Journal of Data Mining and Digital Humanities

EPI sciences overlay journals

Fractal Sentiments and Fairy Tales
Fractal scaling of narrative arcs as predictor of the perceived quality of Andersen's fairy tales

Yuri Bizzoni[1,2], **Telma Peura**[1,2], **Mads Rosendahl Thomsen**[2], **Kristoffer L. Nielbo**[1]

[1]Center for Humanities Computing Aarhus, Aarhus University, Denmark
[2]Comparative Literature, School of Communication and Culture, Aarhus University, Denmark

Corresponding author: Yuri Bizzoni , yuri.bizzoni@cc.au.dk

Abstract

This article explores the sentiment dynamics present in narratives and their contribution to literary appreciation. More specifically, we investigate whether a certain type of sentiment development in a literary narrative correlates with its quality as perceived by a large number of readers. While we do not expect a story's sentiment arc to relate directly to readers' appreciation, we focus on its internal coherence as measured by its arc's level of fractality as a potential predictor of literary quality. To measure the arcs' fractality we use the Hurst exponent, a popular measure of fractal patterns that reflects the predictability or self-similarity of a time series. We apply this measure to the fairy tales of H.C. Andersen, using GoodReads' scores to approximate their level of appreciation. Based on our results we suggest that there might be an optimal balance between predictability and surprise in a sentiment arc's structure that contributes to the perceived quality of a narrative text.

Keywords

computational narratology; sentiment analysis; fractal analysis; literary quality assessment; stylometry

I INTRODUCTION

1.1 Computational approaches to literature and literary appreciation

Since the development of computational methods for literary analysis, it has been possible to explore the concept of literary quality and its underlying assumptions with the support of large scale measurements. The intuition that part of what readers perceive as pleasant, memorable or engaging may be found in hidden statistical patterns has given rise to a number of attempts to start modeling literary quality from a quantitative standpoint (Moretti [2013]). Although it is impossible to give an objective assessment of something as subjective as literary quality, computational tools can render complex statistical properties of texts measurable and correlate them with large numbers of individual assessments, making the analysis of literary canons and aesthetic theories more reliable, while expanding their scope beyond that of one particular text (Underwood [2019], Wilkens [2012]).

1.2 Sentiments in text

It can be argued that one of the most important aspects contributing to the appreciation of a narrative literary text is its emotional property: its ability to not only describe, but evoke sentiments in readers (Drobot [2013], Gao et al. [2016]). While defining the evocative power of a

text is extremely difficult, there are various types of resources that attempt to quantify the sentiments expressed in a narrative (for a review, see Kim and Klinger [2018]). In the last decades, sentiment analysis (SA) has been a widely investigated and developed area of research, and the possibility of roughly assessing large-scale opinions and reactions to products, policies and events has pushed the realization of a multitude of SA studies, from domain-specific, cognitively motivated analyses of small-scale corpora to the unsupervised clustering of large data to detect their predominant emotional value, to the point that the notion of sentiment analysis has sometimes come to be seen as a synonym of opinion mining (Mäntylä et al. [2018], Cambria et al. [2017]). Different types of resources often capture different aspects of the "textual sentiment" and suffer from different types of limitations. Methods based on sentiment dictionaries (Mohammad and Turney [2013]) have more recently been complemented with methods that apply statistical, feature-based machine learning (Jain et al. [2017]), and with neural machine learning approaches (Li et al. [2019]). While such approaches have the significant benefit of allowing the automatic labeling of a large amount of data, they suffer from the necessary limitations of their own training set. For example, many of these tools have been developed for specific domains, such as social media or market analyses, which might make them less reliable when scoring texts outside of their area of "expertise" (Islam et al. [2020] Xu et al. [2020]). Depending on the method, the results of an analysis and its emphasized aspects are bound to be different. Thus, it is crucial to carefully define how to measure sentiments in the context of computational literary studies. Here we focus on the valence of negative and positive sentiments expressed in words (Section 2.2).

1.3 Sentiment dynamics and literary quality

Previous explorations of the temporal dimension of literary narratives suggest that fractal analysis can be used to characterize the sentiment dynamics in a story (Gao et al. [2016]), and the fractal nature of texts has already been studied from other linguistic perspectives, implying that human-produced texts exhibit fractal patterns that are related to their perceived literary quality. Cordeiro et al. [2015] finds correlations between the level of fractality of some surface-level stylometric features and the how beautiful a text was perceived to be by a pool of readers. Mohseni et al. [2021] builds on such findings to analyse the different degrees of fractal self-similarity in canonincal and non-canonical fiction and non-fiction, again using several classic stylistic metrics, finding that they do not explain the difference between canonical and non-canonical fiction, but they do display a significant difference between fiction and non-fiction. As to the sentiments of a story, previous qualitative studies measuring their change at different time scales suggest to use the Hurst exponent, a way to capture and index narrative coherence, and suggest that an optimal Hurst value of the sentiment arc might be between 0.5 and 1 for a story (Hu et al. [2021]).

In this article, we hypothesize that it might not be the sentiments in a text *per se* that contribute to the appreciation of a story, but rather their relative change throughout the narrative. To capture these more complex properties of narrative arcs, we opt for fractal analysis (Section 3). We aim at testing this hypothesis on a dataset of fairy tales (see Section 2.1), and to compare the sentiment dynamics to readers' perception of quality, we opt for a raw but straightforward measure: GoodReads ratings (see Section 2.3), which give us scores from a wide audience of readers.

II DATA

2.1 Andersen corpus

For a pilot study, we opted for a corpus of all Hans Christian Andersen's fairy tales. Our corpus consists of a collection of 126 H.C. Andersen's fairy tales, translated into English and retrieved from the Project Gutenberg [1]. Depending on the estimate, Andersen's overall production was of 150-180 stories; the corpus covers thus the large majority of his fairy production. The tales' length varies between 1956 characters (*The Princess and the Pea*) and 106 496 (*The Ice Maiden*). We chose the Andersen corpus as an ideal case to test our hypothesis because it satisfies an important number of constraints:

1. Andersen's fairy tales tend to have fairly simple narratives, which allows us to interpret the resulting sentiment arcs with ease;
2. Andersen's tales are well known and widely read in English, which allowed us to collect quality assessments from a large number of readers;
3. Since Andersen's tales are widely read in English, we could use state of the art sentiment lexicons on the English translations of their stories;
4. Andersen's stories are short, which has the advantage of allowing us to read the whole text several times if we need to perform a close inspection of its sentiment scores;
5. Finally, but among the most important characteristics, Andersen's fairy tales are all written by one author and all fall within one genre, so we can expect a relative uniformity in style and convention; but their level of popularity can vary greatly, from several stories that are known only to a minority of readers to few fairy tales that have spun a resounding world popularity.

