling factor, $h(q)$, is calculated by fitting a line to the log-log plot of $F_q(s)$ in step 5. If the series is multifractal, fluctuations are heterogeneous and $h(q)$ varies depending on its parameter, q. However, if fluctuations are homogeneous, changing the value of q results in no difference in

scaling factors (for more discussion, see, Roeske et al., 2018). In our experiments, we change the value of q from -5 to 5 in steps of 0.25.

2.3.4 Singularity Spectrum

Once scaling factors are computed, we need a way to represent the multifractality of a series. The singularity spectrum, $f(\alpha)$, summarizes multifractality information of a series effectively and lends itself to an elegant visualization. It is computed as:

$$\begin{aligned} \alpha &= h(q) + qh'(q) \\ f(\alpha) &= q[\alpha - h(q)] + 1 \end{aligned} \tag{2.6}$$

The width of the singularity spectrum, $\mathcal{D} = \alpha_{max} - \alpha_{min}$, shows the *degree of multifractality* of a series (for more technical details, cf., Kantelhardt et al., 2002). α_{min} and α_{max} denote the leftmost side and the rightmost side of $f(\alpha)$, respectively. If $h(q), \forall q$ are in close proximity to each other, the singularity spectrum is narrow and the series is regarded to be monofractal. In the case of multifractal series, the width of the singularity spectrum, $\mathcal{D}$, expands.

2.3.5 Fractal Asymmetry

The multifractality of series may incline more toward either small or large quantities, leading to a skewness in the singularity spectrum, which can be measured by *fractal asymmetry* (Drozdz and Swiecimka, 2015):

$$\mathcal{A} = \frac{\Delta\alpha_L - \Delta\alpha_R}{\Delta\alpha_L + \Delta\alpha_R} \tag{2.7}$$

$\Delta\alpha_L = \alpha_0 - \alpha_{min}$ and $\Delta\alpha_R = \alpha_{max} - \alpha_0$ are the width of the right and the left side of the singularity spectrum curve, respectively and α_0, corresponding to $q = 0$, usually points to the peak of the $f(\alpha)$ curve.

2.4 Examples

To provide a visual perspective on the MFDFA procedure and to effectively illustrate fractality concepts, we selected three sample texts: (a) *Puck of Pooks Hill* by *Rudyard Kipling* published in 1901, (b) *The Golden Bowl* by *Henry James* published in 1904, and (c) *The Science and Philosophy of the Organism* by *Hans Driesch* published in 1908. The first two texts belong to the category of canonical texts and the latter is a non-fictional text from the JEFP corpus.

Figure 2.2a,b,c show the series of sentence lengths and their profile for the three texts. We applied MFDFA to the series to analyze fractal features of the texts. Based on

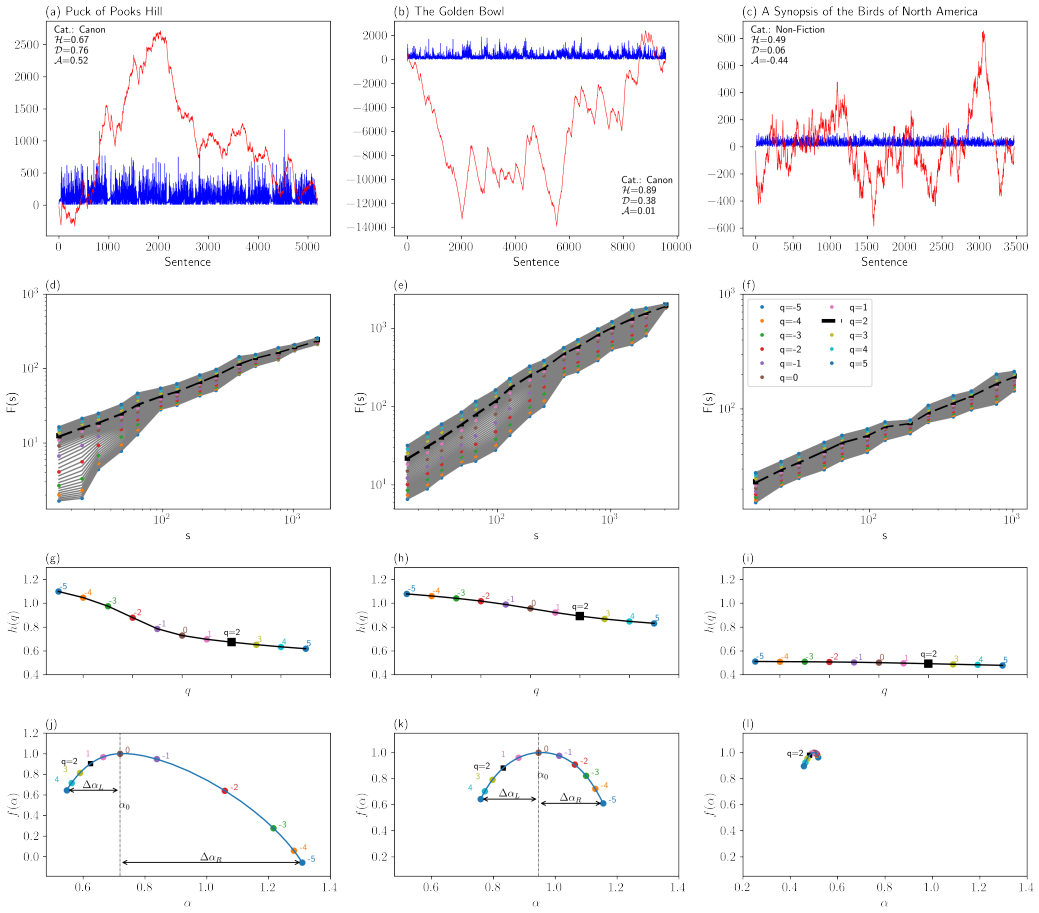

Figure 2.2: Visualization of fractal analysis of three sample texts using MFDFA. Top row: sentence length series (blue) and their profiles (red; cumulative sum of mean centered series; step 1 of the MFDFA procedure) of (a) *Puck of Pooks Hill* by *Rudyard Kipling*, (b) *The Golden Bowl* by *Henry James*, and (c) *The Science and Philosophy of the Organism* by *Hans Driesch*. The series (a) and (b) have been scaled up by a factor of 10 to show more detail. The category (Cat.), the degree of fractality, $\mathcal{H}$, the degree of multifractality, $\mathcal{D}$, and the fractal asymmetry, $\mathcal{A}$, of the series are shown inside each panel. Second row: qth order of mean square fluctuations of the series for integer values of q (step 4 of MFDFA) after segmentation of the profile of series with the window size of s (step 2 of MFDFA) and detrending and computation of the mean square fluctuations of residuals (step 3 of MFDFA). Third row: scaling factor, $h(q)$, for different values of q (step 5 of MFDFA). Values of $h(q)$ varies depending on the value of q for the first two texts, indicating that they are multifractal. For the third text $h(q)$ is almost 0.5, showing that the series is not fractal. Last row: Visualization of the multifractality (Eq. 2.6) and asymmetry (Eq. 2.7) of the series. The width and the skewness of the singularity spectrum in Panel (i) shows that *Puck of Pooks Hill* has a higher degree of multifactality and asymmetry compared to *The Golden Bowl*. Panel (l) shows that the values of $f(\alpha)$ are clustered around one point for *The Science and Philosophy of the Organism*, which is not fractal. In all panels, wherever values correspond to $q = 2$, which is identical to the output of the DFA method, values are shown in bold face.

our discussion on the random walk (Section 2.3.1), if the profile of a series fluctuates abruptly (as in the case of Figure 2.2c), we can expect that the series has no strong long-range correlations and in more extreme cases it might be an anti-persistent series. Conversely, if the profile of a series changes more smoothly, we expect that it may exhibit long-range correlation patterns (as in the case of Figure 2.2a,b).

The profiles of the series go through step 2, the windowing procedure, and step 3 of MFDFA, the detrending and computation of the mean square fluctuations of residuals. In step 4 of the MFDFA procedure, the qth order of mean square fluctuations of the series, $F_q(s)$, are calculated, which are visualized in the second row of the plots, in Figure 2.2d,e,f. The plots only show the computation results for the integer values of q. The curve with bold face represents $F_q(s)$ for $q = 2$, which is used to calculate the degree of fractality, $\mathcal{H}$, of the series and is equivalent to the output of the DFA procedure, if it had been used.

By applying the last step of MFDFA to the mean square fluctuation, the scaling factors, $h(q)$ are obtained (Figure 2.2g,h,i). If a series is monofractal, the curve of $h(q)$, has a slop of zero. For multifractal series values of $h(q)$ differ as its parameter, q, changes. Figure 2.2g,h indicate that the sentence length series of both canonical texts, *Puck of Pooks Hill* and *The Golden Bowl* are highly multifractal, while that of the non-fictional text, *The Science and Philosophy of the Organism* is not multifractal and not even fractal as its degree of fractality is almost 0.5.

Using Eq. 2.6 (Section 2.3.4) we calculate the singularity spectrum, $f(\alpha)$, of the series, which is visualized in Figure 2.2j,k,l for the three texts. The width of $f(\alpha)$ indicates the degree of multifractality, $\mathcal{D}$, of the series. Although from the $h(q)$ curves (Figure 2.2g,h) we realized that the first two sample texts are multifractal, a comparison of the two plots (j) and (k) of Figure 2.2 reveals that *Puck of Pooks Hill* is more multifractal than *The Golden Bowl* with the value of $\mathcal{D}$ =0.76 compared to $\mathcal{D} = 0.38$, respectively. Moreover, The singularity spectrum of the former is highly asymmetrical, while the singularity spectrum of the latter is (almost) symmetrical. By applying Eq. 2.7 (Section 2.3.5) to the plots of $f(\alpha)$, the fractal asymmetry values are computed as 0.52 and 0.01 for the two texts, respectively. As the sentence length series of *The Science and Philosophy of the Organism* is not fractal, the values of $f(\alpha)$ in Figure 2.2l are clustered around one point.

2.5 Unpredictability and Surprise

Any discussion of unpredictability inevitably leads to the concept of entropy. Entropy has been used in visual aesthetics to model preference in artwork and non-art images (see, e.g., Redies et al., 2017; Grebenkina et al., 2018; Stanischewski et al., 2020). In the language domain entropy-based metrics have been used for various purposes, such as analyzing word order (Montemurro and Zanette, 2011; Futrell et al., 2015; Montemurro and Zanette, 2016; Koplenig et al., 2017; Levshina, 2019), ordering preference and complexity of languages (Bentz et al., 2015; Ehret and Szmrecsanyi, 2016; Hernández-Gómez et al., 2017; Bentz et al., 2017) and language acquisition (Fedzechkina et al., 2017). Besides the references mentioned here, we have also discussed studies, which used entropy metrics to analyze preference in the text domain in our previous publications (Chapters 4 and 5 / Mohseni et al., 2022; Mohseni et al., 2023).

By applying entropy metrics in our study, we were interested in measuring the amount of unpredictability and surprise in structures of texts. A higher degree of predictability in a text, which can be indicative of repetition, facilitates text processing. However, it can be associated with a higher level of monotonicity in text, which can potentially lead to an unengaging reading experience. Hence, an author needs to establish a balance between monotony and surprise in order to benefit the reader in his reading experience (for a discussion on this, look at, Chapter 4 / Mohseni et al., 2022, and references there). Unpredictability can be measured globally or in local distributions across the sequence of textual structures. Accordingly we used two entropy metrics in our experiments: Shannon Entropy, which is a global measure of unpredictability, and Approximate Entropy (Pincus, 1991), which measures the amount of surprise in local structures.

2.5.1 Shannon Entropy

Shannon Entropy (ShEn) measures the amount of uncertainty or unpredictability in a random variable. It was originally proposed by Claude Elwood Shannon to determine the amount of information that is transmitted over a communication channel (Shannon, 1948). Shannon defined entropy as the average number of bits required to encode messages, which are generated by an information source.

Given a discrete random variable x and a probability distribution $p(x)$, the ShEn of x, $h(x)$, is computed as

$$h(x) = -\sum_{x \in S_x} p(x) \log_e p(x) \tag{2.8}$$

where S_x is the set of all possible events. Note that the base of logarithm in Eq. 2.8 has no effect on the interpretation of results. Accordingly, we used e, Euler's number, as the base of logarithm, similar to the default base value in most well-known programming libraries. The more unpredictable a system, the higher the ShEn value of observables. ShEn takes its maximum value in a system with a uniform distribution, where all possible events likely to happen equally and, thus, uncertainty is at a maximum.

ShEn is a global measurement of unpredictability and does not reflect local patterns of distribution. As a result, when it is applied to a sequence of text property values, no order is taken into account and no information about local structures is captured.

2.5.2 Approximate Entropy

As mentioned above, ShEn does not capture local patterns of distribution. Moreover, it is only applicable to discrete random variables and not to continuous ones. Pincus (1991) proposed Approximate Entropy (ApEn) to overcome these limitations and to imitate a way of measuring irregularity and uncertainty in time series. Like ShEn, a high ApEn value indicates a low level of predictability, whereas a low ApEn value suggests a higher level of predictability.

Given a time series of $X = x(1), \cdots, x(n)$ and a predefined value m, which is the length of sub-sequences $y_i^m = [x(i), \cdots, x(i + (m - 1))]$ and a tolerance level, r, for computing distances between sub-sequences, Approximate Entropy (ApEn) is computed as follows:

1. Using Chebyshev distance between sub-sequences y_i^m and y_j^m, defined as $d_{i,j}^m = \max_k |y_i^m(k) - y_j^m(k)|$, compute

$$C_i^m(r) = \frac{1}{n - m + 1} \sum_{j=1}^{n-m+1} \mathbb{1}(r - d_{i,j}^m) \tag{2.9}$$

   in which $\mathbb{1}(.)$ is the Heaviside function, whose value is zero for negative arguments and one for positive arguments.

2. Compute

$$\phi^m(r) = \frac{1}{n - m + 1} \sum_{i=1}^{n-m+1} \log(C_i^m(r)) \tag{2.10}$$

3. Repeat step 1 to 3 for sub-sequences of length $m + 1$ to compute $\phi^{m+1}(r)$.

4. Finally, calculate ApEn as

$$\texttt{ApEn}(m, r) = \phi^m(r) - \phi^{m+1}(r) \tag{2.11}$$

Computation of ApEn, as above procedure shows, is based on measuring distances between sub-sequences. Two sub-sequences are regarded similar, i.e. their difference is "tolerated", if the Chebyshev distance between them lies within the tolerance level r (Eq. 2.9). ApEn is thereafter computed by comparing sub-sequence matches of length m to those of length $m + 1$ (Eq. 2.11). If a time series has high levels of variation, longer sub-sequences are less likely to resemble each other. As a result, the time series will have a higher ApEn value, indicating its low predictability. This is where ApEn captures information about local distribution patterns of the sequence.

Unlike ShEn, which advantageously requires no parameter setting, ApEn has two parameters to be set: the length of sub-sequences, m, and the tolerance level, r. Typically, these parameters are set to 2 and 20% of the standard deviation, respectively (see, for example, Li et al., 2008; Hayashi et al., 2012; Lee et al., 2013). The theoretical minimum value for ApEn is 0. For a time series fixed at a specific value, ApEn is always 0, regardless of the parameter setting.

2.5.3 Examples

We compare ShEn and ApEn of the distribution of sentence lengths for two texts in Figure 2.3. The selected sample texts are *My Antonia* by *Willa Cather* and *Ailsa Paige* by *Robert William Chambers* initially published in 1918 and 1910, respectively. The first text is a canonical fictional prose and the second one is a non-canonical fiction. ShEn of both texts is close to each other. This shows that both texts have similar degree of unpredictability in terms of global distribution. However, the higher value of ApEn for *My Antonia* indicates that this canonical prose is less predictable in its sequential organization compared to the non-canonical text, *Alisa Paige*.

2.6 Publication of Code

The python code that we used for fractality analysis and calculating entropy features are published in two Github repositories, `https://github.com/mohsenim/Multifractality` and `https://github.com/mohsenim/Surprise`.

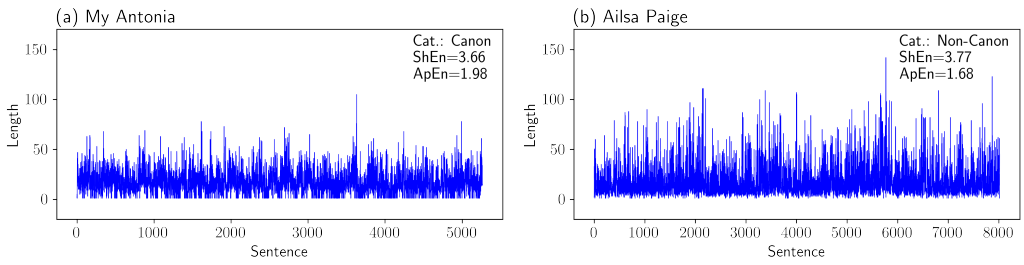

Figure 2.3: Sentence length series of (a) *My Antonia* by *Willa Cather* and (b) *Ailsa Paige* by *Robert William Chambers*. The category (Cat.), Shannon Entropy, ShEn, and Approximate Entropy, ApEn, of each text are shown inside each panel.

References

Arnheim, Rudolf (1974). *Art and Visual Perception: A Psychology of the Creative Eye*. Berkeley: University of California Press.

Bentz, Christian, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i-Cancho (2017). "The Entropy of Words—Learnability and Expressivity across More than 1000 Languages". In: *Entropy* 19.6. DOI: 10.3390/e19060275.

Bentz, Christian, Annemarie Verkerk, Douwe Kiela, Felix Hill, and Paula Buttery (2015). "Adaptive Communication: Languages with More Non-Native Speakers Tend to Have Fewer Word Forms". In: *PLoS ONE*, p. 0128254. DOI: 10.1371/journal.pone.0128254.

Brachmann, Anselm, Erhardt Barth, and Christoph Redies (2017). "Using CNN features to better understand what makes visual artworks special". In: *Frontiers in Psychology* 8, p. 830. DOI: 10.3389/fpsyg.2017.00830.

Chatterjee, Anjan and Oshin Vartanian (2014). "Neuroaesthetics". In: *Trends in Cognitive Sciences* 18.7, pp. 370–375. DOI: 10.1016/j.tics.2014.03.003.

Drozdz, Stanislaw and Pawel Swiecimka (2015). "Detecting and interpreting distortions in hierarchical organization of complex time series". In: *Physical Review E* 91.3, p. 030902. DOI: 10.1103/physreve.91.030902.

Ehret, Katharina and Benedikt Szmrecsanyi (2016). "An Information-Theoretic Approach to Assess Linguistic Complexity". In: *Complexity, Isolation, and Variation*. Ed. by Raffaela Baechler and Guido Seiler. De Gruyter, pp. 71–94. DOI: 10.1515/9783110348965-004.

Fechner, Gustav Theodor (1876). *Vorschule der Ästhetik*. Leipzig: Breitkopf and Härtel.

Fedzechkina, Maryia, Elissa L. Newport, and T. Florian Jaeger (2017). "Balancing Effort and Information Transmission During Language Acquisition: Evidence From Word Order and Case Marking". In: *Cognitive Science* 41.2, pp. 416–446. DOI: 10.1111/cogs.12346.

Futrell, Richard, Kyle Mahowald, and Edward Gibson (2015). "Quantifying Word Order Freedom in Dependency Corpora". In: *Proceedings of the Third International Conference on Dependency Linguistics*. Uppsala, Sweden: Uppsala University, Uppsala, Sweden, pp. 91–100.

Geller, Hannah Alexa, Ralf Bartho, Katja Thömmes, and Christoph Redies (2022). "Statistical image properties predict aesthetic ratings in abstract paintings created by neural style transfer". In: *Frontiers in Neuroscience* 16, p. 999720. DOI: 10.3389/fnins.2022.999720.

Grebenkina, Maria, Anselm Brachmann, Marco Bertamini, Ali Kaduhm, and Christoph Redies (2018). "Edge-Orientation Entropy Predicts Preference for Diverse Types of Man-Made Images". In: *Frontiers in Neuroscience* 12, p. 678. DOI: 10.3389/fnins.2018.00678.

Hayashi, Kazuko, Kenji Shigemi, and Teiji Sawa (2012). "Neonatal Electroencephalography Shows Low Sensitivity to Anesthesia". In: *Neuroscience Letters* 517.2, pp. 87–91. DOI: 10.1016/j.neulet.2012.04.028.

Hernández-Gómez, Candelario, Rogelio Basurto-Flores, Bibiana Obregón-Quintana, and Lev Guzmán-Vargas (2017). "Evaluating the Irregularity of Natural Languages". In: *Entropy* 19, p. 521. DOI: 10.3390/e19100521.

Jacobs, Arthur M. (2015). "Neurocognitive poetics: methods and models for investigating the neuronal and cognitive-affective bases of literature reception". In: *Frontiers in Human Neuroscience* 9, p. 186. DOI: 10.3389/fnhum.2015.00186.

Kantelhardt, Jan W., Stephan A. Zschiegner, Eva Koscielny-Bunde, Shlomo Havlin, Armin Bunde, and H.Eugene Stanley (2002). "Multifractal detrended fluctuation analysis of nonstationary time series". In: *Physica A: Statistical Mechanics and Its Applications* 316.1, pp. 87–114. DOI: 10.1016/S0378-4371(02)01383-3.

Koplenig, Alexander, Peter Meyer, Sascha Wolfer, and Carolin Müller-Spitzer (2017). "The Statistical Trade-off Between Word Order and Word Structure – Large-Scale Evidence for the Principle of Least Effort". In: *PLoS ONE* 12.3, pp. 1–25. DOI: 10.1371/journal.pone.0173614.

Lee, Gerick, Sara Fattinger, Anne-Laure Mouthon, Quentin Noirhomme, and Reto Huber (2013). "Electroencephalogram Approximate Entropy Influenced by Both Age and Sleep". In: *Frontiers in Neuroinformatics* 7, p. 33. DOI: 10.3389/fninf.2013.00033.

Levshina, Natalia (2019). "Token-based typology and word order entropy: A study based on Universal Dependencies". In: *Linguistic Typology* 23.3, pp. 533–572. DOI: 10.1515/lingty-2019-0025.

Li, Xiaoli, Suyuan Cui, and Logan Voss (2008). "Using Permutation Entropy to Measure the Electroencephalographic Effects of Sevoflurane". In: *Anesthesiology* 109, pp. 448–56. DOI: 10.1097/ALN.0b013e318182a91b.

Mandelbrot, Benoit B. and James R. Wallis (1968). "Noah, Joseph, and Operational Hydrology". In: *Water Resources Research* 4.5, pp. 909–918. DOI: 10.1029/WR004i005p00909.

Mohseni, Mahdi, Volker Gast, and Christoph Redies (2021). "Fractality and Variability in Canonical and Non-Canonical English Fiction and in Non-Fictional Texts". In: *Frontiers in Psychology* 12, p. 920. DOI: 10.3389/fpsyg.2021.599063.

Mohseni, Mahdi, Christoph Redies, and Volker Gast (2022). "Approximate Entropy in Canonical and Non-Canonical Fiction". In: *Entropy* 24.2, p. 277. DOI: 10.3390/e24020278.

– (2023). "Comparative Analysis of Preference in Contemporary and Earlier Texts Using Entropy Measures". In: *Entropy* 25.3, p. 486. DOI: 10.3390/e25030486.

Montemurro, Marcelo A. and Damián H. Zanette (2011). "Universal Entropy of Word Ordering Across Linguistic Families". In: *PLoS ONE* 6.5, pp. 1–9. DOI: 10.1371/journal.pone.0019875.

Montemurro, Marcelo A. and Damián H. Zanette (2016). "Complexity and Universality in the Long-Range Order of Words". In: *Creativity and Universality in Language*. Cham: Springer, pp. 27–41. DOI: 10.1007/978-3-319-24403-7_3.

Oświecimka, Paweł, Jarosław Kwapień, and Stanislaw Drozdz (2006). "Wavelet versus Detrended Fluctuation Analysis of multifractal structures". In: *Physical review. E, Statistical, nonlinear, and soft matter physics* 74, p. 016103. DOI: 10.1103/PhysRevE.74.016103.

Pearson, Karl (1905). "The Problem of the Random Walk". In: *Nature* 72.1865, pp. 294–294. DOI: 10.1038/072294b0.

Peng, C.-K., S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger (1994). "Mosaic organization of DNA nucleotides". In: *Physical Review E* 49.2, pp. 1685–1689. DOI: 10.1103/physreve.49.1685.

Pincus, Steven M (1991). "Approximate Entropy as a Measure of System Complexity". In: *Proceedings of the National Academy of Sciences* 88.6, pp. 2297–2301. DOI: 10.1073/pnas.88.6.229.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher Manning (2020). "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, pp. 101–108. DOI: 10.18653/v1/2020.acl-demos.14.

Redies, Christoph (2015). "Combining universal beauty and cultural context in a unifying model of visual aesthetic experience". In: *Frontiers in Human Neuroscience* 9, p. 218. DOI: 10.3389/fnhum.2015.00218.

Redies, Christoph, Anselm Brachmann, and Johan Wagemans (2017). "High Entropy of Edge Orientations Characterizes Visual Artworks From Diverse Cultural Backgrounds". In: *Vision Research* 133, pp. 130–144. DOI: 10.1016/j.visres.2017.02.004.

Roeske, Tina C., Damian Kelty-Stephen, and Sebastian Wallot (2018). "Multifractal analysis reveals music-like dynamic structure in songbird rhythms". In: *Scientific Reports* 8.1. DOI: 10.1038/s41598-018-22933-2.

Shannon, Claude Elwood (1948). "A Mathematical Theory of Communication". In: *Bell System Technical Journal* 27.3, pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

Stanischewski, Sarah, Carolin S. Altmann, Anselm Brachmann, and Christoph Redies (2020). "Aesthetic Perception of Line Patterns: Effect of Edge-Orientation Entropy and Curvilinear Shape". In: *I-Perception* 11.5, p. 2041669520950749. DOI: 10.1177/2041669520950749.

Taylor, Richard, Branka Spehar, Caroline Hagerhall, and Paul Van Donkelaar (2011). "Perceptual and physiological responses to Jackson Pollock's fractals". In: *Frontiers in Human Neuroscience* 5, p. 60. DOI: 10.3389/fnhum.2011.00060.

Chapter 3

FRACTALITY AND VARIABILITY

In our exploration for finding textual correlates of preference we first operationalized preference in terms of canonization. Canonical texts are cultural artifacts which have been regarded as important written works by influential groups inside a culture or society for a long period of time. It is expected that educated members of a society are familiar with canonical texts as they are included in syllabuses of schools and universities. Non-canonical texts are conversely less prestigious texts, which have never obtained a comparable recognition. We analyzed structural features of texts to distinguish the two categories of canonical and non-canonical texts. We also analyzed non-fictional texts in comparison to the two fictional text categories to determine inter- and intra-genre similarities and differences.

Our hypothesis was that global structural design features of the three text categories, i.e. fictional/canonical, fictional/non-canonical and non-fictional texts, are different from each other. We distinguished between lower- and higher-level text properties. The former refer to processing at the level of linguistic decoding and the latter occurs at higher levels of comprehension. We measured two types of text properties as correlates of the lower-level processing: sentence length and frequency of four major POS-tags, i.e. Noun, Verb, Adjective and Adverb, per sentence. We also measured two properties that reflects processing at the higher-level: lexical diversity and topic distribution of text chunks along the text. Using these four types of observables each text were represented in seven series, which were the sequence of textual property values across the text.

For our experiments we compiled our own corpus called the "Jena Corpus of Expository and Fictional Prose (JEFP Corpus)", version 1.0. The corpus contains three text categories: fictional/canonical, fictional/non-canonical and non-fictional texts. Canonical texts, which were written in the 19^{th} and early 20^{th} centuries, were selected from the Corpus of Canonical Western Literature (Green, 2017). Non-canonical fictional texts and non-fictional texts, which were mostly published for the first time between the late 19^{th} and the early 20^{th} centuries, were selected from different e-book publishing websites.

Inspired from research in visual aesthetics, in which variability and fractal analyses

were successfully applied to characterize different categories of images and artworks (Redies and Brachmann, 2017), we analyzed variation and fractal patterns of texts in our corpus. We computed variance of series as our metric of variability (see, Section 2.2). For fractal analysis we used Multi-Fractal Detrended Fluctuation Analysis (MFDFA; Kantelhardt et al., 2002) to compute the degree of fractality, the degree of multifractality and the degree of asymmetry for each text property series (see, Section 2.3). Variance and these three fractal statistics were used to analyze texts in the JEFP corpus.

We compared fictional with non-fictional texts and within the fictional category canonical with non-canonical texts. Our analysis revealed that, generally speaking the lower-level text properties differ for the various text categories more than higher-level text properties. Moreover, variance is more distinctive between the text categories compared to the fractal features. Statistical analysis of variance showed that canonical texts are surprisingly more similar to non-fictional texts than non-canonical fictional texts. However, long-range correlation patterns of the two fictional categories derived from fractal analysis are more similar in comparison to non-fictional texts. In general, non-fictional texts exhibit stronger long-range correlations.

We also conducted classification experiments to determine how effectively variance and the fractal features can separate the text categories. In accordance with our statistical analyses, lower-level properties have more discriminatory power than higher-level properties especially when features are combined together in the classification model. As expected, the results showed that classification of fictional from non-fictional texts is simpler than separation of canonical/fictional from non-canonical/fictional texts using both types of features, i.e. variance and the fractal features. Generally, variability analysis distinguishes classes from each other better than fractal analysis.

Although our study was originally inspired by studies in the field of visual aesthetics, we should pay attention to differences between the two sensory domains. Apparently the most important dissimilarity between image and text perception is that reading is a temporal and slow process as opposed to visual processing which is (almost) instant. Therefore, predictability and expectation of reader can be an important factor to be investigated in text aesthetic analysis. We will discuss this issue in the following chapters.

References

Green, Clarence (2017). "Introducing the Corpus of the Canon of Western Literature: A Corpus for Culturomics and Stylistics". In: *Language and Literature* 26.4, pp. 282–299. DOI: 10.1177/0963947017718996.

Kantelhardt, Jan W., Stephan A. Zschiegner, Eva Koscielny-Bunde, Shlomo Havlin, Armin Bunde, and H.Eugene Stanley (2002). "Multifractal detrended fluctuation analysis of nonstationary time series". In: *Physica A: Statistical Mechanics and Its Applications* 316.1, pp. 87–114. DOI: 10.1016/S0378-4371(02)01383-3.

Redies, Christoph and Anselm Brachmann (2017). "Statistical Image Properties in Large Subsets of Traditional Art, Bad Art, and Abstract Art". In: *Frontiers in Neuroscience* 11, p. 593. DOI: 10.3389/fnins.2017.00593.

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 31 March 2021
doi: 10.3389/fpsyg.2021.599063

Fractality and Variability in Canonical and Non-Canonical English Fiction and in Non-Fictional Texts

Mahdi Mohseni[1,2], *Volker Gast*[2] *and Christoph Redies*[1]*

[1] *Experimental Aesthetics Group, Institute of Anatomy I, Jena University Hospital, University of Jena, Jena, Germany,*
[2] *Department of English and American Studies, University of Jena, Jena, Germany*

This study investigates global properties of three categories of English text: canonical fiction, non-canonical fiction, and non-fictional texts. The central hypothesis of the study is that there are systematic differences with respect to structural design features between canonical and non-canonical fiction, and between fictional and non-fictional texts. To investigate these differences, we compiled a corpus containing texts of the three categories of interest, the Jena Corpus of Expository and Fictional Prose (JEFP Corpus). Two aspects of global structure are investigated, variability and self-similar (fractal) patterns, which reflect long-range correlations along texts. We use four types of basic observations, (i) the frequency of POS-tags per sentence, (ii) sentence length, (iii) lexical diversity, and (iv) the distribution of topic probabilities in segments of texts. These basic observations are grouped into two more general categories, (a) the lower-level properties (i) and (ii), which are observed at the level of the sentence (reflecting linguistic decoding), and (b) the higher-level properties (iii) and (iv), which are observed at the textual level (reflecting comprehension/integration). The observations for each property are transformed into series, which are analyzed in terms of variance and subjected to Multi-Fractal Detrended Fluctuation Analysis (MFDFA), giving rise to three statistics: (i) the degree of fractality ($\mathcal{H}$), (ii) the degree of multifractality ($\mathcal{D}$), i.e., the width of the fractal spectrum, and (iii) the degree of asymmetry ($\mathcal{A}$) of the fractal spectrum. The statistics thus obtained are compared individually across text categories and jointly fed into a classification model (Support Vector Machine). Our results show that there are in fact differences between the three text categories of interest. In general, lower-level text properties are better discriminators than higher-level text properties. Canonical fictional texts differ from non-canonical ones primarily in terms of variability in lower-level text properties. Fractality seems to be a universal feature of text, slightly more pronounced in non-fictional than in fictional texts. On the basis of our results obtained on the basis of corpus data we point out some avenues for future research leading toward a more comprehensive analysis of textual aesthetics, e.g., using experimental methodologies.

Keywords: fractality, self-similarity, multifractal DFA, variability, POS tagging, sentence length, lexical diversity, topic modeling

OPEN ACCESS

Edited by:
Sascha Schroeder,
University of Göttingen, Germany

Reviewed by:
Sebastian Wallot,
Max Planck Society (MPG), Germany
Jana Lüdtke,
Freie Universität Berlin, Germany

***Correspondence:**
Christoph Redies
christoph.redies@med.uni-jena.de

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *26 August 2020*
Accepted: *02 March 2021*
Published: *31 March 2021*

Citation:
Mohseni M, Gast V and Redies C (2021) Fractality and Variability in Canonical and Non-Canonical English Fiction and in Non-Fictional Texts. Front. Psychol. 12:599063. doi: 10.3389/fpsyg.2021.599063

1. INTRODUCTION

Canonical fiction comprises works "which are accepted as legitimate by the dominant circles within a culture and whose conspicuous products are preserved by the community to become part of its historical heritage" (Even-Zohar, 1990, p. 15); they are regarded as "repositories of cultural values" (Guillory, 1987, p. 487). The relevant texts have high prestige ("classics," "high literature") and are often integrated into the school curriculum, so that large parts of a society are familiar with them. In this study we investigate whether English canonical and non-canonical texts from the 19th and early 20th centuries differ in terms of global structural design features. In order to locate the two text categories in the larger space of genres, we moreover compare fictional texts with non-fictional texts. The study is embedded within the field of empirical textual aesthetics insofar as the three text categories under analysis differ in terms of either the presence or absence of an aesthetic function (fictional vs. non-fictional texts), or preferences of societies as reflected in canonization. While canonization is a process driven by a range of social variables, such as "publication mechanisms (i.e., the sale of books, library use, etc.), politics, etc.," it is also based on "the text, its reading, readership, literary history, [and] criticism," i.e., the work of art itself in its cultural context (Tötösy de Zepetnek, 1994, p. 109, cf. also Underwood and Sellers, 2016; Koolen et al., 2020 for a discussion of the relationship between text-intrinsic and text-extrinsic factors in the process of canonization). The question arises whether there are any measurable differences between the text categories of interest. In this article we address this question by analyzing texts in terms of fractality and variability.

The question of objective, measurable correlates of readers' or societies' attitudes to texts has been raised in various contexts, more or less explicitly. The assumption that artistic composition can be measured is most obvious for poetry, with its interplay of meaning and form as manifested in rhythm and rhyme, and other aspects of poetic form, e.g., alliteration (cf. for instance Jakobson, 1960; Leech, 1969; Jacobs, 2015; Jacobs et al., 2016; Vaughan-Evans et al., 2016; König and Pfister, 2017; Menninghaus et al., 2017; Egan et al., 2020; Menninghaus and Wallot, 2021). Relevant studies of prose have mostly used summary statistics of properties extracted from text. Louwerse et al. (2008), one of the earliest relevant studies from the field of computational linguistics, distinguished literary from non-literary texts with distance measures derived from Latent Semantic Analyses and fed into a hierarchical clustering algorithm, and with frequency distributions of unigrams and bigrams. van Cranenburgh and Bod (2017) used frequency distributions of lexical and syntactic features to model human ratings of texts as more or less "literary" (see also Ashok et al., 2013; van Cranenburgh and Koolen, 2015 for similar approaches). In van Cranenburgh et al. (2019), summary statistics derived from topic modeling (Latent Dirichlet Allocation) and paragraph vectors are used to predict degrees of "literariness." Maharjan et al. (2017) explore a wide variety of features (including "readability") that can be used to classify texts in terms of "likability." Other standard methods of computational linguistics used in this context include sentiment and emotion analysis (Alm and Sproat, 2005; Francisco and Gervás, 2006; Kakkonen and Galić Kakkonen, 2011; Mohammad, 2011; Reagan et al., 2016; Maharjan et al., 2018). Global statistical properties such as complexity and entropy have been used to study the regularity (Mehri and Lashkari, 2016; Hernández-Gómez et al., 2017) and the quality of texts (Febres and Jaffe, 2017). Fractal analysis, which figures centrally in our study, has been applied to fictional texts as well (Drożdż and Oświęcimka, 2015; Mehri and Lashkari, 2016; Chatzigeorgiou et al., 2017), and fractal patterns have been observed in both Western (Drożdż et al., 2016) and Chinese literature (Yang et al., 2016; Chen and Liu, 2018). Cordeiro et al. (2015, p. 796) claim that "there is a fractal beauty in the text produced by humans" and "that its quality is directly proportional to the degree of self-similarity."

Our approach to studying structure in texts is inspired by relevant findings from vision, which we take as our starting point. In (cognitive) linguistics there is a widespread assumption that "linguistic structure is shaped by domain-general processes" (Diessel, 2019, p. 23), such as figure-ground segregation and processes of memory retrieval. In other words, linguistic processing is assumed to be based on the same type of brain activity as the processing of other types of sensory input. We therefore use methods that have been successfully applied in vision for the analysis of textual data. This transfer has obvious limitations though. Image data are three-dimensional—two-dimensional matrices with the luminance/color signals as the third dimension—whereas textual data are *prima facie* one-dimensional when regarded as strings of characters (though even silent reading implies prosody, adding a second dimension, cf. Gross et al., 2014). Related to this, the processing of propositional information is an incremental, "piecewise buildup of information, adding bits of information as the reader advances through the text" (Wallot et al., 2014, p. 1748; see also Verhuizen et al., 2019). Reading a text is thus a less immediate experience than contemplating a picture, in the sense that it requires more higher-level activity. Still, the higher-level activity of integrating new information into a "situation model" (Kintsch, 1988; McNamara and Magliano, 2009; Zwaan, 2016) is fed by lower-level processes of linguistic decoding (Cain et al., 2004; Tiffin-Richards and Schroeder, 2015, 2018)[1].

We use vision as our point of reference because there is a long-standing tradition of empirical research on aesthetic perception in this domain (Fechner, 1876; Arnheim, 1974; Chatterjee and Vartanian, 2014; Jacobs, 2015; Redies, 2015), and artworks have been studied in terms of structural properties (for

[1]The "classic" model—the LaBerge/Samuels model of automatic information processing in reading (cf. LaBerge and Samuels, 1974; Samuels, 1994)—assumes four components, (i) visual memory (VM), (ii) phonological memory (PM), (iii) semantic memory (SM), and (iv) episodic memory (EM). VM and PM are closely connected to sensory experience, i.e., visual and acoustic perception, and they are the input gates to processing in reading. Semantic memory is not only the place where "individual word meanings are produced," but also "where the comprehension of written messages occurs" (Samuels, 1994, p. 710). It is thus also responsible for the linguistic process of decoding, including the processing of morphology (word structure) and syntax (sentence structure). Episodic memory is the place where propositional information is stored, and it is "responsible for putting a time, place and context tag on events and knowledge" (Samuels, 1994, p. 710).

reviews, see Taylor et al., 2011; Brachmann and Redies, 2017). In this work, objective image properties were identified that differ between various categories of man-made images, such as traditional visual artworks, other visually preferred images and different types of non-preferred images. A particular focus has been on global properties of preferred stimuli. In contrast to local image properties, such as luminance contrast or color at a given location in an image, global image properties reflect summary statistics of pictorial elements or their relations to each other across an image (Brachmann and Redies, 2017). Global statistical image properties seem particularly suitable for studying visual preferences because aesthetic concepts, such as "balanced composition" (McManus et al., 1985), "good Gestalt" (Arnheim, 1974), or "visual rightness" (Locher et al., 1999) all refer to global image structure (Redies et al., 2017). Examples of global properties that characterize preferred visual stimuli are a scale-invariant (fractal) image structure (Taylor et al., 2011), statistical regularities in the Fourier domain (Graham and Field, 2007; Redies et al., 2007), curved shape (Bar and Neta, 2006; Bertamini et al., 2016), regularities in edge orientation distribution (Redies et al., 2012, 2017), and specific color features (Palmer et al., 2013; Nascimento et al., 2017). Moreover, traditional visual artworks were found to exhibit a high richness and high variability of low-level features of a Convolutional Neural Network (CNN; Brachmann et al., 2017).

Given the time-distributed nature of information processing in reading, aesthetic experience is hard to measure experimentally in this domain (see e.g., Cook and Wei, 2019 for discussion). Studies obtaining real-time measurements (such as reading times) generally investigate smaller windows of text (e.g., O'Brien et al., 2013; Wallot et al., 2014; Blohm et al., 2021; Menninghaus and Wallot, 2021)[2]. The methods used in vision research can thus not easily be transferred to the study of aesthetic experience in reading. In this study we therefore pursue an observational, rather than experimental approach, investigating properties of texts which are classified along the dimensions fictional/non-fictional and (within the fictional texts) canonical/non-canonical.

As a first step, we need to identify measurable properties that differentiate fictional from non-fictional texts, and canonical from non-canonical fictional texts. Moreover, we need to test and validate statistical methods to describe global structural patterns in the distribution of these properties in texts. We will use two text properties that we regard as being relevant to (linguistic) decoding, measurements derived from part-of-speech tags and sentence length, and two properties that we regard as correlates of higher-level comprehension processes, lexical diversity and topic probabilities. For each of these properties, which are represented as series, we determine four statistics reflecting variability and fractality, the most important determinants distinguishing visual artworks of different categories (Redies and Brachmann, 2017)[3]. For our quantitative analysis we have compiled a corpus of fictional and non-fictional texts, the Jena Corpus of Expository and Fictional Prose, JEFP Corpus for short. The fictional texts of this corpus are classified into canonical and non-canonical ones (see section 4 for details). Obviously, our observational approach does not allow us to reach any conclusions concerning cognitive processes during reading (aesthetic experience, e.g., aesthetic emotions as described by Menninghaus et al., 2019), and we abstract away from the role of the reader (see Iser, 1976 for a foundational study of aesthetic responses during reading, and recent empirical studies of the type carried out by Blohm et al., 2021; Menninghaus and Wallot, 2021). We therefore also disregard phonological aspects of texts, which are no doubt important for aesthetic textual perception. Our study is intended to provide the basis for experimental investigations in the future by identifying textual properties, and global patterns in the distribution of such properties, that vary across the text types distinguished in this study.

The article is organized as follows: We start by providing a list of measurable text properties that may contribute to differences between the three categories of texts in the corpus (section 2). Based on these properties, series are derived from the various texts. We then proceed to introduce statistical methods that capture variability and fractal patterns, most importantly Multi-Fractal Detrending Fluctuation Analysis (MFDFA, section 3). The Jena Corpus of Expository and Fictional Prose (JEFP Corpus) is described in section 4. In section 5, we provide the results of individual features relative to the three text categories and we show how well they can distinguish between the categories by feeding them into a binary classifier (Support Vector Machine). In section 6, we discuss the implications of our preliminary findings and outline avenues for future research.

2. MEASURABLE PROPERTIES OF TEXT

The central hypothesis of this study is that texts of the categories fictional/canonical, fictional/non-canonical, and non-fictional differ in terms of measurable structural properties. Such properties can be derived from various types of measurements. While we are ultimately interested in global properties of texts, the basic units of observations are located at different levels of processing. As mentioned in section 1, we distinguish two levels of processing. The lower level of processing concerns the task of linguistic decoding, which is largely automatic and resorts to implicit knowledge. The higher level of processing concerns the integration of propositional information into explicit memory (comprehension).

While the lower-level processes of reading have been studied experimentally in psychological, psycholinguistic and neurolinguistic research, e.g., with eye-tracking and event-related potential measurements (e.g., Kliegl et al., 2004, 2012), comprehension has been studied most extensively in the field of the psychology of learning, specifically in text assessment (e.g., Graesser et al., 2003; McNamara et al., 2013). The Coh-Metrix

[2]An exception is provided by McNerney et al., 2011, who had participants read a 361 pages long novel. For longer texts, human ratings have also been used as behavioral correlates of text structure (e.g., van Cranenburgh and Bod, 2017; van Cranenburgh et al., 2019). A methodological toolbox for measuring reading experience has been proposed by Knoop et al. (2016) and Thissen et al. (2018).

[3]Fractality has also been studied in reading, see Wallot et al. (2014).

tool, which "analyzes texts on over 200 measures of cohesion, language, and readability" (Graesser and Kulikowich, 2011, p. 193) has been developed for the analyses of texts at the higher level of processing, e.g., by focusing on coherence and cohesion. Given the wide range of text properties that have been used as correlates of behavioral measurements in various fields (e.g., computational linguistics and the psychology of learning), in this exploratory study we can only focus on a selection of properties that we expect to be relevant to our research programme. We use two types of lower-level properties (frequencies of part-of-speech tags and sentence length) and two types of higher-level properties (lexical diversity and topic distribution). This is not of course to say that other properties are not potentially relevant to our research programme. Building upon our results we intend to explore additional properties in the future, both from studies on readability (e.g., the measurements delivered by Coh-Metrix) and from Natural Language Processing, e.g., language modeling[4] and embedding vectors[5].

In what follows we briefly characterize the four text properties used for our study, without providing any technical details. The derivation of series on the basis of these properties is described in section 5.1.

Part-of-speech tags, commonly abbreviated as "POS-tags," represent the syntactic class of a word. To some extent, they reflect syntactic structure. At the most general level, POS-tags classify words into major classes, such as "noun," "verb," "adjective," etc., but depending on the specific tagset used, more fine-grained distinctions can be made (e.g., between singular and plural nouns). Parts of speech are considered to be potentially relevant to our research programme because they provide important categorical information at the word level, which is no doubt prominent in reading because text is primarily structured into words, separated by white spaces. Accordingly, "lexical variables are thought to be the main driving force behind the reading process" (Wallot et al., 2014, p. 1746) (note that Wallot et al., 2014, p. 1746 actually reach the conclusion that "lexical features do not play a substantial role in connected text reading," but they only took word length and frequency into account, no categorical information; see also Wallot et al., 2013). Moreover, neurological studies have shown that different parts of speech, e.g., nouns, verbs, and adjectives, are processed at different cortical locations (Perani et al., 1999; Tyler et al., 2004; Scott, 2006; Shapiro et al., 2006; Cappelletti et al., 2008; Sudre et al., 2012; Fyshe et al., 2019). We have no precise expectation with respect to the type of effect that part-of-speech distributions may have on reading processing, or how their distributions may vary across the text categories compared in this study. We do expect them to be potential correlates of reading experience, however, e.g., because they differ in terms of their informativeness (see Seifart et al., 2018 for evidence showing that nouns are more informative than verbs, requiring more cognitive resources), and the type of information that they convey. For our study, we used the Stanford Tagger (version 3.6.0; see section 5.1 for details).

Sentence length, measured in terms of the number of tokens in a sentence, is a very basic indicator of lower-level text structure. In fictional texts, it is potentially informative because it tends to differ between narrative passages (with longer sentences) and passages with dialogues (with shorter sentences). The distribution of sentence length values across a text therefore, to some extent, reflects the text's composition in terms of perspective (external communication with narrative elements vs. internal communication, e.g., dialogs, monologs, thoughts). Sentence length was used in earlier approaches to text assessment (see for instance Petersen, 2007), and it has been used as a measurement for the study of fractality before by Drożdż et al. (2016), though not for a comparison of text types. Even though sentence length is certainly a rough indicator of lower-level text structure, it provides a starting point before we apply more specific measures[6].

Lexical diversity, a derivative of the choice of words in a text, is one of the most perspicuous text properties, and a rich vocabulary is often regarded as a hallmark of good authorship. For example, Simonton (1990) claims that lexical diversity correlates with "aesthetic success." He analyzed Shakespeare's sonnets and showed that there is a vocabulary shift from the more "obscure" to the more popular sonnets. Vocabulary and the richness of lexicon has also been found useful in the assessment of writers' proficiency, for instance in research on second language acquisition (see Laufer and Nation, 1995; Zareva et al., 2005; Yu, 2009). Given the importance of lexical diversity for readability measures, it is natural to include it in a study analyzing text properties that can be expected to have correlates in aesthetic experience. As a measurement of lexical diversity, we have used MTLD (see McCarthy and Jarvis, 2010 and section 5.1).

Topic modeling is a method used to analyze the content of texts by revealing hidden topics of documents in a collection. It has been used in computational studies of literary texts before, though with different objectives and background assumptions (van Cranenburgh et al., 2019). We are interested in the changes of topic distributions along a text, as it can be expected to have an impact on how "a reader progresses through a text with a growing understanding for its content, topics and themes" (Wallot et al., 2014, p. 1749). To extract the distribution of topics from a text, the text is split into segments and then, to infer the topic distribution, a topic modeling method is applied (using Latent Dirichlet Allocation/LDA, see section 5.1).

[4]Language modeling is an essential part of many language processing tasks, such as machine translation, summarization and speech recognition. A language model computes the probability of a sequence of words and predicts the probability of the next word (Jurafsky and Martin, 2009). Language models capture both semantic and structural information, as the probability for a given word to occur is a function of both the surrounding structure and the semantic context.

[5]Embedding vectors—n-dimensional vectors of floats—represent the distribution of a linguistic segment and allow for the computation of (dis)similarities between segments. A wide variety of models have been proposed to represent text at the level of sub-word, word, sentence, etc. (for example, see Pennington et al., 2014; Bojanowski et al., 2017; Devlin et al., 2019).

[6]For example, Coh-Metrix measures complexity in terms of NP-density, the number of higher-level constituents and the presence of logical connectors, see Graesser et al. (2003).

3. GLOBAL MEASURES OF VARIABILITY AND SELF-SIMILARITY

In the present section, we introduce ways of analyzing the series of text properties that were introduced in the previous section. We focus on two global statistical features (variability and self-similarity). These properties were selected because they have previously been used in visual aesthetics and have been shown to be associated with artworks and other visually pleasing stimuli (see section 1).

Variability reflects the degree to which a particular feature (e.g., edge orientation or color) is likely to vary across an image. It can be measured simply by computing the variance of a series. The variance of a random variable X is

$$\mathcal{V}(X)=E[(X-\mu)^2] \quad (1)$$

$E[.]$ denotes the expected value and μ is the population mean. The variance of, for example, the distribution of sentence length reflects the amount of variation in the length of sentences across a text. Despite its mathematical simplicity, we will see that variance performs effectively in the classification of text categories (section 5).

Fractality and self-similarity reflect the degree to which parts of an image have features similar to the image as a whole, i.e., an image is self-similar if it shows similar features at different scales of resolution (scale-invariance). To analyze variability and fractality/self-similarity, several methods are available. The method used in the present study (Multi-Fractal Detrended Fluctuation Analysis/MFDFA) is described below. Alternative methods, such as methods based on entropy, box counting, wavelets and cross-correlation analysis, are described in the **Supplementary Material**.

3.1. Multi-Fractal Detrended Fluctuation Analysis

Self-similarity can be measured with Detrended Fluctuation Analysis (DFA) (Peng et al., 1994) and its extension Multi-Fractal DFA (MFDFA) (Kantelhardt et al., 2002; Oświecimka et al., 2006). These methods have been widely used for studying long-range correlations in a broad range of research fields, such as biology (Das et al., 2016), economics (Caraiani, 2012), music (Sanyal et al., 2016), and animal song (Roeske et al., 2018). MFDFA can be related to Fourier spectral analysis and both methods provide similar results for the degree of fractality (Heneghan and McDarby, 2000). Moreover, MFDFA has a theoretical and practical connection to wavelet-based methods (Leonarduzzi et al., 2016).

In the present work, we will apply MFDFA to the fractal analysis of texts. MFDFA has been used for textual analysis before. For example, Drożdż and Oświęcimka (2015) applied this method to sentence-length series in comparison to other natural series (e.g., the discharge of the Missouri river and sunspot number variability) and non-natural series (e.g., stock market and Forex index prices). The results suggest that natural languages possess a multifractal structure that is comparable to that of other natural and non-natural phenomena. Yang et al. (2016) investigated long-range correlations in sentence-length series in a famous classic Chinese novel, based on the number of characters in each sentence. This study showed that there was a long-range correlation, though it was weak. A diachronic fractality analysis of word-length in Chinese texts spanning 2,000 years revealed two different long-range correlations regimes for short and large scales (Chen and Liu, 2018). An analysis of fractality of sentence-length series in several Western fictional texts revealed that, although most fictional texts show a long-range correlation, the degree of multifractality can vary quite substantially, ranging from monofractal to highly multifractal structure (Drożdż et al., 2016). Although sentence length can be measured in various ways, e.g., as the number of characters or words in unlemmatized and lemmatized texts, the different ways yield robust results that have comparable distributions and similar patterns of long-range correlations (Vieira et al., 2018). MFDFA has also been applied in empirical studies of reading (Wallot et al., 2014).

Given a series $X = x_1, x_2, \cdots, x_N$, MFDFA can be summarized as follows:

1. Subtract the mean and compute the cumulative sum, called the profile, of the series:
   $Y(i)=\sum_{k=1}^{i}[x_k - \langle x \rangle], i=1, \cdots, N$
2. Divide the profile of the signal into $N_s=N/s$ windows for different values of s
3. Compute the local trend, Y', which is the best fitting line (or polynomial), in each window
4. Calculate the mean square fluctuation of the detrended profile in each window v, $v=1, \cdots, N_s$:
   $F^2(s,v)=\frac{1}{s}\sum_{i=1}^{s}[Y(s\times(v-1)+i) - Y'(s\times(v-1)+i)]^2$
5. Calculate the qth order of the mean square fluctuation:
   $F_q(s)=\{\frac{1}{N_s}\sum_{v=1}^{N_s}[F^2(s,v)]^{q/2}\}^{1/q}$
6. Determine the scaling behavior of $F_q(s)$ vs. s: $F_q(s) \sim s^{h(q)}$

In the windowing procedure, as the length of the series, N, is not usually divisible by the chosen window size, s, a part of the series may be ignored. Therefore, it is possible to repeat the windowing procedure, starting from the end. Accordingly, the number of segments rises up to $2 \times N_s$, which is taken into account in the averaging in step 5. In our experiments, we analyzed each series in windows of size s_i; $s_0=16$ and $s_i=s_{i-1}+2^{\lfloor \log(s_{i-1})-1 \rfloor}$, for $i \geq 1$ and $s_i \leq \lfloor N/3 \rfloor$. In other words, the size of the windows is selected from the sequence $16, 24, 32, 48, 64, \ldots$, up to a point where the series is split into three non-overlapping segments. Detrending is accomplished by linear fits, so the fluctuation is computed according to the deviation from the best fitted line in each window. We changed the parameter of the fluctuation function, q, from -5 to 5 with a step size of 0.25.

3.2. The Degree of Fractality

The procedure of MFDFA is equivalent to DFA if q is fixed at 2. For monofractal series, $h(q)$ is independent of q. If a series is stationary, $h(2)$ is equal to the Hurst Exponent, a well-known measure in fractal analysis studies. We refer to this value as $\mathcal{H}$, the degree of fractality of the series. In the remainder of this text, wherever we use "Hurst exponent" we refer to this value,

even though the series may not be stationary. For uncorrelated series, in which each event is independent of other events, $\mathcal{H} \simeq 0.5$. With $\mathcal{H} > 0.5$, the series is more fractal. In the opposite direction, if $\mathcal{H} < 0.5$, the series is called anti-persistent. In such cases a large value in the series is most likely followed by a small value, and vice versa.

To get a more intuitive understanding of $\mathcal{H}$, we show the sentence-length series of a few cases in our corpus (section 4) as well as the profile of each series in **Figure 1** (see step 1 of MFDFA in above). **Figure 1A** represents the series of the *Glossary of Chess Terms* by Gregory Zorzos, which is one of the texts in the non-fictional categories of our corpus. This dictionary-like book consists of a list of terms and their definitions. It represents an example of an anti-persistent text, with $\mathcal{H}$=0.37, and it is an extreme case in the corpus, with the lowest fractal degree. **Figure 1B** corresponds to *The Boats of the "Glen Carrig"* by William Hope Hodgson. With $\mathcal{H}$=0.48, this book has the second lowest $\mathcal{H}$ value and is closest to 0.5, which shows that there is almost no correlation among the elements of its series. This book is categorized as a non-canonical text in our corpus. As a side note, the lower bound of fractality for sentence-length series of canonical texts in the corpus is at $\mathcal{H}$=0.58, which is the value measured for *Old Mortality* by Walter Scott. In **Figures 1C,D** we show the plots of one canonical

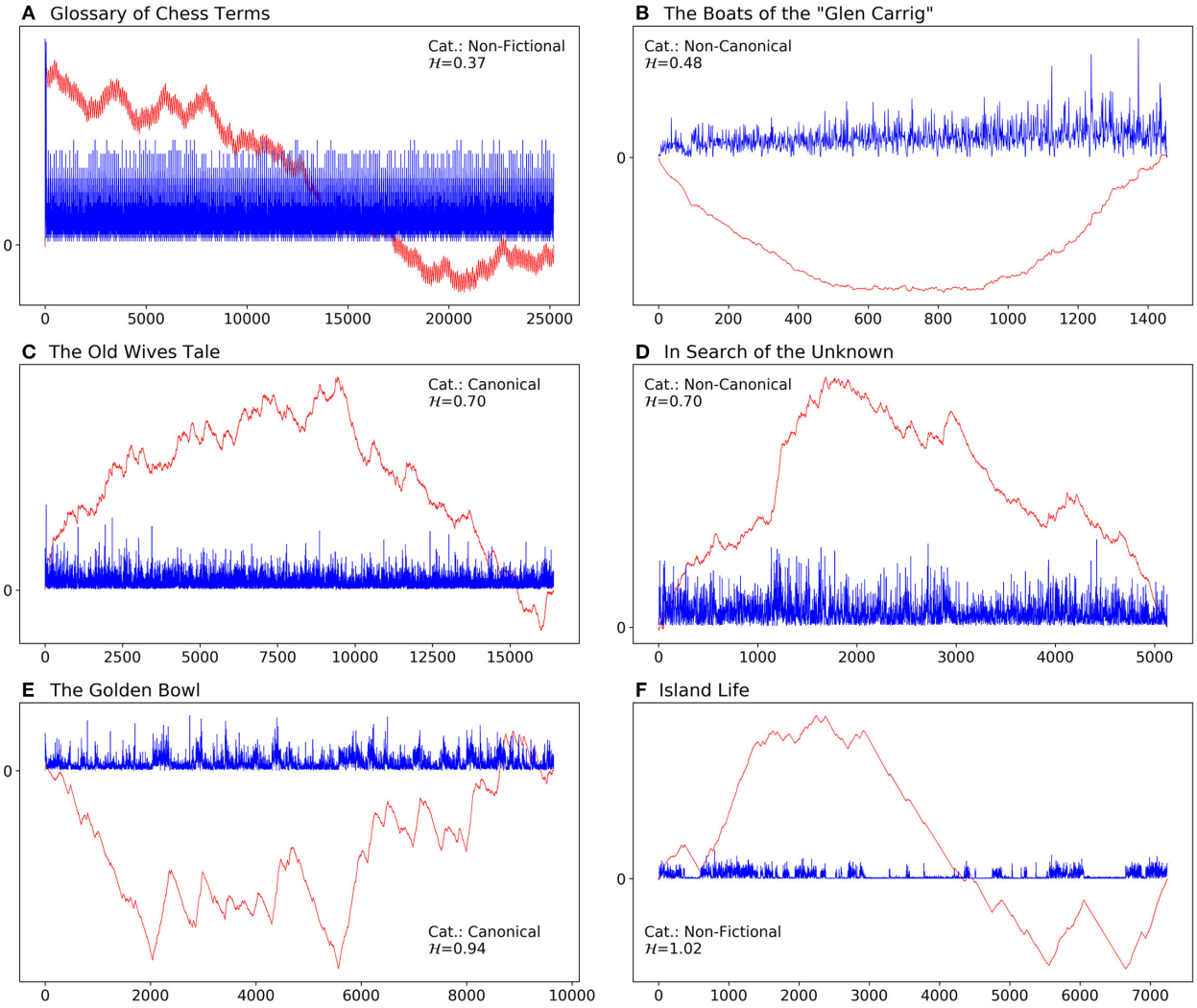

FIGURE 1 | Sentence length series (blue) and their profiles (red; cumulative sum of mean centered series) of some example texts in the corpus. The series have been scaled up by a factor of 20 to show more detail. The category (Cat.) and the fractal degree, $\mathcal{H}$, of each text is shown inside each panel. **(A)** *Glossary of Chess Terms* by Gregory Zorzos, with the lowest fractal degree in our corpus. **(B)** *Boats of the "Glen Carrig"* by William Hope Hodgson with $\mathcal{H}$=0.48, a non-canonical text with the lowest value among the fictional books. **(C)** *Women in Love* by D. H. Lawrence, the median of canonical texts. **(D)** *In Search of the Unknown* by Robert W. Chambers, representing the median of non-canonical texts. **(E)** *The Golden Bowl* by Henry James, with the highest fractal degree among canonical texts. **(F)** *Island Life* by Alfred Russel Wallace, a non-fictional text, with the highest fractal degree in the whole corpus.

and one non-canonical fictional book with a medium degree of fractality, within the relevant category/sub-corpus. For both *The Old Wives' Tale* by Arnold Bennett, a canonical text, and *In Search of the Unknown* by Robert W. Chambers, a non-canonical text, $\mathcal{H}$=0.70. **Figure 1E** represents the series of a canonical text with the highest fractal degree ($\mathcal{H}$=0.94) in the corpus, namely *The Golden Bowl* by Henry James. Finally, the text with the highest value of $\mathcal{H}$ in the entire corpus is *Island Life* by Alfred Russel Wallace, a text from the non-fictional sub-corpus, with $\mathcal{H}$=1.02.