While we initially considered performing our experiment on the original XIX century texts, we opted for an English translation for two reasons. First of all, XIX century Danish is represented in little to no computational linguistics tools. To perform even the simplest sentiment analysis annotation on the original Andersen texts we would have had to either operate completely manually, which was unfeasible, or we would have needed to create a new specific resource. We could have done that either training a sentiment analysis classifier for XIX century Danish from scratch or approximating the results we might get from a contemporary Danish sentimental lexicon. On the other hand, contemporary English is the most resourceful language in NLP, and the highest numbers of ratings on GoodReads for Andersen's fairy tales do not refer to a Danish edition, but an English translation. Computational sentiment analysis is a highly subjective and context-sensitive task, as described in the introduction. Therefore, the arcs need to be carefully evaluated to ensure their accuracy and thus, the validity of our results. For this, we could use a corpus of Andersen stories in English annotated for affect at the sentence level Alm et al. [2005], Alm [2008]. Comparing the computed scores to human annotations allowed us to attain a more objective judgement of the results.

While choosing a reasonable English translation was a satisfying approximation for our pilot, we are aware that different translations do not represent the original text in the same manner, which means that GoodReads' scores based on a different rendering of the original are not fully accounted for. We believe that at least for this kind of fairy tales, characterized by a very simple and linear style, the problem of dealing with different translations is not paramount; nonetheless, it's possible that this variable contributed additional noise to our results.

[1] https://www.gutenberg.org/ebooks/27200

2.2 Sentiment lexicon

To create sentiment arcs, we used the NRC-VAD lexicon: a collection of more than 20 000 English words manually annotated for valence, arousal and dominance (Mohammad [2018]). For the current study we only employed valence. This is a widely used resource covering high frequency words from a number of different sources, from thesauri to tweets, and annotate through the Best-Worst Scaling method by a minimum of six different annotators per word. The annotations were crowd-sourced, and the annotators anonymously interviewed to collect data on their demographics. The lexicon has a high correlation in valence with similar resources. As we will see, the obvious limitation of this kind of resource is its inherent rigidity: annotations performed through similar lexica do not incorporate any contextual information. The valence of the word *joy* is assessed by the annotators out of any textual context, but they simply evaluate it in relation to competing words (e.g., it is more positive than *melancholy*). This means of course that any ironic or equivocal use of the word will not be captured by the lexicon: a sentence like *The joy of doing harm* will return a positive score at the mention of *joy* as high as in *The joy of doing good*. At the same time, this rigidity is the very strength of such approaches: unlike a human annotator, a sentimental lexicon does not override the positive valence of a word given its context. This allows us to observe the "solid ground" on which human readers build their interpretation of the text: *The joy of doing harm* wouldn't sound paradoxical nor sinister if it didn't contain the contrapposition between the natural positive valence of *joy* and that of *harm*. A micro-arc built on this phrase only would show an evident drop in valence at its end. The advantage of this rigidity appears in a maybe even clearer fashion when it comes to less clear-cut sentimental terminology: the myriad of words that tend to have an overall positive or negative connotation without being directly emotion words. For example, *Sun* and *sunlight* result from the Best-Worst Scaling to hold a relatively positive valence, *cancer* and *virus* a relatively negative one. As we will show through the rest of the work, the concatenation of these acontextual valences accounts for both micro- and macroscopical features of a text, since it functions as an interface between the stylistic texture and the narrative structure of the tale.

2.3 GoodReads

Finally, to measure the stories' perceived quality, we used GoodReads' [2] average ratings. GoodReads is a popular web platform used to grade, comment and recommend books. Users can provide a detailed evaluation of a text, explaining their score, but they can also simply score the text - book, poem or story - between 0 (worst) and 5 (best) stars. Like arguably any metric adopted to measure literary quality, GoodReads' scores have important disadvantages and relevant advantages. The most obvious shortcomings of using GoodReads' scores are:

1. GoodReads does not explicitly represent literary quality in its "high brow" acception: readers are not asked to value the finesse of the text, but to give it a holistic score that usually means how much they like it.
2. For this reason, GoodReads' scores necessarily conflate genres and dimensions of reading: an anthology of classical poetry and a contemporary crime novel can receive a similar score, but based on the taste of different types of readers. The same reader could give a similar score to two books based on very different considerations.
3. In the way we use the score in this study, it is a raw average: precious information like standard deviation and representativity are not included.

As a flip side of these very limitations, GoodReads scores constitute an invaluable resource to approximate the literary quality of fairy tales:

[2] https://www.goodreads.com

1. GoodReads' ratings are freely offered by users "in the wild": there is no laboratory bias nor experimental constriction. In the same way, since no definition of literary quality or text value is given, and the readers come to the platform and score the titles in complete freedom, the resource's ratings are unaffected by any theoretical or "canonical" bias: high brow literature does not receive any advantage and authors belonging to canons of any sort are not favoured in the final score (Walsh and Antoniak [2021]).
2. GoodReads' ratings derive from an amount of different readers that could be near impossible to replicate in a controlled setting: popular texts can easily be graded by tens or even hundreds of thousands of different readers, providing an invaluable source of mass annotation (Kousha et al. [2017]).
3. GoodReads' users contribute from all over the world and can belong to all age groups, genders, geographical areas. Most importantly, the raters do not belong to one specific type of reader (e.g., University students).
4. GoodReads' ratings' simplicity provides a straightforward, if artificial, handle to a problem that often becomes too complex to be studied quantitatively.

As can be seen in Figure 1, most stories in our corpus turned out to have less than 100 ratings, but the most known fairy tales reach over 40 thousand different scores. It is interesting (and relevant) to note that stories with more raters also tend to have higher average scores. As we will see later, these stories are also the ones that best fit our hypothesis.