3.3. The Degree of Multifractality and Fractal Asymmetry

From $h(q)$, one can compute the degree of multifractality and the fractal asymmetry, two metrics that represent the fractal complexity of the series. From $h(q)$, the Hölder exponents, α, and the singularity spectrum, $f(\alpha)$, are computed as follows (h' is the derivative of h):

$$\alpha = h(q) + qh'(q) \tag{2}$$

$$f(\alpha) = q[\alpha - h(q)] + 1 \tag{3}$$

Then, the degree of multifractality is defined as $\mathcal{D} = \alpha_{max} - \alpha_{min}$ (cf. Kantelhardt et al., 2002; Drożdż et al., 2016). α_{min} and α_{max} denote the beginning and the end of $f(\alpha)$, respectively. The fractal asymmetry is also computed from $f(\alpha)$:

$$\mathcal{A} = \frac{\Delta\alpha_L - \Delta\alpha_R}{\Delta\alpha_L + \Delta\alpha_R} \tag{4}$$

where $\Delta\alpha_L = \alpha_0 - \alpha_{min}$ and $\Delta\alpha_R = \alpha_{max} - \alpha_0$ (Drożdż and Oświęcimka, 2015). α_0, corresponding to q=0, usually points to the peak of the $f(\alpha)$ curve. It is obvious that $\mathcal{D} = \Delta\alpha_L + \Delta\alpha_R$. In section 5, we will use the three values (fractal degree [$\mathcal{H}$], degree of multifractality [$\mathcal{D}$], and fractal asymmetry [$\mathcal{A}$]) as a basis for classifying the three categories of text (canonical, non-canonical, and non-fictional).

To illustrate these concepts visually, we show the results of the fractal analysis for canonical texts by Charlotte Brontë and D. H. Lawrence in **Figure 2**. The two texts have been converted to series by using the sentence-length property. **Figures 2A,B** show $F_q(s)$ for different values of q ranging from −5 to 5. The slopes of the linear fits to the curves of $F_q(s)$ are represented in **Figures 2C,D** for the two texts, respectively. It is obvious that the slopes of the fits, $h(q)$, change as q changes. This result indicates that the texts are multifractal.

By applying Equations (2) and (3) to these plots, the singularity spectrum of the series is computed as shown in **Figures 2E,F**. *Jane Eyre* by Charlotte Brontë has a high degree of multifractality, $\mathcal{D}$=0.55. The figure also shows that the series has a high fractal asymmetry, $\mathcal{A}$=0.42 (**Figure 2E**). The long right tail of the singularity spectrum indicates that the multifractal structure of the data series is less sensitive to local fluctuations of large magnitudes. Conversely, if a singularity spectrum has a long left tail, this means that its multifractal structure is less affected by local fluctuations with small magnitudes (see Ihlen, 2012). **Figure 2F** presents the singularity spectrum for *The Rainbow* by D. H. Lawrence with $\mathcal{D}$=0.27 and $\mathcal{A}$= − 0.01. The values shown here illustrate that the series of the text has a degree of multifractality smaller than that of *Jane Eyre*, but it is (almost) symmetrical.

4. THE JEFP CORPUS

As mentioned in section 1, our corpus consists of three sub-corpora representing three major text categories: a collection of canonical fictional texts, a corpus of non-canonical fictional texts, and a corpus of non-fictional (expository) texts.

The canonical fictional sub-corpus comprises 77 English prose texts, written by 31 different authors, from Period C (1832–1900) and Period D (20th century) of the *Corpus of Canonical Western Literature* (Green, 2017)[7]. We selected those texts from the corpus that were sufficiently long for our analysis (at least 35K words).

The non-canonical fictional texts were downloaded from e-book publishing sites in the internet. We primarily used www.smashwords.com, an e-book distributor website that is catering to classic texts, independent authors and small press. It offers a large selection of books from several genres and allows downloads in various formats. The books are classified into "Fiction," "Non-fiction," "Essays," "Poetry," and "Screenplays." We selected random books from various prose genres, using the site's filter to make sure that the books had a minimal length comparable to that of canonical texts.

We further supplemented the corpus of non-canonical books with the lowest rated books on www.goodreads.com and www.feedbooks.com, as well as books with the lowest rates of downloads on the Project Gutenberg site. These books are in the public domain, written mostly between 1880 and 1930 and more than 45K words in length. In this way, we obtained 95 books of non-canonical literature (from as many authors in each case). We made sure to collect non-canonical texts from the same time period as for our canonical sub-corpus to minimize the effect of phenomena, such as short-term language change on our analyses. However, collecting "low-quality" non-canonical texts from one century back is not easy, as texts of this category are unlikely to be preserved or even digitized. Those texts that survived are likely of relatively high quality. Therefore, our non-canonical sub-corpus can be regarded as a top-notch non-canonical, and thus, comparatively close to the canonical sub-corpus, which renders the classification tasks more difficult (section 5.5). Nevertheless, the non-canonical texts selected by us are clearly non-canonical in the sense that they currently do not belong to any canon of literature like the one that we used for the selection of canonical texts (Green, 2017).

As another discriminating factor between canonical and non-canonical texts, we counted the number of articles that each author has in the top 30 language editions of Wikipedia. This measure is evidence for the international reputation of an author.

[7]It is an interesting question, beyond the scope of this study, whether a different canon, e.g., a canon of African American Literature (cf. Gates and McKay, 2004)—would yield different results.

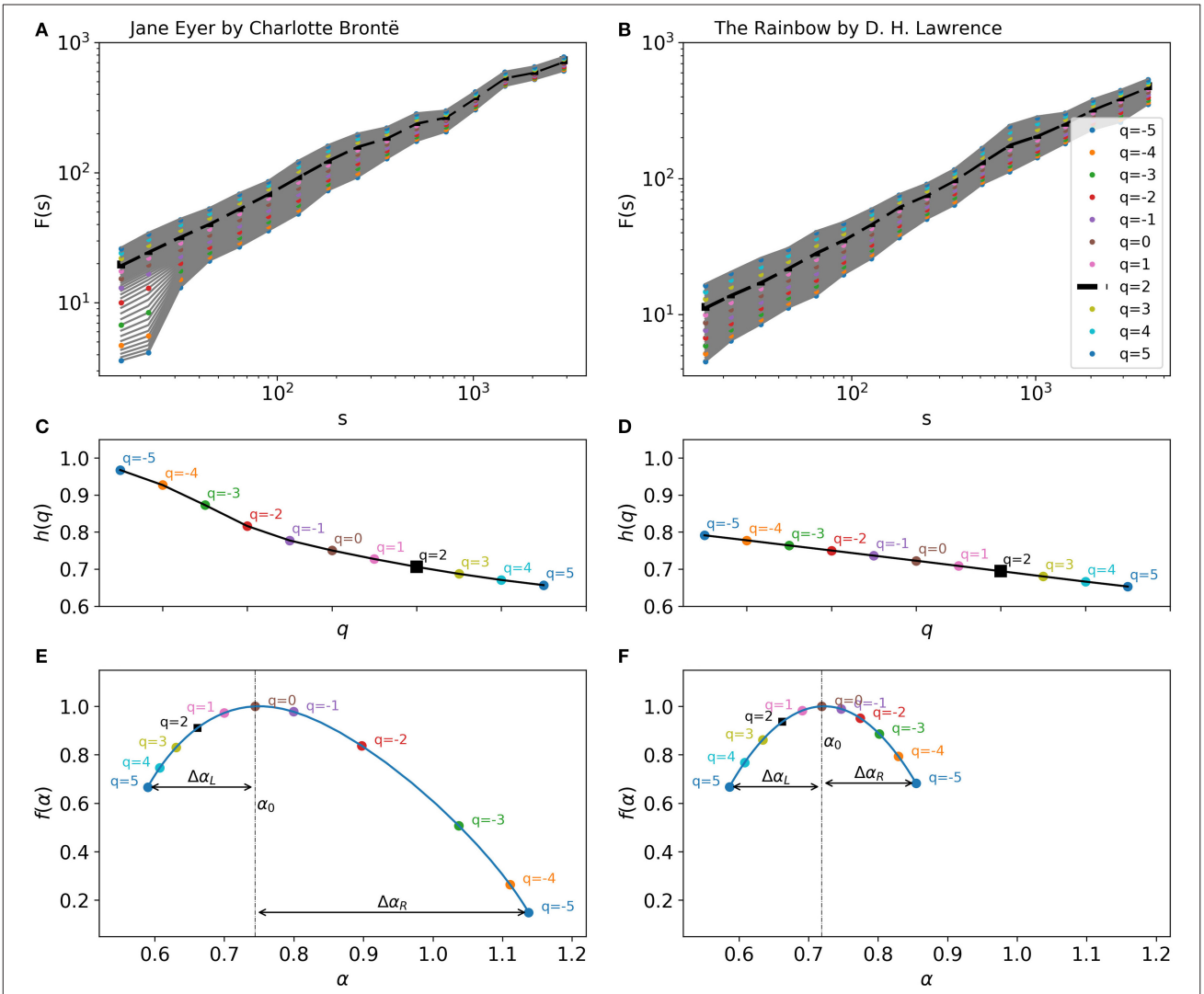

FIGURE 2 | (A,B) qth order of mean square fluctuation of sentence-length series of *Jane Eyre* by Charlotte Brontë **(A)** and *The Rainbow* by D. H. Lawrence **(B)**. The plots show $F_q(s)$ for different scales of s and for different values of q, ranging from −5 to 5 in steps of 0.25. The colored points represent values that correspond to the integer qs. $\mathcal{H}$ is the slope of the best linear fit to the dashed curve, which corresponds to q=2, and which identical to the output of the DFA method. **(C,D)** Slopes of the best linear fits to the fluctuation function in **(A,B)**, respectively. From these plots, singularity spectra are computed by the Legrand transformation. **(E,F)** Singularity spectra of the two texts. α_0 indicates the peak of each curve. The width of the curve, $\Delta\alpha = \Delta\alpha_L + \Delta\alpha_R$, known as the degree of multifractality ($\mathcal{D}$), shows how multifractal a series is. Here, both texts are highly multifractal. The fractal asymmetry ($\mathcal{A}$) of the curve is calculated from $\Delta\alpha_L$ and $\Delta\alpha_R$. The curve is asymmetrical for *Jane Eyre*, but symmetrical for *The Rainbow*.

Figure 3 shows a strip plot for all authors in each category. There is a clear separation between the authors of the two groups. All authors of canonical texts have at least 15 articles each in the 30 Wikipedia editions. In the non-canonical category, each author has up to 13 articles at most; for the majority of authors, the number is <5. These numbers provide independent evidence for the higher degree of prestige (Underwood and Sellers, 2016) of canonical authors, in comparison to non-canonical authors.

To compile the non-fictional sub-corpus we relied on Project Gutenberg. We downloaded all non-fictional books and randomly selected 132 books from different genres, such as architecture, astronomy, geology, geography, philosophy, psychology, and sociology. To increase the diversity, we added the first two volumes of *The Encyclopedia Britannica* published by the University of Cambridge and a text called *Glossary of Chess Terms* by Gregory Zorzos. This text was added to our corpus because of its extreme fractal behavior, as discussed in the previous section and shown in **Figure 1**. The texts of the two fictional categories, with the exception of the last one, were published in similar time periods.

Table 1 contains aggregate information about the length and time of publication of the texts contained in all categories.

Information about the entire JEFP Corpus is provided in **Supplementary Table S1**. The mean lengths of the texts are different for each of the three text categories. It is important to mention that the exact length of a text does not affect the results of our experiments, given that the texts are sufficiently long to be analyzed robustly for their variability and fractal properties (see section 5.2). As far as the year of publication is concerned, the canonical fictional texts span a broader time period than the non-canonical texts. This is not surprising, as canonical literature represents a small selection of texts of a period, and thus constitutes a smaller population per time unit than non-canonical texts. In terms of both language history and literature periodization these differences are negligible.

The texts were tagged manually to eliminate material not belonging to the core text, such as tables of contents and indices. Headers were left in the text, as they are potentially informative. Moreover, the texts were cleaned up semi-automatically using regular expressions to identify (and re-join) hyphenated words at the end of a line.

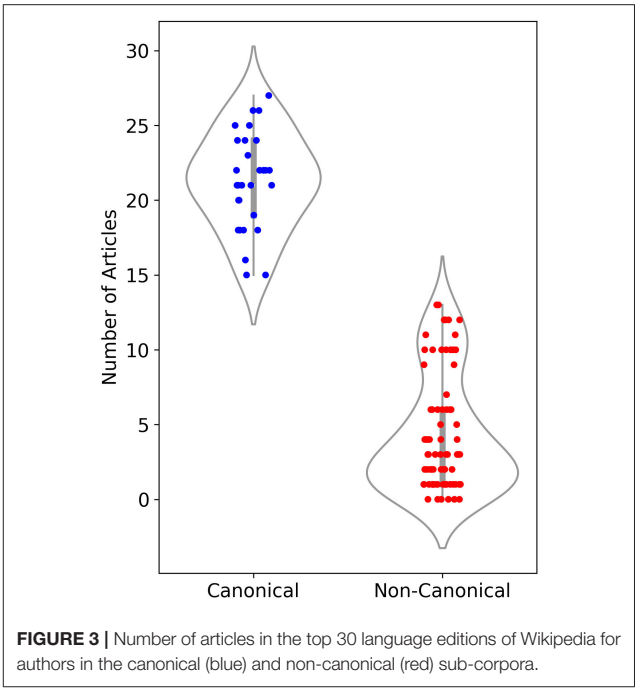

FIGURE 3 | Number of articles in the top 30 language editions of Wikipedia for authors in the canonical (blue) and non-canonical (red) sub-corpora.

TABLE 1 | Number of texts, number of authors, mean text length (number of tokens), and mean year of publication (±SD) for the different text categories of the JEFP Corpus.

	# Texts	# Authors	Length ($\times 10^3$)	Year of publication
Canonical	77	31	196 ± 91	1,870 ± 31
Non-canonical	95	80	102 ± 44	1,905 ± 19
Non-fictional	135	131	168 ± 193	1,902 ± 19

5. ANALYSIS AND CLASSIFICATION RESULTS

The core hypothesis behind the present study is that the three different text categories under analysis—non-fictional texts, fictional/canonical and fictional/non-canonical ones—differ in terms of fractality and variability. The JEFP Corpus allows us to test this hypothesis, as it contains samples of text from the three categories of interest. In order to compare the text categories, we carried out bivariate as well as multivariate analyses. In the bivariate analyses we compare the various statistics across the three categories of text; in order to get an understanding of the interplay between, and relative importance of, the various features, we carried out two binary classification tasks. The first task (Task 1) is to separate the fictional from the non-fictional works. The second task (Task 2) consists in separating the canonical fictional texts from the non-canonical ones.

The series analyzed were derived from the four textual properties described in section 2, POS-tag frequencies, sentence length, lexical diversity, and topic distributions. The first two properties are regarded as correlates of lower-level processing (decoding) while the latter two are taken to correspond to higher-level processing (integration, see sections 1 and 2). In section 5.1 we describe how the basic measurements for these properties were obtained, converting the texts to series. Following some remarks concerning the validation of the methods (section 5.2) we present the results in section 5.3. In section 5.4 the source of the multifractality is discussed before we present the results of the classification tasks in section 5.5.

5.1. Converting Texts Into Series

To convert a text into a series of POS-tag frequencies, we determined the number of each specific tag in the sentences of the text. In our analysis, we focused on the major parts of speech, i.e., nouns, adjectives, verbs, and pronouns. For the annotations we used the Stanford POS-tagger (Toutanova et al., 2003). For the calculations, we included all types of nouns, i.e., singular as well as plural nouns and proper names. Several types of verb forms, e.g., base forms, past tense forms, gerunds, past participles—were all treated as verbs. The category of "adjective" includes simple, comparative as well as superlative adjectives. Pronouns are either personal or possessive. We thus obtained four different series derived from POS-tags.

Sentence length was measured in terms of tokens as delivered by the tokenizers of the NLTK-package for Python (Bird et al., 2009). The texts were first sentence-tokenized (split into sentences), and then each sentence was word-tokenized. The length of each sentence is the number of its tokens. A token is an instance of a word, number or punctuation mark in a text. Punctuation marks were not removed and treated as tokens.

Lexical diversity measures the richness of vocabulary of a text. Several metrics have been proposed for measuring lexical diversity. Type-Token Ratio (TTR) is the simplest one, in which the number of distinct words (types) is divided by the length of the text. However, TTR is highly sensitive to text length. In our experiments (cf. section 5), we therefore use the Measure for Textual Lexical Diversity (MTLD; McCarthy and Jarvis, 2010),

which is more robust because it is less sensitive to text length. To convert a text into a series of lexical diversity values, we first segmented the text into segments of 100 tokens, which seemed like a good compromise between reliability of the calculations, and the required minimal length of series for fractal analysis. We then computed MTLD values for each segment to obtain a series for this feature.

Topic modeling is a high-level analysis of text that focuses on the content conveyed. To extract the topic distribution of a text, we first segmented the text into coherent chunks using the TopicTiling algorithm (Riedl and Biemann, 2012). Then, we applied Latent Dirichlet Allocation (LDA) (Blei et al., 2003; Griffiths and Steyvers, 2004) to all chunks of all texts in the corpus, thus obtaining a topic model. The number of topics, one of the hyperparameters of LDA, was set to 100. The resulting topic model is a statistical model that shows the importance of each word in a topic. Afterwards, the topic model was applied to each chunk of a text to infer the distribution of the 100 topics (the 'topic probabilities'). In order to convert the vector of topic probabilities to a series, we calculated the Jensen–Shannon divergence of the topic representations of adjacent chunks.

5.2. Methodological Validation of Fractal Analysis of the Corpus

As the length of texts varies considerably in our corpus, we conducted an experiment to see whether text length affects the degree of fractality. For the text in the three categories, we chose the maximum scale, i.e., the maximum size of the windows, in such a way that the average of the maximum scales was similar for the three text categories. A statistical test showed no significant difference. Therefore, in our experiments we do not impose any restriction on the maximum scale. In the MFDFA method the scaling behavior of the fluctuation function, $F_q(s)$, is determined vs. the window size, s, i.e., $F_q(s) \sim s^{h(q)}$. By fitting lines to the double-log diagrams of the fluctuation function, $h(q)$ is computed for different values of q and fractal features are then obtained. Looking at the linear fits and how well they have been fitted to the values reveals information about fractal regimes for different values of q in the text properties and for the three text categories. R^2 is a statistical measure that determines how well a linear fit represents the data. We computed mean R^2 values for each text category. For all values of q and for all text properties R^2 is larger than 0.94, which means that linear fits are very precise and close to the observed values. The R^2 values are summarized in **Supplementary Figure S1**.

5.3. Analysis of Variance and Fractality

After generating the series for the seven text properties for all texts, we calculated the variance, $\mathcal{V}$, as a measure of how variable each text property was across each text. Moreover, we used MFDFA to calculate the following fractal features for each text: the degree of fractality ($\mathcal{H}$), the degree of multifractality ($\mathcal{D}$) and the degree of fractal asymmetry ($\mathcal{A}$) (see section 3). As Kolmogorov-Smirnov tests revealed that some of the data were not normally distributed, the data was entered into a Wilcoxon test to assess the differences between the three subcorpora, supplemented by non-parametric Mann-Whitney tests for all (*post-hoc*) pairwise comparisons. The median values of the variances and fractal features are shown in **Table 2** for all three subcorpora of text (canonical, non-canonical, and non-fictional). In addition, we obtained the same statistics for both types of fictional text (canonical and non-canonical texts) together, as we distinguish two classification tasks: the distinction between fictional vs. non-fictional texts (Task 1), and between canonical vs. non-canonical texts (Task 2; see section 4).

Table 2 shows that none of the text properties (four types of POS-tag frequencies, sentence length, lexical diversity and topic probabilities) results in significantly different median values for all features (variance and fractality measures) in both tasks. The higher-level properties (MTLD and topic distributions) do not vary significantly across text types for the fractal features. However, the variance ($\mathcal{V}$) is significantly different for all features in both tasks. Strikingly, $\mathcal{V}$ values are always higher for non-fictional texts than for fictional texts, except for the values obtained from frequencies of pronouns, and from MTLD values. This difference is mainly driven by non-canonical fictional texts. $\mathcal{V}$ values for canonical texts range in between those for non-fictional and non-canonical texts. In some cases (verb frequencies, sentence length and topic distributions), the values for canonical texts are not significantly different from those of non-fictional texts, but higher than the values for non-canonical texts.

In summary, in terms of $\mathcal{V}$, canonical texts are more similar to non-fictional texts than to non-canonical texts. Only for the frequency distribution of pronouns and MTLD values do the canonical texts exhibit the highest values, followed by non-canonical texts and, with even lower values, by non-fictional texts. **Figure 4** shows the differences between the variances of the text categories for all properties. Note that the magnitude of the variances does not reflect the magnitude of the mean values for the text properties (cf. **Supplementary Table S2** for the mean values).

Results for the degree of fractality ($\mathcal{H}$) are listed in **Table 2** and the means are visualized in **Figure 5A**. The degree of fractality is of similar magnitude (closer to 0.5) for all text properties for canonical and non-canonical fictional texts. By contrast, the $\mathcal{H}$ values for non-fictional texts are generally higher than for either type of fictional text (canonical or non-canonical), with the exception of the frequencies of nouns, sentence length and topic distributions. These results suggest that a lower degree of long-range correlations might be a uniform characteristic of fictional texts as opposed to non-fictional texts, regardless of the status of the fictional texts as canonical or non-canonical.

As **Table 2** and **Figure 5B** show, the values for the degree of multifractality, $\mathcal{D}$, are significantly higher for the frequencies of verbs and pronouns as well as sentence length in non-fictional as opposed to fictional texts. A comparison of canonical and non-canonical fictional texts reveals that the $\mathcal{D}$ values of canonical texts are consistently higher than or equal to the values for non-canonical texts, even though this tendency reaches statistical significance only for the frequencies of nouns and verbs, as well as sentence length.

The degree of asymmetry, $\mathcal{A}$, does not differ between canonical and non-canonical fictional texts (**Table 2** and

TABLE 2 | Median values of all text properties.

		Noun	Verb	Adjective	Pronoun	Sentence length	MTLD	Topic distribution
$\mathcal{V}$	Lit.	11 (10, 13)	6.5 (5.6, 7.2)	2.3 (2.1, 2.8)	3.3 (3.0, 3.6)	220 (184, 277)	376 (361, 391)	4.5e-3 (4.4e-3, 4.6e-3)
	Non-Lit.	19 (17, 20)	7.5 (7.0, 8.3)	4.2 (3.9, 4.5)	1.9 (1.4, 2.1)	305 (290, 336)	322 (295, 348)	4.8e-3 (4.6e-3, 5.3e-3)
		***	**	***	***	***	***	***
	Can.	15 (14, 17)[b]	9.0 (7.2, 9.9)	3.3 (3.0, 3.9)[b]	4.3 (3.7, 5.0)[c]	321 (296, 367)	390 (375, 408)[c]	4.8e-3 (4.5e-3, 4.9e-3)
	Non-Can.	9.1 (8.1, 10)[c]	5.0 (4.5, 6.0)[c]	1.9 (1.7, 2.1)[c]	2.7 (2.4, 3.0)[c]	163 (145, 194)[c]	357 (345, 381)[b]	4.2e-3 (4.0e-3, 4.5e-3)[c]
		***	***	***	***	***	**	***
$\mathcal{H}$	Lit.	0.714 (0.706, 0.725)	0.66 (0.65, 0.67)	0.685 (0.677, 0.695)	0.67 (0.66, 0.68)	0.70 (0.69, 0.71)	0.65 (0.64, 0.66)	0.63 (0.62, 0.65)
	Non-Lit.	0.69 (0.67, 0.71)	0.72 (0.70, 0.76)	0.72 (0.70, 0.74)	0.71 (0.70, 0.73)	0.73 (0.70, 0.75)	0.68 (0.66, 0.70)	0.66 (0.61, 0.69)
		*	***	***	***		***	
	Can.	0.72 (0.70, 0.73)	0.67 (0.65, 0.68)[c]	0.69 (0.68, 0.70)[a]	0.67 (0.65, 0.68)[c]	0.70 (0.68, 0.71)	0.64 (0.63, 0.66)[a]	0.64 (0.61, 0.66)
	Non-Can.	0.71 (0.70, 0.73)	0.66 (0.64, 0.67)[c]	0.68 (0.67, 0.69)[b]	0.67 (0.65, 0.68)[c]	0.70 (0.68, 0.71)	0.65 (0.64, 0.66)[b]	0.63 (0.61, 0.65)
$\mathcal{D}$	Lit.	0.30 (0.26, 0.32)	0.20 (0.17, 0.21)	0.31 (0.28, 0.33)	0.67 (0.66, 0.68)	0.26 (0.24, 0.28)	0.20 (0.17, 0.22)	0.19 (0.18, 0.21)
	Non-Lit.	0.26 (0.23, 0.31)	0.37 (0.34, 0.40)	0.30 (0.25, 0.37)	0.71 (0.70, 0.73)	0.34 (0.29, 0.42)	0.20 (0.18, 0.22)	0.25 (0.20, 0.28)
			***		***	***		
	Can.	0.34 (0.32, 0.36)	0.23 (0.19, 0.25)[c]	0.32 (0.28, 0.34)	0.67 (0.65, 0.68)[c]	0.31 (0.26, 0.35)	0.20 (0.16, 0.22)	0.18 (0.15, 0.21)[a]
	Non-Can.	0.26 (0.23, 0.30)	0.16 (0.15, 0.20)[c]	0.29 (0.27, 0.33)	0.67 (0.65, 0.68)[c]	0.23 (0.20, 0.25)[c]	0.21 (0.17, 0.23)	0.20 (0.18, 0.21)
		***	***			***		
$\mathcal{A}$	Lit.	0.03 (−0.05, 0.09)	0.09 (0.02, 0.14)	0.04 (−0.03, 0.07)	0.09 (−0.01, 0.17)	0.08 (0.01, 0.14)	0.11 (0.06, 0.18)	0.15 (0.06, 0.23)
	Non-Lit.	0.24 (0.12, 0.41)	0.61 (0.55, 0.68)	0.55 (0.48, 0.66)	0.36 (0.28, 0.48)	0.56 (0.45, 0.69)	0.15 (0.04, 0.28)	0.09 (−0.04, 0.27)
		***	***	***	***	***		
	Can.	0.09 (0.00, 0.13)[a]	0.10 (0.04, 0.16)[c]	0.04 (−0.03, 0.08)[c]	0.16 (0.05, 0.21)[b]	0.13 (0.04, 0.23)[c]	0.10 (0.03, 0.20)	0.20 (−0.02, 0.27)
	Non-Can.	−0.04 (−0.13, 0.06)[c]	0.07 (−0.03, 0.20)[c]	0.04 (−0.07, 0.11)[c]	−0.01 (−0.07, 0.20)[c]	0.02 (−0.02, 0.12)[c]	0.12 (−0.06, 0.24)	0.15 (0.05, 0.26)

*The 95% confidence intervals are shown in parentheses. The rows represent the features analyzed (variance [$\mathcal{V}$], degree of fractality [$\mathcal{H}$], degree of multifractality [$\mathcal{D}$] and fractal asymmetry [$\mathcal{A}$]). Each feature is analyzed for two tasks: Task 1, fictional (Lit.; N = 172) vs. non-fictional (Non-Lit.; N = 135) texts, and Task 2, canonical (Can.; N = 77) vs. non-canonical (Non-Can.; N = 95) texts. The asterisks indicate whether the differences between the two text categories of a given task are statistically significant (Mann-Whitney test; *$p \leq 0.05$; **$p \leq 0.01$; and ***$p \leq 0.001$). In addition, canonical and non-canonical texts are compared separately to non-fictional texts; the superscript numbers indicate significances (Mann-Whitney test; [a]$p \leq 0.05$; [b]$p \leq 0.01$; and [c]$p \leq 0.001$).*

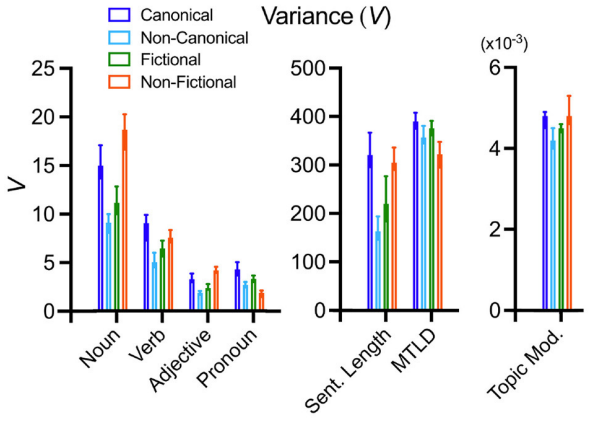

FIGURE 4 | Plots of the median variances for all text properties. The colors indicate the different text categories, as indicated at the upper left-hand side of the figure. The whiskers indicate 95% confidence intervals. For significance levels of the differences, see **Table 2**. Sent. Length, sentence length; Topic Mod., topic modeling.

Figure 5C). For lower-level properties, fictional texts are rather symmetrical (i.e., $\mathcal{A}$ is close to 0), and $\mathcal{A}$ is much higher for non-fictional texts than for fictional texts. For the higher-level properties (MTLD values and topic distributions), $\mathcal{A}$ values do not vary across the three sub-corpora.

To summarize the observations made above, canonical fictional texts show more variability with respect to the properties measured in our study than non-canonical texts, and are, in this respect, more similar to non-fictional texts. However, the lower degree of fractality ($\mathcal{H}$) suggests that the two types of fictional texts display a lower degree of long-range correlations than non-fictional texts do. Moreover, canonical texts tend to be more multifractal than non-canonical texts in terms of the frequencies of nouns and verbs, as well as for sentence length (higher $\mathcal{D}$). Unlike in the case of non-fictional texts, the fractal spectra of fictional texts are rather symmetrical ($\mathcal{A}$ is closer to 0).

The individual values for the variance (y-axis) and fractal features (x-axis) for selected text properties are visualized as scatter plots in **Figure 6** to illustrate the separation and overlap between the different text categories. For this figure, we chose

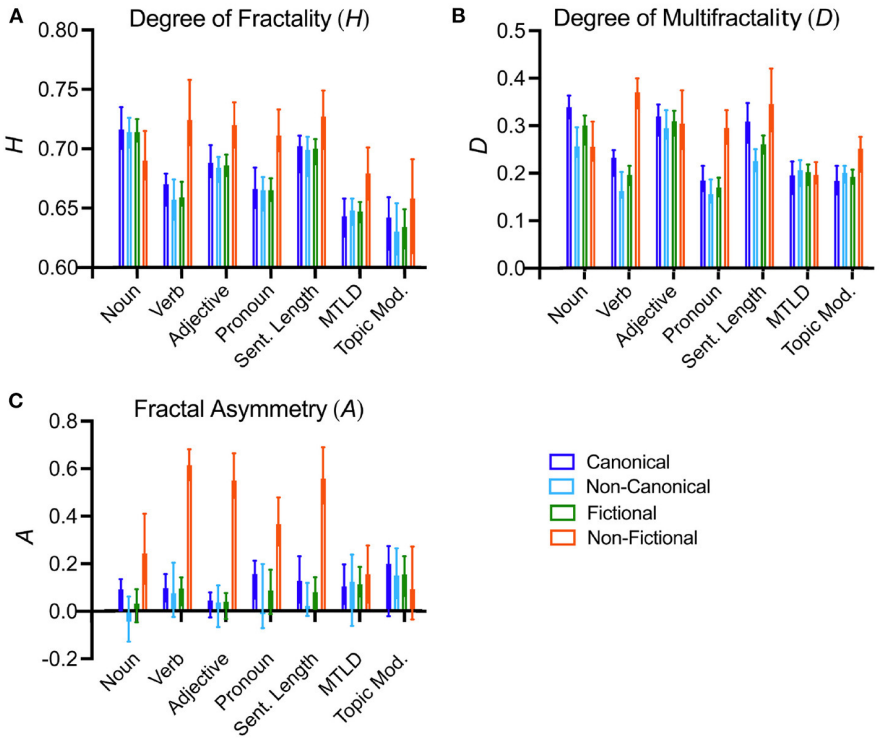

FIGURE 5 | Plots of the median values for the degree of fractality **(A)**, the degree of multifractality **(B)**, and fractal asymmetry **(C)** for all text properties. The colors indicate the different text categories, as indicated on the down right of the figure. The whiskers indicate 95% confidence intervals. For significances of the differences, see **Table 2**. Sent. Length, sentence length; Topic Mod., topic modeling.

plots that showed a relatively clear separation of the text categories by subjective visual inspection. **Figure 6A** shows the degree of multifractality and the variance of noun series. As stated above (**Table 2**), the variances for non-canonical texts tend to be lower than those of the other two categories. **Figure 6B** depicts the degree of multifractality and the variance of pronoun frequencies; it shows that fictional texts tend to have a higher variance compared to non-fictional texts. Both **Figures 6A,B** confirm that non-fictional texts scatter in a wider range of the degree of multifractality. In **Figures 6C,D**, the variances of verb and adjective series are plotted as a function of the degree of asymmetry. Fractal patterns of non-fictional texts are more asymmetrical (higher $\mathcal{A}$). Again, canonical fictional texts exhibit a wider scatter, as variance is higher compared to non-canonical texts, which suggests a more diverse usage of language structures in the former category. The behavior of non-fictional texts varies across the tags. For example, the texts scatter more widely in the plot of adjectives (**Figure 6C**), while their pronoun variances cover a narrower range (**Figure 6D**), since pronouns are not so frequent in non-fictional texts (**Supplementary Table S2**). **Figure 6** also illustrates that non-fictional texts have more complex fractal patterns and spread more broadly along the fractal feature (x-)axes. Non-fictional texts tend to show a higher fractal degree and more fractal asymmetry than fictional texts.

5.4. The Source of Multifractality

Both fictional and non-fictional texts are multifractal up to a certain degree, as can be seen from **Table 2**. It is therefore important to analyze the source of multifractality in the texts of the corpus. Multifractality in a series can be caused by (i) the presence of long-range correlations of small and large fluctuations, or (ii) a broad probability distribution (Kantelhardt, 2011). Therefore, we used the Iterative Amplitude Adjusted Fourier Transform (IAFFT) surrogate test to investigate the source of multifractality in the text property series. IAAFT retains the distribution and linear structures of series while destroying non-linear correlations. If the multifractality of a series is not due to non-linear correlations, IAAFT has no effect on the multifractality (for a comprehensive discussion on surrogate methods, see Lancaster et al., 2018).

We applied IAAFT surrogate tests to the series derived from all text properties, for all books in our corpus. We allowed the IAAFT algorithm to iterate up to 500 times to generate a surrogate. For each series we generated an ensemble of 100 surrogates and compared the degrees of multifractality of the series with those of the surrogates. For all texts in canonical, non-canonical and non-fictional categories as well as all 307 books in the corpus, we computed the percentage of the texts whose degree of multifractality is significantly larger than the mean degree of multifractality of the surrogates ($p < 0.05$).

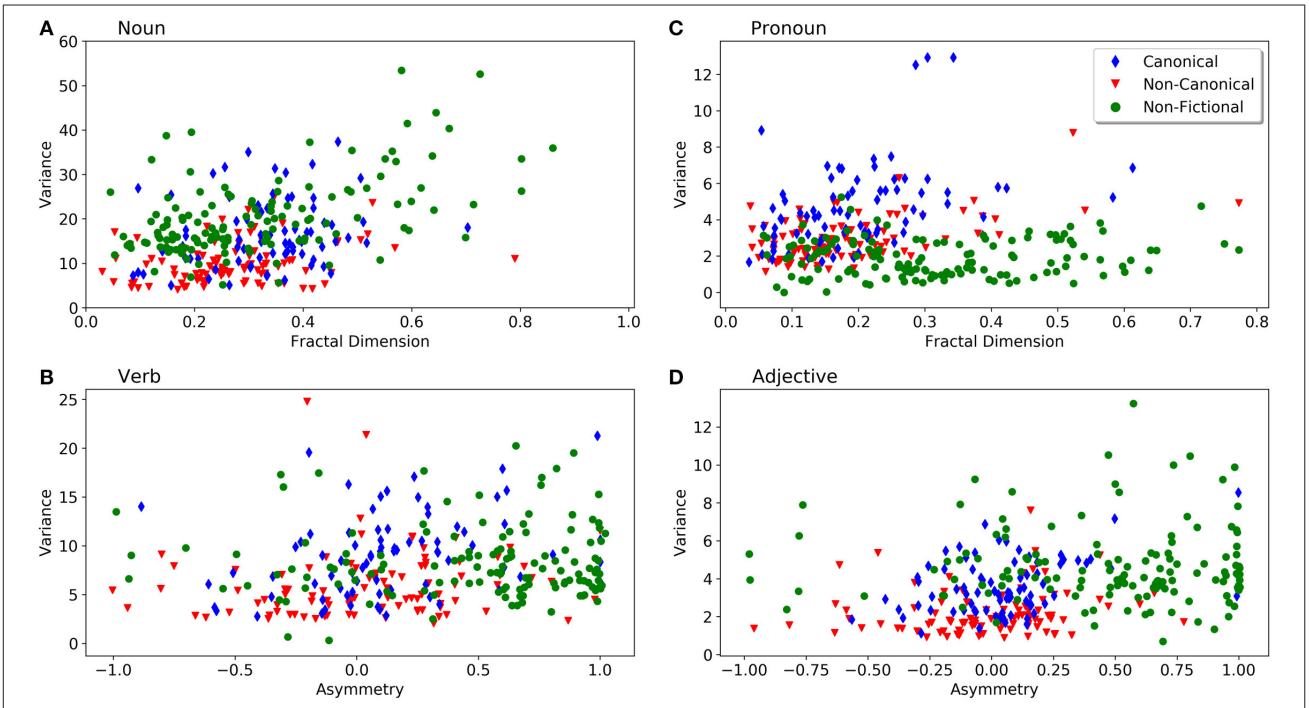

FIGURE 6 | Scatter plots of variance (y-axis) and fractal features of POS-tags (x-axis). **(A)** Degree of fractality ($\mathcal{H}$) and the variance of noun series. **(B)** Degree of multifractality ($\mathcal{D}$) and the variance of verb series. **(C)** Degree of multifractality ($\mathcal{D}$) and the variance of pronoun series. **(D)** Fractal asymmetry ($\mathcal{A}$) and the variance of adjective series. Each dot represents one text from our corpus. For color coding of the text categories, see insert in **(B)**.

Before summarizing the results it is important to note that we do not expect all texts to exhibit multifractality. Our hypothesis says that texts from different categories may differ in terms of their degrees of multifractality. This implies that some texts will be more multifractal than others, and in fact, some texts are expected not to be multifractal at all. Nonetheless, we want to compare the results summarized in **Table 2** with the results of the surrogate tests. We only provide a rough summary here. The results of the IAAFT surrogate tests are summarized in **Supplementary Figure S2**.

The results show that across the three text categories, more than 90% of all texts have a significantly higher degree of multifractality than their surrogates (on average), for all text properties. The only text property for which the average number is lower is topic distribution, with a value of 88%. For lower-level text properties and MTLD, more than 90% of texts have a higher degree of multifractality than their surrogates in all text categories, with the exception of sentence length in the non-canonical texts. More than 88% of the sentence-length series derived from the non-canonical texts and the topic distribution series derived from the fictional/non-canonical and non-fictional texts show a significant difference from their surrogates. The lowest value is the one for the topic distribution series of canonical literary texts (83%). We will see below (section 5.5) that in fact, the fractal features of topic distribution cannot classify the text categories with a high accuracy, either.

While we cannot offer a detailed assessment of surrogate analyses for all individual texts and all individual features, from our point of view the aggregate results make it very unlikely that observed instances of multifractality are not due to long-range correlations, though we cannot, of course, exclude that in individual cases they are caused by a broad probability distribution.

5.5. Classification

While a statistical analysis of features gives insights into the distribution of a single feature (cf. section 5.3), classification separates classes from each other, potentially in a non-linear fashion, which is a better way to detect differences between the text categories than a linear analysis of single properties. In this section, we describe the results for the classification of the text categories. As mentioned before, we distinguish two classification tasks: fictional texts are classified against non-fictional texts (Task 1), and canonical fictional texts against non-canonical fictional texts (Task 2).

For a better understanding of the postulated level of text processing, we present results for the lower-level and higher-level properties separately as well as in combination (**Table 3**). For classification, we used a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel. As the features have varying scales, we normalized them to a mean of 0 and standard deviation of 1. The evaluation measure is balanced accuracy, which is a

TABLE 3 | Accuracy of classification (in %) for the non-fictional/fictional distinction (Task 1) and the canonical/non-canonical distinction (Task 2).

	Task 1		Task 2	
	Variability	**Fractal features**	**Variability**	**Fractal features**
Noun	71.0 ± 2.5	75.3 ± 2.3	69.5 ± 3.8	62.4 ± 3.0
Verb	56.8 ± 3.7	75.1 ± 1.5	68.3 ± 2.1	55.5 ± 3.0
Adjective	74.1 ± 2.7	80.4 ± 2.3	69.7 ± 4.0	51.6 ± 3.7†
Pronoun	69.5 ± 0.9	72.1 ± 1.8	68.0 ± 1.9	52.2 ± 4.7†
Sentence-length	65.0 ± 2.2	74.0 ± 2.0	69.3 ± 2.9	59.7 ± 3.2
MTLD	63.7 ± 2.3	56.9 ± 3.2	52.3 ± 3.3†	55.5 ± 3.1
Topic distribution	62.8 ± 2.3	64.0 ± 3.3	60.6 ± 3.4	49.2 ± 3.5†
Lower-level	92.4 ± 2.1	86.0 ± 2.0	71.6 ± 2.6	62.9 ± 3.9
Lower-level, Combined	94.9 ± 1.0		71.4 ± 4.8	
Higher-level	72.4 ± 1.9	63.2 ± 3.3	63.5 ± 3.2	57.1 ± 1.8
Higher-level, Combined	71.8 ± 2.9		61.9 ± 4.5	
Lower- & Higher-level	93.6 ± 1.3	84.9 ± 1.6	73.6 ± 2.3	65.0 ± 1.7
Lower- & Higher-level, Combined	94.7 ± 1.3		71.6 ± 3.6	

Means±SD are listed (N = 10). All values are significantly different ($p \leq 0.05$) from random accuracy (50%), except where indicated by a dagger (†).

weighted average accuracy value that is proportional to the size of each class, and therefore, does not favor larger classes. We assessed the statistical significance of differences between settings by using a 5 × 2 cv paired *t* test (Dietterich, 1998) (significance level at $p \leq 0.05$). In this test, 2-fold cross validation is repeated 5 times and the dataset is shuffled each time.

Before we present our results, it is important to note that the objective of the classification task is not to obtain a maximum degree of accuracy in absolute terms. We are interested in a comparison of the relative discriminatory power of statistics capturing specific global structural properties (variability, [multi]fractality) obtained from specific observables (four types of POS-tags, sentence length, lexical diversity and topic distributions). While in computational linguistics it is customary to compare classification models to alternative models classifying the same textual material, such a comparison does not seem very informative to us for the purpose of our specific research question. It is to be expected that state-of-the-art language models, such as BERT (Devlin et al., 2019), will achieve much higher classification results than any of the models trained by us. In fact, Louwerse et al. (2008) already achieved 100% accuracy of text classification with a bigram model distinguishing "Literature" from "Non-literature." We use the classification procedure as a way of understanding the relationship between the various predictor variables, i.e., as a tool for multivariate analysis, in an empirical study motivated by theoretical research questions. Note also that even in absolute terms, a comparison with other models would only make sense if the models used a comparable number of features. Readers interested in accuracy scores obtained in the classification of fictional/literary texts are referred to van Cranenburgh and Bod (2017) (see for instance the table on p. 1234).

In **Table 3**, we report the mean and the standard deviation for the 10 runs for each setting. The top part of **Table 3** shows the classification results for individual properties. The analysis of variability provides comparable accuracies in Task 1 and Task 2. Exceptions are provided by verb frequency, which leads to much higher classification rates in Task 2 than in Task 1, and MTLD values, which are better predictors in Task 1. The best performance is observed for adjective frequency, which yields the highest accuracy of all predictors in Task 1, and which provides the best results in Task 2 as well (see also **Table 2**). The variance of MTLD values is more powerful in distinguishing fictional texts from non-fictional text (Task 1), but it cannot separate canonical from non-canonical texts in Task 2. As a lexical diversity measure, MTLD reflects the richness of vocabulary of a text. To get a better understanding of lexical diversity of fictional and non-fictional texts, we submitted the global MTLD-values of the texts, grouped into the categories "non-fictional," "fictional/canonical," and "fictional/non-canonical," to an ANOVA. The test did not reveal a significant difference between the lexical diversity of the text categories (p=0.68). This finding is surprising, as lexical diversity is often regarded as a hallmark of good authorship, and can thus be expected to vary across the sub-corpora of interest.

The fractal features result in better accuracies in Task 1 than in Task 2 for all properties, with the exception of MTLD, which performs similarly in both tasks. The highest classification rate for Task 1 is, again, obtained for adjective series (80.4%). The series of lower-level properties, i.e., POS-tags frequencies and sentence length, perform well in Task 1. By contrast, the fractal features cannot distinguish well between canonical and non-canonical fictional texts (Task 2). This result is in accordance with the finding that the degree of fractality ($\mathcal{H}$) and the degree of asymmetry ($\mathcal{A}$) are of similar magnitude for canonical and non-canonical texts for almost all text properties (cf. **Table 2**).

The POS-tag frequencies and sentence length are regarded as lower-level properties and MTLD and topic distribution as higher-level properties. The top part of **Table 3** presents the classification results for variance and the fractal features separately. When combining the two feature groups for all lower-level and all higher-level properties, as shown in the middle part of the table, a considerably improved accuracy is achieved in Task 1. Although the variance of each property alone does not provide a classification accuracy higher than 74% (for adjective frequencies), their combination effectively raises the accuracy up to 92%. Using all fractal features together for the classification task also increases the performance considerably. Finally, when all variances and fractal features are combined, the performance gets even better. A 5 × 2 cv paired *t* test confirms that all of these improvements are significant. In Task 2, we do not observe such a large improvement by accumulating the variances or the fractal features. For example, the performance of a model combining all variances of lower-level features is only slightly better than the performance of the variance of noun or adjective frequencies. For the fractal features, the classification accuracy of the combined

model is similar to that of noun series only. The combination of all features does not offer any improvement either.

We also ran the classification task using all higher-level properties. In Task 1 (cf. the middle part of **Table 3**), the combination of the variances of two higher-level properties results in a considerable improvement. By contrast, a combination of the fractal features leads to no improvement. It is therefore expected that the combination of all variances and fractal features does not improve classification. Adding more features to an SVM classifier may actually decrease the classification result, because the SVM classifier tries to maximize generalization. Such a decrease is observed if all features are combined together. In Task 2, we can see that the combination of variances of the higher-level features improves the classification results, though not for the fractal features. The accumulation of all features does not provide any obvious improvement either.

Lower-level and higher-level properties can be combined to analyze the different classes of text, as shown at the bottom of **Table 3**. In Task 1, we observe no improvement when combining all variances or all fractal features. Finally, the result obtained by combining all features is not significantly different from the classifier that was trained on all features (variances and fractal features) of lower-level properties. In Task 2, when all variances or all fractal features are taken into account, an improvement can be observed. The combination of all features does not, however, improve the accuracy of the model compared to the model trained on all variances.

In summary, the results of the classification experiment show that lower-level properties are more effective in distinguishing fictional text from non-fictional text (Task 1) than higher-level properties. Even individual properties—the frequencies of nouns and verbs—reach accuracies higher than 70%, or even 80% in the case of the fractal features for adjectives. By combining lower-level features in the classification task, the accuracy reaches 95%. The accuracy values for Task 2 range between 68 and 70% for individual lower-level features, and are much lower for higher-level features. The performance of the classifier does not improve significantly if the lower-level features are combined, and the resulting accuracy score (71.6%) is not significantly higher than the score for adjective frequencies (69.7%). This finding points to a strong correlation of the lower-level features in Task 2[8].

6. DISCUSSION AND CONCLUSIONS

The starting point of this article was the question of whether canonical and non-canonical fictional texts exhibit systematic differences in terms of structural design features. In order to put any observable differences into perspective, we also included non-fictional (expository) texts for comparison. Our study was inspired by findings from the field of vision, where aesthetic experience has been linked to the structural features of variability (measured in terms of variance) and fractality or self-similarity. As pointed out in section 1, the transfer from vision to reading has obvious limitations. Still, given the widespread assumption of domain-general processes in the processing of language (Diessel, 2019), we tested to what extent the features that have been observed to correlate with observers' preferences in vision differentiate canonical from non-canonical fictional texts, and fictional from non-fictional texts.

We used four features as the basic measurements, classified into lower-level features (frequencies of POS-tags and sentence length) and higher-level features (lexical diversity and topic distributions). The global structural design features that we investigated were those that have been shown to be prominent in vision (variability, fractality). By applying the relevant statistical methods to series derived from the four types of text properties we generated global statistics of various types. In our analysis we proceeded in two steps: First, we carried out bivariate comparisons between the three text categories under analysis, for each feature separately. Second, we used the features to classify the three text categories in question, thus determining the relative importance of each feature as well as their combined discriminatory power.

In what follows we discuss our findings and their implications with a focus on the central questions addressed in this article.

6.1. Lower-Level and Higher-Level Text Properties

Our results have shown that generally speaking, the lower-level properties from which we derived series are better discriminators than the higher-level features, for the three text categories of interest. The differences between the text categories are more pronounced in bivariate comparisons, and the accuracy levels reached in the classification tasks are significantly higher for lower-level properties than for higher-level properties. This finding has some parallels obtained in research on other sensory domains. In the visual domain, the global spatial distribution of several low-level properties (for example, luminance changes, edge orientations, curvilinear shape and color features; see section 1) has been related to the global structure of traditional artworks and other preferred visual stimuli. In the auditory domain, music has been shown to be characterized by fluctuations in low-level features, such as loudness and pitch (Voss and Clarke, 1975), frequency intervals (Hsü and Hsü, 1991), sound amplitude (Kello et al., 2017; Roeske et al., 2018), and other simple metrices, such as measures of pitch, duration, melodic intervals, and harmonic intervals (Manaris et al., 2005), as well as patterns of consonance (Wu et al., 2015). These and many other studies indicate that low-level properties of music show long-range correlations that are scale-invariant and obey a power law. Interestingly, similar results were obtained for animal songs (Kello et al., 2017; Roeske et al., 2018).

Why are lower-level text properties informative with respect to the three text categories under analysis? We surmise that lower-level properties of text to some extent

[8]Instead of the variance, which is the second moment of the sample data, one can run classification using the first moment, i.e., the mean. Although this is not the focus of our study, we also carried out the classification tasks using the means of the text properties. The results are provided in **Supplementary Table S3**. The results show that the performance of the models is slightly better using means rather than variances in distinguishing fictional from non-fictional texts (Task 1), but the variances provide better classification results in separation of the canonical from the non-canonical texts (Task 2), which is the more difficult task.

reflect discourse modes (Smith, 2003). These modes—Narrative, Report, Description (temporal), Information and Argument (atemporal)—are associated with different frequency distributions of POS-tags (cf. also Biber, 1995, who uses more specific categories in his multi-dimensional register analysis, however). For example, the Narrative mode is associated with verbs, while Description requires more adjectives. In a comparison of fictional and non-fictional text, it is moreover important to bear in mind that fictional text implies both external communication (between the narrator and the reader) and internal communication (between the protagonists, in the form of dialogues) as well as internal monologs and thoughts. Our results suggest that non-fictional texts show more global variability between discourse modes than fictional texts. Canonical fictional texts seem to pattern with non-fictional texts in terms of their higher global variability, in comparison to non-canonical fiction. While this hypothesis requires more (qualitative as well as quantitative) in-depth studies, it suggests that canonical authors may use a richer variety of discourse modes (or narrative techniques) than non-canonical authors. We intend to test this hypothesis in future studies.

Considering the higher-level properties, only one of the four features studied, the variance $\mathcal{V}$, showed differences between all of the three text categories. No significant differences were observed for any of the fractal features, with the exception of the Hurst exponent $\mathcal{H}$) determined on the basis of MTLD measurements, which is higher for non-fictional than for fictional texts (cf. **Table 2**). Accordingly, the classification rates obtained by using higher-level features only are relatively low (up to 71.8% for the classification of fictional vs. non-fictional texts, and 61.9% for the classification of canonical fictional vs. non-canonical fictional texts; cf. **Table 3**). However, when comparing lower-level and higher-level properties and their distributions in different text types it should be borne in mind that higher-level properties, in particular thematic structure across a text, cannot easily be measured. We have used the distribution of topic probabilities across texts as an indicator of thematic organization. It is of course conceivable that this way of operationalizing thematic structure is imperfect, or at least does not measure properties that have correlates in reading comprehension. In future studies, we will therefore experiment with a broader range of properties, including measurements of cohesion like those provided by Coh-Metrix (Graesser and Kulikowich, 2011; McNamara and Graesser, 2020).

6.2. Variability and Fractality/Long-Range Correlations

In vision, variability and fractality have both been shown to be important discriminators of stimuli, correlating with observers' preferences (see section 1). Our results show variability of the text properties to discriminate better between the three text types under analysis than statistics derived from MFDFA (93.6 vs. 84.9% for fictional vs. non-fictional texts, and 73.6 vs. 65% for canonical vs. non-canonical fiction). This suggests that long-range correlations play a minor role in the distinction between the three text categories under study.

In general, the variability of canonical fictional texts is higher than the variability of non-canonical texts, for all properties investigated by us. The results concerning the variability of non-fictional texts in comparison to fictional texts are less clear. For most properties, variability is higher for non-fictional than for fictional texts. As a result, the variability of canonical fictional texts is closer to (or the same as) that of non-fictional texts. Only for pronoun frequencies and MTLD values can a different pattern be observed. Here, canonical texts are more variable than both non-canonical and non-fictional texts.

It may be surprising to find that canonical fictional texts are, in some respects, more similar to non-fictional texts than they are to non-canonical fictional texts. However, in studies on reading difficulty it has been found that "[n]arrative texts are more easily understood than expository texts," and "read nearly twice as fast" (McNamara et al., 2013, p. 93). Canonical texts are often regarded as being more demanding than popular literature, and have often been written with a different purpose, and for a different readership (learned/educated readers). What McNamara et al. (2013, p. 93) write about expository texts—"they tend to include less familiar concepts and words and require more inferences on the part of the reader"—may apply to canonical fiction to a greater extent than it applies to non-canonical fiction.

Long-range correlations in general seem to be slightly more pronounced in non-fictional texts than in fictional texts. The Hurst exponent, $\mathcal{H}$, for the frequencies of verbs, adjectives and pronouns (as well as for MTLD) is significantly higher for non-fictional texts than for fictional texts (Task 1), and non-fictional texts display higher degrees of fractal asymmetry than fictional text (Task 1), for all lower-level properties. The classification experiments, however, show that fractality features do not discriminate as well as variability features (86.0 vs. 92.4% for Task 1, and 62.9 vs. 71.6% for the lower-level features).

In the visual domain, traditional artworks can be characterized by an intermediate to high degree of self-similarity (Braun et al., 2013; Brachmann and Redies, 2017). In the Fourier domain, large subsets of traditional artworks have spectral properties similar to pink noise, with a power ($1/f^p$) spectral exponent around p=1 (Graham and Field, 2007; Redies et al., 2007), which is also characteristic of many (but not all) natural patterns and scenes (Tolhurst et al., 1992). In MFDFA, this corresponds to a Hurst exponent of 1, while $\mathcal{H}$=0.5 indicates white noise (no long-range correlations, corresponding to a Fourier power spectral exponent of 0). The median $\mathcal{H}$ value for the different text properties ranges from 0.63 to 0.73 in our study, confirming previous results for sentence length by Drożdż et al. (2016). This degree of self-similarity thus lies in between that of most natural signals and random (white) noise. The relevance of this finding requires further exploration.

6.3. Outlook

This study has been exploratory in several respects. It is based on a limited selection of text properties (frequency distributions of POS-tags, sentence length, lexical diversity, topic distributions) whose use was motivated by general considerations and assumptions concerning language processing

and comprehension (cf. section 2), with the intention of identifying those features that are potentially relevant to an understanding of the differences between canonical and non-canonical fiction in terms of global structural design features. There are, of course, many other text properties that are potentially relevant to our endeavor, e.g., those used for text assessment. In future studies, we intend to use a broader set of text properties from which we can derive series, specifically taking into account additional features reflecting cohesion (Graesser et al., 2003; McNamara et al., 2013).

Originally inspired by results from vision, our study has also shown that—valuable though this inspiration has been—there are a number of limitations to the analogy, and for the study of global text design a methodological toolbox of its own is needed. Given that reading has a temporal dimension, the question of predictability may play an important role. In section 2 we mentioned that language modeling could be used to statistically analyze a text. We intend to explore such methods in the future.

Another direction in which the research programme of empirical textual aesthetics should be extended concerns the textual material. We investigated English texts only, and these texts were taken from a restricted time period (19th and early 20th centuries). In order to see whether any of the present findings can be generalized to other types of fictional texts, other languages or other time periods would have to be investigated separately.

Finally, a major challenge for the future concerns the relationship between structure observed in series derived from texts on the one hand, and aesthetic experience on the other. By studying structural differences between text categories that reflect preferences of societies—canonical texts are "privileged" because they are attributed a high cultural value—we have taken a first step in this direction, but aesthetic experience itself can only be studied experimentally. Before experiments can be run, however, it will be necessary to gain a better understanding of the (measurable) text properties, and the types of patterns exhibited by these properties, that can reasonably be assumed to have behavioral or neural correlates. Further observational (corpus) studies, with extensions of the type pointed out above, are good way of gaining such insights.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because, due to copyright restrictions, the corpus cannot be made public. Information to download the texts in the corpus is provided in the **Supplementary Table S1**. Requests to access the datasets should be directed to mahdi.mohseni@uni-jena.de.

AUTHOR CONTRIBUTIONS

MM, VG, and CR developed the idea for the present work and wrote the manuscript. MM designed the experiments, wrote the code, carried out the experiments, and analyzed the data. VG and MM collected the datasets and prepared them for analysis. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by institute funds from the Institute of Anatomy I, Jena University Hospital, and the German Research Council (grant number 380283145). The funders played no role in the study. Open access publication fees were provided by the University of Jena Library.

ACKNOWLEDGMENTS

An earlier version of this article has been published on the arXiv repository (Mohseni et al., 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.599063/full#supplementary-material

REFERENCES

Alm, C. O., and Sproat, R. (2005). "Emotional sequencing and development in fairy tales," in *Affective Computing and Intelligent Interaction*, eds J. Tao, T. Tan, and R. W. Picard (Berlin: Springer), 668–674. doi: 10.1007/11573548_86

Arnheim, R. (1974). *Art and Visual Perception: A Psychology of the Creative Eye*. Berkeley, CA: University of California Press.

Ashok, V. G., Feng, S., and Choi, Y. (2013). "Success with style: using writing style to predict the success of novels," in *Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, Seattle, WA, 1753–1764.

Bar, M., and Neta, M. (2006). Humans prefer curved visual objects. *Psychol. Sci.* 17, 645–648. doi: 10.1111/j.1467-9280.2006.01759.x

Bertamini, M., Palumbo, L., Gheorghes, T. N., and Galatsidas, M. (2016). Do observers like curvature or do they dislike angularity? *Br. J. Psychol.* 107, 154–178. doi: 10.1111/bjop.12132

Biber, D. (1995). *Dimensions of Register Variation. A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing With Python*. Sebastopol, CA: O'Reilly Media.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022. doi: 10.5555/944919.944937

Blohm, S., Schlesewsky, M., Menninghaus, W., and Scharinger, M. (2021). Text type attribution modulates pre-stimulus alpha power in sentence reading. *Brain Lang.* 214:104894. doi: 10.1016/j.bandl.2020.104894

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* 5, 135–146. doi: 10.1162/tacl_a_00051

Brachmann, A., Barth, E., and Redies, C. (2017). Using CNN features to better understand what makes visual artworks special. *Front. Psychol.* 8:830. doi: 10.3389/fpsyg.2017.00830

Brachmann, A., and Redies, C. (2017). Computational and experimental approaches to visual aesthetics. *Front. Comput. Neurosci.* 11:102. doi: 10.3389/fncom.2017.00102

Braun, J., Amirshahi, S. A., Denzler, J., and Redies, C. (2013). Statistical image properties of print advertisements, visual artworks and images of architecture. *Front. Psychol.* 4:808. doi: 10.3389/fpsyg.2013.00808

Cain, K., Oakhill, J., and Bryant, P. (2004). Children's reading comprehension ability: concurrent prediction by working memory, verbal ability, and

component skills. *J. Educ. Psychol.* 96, 31–42. doi: 10.1037/0022-0663.96.1.31

Cappelletti, M., Fregni, F., Shapiro, K., Pascual-Leone, A., and Caramazza, A. (2008). Processing nouns and verbs in the left frontal cortex: a transcranial magnetic stimulation study. *J. Cogn. Neurosci.* 20, 707–720. doi: 10.1162/jocn.2008.20045

Caraiani, P. (2012). Evidence of multifractality from emerging European stock markets. *PLoS ONE* 7:e40693. doi: 10.1371/journal.pone.0040693

Chatterjee, A., and Vartanian, O. (2014). Neuroaesthetics. *Trends Cogn. Sci.* 18, 370–375. doi: 10.1016/j.tics.2014.03.003

Chatzigeorgiou, M., Constantoudis, V., Diakonos, F., Karamanos, K., Papadimitriou, C., Kalimeri, M., et al. (2017). Multifractal correlations in natural language written texts: effects of language family and long word statistics. *Phys. A Stat. Mech. Appl.* 469, 173–182. doi: 10.1016/j.physa.2016.11.028

Chen, H., and Liu, H. (2018). Quantifying evolution of short and long-range correlations in chinese narrative texts across 2000 years. *Complexity* 2018:9362468. doi: 10.1155/2018/9362468

Cook, A. E., and Wei, W. (2019). What can eye movements tell us about higher level comprehension? *Vision* 3, 45–61. doi: 10.3390/vision3030045

Cordeiro, J., Inácio, P. R. M., and Fernandes, D. A. B. (2015). "Fractal beauty in text," in *Progress in Artificial Intelligence*, eds F. Pereira, P. Machado, E. Costa, and A. Cardoso (Cham: Springer), 796–802. doi: 10.1007/978-3-319-23485-4_80

Das, N. K., Dey, R., Chakraborty, S., Panigrahi, P., and Ghosh, N. (2016). Probing multifractality in depth-resolved refractive index fluctuations in biological tissues using backscattering spectral interferometry. *J. Opt.* 18:125301. doi: 10.1088/2040-8978/18/12/125301

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, MN: Association for Computational Linguistics), 4171–4186.