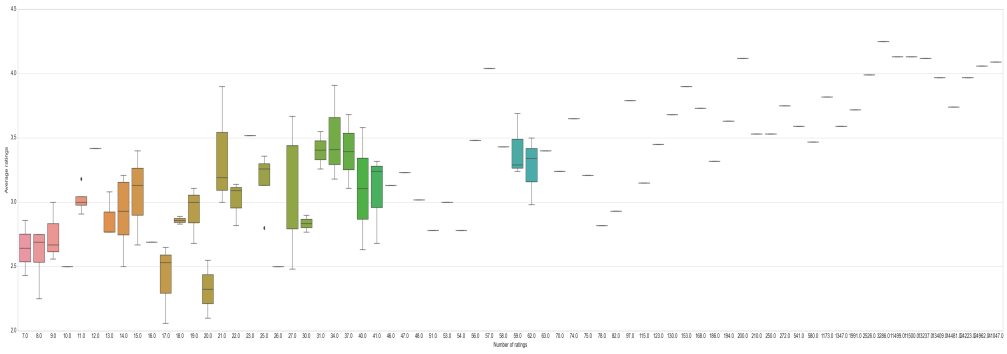

Figure 1: Number and magnitude of ratings. Most stories have less than 100 ratings, but the most known tales reach up to 40 thousand individual scores; stories with more raters tend to also receive higher average scores.

III METHODS AND ANALYSIS

Once the fairy tales and the GoodReads scores were retrieved, we proceeded the analysis in the following steps. First, sentiment arcs were computed for each text. Second, every sentiment arc was treated as a time series on which an adaptive fractal analysis was carried out to get their Hurst exponents. Third, we conducted correlation tests between the Hurst exponents and the GoodReads scores.

3.1 Sentiment arcs

To compute a sentiment arc for each story, we tokenized the corpus and retrieved the sentiment value for each word present in the VAD lexicon. On average, 31 % of the tokens in a story were found in the lexicon. Since the lexicon only contains words with non-neutral scores, we assumed that most of the missing words were close to neutral, and gave them a neutral score of 0.5.

3.2 Quality assessment

To check the quality of the computed narrative arcs, we conducted two kinds of explorations. First, we selected a subset of 12 popular stories, and clustered them in hierarchical sets (Murtagh and Legendre [2014]). These clusters were then evaluated by two of the authors, ensuring that the sentimental clusters made sense. 71 of the stories were also found in the affect annotation corpus by Alm [2008]. In the corpus, each sentence was annotated by two annotators at two aspects: for the primary emotion and the mood. The sum of these values was divided by 4 to scale the conversion between -1 and 1. As our sentiment arcs were computed at the token level, each sentence score was then repeated by the number of words in it to correct for length. The resulting human sentiment "arcs" were smoothed by taking a mean of each x scores, x corresponding to the story length divided by 30. Finally, these human-retrieved dynamics were plotted against the computed, smoothed arcs. A visual inspection was carried out to ensure that the computed dynamics followed the trends in the human judgments (see Figure 5 for an example). We found a strong positivity bias in the human-annotated scores, but the dynamics in both arcs were similar. Since we are interested in the dynamics, and not the scores as such, and since the inter-annotator agreement between the two human annotators was originally quite low, the correlation of the dynamics in these plots was found satisfactory. As to the missing words, a subset of the stories with the lowest amount of hits in the VAD lexicon were examined by one of the authors, and it was observed that most of the missing words were proper nouns, pronouns or prepositions that are expected to be neutral. At first sight, having 31 % of the lexicon can sound like a low number, but that actually means more than 1 word every 4 receives a sentiment label. For comparison, the popular sentiment annotator VADER uses a vocabulary of less than 8000 words to infer its sentence-level scores (Hutto and Gilbert [2014]). Naturally, missing out on sentimentally loaded words is a real risk, and having a more comprehensive method to annotate the texts would be desirable in the future.

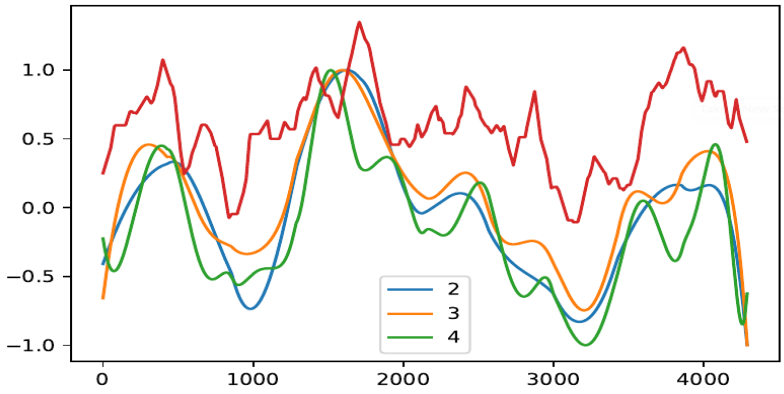

Figure 2: An example of sentiment dynamics in The Nightingale. The averaged human ratings (in red) are plotted together with the smoothed sentiment arcs at different polynomial orders m. A similar dynamic shape is exhibited both in the human-annotated and computationally retrieved sentiments, although there is a positivity bias in the human annotations.

3.3 Sentiment dynamics

Fractal Dynamics

Many complex dynamic systems are fractal, that is, they display self-similar (i.e., the system's fluctuation patterns at faster time-scales resemble fluctuation patterns at slower time scales) and scale-invariant behavior (i.e., the measurement of the fluctuation patterns does not depend on the resolution of the time scale of the measurement) (Riley et al. [2012]). Fractal analysis

therefore inspects the relationship between the measurement and it's time scale, specifically, whether this relationship is characterized by power-law scaling, see Fig. 3.