Diessel, H. (2019). *The Grammar Network. How Linguistic Structure Is Shaped by Language Use*. Cambridge: Cambridge University Press.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* 10, 1895–1923. doi: 10.1162/089976698300017197

Drożdż, S., and Oświęcimka, P. (2015). Detecting and interpreting distortions in hierarchical organization of complex time series. *Phys. Rev. E* 91:030902. doi: 10.1103/PhysRevE.91.030902

Drożdż, S., Oświęcimka, P., Kulig, A., Kwapień, J., Bazarnik, K., Grabska-Gradzińska, I., et al. (2016). Quantifying origin and character of long-range correlations in narrative texts. *Inform. Sci.* 331, 32–44. doi: 10.1016/j.ins.2015.10.023

Egan, C., Cristino, F., Payne, J. S., Thierry, G., and Jones, M. W. (2020). How alliteration enhances conceptual-attentional interactions in reading. *Cortex* 124, 111–118. doi: 10.1016/j.cortex.2019.11.005

Even-Zohar, I. (1990). Polysystem studies. *Poetics Today* 11, 9–26. doi: 10.2307/1772666

Febres, G., and Jaffe, K. (2017). Quantifying structure differences in literature using symbolic diversity and entropy criteria. *J. Quant. Linguist.* 24, 16–53. doi: 10.1080/09296174.2016.1169847

Fechner, G. T. (1876). *Vorschule der Ästhetik*. Leipzig: Breitkopf and Härtel.

Francisco, V., and Gervás, P. (2006). "Exploring the compositionality of emotions in text: word emotions, sentence emotions and automated tagging," in *AAAI-06 Workshop on Computational Aesthetics: Artificial Intelligence Approaches to Beauty and Happiness* (Boston, MA).

Fyshe, A., Sudre, G., Wehbe, L., Rafidi, N., and Mitchell, T. M. (2019). The lexical semantics of adjective-noun phrases in the human brain. *Hum. Brain Mapp.* 40, 4457–4469. doi: 10.1002/hbm.24714

Gates, H. L., and McKay, N. Y. (Eds.). (2004). *The Norton Anthology of African American Literature, 2nd Edn.* New York, NY: W.W. Norton & Co.

Graesser, A. C., D. M., and Kulikowich, J. (2011). Coh-metrix: providing multilevel analyses of text characteristics. *Educ. Res.* 40, 223–234. doi: 10.3102/0013189X11413260

Graesser, A. C., McNamara, D., and Louwerse, M. (2003). "What do readers need to learn in order to process coherence relations in narrative and expository text?" in *Rethinking Reading Comprehension*, eds A. Sweet and C. E. Snow (New York, NY: Guilford), 82–98.

Graham, D., and Field, D. (2007). Statistical regularities of art images and natural scenes: spectra, sparseness and nonlinearities. *Spat. Vision* 21, 149–164. doi: 10.1163/156856807782753877

Green, C. (2017). Introducing the corpus of the canon of western literature: a corpus for culturomics and stylistics. *Lang. Liter.* 26, 282–299. doi: 10.1177/0963947017718996

Griffiths, T. L., and Steyvers, M. (2004). Finding scientific topics. *Proc. Natl. Acad. Sci. U.S.A.* 101, 5228–5235. doi: 10.1073/pnas.0307752101

Gross, J., Millett, A., Bartek, B., Bredell, K., and Winegard, B. (2014). Evidence for prosody in silent reading. *Read. Res. Q.* 49, 189–208. doi: 10.1002/rrq.67

Guillory, J. (1987). Canonical and non-canonical: a critique of the current debate. *ELH* 54, 483–452. doi: 10.2307/2873219

Heneghan, C., and McDarby, G. (2000). Establishing the relation between detrended fluctuation analysis and power spectral density analysis for stochastic processes. *Phys. Rev. E* 62, 6103–6110. doi: 10.1103/PhysRevE.62.6103

Hernández-Gómez, C., Basurto-Flores, R., Obregón-Quintana, B., and Guzmán-Vargas, L. (2017). Evaluating the irregularity of natural languages. *Entropy* 19:521. doi: 10.3390/e19100521

Hsü, K., and Hsü, A. (1991). Self-similarity of the "1/f noise"' called music. *Proc. Natl. Acad. Sci. U.S.A.* 88, 3507–3509. doi: 10.1073/pnas.88.8.3507

Ihlen, E. A. (2012). Introduction to multifractal detrended fluctuation analysis in matlab. *Front. Physiol.* 3:141. doi: 10.3389/fphys.2012.00141

Iser, W. (1976). *Der Akt des Lesens. Theorie ästhetischer Wirkung*. München: Wilhelm Fink.

Jacobs, A. M. (2015). Neurocognitive poetics: methods and models for investigating the neuronal and cognitive-affective bases of literature reception. *Front. Hum. Neurosci.* 9:186. doi: 10.3389/fnhum.2015.00186

Jacobs, A. M., Lüdtke, J., Aryani, A., Meyer-Sickendieck, B., and Conrad, M. (2016). *Mood-Empathic and Aesthetic Responses in Poetry Reception. A Model-Guided, Multilevel, Multimethod Approach*. Amsterdam: John Benjamins.

Jakobson, R. (1960). "Linguistics and poetics," in *Style in Language*, ed Newton, K. M., (Cambridge, MA: MIT Press), 350–377.

Jurafsky, D., and Martin, J. H. (2009). *Speech and Language Processing, 2nd Edn.* Upper Saddle River, NJ: Prentice Hall.

Kakkonen, T., and Galić Kakkonen, G. (2011). "SentiProfiler: creating comparable visual profiles of sentimental content in texts," in *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage* (Hissar), 62–69.

Kantelhardt, J. W. (2011). "Fractal and multifractal time series," in *Mathematics of Complexity and Dynamical Systems*, ed R. A. Meyers (New York, NY: Springer), 463–487. doi: 10.1007/978-1-4614-1806-1_30

Kantelhardt, J. W., Zschiegner, S. A., Koscielny-Bunde, E., Havlin, S., Bunde, A., and Stanley, H. (2002). Multifractal detrended fluctuation analysis of nonstationary time series. *Phys. A Stat. Mech. Appl.* 316, 87–114. doi: 10.1016/S0378-4371(02)01383-3

Kello, C. T., Bella, S. D., Médé, B., and Balasubramaniam, R. (2017). Hierarchical temporal structure in music, speech and animal vocalizations: jazz is like a conversation, humpbacks sing like hermit thrushes. *J. R. Soc. Interface* 14:20170231. doi: 10.1098/rsif.2017.0231

Kintsch, W. (1988). The role of knowledge in discourse comprehension construction: a construction-integration model. *Psychol. Rev.* 95, 163–182. doi: 10.1037/0033-295X.95.2.163

Kliegl, R., Dambacher, M., Dimigen, O., Jacobs, A. M., and Sommer, W. (2012). Eye movements and brain electric potentials during reading. *Psychol. Res.* 76, 145–158. doi: 10.1007/s00426-011-0376-x

Kliegl, R., Grabner, E., Rolfs, M., and Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *Eur. J. Cogn. Psychol.* 16, 262–284. doi: 10.1080/09541440340000213

Knoop, C. A., Wagner, V., Jacobsen, T., and Menninghaus, W. (2016). Mapping the aesthetic space of literature "from below". *Poetics* 56, 35–49. doi: 10.1016/j.poetic.2016.02.001

König, E., and Pfister, M. (2017). *Literary Analysis and Linguistics*. Berlin: Erich Schmidt.

Koolen, C., van Dalen-Oskam, K., van Cranenburgh, A., and Nagelhout, E. (2020). Literary quality in the eye of the dutch reader: the national reader survey. *Poetics* 79:101439. doi: 10.1016/j.poetic.2020.101439

LaBerge, D., and Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cogn. Psychol.* 6, 293–323. doi: 10.1016/0010-0285(74)90015-2

Lancaster, G., Iatsenko, D., Pidde, A., Ticcinelli, V., and Stefanovska, A. (2018). Surrogate data for hypothesis testing of physical systems. *Phys. Rep.* 748, 1–60. doi: 10.1016/j.physrep.2018.06.001

Laufer, B., and Nation, P. (1995). Vocabulary size and use: lexical richness in L2 written production. *Appl. Linguist.* 16, 307–322. doi: 10.1093/applin/16.3.307

Leech, G. N. (1969). *A Linguistic Guide to English Poetry*. London: Longmans Harlow.

Leonarduzzi, R., Wendt, H., Abry, P., Jaffard, S., Melot, C., Roux, S., et al. (2016). P-exponent and P-leaders, part II: multifractal analysis. Relations to detrended fluctuation analysis. *Phys. A Stat. Mech. Appl.* 448, 319–339. doi: 10.1016/j.physa.2015.12.035

Locher, P. J., Stappers, P. J., and Overbeeke, K. (1999). An empirical evaluation of the visual rightness theory of pictorial composition. *Acta Psychol.* 103, 261–280. doi: 10.1016/S0001-6918(99)00044-X

Louwerse, M., Benesh, N., and Zhang, B. (2008). "Computationally discriminating literary from non-literary texts," in *Directions in Empirical Literary Studies: In Honor of Willie Van Peer*, eds S. Zyngier, M. Bortolussi, A. Chesnokova, and J. Auracher (Amsterdam: John Benjamins), 175–191. doi: 10.1075/lal.5.16lou

Maharjan, S., Arevalo, J., Montes, M., González, F., and Solorio, T. (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* (Valencia), 1217–1227. doi: 10.18653/v1/E17-1114

Maharjan, S., Kar, S., Montes, M., González, F. A., and Solorio, T. (2018). "Letting emotions flow: success prediction by modeling the flow of emotions in books," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (New Orleans, LA), 259–265. doi: 10.18653/v1/N18-2042

Manaris, B., Romero, J., Machado, P., Krehbiel, D., Hirzel, T., Pharr, W., et al. (2005). Zipf's law, music classification, and aesthetics. *Comput. Music J.* 29, 55–69. doi: 10.1162/comj.2005.29.1.55

McCarthy, P. M., and Jarvis, S. (2010). MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. *Behav. Res. Methods* 42, 381–392. doi: 10.3758/BRM.42.2.381

McManus, I., Edmondson, D., and Rodger, J. (1985). Balance in pictures. *Br. J. Psychol.* 76, 311–324. doi: 10.1111/j.2044-8295.1985.tb01955.x

McNamara, D., and Graesser, A. (2020). "Coh-metrix: an automated tool for theoretical and applied natural language processing," in *Applied Natural Language Processing and Content Analysis: Identification, Investigation, and Resolution*, eds P. McCarthy and C. Boonthum (Hershey, PA: IGI Global), 188–205. doi: 10.4018/978-1-60960-741-8.ch011

McNamara, D., and Magliano, J. P. (2009). Toward a comprehensive model of comprehension. *Psychol. Learn. Motiv.* 51, 297–384. doi: 10.1016/S0079-7421(09)51009-2

McNamara, D. S., Graesser, A., and Louwerse, M. (2013). "Sources of text difficulty: across genres and grades," in *Measuring Up: Advances in How We Assess Reading Ability*, eds J. Sabatini, E. Albro, and T. O'Reilly (Lanham, MD: R & L Education), 89–116.

McNerney, M. W., Goodwin, K. A., and Radvansky, G. A. (2011). A novel study: a situation model analysis of reading times. *Discour. Process.* 48, 453–474. doi: 10.1080/0163853X.2011.582348

Mehri, A., and Lashkari, S. M. (2016). Power-law regularities in human language. *Eur. Phys. J. B* 89:241. doi: 10.1140/epjb/e2016-70423-9

Menninghaus, W., Wagner, V., Wassiliwizky, E., Jacobsen, T., and Knoop, C. (2017). The emotional and aesthetic powers of parallelistic diction. *Poetics* 63, 47–59. doi: 10.1016/j.poetic.2016.12.001

Menninghaus, W., Wagner, V., Wassiliwizky, E., Schindler, I., Hanich, J., Jacobsen, T., et al. (2019). What are aesthetic emotions? *Psychol. Rev.* 126, 171–195. doi: 10.1037/rev0000135

Menninghaus, W., and Wallot, S. (2021). What the eyes reveal about (reading) poetry. *Poetics*. doi: 10.1016/j.poetic.2020.101526 (in press).

Mohammad, S. (2011). "From once upon a time to happily ever after: tracking emotions in novels and fairy tales," in *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (Portland, OR), 105–114.

Mohseni, M., Gast, V., and Redies, C. (2020). Comparative computational analysis of global structure in canonical, non-canonical and non-literary texts. *arXiv* 2008.10906.

Nascimento, S. M., Linhares, J. M., Montagner, C., João, C. A., Amano, K., Alfaro, C., et al. (2017). The colors of paintings and viewers' preferences. *Vision Res.* 130, 76–84. doi: 10.1016/j.visres.2016.11.006

O'Brien, B. A., Wallot, S., Haussmann, A., and Kloos, H. (2013). Using complexity metrics to assess silent reading fluency: a cross-sectional study comparing oral and silent reading. *Sci. Stud. Read.* 18, 235–254. doi: 10.1080/10888438.2013.862248

Oświecimka, P., Kwapień, J., and Drozdz, S. (2006). Wavelet versus detrended fluctuation analysis of multifractal structures. *Phys. Rev. E* 74:016103. doi: 10.1103/PhysRevE.74.016103

Palmer, S. E., Schloss, K. B., and Sammartino, J. (2013). Visual aesthetics and human preference. *Annu. Rev. Psychol.* 64, 77–107. doi: 10.1146/annurev-psych-120710-100504

Peng, C. K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., and Goldberger, A. L. (1994). Mosaic organization of DNA nucleotides. *Phys. Rev. E* 49, 1685–1689. doi: 10.1103/PhysRevE.49.1685

Pennington, J., Socher, R., and Manning, C. (2014). "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha), 1532–1543. doi: 10.3115/v1/D14-1162

Perani, D., Cappa, S., Schnur, T., Tettamanti, M., Collina, S., Rosa, M., et al. (1999). The neural correlates of verb and noun processing. A PET study. *Brain* 122, 2337–2344. doi: 10.1093/brain/122.12.2337

Petersen, S. E. (2007). *Natural language processing tools for reading level assessment and text simplification for bilingual education* (Ph.D. thesis), University of Washington, Seattle, WA, United States.

Reagan, A., Mitchell, L., Kiley, D., Danforth, C., and Dodds, P. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Sci.* 5:31. doi: 10.1140/epjds/s13688-016-0093-1

Redies, C. (2015). Combining universal beauty and cultural context in a unifying model of visual aesthetic experience. *Front. Hum. Neurosci.* 9:218. doi: 10.3389/fnhum.2015.00218

Redies, C., Amirshahi, S. A., Koch, M., and Denzler, J. (2012). "PHOG-derived aesthetic measures applied to color photographs of artworks, natural scenes and objects," in *ECCV 2012 Ws/Demos, Part I, Lecture Notes in Computer Science*, Vol. 7583, eds A. Fusiello, V. Murino, and R. Cucchiara (Berlin: Springer-Verlag), 522–531. doi: 10.1007/978-3-642-33863-2_54

Redies, C., and Brachmann, A. (2017). Statistical image properties in large subsets of traditional art, bad art, and abstract art. *Front. Neurosci.* 11:593. doi: 10.3389/fnins.2017.00593

Redies, C., Brachmann, A., and Wagemans, J. (2017). High entropy of edge orientations characterizes visual artworks from diverse cultural backgrounds. *Vision Res.* 133, 130–144. doi: 10.1016/j.visres.2017.02.004

Redies, C., Hasenstein, J., and Denzler, J. (2007). Fractal-like image statistics in visual art: similarity to natural scenes. *Spat. Vision* 21, 137–148. doi: 10.1163/156856807782753921

Riedl, M., and Biemann, C. (2012). Text segmentation with topic models. *J. Lang. Technol. Comput. Linguist.* 27, 13–24. doi: 10.1145/1645953.1646170

Roeske, T. C., Kelty-Stephen, D., and Wallot, S. (2018). Multifractal analysis reveals music-like dynamic structure in songbird rhythms. *Sci. Rep.* 8:4570. doi: 10.1038/s41598-018-22933-2

Samuels, S. J. (1994). "Toward a theory of automatic information processing in reading, revisited," in *Theoretical Models and Processes of Reading*, eds R. B. Ruddell, M. R. Ruddell, and H. Singer (Newark, DE: International Reading Association), 816–837.

Sanyal, S., Banerjee, A., Patranabis, A., Banerjee, K., Sengupta, R., and Ghosh, D. (2016). A study on improvisation in a musical performance using multifractal detrended cross correlation analysis. *Phys. A Stat. Mech. Appl.* 462, 67–83. doi: 10.1016/j.physa.2016.06.013

Scott, S. K. (2006). Language processing: the neural basis of nouns and verbs. *Curr. Biol.* 16, R295–R296. doi: 10.1016/j.cub.2006.03.042

Seifart, F., Strunk, J., Danielsen, S., Hartmann, I., Pakendorf, B., Wichmann, S., et al. (2018). Nouns slow down speech across structurally and culturally diverse languages. *Proc. Natl. Acad. Sci. U.S.A.* 115, 5720–5725. doi: 10.1073/pnas.1800708115

Shapiro, K. A., Moo, L. R., and Caramazza, A. (2006). Cortical signatures of noun and verb production. *Proc. Natl. Acad. Sci. U.S.A.* 103, 1644–1649. doi: 10.1073/pnas.0504142103

Simonton, D. K. (1990). Lexical choices and aesthetic success: a computer content analysis of 154 shakespeare sonnets. *Comput. Hum.* 24, 251–264.

Smith, C. (2003). *Modes of Discourse. The Local Structure of Texts*. Cambridge: Cambridge University Press.

Sudre, G., Pomerleau, D., Palatucci, M., Wehbe, L., Fyshe, A., Salmelin, R., et al. (2012). Tracking neural coding of perceptual and semantic features of concrete nouns. *Neuroimage* 62, 451–463. doi: 10.1016/j.neuroimage.2012.04.048

Taylor, R., Spehar, B., Hagerhall, C., and Van Donkelaar, P. (2011). Perceptual and physiological responses to Jackson Pollock's fractals. *Front. Hum. Neurosci.* 5:60. doi: 10.3389/fnhum.2011.00060

Thissen, B. A. K., Menninghaus, W., and Schlotz, W. (2018). Measuring optimal reading experiences: the reading flow short scale. *Front. Psychol.* 9:2542. doi: 10.3389/fpsyg.2018.02542

Tiffin-Richards, S. P., and Schroeder, S. (2015). The component processes of reading in comprehension in adolescents. *Learn. Indiv. Differ.* 42, 1–9. doi: 10.1016/j.lindif.2015.07.016

Tiffin-Richards, S. P., and Schroeder, S. (2018). The development of wrap-up processes in text reading: a study of children's eye movements. *J. Exp. Psychol.* 44, 1051–1063. doi: 10.1037/xlm0000506

Tolhurst, D. J., Tadmor, Y., and Chao, T. (1992). Amplitude spectra of natural images. *Ophthal. Physiol. Opt.* 12, 229–232. doi: 10.1111/j.1475-1313.1992.tb00296.x

Tötösy de Zepetnek, S. (1994). Toward a theory of cumulative canon formation: readership in English Canada. *Mosaic* 27, 107–119.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology–Volume 1, NAACL '03* (Edmonton, AB: Association for Computational Linguistics), 173–180. doi: 10.3115/1073445.1073478

Tyler, L. K., Bright, P., Fletcher, P., and Stamatakis, E. (2004). Neural processing of nouns and verbs: the role of inflectional morphology. *Neuropsychologia* 42, 512–523. doi: 10.1016/j.neuropsychologia.2003.10.001

Underwood, T., and Sellers, J. (2016). The long durée of literary prestige. *Mod. Lang. Q.* 77, 321–344. doi: 10.1215/00267929-3570634

van Cranenburgh, A., and Bod, R. (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* (Valencia: Association for Computational Linguistics), 1228–1238. doi: 10.18653/v1/E17-1115

van Cranenburgh, A., and Koolen, C. (2015). "Identifying literary texts with bigrams," in *Proceedings of the Fourth Workshop on Computational Linguistics for Literature* (Denver, CO: Association for Computational Linguistics), 58–67. doi: 10.3115/v1/W15-0707

van Cranenburgh, A., van Dalen-Oskam, K., and van Zundert, J. (2019). Vector space explorations of literary language. *Lang. Resour. Eval.* 53, 625–650. doi: 10.1007/s10579-018-09442-4

Vaughan-Evans, A., Trefor, R., Jones, L., Lynch, P., Jones, M. W., and Thierry, G. (2016). Implicit detection of poetic harmony by the naïve brain. *Front. Psychol.* 7:1859. doi: 10.3389/fpsyg.2016.01859

Verhuizen, N. J., Crocker, M. W., and Brower, H. (2019). Expectation-based comprehension: modeling the interaction of world knowledge and linguistic experience. *Discour. Process.* 56, 229–255. doi: 10.1080/0163853X.2018.1448677

Vieira, D. S., Picoli, S., and Mendes, R. S. (2018). Robustness of sentence length measures in written texts. *Phys. A Stat. Mech. Appl.* 506, 749–754. doi: 10.1016/j.physa.2018.04.104

Voss, R., and Clarke, J. (1975). '1/f' noise in music and speech. *Nature* 258, 317–318. doi: 10.1038/258317a0

Wallot, S., Hollis, G., and van Rooij, M. (2013). Connected text reading and differences in text reading fluency in adult readers. *PLoS ONE* 8:e71914. doi: 10.1371/journal.pone.0071914

Wallot, S., O'Brian, B. A., Haussmann, A., Kloos, H., and Lyby, M. S. (2014). The role of reading time complexity and reading speed in text comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* 40, 1745–1765. doi: 10.1037/xlm0000030

Wu, D., Kendrick, K., Levitin, D., Chaoyi, L., and Dezhong, Y. (2015). Bach is the father of harmony: revealed by a 1/f fluctuation analysis across musical genres. *PLoS ONE* 10:e0142431. doi: 10.1371/journal.pone.0142431

Yang, T., Gu, C., and Yang, H. (2016). Long-range correlations in sentence series from a story of the stone. *PLoS ONE* 11:e0162423. doi: 10.1371/journal.pone.0162423

Yu, G. (2009). Lexical diversity in writing and speaking task performances. *Appl. Linguist.* 31, 236–259. doi: 10.1093/applin/amp024

Zareva, A., Schwanenflugel, P., and Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: variable sensitivity. *Stud. Sec. Lang. Acquisit.* 27, 567–595. doi: 10.1017/S0272263105050254

Zwaan, R. A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychon. Bull. Rev.* 23, 1028–1034. doi: 10.3758/s13423-015-0864-x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers

Supplementary Material

ALTERNATIVE GLOBAL MEASURES OF VARIABILITY AND SELF-SIMILARITY

1 ENTROPY-BASED METHODS

Entropy, which is related to variability, measures uncertainty or (ir)regularity of a state or phenomenon represented by a random variable. If X is a discrete random variable with a set of possible values $\{x_1, x_2, \cdots, x_n\}$ and a corresponding probability function $P(X) = \{P(x_i), P(x_2), \cdots, P(x_n)\}$, the entropy of X is defined as:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i)$$

Entropy is zero when the state is certain and it is highest when the all possibilities are equally likely to occur, i.e. when uncertainty is maximal. The basic formula of entropy or its extensions have been utilized for text analysis previously.

Rosso et al. (2009) applied statistical complexity and entropy quantifiers to a collection of poems and plays. Their analyses revealed that poems have a higher complexity than plays and Shakepeare's work is interestingly more homogeneous than that of his contemporaries and is exceptionally close to the average use of words in that time period. Chang et al. (2017) defined the information-based energy, combined from the relative temperature and information Shannon entropy, to quantify text complexity and an author's performance. Applying this method to texts of an English and an Chinese author, Shakespeare and Jin Yong, they showed that their more popular works have higher information-based energy. Hernández-Gómez et al. (2017) used an entropy-based method, called approximate entropy, to measure the degree of irregularity or randomness in a series. They applied this method to 14 different languages which belong to four linguistic families: Romance, Germanic, Slavic and Uralic. They showed that the languages exhibit different levels of irregularity which were similar for languages that belonged to the same family. The entropy of word distributions can also be informative for comparing different types of languages in term of word ordering. Montemurro and Zanette (2016) used entropy-based measures to show that word ordering is highly similar over several language families. Febres and Jaffe (2017) studied entropy and symbolic diversity of fictional texts of Nobel and non-Nobel laureates in English and Spanish. While they presented some results to show that there is a correlation between these global statistical properties and the quality of writing, they did not classify different groups of texts.

2 BOX COUNTING

There are several methods to measure fractality and the scaling behavior of structures. These methods typically represent measurements at different scales. Fractal analysis techniques have been widely applied to images (Wendt and Abry, 2007; Li et al., 2009; Wendt et al., 2009; Ji et al., 2013), including artworks (Taylor, 2002; Redies et al., 2007; Spehar et al., 2016). They are therefore of special interest for analyzing aesthetic phenomena.

One of the most widely used fractal analysis methods is box counting, which is mathematically straightforward and easy to apply. Given an object S, for a $\delta > 0$ the smallest possible number of

subsets with a diameter of at most δ, $N_\delta(S)$, which covers S, is found. For 1d objects, subsets are rulers and δ is their length. For 2d objects, subsets are boxes and δ is their area, and so forth. The growth ratio of $N_\delta(S)$, as $\delta \to 0$, reflects the degree of fractality of S. If $N_\delta(S)$ can be approximated by

$$N_\delta(S) \simeq c\delta^{D_B}$$

for a constant c, then D_B is called the box-counting dimension and shows how complex S is.

Mehri and Lashkari (2016) applied this method to seven famous text books and computed their degree of fractality by averaging the fractality degrees of word occurrences. The results revealed that all texts are fractal and their fractal dimensions differed slightly. Fractality patterns of series sometimes do not lend themselves to analysis with a single scaling measure. If different subsets of a series exhibit different types of scaling behavior, the series is multifractal. Chatzigeorgiou et al. (2017) used box counting to find the origin of multifractality in the word-length representation of texts in several Western languages. They showed that the long-range correlations in natural language are related to the clustering feature of long words, i.e. rare and often highly informative content words.

3 WAVELET-BASED METHODS

Fractal analysis methods based on wavelets are another family of techniques for studying scale-invariant properties of signals (Muzy et al., 1993; Wendt and Abry, 2007; Leonarduzzi et al., 2016). The wavelet transform (WT) is a method to analyze non-stationary signals. The WT of a signal X is defined as (Mallat, 1999):

$$T_\psi[X](a, t_0) = \frac{1}{a}\int_{-\infty}^{+\infty} X(t)\psi(\frac{t - t_0}{a})dt,$$

and it describes the content of X around a time parameter t_0 and a scale parameter a. ψ is the analyzing wavelet whose $n + 1$ first moments are zero, i.e. $\int_{\mathbb{R}} t^n\psi(t)dt = 0$, which makes the WT insensitive to possible polynomial trends of order n in the signal, something which is necessary for multifractal analysis (Muzy et al., 1994; Arneodo et al., 1995). The WT modulus maxima (WTMM) is a well-known method for analyzing multifractality and it is based on the WT coefficients. WTMM is defined by the local maxima $\mathcal{L}(a)$ of $|T_\psi[X](a, t)|$ according to a given scale a. Then the following partition function is defined:

$$Z(q, a) = \sum_{l\in\mathcal{L}(a)} |T_\psi[X](a, t)|^q \sim a^{\tau(q)}$$

If the signal is monofractal, $\tau(q)$ is independent of q. For multifractal signals, the scaling behavior cannot be explained with one value, so, $\tau(q)$ changes for different values of q. Based on WT and WTMM, other methods have been extended for discrete and multi-dimensional series (for example, see Wendt and Abry, 2007; Leonarduzzi et al., 2016). Although wavelet-based methods have been applied to a variety of fields, they have been rarely used in text processing. Leonarduzzi et al. (2017) applied the wavelet p-leader method to the sentence-length series of novels that were written either for young people or adults. The authors showed that the latter category is more diverse in terms of its degree of multifractality.

4 FRACTALITY AND CROSS-CORRELATION ANALYSIS

Fractal analysis can be extended to analyzing more than one series, in order to find relations between fractal behaviors of multiple series. Detrended Cross-Correlation Analysis (DCCA) (Podobnik and Stanley, 2008) and Multi-Fractal Detrended Cross-Correlation Analysis (MFDCCA) (Jiang and Zhou, 2011) are two methods for analyzing correlations between two series. Ghosh et al. (2019) applied MFDCCA, also known as MFDXA, to study correlations between two Tagore's poems, one written in Bengali and one in English. They found a nonlinear correlation between the poems. In a similar study, birdsong and human speech were compared by computing the mutual information decay of signals and it was concluded that the two vocal communication signals have similar dynamics (Sainburg et al., 2019).

REFERENCES

Arneodo, A., Bacry, E., and Muzy, J. (1995). The thermodynamics of fractals revisited with wavelets. *Physica A: Statistical Mechanics and Its Applications* 213, 232–275. doi:10.1016/0378-4371(94)00163-N

Chang, M.-C., Yang, A. C.-C., Stanley, H. E., and Peng, C.-K. (2017). Measuring information-based energy and temperature of literary texts. *Physica A: Statistical Mechanics and Its Applications* 468, 783–789. doi:10.1016/j.physa.2016.11.106

Chatzigeorgiou, M., Constantoudis, V., Diakonos, F., Karamanos, K., Papadimitriou, C., Kalimeri, M., et al. (2017). Multifractal correlations in natural language written texts: Effects of language family and long word statistics. *Physica A: Statistical Mechanics and Its Applications* 469, 173–182. doi:10.1016/j.physa.2016.11.028

Febres, G. and Jaffe, K. (2017). Quantifying structure differences in literature using symbolic diversity and entropy criteria. *Journal of Quantitative Linguistics* 24, 16–53. doi:10.1080/09296174.2016.1169847

Ghosh, D., Chakraborty, S., and Samanta, S. (2019). Study of translational effect in tagore's gitanjali using chaos based multifractal analysis technique. *Physica A: Statistical Mechanics and Its Applications* 523, 1343–1354. doi:10.1016/j.physa.2019.04.171

Hernández-Gómez, C., Basurto-Flores, R., Obregón-Quintana, B., and Guzmán-Vargas, L. (2017). Evaluating the irregularity of natural languages. *Entropy* 19, 521. doi:10.3390/e19100521

Ji, H., Yang, X., Ling, H., and Xu, Y. (2013). Wavelet domain multifractal analysis for static and dynamic texture classification. *IEEE Transactions on Image Processing* 22, 286–299. doi:10.1109/TIP.2012.2214040

Jiang, Z.-Q. and Zhou, W.-X. (2011). Multifractal detrending moving-average cross-correlation analysis. *Physical Review E* 84, 016106. doi:10.1103/PhysRevE.84.016106

Leonarduzzi, R., Abry, P., Jaffard, S., Wendt, H., Gournay, L., Kyriacopoulou, T., et al. (2017). P-leader multifractal analysis for text type identification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4661–4665. doi:10.1109/ICASSP.2017.7953040

Leonarduzzi, R., Wendt, H., Abry, P., Jaffard, S., Melot, C., Roux, S., et al. (2016). p-exponent and p-leaders, Part II: Multifractal analysis. Relations to detrended fluctuation analysis. *Physica A: Statistical Mechanics and Its Applications* 448, 319–339. doi:10.1016/j.physa.2015.12.035

Li, J., Du, Q., and Sun, C. (2009). An improved box-counting method for image fractal dimension estimation. *Pattern Recognition* 42, 2460–2469. doi:10.1016/j.patcog.2009.03.001

Mallat, S. (1999). *A Wavelet Tour of Signal Processing (2. ed.).* (Cambridge: Academic Press)

Mehri, A. and Lashkari, S. M. (2016). Power-law regularities in human language. *The European Physical Journal B* 89, 241. doi:10.1140/epjb/e2016-70423-9

Montemurro, M. A. and Zanette, D. H. (2016). Complexity and universality in the long-range order of words. In *Creativity and Universality in Language*, eds. M. Degli Esposti, E. G. Altmann, and F. Pachet (Cham: Springer). 27–41. doi:10.1007/978-3-319-24403-7_3

Muzy, J.-F., Bacry, E., and Arneodo, A. (1993). Multifractal formalism for fractal signals: The structure-function approach versus the wavelet-transform modulus-maxima method. *Physical Review E* 47, 875–884. doi:10.1103/PhysRevE.47.875

Muzy, J.-F., Bacry, E., and Arneodo, A. (1994). The multifractal formalism revisited with wavelets. *International Journal of Bifurcation and Chaos* 4, 245–302. doi:10.1142/S0218127494000204

Podobnik, B. and Stanley, H. E. (2008). Detrended cross-correlation analysis: a new method for analyzing two nonstationary time series. *Physical Review Letters* 100, 084102. doi:10.1103/PhysRevLett.100.084102

Redies, C., Hasenstein, J., and Denzler, J. (2007). Fractal-like image statistics in visual art: similarity to natural scenes. *Spatial Vision* 21, 137–148. doi:10.1163/156856807782753921

Rosso, O. A., Craig, H., and Moscato, P. (2009). Shakespeare and other english renaissance authors as characterized by information theory complexity quantifiers. *Physica A: Statistical Mechanics and Its Applications* 388, 916–926. doi:10.1016/j.physa.2008.11.018

Sainburg, T., Theilman, B., Thielk, M., and Gentner, T. Q. (2019). Parallels in the sequential organization of birdsong and human speech. *Nature Communications* 10, 3636. doi:10.1038/s41467-019-11605-y

Spehar, B., Walker, N., and Taylor, R. (2016). Taxonomy of individual variations in aesthetic responses to fractal patterns. *Frontiers in Human Neuroscience* 10, 350. doi:10.3389/fnhum.2016.00350

Taylor, R. (2002). Order in Pollock's chaos - computer analysis is helping to explain the appeal of Jackson Pollock's paintings. *Scientific American* 287, 116–121

Wendt, H. and Abry, P. (2007). Multifractality tests using bootstrapped wavelet leaders. *IEEE Transactions on Signal Processing* 55, 4811–4820. doi:10.1109/TSP.2007.896269

Wendt, H., Roux, S. G., Jaffard, S., and Abry, P. (2009). Wavelet leaders and bootstrap for multifractal analysis of images. *Signal Processing* 89, 1100–1114. doi:10.1016/j.sigpro.2008.12.015

SUPPLEMENTARY TABLES AND FIGURES

Table S1: List of texts in Jena Corpus of Expository and Fictional Prose (JEFP Corpus). Canonical texts were selected from the Corpus of Canonical Western Literature. Non-canonical texts were downloaded from www.smashwords.com, www.goodreads.com, www.feedbooks.com, or Project Gutenberg. Non-fictional texts were sampled from Project Gutenberg.

	Title	**Author(s)**	**Year of Publication**	**Category**
1	Little Dorrit	Charles Dickens	1857	Canonical
2	Oliver Twist	Charles Dickens	1839	Canonical
3	The Life and Adventures of Nicholas Nickleby	Charles Dickens	1839	Canonical
4	The Mystery of Edwin Drood	Charles Dickens	1870	Canonical
5	The Pickwick Papers	Charles Dickens	1836	Canonical
6	Jane Eyre	Charlotte Bronte	1847	Canonical
7	Villette	Charlotte Bronte	1853	Canonical
8	Cranford	Elizabeth Gaskell	1853	Canonical
9	Mary Barton	Elizabeth Gaskell	1848	Canonical
10	North and South	Elizabeth Gaskell	1854	Canonical
11	Agnes Grey	Anne Bronte	1847	Canonical
12	Adam Bede	George Eliot	1859	Canonical
13	Daniel Deronda	George Eliot	1876	Canonical
14	Middlemarch	George Eliot	1872	Canonical
15	Silas Marner	George Eliot	1861	Canonical
16	The Mill on the Floss	George Eliot	1860	Canonical
17	Emma	Jane Austen	1815	Canonical
18	Mansfield Park	Jane Austen	1814	Canonical
19	Persuasion	Jane Austen	1818	Canonical
20	Pride and Prejudice	Jane Austen	1813	Canonical
21	The Picture of Dorian Gray	Oscar Wilde	1890	Canonical
22	The Tenant of Wildfell Hall	Anne Bronte	1848	Canonical
23	Sartor Resartus	Thomas Carlyle	1834	Canonical
24	Old Mortality	Walter Scott	1816	Canonical
25	Redgauntlet	Walter Scott	1824	Canonical
26	The Heart of Midlothian	Walter Scott	1818	Canonical
27	Waverley	Walter Scott	1814	Canonical
28	No Name	Wilkie Collins	1862	Canonical
29	The Moonstone	Wilkie Collins	1868	Canonical
30	The Woman in White	Wilkie Collins	1859	Canonical
31	The History of Henry Esmond	William Makepeace Thackeray	1852	Canonical
32	Vanity Fair	William Makepeace Thackeray	1847	Canonical
33	Dracula	Bram Stoker	1897	Canonical
34	The Well at the World's end	William Morris	1896	Canonical
35	The Narrative of Arthur Gordon Pym	Edgar Allan Poe	1838	Canonical
36	The Ambassadors	Henry James	1903	Canonical
37	The Awkward Age	Henry James	1899	Canonical
38	The Bostonians	Henry James	1886	Canonical
39	The Golden Bowl	Henry James	1904	Canonical
40	The Portrait of a Lady	Henry James	1881	Canonical
41	The Wings of Dove	Henry James	1902	Canonical

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
42	Moby Dick	Herman Melville	1851	Canonical
43	The Deerslayers	James Fenimore Cooper	1841	Canonical
44	A Christmas Carol	Charles Dickens	1843	Canonical
45	Little Women	Louisa May Alcott	1868	Canonical
46	Puddnhead Wilson	Mark Twain	1893	Canonical
47	The Adventures of Finn	Mark Twain	1884	Canonical
48	The Mysterious Stranger	Mark Twain	1916	Canonical
49	The Marble Faun	Nathaniel Hawthorne	1859	Canonical
50	The Scarlet Letter	Nathaniel Hawthorne	1850	Canonical
51	The Education of Adams	Henry Adams	1907	Canonical
52	Walden	Henry David Thoreau	1854	Canonical
53	A Conneticut Yankee in King Arthurs	Mark Twain	1889	Canonical
54	Babbitt	Sinclair Lewis	1922	Canonical
55	A Tale of Two Cities	Charles Dickens	1859	Canonical
56	Sister Carrie	Theodore Dreiser	1900	Canonical
57	My Antonia	Willa Cather	1918	Canonical
58	The Old Wives Tale	Arnold Bennett	1908	Canonical
59	Portrait of the Artist as a Young Man	James Joyce	1916	Canonical
60	Ulysses	James Joyce	1922	Canonical
61	Lord Jim	Joseph Conrad	1900	Canonical
62	Nostromo	Joseph Conrad	1904	Canonical
63	The Secret Agent	Joseph Conrad	1907	Canonical
64	Under Western Eyes	Joseph Conrad	1911	Canonical
65	Victory: An Island Tale	Joseph Conrad	1915	Canonical
66	Bleak House	Charles Dickens	1853	Canonical
67	The Rainbow	Lawrence D.H	1915	Canonical
68	Women in Love	Lawrence D.H	1920	Canonical
69	Kim	Rudyard Kipling	1901	Canonical
70	Puck of Pooks Hill	Rudyard Kipling	1906	Canonical
71	Jude the Obscure	Thomas Hardy	1895	Canonical
72	Tess of the dUrbervilles	Thomas Hardy	1891	Canonical
73	The Mayor of Casterbridge	Thomas Hardy	1886	Canonical
74	The Return of the Native	Thomas Hardy	1878	Canonical
75	David Copperfield	Charles Dickens	1850	Canonical
76	Great Expectations	Charles Dickens	1860	Canonical
77	Hard Times	Charles Dickens	1854	Canonical
78	The Face in the Abyss	Abraham Merritt	1923	Non-Canonical
79	A Prisoner in Fairyland	Algernon Blackwood	1913	Non-Canonical
80	The Centaur	Algernon Blackwood	1911	Non-Canonical
81	Ruth Fielding at the War Front	Alice B. Emerson	1918	Non-Canonical
82	The International Spy	Allen Upward	1904	Non-Canonical
83	A Texas Matchmaker	Andy Adams	1904	Non-Canonical
84	The Filigree Ball	Anna Katharine Green	1903	Non-Canonical
85	Looking Further Backward	Arthur Dudley Vinton	1890	Non-Canonical
86	The Hill Of Dreams	Arthur Machen	1907	Non-Canonical
87	The Elusive Pimpernel	Baroness Emma Orczy	1908	Non-Canonical
88	The Gloved Hand	Burton E. Stevenson	1913	Non-Canonical
89	Jean of the Lazy A	B.M . Bower	1915	Non-Canonical
90	Wieland : or , The Transformation	Charles Brockden Brown	1798	Non-Canonical

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
91	The Great Quest	Charles Hawes	1921	Non-Canonical
92	The Filibusters	Charles John Cutcliffe Wright Hyne	1900	Non-Canonical
93	Bar-20 Days	Clarence E. Mulford	1911	Non-Canonical
94	Wunpost	Dane Coolidge	1920	Non-Canonical
95	The Girl of the Golden West	David Belasco	1911	Non-Canonical
96	Love Insurance	Earl Derr Biggers	1914	Non-Canonical
97	The Wouldbegoods	Edith Nesbit	1899	Non-Canonical
98	Wet Magic	Edith Nesbit	1913	Non-Canonical
99	Philip Dru : Administrator	Edward Mandell House	1912	Non-Canonical
100	An Amiable Charlatan	Edward Phillips Oppenheim	1916	Non-Canonical
101	The Double Traitor	Edward Phillips Oppenheim	1915	Non-Canonical
102	The Zeppelin 's Passenger	Edward Phillips Oppenheim	1918	Non-Canonical
103	The People of the Ruins	Edward Shanks	1920	Non-Canonical
104	The Honor of the Name	mile Gaboriau	1891	Non-Canonical
105	Kai Lung's Golden Hours	Ernest Bramah Smith	1922	Non-Canonical
106	The Riddle of the Sands	Erskine Childers	1903	Non-Canonical
107	The Missourian	Eugene Percy Lyle	1905	Non-Canonical
108	Privy Seal	Ford Madox Ford	1907	Non-Canonical
109	The Ivory Snuff Box	Frederic Arnold Kummer	1912	Non-Canonical
110	The Afterglow	George Allan England	1913	Non-Canonical
111	The Flying Legion	George Allan England	1920	Non-Canonical
112	West Wind Drift	George Barr McCutcheon	1920	Non-Canonical
113	Peter the Brazen	George F. Worts	1919	Non-Canonical
114	Olga Romanoff or , The Syren of the Skies	George Griffith	1894	Non-Canonical
115	The Princess and Curdie	George MacDonald	1883	Non-Canonical
116	The Adventures of Don Lavington	George Manville Fenn	1896	Non-Canonical
117	A Voyage to the Moon	George Tucker	1827	Non-Canonical
118	Claim Number One	George W. Ogden	1922	Non-Canonical
119	The Flockmaster of Poison Creek	George W. Ogden	1921	Non-Canonical
120	Trilby	George du Maurier	1894	Non-Canonical
121	Rose O'Paradise	Grace Miller White	1915	Non-Canonical
122	Condemned as a Nihilist	G. A. Henty	1893	Non-Canonical
123	Man on the Box	Harold MacGrath	1904	Non-Canonical
124	The Puppet Crown	Harold MacGrath	1901	Non-Canonical
125	The Blind Spot	Homer Eon Flint	1921	Non-Canonical
126	Men of Iron	Howard Pyle	1891	Non-Canonical
127	The Dark House	Ida Alexa Ross Wylie	1922	Non-Canonical
128	The Daughter of Brahma	Ida Alexa Ross Wylie	1912	Non-Canonical
129	Towards Morning	Ida Alexa Ross Wylie	1918	Non-Canonical
130	Jurgen : A Comedy of Justice	James Branch Cabell	1919	Non-Canonical
131	A Strange Manuscript Found in a Copper Cylinder	James De Mille	1888	Non-Canonical
132	Lost in the Fog	James De Mille	1870	Non-Canonical
133	Varney the Vampire	James Malcom Rymer	1847	Non-Canonical
134	The Danger Trail	James Oliver Curwood	1910	Non-Canonical
135	The Lost Stradivarius	John Meade Falkner	1895	Non-Canonical
136	The Nebuly Coat	John Meade Falkner	1903	Non-Canonical
137	The Weapons of Mystery	Joseph Hocking	1890	Non-Canonical
138	The Chestermarke Instinct	Joseph Smith Fletcher	1921	Non-Canonical

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
139	Afloat On The Flood	Lawrence J. Leslie	1915	Non-Canonical
140	Diane of the Green Van	Leona Dalrymple	1914	Non-Canonical
141	Don Rodriguez : Chronicles of Shadow Valley	Lord Dunsany	1922	Non-Canonical
142	The Treasure Trail	Marah Ellis Ryan	1918	Non-Canonical
143	Mizora : A Prophecy	Mary E. Bradley	1889	Non-Canonical
144	Dangerous Days	Mary Roberts Rinehart	1919	Non-Canonical
145	The Blue Germ	Maurice Nicoll	1918	Non-Canonical
146	The Night Horseman	Max Brand	1920	Non-Canonical
147	The Sleuth of St. James 's Square	Melville Davisson Post	1920	Non-Canonical
148	Across the Zodiac	Percy Greg	1880	Non-Canonical
149	Bardelys the Magnificent	Rafael Sabatini	1905	Non-Canonical
150	Soldiers of Fortune	Richard Harding Davis	1897	Non-Canonical
151	The Beetle	Richard Marsh	1897	Non-Canonical
152	The Triumphs of Eugne Valmont	Robert Barr	1906	Non-Canonical
153	Dawn of All	Robert Hugh Benson	1911	Non-Canonical
154	Erling the Bold	Robert Michael Ballantyne	1869	Non-Canonical
155	The Dog Crusoe and His Master	Robert Michael Ballantyne	1894	Non-Canonical
156	Ailsa Paige	Robert William Chambers	1910	Non-Canonical
157	In Search of the Unknown	Robert William Chambers	1904	Non-Canonical
158	In the Quarter	Robert William Chambers	1894	Non-Canonical
159	Under the Ocean to the South Pole	Roy Rockwood	1907	Non-Canonical
160	Erewhon , or Over The Range	Samuel Butler	1910	Non-Canonical
161	The road to Frontenac	Samuel Merwin	1901	Non-Canonical
162	Brood of the Witch-Queen	Sax Rohmer	1918	Non-Canonical
163	The Revolt of Man	Walter Besant	1882	Non-Canonical
164	The Brass Bottle	Thomas Anstey Guthrie	1900	Non-Canonical
165	The Stray Lamb	Thorne Smith	1929	Non-Canonical
166	The Doomsman	Van Tassel Sutphen	1906	Non-Canonical
167	The Song of the Lark	Willa Cather	1915	Non-Canonical
168	The Old Tobacco Shop	William Bowen	1921	Non-Canonical
169	The Boats of the 'Glen-Carrig '	William Hope Hodgson	1907	Non-Canonical
170	Hushed Up!	William Le Queux	1911	Non-Canonical
171	The Border Legion	Zane Grey	1916	Non-Canonical
172	The Desert of Wheat	Zane Grey	1919	Non-Canonical
173	Scottish Cathedrals and Abbeys	Dugald Butler	1901	Non-Fictional
174	A Text-Book of the History of Architecture: Seventh Edition, revised	A. D. F. Hamlin	1896	Non-Fictional
175	Some Account of Gothic Architecture in Spain	George Edmund Street	1865	Non-Fictional
176	Japanese Homes and Their Surroundings	Edward Sylvester Morse	1885	Non-Fictional
177	The Architecture of Provence and the Riviera	David MacGibbon	1888	Non-Fictional
178	Historic Ornament, Vol. 2: Treatise on decorative art and architectural ornament	James Ward	1897	Non-Fictional
179	Military Architecture in England During the Middle Ages	A. Hamilton Thompson	1912	Non-Fictional
180	How to Study Architecture	Charles H. Caffin	1917	Non-Fictional
181	Cakes & Ale: A Dissertation on Banquets Interspersed with Various Recipes, More or Less Original, and anecdotes, mainly veracious	Edward Spencer	1897	Non-Fictional
182	Food and Flavor: A Gastronomic Guide to Health and Good Living	Henry T. Finck	1913	Non-Fictional

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
183	A Concise Dictionary of Middle English from A.D. 1150 to 1580	A. L. Mayhew, Walter William Skeat	1888	Non-Fictional
184	A Dictionary of Slang, Cant, and Vulgar Words: Used at the Present Day in the Streets of London	John Camden Hotten	1860	Non-Fictional
185	The Devil's Dictionary	Ambrose Bierce	1906	Non-Fictional
186	The Encyclopedia Britannica Vol. 1	University of Cambridge	1910	Non-Fictional
187	The Encyclopedia Britannica Vol. 2	University of Cambridge	1910	Non-Fictional
188	Glossary of Chess terms	Gregory Zorzos	2017	Non-Fictional
189	Through the Brazilian Wilderness	Roosevelt	1914	Non-Fictional
190	Gold, Sport, and Coffee Planting in Mysore	Robert H. Elliot	1898	Non-Fictional
191	The Economic Aspect of Geology	C. K. Leith	1921	Non-Fictional
192	The Shores of the Adriatic: The Austrian Side, The Kustenlande, Istria, and Dalmatia	F. Hamilton Jackson	1906	Non-Fictional
193	Island Life; Or, The Phenomena and Causes of Insular Faunas and Floras	Alfred Russel Wallace	1880	Non-Fictional
194	Sea and Sardinia	D. H. Lawrence	1921	Non-Fictional
195	Sketches from the Subject and Neighbour Lands of Venice	Edward A. Freeman	1881	Non-Fictional
196	The Elements of Geology	William Harmon Norton	1905	Non-Fictional
197	The Principles of Stratigraphical Geology	J. E. Marr	1898	Non-Fictional
198	Fragments of Earth Lore: Sketches & Addresses Geological and Geographical	James Geikie	1893	Non-Fictional
199	Earth Features and Their Meaning: An Introduction to Geology for the Student and the General Reader	William Herbert Hobbs	1912	Non-Fictional
200	The Common Law	Oliver Wendell Holmes	1881	Non-Fictional
201	Babylonian and Assyrian Laws, Contracts and Letters	C. H. W. Johns	1904	Non-Fictional
202	Putnam's Handy Law Book for the Layman	Albert Sidney Bolles	1921	Non-Fictional
203	Marriage and Divorce Laws of the World	Hyacinthe Ringrose	1911	Non-Fictional
204	The Law and the Poor	Edward Abbott Parry	1914	Non-Fictional
205	International Law. A Treatise. Vol. 1: Peace. Second Edition	L. Oppenheim	1905	Non-Fictional
206	International Law. A Treatise. Vol. 2: War and Neutrality. Second Edition	L. Oppenheim	1905	Non-Fictional
207	International Law	George Grafton Wilson, George Fox Tucker	1901	Non-Fictional
208	The Criminal Prosecution and Capital Punishment of Animals	E. P. Evans	1906	Non-Fictional
209	The English Constitution	Walter Bagehot	1867	Non-Fictional
210	The Law of the Sea: A Manual of the Principles of Admiralty Law for Students, Mariners, and Ship Operators	George L. Canfield, George W. Dalzell, J. Y. Brinton	1921	Non-Fictional
211	Woman and the Republic: A Survey of the Woman-Suffrage Movement in the United States and a Discussion of the Claims and Arguments of Its Foremost Advocates	Helen Kendrick Johnson	1897	Non-Fictional
212	The American Judiciary	Simeon E. Baldwin	1905	Non-Fictional
213	The Story of Evolution	Joseph McCabe	1912	Non-Fictional
214	A Practical Physiology: A Text-Book for Higher Schools	Albert F. Blaisdell	1897	Non-Fictional
215	Our Vanishing Wild Life: Its Extermination and Preservation	William T. Hornaday	1913	Non-Fictional
216	Amusements in Mathematics	Henry Ernest Dudeney	1917	Non-Fictional
217	On the Genesis of Species	St. George Jackson Mivart	1871	Non-Fictional
218	An Elementary Study of Chemistry	William McPherson, William Edwards Henderson	1906	Non-Fictional
219	Great Astronomers	Robert S. Ball	1895	Non-Fictional
220	Evolution, Old & New	Samuel Butler	1879	Non-Fictional

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
221	Darwin, and After Darwin, Vol. 1: An Exposition of the Darwinian Theory and a Discussion of Post-Darwinian Questions	George John Romanes	1982	Non-Fictional
222	Creative Evolution	Henri Bergson	1907	Non-Fictional
223	Myths and Marvels of Astronomy	Richard A. Proctor	1877	Non-Fictional
224	A Popular History of Astronomy During the Nineteenth Century: Fourth Edition	Agnes M. Clerke	1887	Non-Fictional
225	Pioneers of Science	Oliver Lodge	1893	Non-Fictional
226	A Text-Book of Astronomy	George C. Comstock	1901	Non-Fictional
227	Astronomical Myths: Based on Flammarions's "History of the Heavens"	Camille Flammarion, J. F. Blake	1877	Non-Fictional
228	Darwin, and After Darwin, Vol. 2: Post-Darwinian Questions, Heredity and Utility	George John Romanes	1892	Non-Fictional
229	Astronomy: The Science of the Heavenly Bodies	David P. Todd	1922	Non-Fictional
230	The Foundations of Science: Science and Hypothesis, The Value of Science, Science and Method	Henri Poincaré	1913	Non-Fictional
231	A Civic Biology, Presented in Problems	George W. Hunter	1914	Non-Fictional
232	Physics	Willis E. Tower, Charles M. Turton, Charles H. Smith, Thomas D. Cope	1920	Non-Fictional
233	A Century of Science, and Other Essays	John Fiske	1899	Non-Fictional
234	Side-Lights on Astronomy and Kindred Fields of Popular Science	Simon Newcomb	1906	Non-Fictional
235	Elementary Zoology, Second Edition	Vernon L. Kellogg	1901	Non-Fictional
236	Experiments on Animals	Stephen Paget	1888	Non-Fictional
237	The Sea-beach at Ebb-tide: A Guide to the Study of the Seaweeds and the Lower Animal Life Found Between Tide-marks	Augusta Foote Arnold	1901	Non-Fictional
238	The Making of Species	Douglas Dewar, Frank Finn	1909	Non-Fictional
239	The Science and Philosophy of the Organism	Hans Driesch	1908	Non-Fictional
240	Problems of Genetics	William Bateson	1913	Non-Fictional
241	The Organism as a Whole, from a Physicochemical Viewpoint	Jacques Loeb	1916	Non-Fictional
242	A Guide to the Study of Fishes, Vol. 1	David Starr Jordan	1905	Non-Fictional
243	Evolution: Its nature, its evidence, and its relation to religious thought	Joseph LeConte	1888	Non-Fictional
244	The Races of Man: An Outline of Anthropology and Ethnography	Joseph Deniker	1900	Non-Fictional
245	Physiology: The Science of the Body	Ernest G. Martin	1922	Non-Fictional
246	Observations of a Naturalist in the Pacific Between 1896 and 1899, Vol. 1	H. B. Guppy	1903	Non-Fictional
247	Animal Life and Intelligence	C. Lloyd Morgan	1890	Non-Fictional
248	A Guide to the Study of Fishes, Vol. 2	David Starr Jordan	1905	Non-Fictional
249	Stargazing: Past and Present	Norman Lockyer	1878	Non-Fictional
250	Observations of a Naturalist in the Pacific Between 1896 and 1899, Vol. 2	H. B. Guppy	1903	Non-Fictional
251	Regeneration	Thomas Hunt Morgan	1901	Non-Fictional
252	Telescopic Work for Starlight Evenings	William F. Denning	1891	Non-Fictional
253	The Logic of Chance, 3rd edition	John Venn	1888	Non-Fictional
254	Biology and Its Makers: With Portraits and Other Illustrations	William A. Locy	1908	Non-Fictional
255	The Crayfish: An Introduction to the Study of Zoology	Thomas Henry Huxley	1880	Non-Fictional

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
256	History of Botany (1530-1860)	Julius Sachs	1875	Non-Fictional
257	The Universal Kinship	J. Howard Moore	1906	Non-Fictional
258	The philosophy of biology	James Johnstone	1914	Non-Fictional
259	Hygienic Physiology: with Special Reference to the Use of Alcoholic Drinks and Narcotics	Joel Dorman Steele	1884	Non-Fictional
260	Species and Varieties, Their Origin by Mutation	Hugo de Vries	1905	Non-Fictional
261	The Naturalist in La Plata	W. H. Hudson	1892	Non-Fictional
262	Studies in the Psychology of Sex, Vol. 1	Havelock Ellis	1900	Non-Fictional
263	Studies in the Psychology of Sex, Vol. 2	Havelock Ellis	1900	Non-Fictional
264	The Mind of the Child, Part II: The Development of the Intellect	William T. Preyer	1888	Non-Fictional
265	The Measurement of Intelligence	Lewis M. Terman	1916	Non-Fictional
266	Human Traits and their Social Significance	Irwin Edman	1919	Non-Fictional
267	Chapters in the History of the Insane in the British Isles	Daniel Hack Tuke	1882	Non-Fictional
268	Human Personality and Its Survival of Bodily Death	F. W. H. Myers	1903	Non-Fictional
269	Mysterious Psychic Forces: An Account of the Author's Investigations in Psychical Research, Together with Those of Other European Savants	Camille Flammarion	1907	Non-Fictional
270	The Group Mind: A Sketch of the Principles of Collective Psychology	William McDougall	1920	Non-Fictional
271	On the State of Lunacy and the Legal Provision for the Insane: With Observations on the Construction and Organization of Asylums	J. T. Arlidge	1859	Non-Fictional
272	The Criminal	Havelock Ellis	1890	Non-Fictional
273	Fact and Fable in Psychology	Joseph Jastrow	1900	Non-Fictional
274	Mental Evolution in Man: Origin of Human Faculty	George John Romanes	1888	Non-Fictional
275	A Beginner's Psychology	Edward Bradford Titchener	1915	Non-Fictional
276	Mental diseases: A Public Health Problem	James Vance May	1922	Non-Fictional
277	The Law of Psychic Phenomena	Thomson Jay Hudson	1893	Non-Fictional
278	Psychology: Briefer Course	William James	1892	Non-Fictional
279	The Principles of Psychology, Vol. 1	William James	1890	Non-Fictional
280	The Principles of Psychology, Vol. 2	William James	1890	Non-Fictional
281	Sex & Character	Otto Weininger	1906	Non-Fictional
282	Youth: Its Education, Regimen, and Hygiene	G. Stanley Hall	1906	Non-Fictional
283	Ten Thousand Dreams Interpreted; Or, What's in a Dream: A Scientific and Practical Exposition	Gustavus Hindman Miller	1906	Non-Fictional
284	Browning as a Philosophical and Religious Teacher	Henry Jones	1891	Non-Fictional
285	The Life of Reason: The Phases of Human Progress	George Santayana	1905	Non-Fictional
286	An Introduction to Philosophy	George Stuart Fullerton	1906	Non-Fictional
287	The Approach to Philosophy	Ralph Barton Perry	1905	Non-Fictional
288	The Will to Believe, and Other Essays in Popular Philosophy	William James	1896	Non-Fictional
289	Christianity and Greek Philosophy	B. F. Cocker	1870	Non-Fictional
290	A History of Mediaeval Jewish Philosophy	Isaac Husik	1916	Non-Fictional
291	The Mediaeval Mind (Vol. 1 of 2): A History of the Development of Thought and Emotion in the Middle Ages	Henry Osborn Taylor	1911	Non-Fictional
292	The Mediaeval Mind (Vol. 2 of 2): A History of the Development of Thought and Emotion in the Middle Ages	Henry Osborn Taylor	1911	Non-Fictional
293	The Philosophy of Friedrich Nietzsche	H. L. Mencken	1908	Non-Fictional
294	Philosophical Studies	G. E. Moore	1883	Non-Fictional
295	What Nietzsche Taught	Willard Huntington Wright	1915	Non-Fictional
296	The Greek Philosophers, Vol. 1	Alfred William Benn	1882	Non-Fictional

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
297	The Greek Philosophers, Vol. 2	Alfred William Benn	1882	Non-Fictional
298	An Ethical Philosophy of Life Presented in Its Main Outlines	Felix Adler	1918	Non-Fictional
299	A Beginner's History of Philosophy, Vol. 1	Herbert Ernest Cushman	1910	Non-Fictional
300	Towards the Great Peace	Ralph Adams Cram	1922	Non-Fictional
301	Society: Its Origin and Development	Henry K. Rowe	1916	Non-Fictional
302	Criminal Man, According to the Classification of Cesare Lombroso	Gina Lombroso	1880	Non-Fictional
303	The Challenge of the Country: A Study of Country Life Opportunity	George Walter Fiske	1912	Non-Fictional
304	Criminal Sociology	Enrico Ferri	1895	Non-Fictional
305	Community Civics and Rural Life	Arthur William Dunn	1920	Non-Fictional
306	Sociology and Modern Social Problems	Charles A. Ellwood	1910	Non-Fictional
307	The Theory of the Leisure Class	Thorstein Veblen	1899	Non-Fictional

Table S2: $Means(\pm SD)$ of the different text properties analyzed in the present study. Abbreviation: MTLD, Measure for Textual Lexical Diversity. Asterisks indicate that results are different from fictional and canonical texts, respectively, at *, $p < 0.05$; **, $p < 0.01$; and ***, $p < 0.001$ (paired t-tests).

	Fictional	Non-Fictional	Canonical	Non-Canonical
Noun	3.80 (1.21)	5.76 (1.21)***	4.11 (1.24)	3.55 (1.11)
Verb	3.30 (0.90)	3.22 (0.96)	3.54 (0.83)	3.10 (0.90)
Adjective	1.28 (0.48)	2.06 (0.64)***	1.45 (0.45)	1.13 (0.45)*
Pronoun	2.04 (0.57)	1.01 (0.49)***	2.25 (0.57)	1.86 (0.50)*
Sentence-Length	20.91 (6.15)	25.13 (6.18)***	22.82 (5.83)	19.37 (5.96)*
MTLD	48.34 (7.44)	45.43 (8.56)**	47.81 (6.08)	48.76 (8.36)
Topic Distribution	0.70 (0.02)	0.68 (0.03)***	0.70 (0.02)	0.71 (0.03)

Table S3: Accuracy of classification (in %) using the mean value of the text properties for the non-fictional/fictional distinction (Task 1) and the canonical/non-canonical distinction (Task 2). $Means \pm SD$ are listed (N = 10). All values are significantly different ($p \leq 0.05$) from random accuracy (50%), except where indicated by a †.