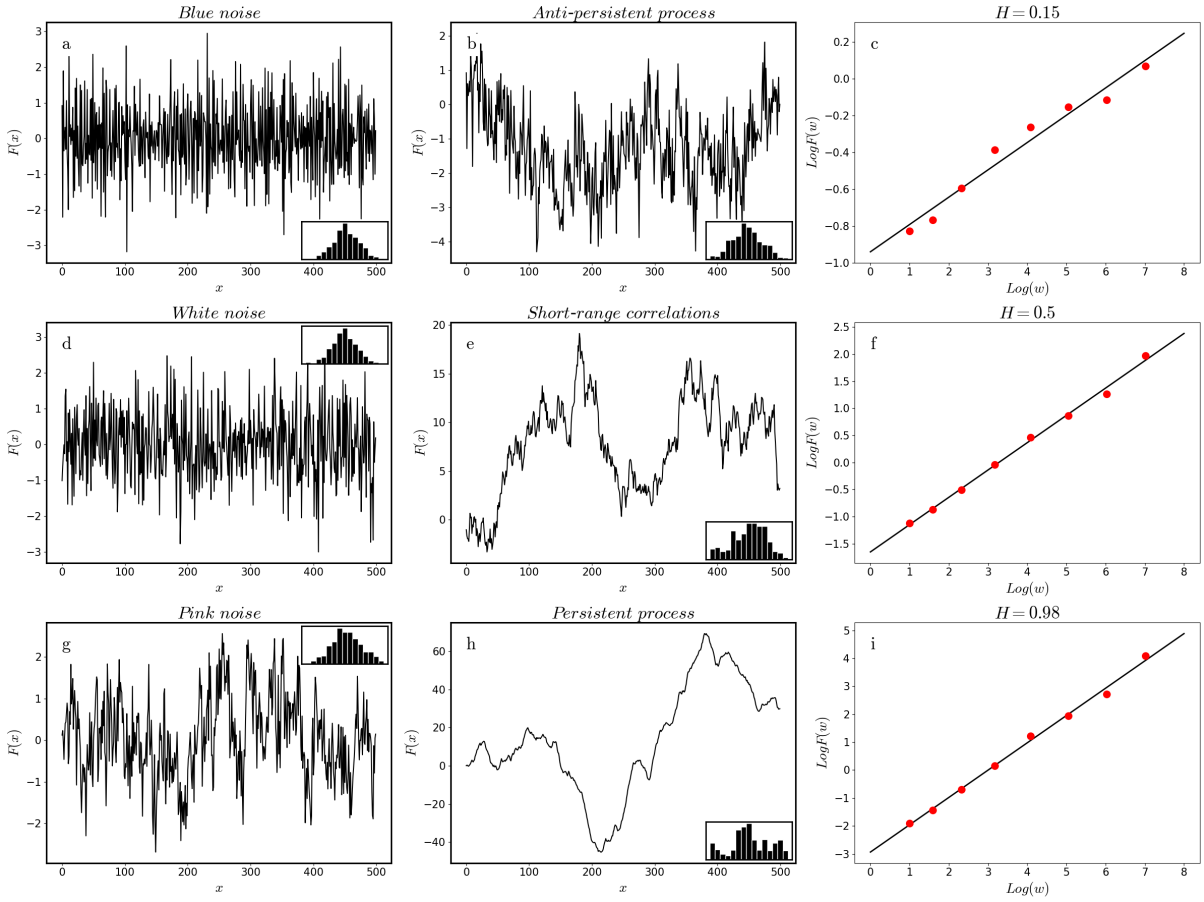

Figure 3: From upper left corner moving right, processes that exhibit anti-persistent (a-c), short memory (d-f), and persistent (g-i) behavior. Left to right shows noise-like processes (a,d,g), random walks obtained from the noise-like processes (b,e,h), and, finally, estimation of the Hurst exponent for matching process (c,f,i). Many cultural processes are noise-like (a,d,g), but in order to estimate the Hurst exponent it is necessary to transform them to random walks through first-order integration, see Eq. 2. In most cases, we are therefore not studying the process directly, but its incremental structure (b,e,h). The behavior of the process (anti-persistent, short-term, persistent) can often be observed at the initial plot (a,d,g). The anti-persistent process (a) oscillates rapidly around its average at $F(X) = 0$ and short memory process (b) contains short wave-like spikes. The persistent process (c) on the other hand contains copies of itself across multiple time-scales. The last feature is an expression of self-similarity, which becomes more apparent for the persistent random walk process (h), where rounded mountain-like structures are embedded in the process at short, medium and long time scales. The degree of self-similarity in a process can be estimated by Adaptive Fractal Analysis (c,f,i),(see appendix A, equation 5-6), where a line is fitted to the residuals $F(w)$ of the detrended process over multiple time-scales or windows w. The Hurst parameter H is estimated as the slope of the linear fit and describes how fast the overall average amplitude $F(w)$ grows with increasing window size w. For the anti-persistent process, the slope < 0.5, for the process with only short-range correlations (i.e., short memory) it is approximately 0.5, and for the persistent process the slope is > 0.5. Notice that a linear fit is not a very accurate description of the anti-persistent process (c), which at short time scales ($w < 16$) is steeper, while at longer time-scale ($w \geqslant 16$) is more flat. This is an indication of a multi-fractal process. In this example, the process still remains anti-persistent because the slope is < 0.5 at both time scales.

A $1/f$ fractal process has a power-law decaying spectral density and it can therefore not be adequately modelled by standard techniques for time series analysis (e.g., an ARIMA model or a Markov process) because they have distinctly different spectral densities. A $1/f^{2H+1}$ process where $0 < H < 1$ is a non-stationary random-walk process, the differentiation of which is a covariance stationary stochastic process with mean μ, variance σ^2, and autocorrelation function Cox [1984]:

$$r(w) = E(X_t X_{t+w})/E(X_t^2) \sim w^{2H-2}, \;\; as \;\; w \to \infty. \tag{1}$$

where w is the time lag. In this case X has a power spectral density $1/f^{2H-1}$. To adequately model a $1/f$ process, a fractional order process has to be used such as the fractional Brownian motion model (Mandelbrot [1982]). For fractal analysis it is helpful to understand the difference between fractional Brownian motion (fBm) and fractional Gaussian noise (fGn). Both types of signals are characterized by long-memory, that is, they exhibit correlations over longer time scales. But where fGn is a stationary process (i.e., its mean or variance do not change over time), fBn is non-stationary (i.e., its mean or variance show a time dependent trend) and has a power-law increasing variance (t^{2H}), and power-law decaying power spectral density ($1/f^{2H+1}$) Mandelbrot and Ness [1968], Beran [1994], Mandelbrot [1997], Kuznetsov et al. [2013]. The two types of signals are related because fBn process can be created from a fGn through integration and fGn from fBn through differentiation (Eke et al. [2002]). The Hurst H exponent quantifies persistence or memory in time series, where $0 < H < 0.5$ is an anti-persistent process, $H = 0.5$ is a short-memory process, and $0.5 < H < 1$ is a persistent process (Gao et al. [2011]). The interpretation of H, however, depends on characterizing the signal as fGn or fBn. H describes the correlation structure for fGn, while it describes the correlation structure for increments for fBm (Riley et al. [2012], Cannon et al. [1997]). For the present study of $1/f^{2H+1}$ processes should be interpreted as the latter. A persistent process indicates continuity of text complexity (i.e., entropy levels will last for a long time). An anti-persistent indicates rigidity (i.e., entropy will rapidly decay to a mean state), and finally, short memory indicates a lack of continuity (entropy will only be correlated at short time scales).