	Task 1	Task 2
Noun	83.0 ± 2.2	56.5 ± 3.8
Verb	53.1 ± 2.0	63.8 ± 3.8
Adjective	76.3 ± 2.3	70.5 ± 3.2
Pronoun	81.3 ± 2.8	65.5 ± 3.0
Sentence-Length	66.5 ± 2.8	62.6 ± 3.5
MTLD	$51.2 \pm 1.7^{\dagger}$	55.6 ± 4.0
Topic Distribution	76.3 ± 2.0	$51.7 \pm 2.4^{\dagger}$
Low-Level	96.5 ± 1.0	66.9 ± 3.6
High-Level	79.5 ± 2.0	57.2 ± 3.3
Low- & High-Level	96.7 ± 1.0	73.5 ± 1.7

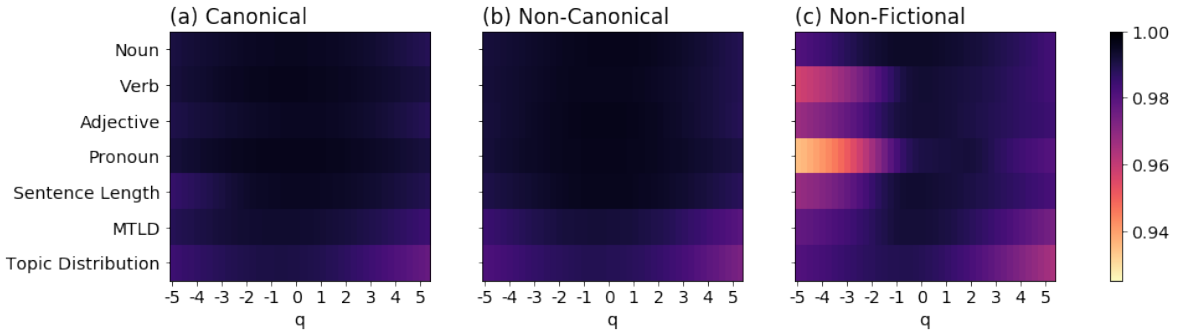

Figure S1: Mean R^2 of the linear fits to the fluctuation function of the MFDFA method for different values of q and for different text properties in canonical (a), non-canonical (b) and non-fictional texts (c) in the corpus. The color coding is shown on the right hand side.

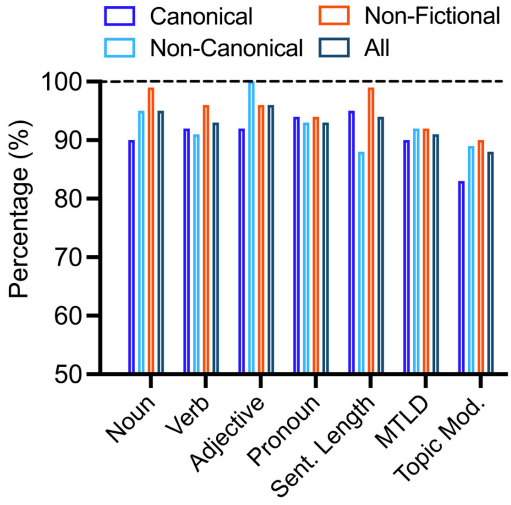

Figure S2: The percentage of the texts for which the degree of multifractality is significantly larger than the degree of multifractality of their surrogates ($p < 0.05$). The different text properties are indicated at the bottom. The color of the bars represent the categories of text separately and all together, as shown on top of the figure.

Chapter 4

PREDICTABILITY AND SURPRISE

Reading is a temporal (sequential) and slow process. It stimulates the reader to – sometimes unconsciously– speculate about what comes next. We thus hypothesized that the reader's expectation plays an important role in his reading experience. The reader's experience alternates between more and less expected situations. Accordingly, texts reflect a trade-off between predictability and surprise. On the one hand, more predictability increases monotonicity, which in turn eases the processing of text. On the other hand, more planning and more complex structural designs may increase the degree of surprise and, consequently, make reading a more demanding process. Therefore, we expect that canonical texts, written by more professional writers, potentially represent less predictable structures compared to non-canonical texts. Canonicity is taken to reflect preference at the level of a community or a society.

To investigate this hypothesis, we compared canonical with non-canonical fictional texts. We also compared these two fictional categories with non-fictional texts to obtain a more broader picture of inter-genre differences. To carry out our experiments, we extended the Jena Corpus of Expository and Fictional Prose (JEFP Corpus; Chapter 3 / Mohseni et al., 2021). Text categories in the updated version are more balanced in terms of the years of publication. The JEFP Corpus, version 2.0, similarly contains three text categories, i.e. fictional/canonical, fictional/non-canonical and non-fictional, which allow us to conduct intra- and inter-genre analyses. In our previous anlalyes, we used a group of lower-level and high-level text properties (see, Chapter 3 / Mohseni et al., 2021). The results showed that lower-text properties are more effective in discrimination of various text categories than higher-level features. Therefore, in the current study we used only two types of lower-level text properties (cf. Chapter 3) as our observables: the frequency distributions of six POS-Tags (Noun, Verb, Adjective, Adverb, Pronoun and Preposition) in fixed-size windows of texts and the lengths of sentences.

We analyzed the sequence of the text properties using two entropy metrics, Shanon Entropy (ShEn; Section 2.5.1) and Approximate Entropy (ApEn; Section 2.5.2; Pincus, 1991). ShEn determines the uncertainty/unpredictability in the global dis-

tribution of observables and does not reflect local patterns. Conversely, ApEn was originally proposed to measure irregularity in time series and is able to capture the level of unpredictability in local structures. Using both entropy metrics we could differentiate between structural patterns of text organization and global distributional features.

Our statistical analysis showed that both ApEn and ShEn take on higher values in canonical than non-canonical fiction for all POS-tags. In other words, the degree of uncertainty in canonical texts is higher compared to non-canonical texts. Classification results, from another angle, determined that the difference between the two fictional categories, i.e. canonical and non-canonical, is a matter of not only global distribution but also sequential organization. This means that in canonical fiction, the reader has less certainty (a higher degree of unpredictability) about what comes next compared to when s/he reads a non-canonical texts. Comparing fictional and non-fictional prose does not show any uniform pattern for all text properties nor for the two entropy metrics. One of the most fascinating observation was the distribution of Pronoun tags, which exhibits a higher degree of unpredictability in fictional prose compared to non-fictional expository prose and its ApEn and ShEn values separate the two text categories with a very high accuracy of 95%, similar to the result of all text properties included.

References

Mohseni, Mahdi, Volker Gast, and Christoph Redies (2021). "Fractality and Variability in Canonical and Non-Canonical English Fiction and in Non-Fictional Texts". In: *Frontiers in Psychology* 12, p. 920. DOI: 10.3389/fpsyg.2021.599063.

Pincus, Steven M (1991). "Approximate Entropy as a Measure of System Complexity". In: *Proceedings of the National Academy of Sciences* 88.6, pp. 2297–2301. DOI: 10.1073/pnas.88.6.229.

entropy

MDPI

Article

Approximate Entropy in Canonical and Non-Canonical Fiction

Mahdi Mohseni [1,2], Christoph Redies [2,*] and Volker Gast [1]

[1] Department of English and American Studies, University of Jena, 07743 Jena, Germany; mahdi.mohseni@uni-jena.de (M.M.); volker.gast@uni-jena.de (V.G.)
[2] Experimental Aesthetics Group, Institute of Anatomy I, Jena University Hospital, University of Jena, 07743 Jena, Germany
* Correspondence: christoph.redies@med.uni-jena.de; Tel.: +49-3641-9396-120

Abstract: Computational textual aesthetics aims at studying observable differences between aesthetic categories of text. We use Approximate Entropy to measure the (un)predictability in two aesthetic text categories, i.e., canonical fiction ('classics') and non-canonical fiction (with lower prestige). Approximate Entropy is determined for series derived from sentence-length values and the distribution of part-of-speech-tags in windows of texts. For comparison, we also include a sample of non-fictional texts. Moreover, we use Shannon Entropy to estimate degrees of (un)predictability due to frequency distributions in the entire text. Our results show that the Approximate Entropy values can better differentiate canonical from non-canonical texts compared with Shannon Entropy, which is not true for the classification of fictional vs. expository prose. Canonical and non-canonical texts thus differ in sequential structure, while inter-genre differences are a matter of the overall distribution of local frequencies. We conclude that canonical fictional texts exhibit a higher degree of (sequential) unpredictability compared with non-canonical texts, corresponding to the popular assumption that they are more 'demanding' and 'richer'. In using Approximate Entropy, we propose a new method for text classification in the context of computational textual aesthetics.

Keywords: Approximate Entropy; Shannon Entropy; fictional texts; non-fictional texts; canonical texts; non-canonical texts; POS-tags; text classification

Citation: Mohseni, M.; Redies, C.; Gast, V. Approximate Entropy in Canonical and Non-Canonical Fiction. *Entropy* **2022**, *24*, 278. https://doi.org/10.3390/e24020278

Academic Editor: Ernestina Menasalvas

Received: 12 January 2022
Accepted: 9 February 2022
Published: 15 February 2022

1. Introduction

Computational textual aesthetics is an emerging field at the interface of literary studies and linguistics. This field aims at identifying the statistical properties of texts to reflect categorizations of different types, e.g., authorship [1,2] and genre [3,4]. From the perspective of empirical aesthetics, properties that can potentially be associated with aesthetic categories and with perceptual responses during reading are of particular interest, as they can provide a basis for formulating specific hypotheses for experimental studies. The present study was inspired by research in (experimental) visual aesthetics, a well-established field with a tradition reaching back to the 19th century [5,6].

More recently, several computational algorithms have been proposed for the analysis of statistical properties in visually pleasing images, including visual artworks, in comparison to images with less aesthetic appeal. Particular emphasis in the studies on artworks has been on global image properties that reflect artistic composition [7]. Many of these properties reflect various aspects of fractality/self-similarity, predictability and variability in the distribution of pictorial elements across individual images. Such properties are believed to form a perceptual basis of aesthetic responses and, hence, of judgments concerning the aesthetic value of an image [8].

The question arises whether texts, like images, are characterized by global properties correlating with the aesthetic responses to those texts during reading. This question is motivated by the hypothesis of an analogy between visual processing and reading [9] on the basis of the assumption of domain-general perceptual and cognitive components

in linguistic processing [10]. Studying the aesthetic responses to texts directly would require comprehensive investigations including the observation of reader behaviour during reading (see for instance [11]). As a first step towards this program, we study the structural properties of texts grouped into different aesthetic categories. Such studies can form the basis of experimental investigations at a later stage and provide important cues concerning the experimental design, e.g., with respect to the stimulus material used and the variables analysed.

Previous observational research in textual aesthetics has often focused on poetry. While most of this research is exploratory and there is still work to be done, a number of interesting observations have been made. For example, Simonton [12] compared the vocabulary of the more "obscure" and the more popular sonnets of Shakespeare. He found a correlation between the lexical diversity and the "aesthetic success" of the sonnets. Forsyth [13] analysed the lexical features, vocabulary richness and the frequency distribution of syntactic tags in poems. He showed that the more popular poems generally used shorter words, fewer rare words, more coordinating conjunctions and more personal pronouns. Kao and Jurafsky [14] studied the style and content of poems written by professional and amateur poets to identify textual features associated with poetic beauty. Their analysis showed that more prestigious poets tended to refer more frequently to natural objects. Moreover, they made less reference to abstract concepts and used more ordinary and common words, though their vocabulary was richer.

The aesthetics of prose texts has been studied by relying on data from websites or social networks. Ashok et al. [15] attributed the success of novels to the writing style. They operationalized 'success' as the number of downloads from the Project Gutenberg site, using the distribution of POS-tags, grammatical rules, constituents and sentiments as basic measurements. In this way, they managed to classify more successful and less successful novels of different genres with acceptable accuracy.

Maharjan et al. [16] operationalized the success of a novel in terms of the average ratings on Goodreads, a social network for book lovers. They used "hand-crafted" textual features, such as the lexical and syntagmatic properties, sentiments and readability measures to predict the success of novels. Maharjan et al. [17] approached the classification of (un)successful novels by modelling the flow of emotion along a book. They showed that emotional information predicted the success of a text with relatively high accuracy.

While entropy measures have mostly been used to analyse the distributional laws of linguistics, e.g., concerning word order [18–21] and word length [22–25], or for a comparison of languages in terms of ordering preferences and complexity [26–30], there are also studies that investigated the aesthetic preference and popularity of texts using entropy metrics.

Febres and Jaffe [31] analysed the entropy and symbolic diversity of literary texts written by English and Spanish Nobel laureates and non-Nobel laureates. Their analyses confirmed that there was a correlation between the global statistical properties of texts and the two categories of authors. Chang et al. [32] analysed Shakespeare's and Jin Yong's works using a metric called "information-based energy". They showed that the more popular works had a higher "energy".

One of the main challenges of textual aesthetics is the question of how we can capture the global properties of longer prose texts, such as novels. Previous studies have used Multifractal Detrended Fluctuation Analysis as a way of measuring fractality or long-range correlations in texts. Drożdż et al. [33] analysed the fractality of sentence-length series in a corpus of Western fictional texts. Mohseni et al. [9] used a number of textual properties (sentence length, frequencies of specific POS-tags per sentence, lexical diversity measured with MTLD and topic probabilities) to generate series. They analysed these series in terms of variance and long-range correlations.

The numerical results of these methods were used as features in a classification task, intending to distinguish fiction from non-fiction and, within the fictional category, canonical vs. non-canonical English texts. The accuracy of classification was relatively high. This

finding demonstrates the feasibility and usefulness of analysing the global structural design patterns of text. Of particular interest in this context are features that are amenable to experimental studies, specifically if they allow for an interpretation in terms of perception and processing, as has been hypothesized for fractality and long-range correlations [9].

Another important aspect of aesthetic perception is the degree of (ir)regularity in a text and, related to this, the degree of predictability or surprise in the signal—cf. Zipf's principles of 'unification' and 'diversification'. Zipf [34] distinguished between the two antagonistic forces of 'unification' of the vocabulary, an economy principle from the speaker's point of view of minimizing the number of word types used, and 'diversification', maximizing the fit between words and meanings and thus benefiting the listener (see also [35]). While unification and diversification in this sense are clearly related to predictability and surprise, at least from the point of view of the specific words used (but not the meanings), we assume that literary writing is not primarily driven by the principle of unification from the author's point of view.

From an aesthetic point of view, a high degree of regularity/predictability is likely to facilitate processing, with a potentially positive effect on aesthetic perceptions. However, too much regularity may cause an impression of monotonicity. We therefore expect prose texts to reflect a trade-off between predictability and surprise. Moreover, we expect different text categories to assign different weights to two antagonistic design principles: "Keep it simple" and "Avoid monotonicity". In other words, we expect different types of balance between predictability and uncertainty in canonical and non-canonical texts. Trade-offs of this type have also been observed in music perception [36,37].

In the present study, we are primarily concerned with fictional prose. The main objective is to identify the global structural properties of texts that we have classified into the categories of 'canonical' vs. 'non-canonical'. This categorization is intended as an operationalization of aesthetic preference at a community level. While there is clearly a considerable degree of variation in individual taste, canonization—a process that involves a range of stakeholders from various sectors of society, such as literary scholars and publishers—reflects the taste of an 'average educated reader', and it has high prestige [38–40].

Canonical texts were written by skilled, mostly professional writers targeting an educated audience. Canonical literature is read in school, and educated members of societies are expected to be familiar with the major canonical works of their culture. In some countries, literary canons play an important role in the constitution of national identity (e.g., 'national poets', such as William Shakespeare ('The Bard') in the UK, Goethe and Schiller in Germany, Pushkin in Russia, etc.). Non-canonical texts do not have any of the prestige characteristics of canonical texts.

The central question of this study is whether, or to what extent, canonicity as a social attribute has structural correlates in the relevant texts. We focus on predictability and surprise, for the reasons mentioned above. As reading is a learned skill, we expect canonical texts to lean in the direction of surprise ("Avoid monotonicity"), at the expense of ease of processing. Non-canonical texts, by contrast, are (supposedly) written by less skilled writers, and not necessarily for a 'trained' audience. In this case, we expect ease of processing to prevail ("Keep it simple"). Note that we do not expect the relationship between predictability/surprise and the text categories 'canonical' and 'non-canonical' to be consistent. Our hypothesis is more general, in the sense that we expect the two classes to be associated with different balances between the two design principles of "Keep it simple" and "Avoid monotonicity", potentially in different aspects of structural design.

The canonical and non-canonical texts of our corpus belong to the same genre, i.e., fictional prose. In order to gauge the degree of register specificity of the observed patterns, we included texts from a different register as well, i.e., non-fictional (expository) prose (see Section 2.1). We used two types of observables, the length of sentences and frequency distributions of part-of-speech (POS) tags. The frequencies of POS-tags were determined in fixed-size windows of text, which we call 'boxes' (see Section 2.2). To measure the (ir)regularity and predictability in a text, we used two types of entropy measures, Approxi-

mate Entropy (ApEn) and Shannon Entropy (ShEn) (see Section 2.3). Section 3 presents the results, which are then discussed in Section 4.

2. Data and Methods

2.1. The JEFP Corpus 2.0

For our computational textual aesthetics studies, we needed a corpus that was tailor-made for the purpose of the project, the comparison of canonical and non-canonical fiction. While there are several corpora of literary texts available (e.g., the Standardized Project Gutenberg Corpus/SPGC [41]), we compiled a corpus of our own with a certain balance across text types and the time of publication: the *Jena Expository and Fictional Prose* (JEFP) corpus. This corpus contains canonical and non-canonical fictional as well as non-fictional texts.

In a previous study [9], we used version 1.0 of this corpus. For the present study, we extended the corpus and included more texts, primarily in order to achieve a better balance in terms of the years of publication.

The canonical texts of the JEFP corpus 2.0 are the same as those contained in version 1.0. The corpus comprises 76 canonical literary texts from 30 authors, which were taken from the *Corpus of the Canon of Western Literature* (CCWL) [42], which, in turn, relies on Bloom [43] (*The Western Canon: The Books and School of the Ages*). As an additional criterion of canonicity (of authors), we used evidence from Wikipedia sites. We determined the number of articles for authors in the top 30 language editions of Wikipedia, as an approximate indication of their international reputation.

In order to obtain a sample of non-canonical fictional texts, we used the websites www.goodreads.com (accessed on 8 February 2022), feedbooks.com as well as Project Gutenberg (www.gutenberg.org, accessed on 8 February 2022). The raw texts were all extracted from the Project Gutenberg site. We selected only long books that comprised at least 35,000 words, as a critical number of words is required for analysis of the global properties of text using methods such as Multifractal Detrended Fluctuation Analysis (MFDFA; see [9]).

At the time of compilation of the corpus (May 2020), none of the books classified as "non-canonical" by us had a download number higher than 40. By thresholding the download count, we avoided including non-canonical popular literature. In previous studies, download counts at the Project Gutenberg site have been used as a surrogate to gauge the success of books [15,44].

We made sure that the relevant (non-canonical) authors were not listed in the canon underlying our study, the Canon of Western Literature [43]. Moreover, none of the non-canonical authors has as many Wikipedia pages as the canonical author with the lowest number of pages (14). The authors classified as 'non-canonical' thus did not have the international prestige that is characteristic of canonical authors. The sample of non-canonical texts thus compiled contained 130 texts.

Non-fictional texts were also taken from the Project Gutenberg site. The sample contained in version 1.0 of the corpus was extended with texts from different genres, such as architecture, astronomy, geology, geography, philosophy, psychology and sociology. The extended corpus contained 185 texts of this category. Table 1 provides summary statistics for the texts of the corpus. The texts with metadata are listed in Supplementary Table S1.

Table 1. Text categories in the Jena Expository and Fictional Prose (JEFP), version 2.0. The table shows, for each text category, the number of texts and the mean text length, measured in tokens, ± standard deviation.

Category	Number of Texts	Mean Length ($\times 10^3$)
Canonical	76	199 ± 96
Non-Canonical	130	111 ± 56
Non-Fictional	185	171 ± 178

For preprocessing of the texts, we removed the tables of contents and indices as well as any other material not belonging to the core text from each document. We cleaned up the texts semi-automatically using regular expressions, e.g., in order to rejoin hyphenated words and fix broken lines. We used the `Stanza` package for Python [45], an up-to-date neural-based text processing toolbox, to sentencize, tokenize and POS-tag all texts.

The three text categories of the JEFP corpus allowed us to carry out intra-genre comparison, i.e., canonical vs. non-canonical fictional texts, which is the main focus of our study, as well as inter-genre comparison, i.e., fictional vs. non-fictional texts. The inter-genre comparison is intended to give us an idea of the degree of genre specificity of any observed effects (see Section 3).

2.2. Properties Underlying Textual Structure

As reflexes of the structural organization of the texts, we used the length of sentences and part-of-speech tags (POS-tags) as assigned by the `Stanza` package for Python [45]. The distributions of POS-tags reflect grammatical structure as well as register and discourse modes [46]. For example, pronouns are associated with interactive communication, such as face-to-face conversation, verbs are typical of narration, and adjectives are characteristic of description. Regularity or irregularity in the organization of discourse modes can thus be measured in terms of the sequential distribution of POS-tags in a text.

In our study, we focused on six major parts of speech: nouns, verbs, adjectives, adverbs, pronouns and prepositions. We only took the top-level categories into account. For example, the tag 'Noun' covers singular as well as plural nouns and common nouns as well as proper names; different forms of verbs, such as the base forms, past tense forms and gerunds, are treated as a single class, 'Verb'; simple, comparative and superlative adjectives are all subsumed under 'Adjective'; and so on. We capitalize these general POS-tags in order to distinguish them from elements of the relevant classes (nouns, verbs, etc.).

We determined the frequencies of POS-tags per fixed-length segments, i.e., windows, of text. We did not use sentences as the scope of measurement because sentence length figured as a separate explanatory variable in our study and because we wanted to obtain measurements that were independent of punctuation practice. We therefore split the texts into windows of 25 tokens, which is the approximate average sentence length of the corpus (in fact, 23.3 tokens). It is important to mention, however, that the window size, within reasonable limits, did not have a noticeable effect on the results. We experimented with segments of 10 to 50 tokens in steps of 5 tokens but did not observe any major differences.

By windowing, each text is converted into a sequence of small bags of words—'boxes', as we call them—in which POS-tag frequencies are determined regardless of the position of the individual words. The linear order of the values obtained from the 25-words boxes was important as it was regarded as a reflex of the structural organization of the texts.

If the linear order of the counts is taken into account, as in the case of Approximate Entropy, we will speak of a 'sequence of boxes'; if linear order does not matter, as in the case of Shannon Entropy, we will speak of a 'bag of boxes'. Our approach is thus neither a bag-of-words nor a word-sequence approach. Word-sequence approaches—specifically, function–word-adjacency networks (WANs)—have been used in authorship attribution [47,48] and gender classification [48] (for a detailed description of WAN, see [49]).

As we used six parts of speech, we obtained six series based on counts of POS-tags in boxes. A series $X_{POS} = x(1), x(2), \cdots, x(n)$ for a specific POS-tag thus contains the frequencies of the relevant tag in subsequent windows of 25 tokens. If L is the text length, the length of the series $n = \lfloor L/25 \rfloor$. In the same way, we generated series of integers representing the length of the sentences in a text. Sentence length was measured as the number of tokens (including punctuation marks) in a sentence as sentencized by the `Stanza` package.

2.3. Computation of Unpredictability in Text

Each series generated as described in Section 2.2 is a sequence of events that are not independent from each other. As was shown in Mohseni et al. [9], they exhibit long-range

correlations (though the method used to generate the series was slightly different in this publication). As an operationalization of (ir)regularity and predictability in a text, we used Approximate Entropy (ApEn), which measures predictability in linearly ordered random variables. We analysed sequences of POS-tag counts observed in 'boxes' in terms of Approximate Entropy (the 'sequence-of-bags approach').

In order to determine to what extent observed degrees of (ir)regularity are a property of the global (bag-of-boxes) distribution of structural features, rather than their linear arrangement, we also calculated summary statistics by using standard Shannon Entropy (ShEn). Associations of entropy values (ShEn and ApEn) with text categories were determined with a classification task, using a Support Vector Machine (Section 3.2). In what follows, we briefly describe both entropy measures, starting with Shannon Entropy.

2.3.1. Shannon Entropy

Shannon Entropy (ShEn) is a well-known concept in information theory that measures uncertainty in a random variable. Given a discrete random variable x and a probability distribution $p(x)$, the ShEn of x, $h(x)$, is computed as

$$h(x) = -\sum_{x \in S_x} p(x) \log_e p(x) \tag{1}$$

where S_x is the set of all possible events. In a system with all possible events being equally likely to happen, uncertainty and, hence, ShEn, is at a maximum. A major advantage of ShEn is that it is parameter-free, straightforward and easily interpretable.

We can determine the ShEn for the six POS-series as well as the series of sentence length measurements and treat them as a global measurements of predictability. Once again, it should be stressed, however, that ShEn does not capture local patterns of distribution but is a function of the probability distribution as a whole. We therefore use it in conjunction with the Approximate Entropy as described in Section 2.3.2.

2.3.2. Approximate Entropy

Approximate Entropy (ApEn) was first proposed by Pincus [50] as a way of measuring the degrees of regularity in times series. A high value of ApEn means a low degree of predictability and vice versa.

ApEn is computed according to sub-sequence matches of length m compared with sub-sequence matches of length $m+1$. The match between sub-sequences of a series is a function of a distance metric in relation to a predefined threshold value r. Let $X = x(1), \cdots, x(n)$ be a time series, m be the length of a sub-sequence and r be a positive value. ApEn is computed as follows:

1. Create sub-sequences $y_i^m = [x(i), \cdots, x(i+(m-1))]$ for $i = 1, \cdots, n-m+1$.
2. Using the distance between y_i^m and y_j^m, defined as $d_{i,j}^m = \max_k |y_i^m(k) - y_j^m(k)|$, compute

$$C_i^m(r) = \frac{1}{n-m+1} \sum_{j=1}^{n-m+1} \mathbb{1}(r - d_{i,j}^m)$$

   in which $\mathbb{1}(.)$ is the Heaviside function whose value is 1 when its parameter is positive and otherwise 0.
3. Compute

$$\phi^m(r) = \frac{1}{n-m+1} \sum_{i=1}^{n-m+1} \log(C_i^m(r))$$

4. Finally, calculate ApEn as

$$\texttt{ApEn}(m, r) = \phi^m(r) - \phi^{m+1}(r)$$

If the series is fixed at some value and is thus fully predictable, ApEn is 0. The value of ApEn depends on its two parameters, m and r. m is usually set to 2, and the value of r, which should be related to the standard deviation (SD) of the series, is set to $0.2 \times$ SD (see, for example, [51–53]). In our experiments, we also applied this parameter setting.

ApEn has been subject to a broad range of research, and its behaviour has been studied under various types of circumstances. Researchers have proposed extensions of ApEn, such as Sample Entropy [54], Multi-Scale Entropy [55] and Multivariate Multi-Scale Entropy [56], which may provide more accurate analyses for certain time series. In order to compare ApEn with these extensions, we conducted experiments using the `NeuroKit2` python package [57], which implements these extensions. We observed that none of these methods provided a better discrimination power compared to ApEn. Therefore, we only report the experimental results of ApEn in Section 3.

3. Results

In this section, we first present the results of the statistical analyses (Section 3.1) and then turn to the results of our classification experiment (Section 3.2). As pointed out in Section 2.1, the JEFP corpus contains texts from three categories: fiction/canonical, fiction/non-canonical, and non-fiction. Our main focus is on the difference between canonical and non-canonical fiction. As we wish to determine to what extent any observed differences are genre-related, we also included non-fictional texts in our comparison.

3.1. Statistical Analysis of Features

For each text in the corpus and for each text property, ApEn (Table 2) and ShEn (Table 3) were computed. As some features were not normally distributed (confirmed by a Kolmogorov–Smirnov test), we used the median values and compared them with the Mann–Whitney U test. In Tables 2 and 3, each pair of columns shows a comparison of ApEn and ShEn values for each text category/feature combination. Whenever a value is significantly higher than the corresponding value for the other text category, the higher value is shown in bold face. Levels of significance are indicated by the superscripts on the right value within each pair of columns.

Table 2. Median values of Approximate Entropy (ApEn) for all text properties. ApEn values were analysed for two tasks: canonical ($N = 76$) vs. non-canonical ($N = 130$) texts and fictional ($N = 206$) vs. non-fictional ($N = 185$) texts. The asterisks indicate whether the differences between the two text categories of a given task are statistically significant (Mann–Whitney U test; ns, not significant; * $p \leq 0.05$; ** $p \leq 0.01$; and *** $p \leq 0.001$). Values that are significantly higher within a pair of columns are shown in boldface. 95% confidence intervals for the median (according to [58]) are shown in parentheses.

Text Property	Canonical	Non-Canonical	Fictional	Non-Fictional
Sentence Length	1.86 (1.83, 1.89)	1.87 (1.86, 1.90) ^{ns}	1.87 (1.86, 1.88)	1.90 (1.88, 1.92) ^{ns}
Noun	**1.89 (1.88, 1.91)**	1.83 (1.81, 1.84) ***	**1.85 (1.84, 1.86)**	1.82 (1.81, 1.84) **
Verb	**1.75 (1.73, 1.76)**	1.70 (1.69, 1.71) ***	1.714 (1.706, 1.723)	**1.756 (1.745, 1.764)** ***
Adjective	**1.50 (1.49, 1.52)**	1.45 (1.43, 1.48) ***	1.488 (1.469, 1.494)	**1.58 (1.55, 1.60)** ***
Adverb	**1.51 (1.49, 1.53)**	1.48 (1.46, 1.49) **	**1.49 (1.48, 1.50)**	1.36 (1.34, 1.39) ***
Pronoun	**1.74 (1.71, 1.76)**	1.681 (1.675, 1.691) ***	**1.695 (1.685, 1.704)**	1.31 (1.28, 1.36) ***
Preposition	**1.71 (1.70, 1.72)**	1.67 (1.66, 1.68) ***	1.678 (1.672, 1.683)	**1.691 (1.686, 1.697)** ***

Table 3. Median values of Shannon Entropy (ShEn) for all text properties. ApEn values were analysed for two tasks: canonical (N = 76) vs. non-canonical (N = 130) texts and fictional (N = 206) vs. non-fictional (N = 185) texts. The asterisks indicate whether the differences between the two text categories of a given task are statistically significant (Mann–Whitney U test; ns, not significant; * $p \leq 0.05$; ** $p \leq 0.01$; and *** $p \leq 0.001$). Values that are significantly higher within a pair of columns are shown in boldface. 95% confidence intervals for the median (according to [58]) are shown in parentheses.

Text Property	Canonical	Non-Canonical	Fictional	Non-Fictional
Sentence Length	3.96 (3.88, 4.05)	3.96 (3.87, 4.08) ^{ns}	3.96 (3.91, 4.03)	**4.10 (4.07, 4.16)** ***
Noun	**2.00 (1.99, 2.02)**	1.97 (1.95, 1.98) ***	1.98 (1.97, 1.99)	1.97 (1.95, 1.99) ^{ns}
Verb	**1.80 (1.79, 1.81)**	1.777 (1.772, 1.783) ***	1.785 (1.779, 1.792)	**1.844 (1.836, 1.853)** ***
Adjective	**1.54 (1.53, 1.55)**	1.49 (1.47, 1.53) ***	1.52 (1.51, 1.53)	**1.63 (1.61, 1.66)** ***
Adverb	**1.54 (1.51, 1.55)**	1.51 (1.49, 1.53) *	**1.52 (1.51, 1.53)**	1.40 (1.37, 1.42) ***
Pronoun	**1.83 (1.80, 1.84)**	1.78 (1.77, 1.79) ***	**1.79 (1.78, 1.80)**	1.37 (1.33, 1.42) ***
Preposition	**1.75 (1.74, 1.77)**	1.73 (1.72, 1.74) ***	1.736 (1.729, 1.744)	**1.76 (1.75, 1.77)** ***

The most important observation that stands out from a superficial inspection of Tables 2 and 3 is that the left two columns, which show the values for canonical and non-canonical fiction, exhibit a rather uniform pattern: while there are no significant differences between the values for sentence length (in the top row), the ApEn as well as the ShEn values for all series derived from POS-frequencies within boxes are higher for canonical than for non-canonical texts.

In contrast, in the right pair of columns, showing the comparison between fictional and non-fictional texts, there is no uniform pattern. Fictional texts have higher ApEn and ShEn values than non-fictional texts for Adverb and Pronoun, and the ApEn value for Noun is higher in fictional than in non-fictional texts. Non-fictional texts have higher ApEn and ShEn values for Verb, Adjective and Preposition, and the ShEn value for Sentence Length is higher than for fictional texts.

In conclusion, Tables 2 and 3 thus show that entropy values—both ApEn and ShEn—are consistently higher in canonical than in non-canonical fiction for POS-tag frequencies within boxes, whereas there is no such clear tendency in the comparison between fictional and non-fictional prose (though there are also significant differences).

Stated differently, the results shown in Tables 2 and 3 suggest that canonical fictional texts are characterized by a higher degree of uncertainty than non-fictional texts, when treated either as a bag-of-boxes distribution (with ShEn) or a sequence-of-boxes distribution (ApEn). Fictional texts differ from non-fictional texts in terms of the uncertainty associated with specific POS-tags; however, there is no uniform pattern. It appears that, in fictional prose, pronouns and adverbs are distributed less predictably than in non-fictional prose, while in non-fictional texts, the distribution of verbs, adjectives and prepositions is less predictable in comparison with fictional texts.

Visual inspection of the data in Tables 2 and 3 does not *prima facie* show any clear patterns with respect to the differences in magnitude of the ApEn values (Table 2) and the ShEn values (Table 3), for each pair of columns. In order to determine whether the degrees of uncertainty observed for the various text category/feature combinations are a property of the texts as bags of boxes or as a function of the linear sequence of the boxes, we used classification tasks with a Support Vector Machine, which allows us to estimate the discriminatory power of each feature.

3.2. Classification

In two classification tasks, we determined what features can most efficiently classify or separate the categories of text under analysis—canonical vs. non-canonical fiction and fictional vs. non-fictional (expository) prose. We refer to the task of classifying canonical vs. non-canonical texts as 'Task 1' and the task of classifying non-fictional vs. fictional texts as 'Task 2'. We used a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel for the two tasks. As the categories to be classified are of different size, we used

balanced accuracy as our evaluation measure. Wherever we compare classification results, we used the 5×2CV paired t-test [59] with a significance level of $\alpha = 0.05$. We report the mean of the 10 runs, 5 times 2-fold cross-validation, for each setting.

Table 4 shows the classification results for the two tasks using ApEn values and ShEn values calculated for each text as features for the classification task. As in Tables 2 and 3, values that are significantly higher than their counterparts are highlighted with boldface. The top section of the table shows the results for each individual property. The most important observation is that ApEn separates canonical from non-canonical fictional texts better than ShEn does (Task 1).

Wherever the results are significantly better than random accuracy (50%), ApEn is more effective than ShEn. Moreover, for ApEn, classification is significantly different from random accuracy for all but one text property, i.e., Adverb, while the differences are not statistically significant for three text properties for ShEn, i.e., Sentence Length, Adjective and Adverb (indicated by a dagger† in Table 4). Table 4 also shows the classification results when all text properties are taken into account. In Task 1, ApEn outperformed ShEn by a large margin (77.3% vs. 68.5%).

Table 4. Balanced accuracy of classification (in %) for the single features for the canonical/non-canonical distinction (Task 1) and the non-fictional/fictional distinction (Task 2). To compare classification results, we used the 5×2CV paired t-test [59] with a significance level of $\alpha = 0.05$. Values that are significantly higher within a pair of columns are shown in boldface. All values are significantly different ($p \leq 0.05$) from random accuracy (50%), except where indicated by a dagger (†).

	Task 1		**Task 2**	
	ApEn	**ShEn**	**ApEn**	**ShEn**
Sentence Length	**54.0 ± 1.6**	50.0 ± 1.0 †	53.6 ± 2.9	**61.7 ± 2.3**
Noun	**73.6 ± 2.9**	60.0 ± 4.5	57.4 ± 1.9	**64.2 ± 1.8**
Verb	**71.3 ± 3.4**	56.2 ± 3.8	65.5 ± 2.4	**74.0 ± 1.6**
Adjective	**55.2 ± 2.5**	51.5 ± 2.7 †	71.7 ± 2.1	74.3 ± 1.0
Adverb	51.6 ± 1.4 †	51.0 ± 1.5 †	72.8 ± 2.2	73.0 ± 2.9
Pronoun	**68.0 ± 1.7**	63.8 ± 1.8	95.1 ± 1.5	95.0 ± 1.7
Preposition	**69.1 ± 2.4**	59.7 ± 1.7	56.9 ± 2.6	**61.4 ± 1.3**
All	**77.3 ± 2.6**	68.5 ± 2.3	95.4 ± 1.8	96.5 ± 1.9

While the overall accuracy measures for ApEn may seem moderate in Task 1—77.3% using all features, with the POS-tag Noun alone reaching 73.6%—it should be borne in mind that this task is particularly difficult. Canonical and non-canonical texts belong to the same genre—fictional prose—and the differences between them can be expected to be subtle. The accuracy values for ShEn, which are significantly lower than those for ApEn, show that the difference between canonical and non-canonical fiction is not so much a matter of global (bag-of-boxes) distributions as it is a matter of sequential organization (sequence-of-boxes distribution).

The results for Task 2 differ strikingly from those for Task 1. Importantly, ShEn overall appeared to perform better than ApEn in this task. The results are significantly higher for Sentence Length, Noun and Verb. For three of the features—Adjective, Adverb and Pronoun— ApEn and ShEn values do not differ significantly. Concerning the results based on all features, the accuracy values of ShEn and ApEn are also similar, with values of $> 95\%$, and the observed difference is not significant. This result suggests that the differences between fictional and non-fictional texts are a matter of global distribution rather than sequential organization.

The right column of Table 4 shows another interesting result: The feature Pronoun alone classifies fictional vs. non-fictional texts with very high accuracy ($\approx$95%), for both ApEn and ShEn. In fact, using all features does not lead to a significantly better performance than using Pronoun alone.

Given the prevalence of the feature Pronoun in the classification of fictional vs. non-fictional texts (Task 2), we repeated the task using all features except Pronoun, to gain a better understanding of the role of the remaining text properties. Without Pronoun, the performance of classification dropped to 89.7% and 91.0% for ApEn and ShEn, respectively, a considerable decrease for both features.

In comparison with other classification studies, the accuracy scores obtained in our study may appear to be rather moderate overall. Studies based on lexical material or n-grams may be more successful in text classification (see, for instance [60] on novels by Stephen King). We would like to emphasize, however, that we are interested in understanding the higher-level design features of texts, not their make-up in terms of low-level features, such as words or n-grams.

Our endeavour is thus more comparable to studies that aim to classify texts in terms of parameters associated with linguistic laws, such as Zipf's law [34,35] and the Menzerath–Altmann law [61–64]. For comparison, we therefore ran classification tasks using parameters of these laws as input features (as suggested by a reviewer). The lambda-values of Zipfian distributions fitted to lemma counts delivered accuracy scores of 64.8% (Task 1) and 56.8% (Task 2). The two parameters b and c of a Menzerath–Altmann distribution ($y = ax^b e^{-cx}$) fitted to the average length of clauses and measured in tokens as a function of the number of clauses in a sentence, yielded accuracy scores of 55.8% (Task 1) and 68.6% (Task 2) (we used the package 'menzerath' for R [65] to extract the parameters with the function 'menzerath()' [method 'MAL']).

This illustrates, again, how difficult Task 1 is. Our experiments with the parameters of linguistic laws were only preliminary, and there are certainly ways of optimizing the classification process, e.g., by applying a more precise definition of 'clause' (we split sentences into clauses by relying on punctuation). In any case, they confirm that classification with a low number of features that describe a text as a whole is a difficult undertaking and that accuracy scores in the range of 75–80% as obtained with Approximate Entropy for Task 1 are less disappointing than they might appear to be on first sight. The lambda parameters of Zipf's law and the two parameters of the Menzerath–Altmann law (b and c) are shown in Supplementary Figures S5 and S6, respectively.

3.3. Most Discriminative Features

As mentioned above, the discrimination of canonical vs. non-canonical texts (Task 1) is much more difficult than that of fictional vs. non-fictional texts (Task 2). While in Task 2 there is one prominent feature—Pronoun—the contributions of the features in Task 1 are more evenly distributed. In order to determine degrees of feature importance, we applied two methods.

First, we used sensitivity analysis [66]; the results are shown in Supplementary Figures S7–S8. This analysis confirms the impression given by Table 4 that Noun and Verb are the most important discriminators for ApEn in Task 1, while Pronoun is the most important discriminator for ShEn. Second, we ran a brute-force search on the ApEn features as well as the ShEn ones (to give readers a visual impression of the discriminatory power of pairs of features, we provide pair plots of all features for ApEn and ShEn in the Supplementary Figures S1–S4).

Again, the most effective pair of ApEn features was that of Noun and Verb. Figure 1a visualizes the values of the two features for all fictional texts. The ApEn values of fictional texts in both the Noun and Verb series tended to be higher for canonical than for non-canonical texts. Moreover, the correlation between these two features, i.e., the ApEn of Noun and of Verb, was higher for canonical texts (Pearson coefficient 0.75) than for non-canonical texts (0.49). For comparison, Figure 1b shows the ShEn values for Noun and Verb. The figure demonstrates that the discriminative power of the two features is significantly lower than that of the corresponding ApEn values as shown in Figure 1a.

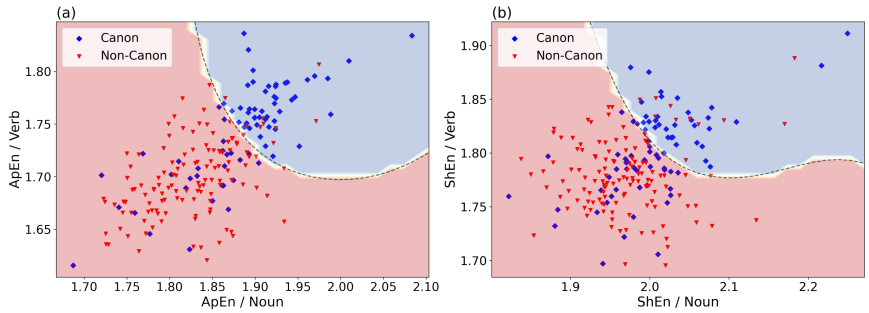

Figure 1. ApEn (**a**) and ShEn (**b**) of Noun and Verb, the two best features for classification of canonical vs. non-canonical texts (Task 1). ApEn and ShEn values of these two features provide an accuracy of 75.9% and 68.4%, respectively. The coloured regions and the border (dashed) line show the decision space of the Support Vector Machine.

In Task 2, ApEn and ShEn of Pronoun were the most effective features in discriminating fictional from non-fictional texts with an accuracy of >95%. As Table 4 shows, adding more features does not improve the classification results significantly. In Figure 2a,b, the distributions of ApEn and ShEn values for Pronoun are visualized in the form of violin plots. The figures show that the values are clearly higher for fictional than for non-fictional texts, while the ranges of values for canonical and non-canonical texts largely overlap.

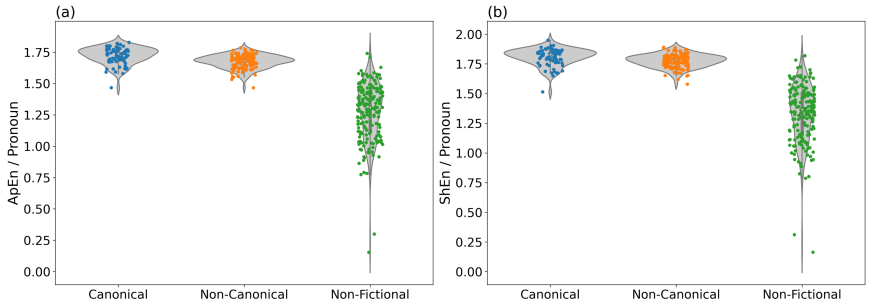

Figure 2. Values for ApEn (**a**) and ShEn (**b**) of Pronoun. These two features yield high accuracy for the classification of fictional vs. non-fictional texts (Task 2).

Note also that the values for non-fictional texts are very broadly distributed, while the values for fictional (canonical and non-canonical) texts are consistently very high. As there is hardly any difference between the plots for ApEn and ShEn, we can assume that the uncertainty due to the distribution of pronouns is a matter of global distribution, rather than sequential organization, as mentioned above.

4. Discussion and Conclusions

The most important result of our study can be summarized as follows: Canonical and non-canonical fictional texts differ in their degrees of predictability regarding the sequential distribution of the major parts of speech Noun, Pronoun, Verb, Adjective, Adverb and Preposition in windows of 25 tokens, and this was reflected in their higher Approximate Entropy (ApEn) values (cf. Table 2).

In other words, following a given window of text, there is less certainty about the frequency of specific parts of speech in the next window in canonical fictional texts, in comparison to non-canonical fictional texts. This result confirms our expectation that canonical fictional texts may be less predictable than non-canonical texts in terms of their textual structure. Whether or not this is perceived by a reader remains to be investigated. We assume that the observed differences are located at a medium level of text or discourse

organization. They are probably not so much a matter of sentence-level syntax as they are of textual organization at the paragraph level.

Specifically, we suspect that frequency distributions of part-of-speech tags reflect discourse modes where the less predictable structural organization of canonical texts is due to (more unpredictable) shifts between discourse modes. The most important discourse modes in (traditional) fictional prose are those of narration and dialogue, followed perhaps by description. Verbs and nouns are important discriminators of discourse modes, insofar as verbs are prevalent in narration and dialogue, while nouns are more frequent in description and are particularly rare in dialogue.

In order to test this hypothesis, more detailed and thorough investigations will be needed. One way of approaching this task could be with Latent Dirichlet Allocation (LDA), which is commonly used for Topic Modelling [67]). If rhetorical modes are associated with multinomial distributions over parts of speech, as we assume, LDA-models (potentially supervised/labelled) could be trained on mixed-genre corpora. The models trained in this way could be used to assign to each window of text a distribution of discourse modes, and the resulting distributions could be analysed using methods like the ones applied in the present study, or other ways of capturing the global structural properties of texts (e.g., MFDFA [9]).

As the discriminative power of Shannon Entropy (ShEn) was lower than that of ApEn in Task 1, we assume that our results concerning the difference between canonical and non-canonical fictional texts do not reflect bag-of-boxes distributions but rather sequential organization within individual texts as reflected in sequences of boxes.

The results of our comparison between fictional and non-fictional prose were very different. The task of discriminating fictional from non-fictional texts was overall much easier than the classification of canonical vs. non-canonical fictional texts, as shown by the (balanced) accuracy scores of the classification tasks. This is not surprising, as we are here dealing with a question of genre classification, whereas canonical and non-canonical texts belong to the same genre and (by hypothesis) differ in terms of the textual structure.

Since ApEn did not fare better than ShEn in the fictional/non-fictional classification task, we assume that this is a matter of the bag-of-boxes distributions of text features, rather than of their sequential structure. Note also that there was no consistent pattern in the distribution of ShEn values across text properties. It appears that fictional and non-fictional texts differ in the ways parts of speech are distributed, with some of them showing flatter distributions (with higher entropy) and others showing steeper distributions (with lower entropy values) without a general trend.

An interesting observation that emerged from Task 2 was the central role of pronoun frequencies, which showed high performance. Pronouns are often not analysed in text classification and are often ignored as they are filtered out as stopwords. However, there are also studies acknowledging the importance of pronouns. For example, Kernot [68] showed that data taken from 30 articles written by three female and two male authors could be classified into gender categories by using only three pronouns, i.e., *my*, *her* and *its*.

Similarly, a study of sentimentalism in literature, Yu [69], found that pronouns are particularly valuable discriminators. In the context of register classification, the discriminatory power of pronouns is plausible. Qureshi et al. [70] found that the ratio of the number of adjectives to the number of pronouns is a good discriminator for distinguishing fictional from non-fictional texts.

Our finding that pronouns are informative when their predictability of occurrence is studied fits into this picture. While fictional texts are characterized by alternations between narrative passages and dialogue, the latter mode being associated with deictic pronouns (*I*, *you*), non-fictional prose can be expected to exhibit a more even distribution of anaphoric pronouns (*she*, *he*, *they*).

Our finding that the sequential structure of canonical texts is less predictable than that of non-canonical texts can be compared to results from vision studies. The basic perceptual features of visual images include, for example, oriented gradients of luminance or colour

(edges). It has been shown that the distribution of edge orientations is less predictable across individual images of traditional artworks than in several types of non-art images [7]. In analogy to the present results for texts, the entropy of edge orientations is relatively high in visual artworks. High entropy of edge orientations can also be observed in other stimuli that beholders like more, including artificially generated visual patterns [71,72].

In the auditory domain, an intermediate degree of unpredictability and its resolution during listening are thought to evoke musical pleasure [36] in agreement with predictive coding accounts of brain function [37,73] (for a review of possible neural correlates of musical expectations in the human brain, see [74]). We speculate that a certain degree of unpredictability in the distribution of basic structural (perceptual) features is one of the hallmarks of aesthetically appreciated stimuli. Whether this hypothesis can be generalized to other types of text and whether this reflects domain-general perceptual and cognitive processes across sensory domains remains to be investigated.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/e24020278/s1: Table S1: List of texts in the Jena Corpus of Expository and Fictional Prose (JEFP Corpus, Version 2.0). Figure S1: Pair-plot of all Approximate Entropy (ApEn) features in fictional/canonical and fictional/non-canonical texts. Figure S2: Pair-plot of all Shannon Entropy (ShEn) features in fictional/canonical and fictional/non-canonical texts. Figure S3: Pair-plot of all Approximate Entropy (ApEn) features in fictional/canonical, fictional/non-canonical and non-fictional texts. Figure S4: Pair-plot of all Shannon Entropy (ShEn) features in fictional/canonical, fictional/non-canonical and non-fictional texts. Figure S5: Zipf's law coefficient (lambda) of fictional/canonical, fictional/non-canonical and non-fictional texts. Figure S6: The two parameters (b and c) of the Menzerath–Altmann law in fictional/canonical, fictional/non-canonical and non-fictional texts. Figure S7: Sensitivity analysis of ApEn features and ShEn features in classification of fictional/canonical and fictional/non-canonical texts. Figure S8: Sensitivity analysis of ApEn features and ShEn features in classification of fictional and non-fictional texts.

Author Contributions: Conceptualization, M.M., C.R. and V.G.; Methodology, M.M.; Software, M.M.; Validation, M.M., C.R. and V.G.; Formal Analysis, M.M.; Investigation, M.M. and V.G.; Resources, C.R. and V.G.; Data Curation, M.M.; Writing—Original Draft Preparation, M.M., C.R. and V.G.; Writing—Review and Editing, M.M., C.R. and V.G.; Visualization, M.M.; Supervision, C.R. and V.G.; Project Administration, C.R. and V.G.; Funding Acquisition, C.R. and V.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the German Science Foundation grant number 380283145.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to copyright restrictions.

Acknowledgments: We thank members of the Experimental Aesthetic Group for helpful discussions and two anonymous reviewers for their constructive criticism and helpful suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Craig, H.; Kinney, A.F. *Shakespeare, Computers, and the Mystery of Authorship*; Cambridge University Press: Cambridge, UK, 2009. doi:10.1017/CBO9780511605437.
2. Koppel, M.; Schler, J.; Argamon, S. Computational Methods in Authorship Attribution. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 9–26. doi:10.1002/asi.20961.
3. Biber, D. *Dimensions of Register Variation. A Cross-linguistic Comparison*; Cambridge University Press: Cambridge, UK, 1995.
4. Lee, D. Genres, Registers, Text Types, Domains and Styles: Clarifying the Concepts and Navigating a Path Through the BNC Jungle. *Technology* **2001**, *5*, 37–72. doi:10.1163/9789004334236_021.
5. Fechner, G.T. *Vorschule der Ästhetik*; Breitkopf and Härtel: Leipzig, Germany, 1876.
6. Bell, C. *Art*; Chatoo & Windus: London, UK, 1914.
7. Redies, C.; Brachmann, A.; Wagemans, J. High Entropy of Edge Orientations Characterizes Visual Artworks From Diverse Cultural Backgrounds. *Vis. Res.* **2017**, *133*, 130–144. doi:10.1016/j.visres.2017.02.004.
8. Brachmann, A.; Redies, C. Computational and Experimental Approaches to Visual Aesthetics. *Front. Comput. Neurosci.* **2017**, *11*, 102. doi:10.3389/fncom.2017.00102.

9. Mohseni, M.; Gast, V.; Redies, C. Fractality and Variability in Canonical and Non-Canonical English Fiction and in Non-Fictional Texts. *Front. Psychol.* **2021**, *12*, 920. doi:10.3389/fpsyg.2021.599063.
10. Diessel, H. *The Grammar Network. How Linguistic Structure is Shaped by Language Use*; Cambridge University Press: Cambridge, UK, 2019.
11. Hartung, F.; Wang, Y.; Mak, M.; Willems, R.; Chatterjee, A. Aesthetic Appraisals of Literary Style and Emotional Intensity in Narrative Engagement Are Neurally Dissociable. *Commun. Biol.* **2021**, *4*, 1401. doi:10.1038/s42003-021-02926-0.
12. Simonton, D.K. Lexical Choices and Aesthetic Success: A Computer Content Analysis of 154 Shakespeare Sonnets. *Comput. Humanit.* **1990**, *24*, 251–264. doi:10.1007/BF00123412.
13. Forsyth, R.S. Pops and Flops: Some Properties of Famous English Poems. *Empir. Stud. Arts* **2000**, *18*, 49–67. doi:10.2190/E7Q8-6062-K6H4-XFRW.
14. Kao, J.; Jurafsky, D. A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry. In Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature, Montreal, QC, Canada, 8 June 2012.
15. Ashok, V.; Feng, S.; Choi, Y. 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, Seattle, WA, USA, 18–21 October 2013.
16. Maharjan, S.; Arevalo, J.; Montes, M.; González, F.; Solorio, T. 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*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 1217–1227. doi:10.18653/v1/E17-1114.
17. Maharjan, S.; Kar, S.; Montes, M.; González, F.A.; Solorio, T. 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)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 259–265. doi:10.18653/v1/N18-2042.
18. Montemurro, M.A.; Zanette, D.H. Universal Entropy of Word Ordering Across Linguistic Families. *PLoS ONE* **2011**, *6*, e19875. doi:10.1371/journal.pone.0019875.
19. Montemurro, M.A.; Zanette, D.H. Complexity and Universality in the Long-Range Order of Words. In *Creativity and Universality in Language*; Degli Esposti, M., Altmann, E.G., Pachet, F., Eds.; Springer: Cham, Germany, 2016; pp. 27–41. doi:10.1007/978-3-319-24403-7_3.
20. Futrell, R.; Mahowald, K.; Gibson, E. Quantifying Word Order Freedom in Dependency Corpora. In *Proceedings of the Third International Conference on Dependency Linguistics*; Uppsala University Press: Uppsala, Sweden, 2015; pp. 91–100.
21. Koplenig, A.; Meyer, P.; Wolfer, S.; Müller-Spitzer, C. The Statistical Trade-off Between Word Order and Word Structure – Large-Scale Evidence for the Principle of Least Effort. *PLOS ONE* **2017**, *12*, e0173614. doi:10.1371/journal.pone.0173614.
22. Piantadosi, S.T.; Tily, H.; Gibson, E. Word Lengths Are Optimized for Efficient Communication. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 3526–3529. doi:10.1073/pnas.1012551108.
23. Mahowald, K.; Fedorenko, E.; Piantadosi, S.T.; Gibson, E. Info/Information Theory: Speakers Choose Shorter Words in Predictive Contexts. *Cognition* **2013**, *126*, 313–318. doi:10.1016/j.cognition.2012.09.010.
24. Ferrer-i-Cancho, R.; Bentz, C.; Seguin, C. Compression and the Origins of Zipf's Law of Abbreviation. *arXiv* **2015**, arXiv:1504.04884.
25. Kanwal, J.; Smith, K.; Culbertson, J.; Kirby, S. Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication. *Cognition* **2017**, *165*, 45–52. doi:10.1016/j.cognition.2017.05.001.
26. Bentz, C.; Verkerk, A.; Kiela, D.; Hill, F.; Buttery, P. Adaptive Communication: Languages with More Non-Native Speakers Tend to Have Fewer Word Forms. *PLoS ONE* **2015**, *10*, e0128254. doi:10.1371/journal.pone.0128254.
27. Kalimeri, M.; Constantoudis, V.; Papadimitriou, C.; Karamanos, K.; Diakonos, F.K.; Papageorgiou, H. Word-length Entropies and Correlations of Natural Language Written Texts. *J. Quant. Linguist.* **2015**, *22*, 101–118. doi:10.1080/09296174.2014.1001636.
28. Ehret, K.; Szmrecsanyi, B. An Information-Theoretic Approach to Assess Linguistic Complexity. In *Complexity, Isolation, and Variation*; Baechler, R., Seiler, G., Eds.; De Gruyter: Berlin, Germany, 2016; pp. 71–94. doi:doi:10.1515/9783110348965-004.
29. Hernández-Gómez, C.; Basurto-Flores, R.; Obregón-Quintana, B.; Guzmán-Vargas, L. Evaluating the Irregularity of Natural Languages. *Entropy* **2017**, *19*, 521. doi:10.3390/e19100521.
30. Bentz, C.; Alikaniotis, D.; Cysouw, M.; Ferrer-i Cancho, R. The Entropy of Words—Learnability and Expressivity across More than 1000 Languages. *Entropy* **2017**, *19*. doi:10.3390/e19060275.
31. Febres, G.; Jaffe, K. Quantifying Structure Differences in Literature Using Symbolic Diversity and Entropy Criteria. *J. Quant. Linguist.* **2017**, *24*, 16–53. doi:10.1080/09296174.2016.1169847.
32. Chang, M.C.; Yang, A.C.C.; Stanley, H.E.; Peng, C.K. Measuring Information-Based Energy and Temperature of Literary Texts. *Phys. A Stat. Mech. Appl.* **2017**, *468*, 783–789. doi:10.1016/j.physa.2016.11.106.
33. Drożdż, S.; Oświęcimka, P.; Kulig, A.; Kwapień, J.; Bazarnik, K.; Grabska-Gradzińska, I.; Rybicki, J.; Stanuszek, M. Quantifying Origin and Character of Long-Range Correlations in Narrative Texts. *Inf. Sci.* **2016**, *331*, 32–44. doi:10.1016/j.ins.2015.10.023.
34. Zipf, G.K. *Human Behavior and the Principle of Least Effort*; Addison-Wesley Press: Cambridge, MA, USA, 1949.
35. Ferrer i Cancho, R.; Solé, R. Least Effort and the Origins of Scaling in Human Language. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 788–791. doi:10.1073/pnas.0335980100.
36. Gold, B.P.; Pearce, M.T.; Mas-Herrero, E.; Dagher, A.; Zatorre, R.J. Predictability and Uncertainty in the Pleasure of Music: A Reward for Learning? *J. Neurosci.* **2019**, *39*, 9397–9409. doi:10.1523/JNEUROSCI.0428-19.2019.

37. Koelsch, S.; Vuust, P.; Friston, K. Predictive Processes and the Peculiar Case of Music. *Trends Cogn. Sci.* **2019**, *23*, 63–77. doi:10.1016/j.tics.2018.10.006.
38. Guillory, J. Canonical and Non-canonical: A Critique of the Current Debate. *ELH* **1987**, *54*, 483–452. doi:doi.org/10.2307/2873219.
39. Even-Zohar, I. Polysystem Studies. *Poet. Today* **1990**, *11*, 9–26. doi:10.2307/1772666.
40. Underwood, T.; Sellers, J. The Long Durée of Literary Prestige. *Mod. Lang. Q.* **2016**, *77*, 321–344. doi:10.1215/00267929-3570634.
41. Gerlach, M.; Font-Clos, F. A Standardized Project Gutenberg Corpus for Statistical Analysis of Natural Language and Quantitative Linguistics. *Entropy* **2020**, *22*, 126. doi:10.3390/e22010126.
42. Green, C. Introducing the Corpus of the Canon of Western Literature: A Corpus for Culturomics and Stylistics. *Lang. Lit.* **2017**, *26*, 282–299. doi:10.1177/0963947017718996.
43. Bloom, H. *The Western Canon: The Books and School of the Ages*; Harcourt: New York, NY, USA, 1994.
44. Reagan, A.J.; Mitchell, L.; Kiley, D.; Danforth, C.M.; Dodds, P.S. The Emotional Arcs of Stories Are Dominated by Six Basic Shapes. *EPJ Data Sci.* **2016**, *5*, 31. doi:10.1140/epjds/s13688-016-0093-1.
45. Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; Manning, C.D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020, pp. 101–108. doi:10.18653/v1/2020.acl-demos.14.
46. Smith, C. *Modes of Discourse. The Local Structure of Texts*; Cambridge University Press: Cambridge, UK, 2003.
47. Eisen, M.; Ribeiro, A.; Segarra, S.; Egan, G. Stylometric analysis of Early Modern period English plays. *Digit. Scholarsh. Humanit.* **2017**, *33*, 500–528. doi:10.1093/llc/fqx059.
48. Segarra, S.; Eisen, M.; Ribeiro, A. Authorship Attribution Through Function Word Adjacency Networks. *IEEE Trans. Signal Process.* **2015**, *63*, 5464–5478. doi:10.1109/TSP.2015.2451111.
49. Brown, P.; Eisen, M.; Segarra, S.; Ribeiro, A.; Egan, G. How the Word Adjacency Network Algorithm Works. *Digit. Scholarsh. Humanit.* **2021**. doi:10.1093/llc/fqab002.
50. Pincus, S.M. Approximate Entropy as a Measure of System Complexity. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 2297–2301.
51. Li, X.; Cui, S.; Voss, L. Using Permutation Entropy to Measure the Electroencephalographic Effects of Sevoflurane. *Anesthesiology* **2008**, *109*, 448–456. doi:10.1097/ALN.0b013e318182a91b.
52. Hayashi, K.; Shigemi, K.; Sawa, T. Neonatal Electroencephalography Shows Low Sensitivity to Anesthesia. *Neurosci. Lett.* **2012**, *517*, 87–91. doi:10.1016/j.neulet.2012.04.028.
53. Lee, G.; Fattinger, S.; Mouthon, A.L.; Noirhomme, Q.; Huber, R. Electroencephalogram Approximate Entropy Influenced by Both Age and Sleep. *Front. Neuroinformatics* **2013**, *7*, 33. doi:10.3389/fninf.2013.00033.
54. Richman, J.; Moorman, J. Physiological Time-Series Analysis Using Approximate Entropy and Sample Entropy. *Am. J. Physiol. Heart Circ. Physiol.* **2000**, *278*, H2039–2049. doi:10.1152/ajpheart.2000.278.6.H2039.
55. Costa, M.; Goldberger, A.L.; Peng, C.K. Multiscale Entropy Analysis of Biological Signals. *Phys. Rev. E* **2005**, *71*, 021906. doi:10.1103/physreve.71.021906.
56. Ahmed, M.; Mandic, D. Multivariate Multiscale Entropy: A Tool for Complexity Analysis of Multichannel Data. *Phys. Rev. Stat. Nonlinear Soft Matter Phys.* **2011**, *84*, 061918. doi:10.1103/PhysRevE.84.061918.
57. Makowski, D.; Pham, T.; Lau, Z.J.; Brammer, J.C.; Lespinasse, F.; Pham, H.; Schölzel, C.; Chen, S.H.A. NeuroKit2: A Python Toolbox for Neurophysiological Signal Processing. *Behav. Res. Methods* **2021**, *53*, 1689–1696. doi:10.3758/s13428-020-01516-y.
58. Zar, J.H. *Biostatistical Analysis*, 5 ed.; Pearson: Upper Saddle River, NJ, USA, 2010.
59. Dietterich, T.G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* **1998**, *10*, 1895–1923. doi:10.1162/089976698300017197.
60. van Cranenburgh, A.; Ketzan, E. Stylometric Literariness Classification: The Case of Stephen King. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 189–197.
61. Menzerath, P.; de Oleza, J. *Spanische Lautdauer. Eine experimentelle Untersuchung*; de Gruyter: Berlin, Germany, 1928.
62. Menzerath, P. *Die Architektonik des deutschen Wortschatzes*; Dümmler: Bonn, Germany, 1954.
63. Altmann, G. Prolegomena to Menzerath's law. *Glottometrika* **1980**, *2*, 1–10.
64. Semple, S.; i Cancho, R.F.; Gustison, M.L. Linguistic laws in biology. *Trends Ecol. Evol.* **2020**, *37*, 53–66. doi:10.1016/j.tree.2021.08.012.
65. Sellis, D. menzerath: Explore Data Following The Menzerath–Altmann Law. R Package Version 0.1.2. 2022. Available online: http://cran.r-project.org/web/packages/mvnfast/vignettes/mvnfast.html (accessed on 8 February 2022)
66. Cortez, P.; Embrechts, M.J. Using Sensitivity Analysis and Visualization Techniques to Open Black Box Data Mining Models. *Inf. Sci.* **2013**, *225*, 1–17. doi:10.1016/j.ins.2012.10.039.
67. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
68. Kernot, D. Can Three Pronouns Discriminate Identity in Writing? In *Data and Decision Sciences in Action*; Springer: Cham, Switzerland, 2018; pp. 397–411. doi:10.1007/978-3-319-55914-8_29.
69. Yu, B. An Evaluation of Text Classification Methods for Literary Study. *Lit. Linguist. Comput.* **2008**, *23*, 327–343. doi:10.1093/llc/fqn015.
70. Qureshi, M.R.; Ranjan, S.; Rajkumar, R.; Shah, K. A Simple Approach to Classify Fictional and Non-Fictional Genres. In *Proceedings of the Second Workshop on Storytelling*; Association for Computational Linguistics: Florence, Italy, 2019; pp. 81–89. doi:10.18653/v1/W19-3409.