Adaptive Fractal Analysis

Detrended fluctuation analysis (DFA) (Peng et al. [1994]) is a widely used method for estimating the Hurst parameter for a time series. DFA consists of five steps, 1) initially a random walk process is constructed from the time series:

$$u(n) = \sum_{k=1}^{n} (x_k - \overline{x}), \;\; n = 1, 2, \cdots, N, \tag{2}$$

where $\overline{x}$ is the mean of the series $x(k), \;\; k = 1, 2, \cdots, N$; 2) divide the constructed random walk process into non-overlapping segments; 3) determine the local trends of each segment as the best polynomial fit; 4) compute the variance of the differences between the random walk process and the local trends; and 5) determine the average variance over all the segments. DFA may involve discontinuities at the boundaries of adjacent segments. Such discontinuities can be detrimental when the data contain trends (Hu et al. [2001]), non-stationarity (Kantelhardt et al. [2002]), or nonlinear oscillatory components (Chen et al. [2005], Hu et al. [2009]). Adaptive

fractal analysis (AFA) is an alternative to DFA that solves these problems (Gao et al. [2011]). The main advantage of AFA over DFA is that it identifies a global smooth trend, which is obtained by optimally combining local linear or polynomial fitting, and thus no longer suffers from DFA's problem of discontinuities of adjacent segments. As a result, AFA can automatically deal with arbitrary, strong nonlinear trends (Gao et al. [2011], Hu et al. [2009]).

AFA is based on a nonlinear adaptive multi-scale decomposition algorithm (Gao et al. [2011]). The first step involves partitioning an arbitrary time series under study into overlapping segments of length $w = 2n + 1$, where neighboring segments overlap by $n + 1$ points. In each segment, the time series is fitted with the best polynomial of order M, obtained by using the standard least-squares regression; the fitted polynomials in overlapped regions are then combined to yield a single global smooth trend. Denoting the fitted polynomials for the $i - th$ and $(i + 1) - th$ segments by $y^i(l_1)$ and $y^{(i+1)}(l_2)$, respectively, where $l_1, l_2 = 1, \cdots, 2n + 1$, we define the fitting for the overlapped region as

$$y^{(c)}(l) = w_1 y^{(i)}(l + n) + w_2 y^{(i+1)}(l), \quad l = 1, 2, \cdots, n + 1, \tag{3}$$

where $w_1 = \left(1 - \frac{l-1}{n}\right)$ and $w_2 = \frac{l-1}{n}$ can be written as $(1 - d_j/n)$ for $j = 1, 2$, and where d_j denotes the distances between the point and the centers of $y^{(i)}$ and $y^{(i+1)}$, respectively. Note that the weights decrease linearly with the distance between the point and the center of the segment. Such weighting ensures symmetry and effectively eliminates any discontinuities around the boundaries of neighboring segments. The global trend therefore is smooth at non-boundary points, and has right and left derivatives at the boundary (Riley et al. [2012]). The global trend can be used to suppress effects of complex nonlinear trends in fractal analysis. The parameters of each local fit are determined by maximizing the goodness of fit in each segment. The different polynomials in overlapped part of each segment are combined using Eq. 3 in order for the global fit will be the smoothest fit of the overall time series. Note that, even if $M = 1$ is selected, i.e., the local fits are linear, the global trend signal will still be nonlinear. With the above procedure, AFA can be readily described. For an arbitrary window size w, we determine, for the random walk process $u(i)$, a global trend $v(i), i = 1, 2, \cdots, N$, where N is the length of the walk. The residual of the fit, $u(i) - v(i)$, characterizes fluctuations around the global trend, and its variance yields the Hurst parameter H according to the following scaling equation:

$$F(w) = \left[\frac{1}{N}\sum_{i=1}^{N}(u(i) - v(i))^2\right]^{1/2} \sim w^H. \tag{4}$$

By computing the global fits, the residual, and the variance between original random walk process and the fitted trend for each window size w, we can plot $\log_2 F(w)$ as a function of $\log_2 w$. The presence of fractal scaling amounts to a linear relation in the plot, with the slope of the relation providing an estimate of H, c.f., Fig. 3 column three.

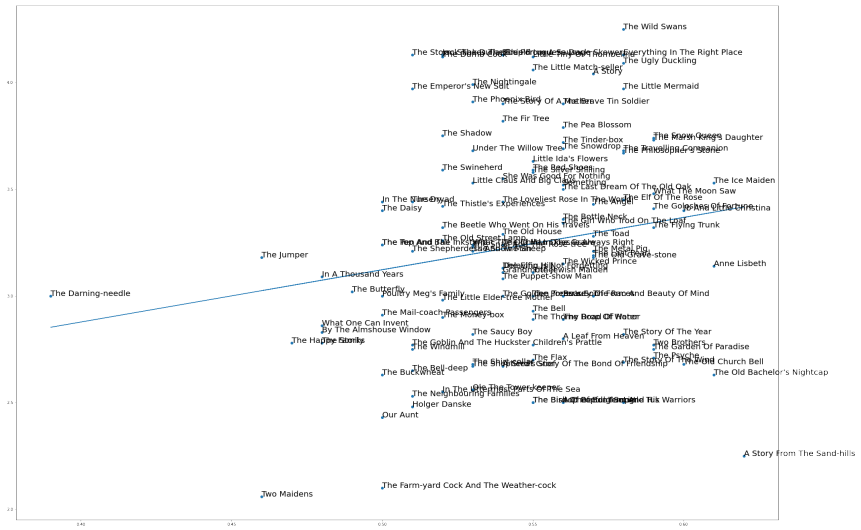

Figure 4: Correlation between Hurst exponent and average rating for all tales.