71. Grebenkina, M.; Brachmann, A.; Bertamini, M.; Kaduhm, A.; Redies, C. Edge-Orientation Entropy Predicts Preference for Diverse Types of Man-Made Images. *Front. Neurosci.* **2018**, *12*, 678. doi:10.3389/fnins.2018.00678.
72. Stanischewski, S.; Altmann, C.S.; Brachmann, A.; Redies, C. Aesthetic Perception of Line Patterns: Effect of Edge-Orientation Entropy and Curvilinear Shape. *i-Perception* **2020**, *11*. doi:10.1177/2041669520950749.
73. Kraus, N. The Joyful Reduction of Uncertainty: Music Perception as a Window to Predictive Neuronal Processing. *J. Neurosci.* **2020**, *40*, 2790–2792. doi:10.1523/JNEUROSCI.0072-20.2020.
74. Salimpoor, V.N.; Zald, D.H.; Zatorre, R.J.; Dagher, A.; McIntosh, A.R. Predictions and the Brain: How Musical Sounds Become Rewarding. *Trends Cogn. Sci.* **2015**, *19*, 86–91. doi:10.1016/j.tics.2014.12.001.

Supplementary Materials: Approximate Entropy in Canonical and Non-Canonical Fiction

Table S1: List of texts in the Jena Corpus of Expository and Fictional Prose (JEFP Corpus), Version 2.0. Canonical texts were selected from the Corpus of Canonical Western Literature. Non-canonical texts were downloaded from www.smashwords.com, www.goodreads.com, www.feedbooks.com, or Project Gutenberg. Non-fictional texts were sampled from Project Gutenberg.

	Title	**Author(s)**	**Year of Publication**	**Category**
1	Little Dorrit	Charles Dickens	1857	Canonical
2	Oliver Twist	Charles Dickens	1839	Canonical
3	The Life and Adventures of Nicholas Nickleby	Charles Dickens	1839	Canonical
4	The Mystery of Edwin Drood	Charles Dickens	1870	Canonical
5	The Pickwick Papers	Charles Dickens	1836	Canonical
6	Jane Eyre	Charlotte Bronte	1847	Canonical
7	Villette	Charlotte Bronte	1853	Canonical
8	Cranford	Elizabeth Gaskell	1853	Canonical
9	Mary Barton	Elizabeth Gaskell	1848	Canonical
10	North and South	Elizabeth Gaskell	1854	Canonical
11	Agnes Grey	Anne Bronte	1847	Canonical
12	Adam Bede	George Eliot	1859	Canonical
13	Daniel Deronda	George Eliot	1876	Canonical
14	Middlemarch	George Eliot	1872	Canonical
15	Silas Marner	George Eliot	1861	Canonical
16	The Mill on the Floss	George Eliot	1860	Canonical
17	Emma	Jane Austen	1815	Canonical
18	Mansfield Park	Jane Austen	1814	Canonical
19	Persuasion	Jane Austen	1818	Canonical
20	Pride and Prejudice	Jane Austen	1813	Canonical
21	The Picture of Dorian Gray	Oscar Wilde	1890	Canonical
22	The Tenant of Wildfell Hall	Anne Bronte	1848	Canonical
23	Sartor Resartus	Thomas Carlyle	1834	Canonical
24	Old Mortality	Walter Scott	1816	Canonical
25	Redgauntlet	Walter Scott	1824	Canonical
26	The Heart of Midlothian	Walter Scott	1818	Canonical
27	Waverley	Walter Scott	1814	Canonical
28	No Name	Wilkie Collins	1862	Canonical
29	The Moonstone	Wilkie Collins	1868	Canonical
30	The Woman in White	Wilkie Collins	1859	Canonical
31	The History of Henry Esmond	William Makepeace Thackeray	1852	Canonical
32	Vanity Fair	William Makepeace Thackeray	1847	Canonical
33	Dracula	Bram Stoker	1897	Canonical
34	The Well at the World's end	William Morris	1896	Canonical
35	The Narrative of Arthur Gordon Pym	Edgar Allan Poe	1838	Canonical
36	The Ambassadors	Henry James	1903	Canonical
37	The Awkward Age	Henry James	1899	Canonical
38	The Bostonians	Henry James	1886	Canonical
39	The Golden Bowl	Henry James	1904	Canonical
40	The Portrait of a Lady	Henry James	1881	Canonical
41	The Wings of Dove	Henry James	1902	Canonical
42	Moby Dick	Herman Melville	1851	Canonical
43	The Deerslayers	James Fenimore Cooper	1841	Canonical
44	A Christmas Carol	Charles Dickens	1843	Canonical
45	Little Women	Louisa May Alcott	1868	Canonical
46	Puddnhead Wilson	Mark Twain	1893	Canonical
47	The Adventures of Finn	Mark Twain	1884	Canonical
48	The Mysterious Stranger	Mark Twain	1916	Canonical
49	The Marble Faun	Nathaniel Hawthorne	1859	Canonical
50	The Scarlet Letter	Nathaniel Hawthorne	1850	Canonical
51	Walden	Henry David Thoreau	1854	Canonical
52	A Conneticut Yankee in King Arthurs	Mark Twain	1889	Canonical
53	Babbitt	Sinclair Lewis	1922	Canonical
54	A Tale of Two Cities	Charles Dickens	1859	Canonical

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
55	Sister Carrie	Theodore Dreiser	1900	Canonical
56	My Antonia	Willa Cather	1918	Canonical
57	The Old Wives Tale	Arnold Bennett	1908	Canonical
58	Portrait of the Artist as a Young Man	James Joyce	1916	Canonical
59	Ulysses	James Joyce	1922	Canonical
60	Lord Jim	Joseph Conrad	1900	Canonical
61	Nostromo	Joseph Conrad	1904	Canonical
62	The Secret Agent	Joseph Conrad	1907	Canonical
63	Under Western Eyes	Joseph Conrad	1911	Canonical
64	Victory: An Island Tale	Joseph Conrad	1915	Canonical
65	Bleak House	Charles Dickens	1853	Canonical
66	The Rainbow	D. H. Lawrence	1915	Canonical
67	Women in Love	D. H. Lawrence	1920	Canonical
68	Kim	Rudyard Kipling	1901	Canonical
69	Puck of Pooks Hill	Rudyard Kipling	1906	Canonical
70	Jude the Obscure	Thomas Hardy	1895	Canonical
71	Tess of the dUrbervilles	Thomas Hardy	1891	Canonical
72	The Mayor of Casterbridge	Thomas Hardy	1886	Canonical
73	The Return of the Native	Thomas Hardy	1878	Canonical
74	David Copperfield	Charles Dickens	1850	Canonical
75	Great Expectations	Charles Dickens	1860	Canonical
76	Hard Times	Charles Dickens	1854	Canonical
77	A Prisoner in Fairyland	Algernon Blackwood	1913	Non-Canonical
78	The Centaur	Algernon Blackwood	1911	Non-Canonical
79	Ruth Fielding at the War Front	Alice B. Emerson	1918	Non-Canonical
80	The International Spy	Allen Upward	1904	Non-Canonical
81	A Texas Matchmaker	Andy Adams	1904	Non-Canonical
82	The Filigree Ball	Anna Katharine Green	1903	Non-Canonical
83	Looking Further Backward	Arthur Dudley Vinton	1890	Non-Canonical
84	The Hill Of Dreams	Arthur Machen	1907	Non-Canonical
85	Jean of the Lazy A	Burton E. Stevenson	1915	Non-Canonical
86	The Gloved Hand	Baroness Emma Orczy	1913	Non-Canonical
87	The Filibusters	Charles John Cutcliffe Wright Hyne	1900	Non-Canonical
88	Wunpost	Dane Coolidge	1920	Non-Canonical
89	Love Insurance	Earl Derr Biggers	1914	Non-Canonical
90	The Wouldbegoods	Edith Nesbit	1899	Non-Canonical
91	Wet Magic	Edith Nesbit	1913	Non-Canonical
92	An Amiable Charlatan	Edward Phillips Oppenheim	1916	Non-Canonical
93	The Double Traitor	Edward Phillips Oppenheim	1915	Non-Canonical
94	The Zeppelin 's Passenger	Edward Phillips Oppenheim	1918	Non-Canonical
95	The People of the Ruins	Edward Shanks	1920	Non-Canonical
96	The Honor of the Name	Ernest Bramah Smith	1891	Non-Canonical
97	The Riddle of the Sands	Eugene Percy Lyle	1903	Non-Canonical
98	The Missourian	Ford Madox Ford	1905	Non-Canonical
99	Privy Seal	Frederic Arnold Kummer	1907	Non-Canonical
100	Condemned as a Nihilist	Harold MacGrath	1893	Non-Canonical
101	The Afterglow	George Allan England	1913	Non-Canonical
102	The Flying Legion	George Allan England	1920	Non-Canonical
103	West Wind Drift	George Barr McCutcheon	1920	Non-Canonical
104	Trilby	George W. Ogden	1894	Non-Canonical
105	Olga Romanoff or , The Syren of the Skies	George F. Worts	1894	Non-Canonical
106	The Princess and Curdie	George Griffith	1883	Non-Canonical
107	The Adventures of Don Lavington	George MacDonald	1896	Non-Canonical
108	A Voyage to the Moon	George Manville Fenn	1827	Non-Canonical
109	Man on the Box	Harold MacGrath	1904	Non-Canonical
110	The Puppet Crown	Homer Eon Flint	1901	Non-Canonical
111	Men of Iron	Ida Alexa Ross Wylie	1891	Non-Canonical
112	Towards Morning	James Branch Cabell	1918	Non-Canonical
113	A Strange Manuscript Found in a Copper Cylinder	James De Mille	1888	Non-Canonical
114	Lost in the Fog	James Malcom Rymer	1870	Non-Canonical
115	Varney the Vampire	James Oliver Curwood	1847	Non-Canonical
116	The Danger Trail	John Meade Falkner	1910	Non-Canonical

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
117	The Lost Stradivarius	John Meade Falkner	1895	Non-Canonical
118	The Nebuly Coat	Joseph Hocking	1903	Non-Canonical
119	The Weapons of Mystery	Joseph Smith Fletcher	1890	Non-Canonical
120	Diane of the Green Van	Lord Dunsany	1914	Non-Canonical
121	The Treasure Trail	Mary E. Bradley	1918	Non-Canonical
122	Mizora : A Prophecy	Mary Roberts Rinehart	1889	Non-Canonical
123	Across the Zodiac	Percy Greg	1880	Non-Canonical
124	Bardelys the Magnificent	Rafael Sabatini	1905	Non-Canonical
125	Soldiers of Fortune	Richard Harding Davis	1897	Non-Canonical
126	The Beetle	Richard Marsh	1897	Non-Canonical
127	The Triumphs of Eugne Valmont	Robert Barr	1906	Non-Canonical
128	Erling the Bold	Robert Michael Ballantyne	1869	Non-Canonical
129	The Dog Crusoe and His Master	Robert Michael Ballantyne	1894	Non-Canonical
130	Ailsa Paige	Robert William Chambers	1910	Non-Canonical
131	In Search of the Unknown	Robert William Chambers	1904	Non-Canonical
132	In the Quarter	Robert William Chambers	1894	Non-Canonical
133	Erewhon , or Over The Range	Samuel Butler	1910	Non-Canonical
134	The road to Frontenac	Samuel Merwin	1901	Non-Canonical
135	The Revolt of Man	Sir Walter Besant	1882	Non-Canonical
136	The Brass Bottle	Thomas Anstey Guthrie	1900	Non-Canonical
137	The Doomsman	Van Tassel Sutphen	1906	Non-Canonical
138	The Border Legion	Zane Grey	1916	Non-Canonical
139	The Daltons; Or, Three Roads In Life. Volume I (of II)	Charles James Lever	1850	Non-Canonical
140	Melmoth the Wanderer, Vol. 3	Charles Robert Maturin	1820	Non-Canonical
141	Melmoth the Wanderer, Vol. 2	Charles Robert Maturin	1820	Non-Canonical
142	The Wanderer; or, Female Difficulties (Volume 4 of 5)	Fanny Burney	1814	Non-Canonical
143	The Wanderer; or, Female Difficulties (Volume 2 of 5)	Fanny Burney	1814	Non-Canonical
144	The Wanderer; or, Female Difficulties (Volume 3 of 5)	Fanny Burney	1814	Non-Canonical
145	The Vicar of Wrexhill	Frances Milton Trollope	1837	Non-Canonical
146	The Fortunes of the Colville Family; or, A Cloud with its Silver Lining	Frank Edward Smedley	1853	Non-Canonical
147	Harry Coverdale's Courtship, and All That Came of It	Frank Edward Smedley	1854	Non-Canonical
148	Lewis Arundel; Or, The Railroad Of Life	Frank Edward Smedley	1852	Non-Canonical
149	The Little Savage	Frederick Marryat	1848	Non-Canonical
150	Newton Forster	Frederick Marryat	1832	Non-Canonical
151	Snarleyyow; or, The Dog Fiend	Frederick Marryat	1837	Non-Canonical
152	Travels and Adventures of Monsieur Violet	Frederick Marryat	1843	Non-Canonical
153	The Privateer's-Man, One hundred Years Ago	Frederick Marryat	1846	Non-Canonical
154	The Little Savage	Frederick Marryat	1848	Non-Canonical
155	Snarleyyow, or, the Dog Fiend	Frederick Marryat	1837	Non-Canonical
156	Newton Forster; Or, The Merchant Service	Frederick Marryat	1832	Non-Canonical
157	Mr. Midshipman Easy	Frederick Marryat	1836	Non-Canonical
158	Arrah Neil; or, Times of Old	George Payne Rainsford James	1843	Non-Canonical
159	The Castle of Ehrenstein: Its Lords Spiritual and Temporal; Its Inhabitants Earthly and Unearthly	George Payne Rainsford James	1847	Non-Canonical
160	Forest Days: A Romance of Old Times	George Payne Rainsford James	1843	Non-Canonical
161	A Voyage to the Moon: With Some Account of the Manners and Customs, Science and Philosophy, of the People of Morosofia, and Other Lunarians	George Tucker	1827	Non-Canonical
162	Market Harborough, and Inside the Bar	George John Whyte-Melville	1858	Non-Canonical
163	The Gladiators. A Tale of Rome and Judæa	George John Whyte-Melville	1863	Non-Canonical
164	The Mother's Recompense, Volume 2: A Sequel to Home Influence	Grace Aguilar	1874	Non-Canonical
165	The Mother's Recompense, Volume 1: A Sequel to Home Influence	Grace Aguilar	1874	Non-Canonical
166	The Vale of Cedars; Or, The Martyr	Grace Aguilar	1850	Non-Canonical
167	Jasper Lyle	Mary Augusta Ward	1851	Non-Canonical
168	The Eskdale Herd-boy: A Scottish Tale for the Instruction and Amusement of Young People	Martha Blackford	1819	Non-Canonical
169	Mary Erskine	Jacob Abbott	1850	Non-Canonical
170	Bruno; or, lessons of fidelity, patience, and self-denial taught by a dog	Jacob Abbott	1854	Non-Canonical
171	Rollo in Rome	Jacob Abbott	1858	Non-Canonical

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
172	Rollo in London	Jacob Abbott	1850	Non-Canonical
173	Among the Brigands	James De Mille	1871	Non-Canonical
174	The Lily and the Cross: A Tale of Acadia	James De Mille	1875	Non-Canonical
175	Fire in the Woods: Illustrated	James De Mille	1872	Non-Canonical
176	Cord and Creese	James De Mille	1869	Non-Canonical
177	The Three Perils of Man; or, War, Women, and Witchcraft, Vol. 3 (of 3)	James Hogg	1822	Non-Canonical
178	The Three Perils of Man; or, War, Women, and Witchcraft, Vol. 2 (of 3)	James Hogg	1822	Non-Canonical
179	The Provost	John Galt	1822	Non-Canonical
180	Ringan Gilhaize, or, The Covenanters	John Galt	1823	Non-Canonical
181	Valerius. A Roman Story	John Gibson Lockhart	1821	Non-Canonical
182	The Manoeuvring Mother (vol. 3 of 3)	Lady Charlotte Susan Maria Bury	1842	Non-Canonical
183	The Manoeuvring Mother (vol. 2 of 3)	Lady Charlotte Susan Maria Bury	1842	Non-Canonical
184	The Manoeuvring Mother (vol. 1 of 3)	Lady Charlotte Susan Maria Bury	1842	Non-Canonical
185	The Annals of the Poor	Legh Richmond	1814	Non-Canonical
186	Aunt Kitty's Tales	Maria Jane McIntosh	1847	Non-Canonical
187	Camperdown; or, News from our neighbourhood	Mary Griffith	1836	Non-Canonical
188	The Actress' Daughter: A Novel	May Agnes Fleming	1879	Non-Canonical
189	Emilie the Peacemaker	Mrs. Thomas Geldart	1851	Non-Canonical
190	The Old Church Clock	Richard Parkinson	1843	Non-Canonical
191	Sheppard Lee, Written by Himself. Vol. 1 (of 2)	Robert Montgomery Bird	1836	Non-Canonical
192	The Young Trail Hunters: Or, the Wild Riders of the Plains. The Veritable Adventures of Hal Hyde and Ned Brown, on Their Journey Across the Great Plains of the South-West	Samuel Woodworth Cozzens	1876	Non-Canonical
193	Watch—Work—Wait: Or, The Orphan's Victory	Sarah Ann Myers	1859	Non-Canonical
194	Pine Needles	Susan Bogert Warner	1877	Non-Canonical
195	Confession; Or, The Blind Heart. A Domestic Story	William Gilmore Simms	1841	Non-Canonical
196	Antony Waymouth; Or, The Gentlemen Adventurers	William Henry Giles Kingston	1865	Non-Canonical
197	Clara Maynard; Or, The True and the False: A Tale of the Times	William Henry Giles Kingston	1877	Non-Canonical
198	The Story of Nelson: also "The Grateful Indian", "The Boatswain's Son"	William Henry Giles Kingston	1860	Non-Canonical
199	Off to Sea: The Adventures of Jovial Jack Junker on his Road to Fame	William Henry Giles Kingston	1870	Non-Canonical
200	Fred Markham in Russia; Or, The Boy Travellers in the Land of the Czar	William Henry Giles Kingston	1858	Non-Canonical
201	In New Granada; Or, Heroes and Patriots	William Henry Giles Kingston	1879	Non-Canonical
202	Roger Kyffin's Ward	William Henry Giles Kingston	1874	Non-Canonical
203	Caxton's Book: A Collection of Essays, Poems, Tales, and Sketches.	William Henry Rhodes	1876	Non-Canonical
204	Blue-Stocking Hall, (Vol. 2 of 3)	William Pitt Scargill	1827	Non-Canonical
205	Blue-Stocking Hall, (Vol. 3 of 3)	William Pitt Scargill	1827	Non-Canonical
206	Aurelian; or, Rome in the Third Century	William Ware	1838	Non-Canonical
207	Scottish Cathedrals and Abbeys	Dugald Butler	1901	Non-Fictional
208	A Text-Book of the History of Architecture: Seventh Edition, revised	Alfred Dwight Foster Hamlin	1896	Non-Fictional
209	Japanese Homes and Their Surroundings	Edward Sylvester Morse	1885	Non-Fictional
210	The Architecture of Provence and the Riviera	David MacGibbon	1888	Non-Fictional
211	Historic Ornament, Vol. 2 (of 2): Treatise on decorative art and architectural ornament	James Ward	1897	Non-Fictional
212	How to Study Architecture	Charles Henry Caffin	1917	Non-Fictional
213	A Dictionary of Slang, Cant, and Vulgar Words: Used at the Present Day in the Streets of London; the Universities of Oxford and Cambridge; the Houses of Parliament; the Dens of St. Giles; and the Palaces of St. James.	John Camden Hotten	1860	Non-Fictional
214	THE ENCYCLOPAEDIA BRITANNICA-Vol 1	University of Cambridge	1910	Non-Fictional
215	THE ENCYCLOPAEDIA BRITANNICA-Vol 2	University of Cambridge	1910	Non-Fictional
216	Through the Brazilian Wilderness	Theodore Roosevelt	1914	Non-Fictional

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
217	Gold, Sport, and Coffee Planting in Mysore: With chapters on coffee planting in Coorg, the Mysore representative assembly, the Indian congress, caste and the Indian silver question, being the 38 years' experiences of a Mysore planter	Robert Henry Elliot	1898	Non-Fictional
218	The Economic Aspect of Geology	Charles Kenneth Leith	1921	Non-Fictional
219	The Shores of the Adriatic: The Austrian Side, The Küstenlande, Istria, and Dalmatia	Frederick Hamilton Jackson	1906	Non-Fictional
220	Island Life; Or, The Phenomena and Causes of Insular Faunas and Floras	Alfred Russel Wallace	1880	Non-Fictional
221	Sea and Sardinia	David Herbert Lawrence	1921	Non-Fictional
222	Sketches from the Subject and Neighbour Lands of Venice	Edward Augustus Freeman	1881	Non-Fictional
223	The Principles of Stratigraphical Geology	John Edward Marr	1898	Non-Fictional
224	Babylonian and Assyrian Laws, Contracts and Letters	Claude Hermann Walter Johns	1904	Non-Fictional
225	Putnam's Handy Law Book for the Layman	Albert Sidney Bolles	1921	Non-Fictional
226	Marriage and Divorce Laws of the World	Hyacinthe Ringrose	1911	Non-Fictional
227	The Law and the Poor	Sir Edward Abbott Parry	1914	Non-Fictional
228	International Law	George Grafton Wilson, George Fox Tucker	1901	Non-Fictional
229	The Criminal Prosecution and Capital Punishment of Animals	Edward Payson Evans	1906	Non-Fictional
230	The English Constitution	Walter Bagehot	1867	Non-Fictional
231	The Law of the Sea: A manual of the principles of admiralty law for students, mariners, and ship operators	George L. Canfield, George W. Dalzell, J. Y. Brinton	1921	Non-Fictional
232	Woman and the Republic: A Survey of the Woman-Suffrage Movement in the United States and a Discussion of the Claims and Arguments of Its Foremost Advocates	Helen Kendrick Johnson	1897	Non-Fictional
233	The American Judiciary	Simeon Eben Baldwin	1905	Non-Fictional
234	A Practical Physiology: A Text-Book for Higher Schools	Albert Franklin Blaisdell	1897	Non-Fictional
235	Amusements in Mathematics	Henry Ernest Dudeney	1917	Non-Fictional
236	On the Genesis of Species	St. George Jackson Mivart	1871	Non-Fictional
237	Great Astronomers	Robert Stawell Ball	1895	Non-Fictional
238	Evolution, Old & New: Or, the Theories of Buffon, Dr. Erasmus Darwin and Lamarck; as compared with that of Charles Darwin	Samuel Butler	1879	Non-Fictional
239	Darwin, and After Darwin, Volumes 1 and 3: An Exposition of the Darwinian Theory and a Discussion of Post-Darwinian Questions	George John Romanes	1892	Non-Fictional
240	Creative Evolution	Henri Bergson	1907	Non-Fictional
241	Myths and Marvels of Astronomy	Richard Anthony Proctor	1877	Non-Fictional
242	A Popular History of Astronomy During the Nineteenth Century: Fourth Edition	Agnes Mary AClerke	1887	Non-Fictional
243	A Text-Book of Astronomy	George Cary Comstock	1901	Non-Fictional
244	Astronomical Myths: Based on Flammarions's Ḧistory of the Heavens´	Camille Flammarion, John Frederick Blake	1877	Non-Fictional
245	Darwin, and After Darwin, Volume 2 of 3: Post-Darwinian Questions: Heredity and Utility	George John Romanes	1892	Non-Fictional
246	A Civic Biology, Presented in Problems	George William Hunter	1914	Non-Fictional
247	Physics	Willis E. Tower, Charles M. Turton, Charles H. Smith, Thomas D. Cope	1920	Non-Fictional
248	A Century of Science, and Other Essays	John Fiske	1899	Non-Fictional
249	Side-Lights on Astronomy and Kindred Fields of Popular Science	Simon Newcomb	1906	Non-Fictional
250	Elementary Zoology, Second Edition	Vernon Lyman Kellogg	1901	Non-Fictional
251	Experiments on Animals	Stephen Paget	1888	Non-Fictional
252	The Sea-beach at Ebb-tide: A Guide to the Study of the Seaweeds and the Lower Animal Life Found Between Tide-marks	Augusta Foote Arnold	1901	Non-Fictional
253	The Science and Philosophy of the Organism	Hans Driesch	1908	Non-Fictional

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
254	The Organism as a Whole, from a Physicochemical Viewpoint	Jacques Loeb	1916	Non-Fictional
255	A Guide to the Study of Fishes, Volume 1 (of 2)	David Starr Jordan	1905	Non-Fictional
256	Evolution: Its nature, its evidence, and its relation to religious thought	Joseph LeConte	1888	Non-Fictional
257	The Races of Man: An Outline of Anthropology and Ethnography	Joseph Deniker	1900	Non-Fictional
258	Observations of a Naturalist in the Pacific Between 1896 and 1899, Volume 1: Vanua Levu, Fiji	Henry Brougham Guppy	1903	Non-Fictional
259	Animal Life and Intelligence	Conwy Lloyd Morgan	1890	Non-Fictional
260	Stargazing: Past and Present	Sir Joseph Norman Lockyer	1878	Non-Fictional
261	Observations of a Naturalist in the Pacific Between 1896 and 1899, Volume 2: Plant-Dispersal	Henry Brougham Guppy	1903	Non-Fictional
262	The Logic of Chance, 3rd edition: An Essay on the Foundations and Province of the Theory of Probability, With Especial Reference to Its Logical Bearings and Its Application to Moral and Social Science and to Statistics	John Venn	1888	Non-Fictional
263	Biology and Its Makers: With Portraits and Other Illustrations	William Albert Locy	1908	Non-Fictional
264	The Crayfish: An Introduction to the Study of Zoology.	Thomas Henry Huxley	1880	Non-Fictional
265	History of Botany (1530-1860)	Julius Sachs	1875	Non-Fictional
266	The Universal Kinship	John Howard Moore	1906	Non-Fictional
267	The philosophy of biology	James Johnstone	1914	Non-Fictional
268	Hygienic Physiology : with Special Reference to the Use of Alcoholic Drinks and Narcotics	Joel Dorman Steele	1884	Non-Fictional
269	Species and Varieties, Their Origin by Mutation	Hugo de Vries	1905	Non-Fictional
270	The Naturalist in La Plata	William Henry Hudson	1892	Non-Fictional
271	Studies in the Psychology of Sex, Volume 1: The Evolution of Modesty; The Phenomena of Sexual Periodicity; Auto-Erotism	Havelock Ellis	1900	Non-Fictional
272	Studies in the Psychology of Sex, Volume 2: Sexual Inversion	Havelock Ellis	1900	Non-Fictional
273	The Mind of the Child, Part II: The Development of the Intellect, International Education; Series Edited By William T. Harris, Volume IX.	William T. Preyer	1888	Non-Fictional
274	The Measurement of Intelligence: An Explanation of and a Complete Guide for the Use of the; Stanford Revision and Extension of the Binet-Simon; Intelligence Scale	Lewis Madison Terman	1916	Non-Fictional
275	Human Traits and their Social Significance	Irwin Edman	1919	Non-Fictional
276	Human Personality and Its Survival of Bodily Death	Frederic William Henry Myers	1903	Non-Fictional
277	Mysterious Psychic Forces: An Account of the Author's Investigations in Psychical Research, Together with Those of Other European Savants	Camille Flammarion	1907	Non-Fictional
278	The Group Mind: A Sketch of the Principles of Collective Psychology: With Some Attempt to Apply Them to the Interpretation of National Life and Character	William McDougall	1920	Non-Fictional
279	On the State of Lunacy and the Legal Provision for the Insane: With Observations on the Construction and Organization of Asylums	John Thomas Arlidge	1859	Non-Fictional
280	The Criminal	Havelock Ellis	1890	Non-Fictional
281	Fact and Fable in Psychology	Joseph Jastrow	1900	Non-Fictional
282	A Beginner's Psychology	Edward Bradford Titchener	1915	Non-Fictional
283	The Law of Psychic Phenomena: A working hypothesis for the systematic study of hypnotism, spiritism, mental therapeutics, etc.	Thomson Jay Hudson	1893	Non-Fictional
284	Psychology: Briefer Course	William James	1892	Non-Fictional
285	The Principles of Psychology, Volume 1 (of 2)	William James	1890	Non-Fictional
286	The Principles of Psychology, Volume 2 (of 2)	William James	1890	Non-Fictional
287	Browning as a Philosophical and Religious Teacher	Sir Jones, Henry	1891	Non-Fictional
288	The Life of Reason: The Phases of Human Progress	George Santayana	1905	Non-Fictional
289	An Introduction to Philosophy	George Stuart Fullerton	1906	Non-Fictional
290	The Approach to Philosophy	Ralph Barton Perry	1905	Non-Fictional

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
291	The Will to Believe, and Other Essays in Popular Philosophy	William James	1896	Non-Fictional
292	Christianity and Greek Philosophy: or, the relation between spontaneous and reflective thought in Greece and the positive teaching of Christ and His Apostles	Benjamin Franklin Cocker	1870	Non-Fictional
293	A History of Mediaeval Jewish Philosophy	Isaac Husik	1916	Non-Fictional
294	The Philosophy of Friedrich Nietzsche	Henry Louis Mencken	1908	Non-Fictional
295	Philosophical Studies	George Edward Moore	1883	Non-Fictional
296	What Nietzsche Taught	Willard Huntington Wright	1915	Non-Fictional
297	An ethical philosophy of life presented in its main outlines	Felix Adler	1918	Non-Fictional
298	A Beginner's History of Philosophy, Vol. 1: Ancient and Medieval Philosophy	Herbert Ernest Cushman	1910	Non-Fictional
299	Towards the Great Peace	Ralph Adams Cram	1922	Non-Fictional
300	Criminal Man, According to the Classification of Cesare Lombroso	Gina Lombroso	1880	Non-Fictional
301	Criminal Sociology	Enrico Ferri	1895	Non-Fictional
302	Community Civics and Rural Life	Arthur William Dunn	1920	Non-Fictional
303	Sociology and Modern Social Problems	Charles Abram Ellwood	1910	Non-Fictional
304	The Theory of the Leisure Class	Thorstein Veblen	1899	Non-Fictional
305	An Historical View of the Philippine Islands, Vol 1 (of 2): Exhibiting their discovery, population, language, government, manners, customs, productions and commerce.	Joaquin Martinez De Zugniga	1814	Non-Fictional
306	An Historical View of the Philippine Islands, Vol 2 (of 2): Exhibiting their discovery, population, language, government, manners, customs, productions and commerce.	Joaquin Martinez De Zugniga	1814	Non-Fictional
307	History of the Buccaneers of America	James Burney	1816	Non-Fictional
308	The Natural History of Cage Birds: Their Management, Habits, Food, Diseases, Treatment, Breeding, and the Methods of Catching Them.	Johann Matthäus Bechstein	1838	Non-Fictional
309	A System of Pyrotechny: Comprehending the theory and practice, with the application of chemistry; designed for exhibition and for war.	James Cutbush	1825	Non-Fictional
310	The History of the Inquisition of Spain from the Time of its Establishment to the Reign of Ferdinand VII.	Juan Antonio Llorente	1825	Non-Fictional
311	History, Manners, and Customs of the Indian Nations Who Once Inhabited Pennsylvania and the Neighbouring States.	John Gottlieb Ernestus Heckewelder	1818	Non-Fictional
312	On The Principles of Political Economy, and Taxation	David Ricardo	1819	Non-Fictional
313	Pedestrianism; or, An Account of the Performances of Celebrated Pedestrians During the Last and Present Century.: With a full narrative of Captain Barclay's public and private matches; and an essay on training.	Walter Thom	1813	Non-Fictional
314	The Grounds of Christianity Examined by Comparing The New Testament with the Old	George Bethune English	1813	Non-Fictional
315	An Account of The Kingdom of Nepal: And of the Territories Annexed to this Dominion by the House of Gorkha	Francis Hamilton	1819	Non-Fictional
316	The Logic of Hegel	Georg Wilhelm Friedrich Hegel	1812	Non-Fictional
317	Hegel's Philosophy of Mind	Georg Wilhelm Friedrich Hegel	1817	Non-Fictional
318	A Historical Survey of the Customs, Habits, & Present State of the Gypsies	John Hoyland	1816	Non-Fictional
319	Not Paul, But Jesus	Jeremy Bentham	1823	Non-Fictional
320	Aids to Reflection; and, The Confessions of an Inquiring Spirit	Samuel Taylor Coleridge	1825	Non-Fictional
321	The Dance of Death: Exhibited in Elegant Engravings on Wood with a Dissertation on the Several Representations of that Subject but More Particularly on Those Ascribed to Macaber and Hans Holbein	Francis Douce	1833	Non-Fictional

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
322	Definitions in Political Economy,: Preceded by an Inquiry Into the Rules which Ought to Guide Political Economists in the Definition and Use of Their Terms; with Remarks on the Deviation from These Rules in Their Writings	Thomas Robert Malthus	1853	Non-Fictional
323	Cottage Economy, to Which is Added The Poor Man's Friend	William Cobbett	1833	Non-Fictional
324	Indian Nullification of the Unconstitutional Laws of Massachusetts Relative to the Marshpee Tribe: Or, the Pretended Riot Explained	William Apess	1835	Non-Fictional
325	Slavery	William Ellery Channing	1835	Non-Fictional
326	Thoughts on Missions	Sheldon Dibble	1850	Non-Fictional
327	A Portraiture of Quakerism, Volume 2: Taken from a View of the Education and Discipline, Social Manners, Civil and Political Economy, Religious Principles and Character, of the Society of Friends	Thomas Clarkson	1841	Non-Fictional
328	The Field Book: or, Sports and pastimes of the United Kingdom: compiled from the best authorities, ancient and modern	William Hamilton Maxwell	1833	Non-Fictional
329	Cosmos: A Sketch of a Physical Description of the Universe	Frédéric Bastiat	1853	Non-Fictional
330	Elements of Physiophilosophy	Lorenz Oken	1847	Non-Fictional
331	A Synopsis of the Birds of North America	Alexander von Humboldt	1845	Non-Fictional
332	The Practical Astronomer: Comprising illustrations of light and colours–practical descriptions of all kinds of telescopes–the use of the equatorial-transit–circular, and other astronomical instruments, a particular account of the Earl of Rosse's large telescopes, and other topics connected with astronomy	Thomas Dick	1850	Non-Fictional
333	The Gastronomic Regenerator: A Simplified and Entirely New System of Cookery: With Nearly Two Thousand Practical Receipts Suited to the Income of All Classes	Alexis Soyer	1846	Non-Fictional
334	The Pantropheon; Or, History of Food, Its Preparation, from the Earliest Ages of the World	Alexis Soyer	1850	Non-Fictional
335	Miss Leslie's New Cookery Book	Eliza Leslie	1867	Non-Fictional
336	Conversations on Chemistry, V. 1-2: In Which the Elements of that Science Are Familiarly Explained and Illustrated by Experiments	Jane Haldimand Marcet	1847	Non-Fictional
337	Conversations on Natural Philosophy, in which the Elements of that Science are Familiarly Explained	Jane Haldimand Marcet	1836	Non-Fictional
338	Botany for Ladies: or, A Popular Introduction to the Natural System of Plants, According to the Classification of De Candolle.	Jane Loudon	1815	Non-Fictional
339	American Institutions and Their Influence	Alexis de Tocqueville	1851	Non-Fictional
340	The Steam Engine Explained and Illustrated (Seventh Edition): With an Account of Its Invention and Progressive Improvement, and Its Application to Navigation and Railways; Including Also a Memoir of Watt	Dionysius Lardner	1840	Non-Fictional
341	History of the State of California: From the Period of the Conquest by Spain to Her Occupation by the United States of America	John Frost	1851	Non-Fictional
342	History of the Conquest of Mexico; vol. 3/4	William Hickling Prescott	1857	Non-Fictional
343	Norman's New Orleans and Environs: Containing a Brief Historical Sketch of the Territory and State of Louisiana and the City of New Orleans, from the Earliest Period to the Present Time	Benjamin Moore Norman	1842	Non-Fictional
344	The Philosophy of Health; Volume 1 (of 2): or, an exposition of the physical and mental constitution of man	Southwood Smith	1847	Non-Fictional
345	History of Brighthelmston; or, Brighton as I View it and Others Knew It: With a Chronological Table of Local Events	John Ackerson Erredge	1851	Non-Fictional

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
346	Summary Narrative of an Exploratory Expedition to the Sources of the Mississippi River, in 1820: Resumed and Completed, by the Discovery of its Origin in Itasca Lake, in 1832	Henry Rowe Schoolcraft	1836	Non-Fictional
347	The Infant System: For Developing the Intellectual and Moral Powers of all Children, from One to Seven years of Age	Samuel Wilderspin	1850	Non-Fictional
348	Bulfinch's Mythology: The Age of Fable; The Age of Chivalry; Legends of Charlemagne	Thomas Bulfinch	1862	Non-Fictional
349	Dealings with the Dead, Volume 2 (of 2)	Lucius Manlius Sargent	1846	Non-Fictional
350	Science for the School and Family, Part I. Natural Philosophy	Worthington Hooker	1853	Non-Fictional
351	On the various forces of nature and their relations to each other	Michael Faraday	1847	Non-Fictional
352	British Bees: An Introduction into the Studies of the Natural History and Economy of the Bees Indigenous to the British Isles	William Edward Shuckard	1859	Non-Fictional
353	Gunnery in 1858: Being a Treatise on Rifles, Cannon, and Sporting Arms: Explaining the Principles of the Science of Gunnery, and Describing the Newest Improvements in Fire-Arms	William Greener	1846	Non-Fictional
354	The Sabbath-School Index: Pointing out the history and progress of Sunday-schools, with approved modes of instruction.	Richard Gay Pardee	1842	Non-Fictional
355	The Opium Habit	Horace B. Day	1860	Non-Fictional
356	History of Greece, Volume 12 (of 12)	George Grote	1844	Non-Fictional
357	History of Greece, Volume 03 (of 12)	George Grote	1852	Non-Fictional
358	History of Greece, Volume 05 (of 12)	George Grote	1852	Non-Fictional
359	History of Greece, Volume 04 (of 12)	George Grote	1838	Non-Fictional
360	Reflections on the Decline of Science in England, and on Some of Its Causes	Charles Babbage	1857	Non-Fictional
361	On the Connexion of the Physical Sciences	Mary Somerville	1880	Non-Fictional
362	Knowledge Is Power:: A View of the Productive Forces of Modern Society and the Results of Labor, Capital and Skill.	Charles Knight	1825	Non-Fictional
363	Athens: Its Rise and Fall, Book II	Edward George Bulwer-Lytton	1870	Non-Fictional
364	A Popular History of England, From the Earliest Times to the Reign of Queen Victoria; Vol. I	François Guizot	1837	Non-Fictional
365	Elements of Agricultural Chemistry	Thomas Anderson	1852	Non-Fictional
366	Practical Guide to English Versification: With a Compendious Dictionary of Rhymes, an Examination; of Classical Measures, and Comments Upon Burlesque and; Comic Verse, Vers de Société, and Song-writing	Tom Hood	1838	Non-Fictional
367	A Manual of Elementary Geology: or, The Ancient Changes of the Earth and its Inhabitants as Illustrated by Geological Monuments	Sir Charles Lyell	1844	Non-Fictional
368	Health and Education	Charles Kingsley	1837	Non-Fictional
369	Stones of the Temple; Or, Lessons from the Fabric and Furniture of the Church	Walter Field	1847	Non-Fictional
370	History of the United Netherlands from the Death of William the Silent to the Twelve Year's Truce — Complete (1600-1609)	John Lothrop Motley	1849	Non-Fictional
371	A Dictionary of English Synonymes and Synonymous or Parallel Expressions: Designed as a Practical Guide to Aptness and Variety of Phraseology	Richard Soule	1838	Non-Fictional
372	The Rise of the Dutch Republic — Complete (1566-74)	John Lothrop Motley	1871	Non-Fictional
373	A History of Domestic Manners and Sentiments in England During the Middle Ages	Thomas Wright	1859	Non-Fictional
374	The History, Theory, and Practice of Illuminating: Condensed from 'The Art of Illuminating' by the same illustrator and author	Matthew Digby Wyatt	1860	Non-Fictional
375	An Architect's Note-Book in Spain: principally illustrating the domestic architecture of that country.	Matthew Digby Wyatt	1858	Non-Fictional
376	History of Lace	Fanny Bury Palliser	1862	Non-Fictional

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
377	The Physical Basis of Mind: Being the Second Series of Problems of Life and Mind.	George Henry Lewes	1853	Non-Fictional
378	Lectures on the rise and development of medieval architecture; vol. 2	George Gilbert Scott	1872	Non-Fictional
379	The History of Ancient America, Anterior to the Time of Columbus: Proving the Identity of the Aborigines with the Tyrians and Israelites; and the Introduction of Christianity into the Western Hemisphere By The Apostle St. Thomas	George Jones	1838	Non-Fictional
380	The Subterranean World	Georg Hartwig	1871	Non-Fictional
381	Guano: A Treatise of Practical Information for Farmers	Solon Robinson	1852	Non-Fictional
382	Wild Wales: The People, Language, & Scenery	George Borrow	1862	Non-Fictional
383	History of Indian and Eastern Architecture	James Fergusson	1876	Non-Fictional
384	Parasites: A Treatise on the Entozoa of Man and Animals: Including Some Account of the Ectozoa	Thomas Spencer Cobbold	1879	Non-Fictional
385	History of American Socialisms	John Humphrey Noyes	1869	Non-Fictional
386	London Labour and the London Poor, Vol. 2	Henry Mayhew	1851	Non-Fictional
387	Companion to the Bible	Elijah Porter Barrows	1867	Non-Fictional
388	The Non-religion of the Future: A Sociological Study	Jean-Marie Guyau	1887	Non-Fictional
389	Ten Great Religions: An Essay in Comparative Theology	James Freeman Clarke	1871	Non-Fictional
390	A History of Oregon, 1792-1849: Drawn From Personal Observation and Authentic Information	William Henry Gray	1870	Non-Fictional
391	Bible Animals;: Being a Description of Every Living Creature Mentioned in the Scripture, from the Ape to the Coral.	John George Wood	1869	Non-Fictional

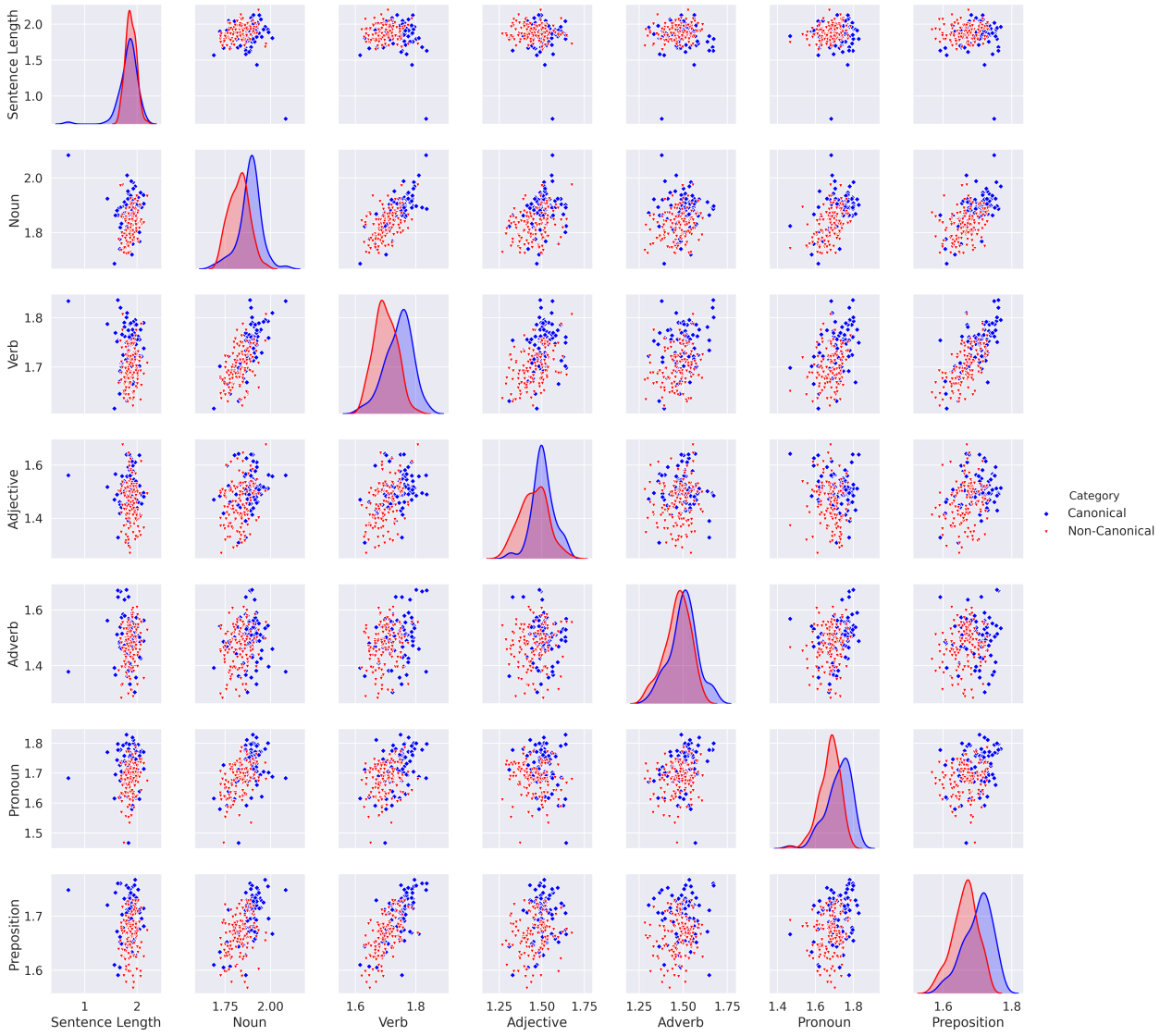

Figure S1. Pair-plot of all Approximate Entropy (ApEn) features in fictional/canonical and fictional/non-canonical texts. While each non-diagonal plot shows the relationships between two features, the main-diagonal subplots visualize the univariate distributions of each feature.

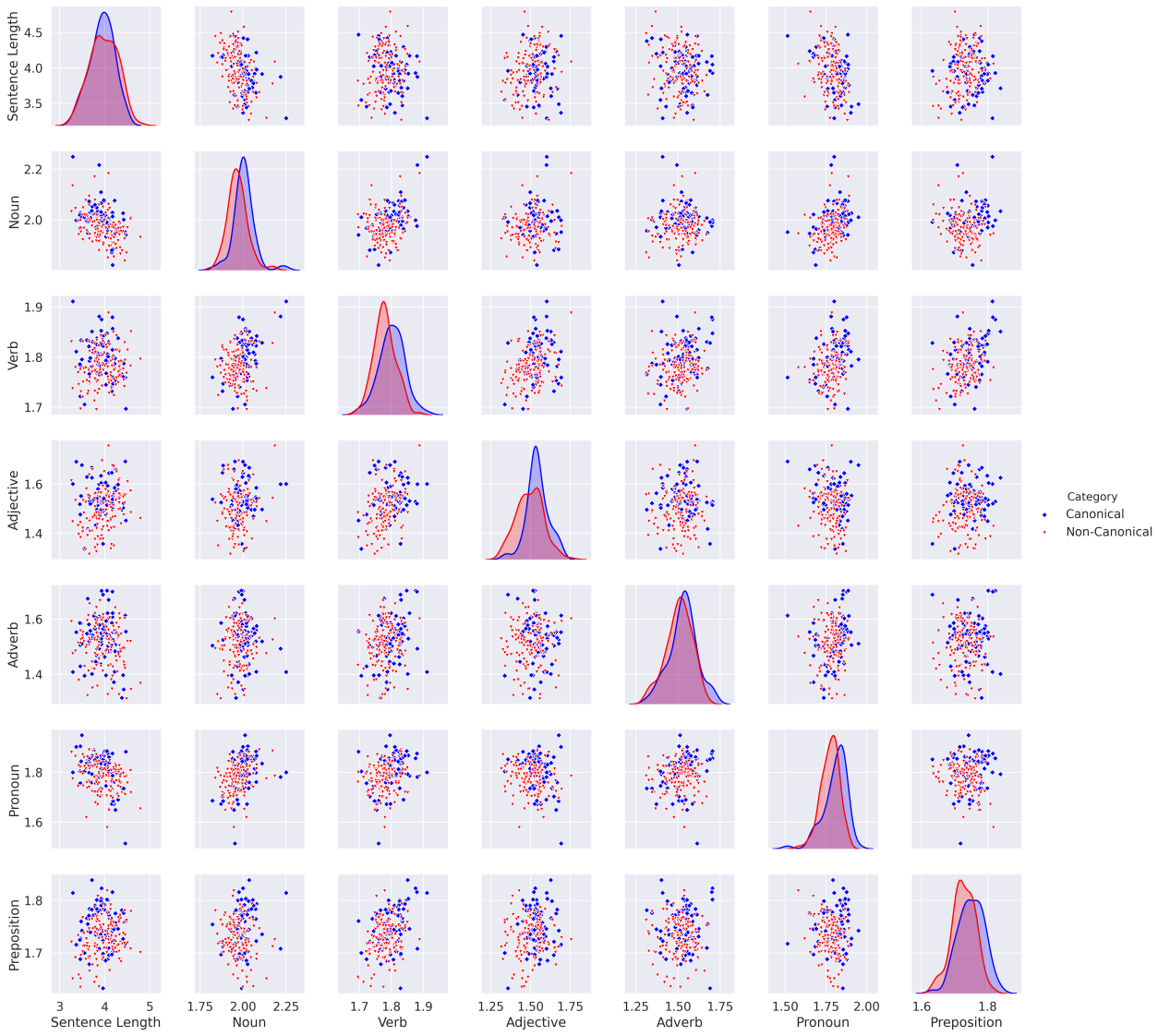

Figure S2. Pair-plot of all Shannon Entropy (ShEn) features in fictional/canonical and fictional/non-canonical texts. While each non-diagonal plot shows the relationships between two features, the main-diagonal subplots visualize the univariate distributions of each feature.

Figure S3. Pair-plot of all Approximate Entropy (ApEn) features in fictional/canonical, fictional/non-canonical and non-fictional texts. For better visibility of the data for canonical texts vs. non-canonical texts, see Figure S1. While each non-diagonal plot shows the relationships between two features, the main-diagonal subplots visualize the univariate distributions of each feature.

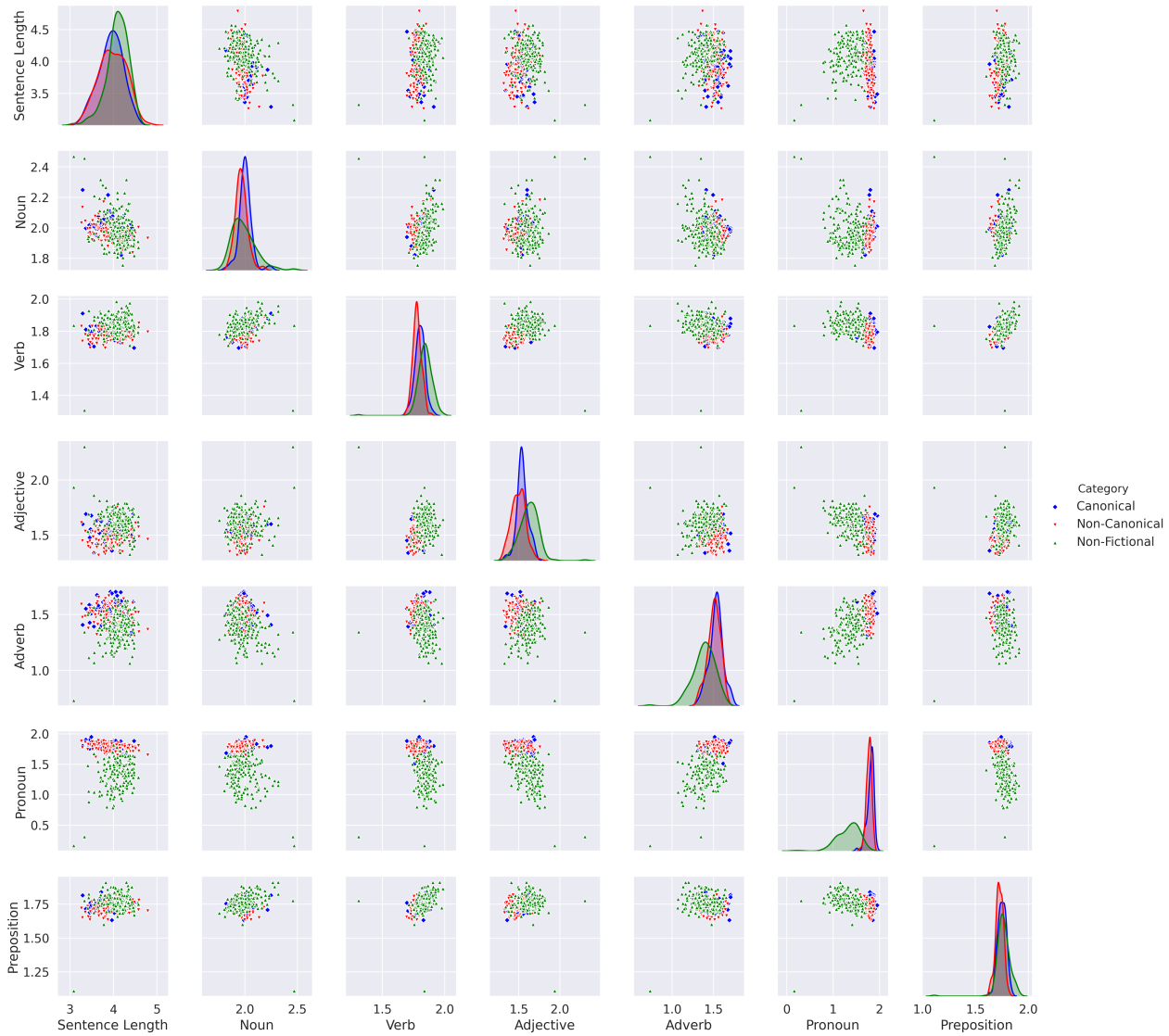

Figure S4. Pair-plot of all Shannon Entropy (ShEn) features in fictional/canonical, fictional/non-canonical and non-fictional texts. For better visibility of the data for canonical texts vs. non-canonical texts, see Figure S2. While each non-diagonal plot shows the relationships between two features, the main-diagonal subplots visualize the univariate distributions of each feature.

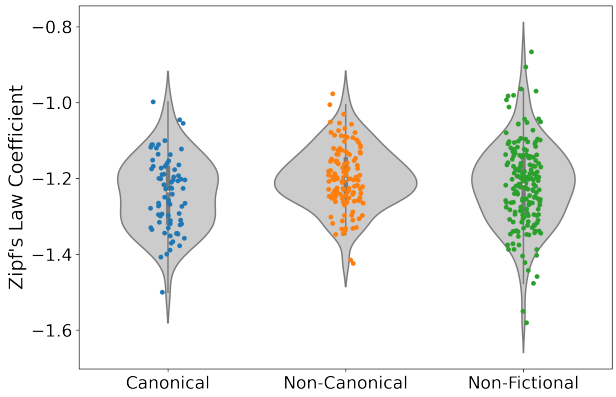

Figure S5. Zipf's law coefficient (lambda) of fictional/canonical, fictional/non-canonical and non-fictional texts. The high overlap of values between the text categories results in a poor classification accuracy.

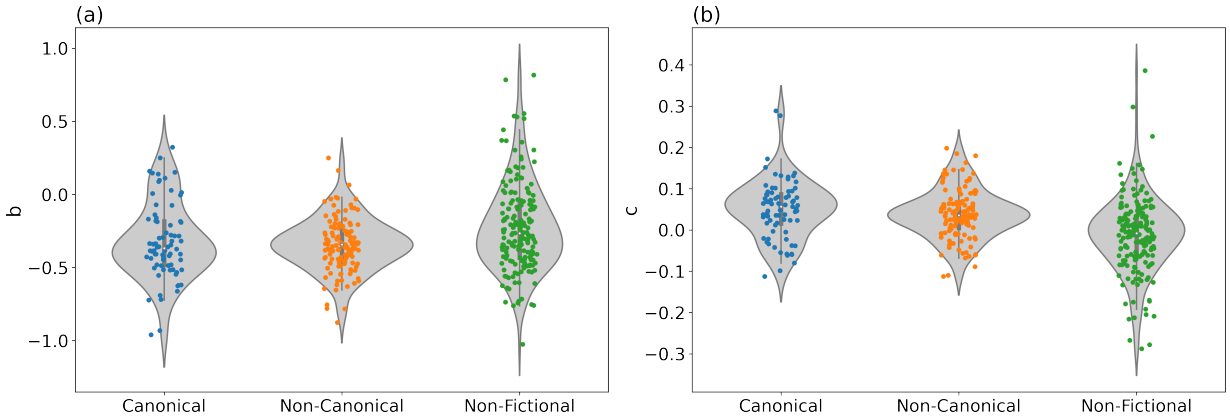

Figure S6. Menzerath–Altmann law assumes a relation between the size of constituents, y, of a linguistic construct with the size of the construct, x: $y = ax^b e^{-cx}$. The plots in (a) and (b) represent the two parameters of the Manzerath-Altmann law, b and c, respectively, in fictional/canonical, fictional/non-canonical and non-fictional texts.

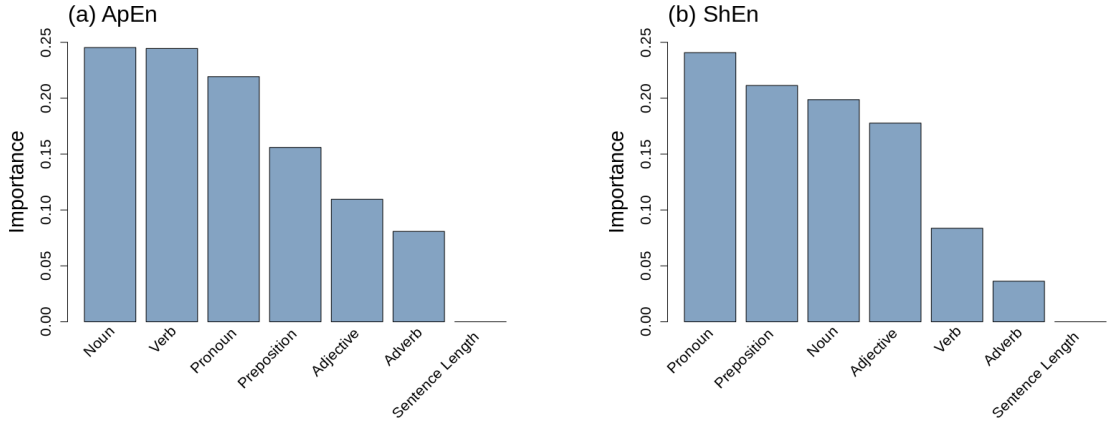

Figure S7. Sensitivity analysis of ApEn features (a) and ShEn features (b) in classification of fictional/canonical and fictional/non-canonical texts.

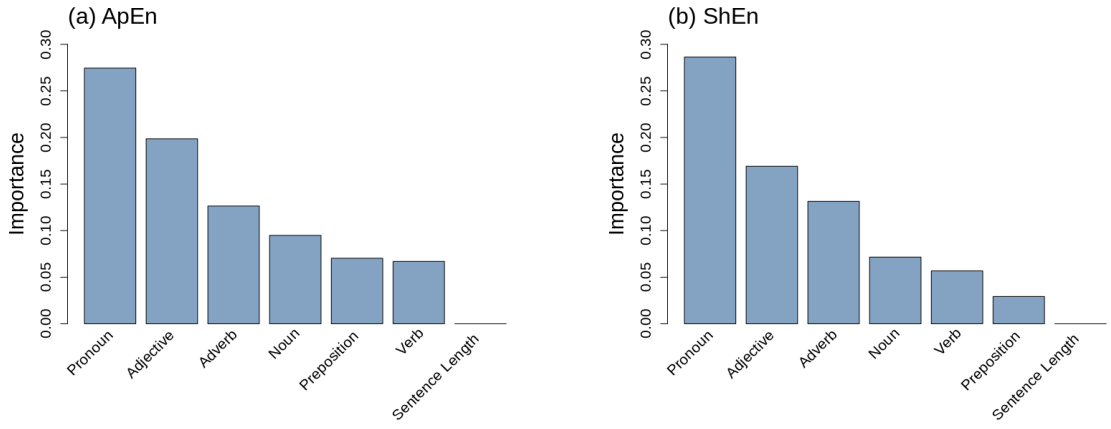

Figure S8. Sensitivity analysis of ApEn features (a) and ShEn features (b) in classification of fictional and non-fictional texts.

Chapter 5

PREDICTABILITY AND SURPRISE ACROSS TIME PERIODS

In Chapters 3 and 4, we operationalized preference in terms of canonization and investigated global structural patters of various texts categories using variability, fractality and predictability analyses. The purpose of the current chapter is to answer the question whether textual correlates of preference are time-invariant or change with time. More specifically, we ask what the structural characteristics of preferred and less-preferred texts are in the contemporary and in the earlier time periods.

Contemporary categories of prose were selected from published texts from 2000 to 2020. For an earlier period we used texts in the JEFP Corpus, version 2.0 (Chapter 4 / Mohseni et al., 2022), which were written in the 19th and early 20th centuries. Accordingly, we dealt with two different types of preference. In the earlier period, preference was operationalized based on canonicity. Canonical texts were categorized as preferred texts, while non-canonical texts were regarded as less-preferred prose. For the contemporary period, preference was operationalized in terms of commercial success using sale figures. We compiled a corpus of contemporary texts called the "Jena Corpus of Contemporary Expository and Fictional Prose (JCEFP Corpus)"; it includes texts published from 2000 to 2020. Similar to the JEFP corpus, three text categories were incorporated in the JCEFP corpus: fictional/preferred, fictional/non-preferred and non-fictional. Preferred contemporary fictional texts were selected from the bestseller lists of the New York Times Book Review and non-preferred texts were collected from free published books on the website www.smashwords.com that never gained a sale success. The non-fictional category was compiled by randomly selecting texts from diverse genres.

The comparative nature of the present study demanded the same methodological approach for analyzing texts from both time periods. Promising results of (un)predictability analysis of texts (cf. Chapter 3) encouraged us to use Shannon Entropy (ShEn, Section 2.5.1) and Approximate Entropy (ApEn; Section 2.5.2; Pincus, 1991) as our analysis methods. The two entropy metrics were applied to series derived from the lengths of sentences and the frequency distributions of six POS-tags (Noun, Verb, Adjective, Adverb, Pronoun and Preposition) in windows of

25 tokens.

Our statistical and classification results showed that, in general, preferred texts in both time periods, i.e. contemporary bestsellers and canonical earlier texts, exhibit higher degrees of unpredictability. However, the distinction between preferred and non-preferred earlier texts (canonical and non-canonical) is more discernible than that of preferred and non-preferred contemporary texts (bestseller and non-bestseller). Moreover, the difference between preferred and non-preferred contemporary texts is more a matter of global distribution (ShEn) than the degree of surprise in local structures (ApEn). In only two major classes of POS-tag, Noun and Verb, ApEn can classify the two contemporary fictional texts with a higher accuracy than ShEn. Although different factors are involved in commercial success –as in the case of contemporary bestsellers– and canonization –as in the case of canonical earlier texts– our results show that surprise and (un)predictability analysis is a promising method in text aesthetics studies as it effectively model textual preference and distinguishes different categories of texts.

References

Mohseni, Mahdi, Christoph Redies, and Volker Gast (2022). "Approximate Entropy in Canonical and Non-Canonical Fiction". In: *Entropy* 24.2, p. 277. DOI: 10.3390/e24020278.

Pincus, Steven M (1991). "Approximate Entropy as a Measure of System Complexity". In: *Proceedings of the National Academy of Sciences* 88.6, pp. 2297–2301. DOI: 10.1073/pnas.88.6.229.

entropy

MDPI

Article

Comparative Analysis of Preference in Contemporary and Earlier Texts Using Entropy Measures

Mahdi Mohseni [1,2,*], Christoph Redies [2] and Volker Gast [1]

[1] Department of English and American Studies, University of Jena, 07743 Jena, Germany; volker.gast@uni-jena.de
[2] Experimental Aesthetics Group, Institute of Anatomy I, Jena University Hospital, University of Jena, 07740 Jena, Germany; christoph.redies@med.uni-jena.de
* Correspondence: mahdi.mohseni@uni-jena.de; Tel.: +49-3641-9396-123

Abstract: Research in computational textual aesthetics has shown that there are textual correlates of preference in prose texts. The present study investigates whether textual correlates of preference vary across different time periods (contemporary texts versus texts from the 19th and early 20th centuries). Preference is operationalized in different ways for the two periods, in terms of canonization for the earlier texts, and through sales figures for the contemporary texts. As potential textual correlates of preference, we measure degrees of (un)predictability in the distributions of two types of low-level observables, parts of speech and sentence length. Specifically, we calculate two entropy measures, Shannon Entropy as a global measure of unpredictability, and Approximate Entropy as a local measure of surprise (unpredictability in a specific context). Preferred texts from both periods (contemporary bestsellers and canonical earlier texts) are characterized by higher degrees of unpredictability. However, unlike canonicity in the earlier texts, sales figures in contemporary texts are reflected in global (text-level) distributions only (as measured with Shannon Entropy), while surprise in local distributions (as measured with Approximate Entropy) does not have an additional discriminating effect. Our findings thus suggest that there are both time-invariant correlates of preference, and period-specific correlates.

Keywords: Approximate Entropy; Shannon Entropy; fictional texts; non-fictional texts; canonical texts; non-canonical texts; contemporary texts; bestseller books; POS tags; text classification

Citation: Mohseni, M.; Redies, C.; Gast, V. Comparative Analysis of Preference in Contemporary and Earlier Texts Using Entropy Measures. *Entropy* **2023**, *25*, 486. https://doi.org/10.3390/e25030486

Academic Editor: Boris Ryabko

Received: 6 February 2023
Revised: 4 March 2023
Accepted: 7 March 2023
Published: 10 March 2023

1. Introduction

What makes a text "successful", in the sense that it sells well, reaches a broad readership and/or acquires prestige among educated readers and critics? Is it promotion, network effects, economic or social circumstances—or perhaps the "quality" of the text itself? These questions have recently been addressed in a variety of studies in the field of computational aesthetics, aiming to identify observable correlates of preference in the structure of a text [1–7]. In empirical aesthetics the term "preference" is used to capture aesthetic attitudes towards cultural artefacts [8]. Such attitudes can be held both at an individual level—specific readers enjoy specific (types of) books—and at a community level—specific types of texts, and their authors, may obtain recognition and acquire prestige [9–11].

On the assumption that aesthetic experience can have a foundation in the cultural artifact itself, a natural question to ask is whether, or to what extent, correlations between properties of a work of art, such as a literary text, and the aesthetic response in readers, are invariant across time, space, and cultural environments, or whether they are dependent on such variables. In the present study we address this question by studying correlations between structural properties of texts and degrees of (community-level) preference across two time periods. Specifically, the central question is to what extent the textual determinants of preference in the 19th and early 20th centuries were the same as, or different from, the

textual determinants of preference today. As we operationalize preference differently for the two time periods (canonization and sales figures), the notion of 'preference' itself, in the context of prose texts, is under scrutiny as well.

The study of correlations between measurable properties of cultural artefacts on the one hand, and preference on the other, is obviously non-trivial. There are two major challenges: first, preference for texts is not immediately measurable, as it is for, say, visual stimuli, where preference for large numbers of images can be recorded directly and in real time [12]. We thus need to operationalize that concept in a reasonable (valid) way. Second, we do not know at present what types of structural properties will show correlations with our operationalizations of preference. The exploratory nature of this study (as well as other studies carried out in this spirit) should thus be obvious. Correspondingly, even minor correlations (small effect sizes) are of interest to us, as long as they are statistically significant.

In the present study we operationalize preference in terms of reception or, put differently, the scope of the readership. We deal with texts from two time periods: with the contemporary period, spanning the time between 2000 and 2020, and from an earlier period, covering the time between 1813 and 1922. For each period we (necessarily) use different operationalizations of preference: The earlier texts are divided into canonical and non-canonical texts, using the canon of Western literature [13,14] as a criterion of classification (see also [7,15]). Canonical texts form part of the cultural backbone and historical memory of a society [16]. They are often time-honoured and are included in syllabuses at schools and universities. Given their prestige and institutional support, they reach a broad readership distributed over a large time span.

The reception of contemporary texts cannot be measured via canonization, which is a process that takes some time and involves several stakeholders, such as publishing houses, academics and government departments. These texts are therefore classified according to their commercial success, and thus divided into bestsellers and non-bestsellers (see also [17–19]). Like canonical texts, bestsellers have reached a broad readership. This readership is not distributed over time, however, but constituted by a single 'cohort' at the time of publication. Obviously, the two operationalizations of preference (canonization and sales figures) are not identical. What they share is that they measure reception; they (may) differ in the type of readership. In interpreting our results, that difference of course needs to be taken into consideration.

As for the structural properties of texts that are potential correlates of preference, there are two central challenges. The first question is what type of observable properties we measure. We assume that the aesthetic experience in reading is a function of both what a text is about—for instance, the plot and the characters in a novel—and how it is written. Figures may be characterized in a specific way (explicit vs. implicit characterization) [20], and the state of affairs can be described by the narrative voice, through dialogues or interior monologue, etc. [21]. While such elements of style are hard to measure directly, they have structural reflexes in texts, for instance, insofar as they imply the use of different discourse modes [22] which, in turn, come with different distributions of parts of speech [23,24]. For example, the mode of narration typically implies the use of past tense verbs, dialogue comes with a high proportion of pronouns and verbs, description requires adjectives, etc. Given the association between register, style, and discourse modes and the distribution of parts of speech [23–26], the latter category, which is observable and measurable, figures prominently in our work.

The second major challenge of our work concerns the type of statistic that may be informative with respect to the degree of preference. Previous studies, inspired by research in the domain of vision (for a review, see [12]), have focused on global statistical properties such as the variability (of observables) in a text [15,27,28], long-range correlations [6,15,29] and various indicators of predictability or surprise [4,7]. In the present study we use two measures of surprise—Shannon Entropy and Approximate Entropy—as the aesthetic experience has been shown in previous work to be driven by the interplay between expec-

tation and surprise [7,30,31] (moreover, see [32,33] for a discussion on "unification" and "diversification" in text).

Specifically, the present study investigates the differences and similarities in the relationship between the degree of surprise in the textual structure and (community-level) preference, in two time periods, the 19th and early 20th centuries, and the contemporary period. Preference is operationalized as canonization for the earlier texts, and in terms of sales figures for the contemporary texts. Similarities can be expected on the assumption that certain determinants of preference are time-invariant (universal), and do not vary significantly with the readership. Differences can be expected because writing styles are known to vary from one period to the next [25], and because literature is embedded into socio-cultural contexts, with changing aesthetic preferences in all domains of culture (music, painting, architecture, etc.). Moreover, the two operationalizations of preference can be expected to have different types of reflexes.

As reviewed in more detail in [7], preference has been operationalized in terms of the scope of the readership in previous studies under different terms, such as "success" [27], "popularity" [34], being "professional" vs. "amateur" [1], or "information-based energy" [35]. Data from websites and social networks have been used in some previous studies to model readers' preference, for instance, the download counts from the website of the Gutenberg Project [2], or ratings of readers on the website Goodreads [3,5].

Some previous studies have referred to the Nobel Prize as a gauge for high quality or success. For example, Febres and Jaffe [4] analysed the categories of Nobel laureates and non-Nobel laureates in two languages, English and Spanish, using global properties, such as entropy, lexical diversity and word frequency distribution in texts. Their results showed that statistical measures can be predictive of the category of texts, with a higher performance for Spanish compared to English texts. Bizzoni et al. [6] classified Nobel prize winners from other texts using the fractality of sentiment arcs. They showed that the distribution of self-similarity measures in the two text categories under analysis differed, and that the degree of fractality of higher-quality texts is likely to be located in a specific range of values.

Mohseni et al. [15] approached the discrimination of canonical from non-canonical texts using textual properties of texts represented in the form of a series. They used sentence length, the frequencies of POS tags per sentence, the lexical diversity metric MTLD, and topic probabilities to numerically represent the structure of a text, and determined the variance and long-range correlations of the series corresponding to the texts. Training a classifier with the calculated values, they were able to distinguish fictional from non-fictional and, within the fictional category, canonical vs. non-canonical English texts with acceptable accuracy.

Success has also been defined based on sales figures. Yucesoy et al. [17] and Wang et al. [18] analysed texts in the New York Times Bestseller lists and Vasyliuk et al. [19] investigated bestseller books on Amazon. However, they did not analyse the texts of books, but rather restricted their analyses to more straightforward statistical information and metadata, such as the time of publication, number of reviews, genre, and price, and related the success of the texts to non-textual factors.

In the present study we adopt the approach to textual aesthetics proposed by Mohseni et al. [7,15]. We assume that a pleasant reading experience emerges from an interplay of predictability and surprise. Previous work has shown that canonical literature differs from non-canonical literature in its degree of predictability. Mohseni et al. [7] analysed two types of series derived from texts, sequences of sentence lengths and of frequencies of part-of-speech (POS) tags in fixed-size windows of text (see Section 2.3). Two entropic metrics were computed, Shannon Entropy (ShEn) and Approximate Entropy (ApEn), for the distributions of relevant text properties. ShEn measures (ir)regularity as a global structural property. ApEn determines (un)predictability as a sequential characteristic of underlying text property series (see Section 2.4). This method was also applied in the present study, with a different dataset. Note that the present study primarily focuses on

the classification of texts on the basis of preference levels. The temporal dimension comes into play insofar as we compare texts of preference levels from two periods of time. We do not perform temporal classification, in the sense that the time of writing is the category of classification. Approaches to temporal classification are nevertheless summarized in the Supplementary Materials, Section S1.

The paper is organized into three sections. Section 2 contains a description of the data and methods. Section 3 presents the results, which are discussed in Section 4.

2. Data and Methods

2.1. The Jena Corpus of Expository and Fictional Prose

The present study is based on the Jena Corpus of Expository and Fictional Prose (JEFP), version 2.0. The corpus was compiled for a comparison of different types of fictional and non-fictional texts from the 19th and early 20th centuries, here called "earlier" texts, and it has been used for the study of questions relating to empirical aesthetics [7,15]. The JEFP Corpus comprises three sub-corpora: canonical/fictional, non-canonical/fictional and non-fictional (Table 1). The canonical sub-corpus consists of 76 texts that form part of the Western literature canon Bloom [13]. It represents a collection of fictional texts that are widely known among the educated population, often taught in school and discussed in academic discourse. The category of non-canonical fictional texts comprises 130 texts that were obtained from the Project Gutenberg website. It represents the non-preferred earlier texts. Finally, the sub-corpus of non-fictional texts contains 185 texts from different genres such as architecture, astronomy, geology, geography, philosophy, psychology and sociology. These texts were also obtained from the Project Gutenberg website.

Table 1. Text categories in the Jena Corpus of Expository and Fictional Prose (JEFP), version 2.0. The table shows, for each text category, the number of texts and the mean text length, measured in tokens, ±standard deviation. Data are from the study by Mohseni et al. [7].

Category	Number of Texts	Length ($\times 10^3$)
Canonical (preferred)	76	199 ± 96
Non-Canonical (non-preferred)	130	111 ± 56
Non-Fictional	185	171 ± 178

2.2. The Jena Corpus of Contemporary Expository and Fictional Prose

For our comparative study, we also needed a corpus of contemporary texts to compare them with the earlier texts in the JEFP corpus. Thus, we compiled a corpus which contained categories analogous to those of the JEFP corpus (preferred/fictional, non-preferred/fictional and non-fictional). We called this corpus the "Jena Corpus of Contemporary Expository and Fictional Prose" (JCEFP).

To compile the list of preferred contemporary texts, we used the New York Times Bestseller list, which is published weekly in the New York Times Book Review. Some books manage to appear on the list for several weeks, and some lose their rank after only one week in competition with other books. We selected ninety-three texts from lists of the New York Times Fiction Best Sellers published from 2000 to 2020. Our selection was based on lists taken from Wikipedia for each year.

To build the category of non-preferred contemporary texts, we used the website www.smashwords.com (accessed on 11 March 2021), which allowed us to search for texts based on various criteria, such as genre, length and price. In this part of the corpus, we only included freely available fictional texts, assuming that texts promising commercial success will not be distributed for free by a publisher. This part of the corpus consequently contains no bestsellers, as bestsellers would have to be bought. For a book to be free does not of course mean that the book is not read by anyone. In fact, free distribution could be an incentive for people interested in popular literature to read the texts. Moreover, if an author manages to publish a successful text later, their previous, less-successful texts

may find more readers (as in the case of B. Obama's first book *Dreams from my Father*, for instance). Still, at the time of publication the texts are clearly non-bestsellers, and books that are not promoted by publishers. The non-preferred sub-corpus thus compiled by us contained 110 texts.

Non-fictional texts were randomly selected from different genres, e.g., philosophy, psychology, sociology and natural science, similar to the genres that we included for texts in the JEFP corpus. The contemporary version of the non-fictional sub-corpus contained 122 texts. Table 2 presents the summary statistics for the JCEFP corpus. As we selected best-selling books from lists from 2000 to 2020, the category of non-preferred non-fictional texts was also restricted to texts that were published after 2000. Table S1 in the Supplementary Materials lists all texts with the metadata.

Table 2. Text categories in the Jena Corpus of Contemporary Expository and Fictional Prose (JCEFP). The table shows, for each text category, the number of texts and the mean text length, measured in tokens, ±standard deviation.

Category	Number of Texts	Length ($\times 10^3$)
Bestseller	93	153 ± 90
Non-Bestseller	110	105 ± 39
Non-Fictional	122	142 ± 84

All texts in both the JEFP corpus and the JCEFP corpus were pre-processed in the same way. We removed the tables of contents and indices and cleaned up the texts partly manually and partly automatically using regular expressions to fix broken lines and hyphenated words.

To segment texts into sentences and to assign POS tags to tokens, we used the `Stanza` package for Python [36], a neural-based text processing toolbox with high accuracy. We used the toolbox with the default pre-trained model for English (UD English EWT, version 1.0.0 [37]).

Note that previous studies have shown no underperformance of taggers for texts from the 19th century. This is probably due to the fact that orthography was already standardized at that time. For instance, Schneider et al. [38] showed that if a POS tagger was trained on contemporary texts and applied to historical texts written after 1800, the performance would not drop. They also analysed the tagging errors and showed that most POS tagging mistakes were found in lower-level categories within the major classes; for example, between NN (noun, singular or mass) and NNP (proper noun, singular), and between VB (verb, base form) and VBP (verb, non-third person singular present). Such errors would not affect our results because we analysed the distribution of major word categories (see Section 2.3).

2.3. Properties Underlying Textual Structure

To analyse the structural organization of texts, we took the same approach as Mohseni et al. [7]. We represented and analysed texts by seven text properties: sentence length and the frequencies of six major parts of speech in fixed-size windows: Noun, Verb, Adjective, Adverb, Pronoun and Preposition. Sentence length was measured as the number of tokens in a sentence, including all words and punctuation marks. Each major part-of-speech (POS) included all relevant sub-categories. For example, plural, singular, common and proper nouns all were counted as Noun. All forms of verbs, base form, past tense, past participle and gerund, were treated similarly as Verb. Adjective and Adverb included simple, comparative and superlative types. Pronoun covered personal and possessive pronouns.

To build series of part-of-speech (POS) tags, we counted the number of each POS tag in subsequent windows of 25 tokens of text. As mentioned in Mohseni et al. [7], the window size does not have a significant effect on the results as long as it is within reasonable limits.

By windowing, we split each text into a sequence of fixed-length segments. Fixed-length segmentation eliminates undesirable effects of correlation between sentence length and frequencies of POS tags. Each window of text is called a "box". Each box is like a small bag of words, in which the internal structure of the texts is ignored and only the frequency of POS tags is determined. We therefore call this approach a 'sequence of boxes' approach. If the order of the boxes in the sequence was taken into account, we analysed the underlying structural design of a text (as in the case of Approximate Entropy; Section 2.4). If we ignored the linear order of the boxes, we analysed the global distribution of POS tags in a text (as in the case of Shannon sntropy; Section 2.4).

2.4. Approximate Entropy and Shannon Entropy

To measure the degrees of (ir)regularity and (up)predictability in a series of text properties (Section 2.3) we used two entropy measures: Shannon Entropy (ShEn) and Approximate Entropy (ApEn) [39]. ShEn is a measure of global distribution and is computed as

$$-\sum_{x \in S_x} p(x) \log p(x)$$

where S_x is the set of all possible events x. ShEn assumes that events happen independent of each other. This metric measures the degree of uncertainty. If the probability of all events is equal, the system has the highest uncertainty, and as a result, ShEn takes its maximum value.

Conversely, ApEn is a measure of sequential organization (cf. Supplementary Materials, Section S2). It was proposed to measure the degree of (ir)regularity in a series according to the distance (dissimilarity) of sub-sequences to each other. As variation is an intrinsic characteristics of a series, in ApEn some level of fluctuation is "tolerated". If the difference between two sub-sequences lies within the "tolerance" level, it is assumed that "similarity" is not violated. In the computation of ApEn, the sub-sequence matches of length m are compared with sub-sequence matches of length $m + 1$. In a sequence with a high level of fluctuation, longer sub-sequences are less-likely to be similar to each other, which in turn leads to a higher ApEn value. In exploratory studies, the parameters of ApEn, i.e., m and r, are usually set to 2 and 20% of the standard deviation, respectively, (see, for example, [7,40–42]).

In our experiments we used both ShEn and ApEn. ShEn measures surprise based on global distributions. AppEn measures surprise based on (ir)regularities in the series. Note that a high degree of AppEn implies a high degree of ShEn but not vice versa. We first calculated the degree of irregularity (or unpredictability) in a series of text properties. On this basis we determined to what extent any observed difference originate from the global distribution of the features (ShEn), or from their sequential organization (ApEn). The code that we used to calculate features is accessible at https://github.com/mohsenim/Surprise (accessed on 5 February 2023).

3. Results

Our analyses implied a two-dimensional comparison. We carried out (i) a comparison of preferred and non-preferred fictional texts, for each period, and (ii) a comparison of the differences for each period. We used our two corpora, JEFP and JCEFP, which, as explained in Section 2.1, contained preferred texts (canonical texts in JEFP; bestselling contemporary texts in JCEFP), and non-preferred texts (non-canonical texts in JEFP; non-bestselling contemporary texts in JCEFP). In the following subsections, we start by presenting the results of the statistical analyses (Section 3.1) before turning to the results from classification (Section 3.2).

3.1. Statistical Analysis of Features

For the category of earlier texts we used the data published in Mohseni et al. [7], where the texts of the JEFP corpus were analysed. For contemporary texts we created a

series of seven observables for each text in the JCEFP corpus, following the procedure of Mohseni et al. [7]. We determined sentence lengths and the number of specific POS tags in windows of 25 tokens for six POS tags (see Section 2.3). For each series we computed ApEn and ShEn values (Section 2.4). We then compared the text categories using their median values because a Kolmogorov–Smirnov test indicated that some features were not normally distributed. For our statistical comparison we used the non-parametric Mann–Whitney U test.

Tables 3 and 4 (left-hand side) compare the contemporary bestselling and non-bestselling texts in terms of ApEn and ShEn, respectively. The values of the features for earlier canonical and non-canonical texts are shown on the right-hand side. These data have been taken from Mohseni et al. [7]. To facilitate the comparison of values for each text category/feature combination, the (significantly) higher value of each pair is shown in boldface. For Noun, Verb, Adjective and Preposition, the category of bestseller has higher values than the non-bestselling texts in the contemporary corpus. In both categories, the values for sentence length are not significantly different from each other. Only in one major POS category, i.e., Pronoun, are the values for ApEn and ShEn higher for contemporary non-bestselling texts than for the bestsellers.

Table 3. Median values of Approximate Entropy (ApEn) for all text properties and for all fictional text categories. ApEn values for contemporary bestselling ($N = 94$) vs. non-bestselling ($N = 110$) texts, and for canonical ($N = 76$) vs. non-canonical ($N = 130$) texts. The asterisks indicate whether the differences between the two text categories in the earlier or contemporary periods are statistically significant (Mann–Whitney U test; ns, not significant; *, $p \leq 0.05$; **, $p \leq 0.01$; and ***, $p \leq 0.001$). Values that are significantly higher within a pair of columns are shown in boldface. The 95% confidence intervals for the median (according to [43]) are shown in parentheses. The data for earlier texts are from the study by Mohseni et al. [7].

	Contemporary		**Earlier**	
Text Property	**Bestseller**	**Non-Bestseller**	**Canonical**	**Non-Canonical**
Sentence Length	1.99 (1.95, 2.02)	2.01 (1.99, 2.04) ^{ns}	1.86 (1.83, 1.89)	1.87 (1.86, 1.90) ^{ns}
Noun	**1.93 (1.921, 1.934)**	1.85 (1.84, 1.86) ***	**1.89 (1.88, 1.91)**	1.83 (1.81, 1.84) ***
Verb	**1.74 (1.730, 1.742)**	1.70 (1.68, 1.71) ***	**1.75 (1.73, 1.76)**	1.70 (1.69, 1.71) ***
Adjective	**1.40 (1.38, 1.41)**	1.36 (1.34, 1.38) **	**1.50 (1.49, 1.52)**	1.45 (1.43, 1.48) ***
Adverb	1.50 (1.47, 1.53)	1.51 (1.50, 1.52) ^{ns}	**1.51 (1.49, 1.53)**	1.48 (1.46, 1.49) **
Pronoun	1.71 (1.69, 1.73)	**1.73 (1.71, 1.74)** *	**1.74 (1.71, 1.76)**	1.681 (1.675, 1.691) ***
Preposition	**1.63 (1.62, 1.64)**	1.61 (1.60, 1.62) ***	**1.71 (1.70, 1.72)**	1.67 (1.66, 1.68) ***

If we compare earlier and contemporary texts in the fictional categories, we observe both differences and similarities. In earlier texts the values for all POS tags are higher for canonical texts than for non-canonical texts. Contemporary texts do not show any difference for the category of Adverb. For Pronoun, the value is higher for the non-bestselling texts. In summary, we observe a similar pattern for prepositions and the three POS tags representing major classes of content words, i.e., Noun, Verb and Adjective. Thus, the biggest difference in the comparison of preferred vs. non-preferred texts in the earlier and contemporary periods lies in the distribution of pronouns. Notably, ApEn and ShEn exhibit similar patterns of differences for all comparisons.

Examples of texts with a high degree of unpredictability in the JEFP corpus are *Ulysses* by James Joyce, *The Golden Bowl* by Henry James and *Sartor Resartus* by Thomas Carlyle, showing the highest ApEn values in the category of earlier canonical texts for Noun, Verb and Adjective, respectively. In the bestsellers category among the contemporary texts, *Port Mortuary* by Patricia Cornwell has the highest ApEn value for Noun and the highest ShEn

value for Verb. Another prominent example is *Freedom* by Jonathan Franzen, which is the bestseller with the highest ApEn value for Adjective in the corpus.

Table 4. Median values of Shannon Entropy (ShEn) for all text properties and for all fictional text categories. ShEn values for contemporary bestselling (N = 94) vs. non-bestselling (N = 110) texts, and for canonical (N = 76) vs. non-canonical (N = 130) texts. The asterisks indicate whether the differences between the two text categories in the earlier or contemporary periods are statistically significant (Mann–Whitney U test; ns, not significant; *, $p \leq 0.05$; **, $p \leq 0.01$; and ***, $p \leq 0.001$). Values that are significantly higher within a pair of columns are shown in boldface. The 95% confidence intervals for the median (according to [43]) are shown in parentheses. The data for earlier texts are from the study by Mohseni et al. [7].

	Contemporary		Earlier	
Text Property	**Bestseller**	**Non-Bestseller**	**Canonical**	**Non-Canonical**
Sentence Length	3.42 (3.39, 3.46)	3.36 (3.31, 3.39) ^{ns}	3.96 (3.88, 4.05)	3.96 (3.87, 4.08) ^{ns}
Noun	**2.09 (2.08, 2.11)**	1.99 (1.77, 2.02) ***	**2.00 (1.99, 2.02)**	1.97 (1.95, 1.98) ***
Verb	**1.80 (1.78, 1.81)**	1.77 (1.767, 1.789) ***	**1.80 (1.79, 1.81)**	1.777 (1.772, 1.783) ***
Adjective	**1.43 (1.41, 1.45)**	1.39 (1.37, 1.42) **	**1.54 (1.53, 1.55)**	1.49 (1.47, 1.53) ***
Adverb	1.53 (1.49, 1.56)	1.54 (1.53, 1.57) ^{ns}	**1.54 (1.51, 1.55)**	1.51 (1.49, 1.53) *
Pronoun	1.80 (1.79, 1.81)	**1.82 (1.81, 1.84)** ***	**1.83 (1.80, 1.84)**	1.78 (1.77, 1.79) ***
Preposition	**1.67 (1.66, 1.68)**	1.66 (1.64, 1.67) *	**1.75 (1.74, 1.77)**	1.73 (1.72, 1.74) ***

Both corpora (JEFP and JCEFP) contained non-fictional texts as well. In the Supplementary Materials, Tables S2 and S3 show the results for fictional and non-fictional texts for ApEn and ShEn, respectively. We refer the interested reader to these two supplementary tables, to gain an impression of the comparison between fictional and non-fictional texts. Summarizing the results, there is no uniform pattern in the degree of (un)predictability in fictional or non-fictional texts. For some text properties, such as Verb and Adjective, the values of ApEn and ShEn are higher for fictional than non-fictional texts, while for other text properties, such as Adverb and Pronoun, the opposite pattern can be observed. Moreover, the values of ApEn and ShEn do not correspond to each other in measuring the degree of (un)predictability in the fictional or non-fictional text categories.

Figure S3 in the Supplementary Materials shows a correlation plot for ApEn and ShEn values for all earlier and contemporary text categories, for all text properties. For some text properties, such as Adjective and Adverb, the correlation coefficients are very high, while for others, such as Noun and Verb, they are lower. This finding is related to the difference between the discrimination power of ApEn and ShEn, which becomes visible when we look at the classification results in the next section.

3.2. Classification

We extend our analysis of preferred vs. non-preferred texts with a classification tasks. Classification determines the performance of each property/feature in distinguishing the text categories under analysis. For each setting we trained a support vector machine (SVM) with a radial basis function (RBF) kernel. To report the performance of the classification models, we used balanced accuracy, which eliminates the undesired effect of different class sizes in the input data. In the comparison of the classification results we rely on the 5×2CV paired t-test [44] with a significance level of $\alpha = 0.05$.

Table 5 shows the balanced accuracy scores for bestselling vs. non-bestselling contemporary texts, for each text property/feature combination. To compare contemporary and earlier texts, we also include the classification results of canonical vs. non-canonical earlier texts, which were published in Mohseni et al. [7] (right-hand side of Table 5).

Table 5. Balanced accuracy of classification (%) for the single features for the bestselling/non-bestselling contemporary texts distinction and for the canonical/non-canonical early texts distinction. Values that are significantly higher within a pair of columns are shown in boldface. Wherever the results are not significantly better than random accuracy (50%), we mark the result with a dagger †. The data for earlier texts are from the study by Mohseni et al. [7].

	Bestselling vs. Non-Bestselling		**Canonical vs. Non-Canonical**	
	ApEn	**ShEn**	**ApEn**	**ShEn**
Sentence Length	53.6 ± 3.1	53.8 ± 3.0	$\mathbf{54.0 \pm 1.6}$	50.0 ± 1.0 †
Noun	$\mathbf{80.4 \pm 3.4}$	72.9 ± 2.7	$\mathbf{73.6 \pm 2.9}$	60.0 ± 4.5
Verb	$\mathbf{67.7 \pm 3.7}$	62.7 ± 2.5	$\mathbf{71.3 \pm 3.4}$	56.2 ± 3.8
Adjective	56.2 ± 3.2	57.4 ± 3.3	$\mathbf{55.2 \pm 2.5}$	51.5 ± 2.7 †
Adverb	53.6 ± 2.2	51.3 ± 2.6 †	51.6 ± 1.4 †	51.0 ± 1.5 †
Pronoun	57.6 ± 1.8	58.1 ± 1.9	$\mathbf{68.0 \pm 1.7}$	63.8 ± 1.8
Preposition	57.8 ± 2.6	53.5 ± 2.2	$\mathbf{69.1 \pm 2.4}$	59.7 ± 1.7
All	79.4 ± 4.2	77.6 ± 2.4	$\mathbf{77.3 \pm 2.6}$	68.5 ± 2.3

In the task of classifying bestselling vs. non-bestselling contemporary texts, both ApEn and ShEn perform comparably well, except for Noun and Verb, where ApEn provides a significantly higher accuracy compared to ShEn. Comparing accuracy scores for the two time periods, we observe a shift in the performance of individual text properties, while ApEn of all text properties except Adverb distinguishes canonical from non-canonical earlier texts better than ShEn, the ApEn values of only two text properties in the contemporary texts, i.e., Noun and Verb, provide a better performance compared to ShEn. For other text properties, no significant difference was observed.

The last row of Table 5 shows the performance of classification using all features. No significant difference between the discriminative power of ApEn and ShEn for the bestselling/non-bestselling contemporary texts distinction can be observed. Moreover, the results show that classification using the ApEn values of all text properties cannot distinguish the text categories under study better than ApEn of Noun alone. The difference between the two values is not statistically significant. Using ShEn of all text properties surpasses the performance of all individual ShEn features.

Concerning the results based on all features for earlier texts, ApEn outperforms ShEn with a high margin in the classification of canonical versus non-canonical texts. Taking all text properties into account, the difference between the performance of ApEn and ShEn in the separation of preferred and non-preferred contemporary texts disappears. Nevertheless, the classification accuracy for both features remains comparably high (79.4 and 77.6%, respectively), which confirms that (un)predictability analysis is a promising approach for analysing texts of different aesthetic categories.

4. Discussion and Conclusions

Confirming the results obtained by Mohseni et al. [7] for texts from the 19th and early 20th centuries, our study shows that the degree of preference associated with a contemporary text also has correlates in global statistical properties of the text. Generally speaking, preferred texts (bestsellers) are characterized by lower degrees of predictability for most features, as reflected in higher values for the two entropy measures, Shannon Entropy and Approximate Entropy (Tables 3 and 4).

However, we also found differences between contemporary and earlier texts. The earlier texts were better distinguished by Approximate Entropy than by Shannon Entropy (Table 5) [7]. This shows that the two text categories not only differ in terms of the unpredictability of the part-of-speech rates across windows of text (Shannon Entropy); the part-of-speech rates are also less predictable along the sequential organization of a text (Approximate Entropy). After reading a window of 25 words, a reader has less informa-

tion about the part-of-speech distribution in the next window of 25 words, in preferred (canonical) texts compared to non-preferred (non-canonical) texts. This is different for the contemporary texts. Approximate Entropy does not globally provide better classification results than Shannon Entropy for this part of the corpus. Only two part-of-speech categories—Noun and Verb—exhibit higher classification accuracy values on the basis of Approximate Entropy than they do based on Shannon Entropy. When all parts of speech as well as sentence length are taken into consideration, there is no significant difference between the classification results (see Table 5). This shows that bestsellers generally exhibit a higher degree of irregularity in the distribution of the linguistic features used for this study than non-bestsellers. The degree of irregularity is not modulated locally, however, and does not depend on the sequential arrangement of structural features.

A second difference between the two time periods is that in the earlier works from the 19th and early 20th centuries, all part-of-speech tags were distributed more unpredictably in the canonical texts than in the non-canonical ones (Tables 3 and 4). For canonical texts, a low degree of predictability seems to be a general design principle. For contemporary texts, one part of speech, Pronoun, had higher entropy values for the non-bestselling texts compared to the bestselling texts. Moreover, there was no significant difference in the distribution of Adverbs. It seems that only the major classes of content words, Nouns, Verbs and Adjectives as well as Prepositions, whose occurrence correlates with that of nouns, are distributed more unpredictably in bestselling texts as opposed to non-bestselling contemporary texts.

There are at least four possible explanations for the observed differences. The first explanation is based on changes in writing styles. It is well known that narrative styles have changed considerably since the 17th century [25]. This concerns, among other things, the narrator's visibility and reliability, and the relationship between the narrator and the reader. Moreover, the inventory of registers used in novels has been broadened. For example, the technique of interior monologuing was introduced in modernism [45]. The high degree of unpredictability of POS tags in modern bestsellers, in comparison to non-bestsellers, points to a higher degree of heterogeneity of discourse modes in the former group of texts (Tables 3 and 4). However, then, the fact that Approximate Entropy does not separate the classes better than Shannon Entropy for all POS tags does seem to show that the sequential arrangement of discourse modes is no less predictable in bestsellers (Table 5). Simplifying this hypothesis, we speculate that bestselling authors draw on a more varied inventory of discourse modes than other authors, but the texts do not exhibit a higher degree of unpredictability as far as the sequential arrangement of these modes is concerned. This hypothesis would require closer inspection of the data, and additional methods that allow us to trace the trajectory of discourse modes across a text.

Related to this first explanation is a second one, which concerns the question of register and genre. Writing styles have not only changed 'locally' [25], but there are also shifts in the frequency of literary genres. Among the contemporary texts, specific genres seem to be particularly successful that are rare in the category of canonical texts (e.g., crime stories). As we have no reliable genre classification for our sample, we cannot test for the effect of genre directly. We did, however, conduct an experiment on another corpus, a large collection of fictional texts from several genres (see the Supplementary Materials, Section S3). The results show that distributions of Approximate Entropy and Shannon Entropy vary significantly between genres. However, there is no general pattern across textual properties: there is no genre that exhibits particularly high or low values for all part-of-speech frequencies and sentence length values, while the effect of genre and register as determinants of preference needs to be taken into account without doubt, the results of our preliminary study suggest that they may have a modulating, rather than a direct effect. Further studies are needed to test this assumption.

A third possible explanation for the observed differences between contemporary and earlier texts is provided by the factor of 'technology'. The process of writing has changed considerably between the earlier period—the 19th and early 20th centuries—and today,

while the earlier texts were written either by hand or with a typewriter, contemporary writers can use computers. Texts can easily be edited, and re-edited, and the process of writing requires less planning than it used to. As a consequence, the difference between preferred and non-preferred texts may have decreased, as far as sequential organization is concerned, as the skills of a writer (as the architect of a story) may be less visible in contemporary texts. The general distributions of discourse modes, however, would not be affected by the process of writing, as they seem to be primarily a function of the author's creativity.

Finally, it is of course conceivable that the two types of preference that we considered—canonization for the earlier texts, sales figures for the contemporary texts—are driven by different forces. The process of canonization is, to a large extent (though not exclusively), driven by academics. It is based on thorough analyses conducted by a community of researchers over an extended period of time. Bestselling books, in contrast, have not gone through this type of filter. For a text to succeed on the book market, it has to be advertised broadly and supported by the media, e.g., with reviews and public discussion. Even though literary critics play an important role in this process, they may have a comparatively small impact on the success of a book (sometimes, negative reviews increase the sales figures, as they lead to controversial public discussion).

From the perspective of empirical aesthetics, it seems conceivable that the design principles of canonical literature—variation both in global distribution and sequential organization—play a less important role in the commercial success of a (contemporary) work. While canonical literature typically targets 'educated readers', contemporary best-sellers have a broader target audience—in fact, they tend to target an audience as broad as possible. Aesthetic pleasure varies from reader to reader (see, for example, [46], and for poetry [47,48]). More experienced (or even professional) readers may take pleasure in reading less predictable texts than less experienced readers do.

Unfortunately, we cannot use the same type of operationalization for preference for contemporary and earlier texts, as sales figures (at the time of publication) are not available for the canonical texts (and today's sales figures are, again, influenced by canonization), and because contemporary texts are too young to be canonized. An alternative way of measuring preference for contemporary texts may be literary prizes. As mentioned in the Section 1, the Nobel Prize has been used as an indicator of preference [4,6]. A comparison between our data and Nobel prize winning books is another project that would broaden our understanding of structural reflexes of preference, and of preference itself.

The program of computational textual aesthetics has been heavily influenced by relevant studies from other domains. For example, statistical properties of (time) series have been analysed for music [31], poetry [49], and even bird song [50]. Measures such as autocorrelation, variability, surprise and predictability have also been used to predict musical preferences in humans [30,51]. As our own work has been influenced by the work on vision, we conclude with a remark on how our results relate to the visual domain. Here, basic perceptual features are also richer and more variable (or less predictable) in artworks than in many types of non-art images. Examples include the spatial distribution of luminance and colour edges across an image [52] and other basic visual features, such as edge orientation, spatial frequency tuning and colour–opponent spatial organization [12,53]. Whether a high degree of variation in such basic perceptual features is universal across aesthetic domains (visual art, literature, dance, music, etc.) is unclear at present.

In relevant studies, perceptual (structural) differences between traditional visual artworks and contemporary art have been observed. With the rise of modern art at the end of the 19th century, the pattern of image properties in visual artworks diversified [54,55]. In parallel, perceptual features that mediate the sensual beauty of artworks became less central for aesthetic judgements. Instead, image content and cultural context emerged as guides of what beholders prefer [56].

We speculate that there are parallels between aesthetic experience in the visual domain and in reading. In both domains, aesthetic preference seems to be related to the interplay

between predictability and surprise. Our results are compatible with the hypothesis that the determinants of aesthetic experience in reading, like those in vision, are partly time-invariant, and partly culturally determined. A certain amount of variability and unpredictability, reflected in Approximate Entropy and Shannon Entropy in the present, study seems to be a good candidate for a time-invariant factor. However, in order to gain a better understanding of the determinants of preference in reading, several follow-up studies as sketched above will be needed.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/e25030486/s1, Section S1: Previous Work on the Temporal Analysis of Language; Section S2: Approximate Entropy; Section S3: Effects of Genre; Figure S1. Boxplot of ApEn for all genres and for all text properties; Figure S2. Boxplot of ShEn for all genres and for all text properties; Table S1: List of texts in the Jena Corpus of Contemporary Expository and Fictional Prose (JCEFP Corpus); Figure S3: Approximate Entropy vs. Shannon Entropy for each text category and for each text property; Table S2: Median values of Approximate Entropy (ApEn) for all text properties and for all fictional text categories; Table S3: Median values of Shannon Entropy (ShEn) for all text properties and for all fictional text categories. Refs. [7,43,57–73] are cited in the Supplementary Materials.

Author Contributions: Conceptualization, M.M., C.R. and V.G.; methodology, M.M.; software, M.M.; validation, M.M., C.R. and V.G.; formal analysis, M.M.; investigation, M.M. and V.G.; resources, C.R. and V.G.; data curation, M.M.; writing—original draft preparation, M.M., C.R. and V.G.; writing—review and editing, M.M., C.R. and V.G.; visualization, M.M.; supervision, C.R. and V.G.; project administration, C.R. and V.G.; funding acquisition, C.R. and V.G. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge support by the German Research Foundation grant number 512648189, the Open Access Publication Fund of the Thueringer Universitaets und Landesbibliothek Jena, and the German Research Foundation grant number 391160252.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to copyright restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kao, J.; Jurafsky, D. A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry. In *Proceedings of the Workshop on Computational Linguistics for Literature*; The Association for Computer Linguistics: Montréal, QC, Canada, 2012; pp. 8–17.
2. Ashok, V.; Feng, S.; Choi, Y. Success with Style: Using Writing Style to Predict the Success of Novels. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 18–21 October 2013; pp. 1753–1764.
3. Maharjan, S.; Arevalo, J.; Montes, M.; González, F.; Solorio, T. 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, Valencia, Spain, 3–7 April 2017; pp. 1217–1227. [CrossRef]
4. Febres, G.; Jaffe, K. Quantifying Structure Differences in Literature Using Symbolic Diversity and Entropy Criteria. *J. Quant. Linguist.* **2017**, *24*, 16–53. [CrossRef]
5. Maharjan, S.; Kar, S.; Montes, M.; González, F.A.; Solorio, T. 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), New Orleans, LA, USA, 1–6 June 2018; pp. 259–265. [CrossRef]
6. Bizzoni, Y.; Nielbo, K.; Thomsen, M. 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—NLP4DH 2022, Taipei, Taiwan, 21–24 November 2022.
7. Mohseni, M.; Redies, C.; Gast, V. Approximate Entropy in Canonical and Non-Canonical Fiction. *Entropy* **2022**, *24*, 278. [CrossRef] [PubMed]
8. Palmer, S.E.; Schloss, K.B.; Sammartino, J. Visual Aesthetics and Human Preference. *Annu. Rev. Psychol.* **2013**, *64*, 77–107. [CrossRef] [PubMed]
9. Guillory, J. Canonical and Non-Canonical: A Critique of the Current Debate. *ELH* **1987**, *54*, 483–527. [CrossRef]

10. Tötösy de Zepetnek, S. Toward a Theory of Cumulative Canon Formation: Readership in English Canada. *Mosaic* **1994**, *27*, 107–119.
11. Underwood, T.; Sellers, J. The Long Durée of Literary Prestige. *Mod. Lang. Q.* **2016**, *77*, 321–344. [CrossRef]
12. Brachmann, A.; Redies, C. Computational and Experimental Approaches to Visual Aesthetics. *Front. Comput. Neurosci.* **2017**, *11*, 102. [CrossRef]
13. Bloom, H. *The Western Canon: The Books and School of the Ages*; Harcourt: New York, NY, USA, 1994.
14. Green, C. Introducing the Corpus of the Canon of Western Literature: A Corpus for Culturomics and Stylistics. *Lang. Lit.* **2017**, *26*, 282–299. [CrossRef]
15. Mohseni, M.; Gast, V.; Redies, C. Fractality and Variability in Canonical and Non-Canonical English Fiction and in Non-Fictional Texts. *Front. Psychol.* **2021**, *12*, 920. [CrossRef]
16. Even-Zohar, I. Polysystem Studies. *Poet. Today* **1990**, *11*, 9–26. [CrossRef]
17. Yucesoy, B.; Wang, X.; Huang, J.; Barabási, A.L. Success in books: A big data approach to bestsellers. *EPJ Data Sci.* **2018**, *7*, 1–25. [CrossRef]
18. Wang, X.; Yucesoy, B.; Varol, O.; Eliassi-Rad, T.; Barabasi, A.L. Success in books: Predicting book sales before publication. *EPJ Data Sci.* **2019**, *8*, 31. [CrossRef]
19. Vasyliuk, A.; Matseliukh, Y.; Batiuk, T.; Luchkevych, M.; Shakleina, I.; Harbuzynska, H.; Kondratiuk, S.; Zelenska, K. Intelligent Analysis of Best-Selling Books Statistics on Amazon. In Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022), Gliwice, Poland, 12–13 May 2022 ; Volume 3171, CEUR Workshop Proceedings, pp. 1432–1462.
20. Pfister, M. *Das Drama: Theorie und Analyse*; utb GmbH: München, Germany, 1988.
21. Genette, G. *Narrative Discourse: An Essay in Method*; Cornell University Press: New York, NY, USA, 1983; Volume 3.
22. Smith, C. *Modes of Discourse. The Local Structure of Texts*; Cambridge University Press: Cambridge, UK, 2003.
23. Biber, D. *Variation across Speech and Writing*; Cambridge University Press: Cambridge, UK, 1991.
24. Biber, D. *Dimensions of Register Variation. A Cross-Linguistic Comparison*; Cambridge University Press: Cambridge, UK, 1995.
25. Biber, D.; Conrad, S. *Register, Genre, and Style*; Cambridge University Press: Cambridge, UK, 2019.
26. Egbert, J.; Mahlberg, M. Fiction—One Register or Two? Speech and Narration in Novels. *Regist. Stud.* **2020**, *2*, 72–101. [CrossRef]
27. Simonton, D.K. Lexical Choices and Aesthetic Success: A Computer Content Analysis of 154 Shakespeare Sonnets. *Comput. Humanit.* **1990**, *24*, 251–264. [CrossRef]
28. Forsythe, A.; Nadal, M.; Sheehy, N.; Cela-Conde, C.J.; Sawey, M. Predicting beauty: Fractal dimension and visual complexity in art. *Br. J. Psychol.* **2011**, *102*, 49–70. [CrossRef]
29. Bizzoni, Y.; Peura, T.; Thomsen, M.R.; Nielbo, K. Sentiment Dynamics of Success: Fractal Scaling of Story Arcs Predicts Reader Preferences. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*; NLP Association of India (NLPAI): Silchar, India, 2021; pp. 1–6.
30. Gold, B.P.; Pearce, M.T.; Mas-Herrero, E.; Dagher, A.; Zatorre, R.J. Predictability and Uncertainty in the Pleasure of Music: A Reward for Learning? *J. Neurosci.* **2019**, *39*, 9397–9409. [CrossRef]
31. Koelsch, S.; Vuust, P.; Friston, K. Predictive Processes and the Peculiar Case of Music. *Trends Cogn. Sci.* **2019**, *23*, 63–77. [CrossRef]
32. Zipf, G.K. *Human Behavior and the Principle of Least Effort*; Addison-Wesley Press: Cambridge, MA, USA, 1949.
33. Ferrer i Cancho, R.; Solé, R. Least Effort and the Origins of Scaling in Human Language. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 788–791. [CrossRef]
34. Forsyth, R.S. Pops and Flops: Some Properties of Famous English Poems. *Empir. Stud. Arts* **2000**, *18*, 49–67. [CrossRef]
35. Chang, M.C.; Yang, A.C.C.; Stanley, H.E.; Peng, C.K. Measuring Information-Based Energy and Temperature of Literary Texts. *Phys. A Stat. Mech. Its Appl.* **2017**, *468*, 783–789. [CrossRef]
36. Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; Manning, C.D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*; Association for Computational Linguistics: Florence, Italy, 2020; pp. 101–108. [CrossRef]
37. Stanza: Available Models & Languages. Available online: https://stanfordnlp.github.io/stanza/available_models.html (accessed on 1 March 2023).
38. Schneider, G.; Hundt, M.; Oppliger, R. Part-Of-Speech in Historical Corpora: Tagger Evaluation and Ensemble Systems on ARCHER. In Proceedings of the 13th Conference on Natural Language Processing, KONVENS 2016, Bochum, Germany, 19–21 September 2016; Volume 16, Bochumer Linguistische Arbeitsberichte.
39. Pincus, S.M. Approximate Entropy as a Measure of System Complexity. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 2297–2301. [CrossRef]
40. Li, X.; Cui, S.; Voss, L. Using Permutation Entropy to Measure the Electroencephalographic Effects of Sevoflurane. *Anesthesiology* **2008**, *109*, 448–456. [CrossRef]
41. Hayashi, K.; Shigemi, K.; Sawa, T. Neonatal Electroencephalography Shows Low Sensitivity to Anesthesia. *Neurosci. Lett.* **2012**, *517*, 87–91. [CrossRef]
42. Lee, G.; Fattinger, S.; Mouthon, A.L.; Noirhomme, Q.; Huber, R. Electroencephalogram Approximate Entropy Influenced by Both Age and Sleep. *Front. Neuroinform.* **2013**, *7*, 33. [CrossRef]
43. Zar, J.H. *Biostatistical Analysis*, 5th ed.; Pearson: Upper Saddle River, NJ, USA, 2010.

44. Dietterich, T.G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* **1998**, *10*, 1895–1923. [CrossRef]
45. Gast, V.; Wehmeier, C.; Vanderbeke, D. A Register-Based Study of Interior Monologue in James Joyce's Ulysses. *Literature* **2023**, *3*, 42–65. .: 10.3390/literature3010004. [CrossRef]
46. Moore, A.T.; Schwitzgebel, E. The experience of reading. *Conscious. Cogn.* **2018**, *62*, 57–68. [CrossRef]
47. Belfi, A.M.; Vessel, E.A.; Starr, G.G. Individual ratings of vividness predict aesthetic appeal in poetry. *Psychol. Aesthet. Creat. Arts* **2018**, *12*, 341. [CrossRef]
48. Piţur, S.; Miu, A.C. Poetry-elicited emotions: Reading experience and psychological mechanisms. *Psychol. Aesthet. Creat. Arts* **2022** . [CrossRef]
49. Scharinger, M.; Wagner, V.; Knoop, C.; Menninghaus, W. Melody in poems and songs: Fundamental statistical properties predict aesthetic evaluation. *Psychol. Aesthet. Creat. Arts* **2022** . [CrossRef]
50. Roeske, T.C.; Kelty-Stephen, D.; Wallot, S. Multifractal analysis reveals music-like dynamic structure in songbird rhythms. *Sci. Rep.* **2018**, *8*, 4570. [CrossRef] [PubMed]
51. Miles, S.A.; Rosen, D.S.; Grzywacz, N.M. A Statistical Analysis of the Relationship between Harmonic Surprise and Preference in Popular Music. *Front. Hum. Neurosci.* **2017**, *11*, 263. [CrossRef] [PubMed]
52. Redies, C.; Brachmann, A.; Wagemans, J. High Entropy of Edge Orientations Characterizes Visual Artworks From Diverse Cultural Backgrounds. *Vis. Res.* **2017**, *133*, 130–144. [CrossRef] [PubMed]
53. Geller, H.A.; Bartho, R.; Thömmes, K.; Redies, C. Statistical image properties predict aesthetic ratings in abstract paintings created by neural style transfer. *Front. Neurosci.* **2022**, *16*, 999720 . [CrossRef] [PubMed]
54. Mather, G. Visual Image Statistics in the History of Western Art. *Art Percept.* **2018**, *6*, 97–115 . [CrossRef]
55. Redies, C.; Brachmann, A. Statistical Image Properties in Large Subsets of Traditional Art, Bad Art, and Abstract Art. *Front. Neurosci.* **2017**, *11*, 593. [CrossRef]
56. Chamberlain, R. The Interplay of Objective and Subjective Factors in Empirical Aesthetics. In *Human Perception of Visual Information: Psychological and Computational Perspectives*; Ionescu, B., Bainbridge, W.A., Murray, N., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 115–132. [CrossRef]
57. Kumar, A.; Lease, M.; Baldridge, J. Supervised Language Modeling for Temporal Resolution of Texts. In *CIKM'11, Proceedings of the 20th ACM International Conference on Information and Knowledge Management*; Association for Computing Machinery: New York, NY, USA, 2011; pp. 2069–2072. [CrossRef]
58. Garcia-Fernandez, A.; Ligozat, A.L.; Dinarelli, M.; Bernhard, D. When Was It Written? Automatically Determining Publication Dates. In *SPIRE'11, Proceedings of the 18th International Conference on String Processing and Information Retrieval*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 221–236.
59. Ciobanu, A.M.; Dinu, L.P.; Şulea, O.M.; Dinu, A.; Niculae, V. Temporal Text Classification for Romanian Novels set in the Past. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*; INCOMA Ltd. Shoumen, BULGARIA: Hissar, Bulgaria, 2013; pp. 136–140.
60. Štajner, S.; Zampieri, M. Stylistic Changes for Temporal Text Classification. *Lect. Notes Comput. Sci.* **2013**, *8082*, 519–526. [CrossRef]
61. Gómez-Adorno, H.; Posadas-Duran, J.P.; Ríos-Toledo, G.; Sidorov, G.; Sierra, G. Stylometry-based approach for detecting writing style changes in literary texts. *Comput. Sist.* **2018**, *22*, 47–53. [CrossRef]
62. Efremova, J.; García, A.M.; Zhang, J.; Calders, T. Effects of evolutionary linguistics in text classification. In *Proceedings of the International Conference on Statistical Language and Speech Processing*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 50–61.
63. Liebeskind, C.; Liebeskind, S. Deep Learning for Period Classification of Historical Hebrew Texts. *J. Data Min. Digit. Humanit.* **2020**, *2020*. [CrossRef]
64. Gopidi, A.; Alam, A. Computational Analysis of the Historical Changes in Poetry and Prose. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*; Association for Computational Linguistics: Florence, Italy, 2019; pp. 14–22. [CrossRef]
65. Lagutina, K.; Poletaev, A.; Lagutina, N.; Boychuk, E.; Paramonov, I. Automatic Extraction of Rhythm Figures and Analysis of Their Dynamics in Prose of 19th-21st Centuries. In Proceedings of the 2020 26th Conference of Open Innovations Association (FRUCT), Yaroslavl, Russia, 20–24 April 2020; pp. 247–255. [CrossRef]
66. Lagutina, K.V.; Manakhova, A.M. Automated Search and Analysis of the Stylometric Features That Describe the Style of the Prose of 19th–21st Centuries. *Autom. Control Comput. Sci.* **2021**, *55*, 866–876. [CrossRef]
67. Degaetano-Ortlieb, S. Stylistic Variation Over 200 Years of Court Proceedings According to Gender and Social Class. In *Proceedings of the Second Workshop on Stylistic Variation*; Association for Computational Linguistics: New Orleans, LA, USA, 2018; pp. 1–10. [CrossRef]
68. Fankhauser, P.; Knappen, J.; Teich, E. Topical Diversification over Time in the Royal Society Corpus; Jagiellonian University; Pedagogical University: Kraków, 2016; Digital Humanities. Available online: https://ids-pub.bsz-bw.de/frontdoor/index/index/year/2016/docId/5474 (accessed on 1 March 2023).
69. Bizzoni, Y.; Degaetano-Ortlieb, S.; Fankhauser, P.; Teich, E. Linguistic Variation and Change in 250 Years of English Scientific Writing: A Data-Driven Approach. *Front. Artif. Intell.* **2020**, *3*. [CrossRef]

70. Wang, G.; Wang, H.; Sun, X.; Nan, W.; Wang, L. Linguistic complexity in scientific writing: A large-scale diachronic study from 1821 to 1920. *Scientometrics* **2022**, *128*, 441–460. [CrossRef]
71. Krielke, M.P.; Fischer, S.; Degaetano-Ortlieb, S.; Teich, E. System and use of wh-relativizers in 200 years of English scientific writing. In Proceedings of the 10th International Corpus Linguistics Conference, Cardiff, Wales, UK, 23–27 July 2019.
72. US Novel Corpus. Available online: https://textual-optics-lab.uchicago.edu/us_novel_corpus (accessed on 1 March 2023).
73. Degaetano-Ortlieb, S.; Strötgen, J. Diachronic Variation of Temporal Expressions in Scientific Writing Through the Lens of Relative Entropy. In *Language Technologies for the Challenges of the Digital Age*; Rehm, G., Declerck, T., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 259–275.

Supplementary Materials: Comparative Analysis of Preference in Contemporary and Earlier Texts Using Entropy Measures

1. Previous Work on Temporal Analysis of Language

Kumar *et al.* [1] applied the Kullback-Leibler divergence to histograms derived from the language model of documents to determine the time period of short stories published between 1798 and 2008. Their approach predicted dates of publication, though within a wide range of temporal difference. Garcia-Fernandez *et al.* [2] used several external resources, e.g. Wikipedia, to detect the publication date of article excerpts from seven French newspapers published from the early 19th century to the middle of the 20th century. Their system correctly detected the year of publication for up to 14% of documents, and the decade of publication for up to 42% of the documents. Ciobanu *et al.* [3] applied temporal text classification to Romanian novels, showing that a high classification accuracy can be achieved with bag-of-word features.

Štajner and Zampieri [4] proposed a supervised method to study stylistic changes in Portuguese historical texts spanning a range from the 17th century to the early 20th century. They revealed a noticeable shift in lexical diversity and lexical richness of texts written in the 19th and 20th centuries, compared to the texts from the two preceding centuries. Gómez-Adorno *et al.* [5] analyzed changes in the writing styles of seven authors using stylometric features. They achieved a high level of accuracy in the detection of writing stages for works by some authors. Efremova *et al.* [6] worked on Dutch historical notary acts spanning a period of more than six centuries. First, they identified time periods based on historical events. Then, using Term Frequency–Inverse Document Frequency (TF-IDF) features and spectral clustering, they classified texts into these periods, reaching a high accuracy level.

Liebeskind and Liebeskind [7] applied neural classifiers to historical Hebrew texts from four different periods. They showed that neural networks outperformed classic machine learning algorithms in the task of period classification. Gopidi and Alam [8] studied stylistic differences between prose and poem in two different time spans of 1870–1920 and 1970–2019. They combined quantitative analyses with interpretation based on close reading. Using features derived from grammatical properties, meter and rhyme, they concluded on the basis of their classification results that modern poetry is more similar to prose, in comparison to older poetry and prose. Lagutina *et al.* [9] studied rhythm in 300 English and Russian prosaic texts dating from the 19th century to the 21st century. They found that rhythm figures change with time and can be regarded as a determinant of an author's style. This approach was extended in Lagutina and Manakhova [10], who analyzed not only rhythm features but also low-level features, i.e. at the level of the word and symbol. They compared stylometric features of texts from each decade and revealed that rhythm features have changed more than other features in the texts under analysis. They also found that the average lengths of sentences decreased in Russian texts in a wave form during the last two centuries, while the average word length increased consistently. Degaetano-Ortlieb [11] used lexical and grammatical models to explore stylistic variation of different groups of language users. She showed that temporal stylistic changes across genders and classes of society can be captured using Relative Entropy (Kullback-Leibler Divergence).

Diachronic analysis of scientific texts have also been addressed in previous studies. For example, Fankhauser *et al.* [12] used topic modeling to monitor topic developments in a corpus of the Royal Society of London. Unsurprisingly, their observations showed that scientific topics have diversified over time, while individual documents have been more specialized in terms of topics. Bizzoni *et al.* [13] proposed to investigate diachronic language changes using Relative Entropy as a measure of diversification. They analyzed scientific English texts published in a period longer than 250 years and showed evidence of register formation and, at the same time, diversification in word usage. Wang *et al.* [14] explored temporal variation of linguistic structures by applying Kolmogorov complexity to different types of scientific texts. Their analysis showed that while the scientific lexicon has been enriched during the time period analyzed, the language complexity has declined in favor of grammatical simplification. There are also studies which focused on more specific language structures in scientific writing, e.g. temporal expressions [15] and *wh*-words [16].