IV RESULTS

The final stage of our research consisted in checking the correlation between the fairy tales' average ratings and their average Hurst coefficients. As can be seen in Figure 4, there is a moderate correlation between the two.

An important dimension to control is the number of ratings each story received. Given the huge difference in popularity that Andersen's fairy tales have enjoyed (see Figure 1), there is a strong imbalance in how many people have read, and rated, different tales. As we discussed in the beginning, several of Andersen's stories are relatively obscure, and thus received a small number of reviews, if compared to the most famous ones. This difference in appreciation within one single author's production was one of the reasons that we found Andersen to be a good case to analyse. The main reason for checking for the number of raters of each stories is that many raters make the rating average more reliable, while stories having less than ten or fifteen raters can be quite sensitive to individual outliers, meaning that the addition of just few more scores could change it significantly. On the other hand, the more raters contribute to the final score the more that score is likely be representative and stable.

Based on the average number of ratings in our dataset, we recomputed the correlation keeping only the tales that received more than 30 different scores, which left us with 63 tales in total, excluding in this manner around half of our corpus. The threshold of 30 raters was somewhat arbitrary, and we set it in order to have a harsh threshold while still preserving a reasonable percentage of our dataset.

With this system we focused only on well-known stories that had a significant number of different annotators, and as such received more robust average scores. These stories naturally include all the best known and classic Andersen tales, such as *The Little Mermaid* and *The Ugly Duckling*, that received thousands of votes on Goodreads; and also less canonical ones that still received a relevant amount of different scores, such as *The Darning Needle*.

The result when using this subset of our data is a much stronger correlation with the stories' Hurst exponent, and a lower p-value overall. This doesn't seem to be a spurious effect of popularity: while the best liked stories tend to attract more readers, there are many tale with over 30 raters receiving overall poor ratings (*The Jumper*, *The Saucy Boy*).

In Table 1 we show a summary of the correlation values we obtained on both datasets - the one containing all of the fairy tales and the one containing only the tales rated by more than 30 readers.

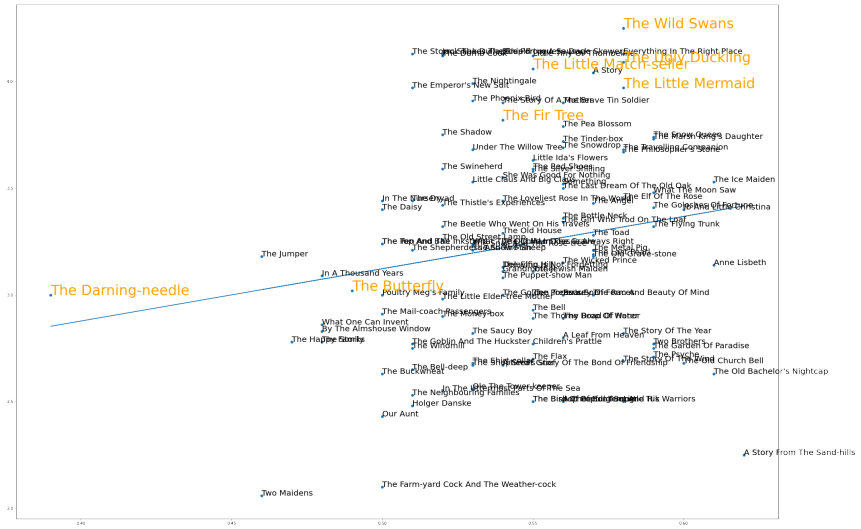

Figure 5: Correlation between Hurst exponent and average rating for tales having more than 30 ratings with some titles in evidence. Most of the very popular stories fall on the upper right corner, independently on whether they tend to be comic (*The Emperor's New Clothes*) or tragic (*The Little Match-Seller*). Tales like *The Butterfly*, based on a simple repetition of the same dynamic, give rise to a mean-reverting arc that elicits a lower Hurst exponent. These stories fall more to the lower left, receiving mediocre ratings. Also notice how without the outlier *The Darning Needle*, the correlation of the remaining data points would be even steeper.

	All tales		**Popular tales**	
	corr.	p value	corr.	p value
Pearson	.19	.03	.4	.001
Spearmann	.18	.04	.35	.005
Kendall Tau	.12	.03	.23	.009
Distance corr.	.81		.6	

Table 1: Correlations for all tales and tales with more than 30 ratings ("popular"). Ratings averaged from a more robust pool correlate more with Hurst exponents. Statistical significance is not directly applicable to standard distance correlation.

In Figure 6 we draw the overall intuition of our study: works with a smaller Hurst exponent feature mean-reverting arcs and lower overall ratings. A very low Hurst exponent represents a series where every step points in the opposite direction than the previous, so that the overall series simply reverts to its mean: after every positive score follows a negative score and so forth.

Stories based on the systematic repetition of a sentiment dynamic, such as The Butterfly, have a predictably lower Hurst and also fall among the less popular titles of the dataset, while many of the best known or canonical works of Andersen (from *The Little Match-Seller* to *The Little Mermaid*) tend to cluster to the top right, displaying more coherent patterns.

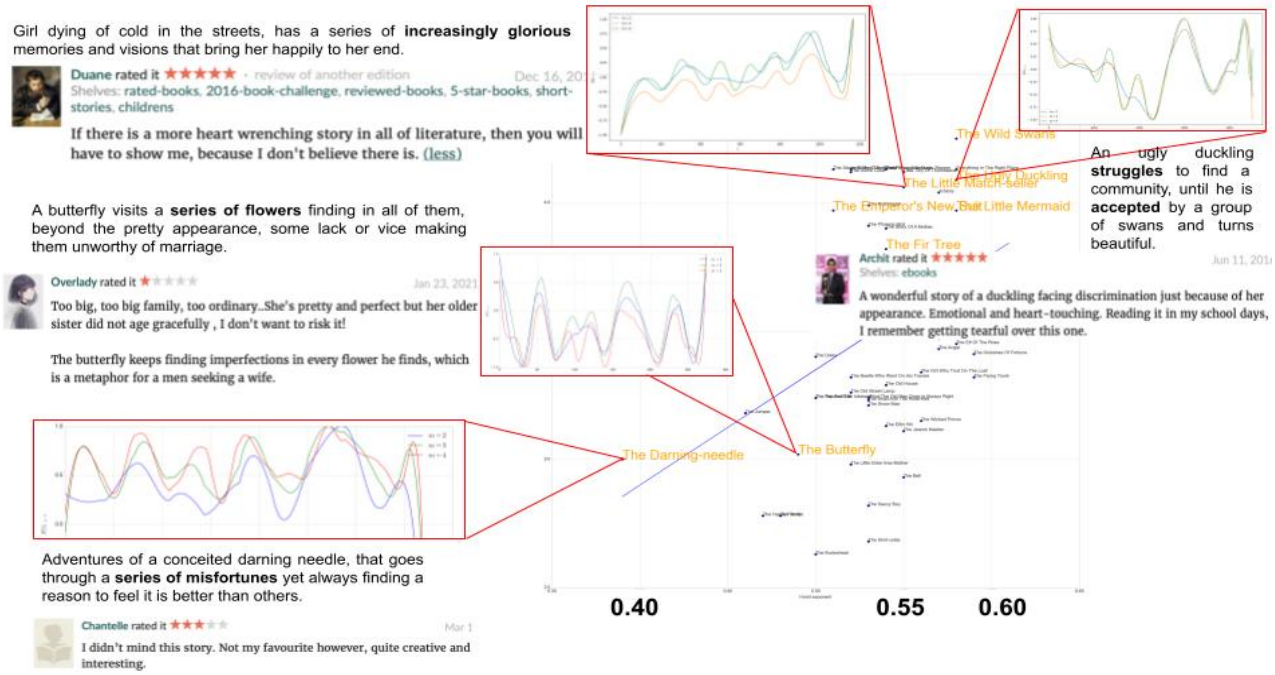

Figure 6: A visual summary of our concept. Hand-picked comments from GoodReads users are accompanied with an essential synopsis of the tale. More "zig-zag" lines reverting to the mean tend to receive less favorable overall reviews than stories having a smoother trendline. The latter ones also include many of the best known Andersen's stories. On the x-axis, the values of Hurst are evidenced: a *sweet spot* seems to lie roughly between 0.54 and 0.58

Independently from their GoodReads' score, many of the best known Andersen tales cluster on the upper right corner, with a Hurst exponent between 0.54 and 0.58.

V DISCUSSION AND FUTURE DIRECTIONS

We have found a correlation between a story's sentimental coherence and its perceived quality by a large number of readers. We find that such correlation is an interesting result and advocates for a more extensive use of multifractal theory in the study of sentimental arcs in literature. The emotional coherence of a story, as represented by the average Hurst exponent of the sentiment arc, explains a part of its perceived quality as measured by its average rating on a general audience website. Furthermore, stories with more ratings tend to show a stronger correlation with their Hurst value, which might mean that the weaker correlations we recorded for the less known stories might be due to an insufficiently large pool of raters (thus having a less robust score). It is interesting to notice that stories with many ratings also tend to have higher ratings overall: this further shows how, at least for fairy tales, fame and popularity tend to go together.

Our approach to H.C. Andersen's work also shows the potential for this approach to account for the high status of works that are mostly praised for other features. In the case of Andersen, there has been – and certainly justified – much focus on the stories' psychological depth, their appeal to both young and adults, and the fascination they have globally, in part due to the use of supernatural and enchanting elements from the popular genre of folklore fairy tales. Less

attention has been given to how Andersen compares on properties such as the cohesion and narrative appeal of writing that we have studied here. We believe this adds another dimension to the understanding of Andersen's fairy tales and we would expect that this approach could be used to show how important narrative appeal disclosed at a level not possible without a computational approach is in other authorships.

At the same time, reducing a story arc to one overall Hurst coefficient, while it proved surprisingly predictive, means losing essential information about the way coherence is distributed through the narrative arc. Exploring the variation of the Hurst coefficient at different points and at different time scales might reveal further insights into a story's sentiment dynamics and its narrative evolution.

In this study, we have been using lexicon-based sentimental arcs; while we provided the reasons for doing so, drawing arcs in this fashion has some important limitations. For example, as we discussed in the beginning, any interpretation of the sentiment value of a word in context is impossible using lexicon-based sentiment analysis. Thus, another aspect for further exploration is evaluating alternative ways of scoring the sentiments in a text. Particularly, when applied to longer stories, adopting a sentence-level approach could help accounting for the context that indeed affects the sentiment interpretation. Furthermore, instead of sentiment analysis, moving to an emotion analysis might allow for more detailed insights about an optimal narrative development.

Another line of direction for our future work is increasing the size of our corpus. Our current dataset is quite reduced and Andersen's fairy tales are short. In the near future we intend to bring our method to a much larger and broader corpus of novels and literary stories.

Finally, as we have been focusing on sentiment analysis, we have naturally discarded any other dimension of the text. Even if we take a highly intrinsic perspective on reader appreciation, it is reasonable to believe that the overall effect of a text on a population depends on the interplay of its many components - as such, we are drawn to believe that other important stylistic elements like a text's entropy and structural predictability, or its words' levels of concreteness and accessibility, would help explaining the effect of each tale on the readers. All these textual aspects, together with the overall content of the story (such as its narrative structure, its allegorical elements, and so forth) are probably constituting a significant part of the "noise" our method did not capture. In this sense, our research is rather the attempt to detect a persistent signal in a complex interplay of components: while it might make sense that readers appreciate some fractal regularity in the way that listeners and observers appreciate fractal patterns in music and images, the art of narrative functions on so many different elements that a similar effect, even if present, could easily become hard to detect. Despite all this, the present results suggest that there might exist a desirable, quantifiable ratio of coherence and unpredictability in the sentiment arcs that contribute to the appreciation of a story.

References

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

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

Jan Beran. *Statistics for Long-Memory Processes*. Chapman and Hall/CRC, New York, 1 edition, October 1994. ISBN 978-0-412-04901-9.

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

Michael J. Cannon, Donald B. Percival, David C. Caccia, Gary M. Raymond, and James B. Bassingthwaighte. Evaluating scaled windowed variance methods for estimating the Hurst coefficient of time series. *Physica A: Statistical Mechanics and its Applications*, 241(3-4):606–626, 1997.