2. Approximate Entropy

For series $X = x(1), \cdots, x(n)$, sub-sequences of length m, $y_i^m = [x(i), \cdots, x(i + (m-1))]$, and tolerance r, Approximate Entropy (ApEn) is computed as follows:

1. Compute Chebyshev distance between each sub-sequence y_i^m and y_j^m:

$$d_{i,j}^m = \max_k |y_i^m(k) - y_j^m(k)|$$

2. Using the Heaviside function, $\mathbb{1}(.)$, whose value is zero for negative arguments and one for positive arguments, for each sub-sequence y_i^m, compute:

$$C_i^m(r) = \frac{1}{n-m+1} \sum_{j=1}^{n-m+1} \mathbb{1}(r - d_{i,j}^m)$$

3. Compute

$$\phi^m(r) = \frac{1}{n-m+1} \sum_{i=1}^{n-m+1} \log(C_i^m(r))$$

4. Repeat step 1 to 3 for sub-sequences of length $m+1$ to compute $\phi^{m+1}(r)$.
5. Calculate ApEn as

$$\texttt{ApEn}(m,r) = \phi^m(r) - \phi^{m+1}(r)$$

3. Effects of Genre

Genre characteristics may have some effects on structural properties of texts. We analyzed Approximate Entropy (ApEn) and Shannon Entropy (ShEn) of series of 6 POS-tags (Noun, Verb, Adjective, Adverb, Pronoun, Preposition) and sentence length in various genres. We used texts from the US Novel corpus [17]. We analyzed texts written from 1980 to 2000, which are more similar to the contemporary texts in our corpus (JCEFP) in terms of publication time. We did not include multi-labeled texts and we analyzed categories which contained at least 50 texts. As a result, 3168 texts from 7 categories were selected. Figures S1 and S2 show the results for ApEn and ShEn, respectively. As the plots demonstrate, the distribution of ApEn and ShEn are different for the text categories. An ANOVA test confirmed that the distribution of ApEn and ShEn differed between the genres for all text properties. However, the ranking of the genres changes for different text properties and none of them consistently shows the highest or lowest values compared to the other genres.

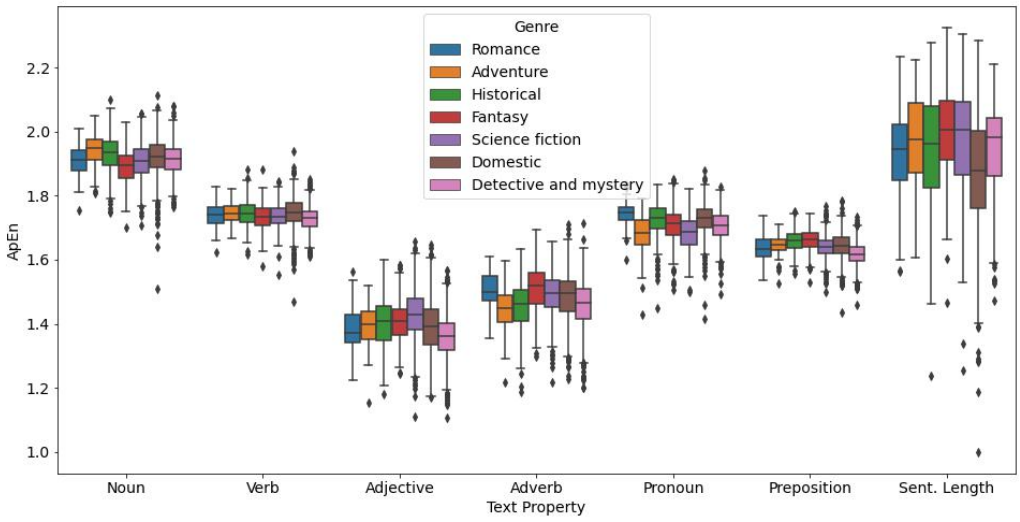

Figure S1. Boxplot of ApEn for all genres and for all text properties.

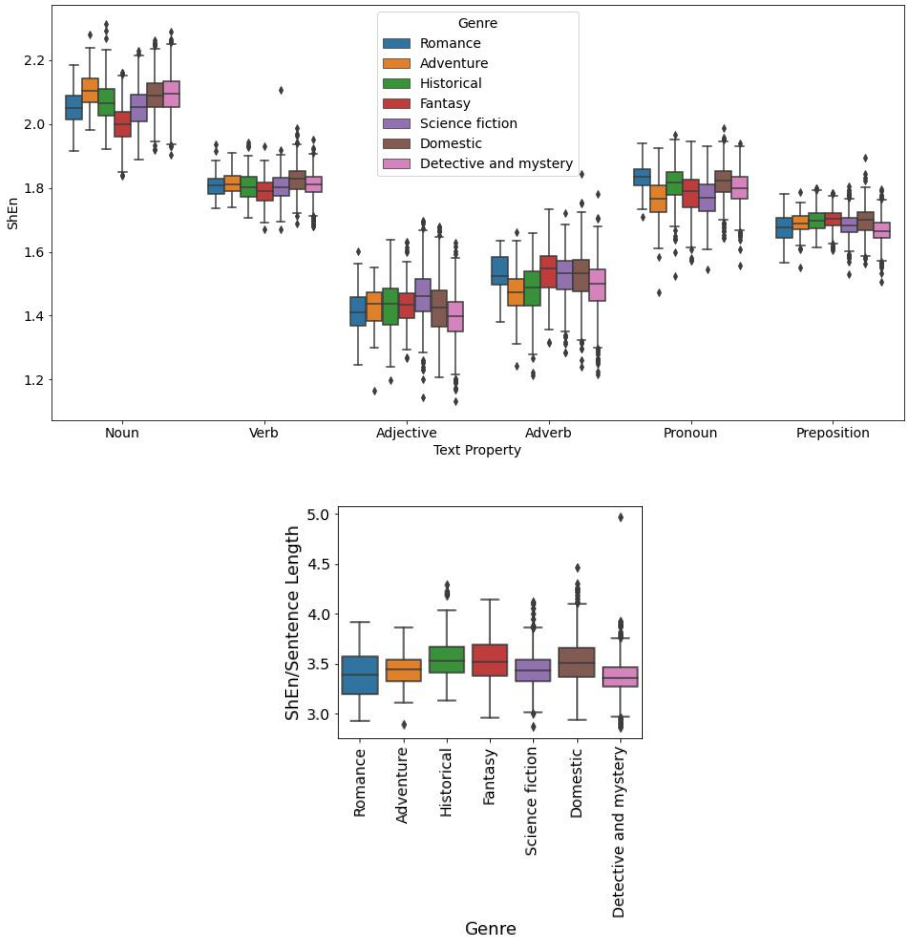

Figure S2. Boxplot of ShEn for all genres and for all text properties. The plot of sentence length is separated as its values are in a different range.

4. Tables and Figures

Table S1: List of texts in the Jena Corpus of Contemporary Expository and Fictional Prose (JCEFP Corpus). Bestseller texts were selected from lists of the New York Times Best Sellers. Corpus of non-bestseller texts were downloaded from www.smashwords.com. Non-fictional texts were selected from various sources.

	Title	**Author(s)**	**Year of Publication**	**Category**
1	11.22.63	Stephen King	2011	Bestseller
2	5th Horseman	James Patterson, Maxine Paetro	2006	Bestseller
3	A Dance with Dragons	George R. R. Martin	2011	Bestseller
4	A Man Called Ove	Fredrik Backman	2012	Bestseller
5	A Thousand Splendid Suns	Khaled Hosseini	2007	Bestseller
6	Alex Cross	James Patterson	2009	Bestseller
7	All the Light We Cannot See	Anthony Doerr	2014	Bestseller
8	Allegiant	Veronica Roth	2013	Bestseller
9	Angels and Demons	Dan Brown	2000	Bestseller
10	At First Sight	Nicholas Sparks	2005	Bestseller
11	Book of the Dead	Patricia Cornwell	2008	Bestseller
12	Catching Fire	Suzanne Collins	2009	Bestseller
13	Cell	Stephen King	2006	Bestseller

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
14	Cross	James Patterson	2006	Bestseller
15	Cross Country	James Patterson	2008	Bestseller
16	Cross Fire	James Patterson	2011	Bestseller
17	Dead or Alive	Tom Clancy, Grant Blackwood	2010	Bestseller
18	Dead Rechoning	Charlaine Harris	2001	Bestseller
19	Dear John	Nicholas Sparks	2006	Bestseller
20	Desecration	Tim F. LaHaye, Jerry B. Jenkins	2001	Bestseller
21	Doctor Sleep	Stephen King	2013	Bestseller
22	Double Cross	James Patterson	2007	Bestseller
23	Dreamcatcher	Stephen King	2001	Bestseller
24	Eleven on Top	Janet Evanovich	2005	Bestseller
25	Everythings Eventual	Stephen King	2002	Bestseller
26	Fearless Fourteen	Janet Evanovich	2008	Bestseller
27	Fifty Shades Darker	EL James	2011	Bestseller
28	Fifty Shades Freed	EL James	2012	Bestseller
29	Fifty Shades of Grey	EL James	2011	Bestseller
30	Fifty Shades of Louisa May	L. M. Anonymous	2012	Bestseller
31	Finger Lickin Fifteen	Janet Evanovich	2009	Bestseller
32	Four Blind Mice	James Patterson	2001	Bestseller
33	Freedom	Jonathan Franzen	2010	Bestseller
34	Full Dark, No Stars	Stephen King	2010	Bestseller
35	Go Set a Watchman	Harper Lee	2015	Bestseller
36	Hannibal Rising	Thomas Harris	2006	Bestseller
37	Inferno	Dan Brown	2013	Bestseller
38	Insurgent	Veronica Roth	2012	Bestseller
39	Jack Ryan12 The Teeth Of The Tiger	Tom Clancy	2003	Bestseller
40	Kill Alex Cross	James Patterson	2011	Bestseller
41	Light from Heaven	Jan Karon	2005	Bestseller
42	Liseys Story	Stephen King	2006	Bestseller
43	Locked On	Tom Clancy, Mark Greaney	2011	Bestseller
44	London Bridges	James Patterson	2004	Bestseller
45	Mary Mary	James Patterson	2004	Bestseller
46	Micro	Michael Crichton, Richard Preston	2011	Bestseller
47	Mockingjay	Suzanne Collins	2010	Bestseller
48	Next	Michael Crichton	2006	Bestseller
49	Pirate Latitudes	Michael Crichton	2009	Bestseller
50	Port Mortuary	Patricia Cornwell	2010	Bestseller
51	Prey	Michael Crichton	2002	Bestseller
52	Red Rabbit	Tom Clancy	2002	Bestseller
53	Safe Haven	Nicholas Sparks	2010	Bestseller
54	Sizzling Sixteen	Janet Evanovich	2010	Bestseller
55	Skipping Christmas	John Grisham	2001	Bestseller
56	Smokin Seventeen	Janet Evanovich	2011	Bestseller
57	State of Fear	Michael Crichton	2004	Bestseller
58	The 6th Target	James Patterson, Maxine Paetro	2007	Bestseller
59	The Appeal	John Grisham	2008	Bestseller
60	The Associate	John Grisham	2009	Bestseller
61	The Best of Me	Nicholas Sparks	2010	Bestseller
62	The Big Bad Wolf	James Patterson	2003	Bestseller
63	The Broker	John Grisham	2005	Bestseller
64	The Choice	Nicholas Sparks	2007	Bestseller
65	The Christmas Sweater	Glenn Beck	2008	Bestseller
66	The Confession	John Grisham	2010	Bestseller
67	The Da Vinci Code	Dan Brown	2003	Bestseller
68	The Darkest Evening of the Year	Dean Koontz	2007	Bestseller
69	The Fault in Our Stars	John Green	2012	Bestseller
70	The Five People You Meet in Heaven	Mitch Albom	2003	Bestseller
71	The Girl on the Train	Paula Hawkins	2015	Bestseller
72	The Girl Who Kicked the Hornets Nest	Stieg Larsson	2007	Bestseller
73	The Help	Kathryn Stockett	2009	Bestseller
74	The Historian	Elizabeth Kostova	2005	Bestseller
75	The Host	Stephenie Meyer	2008	Bestseller
76	The House of Hades	Rick Riordan	2013	Bestseller

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
77	The Hunger Games	Suzanne Collins	2010	Bestseller
78	The King of Torts	John Grisham	2003	Bestseller
79	The Last Song	Nicholas Sparks	2009	Bestseller
80	The Litigators	John Grisham	2011	Bestseller
81	The Lost Symbol	Dan Brown	2009	Bestseller
82	The Lovely Bones	Alice Sebold	2002	Bestseller
83	The Mark of Athena	Rick Riordan	2012	Bestseller
84	The Martian	Andy Weir	2014	Bestseller
85	The Quickie	James Patterson, Michael Ledwidge	2009	Bestseller
86	The Remnant	Tim LaHaye, Jerry B. Jenkins	2002	Bestseller
87	The Rule of Four	Ian Caldwell and Dustin Thomason	2004	Bestseller
88	The Shelters of Stone	Jean M. Auel	2002	Bestseller
89	The Story of Edgar Sawtelle	David Wroblewski	2008	Bestseller
90	The Wedding	Nicholas Sparks	2003	Bestseller
91	True Believer	Nicholas Sparks	2005	Bestseller
92	Twelve Sharp	Janet Evanovich	2006	Bestseller
93	Your Heart Belongs to Me	Dean Ray Koontz	2008	Bestseller
94	A Highland Affair	Richard F. Jones	2018	Non-Bestseller
95	A Long, Cool Rain	Linda Seed	2017	Non-Bestseller
96	A Unicorn's Memoir	Stephanie Menges	2020	Non-Bestseller
97	After the Fire	Kathryn Shay	2003	Non-Bestseller
98	After the Fog: A Novel	Kathleen Shoop	2012	Non-Bestseller
99	An English Visitor	Graham Wilson	2007	Non-Bestseller
100	An Ignorant Witch	E M Graham	2019	Non-Bestseller
101	Ash and Water	Everleigh Miles	2020	Non-Bestseller
102	Awakened	Brenda K. Davies	2012	Non-Bestseller
103	Bad Choices Make Good Stories: Going to New York	Oliver Markus Malloy	2017	Non-Bestseller
104	Beautiful Secret	Claire Raye	2019	Non-Bestseller
105	Beg For You: A Small Town Romance	Sherilee Gray	2019	Non-Bestseller
106	Breaking the Rules	Ruth Ann Nordin	2020	Non-Bestseller
107	Bridge Through Time	Scott Spotson	2014	Non-Bestseller
108	Cactus Island	William Manchee	2006	Non-Bestseller
109	Case of the One-Eyed Tiger	Jeffrey M. Poole	2020	Non-Bestseller
110	Christmas Magic	Alexandra Moody	2018	Non-Bestseller
111	Clocks Locks and Danger	Lizzie Lewis	2020	Non-Bestseller
112	Co-Ed	Rachel Van Dyken	2018	Non-Bestseller
113	Cole	Tory Richards	2019	Non-Bestseller
114	Crocodile Man	Graham Wilson	2017	Non-Bestseller
115	Darkhouse	Karina Halle	2011	Non-Bestseller
116	Delusions	Christina Smith	2012	Non-Bestseller
117	Dragma's Keep	Vance Pumphrey	2015	Non-Bestseller
118	Dreaming of You	S.E. Felida	2020	Non-Bestseller
119	Duly Noted	H.M. Shander	2016	Non-Bestseller
120	Dying for a Living	Kory M. Shrum	2014	Non-Bestseller
121	Elfin	Quinn Loftis	2012	Non-Bestseller
122	Eternally Bound	Brenda K. Davies	2016	Non-Bestseller
123	Ever Onward	Wayne Mee	2011	Non-Bestseller
124	Everything we Lost	Kate Smith	2016	Non-Bestseller
125	Falling For You	Leeanna Morgan	2018	Non-Bestseller
126	Falling Into The Black	Lauren Runow	2017	Non-Bestseller
127	Fated Dreams	Christina Smith	2012	Non-Bestseller
128	Fighting Destiny	Amelia Hutchins	2013	Non-Bestseller
129	Fire Song	Val St. Crowe	2016	Non-Bestseller
130	Frey	Melissa Wright	2019	Non-Bestseller
131	Genesis Code	Eliza Green	2012	Non-Bestseller
132	Girl in a Cage	Graham Wilson	2019	Non-Bestseller
133	Governor	Lesli Richardson	2018	Non-Bestseller
134	Hellfire - Treachery	Simon Goodson	2020	Non-Bestseller
135	Human Intelligence	Klaus Marre	2013	Non-Bestseller
136	I Woke Up Feeling Thailand	D. Bruno Starrs	2012	Non-Bestseller
137	Ice Homme	Vance Pumphrey	2015	Non-Bestseller

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
138	In Defense of Mankind	Ron L. Carter, H.R. Carter	2019	Non-Bestseller
139	Just One Kiss	Jami Rogers	2020	Non-Bestseller
140	Just Visiting	Graham Wilson	2015	Non-Bestseller
141	Killing Me Softly	Bianca Sloane	2012	Non-Bestseller
142	King's Crown	Marie Johnston	2020	Non-Bestseller
143	Last Breath	Greg Tuck	2015	Non-Bestseller
144	Legacy of Darkness: Undercover Mistress	Dai Fuse	2020	Non-Bestseller
145	Like a Memory	Abbi Glines	2017	Non-Bestseller
146	Little Lost Girl	Graham Wilson	2011	Non-Bestseller
147	Lost	Jodi Kae	2016	Non-Bestseller
148	Lost Girl	Chanda Hahn	2016	Non-Bestseller
149	Lost in Me	Lexi Ryan	2014	Non-Bestseller
150	Loveoid	J.L. Morin	2020	Non-Bestseller
151	Moonstone	Linda Seed	2015	Non-Bestseller
152	Mystic Mayhem	Sally J. Smith	2015	Non-Bestseller
153	No More Tears	Sandy Appleyard	2020	Non-Bestseller
154	Nowhere Man	Graham Wilson	2020	Non-Bestseller
155	Of Beast and Beauty	Chanda Hahn	2019	Non-Bestseller
156	Our Broken Pieces	M.E. Clayton	2020	Non-Bestseller
157	Pierced	Sydney Landon	2015	Non-Bestseller
158	Possession	Graham Wilson	2018	Non-Bestseller
159	Prince of Wolves	Quinn Loftis	2013	Non-Bestseller
160	Ragnarok Conspiracy	Rob J. Meijer	2018	Non-Bestseller
161	Red Hot Mama	Reagan McDaniels	2021	Non-Bestseller
162	Redemption Lake	Susan Clayton-Goldner	2017	Non-Bestseller
163	Return of the Breaker	Graham Wilson	2018	Non-Bestseller
164	Riley 's Secret	Christina Smith	2012	Non-Bestseller
165	Riley 's Torment	Christina Smith	2013	Non-Bestseller
166	Rise of the Gladiator	Cheree Alsop	2020	Non-Bestseller
167	Rosebloom	Christine Keleny	2008	Non-Bestseller
168	Safe Haven	Leeanna Morgan	2016	Non-Bestseller
169	Saving Grace	Sandy James	2013	Non-Bestseller
170	Sealed with a Kiss	Leeanna Morgan	2016	Non-Bestseller
171	Seeking Dr. Magic	Scott Spotson	2018	Non-Bestseller
172	Shadow Phantoms	H.P. Mallory	2020	Non-Bestseller
173	Silent Star	James F. David	2014	Non-Bestseller
174	Some Call it Love	Sarah Peis	2018	Non-Bestseller
175	Soul of the Dragon	Natalie J. Damschroder	2012	Non-Bestseller
176	The American Terrorist	Ron L. Carter	2012	Non-Bestseller
177	The Broken	Igor Ljubuncic	2013	Non-Bestseller
178	The Bulldog	Tricia Andersen	2019	Non-Bestseller
179	The Diary	Graham Wilson	2014	Non-Bestseller
180	The Dragon Question	L. Darby Gibbs	2018	Non-Bestseller
181	The Dragon's Slave	Lacey St. Sin	2016	Non-Bestseller
182	The Empty Place	Graham Wilson	2014	Non-Bestseller
183	The Heartbreaker	Tricia Andersen	2015	Non-Bestseller
184	The House on Persimmon Road	Jackie Weger	2014	Non-Bestseller
185	The Library of Antiquity	Vance Pumphrey	2013	Non-Bestseller
186	The Mystery of the Hidden Jewels	Carrie Cross	2014	Non-Bestseller
187	The Old Balmain House	Graham Wilson	2011	Non-Bestseller
188	The Platinum Dragon	Vance Pumphrey	2015	Non-Bestseller
189	The Prophecy	Jeffrey M. Poole	2012	Non-Bestseller
190	The Storm Inside	Alexis Anne	2013	Non-Bestseller
191	The Strange Life of Brandon Chambers	Scott Spotson	2020	Non-Bestseller
192	The Truth About James	Sarah Tork	2014	Non-Bestseller
193	The Valkyrie	L K Walker	2018	Non-Bestseller
194	The Watchers	Lynnie Purcell	2011	Non-Bestseller
195	Theft of the Giant's Soul	Mark Cheverton	2019	Non-Bestseller
196	Trapped	Graham Wilson	2017	Non-Bestseller
197	Trigger	L. P. Dover	2017	Non-Bestseller
198	True Colors	Melissa Pearl	2014	Non-Bestseller
199	Unlucky Charm	Kimberly Gordon	2017	Non-Bestseller
200	Wagon Trail Bride	Ruth Ann Nordin	2016	Non-Bestseller

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
201	Wild On You	Justiss Alliance	2014	Non-Bestseller
202	Witch's Bell	Odette C. Bell	2010	Non-Bestseller
203	Yesterday's Sins	Shirley Wine	2012	Non-Bestseller
204	A Compendium of Philosophical Concepts and Methods	Peter S. Fosl: JULIAN BAGGINI	2020	Nonfictional
205	A Grand Origin for Grand Canyon	Michael Oard	2014	Nonfictional
206	Act Natural: A Cultural History of Misadventures in Parenting	Jennifer Traig	2019	Nonfictional
207	Aesthetics Volume II	Dietrich von Hildebrand	2019	Nonfictional
208	AI Ethics	Mark Coeckelbergh	2020	Nonfictional
209	Anatomy 101: From Muscles and Bones to Organs and Systems Your Guide to How the Human Body Works	Kevin Langford	2015	Nonfictional
210	Aristotle's ladder, Darwin's tree : the evolution of visual metaphors for biological order	J. David Archibald	2014	Nonfictional
211	Art and Architecture of Viceregal Latin America 1521-1821	Kelly Donahue Wallace	2008	Nonfictional
212	Awkward: The Science of Why We re Socially Awkward and Why That s Awesome	Ty Tashiro	2017	Nonfictional
213	Balance: A Dizzying Journey Through the Science of Our Most Delicate Sense	Carol Svec	2017	Nonfictional
214	Biochemistry	Denise R. Ferrier PhD	2013	Nonfictional
215	Biostatistics: The Bare Essentials	Geoffrey R. Norman David L. Sreiner	2014	Nonfictional
216	Bird Sense: What It s Like to Be a Bird	Tim Birkhead	2012	Nonfictional
217	Blind Descent: The Quest to Discover the Deepest Place on Earth	James M. Tabor	2010	Nonfictional
218	Cannibalism: A Perfectly Natural History	Bill Schutt	2017	Nonfictional
219	City Shaped Churches: Planting Churches in a Global Era	Linda Bergquist Michael Crane	2018	Nonfictional
220	Climate Action Planning: A Guide to Creating Low Carbon Resilient Communities	Michael R. Boswell, Adrienne I. Greve Tammy L. Seale	2019	Nonfictional
221	Communication and Capitalism: A Critical Theory	Christian Fuchs	2020	Nonfictional
222	Competition Overdose: How Free Market Mythology Transformed Us from Citizen Kings to Market Servants	Maurice E. Stucke: Ariel Ezrachi	2020	Nonfictional
223	Conflict and Contest in Nietzsche s Philosophy	Herman Siemens: James Pearson	2019	Nonfictional
224	Coping with Trauma Related Dissociation: Skills Training for Patients and Therapists	Suzette Boon, Kathy Steele, Onno van der Hart	2011	Nonfictional
225	Copyright Law for Librarians and Educators	Kenneth D. Crews	2011	Nonfictional
226	Cosmic DNA at the Origin: A Hyperdimension before the Big Bang: the Infinite Spiral Staircase Theory	Chris H. Hardy	2015	Nonfictional
227	Dazzled and Deceived: Mimicry and Camouflage	Peter Forbes	2011	Nonfictional
228	Decriminalizing Domestic Violence: A Balanced Policy Approach to Intimate Partner Violence	Leigh Goodmark	2018	Nonfictional
229	Dog Behaviour Evolution and Cognition	Adam Miklosi	2015	Nonfictional
230	Drawn from Paradise: The Natural History Art and Discovery of the Birds of Paradise with Rare Archival Art	David Attenborough, Errol Fuller	2012	Nonfictional
231	EcoCities: Rebuilding Cities in Balance with Nature	Richard Register	2006	Nonfictional
232	Emotional Intelligence: Emotional Mastery Influence	Modern Psychology Publishing	2019	Nonfictional
233	Enforcement of Maritime Claims	David Jackson	2005	Nonfictional
234	Environmental and Low Temperature Geochemistry	Peter Ryan	2019	Nonfictional
235	Epistemology and the Psychology of Human Judgment	Michael A. Bishop, J. D. Trout	2005	Nonfictional
236	Essentials of Environmental Health	Robert H. Friis	2012	Nonfictional
237	Ethnosociology: The Foundations	Alexander Dugin	2019	Nonfictional
238	Five Pillars of the Mind: Redesigning Education to Suit the Brain	Tracey Tokuhama Espinosa	2019	Nonfictional
239	Fundamentals of Structural Mechanics Dynamics and Stability	A.I. Rusakov	2020	Nonfictional
240	Geology by Design: Interpreting Rocks and their Catastrophic Record	Carl R. Froede	2007	Nonfictional
241	Global Sales and Contract Law	Ingeborg Schwenzer, Christopher Kee, Pascal Hachem	2012	Nonfictional

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
242	Happy City: Transforming Our Lives Through Urban Design	Charles Montgomery	2015	Nonfictional
243	Hegel on Possibility: Modality Perfection and Dialectics	Nahum Brown	2020	Nonfictional
244	Historical Dictionary of Romantic Art and Architecture	Allison Lee Palmer	2019	Nonfictional
245	How Not to Be Wrong: The Power of Mathematical Thinking	Jordan Ellenberg	2014	Nonfictional
246	How Sexual Desire Works: The Enigmatic Urge	Frederick Toates	2014	Nonfictional
247	Human Nature	David Berlinski	2019	Nonfictional
248	Indian Philosophy: A Reader	Jonardon Ganeri	2020	Nonfictional
249	Lessons from Nanoelectronics: A New Perspective on Transport	Supriyo Datta	2012	Nonfictional
250	Leviathan or The Whale	Philip Hoare	2009	Nonfictional
251	Life Unfolding: How the Human Body Creates Itself	Jamie A. Davies	2014	Nonfictional
252	Liquid: The Delightful and Dangerous Substances That Flow Through Our Lives	Mark Miodownik	2018	Nonfictional
253	Marine Insurance Legislation	Robert Merkin, Jennifer Lavelle	2010	Nonfictional
254	Molecules Microbes and Meals: The Surprising Science of Food	Alan Kelly	2019	Nonfictional
255	Natural Law and Human Rights: Toward a Recovery of Practical Reason	Pierre Manent, Ralph C. Hancock, Daniel J. Mahoney	2020	Nonfictional
256	Night School: Wake Up to the Power of Sleep	Richard Wiseman	2014	Nonfictional
257	On Food and Cooking: The Science and Lore of the Kitchen rev. and updated	Harold McGee	2004	Nonfictional
258	Particle Physics: An Introduction	Robert Purdy	2018	Nonfictional
259	Philosophy: A Christian Introduction	James K. Dew, Jr.: Paul M. Gould	2019	Nonfictional
260	Preserving: The canning and freezing guide for all seasons	Pat Crocker	2012	Nonfictional
261	Principles of International Economic Law	Matthias Herdegen	2016	Nonfictional
262	Psychology moment by moment: a guide to enhancing your clinical practice with mindfulness and meditation	Elise E. Labbe	2011	Nonfictional
263	Radically Open Dialectical Behavior Therapy: Theory and Practice for Treating Disorders of Overcontrol	Thomas R. Lynch	2018	Nonfictional
264	Rest Play Grow: Making Sense of Preschoolers (or Anyone Who Acts Like One)	Deborah MacNamara	2016	Nonfictional
265	Revolution in Mind: The Creation of Psychoanalysis	George Makari	2008	Nonfictional
266	Same Sex Parenting Research: A Critical Assessment	Walter R. Schumm	2018	Nonfictional
267	Say What You Mean: A Mindful Approach to Nonviolent Communication	Oren Jay Sofer	2018	Nonfictional
268	Science in Black and White: How Biology and Environment Shape Our Racial Divide	Alondra Oubre	2020	Nonfictional
269	Secrets of Your Cells: Discovering your body s Inner Intelligence	Sondra Barrett	2013	Nonfictional
270	Sex and the Failed Absolute	Slavoj Zizek	2019	Nonfictional
271	Sex on Earth: A Celebration of Animal Reproduction	Jules Howard	2015	Nonfictional
272	Ship Registration: Law and Practice	Richard Coles, Edward Watt	2002	Nonfictional
273	Smells: A Cultural History of Odours in Early Modern Times	Robert Muchembled	2020	Nonfictional
274	Social Problems: Community Policy and Social Action	Anna Leon Guerrero	2018	Nonfictional
275	Social Psychology	Thomas E. Heinzen, Wind Goodfriend	2018	Nonfictional
276	Spheres of Influence: The Social Ecology of Racial and Class Inequality	Douglas S. Massey	2014	Nonfictional
277	Structural Geology	Haakon Fossen	2016	Nonfictional
278	Sustainable Landscape Construction: A Guide to Green Building Outdoors	Kim Sorvig J. William Thompson	2018	Nonfictional
279	The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power	Shoshana Zuboff	2019	Nonfictional
280	The Aquaponic Farmer: A Complete Guide to Building and Operating a Commercial Aquaponic System	Adrian Southern, Whelm King	2017	Nonfictional
281	The Architecture of Learning: Designing Instruction for the Learning Brain	Kevin D. Washburn	2010	Nonfictional

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
282	The Battery: How Portable Power Sparked a Technological Revolution	Henry Schlesinger	2010	Nonfictional
283	The Charisma Myth: How Anyone Can Master the Art and Science of Personal Magnetism	Olivia Fox Cabane	2012	Nonfictional
284	The Chemistry of Wine: From Blossom to Beverage and Beyond	David R. Dalton	2017	Nonfictional
285	The Constitution in Exile: How the Federal Government Has Seized Power by Rewriting the Supreme Law of the Land	Andrew P. Napolitano	2007	Nonfictional
286	The Constitution: An Introduction	Michael Stokes Paulsen, Luke Paulsen	2015	Nonfictional
287	The Copenhagen Conspiracy	David Ferry	2019	Nonfictional
288	The dolphin in the mirror : exploring dolphin minds and saving dolphin lives	Reiss Diana	2011	Nonfictional
289	The Dream Universe: How Fundamental Physics Lost Its Way	David Lindley	2020	Nonfictional
290	The Empathy Advantage: Coaching Children to Be Kind Respectful and Successful	Lynne Azarchi	2021	Nonfictional
291	The End of Ownership: Personal property in the digital economy	Aaron Perzanowski, Jason Schultz	2016	Nonfictional
292	The Equations World	Boris Pritsker	2019	Nonfictional
293	The Evolution of Beauty: How Darwin s Forgotten Theory of Mate Choice Shapes the Animal World and Us	Richard O. Prum	2017	Nonfictional
294	The Greatest Show on Earth: The Evidence for Evolution	Richard Dawkins	2009	Nonfictional
295	The Kingdom of Infinite Number: A Field Guide	Bryan Bunch	2011	Nonfictional
296	The Law of State Immunity	Hazel Fox, QC and Philippa Webb	2013	Nonfictional
297	The Life Organic: The Theoretical Biology Club and the Roots of Epigenetics	Erik L. Peterson	2017	Nonfictional
298	The Limits of Epistemology	Markus Gabriel, Alex Englander	2020	Nonfictional
299	The Longing for Less: Living with Minimalism	Kyle Chayka	2020	Nonfictional
300	The Map That Changed the World: William Smith and the Birth of Modern Geology	Simon Winchester	2009	Nonfictional
301	The Master and His Emissary: The Divided Brain and the Making of the Western World	Iain McGilchrist	2019	Nonfictional
302	The Music of the Primes: Searching to Solve the Greatest Mystery in Mathematics	Marcus du Sautoy	2012	Nonfictional
303	The Narrow Corridor: States Societies and the Fate of Liberty	Daron Acemoglu, James A. Robinson	2019	Nonfictional
304	The New Evil: Understanding the Emergence of Modern Violent Crime	Michael H. Stone: Gary Brucato	2019	Nonfictional
305	The Origins of Political Order: From Prehuman Times to the French Revolution	Francis Fukuyama	2011	Nonfictional
306	The paradox of choice: why more is less	Barry Schwartz	2005	Nonfictional
307	The Pope of Physics: Enrico Fermi and the Birth of the Atomic Age	Gino Segrè Bettina Hoerlin	2016	Nonfictional
308	The Post Traumatic Stress Disorder Sourcebook: A Guide to Healing Recovery and Growth	Glenn R. Schiraldi	2016	Nonfictional
309	The Science of Communicating Science: The Ultimate Guide	Craig Cormick	2020	Nonfictional
310	The Silencing: How the Left is Killing Free Speech	Kirsten Powers	2015	Nonfictional
311	The Sixth Extinction: An Unnatural History	Elizabeth Kolbert	2014	Nonfictional
312	The Social Lens: An Invitation to Social and Sociological Theory	Kenneth Allan	2013	Nonfictional
313	The Story of Light	Ben Bova	2012	Nonfictional
314	The World According to Physics	Jim Al Khalili	2020	Nonfictional
315	Thinking fast and slow	Kahneman Daniel	2015	Nonfictional
316	This Life: Secular Faith and Spiritual Freedom	Martin Hagglund	2019	Nonfictional
317	Unconditional Parenting: Moving from Rewards and Punishments to Love and Reason	Alfie Kohn	2006	Nonfictional
318	Underground: A Human History of the Worlds Beneath Our Feet	Will Hunt	2019	Nonfictional

Continued on next page

Table S1 – *Continued from previous page*

	Title	Author(s)	Year of Publication	Category
319	Unequal Childhoods: Class Race and Family Life Second Edition with an Update a Decade Later	Annette Lareau	2011	Nonfictional
320	Unpeople: Britain s Secret Human Rights Abuses	Mark Curtis	2004	Nonfictional
321	What the Nose Tells the Mind	A. S. Barwich	2020	Nonfictional
322	Why Diets Make Us Fat: The Unintended Consequences of Our Obsession With Weight Loss	Sandra Aamodt	2016	Nonfictional
323	Why You Hear What You Hear: An Experiential Approach to Sound Music and Psychoacoustics	Eric J. Heller	2013	Nonfictional
324	Witcraft: The Invention of Philosophy in English	Jonathan Ree	2019	Nonfictional
325	Words That Change Minds: The 14 Patterns for Mastering the Language of Influence	Shelle Rose Charvet	2019	Nonfictional

Table S2: Median values of Approximate Entropy (ApEn) for all text properties and for all fictional text categories. ApEn values were analyzed for contemporary fictional (N = 204) vs. non-fictional (N = 122) texts, and for earlier fictional (N =206) vs. earlier non-fictional (N = 185) texts. The asterisks indicate whether the differences between the two text categories in contemporary or earlier text categories are statistically significant (Mann-Whitney U test; ns, not significant; *, $p \leq 0.05$; **, $p \leq 0.01$; and ***, $p \leq 0.001$). Values that are significantly higher within a pair of columns are shown in boldface. 95% confidence intervals for the median (according to [18]) are shown in parentheses. Data for earlier texts are from the study by Mohseni *et al.* [19].

	Contemporary		Earlier	
Text Property	Fictional	Non-Fictional	Fictional	Non-Fictional
Sentence Length	2.01 (1.99, 2.02)	1.98 (1.96, 2.00)^{ns}	1.87 (1.86, 1.88)	1.90 (1.88, 1.92)^{ns}
Noun	1.89 (1.88, 1.90)	1.90 (1.88, 1.91)^{ns}	**1.85 (1.84, 1.86)**	1.82 (1.81, 1.84)**
Verb	1.72 (1.71, 1.73)	**1.74 (1.73, 1.75)*****	1.714 (1.706, 1.723)	**1.756 (1.745, 1.764)*****
Adjective	1.38 (1.37, 1.39)	**1.63 (1.61, 1.64)*****	1.488 (1.469, 1.494)	**1.58 (1.55, 1.60)*****
Adverb	**1.51 (1.495, 1.516)**	1.38 (1.35, 1.40)***	**1.49 (1.48, 1.50)**	1.36 (1.34, 1.39)***
Pronoun	**1.72 (1.71, 1.73)**	1.27 (1.19, 1.32)***	**1.695 (1.685, 1.704)**	1.31 (1.28, 1.36)***
Preposition	1.62 (1.61, 1.63)	**1.66 (1.65, 1.65)*****	1.678 (1.672, 1.683)	**1.691 (1.686, 1.697)*****

Table S3: Median values of Shannon Entropy (ShEn) for all text properties and for all fictional text categories. ShEn values were analyzed for contemporary fictional (N = 204) vs. non-fictional (N = 122) texts, and for earlier fictional (N =206) vs. earlier non-fictional (N = 185) texts. The asterisks indicate whether the differences between the two text categories in contemporary or earlier text categories are statistically significant (Mann-Whitney U test; ns, not significant; *, $p \leq 0.05$; **, $p \leq 0.01$; and ***, $p \leq 0.001$). Values that are significantly higher within a pair of columns are shown in boldface. 95% confidence intervals for the median (according to [18]) are shown in parentheses. Data for earlier texts are from the study by Mohseni *et al.* [19].

	Contemporary		Earlier	
Text Property	Fictional	Non-Fictional	Fictional	Non-Fictional
Sentence Length	3.39 (3.37, 3.42)	**3.89 (3.84, 3.91)*****	3.96 (3.91, 4.03)	**4.10 (4.07, 4.16)*****
Noun	2.05 (2.02, 2.06)	**2.07 (2.05, 2.08)****	1.98 (1.97, 1.99)	1.97 (1.95, 1.99)^{ns}
Verb	1.79 (1.78, 1.80)	**1.82 (1.81, 1.83)*****	1.785 (1.779, 1.792)	**1.844 (1.836, 1.853)*****
Adjective	1.41 (1.40, 1.42)	**1.68 (1.66, 1.69)*****	1.52 (1.51, 1.53)	**1.63 (1.61, 1.66)*****
Adverb	**1.54 (1.52, 1.56)**	1.41 (1.38, 1.43)***	**1.52 (1.51, 1.53)**	1.40 (1.37, 1.42)***
Pronoun	**1.81 (1.807, 1.820)**	1.33 (1.24, 1.40)***	**1.79 (1.78, 1.80)**	1.37 (1.33, 1.42)***
Preposition	1.66 (1.657, 1.673)	**1.71 (1.705, 1.720)*****	1.736 (1.729, 1.744)	**1.76 (1.75, 1.77)*****

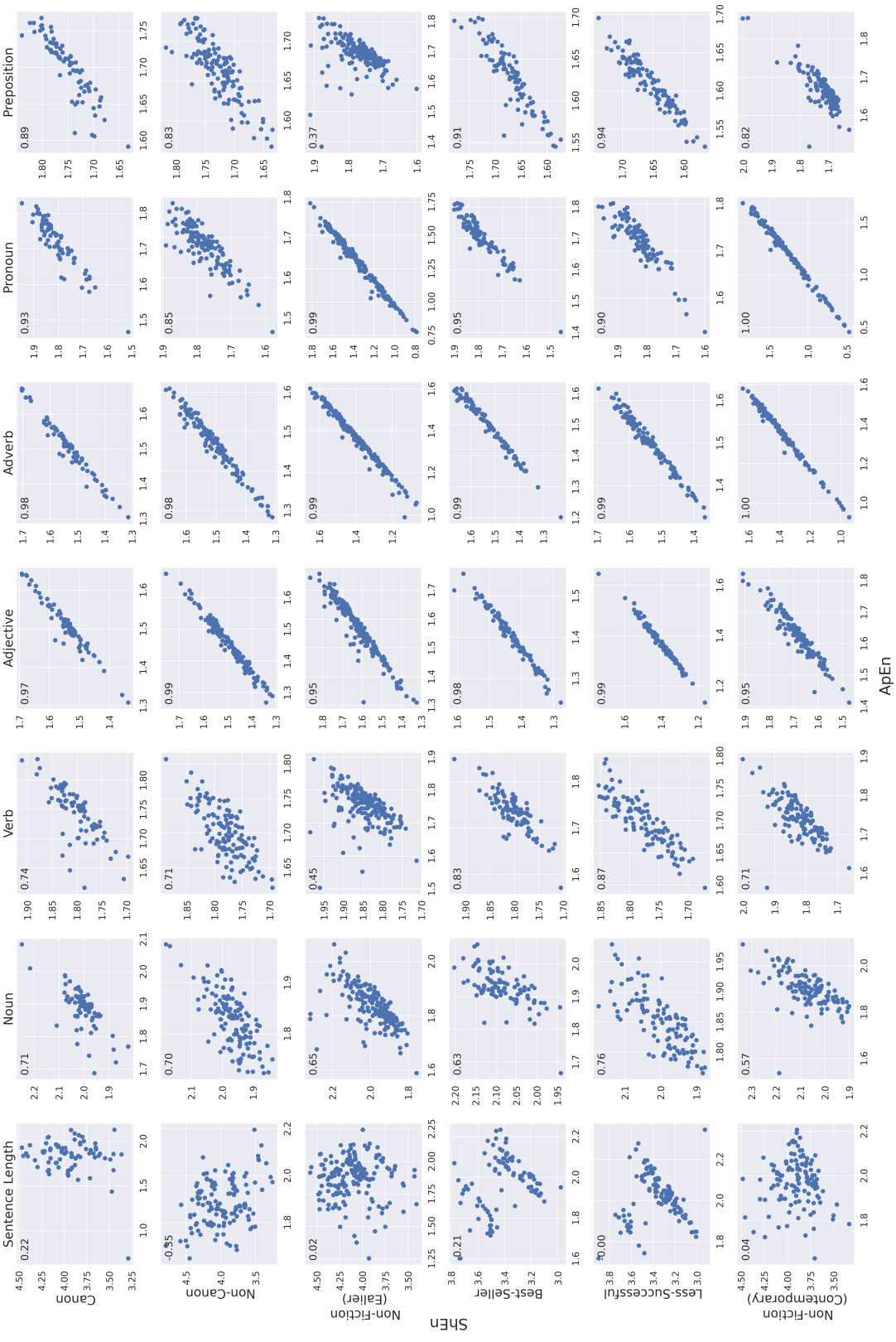

Figure S3. Approximate Entropy (ApEn; x-axis) vs. Shannon Entropy (ShEn; y-axis) for each text category (rows) and for each text property (columns). The correlation value is shown on the top left corner of each plot. To make the plots more readable, two outliers, which were both nonfictional texts written in the 19th century, were removed: A Dictionary of English Synonyms and Synonymous Expressions by Richard Soule and A Synopsis of the Birds of North America by John James Audubon.

References

1. Kumar, A.; Lease, M.; Baldridge, J. Supervised Language Modeling for Temporal Resolution of Texts. Proceedings of the 20th ACM International Conference on Information and Knowledge Management; Association for Computing Machinery: New York, NY, USA, 2011; CIKM '11, p. 2069–2072. doi:10.1145/2063576.2063892.
2. Garcia-Fernandez, A.; Ligozat, A.L.; Dinarelli, M.; Bernhard, D. When Was It Written? Automatically Determining Publication Dates. Proceedings of the 18th International Conference on String Processing and Information Retrieval; Springer-Verlag: Berlin, Heidelberg, 2011; SPIRE'11, p. 221–236.
3. Ciobanu, A.M.; Dinu, L.P.; Şulea, O.M.; Dinu, A.; Niculae, V. Temporal Text Classification for Romanian Novels set in the Past. Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013; INCOMA Ltd. Shoumen, BULGARIA: Hissar, Bulgaria, 2013; pp. 136–140.
4. Štajner, S.; Zampieri, M. Stylistic Changes for Temporal Text Classification. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2013, Vol. 8082, pp. 519–526. doi:10.1007/978-3-642-40585-3_65.
5. Gómez-Adorno, H.; Posadas-Duran, J.P.; Ríos-Toledo, G.; Sidorov, G.; Sierra, G. Stylometry-Based Approach for Detecting Writing Style Changes in Literary Texts. *Computación y Sistemas* **2018**, *22*, 47–53.
6. Efremova, J.; García, A.M.; Zhang, J.; Calders, T. Effects of Evolutionary Linguistics in Text Classification. International Conference on Statistical Language and Speech Processing. Springer, 2015, pp. 50–61. doi:10.46298/jdmdh.5864.
7. Liebeskind, C.; Liebeskind, S. Deep Learning for Period Classification of Historical Hebrew Texts. *Journal of Data Mining & Digital Humanities* **2020**, *2020*.
8. Gopidi, A.; Alam, A. Computational Analysis of the Historical Changes in Poetry and Prose. Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change; Association for Computational Linguistics: Florence, Italy, 2019; pp. 14–22. doi:10.18653/v1/W19-4702.
9. Lagutina, K.; Poletaev, A.; Lagutina, N.; Boychuk, E.; Paramonov, I. Automatic Extraction of Rhythm Figures and Analysis of Their Dynamics in Prose of 19th-21st Centuries. 26th Conference of Open Innovations Association (FRUCT), 2020, pp. 247–255. doi:10.23919/FRUCT48808.2020.9087430.
10. Lagutina, K.V.; Manakhova, A.M. Automated Search and Analysis of the Stylometric Features That Describe the Style of the Prose of 19th–21st Centuries. *Automatic Control and Computer Sciences* **2021**, *55*, 866–876. doi:10.3103/S0146411621070257.
11. Degaetano-Ortlieb, S. Stylistic Variation Over 200 Years of Court Proceedings According to Gender and Social Class. Proceedings of the Second Workshop on Stylistic Variation; Association for Computational Linguistics: New Orleans, 2018; pp. 1–10. doi:10.18653/v1/W18-1601.
12. Fankhauser, P.; Knappen, J.; Teich, E. Topical Diversification Over Time In The Royal Society Corpus; Jagiellonian University; Pedagogical University: Kraków, 2016; Digital Humanities.
13. Bizzoni, Y.; Degaetano-Ortlieb, S.; Fankhauser, P.; Teich, E. Linguistic Variation and Change in 250 Years of English Scientific Writing: A Data-Driven Approach. *Frontiers in Artificial Intelligence* **2020**, *3*. doi:10.3389/frai.2020.00073.
14. Wang, G.; Wang, H.; Sun, X.; Nan, W.; Wang, L. Linguistic complexity in scientific writing: A large-scale diachronic study from 1821 to 1920. *Scientometrics* **2022**, *128*, 441–460. doi:10.1007/s11192-022-04550-z.
15. Degaetano-Ortlieb, S.; Strötgen, J. Diachronic Variation of Temporal Expressions in Scientific Writing Through the Lens of Relative Entropy. Language Technologies for the Challenges of the Digital Age; Rehm, G.; Declerck, T., Eds.; Springer International Publishing: Cham, 2018; pp. 259–275.
16. Krielke, M.P.; Fischer, S.; Degaetano-Ortlieb, S.; Teich, E. System and use of wh-relativizers in 200 years of English scientific writing. Proceedings of the 10th International Corpus Linguistics Conference, Cardiff, Wales, UK, 2019.
17. US Novel Corpus. https://textual-optics-lab.uchicago.edu/us_novel_corpus. Accessed: 01-03-2023.
18. Zar, J.H. *Biostatistical Analysis*, 5 ed.; Pearson: Upper Saddle River, NJ, 2010.
19. Mohseni, M.; Redies, C.; Gast, V. Approximate Entropy in Canonical and Non-Canonical Fiction. *Entropy* **2022**, *24*. doi:10.3390/e24020278.

Chapter 6

DIACHRONIC ANALYSIS OF STRUCTURAL FEATURES

Our analysis of preferred and non-preferred texts in two different periods (Chapter 5) suggested that structural properties of texts are not consistent in all time periods and change in the course of time. We thoroughly investigate this diachronic change in this chapter.

We analyzed global structures of texts in a period of 120 years, from 1880 to 2000. We applied two analysis approaches: long-range correlations, which analyze auto-correlation patterns present in a text (Chapter 3 / Mohseni et al., 2021), and predictability, which measures the amount of surprise in the underlying structures of a text (Chapter 4 / Mohseni et al., 2022).

For our experiments, we used the US Novel corpus[1], which consists of more than 9000 fictional English texts. We converted texts into series of sentence length and frequencies of major POS-tags. Long-range correlations were analyzed using Multi-Fractal Detrended Fluctuation Anlysis (Kantelhardt et al., 2002; Mohseni et al., 2021), which results in two fractal features. The first one is the Hurst exponent (the degree of fractality), which indicates the strength of long-range correlation patterns present in the input series. The second one is the degree of multifractality, which represents the complexity of long-range correlations. Predictability analysis was accomplished using Approximate Entropy, which is a measure of surprise in sequential organization (Pincus, 1991; Mohseni et al., 2022).

Our results show that both fractal features and Approximate Entropy values of texts change over a time period longer than one century. The MFDFA showed that contemporary (late 20^{th} century) texts exhibit weaker long-range correlations for most structural properties (lower degrees of fractality) compared to earlier (late 19^{th} and early 20^{th} century) texts, and the complexity of long-range correlations (the degree of multifractality) decreases in contemporary texts. Predictability analysis also showed that contemporary text are less predictable than earlier texts.

Both long-range correlation and predictability analyses suggest that contemporary texts have less complex structures than earlier texts. Whether this structural changes

[1] `https://textual-optics-lab.uchicago.edu/us_novel_corpus`

have any implication for textual preference remains a question for further analysis.

References

Kantelhardt, Jan W., Stephan A. Zschiegner, Eva Koscielny-Bunde, Shlomo Havlin, Armin Bunde, and H.Eugene Stanley (2002). "Multifractal detrended fluctuation analysis of nonstationary time series". In: *Physica A: Statistical Mechanics and Its Applications* 316.1, pp. 87–114. DOI: 10.1016/S0378-4371(02)01383-3.

Mohseni, Mahdi, Volker Gast, and Christoph Redies (2021). "Fractality and Variability in Canonical and Non-Canonical English Fiction and in Non-Fictional Texts". In: *Frontiers in Psychology* 12, p. 920. DOI: 10.3389/fpsyg.2021.599063.

Mohseni, Mahdi, Christoph Redies, and Volker Gast (2022). "Approximate Entropy in Canonical and Non-Canonical Fiction". In: *Entropy* 24.2, p. 277. DOI: 10.3390/e24020278.

Pincus, Steven M (1991). "Approximate Entropy as a Measure of System Complexity". In: *Proceedings of the National Academy of Sciences* 88.6, pp. 2297–2301. DOI: 10.1073/pnas.88.6.229.

Diachronic Analysis of Global Structural Properties in English Texts From 1880 to 2000

Abstract

In this study, we investigate long-range correlations and degrees of predictability in the sequential organization of American fictional texts from the period between 1880 and 2000. As basic measurements we use sentence lengths and frequencies of POS-tags in 25-words segments of texts. We determine long-range correlations, a type of auto-correlation that emerges from local structural properties, using Multi-Fractal Detrended Fluctuation Analysis (MFDFA). (Un)predictability of structural features is determined using Approximate Entropy, a measure of surprise in sequential organization. These methods are used to identify changes in global structural features of texts in a large corpus. We show that for most structural properties, long-range correlations weakened in the period under study. The earlier texts show stronger as well as more complex long-range correlations than the contemporary texts. Approximate Entropy increases steadily for most text properties, thus showing a higher degree of unpredictability for the contemporary texts. Both results suggest that texts became less clearly structured, and hence, less predictable, during the 20th century. Our study confirms a common intuition of modern texts exhibiting less regularity in their formal make-up. Beyond the general trend towards less structure we observe an anomaly in the period of American modernism, which is associated with particularly high values for the degree of multi-fractality, pointing to multi-layered structures in the texts of this period.

1 Introduction

Texts, specially longer ones, can be analyzed in terms of global structural properties. Long-range correlations and predictability analyses are two approaches that reveal the characteristics of underlying structures of texts. Texts are known to exhibit spatio-temporal auto-correlations, in the sense that the occurrence of any given linguistic item is a function of the context, specifically of the preceding material (see e.g. studies on 'priming', such as Hoey, 2004). From a different point of view, variation in underlying structural elements can be analyzed using (ir)regularity and (un)predictability measures. These types of analyses have only rarely been used for the diachronic study of texts. In this study we analyze structural properties of English texts spanning a period of over a century from the two above-mentioned perspectives, i.e. long-range correlations and (un)predictability.

Long-range correlations have been studied in various domains such as music (Sanyal et al., 2016), animal songs (Roeske et al., 2018) and biological processes (Das et al., 2016). Similar analyses have also been carried out on texts (for example, Drożdż et al., 2016; Yang et al., 2016; Vieira et al., 2018; Mohseni et al., 2021). In the context of diachronic analysis, Chen and Liu (2018) conducted a study to analyze long-range word-length correlations in Chinese texts across 2000 years. They revealed that the Hurst exponent, which shows the strength of correlation patterns, decreased for shorter ranges, but it increased for larger ranges.

The concept of predictability can be regarded as a statistical counterpart of regularity. Consequently, (un)predictability analysis using entropy metrics has been used to measure the amount of (ir)regularity within language structures (e.g. Piantadosi et al., 2011; Montemurro and Zanette, 2011; Mahowald et al., 2013; Ferrer-i-Cancho et al., 2015; Montemurro and Zanette, 2016; Koplenig et al., 2017; Kanwal et al., 2017). In an earlier study (Mohseni et al., 2022), we used entropy measures to determine "surprise" as a global structural property of text, which allowed us to distinguish canonical from non-canonical texts with high accuracy. Expanding on our previous research, we conducted an inter-period analysis by

defining canonical and non-canonical texts as preferred and non-preferred fictional texts, respectively and compared them with bestselling and non-bestselling contemporary texts (Mohseni et al., 2023). The categories '(non-)canonical' and '(non-)best-selling' were intended to reflect 'preference' in the sense of empirical aesthetics. We showed that predictability analysis can effectively differentiate between preferred (canonical, best-selling) and non-preferred (non-canonical, non-best-selling) texts in both time periods. However, the differentiation between the two text categories based on global statistical properties was clearer for the earlier period than for the later period.

While the temporal analysis of global text structure has so far been neglected, diachronic changes with more local scope (at the level of the word and the surrounding text) have figured more prominently in earlier research. Distributional semantic methods, especially ones using neural language representations, have been used to trace semantic and, to some extent, grammatical changes across time (see, for example, Kim et al., 2014; Hamilton et al., 2016; Szymanski, 2017; Del Tredici et al., 2019). Previous studies on related tasks such as time period classification (Kumar et al., 2011; Ciobanu et al., 2013; Efremova et al., 2015; Liebeskind and Liebeskind, 2020), style change analysis (Štajner and Zampieri, 2013; Gómez-Adorno et al., 2018) and temporal analysis of news documents (Garcia-Fernandez et al., 2011; Popescu and Strapparava, 2015) have concentrated on the classification of texts without providing a deeper analysis of underlying global structures of texts. An exception is provided by Gopidi and Alam (2019), who analyzed meter and rhyme, two prominent structural poetic features, and grammatical features of poems in two different time periods and compared them with the prose texts from the same periods. They showed that contemporary poetry is less structured and thus more similar to prose compared to earlier poetry. Similarly, Lagutina and Manakhova (2021) showed a significant shift in the rhythms of British and Russian texts throughout the 19th and 20th centuries using stylometric analysis. Another study relevant to our research, though in the context of academic prose, is the work by Wang and Manning (2012). They analyzed the linguistic complexity of texts published by *Philosophical Transaction of the Royal Society of London* between 1821 and 1920. They showed that – while morphological complexity increased during this time period – syntactic complexity decreased, leading to simplification. However, these studies primarily focused on the analysis of local structures (at the level of words or phrases) and did not analyze global structures, in the sense of the overall organization of texts.

In order to study long-range correlations in texts, we use Multi-Fractal Detrended Fluctuation Analysis (MFDFA; Kantelhardt et al., 2002). MFDFA condenses correlation information of series into two statistics: (I) the Hurst exponent, which reflects the strength of correlations in series, and (II) the width of the singularity spectrum, also known as the degree of multifractality, which represents the complexity of long-range correlations (Mohseni et al., 2021). To analyze the amount of predictability in texts, we use Approximate Entropy (ApEn; Pincus, 1991), which has been proposed to measure (ir)regularity in time series. ApEn is a measure of (un)predictability in linearly ordered random variables (Mohseni et al., 2022, 2023). We apply these methods to English texts from 1880 to 2000. We use the US Novel corpus, which consists of more than 9000 fictional texts written by a wide range of authors in various genres (see, Section 2.1).

The study is organized as follows: we first explain our methods of global structural analysis, i.e. MFDFA and ApEn (Section 3). Section 2.1 focuses on describing the corpus. We discuss how we represent texts in Section 2.2. The results are presented in Section 4, which is followed by a discussion and some conclusions in Section 5.

2 Data

2.1 The corpus

For our study, we used the US Novel corpus[1], a comprehensive collection of fictional English texts written between 1880 and 2000. The corpus aims to represent the diversity and development of American literature throughout the period from 1880 to 2000. The corpus contains fictional works from various authors, genres, and literary movements. The version of the corpus that we used in our experiments consisted of 9089 texts. We removed 35 texts, whose length was less than 10K tokens, as our methods

[1] https://textual-optics-lab.uchicago.edu/us_novel_corpus

require a certain length and the shorter texts would have compromised the validity of our results. The remaining texts were analysed as described in the following.

2.2 Series of structural text properties

We expect to find long-range correlations in the frequency distributions of local structural properties. As in previous studies (Mohseni et al., 2022, 2023), the first structural property that we used was the length of sentences, measured as the number of tokens. Moreover, we split text into segments of 25 tokens and determined frequency distributions of particular part-of-speech tags per segment. We used fixed-length segments rather than, for instance, sentences as windows because there are correlations between the distributions of POS-tags and sentence length measurements. We focused on the major parts of speech 'Noun', 'Verb', 'Adjective', 'Adverb', 'Pronoun' and 'Preposition'. Each text is thus represented by seven series of structural properties: sentence length and frequencies of six parts of speech.

3 Methods

In this section we describe the method of Multi-Fractal Detrended Fluctuation Analysis (MFDFA) and Approximate Entropy (ApEn), which we used to analyze text property series.

3.1 MFDFA

For a temporal or spatial series, the auto-correlation $C(\tau)$ is defined as the correlation of the series with a version of the series lagging by τ periods. For uncorrelated series, $C(\tau)$ is zero for $\tau > 0$. For short-range correlations, $C(\tau)$ declines exponentially, i.e. $C(\tau) \sim \exp^{-\tau/\tau_d}$, in which τ_d is the decorrelation time. Conversely, for long-range correlations $C(\tau)$ is described by a power law function, i.e. $C(\tau) \sim \tau^{-\gamma}, 0 < \gamma < 1$. Long-range correlations introduce a memory to series, which means that each element is influenced by characteristics of preceding elements. Describing correlation properties of series through direct calculations of $C(\tau)$ is not recommended because of noisy sampling, the presence of trends in series, and fluctuations of auto-correlation values around zero for large values of τ (Bashan et al., 2008; Koscielny-Bunde et al., 2006). Therefore, other methods such as MFDFA (Kantelhardt et al., 2002) have been proposed to analyze long-range correlations.

MFDFA is a well-established and numerically stable method for studying long-range correlations in temporal or spatial series. According to Kantelhardt et al. (2002) (see also our previous publication Mohseni et al. 2021), MFDFA measures long-range correlations in the following steps:

1. Given a series $X = x_1, x_2, \cdots, x_N$, the profile of the series, which is the cumulative sum centred around zero, is calculated:

$$Y(i) = \sum_{k=1}^{i} [x_k - \langle x \rangle], i = 1, \cdots, N$$

   The step converts the series into a random walk. Y_i can be seen as the position of a random walker after i steps.

2. The profile is divided into $N_s = N/s$ non-overlapping segments for different lengths of s. As the length of the series, N, is often not divisible by the chosen segment size, s, and a part of the series is ignored, the segmentation is repeated, starting from the end. Therefore, $2N_s$ segments are obtained.

3. In each segment v, the profile is detrended by subtracting the best fitting line, Y', and the mean square fluctuation is computed by:

$$F^2(s, v) = \frac{1}{s} \sum_{i=1}^{s} [Y(s \times (v-1) + i) - Y'(s \times (v-1) + i)]^2$$

   for $v = 1, \cdots, N_s$, and similarly for $v = N_s + 1, \cdots, 2N_s$:

$$F^2(s, v) = \frac{1}{s} \sum_{i=1}^{s} [Y(N - s \times (v - N_s) + i) - Y'(N - s \times (v - N_s) + i)]^2$$

Detrending removes the effects of large-scale trends, which are not of interest to the analysis, from the profile of the series. For example, the recurrence of narrative passages (with longer sentences) and dialogues (with shorter sentences) in a novel leads to increasing and decreasing trends, respectively. Detrending removes such trends and facilitates a more accurate evaluation of long-range correlations among the segments of a series.

4. The qth order of fluctuation function is estimated by:

$$F_q(s) = \{\frac{1}{2N_s}\sum_{v=1}^{2N_s}[F^2(s,v)]^{q/2}\}^{1/q}$$

This step of the procedure amounts to averaging over fluctuations along the series to reach a reliable scaling measure.

5. The exponent of the power law function $F(s) \sim s^{h(q)}$, i.e. $h(q)$, is calculated.

$h(q)$ is called the 'extended Hurst exponent'. If a series is stationary, $h(2)$ is identical to the Hurst exponent, a term initially coined in fractal analysis studies. We simply refer to $h(2)$ as the "Hurst Exponent" and represent it with H. For an uncorrelated random series, $H = 0.5$. For a series with positive long-range correlations, $H > 0.5$. The larger the Hurst exponent is, the stronger the correlation patterns of the series. If $H < 0.5$, the series is called 'anti-persistent', indicating the presence of negative long-range correlations.

The parameter q of the extended Hurst exponent allows us to emphasize smaller or larger fluctuations. If a single paradigm dominates the correlation patterns of the series, the procedure is independent of q. From $h(q)$, the singularity spectrum is computed using a Legendre transformation: $f(\alpha) = q[\alpha - h(q)] + 1$, in which $\alpha = h(q) + qh'(q)$. $f(\alpha)$ condenses information about changes of $h(q)$ values for different qs into a well-interpretable form. The width of the singularity spectrum, which we refer to as D, reflects the 'degree of multifractality' or the complexity of the fractal structure (for more details, see Kantelhardt et al., 2002).

In our experiments, we set the parameters of MFDFA as follows: we select the segment size from the sequence $16, 24, 32, 48, 64, \cdots, M$, where $M \leq \lfloor N/3 \rfloor$, i.e $s_0 = 16$ and $s_i = s_{i-1} + 2^{\lfloor \log(s_{i-1})-1 \rfloor}$. The parameter of the fluctuation function, q, is changed from -5 to 5 with a step size of 0.25 (Mohseni et al., 2021).