Zhi Chen, Kun Hu, Pedro Carpena, Pedro Bernaola-Galvan, H. Eugene Stanley, and Plamen Ch. Ivanov. Effect of nonlinear filters on detrended fluctuation analysis. *Phys. Rev. E*, 71(1):011104, January 2005. doi: 10.1103/PhysRevE.71.011104. URL https://link.aps.org/doi/10.1103/PhysRevE.71.011104.

João Cordeiro, Pedro R. M. Inácio, and Diogo A. B. Fernandes. Fractal beauty in text. In Francisco Pereira, Penousal Machado, Ernesto Costa, and Amílcar Cardoso, editors, *Progress in Artificial Intelligence*, Lecture Notes in Computer Science, pages 796–802. Springer International Publishing, 2015. ISBN 978-3-319-23485-4. doi: 10.1007/978-3-319-23485-4_80.

Timothy M. Cox. Long range dependence: A review. In *Iowa State University*. Press, 1984.

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

A. Eke, P. Herman, L. Kocsis, and L. R. Kozak. Fractal characterization of complexity in temporal physiological signals. *Physiological Measurement*, 23(1):R1, 2002. URL http://stacks.iop.org/0967-3334/23/i=1/a=201.

Jianbo Gao, Jing Hu, and Wen-wen Tung. Facilitating Joint Chaos and Fractal Analysis of Biosignals through Nonlinear Adaptive Filtering. *PLoS ONE*, 6(9):e24331, September 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0024331. URL http://dx.plos.org/10.1371/journal.pone.0024331.

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

Jing Hu, Jianbo Gao, and Xingsong Wang. Multifractal analysis of sunspot time series: the effects of the 11-year cycle and Fourier truncation. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(02):P02066, February 2009. ISSN 1742-5468. doi: 10.1088/1742-5468/2009/02/P02066. URL http://stacks.iop.org/1742-5468/2009/i=02/a=P02066?key=crossref.879d2c42ec8804831202df82da8d7a1a.

Kun Hu, Plamen Ch. Ivanov, Zhi Chen, Pedro Carpena, and H. Eugene Stanley. Effect of trends on detrended fluctuation analysis. *Physical Review E*, 64(1), June 2001. ISSN 1063-651X, 1095-3787. doi: 10.1103/PhysRevE.64.011114. URL https://link.aps.org/doi/10.1103/PhysRevE.64.011114.

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

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

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

Swapnil Jain, Shrikant Malviya, Rohit Mishra, and Uma Shanker Tiwary. Sentiment analysis: An empirical comparative study of various machine learning approaches. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 112–121, Kolkata, India, December 2017. NLP Association of India. URL https://aclanthology.org/W17-7515.

Jan W. Kantelhardt, Stephan A. Zschiegner, Eva Koscielny-Bunde, Shlomo Havlin, Armin Bunde, and H. Eugene Stanley. Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A: Statistical Mechanics and its Applications*, 316(1-4):87–114, 2002.

Evgeny Kim and Roman Klinger. A survey on sentiment and emotion analysis for computational literary studies. 2018.

Kayvan Kousha, Mike Thelwall, and Mahshid Abdoli. Goodreads reviews to assess the wider impacts of books. 68(8):2004–2016, 2017. ISSN 2330-1643. doi: 10.1002/asi.23805. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23805. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23805.

Nikita Kuznetsov, Scott Bonnette, Jianbo Gao, and Michael A. Riley. Adaptive Fractal Analysis Reveals Limits to Fractal Scaling in Center of Pressure Trajectories. *Annals of Biomedical Engineering*, 41(8):1646–1660, August 2013. ISSN 0090-6964, 1573-9686. doi: 10.1007/s10439-012-0646-9. URL http://link.springer.com/10.1007/s10439-012-0646-9.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. Exploiting bert for end-to-end aspect-based sentiment analysis. *arXiv preprint arXiv:1910.00883*, 2019.

Benoit Mandelbrot. *The Fractal Geometry of Nature*. Times Books, San Francisco, updated ed. edition edition, 1982. ISBN 978-0-7167-1186-5.

Benoit B. Mandelbrot. *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk. Selecta Volume E*. Springer, New York, 1997 edition edition, September 1997. ISBN 978-0-387-98363-9.

Benoit B. Mandelbrot and John W. Van Ness. Fractional Brownian Motions, Fractional Noises and Applications. *SIAM Review*, 10(4):422–437, 1968. ISSN 00361445. URL http://www.jstor.org/stable/2027184.

Saif M. Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, 2018.

Saif M Mohammad and Peter D Turney. Nrc emotion lexicon. *National Research Council, Canada*, 2, 2013.

Mahdi Mohseni, Volker Gast, and Christoph Redies. Fractality and variability in canonical and non-canonical english fiction and in non-fictional texts. 12, 2021. ISSN 1664-1078. URL https://www.frontiersin.org/article/10.3389/fpsyg.2021.599063.

Franco Moretti. *Distant reading*. Verso Books, 2013.

Fionn Murtagh and Pierre Legendre. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of classification*, 31(3):274–295, 2014.

Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. 27:16–32, 2018. ISSN 1574-0137. doi: 10.1016/j.cosrev.2017.10.002. URL https://www.sciencedirect.com/science/article/pii/S1574013717300606.

C.-K. Peng, Sergey V. Buldyrev, Shlomo Havlin, Michael Simons, H. Eugene Stanley, and Ary L. Goldberger. Mosaic organization of DNA nucleotides. *Physical review e*, 49(2):1685, 1994.

Michael A. Riley, Scott Bonnette, Nikita Kuznetsov, Sebastian Wallot, and Jianbo Gao. A tutorial introduction to adaptive fractal analysis. *Frontiers in Physiology*, 3, 2012. ISSN 1664-042X. doi: 10.3389/fphys.2012.00371. URL http://journal.frontiersin.org/article/10.3389/fphys.2012.00371/abstract.

Ted Underwood. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, 2019. ISBN 978-0-226-61297-3. doi: 10.7208/9780226612973. URL https://www.degruyter.com/document/doi/10.7208/9780226612973/html. Publication Title: Distant Horizons.

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

Matthew Wilkens. Canons, close reading, and the evolution of method. *Debates in the digital humanities*, pages 249–58, 2012.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. Dombert: Domain-oriented language model for aspect-based sentiment analysis. *arXiv preprint arXiv:2004.13816*, 2020.