3.2 Approximate Entropy

Approximate Entropy (ApEn) was proposed to measure the degree of uncertainty in time series (Pincus, 1991). ApEn has been thoroughly explained in previous work, including our studies (Mohseni et al., 2022, 2023). ApEn is a measure of (ir)regularity or (un)predictability in the sequential organization of series. Its computation is based on dissimilarities of sub-sequences with length m and sub-sequences with length $m + 1$. Sub-sequences of length m are more likely to be similar to each other than sub-sequences of length $m + 1$. Therefore, the more fluctuations a series exhibits, the higher its degree of unpredictability will be. m is a parameter set by the user. As series inherently exhibit variation – except for series fixed at a constant value – ApEn allows for a certain degree of fluctuation. If the difference between two sub-sequences falls within a predefined tolerated range, the two sub-sequences are assumed to be similar.

Given a series $X = x(1), \cdots, x(n)$, the sub-sequence length m, and a tolerance r, ApEn is computed as follows:

1. Compute the Chebyshev distance between sub-sequences of length m, $y_i^m = [x(i), \cdots, x(i + (m - 1))]$:
   $d_{i,j}^m = \max_k |y_i^m(k) - y_j^m(k)|$
2. Count the number of matches using the distance between y_i^m and all sub-sequences:
   $C_i^m(r) = \frac{1}{n-m+1}\sum_{j=1}^{n-m+1} \mathbb{1}(r - d_{i,j}^m)$
   $\mathbb{1}(.)$ is a function that returns zero for negative inputs and one for positive inputs.

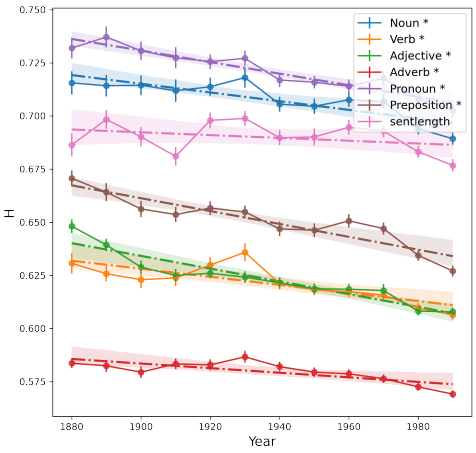

Figure 1: Mean Hurst exponent (H) of texts in each decade from 1880 to 2000. The whiskers show the variance of values for each decade. Wherever the slope of the fitted line is significantly different from zero (p-value<0.05), it is marked by * in the legend.

3. Compute

   $\phi^m(r) = \frac{1}{n-m+1} \sum_{i=1}^{n-m+1} \log(C_i^m(r))$

4. Set m with value $m+1$ and repeat step 1 to 3 to compute $\phi^{m+1}(r)$.

5. ApEn is computed as

   $\texttt{ApEn}(m, r) = \phi^m(r) - \phi^{m+1}(r)$

Similarly to previous studies (Li et al., 2008; Hayashi et al., 2012; Lee et al., 2013; Mohseni et al., 2022, 2023), we set the parameters of ApEn, m and r, to 2 and 20% of the standard deviation of the input series, respectively.

4 Results

4.1 Long-range correlations

To analyze changes in long-range correlations, we arranged the texts of the corpus into decades according to their year of publication and compared the mean Hurst exponent and the mean degree of multifractality. Figure 1 shows the mean Hurst exponents, H, of the texts for each decade (1880 to 2000). We fitted a line to the values of each text property. Although the Hurst exponent increases in some decades, the plot shows a decreasing trend for all POS-tags. Only for sentence length does the slope of the fitted line show any significant difference from zero (flat line).

Figure 2 shows the mean degree of multifractality, D. To make the plot more readable, we represent sentence length separately, as it has a different scale. For all text properties, except Adjective, we observe a decreasing trend. However, Noun and, to some extent, sentence length exhibit two different paradigms for the periods before and after 1930. In Figure 1, a notable irregularity is also observable for texts from the 1930's, in which all properties exhibit a discernible spike.

Figure 1 and Figure 2 show that the Hurst exponent, H, and the degree of multifractality, D, decrease during the time period under study for almost all text properties. As H and D represent the strength and the complexity of long-range correlations, their decreasing trends suggest that during the 20$^{\text{th}}$ century structural properties of texts have relaxed, and contemporary texts tend to be less structured compared with earlier texts.

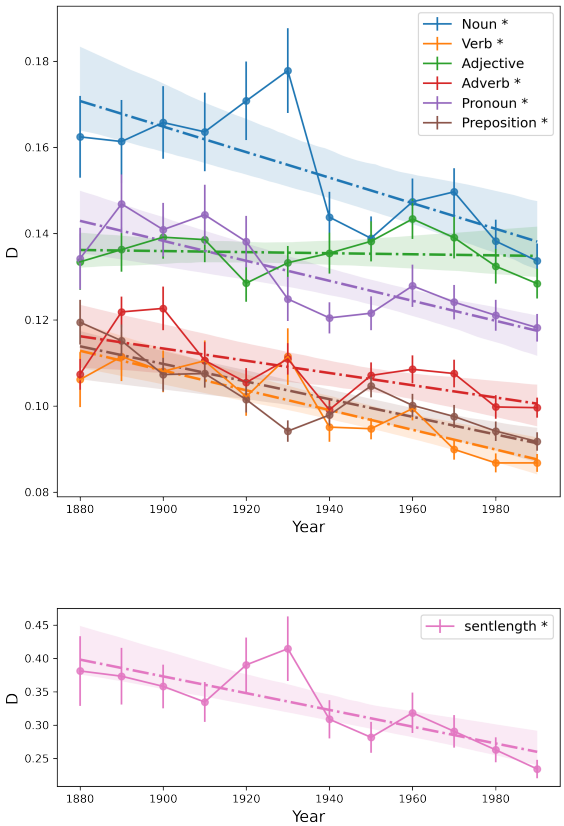

Figure 2: Mean degree of multifractality (D) of texts in each decade from 1880 to 2000. The whiskers show the variance of values for each decade. Wherever the slope of the fitted line is significantly different from zero (p-value$<$0.05), it is marked by * in the legend.

4.2 Predictability

To analyze temporal changes of (ir)regularity/(un)predictability of texts in the US Novel corpus, we similarly grouped texts into decades according to their year of publication and plotted the mean of the ApEn values for all text properties. Figure 3 shows the results. The regression lines demonstrate that the ApEn values for sentence length, Noun, Verb, Adverb and Pronoun increase during the 20th century. Conversely, the slope of the fitted line for Preposition is not significantly different from zero (flat line), and the coefficient for Adjective is negative.

As discussed in Section 3, ApEn is a measure of sequential organization. Thus, as long as the sequential organization of texts is concerned, contemporary texts seems to be less predictable – i.e., they seem to exhibit a looser structure, than earlier texts.

5 Discussion and conclusions

In the present study, we used long-range correlation and predictability analysis to investigate the diachronic evolution of structural text properties. Our results show that long-range correlation patterns of texts have changed in a time period of more than one century, and that contemporary (late 20th century) texts exhibit weaker long-range correlations than earlier (late 19th and early 20th century) texts. The complexity of long-range correlations has also decreased in contemporary texts. Both of these observations confirm that contemporary texts have less complex structures than earlier texts.

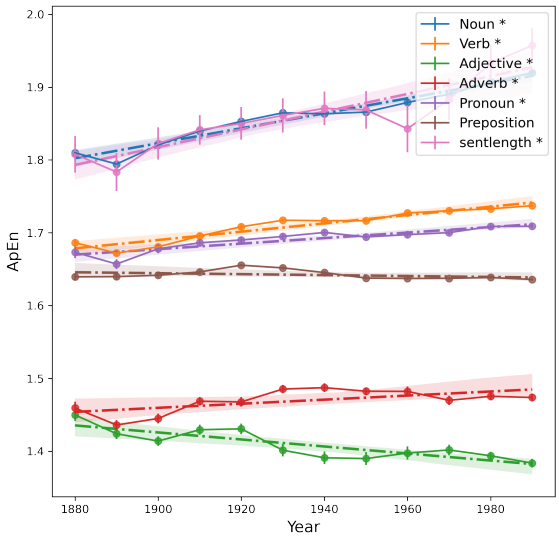

Figure 3: Mean ApEn values of texts in each decade from 1880 to 2000. The whiskers show the variance of values for each decade. Wherever the slope of the fitted line is significantly different from zero (p-value$<$0.05), it is marked by * in the legend.

Although the general pattern of changes in all text properties shows a decrease for both fractal features, an irregularity can be observed for texts from the 1930's. While the lines for ApEn-values are relatively straight, some of the D-values in Figure 2 show a peak at this time. Without being able to provide any reasonable explanation for this peak – without a corresponding development in Approximate Entropy, and only weak anomalies in the H-values – it seems to coincide roughly with the rise and fall of American modernism. Modernism was a literary movement that emerged in the late 19$^{\text{th}}$ century and flourished during the period between World War I and World War II. It is characterized by a notable shift away from traditional literary genres, styles, and themes. Modernist writers aimed to challenge established conventions by experimenting new ways of representing concepts and expressing their ideas (for more discussion see, for example, Childs, 2016). It is not clear to us why American modernism seems to be associated with a peak in the degree of multi-fractality. Multi-fractality points to long-range correlations at different scales. Such correlations suggest a 'layered' design, with fractal structures at various scales of resolution. More detailed studies of individual texts will be required to understand this finding, an endeavour that is beyond the scope of the present study.

Predictability analysis using Approximate Entropy, which is a measure of sequential organization, showed that the irregularity, and therefore, the degree of unpredictability, has increased during the 20$^{\text{th}}$ century for most text properties, suggesting that the later texts are less predictable or structured compared to the earlier texts.

Both types of analyses showed that during 1880 and 2000 structural properties of texts have relaxed and contemporary texts tend to exhibit less structure in comparison to earlier texts. There may be several reasons for these changes. In a previous study (Mohseni et al., 2023), where we analyzed preferred and non-preferred fictional texts in two time periods, we discussed how the impact of technology and the way it changed the process of writing may have influenced the structural design of texts. The invention of the typewriter in the late 19$^{\text{th}}$ century and its subsequent popularity throughout the 20$^{\text{th}}$ century revolutionized the process of writing. While the typewriter lost its prominence after the widespread adoption of personal computers in the 1970s and 1980s, the process of writing became even easier, as writers can edit their

texts without a major effort. Technology may thus have streamlined the writing process, which in turn, potentially alleviated the need for extensive pre-designing and organization.

Moreover, styles and genres have changes during the last century. There are differences between frequency distributions of genres in earlier and contemporary texts. Structural properties may vary across different literary genres and some genres may have less complex structures compared to others, which were more popular in the earlier periods. We leave these questions for future research.

References

Amir Bashan, Ronny Bartsch, Jan W Kantelhardt, and Shlomo Havlin. 2008. Comparison of detrending methods for fluctuation analysis. *Physica A: Statistical Mechanics and its Applications*, 387(21):5080–5090.

Heng Chen and Haitao Liu. 2018. Quantifying evolution of short and long-range correlations in chinese narrative texts across 2000 years. *Complex.*, 2018:9362468.

Peter Childs. 2016. *Modernism*. Routledge, Oxfordshire, England, UK.

Alina Maria Ciobanu, Liviu P. Dinu, Octavia-Maria Şulea, Anca Dinu, and Vlad Niculae. 2013. Temporal text classification for Romanian novels set in the past. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 136–140. Association for Computational Linguistics.

Nandan Kumar Das, Rajib Dey, Semanti Chakraborty, PK Panigrahi, and Nirmalya Ghosh. 2016. Probing multifractality in depth-resolved refractive index fluctuations in biological tissues using backscattering spectral interferometry. *Journal of Optics*, 18(12):125301.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.

StanisŁaw Drożdż, PaweŁ Oświęcimka, Andrzej Kulig, JarosŁaw Kwapień, Katarzyna Bazarnik, Iwona Grabska-Gradzińska, Jan Rybicki, and Marek Stanuszek. 2016. Quantifying origin and character of long-range correlations in narrative texts. *Information Sciences*, 331:32 – 44.

Julia Efremova, Alejandro Montes García, Jianpeng Zhang, and Toon Calders. 2015. Effects of evolutionary linguistics in text classification. In *International Conference on Statistical Language and Speech Processing*, pages 50–61. Springer.

Ramon Ferrer-i-Cancho, Chris Bentz, and Caio Seguin. 2015. Compression and the origins of zipf's law of abbreviation. *ArXiv*, abs/1504.04884:1–36.

Anne Garcia-Fernandez, Anne-Laure Ligozat, Marco Dinarelli, and Delphine Bernhard. 2011. When was it written? automatically determining publication dates. In *Proceedings of the 18th International Conference on String Processing and Information Retrieval*, SPIRE'11, page 221–236, Pisa, Italy. Springer-Verlag.

Helena Gómez-Adorno, Juan-Pablo Posadas-Duran, Germán Ríos-Toledo, Grigori Sidorov, and Gerardo Sierra. 2018. Stylometry-based approach for detecting writing style changes in literary texts. *Computación y Sistemas*, 22(1):47–53.

Amitha Gopidi and Aniket Alam. 2019. Computational analysis of the historical changes in poetry and prose. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 14–22, Florence, Italy. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Kazuko Hayashi, Kenji Shigemi, and Teiji Sawa. 2012. Neonatal electroencephalography shows low sensitivity to anesthesia. *Neuroscience Letters*, 517(2):87 – 91.

Michael Hoey. 2004. *Lexical Priming. A New Theory of Words and Language*. Routledge, London.

Jan W. Kantelhardt, Stephan A. Zschiegner, Eva Koscielny-Bunde, Shlomo Havlin, Armin Bunde, and H.Eugene Stanley. 2002. Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A: Statistical Mechanics and its Applications*, 316(1):87 – 114.

Jasmeen Kanwal, Kenny Smith, Jennifer Culbertson, and Simon Kirby. 2017. Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, 165:45–52.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

Alexander Koplenig, Peter Meyer, Sascha Wolfer, and Carolin Müller-Spitzer. 2017. The statistical trade-off between word order and word structure – large-scale evidence for the principle of least effort. *PLoS ONE*, 12(3):1–25.

Eva Koscielny-Bunde, Jan W Kantelhardt, Peter Braun, Armin Bunde, and Shlomo Havlin. 2006. Long-term persistence and multifractality of river runoff records: Detrended fluctuation studies. *Journal of Hydrology*, 322(1-4):120–137.

Abhimanu Kumar, Matthew Lease, and Jason Baldridge. 2011. Supervised language modeling for temporal resolution of texts. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, page 2069–2072, New York, NY, USA. Association for Computing Machinery.

Ksenia Lagutina and Alla Manakhova. 2021. Automated search and analysis of the stylometric features that describe the style of the prose of 19th–21st centuries. *Automatic Control and Computer Sciences*, 55(7):866–876.

Gerick Lee, Sara Fattinger, Anne-Laure Mouthon, Quentin Noirhomme, and Reto Huber. 2013. Electroencephalogram approximate entropy influenced by both age and sleep. *Frontiers in Neuroinformatics*, 7:33.

Xiaoli Li, Suyuan Cui, and Logan Voss. 2008. Using permutation entropy to measure the electroencephalographic effects of sevoflurane. *Anesthesiology*, 109:448–56.

Chaya Liebeskind and Shmuel Liebeskind. 2020. Deep Learning for Period Classification of Historical Hebrew Texts. *Journal of Data Mining & Digital Humanities*, 2020.

Kyle Mahowald, Evelina Fedorenko, Steven T Piantadosi, and Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126:313–318.

Mahdi Mohseni, Volker Gast, and Christoph Redies. 2021. Fractality and variability in canonical and non-canonical english fiction and in non-fictional texts. *Frontiers in Psychology*, 12:920.

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

Mahdi Mohseni, Christoph Redies, and Volker Gast. 2023. Comparative analysis of preference in contemporary and earlier texts using entropy measures. *Entropy*, 25(3).

Marcelo A. Montemurro and Damián H. Zanette. 2016. Complexity and universality in the long-range order of words. In Mirko Degli Esposti, Eduardo G. Altmann, and François Pachet, editors, *Creativity and Universality in Language*, pages 27–41. Springer, Cham.

Marcelo A. Montemurro and Damián H. Zanette. 2011. Universal entropy of word ordering across linguistic families. *PLoS ONE*, 6(5):1–9.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Steven M Pincus. 1991. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6):2297–2301.

Octavian Popescu and Carlo Strapparava. 2015. SemEval 2015, task 7: Diachronic text evaluation. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 870–878, Denver, Colorado. Association for Computational Linguistics.

Tina C. Roeske, Damian Kelty-Stephen, and Sebastian Wallot. 2018. Multifractal analysis reveals music-like dynamic structure in songbird rhythms. *Scientific Reports*, 8(1).

Shankha Sanyal, Archi Banerjee, Anirban Patranabis, Kaushik Banerjee, Ranjan Sengupta, and Dipak Ghosh. 2016. A study on improvisation in a musical performance using multifractal detrended cross correlation analysis. *Physica A: Statistical Mechanics and Its Applications*, 462:67 – 83.

Terrence Szymanski. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 448–453, Vancouver, Canada. Association for Computational Linguistics.

Denner S Vieira, Sergio Picoli, and Renio S Mendes. 2018. Robustness of sentence length measures in written texts. *Physica A: Statistical Mechanics and Its Applications*, 506:749–754.

Sanja Štajner and Marcos Zampieri. 2013. Stylistic changes for temporal text classification. In *Lecture Notes in Computer Science*, volume 8082, pages 519–526. Springer, Berlin, Heidelberg.

Sida Wang and Christopher D. Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, page 90–94, Jeju Island, Korea. Association for Computational Linguistics.

Tianguang Yang, Changgui Gu, and Huijie Yang. 2016. Long-range correlations in sentence series from a story of the stone. *PLoS ONE*, 11(9):1–11.

Chapter 7

DISCUSSION AND CONCLUSIONS

The central hypothesis of the present study is that preference has textual correlates in structural design features of text. To ensure that our research was based on a strong foundation and to have a more focused and purposeful exploration, we opted to anchor our study to the well-established field of visual aesthetics as a primary reference point. Despite obvious differences between the textual and the visual domains, our approach is supported by the theory of domain-general cognitive processes (Diessel, 2019).

We used different text properties and a group of metrics derived from various analysis methods to investigate preference in text. In the subsequent sections, we discuss our key findings and their implications for future research. Although "failed experiments" are an inherent part of scientific research and may provide valuable insights to fellow researchers, they are usually the missing part of scientific reports. In contrast to this common practice, we will also mention experiments in our study that yielded inconclusive results, regarding our research objectives.

7.1 Textual Correlates of Preference

The main finding of our study is that preference in various text categories has textual correlates in the structural organization of text and that these correlates can be formally defined and objectively measured. We showed that if we represent a text using different text properties and measure global structural features of the text, we can differentiate between preferred and non-preferred fictional text categories. Global structural properties can also distinguish fictional from non-fictional prose texts.

7.2 Text Properties

According to the two levels of language processing, linguistic encoding and comprehension, we distinguished between lower-level and higher-level text properties (cf. Chapter 3). We used two types of lower-level properties, sentence length and frequencies of part-of-speech (POS) tags, and two types of higher-level properties, lexical diversity and topic distribution, in our first experiments (Chapter 3). Our analysis showed that lower-level text properties can characterize various categories

of text more effectively than higher-level properties. While the former group of properties can be formalized more easily, independent of genres, registers and other factors, properties in the latter group can be susceptible to the influence of external factors, such as granularity size and accuracy of the reference model, which is used to extract features, as in the case of topic distributions. Nevertheless, we cannot ignore the potential of text properties at the comprehension level. In our analysis of variability, we showed that variance of the two higher-level text properties are significantly different for canonical/fictional, non-canonical/fictional and non-fictional texts.

Lower-level text properties, particularly frequencies of POS-tags, proved to be successful in distinguishing preferred from non-preferred text categories in other experiments as well (see Chapters 4 and 5). Our hypothesis is that variability and unpredictability analyses of POS-tag frequencies reflect textual preference in fiction because of their associations with distributions of discourse modes. For example, narration and dialogue tend to have a higher frequency of verbs, while description relies more on the use of adjectives and nouns. In dialogue, nouns are notably less common, with a greater emphasis on pronouns. To validate this hypothesis, further investigation using an annotated corpus, which allows a comparative analysis of the multinomial distribution of POS-tags and the distribution of discourse modes, is required.

Another direction for future studies is to investigate other text properties, especially at the higher-level, which represent topical and sentimental information and can potentially reflect aesthetic values of text. In recent studies, long-range correlations of sentiment arcs were used to model Nobel-prize winners and to successfully distinguish them from a control group of non-winners (see, Bizzoni et al., 2022).

In addition to the text properties mentioned above, we conducted experiments using other properties. We used classic readability measures, including the Flesch–Kincaid Grade Level (Kincaid et al., 1975), the Flesch Reading Ease score (Kincaid et al., 1975), the Gunning's fog index (FOG; Robert, 1952), the Coleman–Liau index (Coleman and Liau, 1975), The Dale–Chall readability formula (Dale and Chall, 1948), the automated readability index (ARI; Senter and Smith, 1967), the Linsear Write metric (Klare, 1974) and the Spache readability formula (Spache, 1953). We also used some metrics which reflect reading difficulty at the lexical level, such as word length and the number of syllables per word. Another group of text properties that we investigated were grammatical complexity metrics, such as

left-embeddedness (McNamara et al., 2014), mean dependency length and mean dependency depth. Additionally, we analyzed referring expressions, with a specific focus on pronouns (“I”, “we”, “she”, “he” and “you”) as prevailing surrogates for noun phrases in fictional prose. We also applied a neural-based sentiment analysis model (Tian et al., 2020) to texts in our corpus to extract the sentiment score of sentences. None of these text properties showed a significant discriminatory power compared to frequencies of POS-tags.

We used the JEFP corpus and compared the median values of variance, the three fractal features and two entropy metrics for canonical (N = 76) vs. non-canonical (N = 130) texts and fictional (including both canonical and non-canonical; N = 206) vs. non-fictional (N = 185) texts in Table 7.1 and Table 7.2, respectively. We analyzed values for readability, syntactic complexity, word-level features and sentiment. Referring expressions were not included in the analysis due low frequencies of single pronouns that leads to no significant results. In Table 7.1 the upper row in each cell shows values for canonical texts and the lower row demonstrates values for non-canonical texts. Similarly, in Table 7.2, the upper and lower rows in each cell show values for fictional and non-fictional texts, respectively. As some data are not distributed normally, we used the non-parametric Mann-Whitney U test. The asterisks indicate whether the differences between the two text categories of a given task are statistically significant (ns, not significant; *, $p \leq 0.05$; **, $p \leq 0.01$; and ***, $p \leq 0.001$). 95% confidence intervals for the median (according to Zar, 2010) are shown in parentheses. Although these features did not outperform frequencies of POS-tags in our experiments, both tables demonstrate significant differences in the distribution of some features across the text categories. Therefore, our preliminary results do not negate the potential usefulness of these and similar text properties. More substantial results may be achieved using other models and specific-purposed analysis methods.

Table 7.1: Median values of features for canonical and non-canonical texts. Look at the text for more information.

Property	Variance	Fractal Features			Entropy Metrics	
		Degree of Fractality	Degree of Multifractality	Asymmetry	ApEn	ShEn
ARI	35.9 (31.1, 41.4)***	0.68 (0.66, 0.72)^{ns}	0.31 (0.28, 0.35)^{ns}	0.10 (0.03, 0.20)^{ns}	1.75 (1.72, 1.79)***	7.37 (7.24, 7.47)***
	24.3 (20.9, 27.3)	0.69 (0.68, 0.71)	0.30 (0.27, 0.34)	0.05 (-0.00, 0.12)	1.61 (1.58, 1.64)	6.73 (6.67, 6.83)
Coleman–Liau	6.98 (6.55, 7.47)***	0.73 (0.70, 0.76)^{ns}	0.24 (0.20, 0.27)^{ns}	0.15 (0.02, 0.25)^{ns}	1.83 (1.77, 1.86)***	7.38 (7.23, 7.46)***
	6.17 (5.67, 6.43)	0.72 (0.70, 0.73)	0.27 (0.22, 0.30)	0.07 (-0.01, 0.14)	1.61 (1.59, 1.64)	6.73 (6.67, 6.83)
Dale–Chall	1.12 (1.06, 1.21)^{ns}	0.71 (0.67, 0.74)^{ns}	0.21 (0.18, 0.24)^{ns}	-0.12 (-0.23, 0.03)^{ns}	1.85 (1.79, 1.87)***	6.75 (6.67, 6.83)***
	1.10 (1.04, 1.16)	0.70 (0.69, 0.73)	0.18 (0.16, 0.21)	-0.14 (-0.27, -0.03)	1.64 (1.60, 1.67)	6.33 (6.28, 6.39)
Flesch	249 (237, 260)***	0.71 (0.68, 0.74)^{ns}	0.27 (0.21, 0.28)^{ns}	0.09 (-0.04, 0.12)^{ns}	1.80 (1.77, 1.83)***	7.03 (6.93, 7.13)***
	186 (167, 215)	0.71 (0.70, 0.72)	0.25 (0.22, 0.28)	0.06 (0.01, 0.11)	1.63 (1.58, 1.64)	6.54 (6.48, 6.61)
Flesch–Kincaid	19.7 (17.1, 23.4)***	0.68 (0.66, 0.72)^{ns}	0.32 (0.27, 0.36)^{ns}	0.11 (0.03, 0.19)^{ns}	1.75 (1.72, 1.78)***	7.03 (6.92, 7.12)***
	13.0 (11.3, 15.4)	0.694 (0.685, 0.710)	0.29 (0.25, 0.32)	0.04 (-0.03, 0.08)	1.60 (1.58, 1.64)	6.53 (6.48, 6.61)
FOG	22.5 (19.2, 27.1)***	0.69 (0.67, 0.72)^{ns}	0.32 (0.27, 0.35)^{ns}	0.09 (-0.03, 0.15)^{ns}	1.75 (1.72, 1.79)***	6.55 (6.45, 6.62)***
	15.5 (13.9, 17.7)	0.70 (0.69, 0.72)	0.28 (0.25, 0.31)	-0.01 (-0.09, 0.07)	1.61 (1.58, 1.64)	6.16 (6.10, 6.20)
Linsear Write	44.3 (37.3, 53.2)***	0.67 (0.65, 0.71)^{ns}	0.34 (0.30, 0.39)^{ns}	0.06 (0.00, 0.14)^{ns}	1.67 (1.63, 1.71)***	5.35 (5.31, 5.40)***
	28.5 (25.2, 33.7)	0.68 (0.67, 0.70)	0.33 (0.28, 0.35)	0.01 (-0.06, 0.09)	1.59 (1.56, 1.62)	5.16 (5.11, 5.19)
spache	2.36 (1.78, 2.73)***	0.69 (0.66, 0.71)^{ns}	0.32 (0.28, 0.36)^{ns}	0.00 (-0.08, 0.11)^{ns}	1.74 (1.69, 1.77)***	6.75 (6.65, 6.82)***
	1.49 (1.19, 1.72)	0.69 (0.68, 0.71)	0.31 (0.29, 0.35)	-0.05 (-0.11, 0.01)	1.60 (1.58, 1.63)	6.34 (6.29, 6.38)
Left-Embeddedness	29.9 (26.2, 34.5)^{ns}	0.595 (0.579, 0.604)^{ns}	0.32 (0.30, 0.35)^{ns}	0.02 (-0.10, 0.10)^{ns}	1.63 (1.56, 1.74)^{ns}	2.58 (2.50, 2.65)^{ns}
	29.4 (26.8, 31.3)	0.59 (0.58, 0.60)	0.31 (0.28, 0.34)	0.00 (-0.04, 0.05)	1.58 (1.57, 1.65)	2.63 (2.59, 2.67)
Dependency Length	1.04 (0.96, 1.13)*	0.64 (0.63, 0.66)***	0.10 (0.09, 0.13)^{ns}	-0.13 (-0.21, 0.02)^{ns}	2.10 (2.07, 2.12)***	6.19 (5.99, 6.34)^{ns}
	0.94 (0.90, 1.01)	0.62 (0.61, 0.63)	0.11 (0.10, 0.14)	-0.07 (-0.13, 0.01)	2.02 (2.00, 2.05)	6.21 (6.11, 6.35)
Dependency Depth	0.84 (0.78, 0.90)*	0.69 (0.68, 0.71)^{ns}	0.23 (0.20, 0.27)***	0.01 (-0.05, 0.07)^{ns}	1.95 (1.92, 1.98)^{ns}	5.91 (5.60, 6.04)^{ns}
	0.76 (0.70, 0.81)	0.68 (0.66, 0.70)	0.18 (0.16, 0.20)	-0.04 (-0.07, 0.03)	1.94 (1.92, 1.96)	5.96 (5.86, 6.11)
Word Length	5.75 (5.65, 5.92)^{ns}	0.71 (0.70, 0.72)*	0.03 (0.02, 0.04)^{ns}	-0.27 (-0.41, -0.17)^{ns}	2.04 (2.03, 2.06)^{ns}	2.10 (2.09, 2.11)^{ns}
	5.72 (5.48, 5.96)	0.698 (0.687, 0.703)	0.033 (0.028, 0.038)	-0.28 (-0.37, -0.18)	2.03 (2.02, 2.05)	2.10 (2.09, 2.12)
Word Syllables	0.70 (0.69, 0.73)^{ns}	0.70 (0.69, 0.71)*	0.035 (0.031, 0.043)^{ns}	-0.13 (-0.28, -0.02)^{ns}	1.09 (1.08, 1.11)^{ns}	1.12 (1.11, 1.13)^{ns}
	0.69 (0.68, 0.72)	0.696 (0.689, 0.700)	0.039 (0.033, 0.042)	0.01 (-0.08, 0.11)	1.09 (1.08, 1.11)	1.12 (1.11, 1.13)
Sentiment	0.59 (0.58, 0.60)**	0.61 (0.60, 0.63)^{ns}	0.10 (0.06, 0.12)^{ns}	0.45 (0.16, 0.59)^{ns}	1.24 (1.23, 1.25)***	5.24 (5.16, 5.31)***
	0.58 (0.57, 0.59)	0.61 (0.60, 0.63)	0.09 (0.07, 0.10)	0.26 (0.09, 0.42)	1.25 (1.24, 1.26)	5.33 (5.30, 5.36)

Table 7.2: Median values of features for fictional and non-fictional texts. Look at the text for more information.

Property	Variance	Fractal Features			Entropy Metrics	
		Degree of Fractality	Degree of Multifractality	Asymmetry	ApEn	ShEn
ARI	29.2 (25.1, 32.4)^{ns}	0.69 (0.68, 0.71)^{ns}	0.30 (0.28, 0.33)^{ns}	0.07 (0.03, 0.12)***	1.65 (1.62, 1.69)^{ns}	6.93 (6.82, 7.07)^{ns}
	27.7 (25.7, 32.3)	0.69 (0.67, 0.71)	0.28 (0.26, 0.31)	-0.15 (-0.21, -0.07)	1.66 (1.65, 1.68)	7.00 (6.95, 7.11)
Coleman–Liau	6.45 (6.24, 6.67)***	0.72 (0.70, 0.73)*	0.25 (0.23, 0.28)^{ns}	0.10 (0.02, 0.15)***	1.70 (1.64, 1.73)^{ns}	6.92 (6.81, 7.07)^{ns}
	4.83 (4.48, 5.29)	0.74 (0.72, 0.76)	0.24 (0.21, 0.26)	-0.22 (-0.34, -0.10)	1.70 (1.68, 1.73)	7.00 (6.94, 7.10)
Dale–Chall	1.11 (1.06, 1.15)***	0.71 (0.69, 0.72)**	0.19 (0.18, 0.21)***	-0.13 (-0.20, -0.04)***	1.71 (1.66, 1.74)**	6.43 (6.38, 6.54)^{ns}
	1.72 (1.54, 1.93)	0.74 (0.71, 0.75)	0.33 (0.30, 0.36)	-0.41 (-0.49, -0.29)	1.66 (1.64, 1.69)	6.48 (6.44, 6.53)
Flesch	217 (203, 237)*	0.71 (0.70, 0.72)*	0.25 (0.22, 0.27)*	0.06 (0.02, 0.10)***	1.67 (1.64, 1.73)^{ns}	6.67 (6.58, 6.78)*
	218 (207, 244)	0.73 (0.72, 0.75)	0.22 (0.19, 0.24)	-0.15 (-0.24, -0.10)	1.71 (1.68, 1.73)	6.79 (6.71, 6.84)
Flesch–Kincaid	15.8 (13.5, 17.3)*	0.69 (0.68, 0.71)^{ns}	0.30 (0.27, 0.32)^{ns}	0.06 (0.02, 0.10)***	1.65 (1.62, 1.69)^{ns}	6.66 (6.58, 6.78)*
	16.5 (15.3, 18.5)	0.70 (0.69, 0.72)	0.28 (0.23, 0.30)	-0.16 (-0.23, -0.06)	1.67 (1.66, 1.70)	6.79 (6.70, 6.84)
FOG	18.3 (15.8, 20.5)**	0.70 (0.69, 0.71)^{ns}	0.29 (0.27, 0.32)^{ns}	0.03 (-0.03, 0.08)**	1.66 (1.62, 1.69)*	6.26 (6.20, 6.36)**
	19.7 (18.9, 22.3)	0.70 (0.69, 0.72)	0.25 (0.22, 0.28)	-0.14 (-0.20, -0.05)	1.69 (1.67, 1.71)	6.37 (6.33, 6.44)
Linsear Write	34.9 (29.7, 39.8)***	0.68 (0.67, 0.69)^{ns}	0.33 (0.30, 0.35)^{ns}	0.04 (-0.02, 0.09)**	1.62 (1.59, 1.64)*	5.21 (5.17, 5.25)^{ns}
	45.5 (42.3, 52.9)	0.68 (0.67, 0.71)	0.33 (0.29, 0.37)	-0.11 (-0.19, -0.06)	1.65 (1.63, 1.67)	5.21 (5.17, 5.26)
spache	1.74 (1.57, 1.95)***	0.69 (0.68, 0.71)^{ns}	0.31 (0.29, 0.34)^{ns}	-0.03 (-0.07, 0.02)***	1.65 (1.61, 1.68)^{ns}	6.44 (6.38, 6.53)^{ns}
	2.20 (1.99, 2.39)	0.68 (0.66, 0.69)	0.31 (0.28, 0.36)	-0.23 (-0.33, -0.13)	1.66 (1.64, 1.69)	6.47 (6.43, 6.51)
Left-Embeddedness	29.6 (27.8, 31.5)***	0.592 (0.586, 0.600)***	0.31 (0.30, 0.33)***	0.01 (-0.04, 0.05)***	1.60 (1.57, 1.66)*	2.61 (2.58, 2.65)***
	46.1 (44.0, 49.3)	0.64 (0.61, 0.66)	0.44 (0.39, 0.47)	0.53 (0.47, 0.59)	1.59 (1.57, 1.62)	2.79 (2.75, 2.82)
Dependency Length	0.99 (0.93, 1.03)***	0.63 (0.62, 0.64)***	0.11 (0.10, 0.13)***	-0.08 (-0.14, -0.02)***	2.05 (2.03, 2.07)***	6.20 (6.12, 6.30)**
	1.18 (1.12, 1.25)	0.68 (0.66, 0.71)	0.15 (0.13, 0.17)	0.13 (0.00, 0.22)	2.02 (2.00, 2.04)	6.40 (6.32, 6.47)
Dependency Depth	0.80 (0.76, 0.83)***	0.68 (0.67, 0.70)^{ns}	0.20 (0.19, 0.21)**	-0.01 (-0.05, 0.03)***	1.95 (1.93, 1.96)**	5.94 (5.86, 6.04)***
	1.00 (0.95, 1.03)	0.69 (0.66, 0.73)	0.24 (0.20, 0.27)	0.35 (0.26, 0.48)	1.97 (1.95, 1.99)	6.20 (6.09, 6.27)
Word Length	5.74 (5.60, 5.86)***	0.700 (0.696, 0.707)***	0.031 (0.028, 0.035)***	-0.28 (-0.35, -0.21)***	2.04 (2.03, 2.05)***	2.10 (2.09, 2.11)***
	7.50 (7.36, 7.76)	0.655 (0.646, 0.663)	0.06 (0.05, 0.07)	0.05 (0.01, 0.17)	2.13 (2.12, 2.14)	2.24 (2.22, 2.24)
Word Syllables	0.70 (0.69, 0.72)***	0.697 (0.693, 0.703)***	0.038 (0.033, 0.042)***	-0.04 (-0.13, 0.04)***	1.09 (1.09, 1.10)***	1.12 (1.11, 1.13)***
	1.00 (0.97, 1.05)	0.68 (0.66, 0.69)	0.054 (0.048, 0.063)	0.30 (0.24, 0.40)	1.23 (1.22, 1.24)	1.27 (1.26, 1.29)
Sentiment	0.584 (0.582, 0.588)***	0.61 (0.60, 0.62)***	0.09 (0.08, 0.10)***	0.32 (0.18, 0.47)***	1.247 (1.243, 1.251)***	5.30 (5.27, 5.33)***
	0.545 (0.540, 0.552)	0.66 (0.65, 0.68)	0.19 (0.18, 0.21)	0.77 (0.71, 0.83)	1.23 (1.22, 1.24)	5.56 (5.53, 5.59)

7.3 Analysis Methods

In our experiments we mostly focused on three analysis methods: fractality (long-range correlations), variability and predictability. We applied Multi-Fractal Detrended Fluctuation Analysis (MFDFA; Kantelhardt et al., 2002) to series of text property values along texts and computed three statistics, the degree of fractality, the degree of multifractality and fractal asymmetry, which were used to statistically describe text categories and classify them from each other. Our results revealed a notable distinction in long-range correlations between fictional and non-fictional texts. Non-fictional texts demonstrate more discernible long-range correlations, which indicate that authors of the non-fictional genre tend to use more consistent structural patterns throughout their writing (cf. Section 2.3). Within the fictional category, the ranges of fractal feature values for preferred (canonical) and non-preferred (non-canonical) texts overlap significantly, which results in poor classification results. Although fractal analysis can distinguish fiction and non-fiction with an acceptable accuracy, it is still the least effective approach in classification of the text categories under study compared to the other analysis methods (see, Chapter 3 / Mohseni et al., 2021). This finding suggests that long-range correlations are universal structural characteristics that can be found in different text categories. However, further analysis is required to validate this assumption.

We operationalized variability as variance of text properties (Chapter 3 / Mohseni et al., 2021). Our experiments demonstrated that canonical fictional texts have higher variability across all properties investigated in our study compared to non-canonical texts. The results regarding the variability of non-fictional texts in comparison to fictional texts are less conclusive, with some properties showing greater variability in fictional texts and others in non-fictional texts.

Predictability analysis was another approach that we applied to model textual preference in various text categories from two different time periods (see, Chapters 4 and 5 / Mohseni et al., 2022; Mohseni et al., 2023). Shannon Entropy is a global measure of irregularity/unpredictability that we used to analyze distributions of different text properties. However, text, or more precisely, reading, is a sequential process, in which information accumulates and at each stage, the reader develops certain expectation about what will follow. By establishing a trade-off between the reader's expectations and surprise, an author engages a reader with the text. We used Approximate Entropy as a measure of surprise in the sequential organization of text. We analyzed unpredictability in text from two different time periods, texts from the

19th and early 20th centuries and contemporary texts. We operationalized preference for the two time periods differently: canonization for earlier texts and sales figure using lists of bestseller books for contemporary texts. Our most important observation was that preferred fictional texts exhibit a higher degree of unpredictability than non-preferred texts in both periods. Nevertheless, unpredictability in local structures (measured by Approximate Entropy) is more pronounced where preference is operationalized based on canonization than where it is operationalized based on sales figures. Our results indicate the high potential of predictability and surprise analysis in analyzing preference in the textual domain.

We also conducted experiments by applying other approaches in our study. Inspired by Brachmann et al. (2017), who analyzed variability in the low-level filters of a convolutional neural network to classify artworks from non-art images, we applied BERT (Devlin et al., 2019), a multi-layered transformer-based language model, to texts in the JEFP corpus and measured variability in each layer of the language model as well as changes in the vector representation of text across layers. Although preliminary results were promising in modeling different categories of texts, the inability of BERT and similar language models to process long texts and the lack of interpretability of neural language models discouraged us to pursue this approach (for a review on interpretation of inner processes of neural language models, look at, for example, Rogers et al., 2020).

Series of feature values can be analyzed not only in the time domain but also in the frequency domain. Using the Fourier transform we analyzed the spectral features of text property series in various categories of texts. The Fourier analysis provided no more information than fractal analysis (for a discussion on the relation between fractality and spectral properties of signals, see Chapter 3 / Mohseni et al., 2021).

Fractal analysis, as shown by using MFDFA (Chapter 3), describes long-range correlation patterns of a signal. However, long-range correlations can also be analyzed based on components of complex signals. We used multiple signal decomposition methods, such as Ensemble Empirical Mode Decomposition (Wu and Huang, 2009), to extract components, which are called Intrinsic Mode Functions (IMFs), of text property series and then we analyzed long-range correlations for each IMF using MFDFA. Our idea was to distinguish noisy-like local fluctuations in the sequence of text property values, which are imposed by language structures and are represented in lower IMFs, from global structures, which are exhibited in higher IMFs. Our analysis showed that results depend substantially on the length of the input signals.

This drawback disqualified them for our tasks.

Our analysis of predictability was not limited to applying Approximate Entropy and Shannon Entropy. We also investigated other extensions of Approximate Entropy, such as Sample Entropy (Richman and Moorman, 2000), Multi-Scale Entropy (Costa et al., 2005) and Multivariate Multi-Scale Entropy (Ahmed and Mandic, 2011). In our experiments, none of these metrics performed significantly better than Approximate Entropy. Regarding the interdependency of POS-tags frequencies, we used Conditional Entropy as an additional metric to analyze unpredictability in underlying structures of text. Results were not substantially different from those of Shannon Entropy.

7.4 Operationalization of Preference

Our study is situated within the field of computational textual aesthetics and our main goal was to investigate structural properties of preferred texts *versus* non-preferred texts. As opposed to experimental studies, our exploratory research demanded a way to operationalize preference. Although we do not deny individual preferential factors, we focused on preference at a community level to analyze global and potentially generalizable structural features in texts. To compile the JEFP corpus (Chapters 3 and 4 / Mohseni et al., 2021; Mohseni et al., 2022), we used canonization as our discrimination metric for distinction of preferred (canonical) from non-preferred (non-canonical) fictional texts. Canonization, which involves different sectors of a society, such as educational, social and political departments, is a long and time-distributed process that cannot be used for evaluation of recently published texts. We therefore used sales figures to operationalize preference in contemporary texts. Sales figures, representative of the success of a book, were used to build the JCEFP corpus (Chapter 5 / Mohseni et al., 2023), which includes a category of bestsellers as preferred contemporary texts and a category of non-bestsellers as non-preferred contemporary texts.

Preference can also be operationalized using other factors, such as feedback and ratings in social media (Maharjan et al., 2017; Maharjan et al., 2018) and literature prizes (Febres and Jaffe, 2017; Bizzoni et al., 2022). In an ongoing project, we are analyzing the correlations of features extracted from the above-mentioned methods with ratings extracted from the website www.goodreads.com (more on this in the next chapter).

We should bear in mind that preference may have been derived from not only textual

factors but also non-textual ones, such as social and political circumstances, and each way of operationalization of preference targets different groups of readers. Furthermore, appreciation of text may involve some time-variant factors. Although we have taken the initial steps to address these questions (Chapter 5 / Mohseni et al., 2023), further studies are undoubtedly necessary.

7.5 Limitations of Our Study

Perhaps the most critical limitation of our study is that we did not assess aesthetic experience directly and rather adopted an exploratory approach to investigate textual preference. However, in the design of future experimental studies, in which aesthetic experience is investigated directly, findings of exploratory studies offer valuable insights about potential correlates of preference.

In operationalizing preference, we limited our experiments to two approaches, i.e. canonization and sales figures. Other methods of modeling preference need to be explored as it has already suggested by other studies (Section 7.4).

Additionally, our study has limitations from multiple perspectives, which are related to each other: language, genre, text length, text representation and methodology. We only analyzed English prose texts from two different time periods, texts from the 19th and early 20th centuries and contemporary texts published after 2000. Moreover, we only investigated long prose texts. Our experiments were also confined to a limited selection of analysis methods. The exploration of other methods is another direction for future studies. Specially, we expect that shorter texts and other text classes require different ways of representation and analysis. Furthermore, we did not profoundly analyze any content-related or higher-level text properties, which may complement lower-lever text properties. No general conclusion can be made unless different types of texts from a wide range of languages and from different cultural backgrounds are investigated.

We emphasized that our research was inspired by relevant studies in the field of visual aesthetics, where structural properties of artworks and non-art images were analyzed. This approach aligns with the theory of domain-general cognitive processes (Diessel, 2019). Nevertheless, this analogy between the two sensory domains, textual and visual, might be established based on a (partially) wrong assumption. The immanent implication of this is that we should explore alternative structural design features that are completely distinct from those found in images. For example, the sequential time-distributed nature of text distinguishes reading from image perception. As a

result, other factors, such as memory and incremental processing of information (Venhuizen et al., 2019), expectation and mental imagery (Magyari et al., 2020), may play an important role in reading experience.

Some may argue that the time-distributed nature of music makes it a more suitable analogy to text than an image. Recognizing the significance of research in the auditory domain, we diligently cited relevant studies in our publications, such as analysis of variability and fractality (Voss and Clarke, 1975; Wu et al., 2015; Levitin et al., 2012; Sanyal et al., 2016) and predictability and surprise (Gold et al., 2019; Miles et al., 2017) in music. In our study we used research in the visual and auditory domains as a reference point. To advance our understanding of human aesthetic perception and appreciation, further studies are required to investigate the interconnection between various sensory domains, including text, image, and music, within a multimodal analysis framework and using comparative analyses across these domains.

References

Ahmed, Mosabber and Danilo Mandic (2011). "Multivariate Multiscale Entropy: A Tool for Complexity Analysis of Multichannel Data". In: *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* 84, p. 061918. DOI: 10.1103/PhysRevE.84.061918.

Bizzoni, Yuri, Kristoffer Nielbo, and Mads Thomsen (2022). "Fractality of sentiment arcs for literary quality assessment: The case of Nobel laureates". In: *The 2nd International Workshop on Natural Language Processing for Digital Humanities – NLP4DH 2022*. Taipei, Taiwan: Association for Computational Linguistics, pp. 31–41.

Brachmann, Anselm, Erhardt Barth, and Christoph Redies (2017). "Using CNN features to better understand what makes visual artworks special". In: *Frontiers in Psychology* 8, p. 830. DOI: 10.3389/fpsyg.2017.00830.

Coleman, Meri and Ta Lin Liau (1975). "A computer readability formula designed for machine scoring." In: *Journal of Applied Psychology* 60, pp. 283–284.

Costa, Madalena, Ary L Goldberger, and C-K Peng (2005). "Multiscale Entropy Analysis of Biological Signals". In: *Physical Review E* 71.2, p. 021906. DOI: 10.1103/physreve.71.021906.

Dale, Edgar and Jeanne S. Chall (1948). "A Formula for Predicting Readability". In: *Educational Research Bulletin* 27.1, pp. 11–28.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

Diessel, Holger (2019). *The Grammar Network. How Linguistic Structure is Shaped by Language Use*. Cambridge: Cambridge University Press.

Febres, Gerardo and Klaus Jaffe (2017). "Quantifying Structure Differences in Literature Using Symbolic Diversity and Entropy Criteria". In: *Journal of Quantitative Linguistics* 24.1, pp. 16–53. DOI: 10.1080/09296174.2016.1169847.

Gold, Benjamin P., Marcus T. Pearce, Ernest Mas-Herrero, Alain Dagher, and Robert J. Zatorre (2019). "Predictability and Uncertainty in the Pleasure of Music: A Reward for Learning?" In: *Journal of Neuroscience* 39.47, pp. 9397–9409. DOI: 10.1523/JNEUROSCI.0428-19.2019.

Kantelhardt, Jan W., Stephan A. Zschiegner, Eva Koscielny-Bunde, Shlomo Havlin, Armin Bunde, and H.Eugene Stanley (2002). "Multifractal detrended fluctuation analysis of nonstationary time series". In: *Physica A: Statistical Mechanics and Its Applications* 316.1, pp. 87–114. DOI: 10.1016/S0378-4371(02)01383-3.

Kincaid, J Peter, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom (1975). *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for navy enlisted personnel*. Tech. rep. Naval Technical Training Command Millington TN Research Branch.

Klare, George R. (1974). "Assessing Readability". In: *Reading Research Quarterly* 10.1, pp. 62–102.

Levitin, Daniel J., Parag Chordia, and Vinod Menon (2012). "Musical rhythm spectra from Bach to Joplin obey a 1/f power law". In: *Proceedings of the National Academy of Sciences* 109.10, pp. 3716–3720. DOI: 10.1073/pnas.1113828109.

Magyari, Lilla, Anne Mangen, Anezka Kuzmicova, Arthur Jacobs, and Jana Lüdtke (2020). "Eye movements and mental imagery during reading of literary texts in different narrative styles". In: *Journal of Eye Movement Research* 3, pp. 1–35. DOI: 10.16910/jemr.13.3.3.

Maharjan, Suraj, John Arevalo, Manuel Montes, Fabio 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*. Valencia, Spain: Association for Computational Linguistics, pp. 1217–1227. DOI: 10.18653/v1/E17-1114.

Maharjan, Suraj, Sudipta Kar, Manuel Montes, Fabio A. González, and Thamar Solorio (2018). "Letting Emotions Flow: Success Prediction by Modeling the Flow of Emotions in Books". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 259–265. DOI: 10.18653/v1/N18-2042.

McNamara, Danielle S., Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511894664.

Miles, Scott A., David S. Rosen, and Norberto M. Grzywacz (2017). "A Statistical Analysis of the Relationship between Harmonic Surprise and Preference in Popular Music". In: *Frontiers in Human Neuroscience* 11, pp. 1–13. DOI: 10.3389/fnhum.2017.00263.

Mohseni, Mahdi, Volker Gast, and Christoph Redies (2021). "Fractality and Variability in Canonical and Non-Canonical English Fiction and in Non-Fictional Texts". In: *Frontiers in Psychology* 12, p. 920. DOI: 10.3389/fpsyg.2021.599063.

Mohseni, Mahdi, Christoph Redies, and Volker Gast (2022). "Approximate Entropy in Canonical and Non-Canonical Fiction". In: *Entropy* 24.2, p. 277. DOI: 10.3390/e24020278.

– (2023). "Comparative Analysis of Preference in Contemporary and Earlier Texts Using Entropy Measures". In: *Entropy* 25.3, p. 486. DOI: 10.3390/e25030486.

Richman, Joshua and Joseph Moorman (2000). "Physiological Time-Series Analysis Using Approximate Entropy and Sample Entropy". In: *American Journal of Physiology. Heart and Circulatory Physiology* 278, H2039–49. DOI: 10.1152/ajpheart.2000.278.6.H2039.

Robert, Gunning (1952). *The technique of clear writing*. New York: McGraw-Hill.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky (2020). "A Primer in BERTology: What We Know About How BERT Works". In: *Transactions of the Association for Computational Linguistics* 8, pp. 842–866. DOI: 10.1162/tacl_a_00349.

Sanyal, Shankha, Archi Banerjee, Anirban Patranabis, Kaushik Banerjee, Ranjan Sengupta, and Dipak Ghosh (2016). "A study on Improvisation in a Musical performance using Multifractal Detrended Cross Correlation Analysis". In: *Physica A: Statistical Mechanics and Its Applications* 462, pp. 67–83. DOI: 10.1016/j.physa.2016.06.013.

Senter, RJ and Edgar A Smith (1967). *Automated readability index*. Tech. rep. Cincinnati Univ OH.

Spache, George (1953). "A New Readability Formula for Primary-Grade Reading Materials". In: *The Elementary School Journal* 53.7, pp. 410–413. DOI: 10.1086/458513. (Visited on 05/25/2023).

Tian, Hao et al. (2020). "SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4067–4076. DOI: 10.18653/v1/2020.acl-main.374.

Venhuizen, Noortje J., Matthew W. Crocker, and H. Brower (2019). "Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience". In: *Discourse Processes* 56.3, pp. 229–255. DOI: 10.1080/0163853X.2018.1448677.

Voss, R.F. and J. Clarke (1975). "'1/f' noise in music and speech". In: *Nature* 258, pp. 317–318. DOI: 10.1038/258317a0.

Wu, D., K.M. Kendrick, D. Levitin, L. Chaoyi, and Yao Dezhong (2015). "Bach Is the Father of Harmony: Revealed by a 1/f Fluctuation Analysis across Musical Genres". In: *PLoS ONE* 10.11, pp. 1–17. DOI: 10.1371/journal.pone.0142431.

Wu, Zhaohua and Norden Huang (2009). "Ensemble empirical mode decomposition: a noise-assisted data analysis method". In: *Advances in Adaptive Data Analysis* 01.01, pp. 1–41. DOI: 10.1142/S1793536909000047.

Zar, Jerrold H. (2010). *Biostatistical Analysis*. 5th ed. Upper Saddle River, NJ: Pearson.

Chapter 8

OUTLOOK: ONGOING RESEARCH

In the preceding chapters, we outlined our methodology, presented the results of our experiments, and discussed the key findings and future directions. On top of these accomplishments, we are actively involved in several ongoing projects that further expand our research perspective. In this chapter, we briefly overview these projects and highlight the objectives of each. We also present initial results concisely. By exploring these ongoing endeavors, we aim to deepen our understanding and contribute to our ever-evolving research field.

8.1 Clustering Canonical Texts

Applying fractality and predictability analysis methods, we extracted a group of features, such as the degree of fractality, the degree of multifractality and Approximate Entropy, for each text in the JEFP corpus. These features were used to distinguish canonical from non-canonical fictional texts as well as fictional from non-fictional texts. We were interested to explore whether these features reveal any patterns inside each text category, particularly in the case of canonical texts, which are written by more prestigious authors and are widely recognized within the community.

In the first step, we clustered canonical prose texts using the Agglomerative Clustering, which is a hierarchical clustering method, and groups object based on their similarity. As an example, Figure 8.1 shows the dendogram representation of the clustering result using Approximate Entropy values (see, Chapter 4/ Mohseni et al., 2022). Each data point corresponds to an author and one of his/her works. The heatmap used to represent years of publication helps to characterize clusters. For example, *The Secrete Agent* by *Joseph Conrad* shows greater similarity to *Moby Dick* by *Herman Melville* up to *The Deerslayers* by *James Fenimore*, written over half a century earlier, than its contemporary texts. Four works of *Henry James* form a cluster with those of *Jane Austen* and *Anne Bronte*, despite the time gap of one century and half a century, respectively. Moreover, the visually discernible patterns suggest that Approximate Entropy, as one of structural features, may exhibit information about the writing style of authors. *Ulysses* by *James Joyce* constitutes its own cluster, thanks to its unique written structure. While four works of *Henry James* are categorized in one group, the *Charles Dickens*' works have spread across

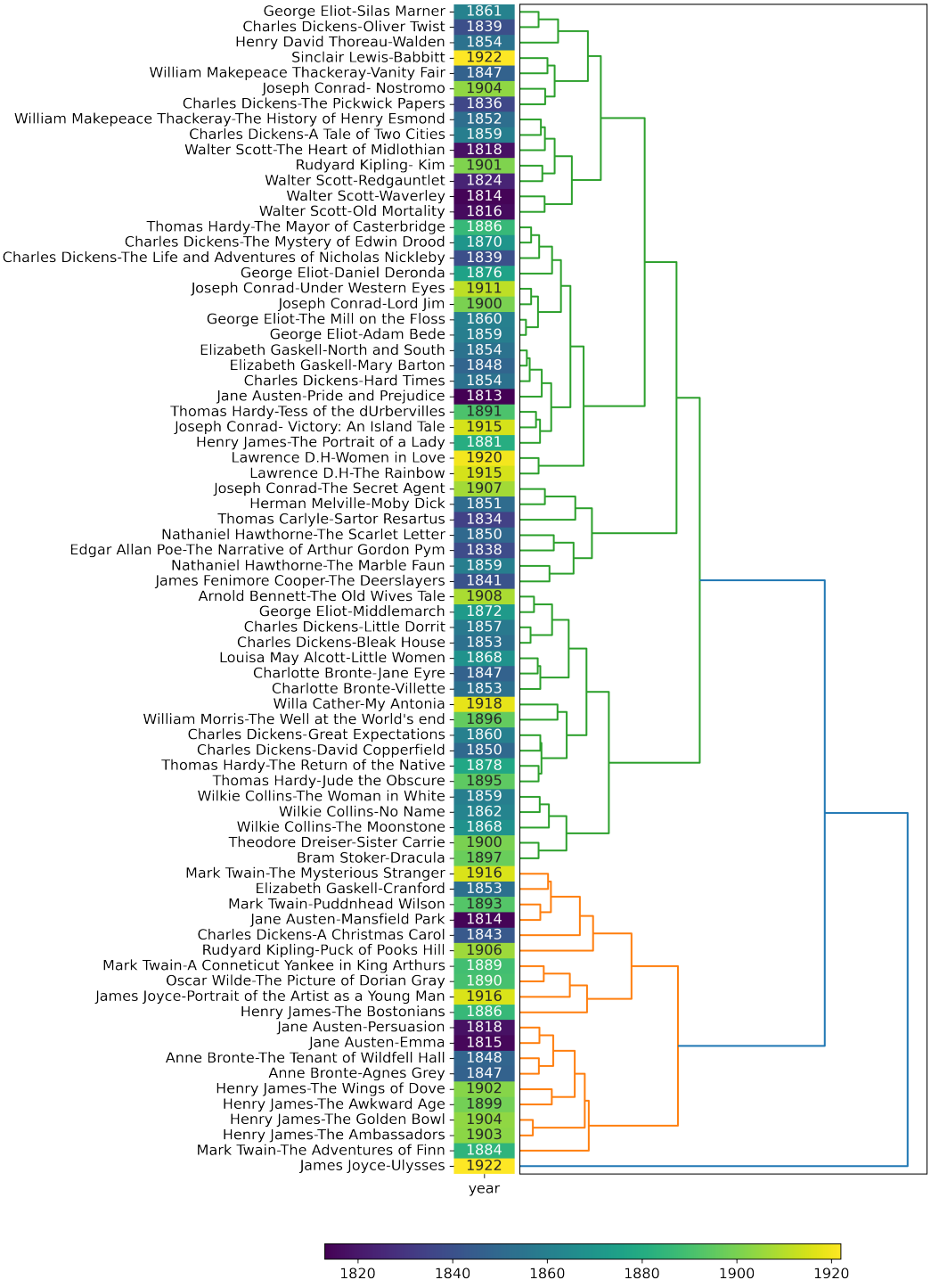

Figure 8.1: Agglomerative hierarchical clustering of canonical texts. The author's name and the title of each text is written for each data point. The heatmap shows years of publication.

various clusters.

Another 'irregularity' that we can see is the chronological orders of texts inside sub-clusters which varies greatly. Therefore, we surmised that analyzing modes and genres of canonical texts explains the structure of clusters. We had a group of literary scholars annotate the texts with various mode categories, such as romanticism, modernism, and realism, as well as genre classes, including industrial novel, Gothic novel, historical novel, satire, etc. The subjectivity of labeling texts, which can result disagreements among annotators, and the possibility of assigning multiple labels to each text are the challenges of the project.

8.2 Genre Effect

What is the effect of genres on structural design patterns of texts and are there any systematic differences between global structural properties of texts in different fictional genres? If the distribution of features varies significantly among genres, and the distribution of genres differs between two text categories, any difference observed between the text categories can be an artifact of genres. We first raised this questions when we compared preferred and non-preferred fictional texts in two time periods, contemporary texts (published after 2000) and texts from 19th and early 20th centuries (see, Chapter 5/ Mohseni et al., 2023).

We conducted an experiment to compare Approximate Entropy and Shannon Entropy of different genres. We used the US Novel corpus[1], which is a very large collection of texts from various genres. Although our results showed that the distributions of features varies among genres, no genre consistently exhibited the highest or lowest values compared to others for all text properties. Consequently, we conjectured that genre may have a modulating effect. Nevertheless, the results obtained from this experiment showed variations in structural features among genres for different text properties. This finding encourages us to further investigate whether these differences can be effectively used in modeling textual preference.

8.3 Modeling Ratings of Readers

Ratings on social media have been previously used to analyze the predictive power of lexical, syntactical and semantic features of texts in determining the likability of readers (e.g, Maharjan et al., 2017; Saba et al., 2021, ; see also Section 1.5.2). Similarly, we initiated a project to correlate global features of texts derived from

[1] https://textual-optics-lab.uchicago.edu/us_novel_corpus

variability, fractality and predictability analysis with readers' ratings on the website Goodreads. In this project, we collaborate with the Fabula-NET research team at the Aarhus University. The dataset that we use is the US Novel corpus. The mean rating and the number of comments have been extracted for each text in the corpus. Our initial findings indicate that not only certain lower-level text properties but also higher-level text properties, such as lexical diversity, can be quite effective in predicting ratings of texts (for a discussion on lower- and higher-level properties, look at Chapter 3/ Mohseni et al., 2021).

References

Maharjan, Suraj, John Arevalo, Manuel Montes, Fabio 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*. Valencia, Spain: Association for Computational Linguistics, pp. 1217–1227. DOI: 10.18653/v1/E17-1114.

Mohseni, Mahdi, Volker Gast, and Christoph Redies (2021). "Fractality and Variability in Canonical and Non-Canonical English Fiction and in Non-Fictional Texts". In: *Frontiers in Psychology* 12, p. 920. DOI: 10.3389/fpsyg.2021.599063.

Mohseni, Mahdi, Christoph Redies, and Volker Gast (2022). "Approximate Entropy in Canonical and Non-Canonical Fiction". In: *Entropy* 24.2, p. 277. DOI: 10.3390/e24020278.

– (2023). "Comparative Analysis of Preference in Contemporary and Earlier Texts Using Entropy Measures". In: *Entropy* 25.3, p. 486. DOI: 10.3390/e25030486.

Saba, Syeda Jannatus, Biddut Sarker Bijoy, Henry Gorelick, Sabir Ismail, Md Saiful Islam, and Mohammad Amin (2021). "A Study on Using Semantic Word Associations to Predict the Success of a Novel". In: *The Tenth Joint Conference on Lexical and Computational Semantics*. Bangkok, Thailand (online): Association for Computational Linguistics, pp. 38–51. DOI: 10.18653/v1/2021.starsem-1.4.